SFNCS:
A Framework for Assessment of Spatio-Temporal Visualization Methods

Kevin Tudal Allain

Department of Computer Sciences
City, University of London

This dissertation is submitted for the degree of
Doctor of Philosophy

April 2022

This thesis is dedicated to my grandmother Liliane who did not have the opportunity to attend school after her early teens.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Kevin Tudal Allain
April 2022

Acknowledgements

I would like to thank my supervisors Jason Dykes and Jo Wood, as well as my external supervisor Cagatay Turkay for their support throughout my PhD. Their passion and aim for excellency were great motivators as unexpected events continuously derailed plans.
I also wish to thank my colleagues at the giCentre, who helped me to put my situation in perspective when doubt hindered my progress.
It is clear that without the support of my friends and family I would not have been able to finish this thesis, for reasons too numerous to detail, and for this they have my eternal thanks.

Abstract

Movement analysis is complex due to many different factors: different forms of data, different levels of precision, strongly influenced by context, for which diverse sets of tasks require different visualizations and algorithmic approaches. There is a vast scope of previous work that researches, for diverse tasks, several approaches to visualization designs and data processing methods. The scope of tasks, potential visualization methods, and data processing that is yet to research is vast. To help reach a higher precision when describing contributions of researchers, we define a framework that characterizes information, from its recording into data to the way it is presented to the user and the terms used to communicate about it for evaluations. Within this thesis, we explain how our original research scope directed us from establishing the current state of the art for visualization methods for movement analysis while accounting for context into a characterization of visualizations, data processing methods, and communication approaches. This results in the framework that is the main contribution of our thesis. This thesis also presents several studies that refine our understanding of the impact of data complexity over diverse tasks, using precise terms. We also discuss how our system can be used to set up and analyze studies based on vague terms. Furthermore, we discuss the strength and weaknesses of existing designs for exploration tasks of contextually rich data movement, and potential design approaches to investigate in future work. These discussions include the tasks for which the designs could be most useful and how they fit within different characterizations of information and data.

Table of contents

Chapter 1

Introduction

The story of this thesis is one where curiosity to evaluate a visualization design encountered a succession of roadblocks. While the successions of pivots were time-consuming, they also helped us generate a unique approach to face a larger problem than first anticipated.
The work for the thesis began on October 2017 with a very vague subject: "Uncertainty". The objective with this approach was to adapt the subject of the thesis according to whatever would grab my interest the strongest. And for my very first day being officially a PhD student at City, University of London, I was in Phoenix Arizona, USA, at the IEEE VIS conference, as a spectator. While it is unconventional to send PhD students to conferences as spectators, we wish to make it as clear as possible that this approach was critical for achieving a feeling of belonging in a community and motivated us to discuss with our peers, both offline and online. The conference showed a variety of different problems and approaches to tackle them, as well as potential research directions. We decided to first investigate similarities and differences within different fields when dealing with uncertainty. Contributions presented at the conference and our own research indicated that diverse fields require to adapt to uncertainty, e.g. data retrieval from written documents and document classification [62, 85], extraction of information prone to interpretation due to language [186], historical documents with potentially contradicting sources [107], predicting and communicating hurricane tracks [50]. Following the conference and after considering several potential thesis subjects relating to uncertainty, our interest stopped on predictions and communications of hurricane tracks. Our choice was strongly influenced by being informed by experts during an IEEE VIS conference workshop that most prediction software were problematically long to compute (the number provided as an indication of time necessary to run a prediction was 3 hours). We considered that a novel and potentially fruitful approach would be to use machine learning to generate relatively imprecise predictions, for which visualizations displaying levels of uncertainty on results generated could help experts evaluate the value of predictions. We thus read literature about uncertainty, machine learning and representation of moving entities, be it natural phenomena or human driven. But an informal discussion with colleagues informed us that a similar approach had been attempted previously, but that machine learning methods proved too unreliable to be reasonably considered as tools for trajectory predictions. That information was known by

experts in the field but not published, as it is uncommon to report on failures. Fortunately, reading the literature on uncertainty and analysis of trajectories showed us a variety of other interesting subjects. The decision was thus made to pivot towards focusing on a field that carries many different types of information, each carrying different types of uncertainties, and dependent on context surrounding the information retrieval: the analysis of space-time attributes related to movement.
As visualization designs were being developed, we faced an important road block: there was no systematic approach for the evaluations of the designs we were working on. There were many studies run to evaluate design contributions, but no overarching structure to compare different approaches systematically. There were frameworks describing tasks, framework to describe visualizations, structures to characterize data, and different approaches to assess answers, but a structure linking those from start to finish had yet to be set up. The necessity and value to generate such a structure seemed very important to us and hopefully for the scientific community.
This PhD is thus motivated by an interest in visualization methods related to space-time and movement-related visualizations, and conceptual frameworks to allow their evaluations.

1.1 There, making a big detour, and back again: the motivation

The analysis of data related to movement is important for tasks that rely on understanding the movements of a variety of moving entities, e.g. aeroplanes, boats, cars, pedestrians, wildlife [95, 14, 175, 151]. Understanding the effect of factors that influence movements is a complicated task due to the necessity to consider up to 3 geographical dimensions over time, as well as evolution of other time dependent attributes. One visualization method seemed particularly promising: the time mask, developed by Andrienko *et al.* [17]. The time mask overlays time series where certain *conditions*, such as those described in a query, are matched, a condition being a status of the data queried by a user. The time mask is illustrated in Fig. 1.1. The overlay is thus indicated before any filtering of the data.

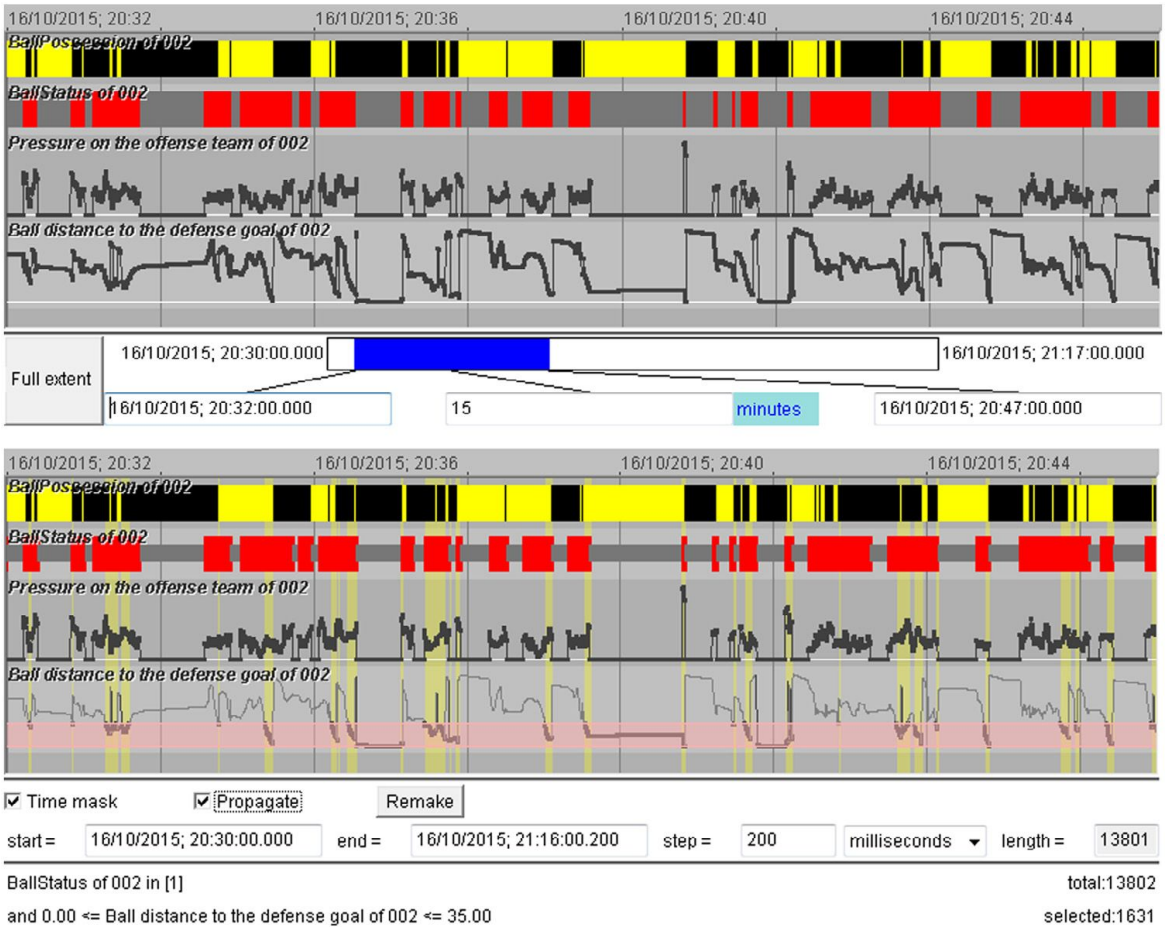


Fig. 1.1 The time mask developed by Andrienko *et al.* [17]. This picture shows a system displaying quantitative and qualitative attributes in a common design on top, and the same data with the overlay of the time mask at the bottom. Their design is set as an overlay of the quantitative and qualitative attributes with rectangles with a low opacity indicating the query entered; the red transparent rectangle over the display of the 'ball distance of the defence goal of 002' indicates a range selection, and the text at the bottom indicates that the selection also includes 'BallStatus of 002 in [1]' represented by the bright opaque red rectangles at the top. The yellow pale rectangles indicate the time frames for which the data presented is within that combination of selections.

As mentioned previously, our analysis of potential designs derived from the time mask was stopped due to a lack of a systematic structure for characterizing all the elements of the evaluation process. We argue that the lack of such structure results in difficulties to assess differences in studies. E.g. The comparison of two similar visualizations designs evaluated in different studies, with different datasets, with different tasks, can make it cumbersome to assess the value of each design for either another task or another dataset.

The influential model linking the steps of the process from information to visualization from Card *et al.* [39] quickly appeared to be the soundest base to reinforce with additional steps if further steps were to be added in order to characterize the set-up of a study to assess the validity of a novel

visualization contribution. The goal is to generate a new structure that would characterize each step of the process to easily highlight differences within studies, and thus more easily build knowledge on top of previous research. To achieve this, we developed a framework, namely the **Systematic Framework for N-scales Characterizations of Studies (SFNCS)**, which extends Card *et al.* 's [39] and characterizes the questions asked of participants, as well as the tools offered to them to answer. We later use the SFNCS as a tool to evaluate a design derived from a novel time-space-attribute visualization concept we developed, the **ATS-ATS Mask**, which broadens the time mask to an offer wider design space for enriching visualizations with overlays to indicate occurrences of data matching user-generated queries. The ATS-ATS Mask is formally defined and detailed in section 4, as well as how its concept can adapt to specific variations, e.g. in our studies described in section 5, the A-ATS Mask.

1.2 Aims and research questions

Our framework, derived from the framework of Card *et al.* [39], is itself populated by previous contributions. The details of these contributions and how they fit together within the SFNCS are detailed in section 3.3, but we need to introduce them briefly to ensure clarity of our research questions. The characterization of tasks within the SFNCS is made using the framework of Andrienko *et al.* [9]. In this framework, an important notion introduced is the difference between elementary and synoptic tasks. Elementary tasks address individual data elements, while synoptic tasks require that the data be analysed as a set. Further details about that notion are provided in section 2.1.2.
Additionally, the evaluation of our design contribution, the ATS-ATS Mask, implies investigating several aspects about it: what are its strengths and weaknesses for tasks of interest, understanding which aspects of it are impactful, and whether a concept that extends to several visualizations at once can and should be evaluated with several visualizations displayed at once.

Our thesis has thus three aims: the generation of a framework to characterize studies from task to answering process of participants, to set up designs to extend the time mask, and to assess their effect. The evaluations of the designs we created represents a small and specific selection of the theoretical range of the SFNCS, which we justify in section 5. Our research questions are thus defined for those three aims.

The first main research question our thesis aims to answer is:
Research question 1: *How can the information visualization reference model be expanded to define the evaluation process when generating a study to assess a novel contribution?*

The second main research question of our thesis is:
Research question 2: *How can the time mask be extended into a theoretical framework that enables filtering according to time, attributes, and space?*

The third main research question that our thesis, through the studies, aims to answer is:
Research question 3: *How does the visualization of conditions over time-space-attributes affect people's capabilities in conducting synoptic comparative tasks within multivariate spatio-temporal data analysis?*

By considering the variations of time-space-attributes and performance in associated analysis tasks, we generate the following more detailed research questions:

- **Research question 3.1:** *How does the visualization of conditions over time-space-attributes affect the ability to conduct synoptic comparative tasks within multivariate spatio-temporal data analysis?*
- **Research question 3.2:** *How does the visualization of conditions over time-space-attributes affect self-reported trust in conducting synoptic comparative tasks within multivariate spatio-temporal data analysis?*
- **Research question 3.3:** *How does the scaling of the axis displaying time variations affect the ability to conduct synoptic comparative tasks within multivariate temporal data analysis?*
- **Research question 3.4:** *How does the scaling of the axis displaying time variations affect self-reported trust for conducting synoptic comparative tasks within multivariate temporal data analysis?*
- **Research question 3.5:** *How does the display of additional unnecessary information affect the ability to conduct synoptic comparative tasks within multivariate spatio-temporal data analysis?*
- **Research question 3.6:** *How does the display of additional unnecessary information affect self-reported trust for conducting synoptic comparative tasks within multivariate spatio-temporal data analysis?*

1.3 Contributions

The contributions of our thesis are as follows:

- **Primary contribution: the Systematic Framework for N-scales Characterizations of Studies (SFNCS)**. The SFNCS, (pronounced "sphinx") is the main contribution from this thesis. Developed to expand upon the Card [39] model, this framework draws upon existing contributions and links them together to characterize the entire workflow, from the tasks asked of participants to the answering method.

- **Second contribution: we defined the ATS-ATS Mask, which extends the time mask to incorporate characterizations of visualizations with overlays displaying conditions based on attributes, time and space**. The letters used to characterize the ATS-ATS Mask are following the characterization of information as presented by Peuquet *et al.* [139] in their framework, where information is considered to be of three types, A for Attributes (WHAT), T for Time (WHEN), and S for Space (WHERE).
- **Third contribution: we operationalized the SFNCS to set up studies to assess the effect of the A-ATS Mask over synoptic comparative tasks**. The A-ATS Mask is a variation of the ATS-ATS Mask. These studies both demonstrate the usage of the SFNCS and generate insight about the use of A-ATS masks for synoptic comparative tasks.

1.4 Thesis scope

The scope of this thesis is defined by the range of our contributions. The SFNCS is a framework that is built with high-level elements, and developed with other frameworks that are themselves exhaustive in the sense of encompassing essential concepts relevant to movement. Thus, evaluating the scope of the SFNCS depends on the new elements we created for it, not the elements from existing frameworks used to build it. We made two additions: the characterization of questions and the characterization of the answer process.
The Question and Response blocks are defined in detail, with the justifications behind how they were built, as well as their range, in section 3. We can not claim that the Response block is exhaustive in its current version.
The Question block, as its elements are directly linked to the evaluation of tasks as defined by [15], can claim the same level of exhaustivity. This statement has to be nuanced with potential difficulties to adapt the phrasing accordingly. Furthermore, its exhaustivity is limited to information we characterize in section 3.2.4 as measurements, i.e. it can incorporate questions for tasks related to judgements, but does not integrate them in an exhaustive taxonomy. Due to the high level of abstraction of the elements we discuss, we can not claim exhaustiveness when discussing details. Details about the blocks are discussed in section 3, and limitations and future work are discussed in section 6.
Our thesis also presents studies set up to evaluate a new design, the ATS-ATS Mask, defined in section 4. The three studies are set to both illustrate the functioning of the SFNCS, and learn about the ATS-ATS Mask, but represent a small fraction out of a virtually infinite potential of combinations of blocks composing the SFNCS. Furthermore, some elements have to be ruled out from the thesis scope due to technical constraints and time:

- An important element that was not incorporated in our work is dynamics. In the scope of this thesis, discussions are limited to static visualizations, due to time and technical constraints.
- Interactions are not within the scope of this thesis, albeit short discussions of future work in section 6.

- Our work mainly discusses time as a linear set of records separated by common units such as seconds, minutes and hours, and does discuss in depth the importance of its recurring characteristics, e.g. cyclicity of days of the week.

1.5 Chapter summary

This chapter presented the origin of this thesis, its aims and the research questions it intends to answer. We discussed and introduced our contributions which will be discussed in details in further sections and assessed the scope of our contributions.
In this chapter, we mention a need for the framework we developed, the SFNCS, that rose as we were working on extending a particularly interesting concept: the time mask from Andrienko *et al.* [17]. Following a literature review in chapter 2 which presents the work that influenced our thinking, we present the SFNCS in chapter 3.
The extension of the time mask we developed, the ATS-ATS Mask, is then discussed in chapter 4. We then present, in chapter 5, a series of studies evaluating the ATS-ATS Mask set up using the SFNCS. We conclude this thesis by considering potential future work and the desired consequences of our contributions in chapter 6.

Chapter 2

Background and Related Work

There is a large amount of work done towards expanding the set of existing designs to represent data related to movement, be it trajectories or attributes that are related to it by time and space. In this section, we discuss the concepts related to movement and space-time information and different frameworks to characterize tasks performed when analysing data related to movement. We also present a representative sample of visualization methods, and discuss different frameworks and systems that exist to characterize data related to movement. The objective of this section is not to be exhaustive, but instead to present a sample large and diverse enough that indicates the steps towards the choices for designs, framework structure and studies set up, presented in the following chapters.

2.1 Concepts and theories for movement and space-time attributes

2.1.1 The conceptualization of information related to movement

The communication of information related to movement necessitates clarity over what is discussed. Within this thesis, when discussing moving entities, we refer to objects moving over space with a single position at each time unit for which its position was recorded. The moving entities discussed in this thesis are sized at 'human scale', e.g. natural phenomena, animal, urban, naval and aerial

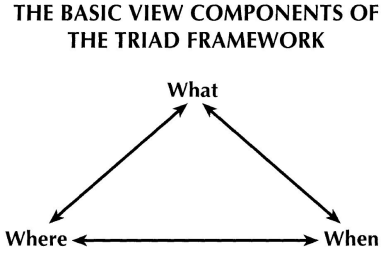


Fig. 2.1 The basic components of the Triad framework from Peuquet *et al.* [139]. The WHAT-WHERE-WHEN basic blocks represent high level concepts from which precise information can be defined.

mover. We do not discuss in this thesis movement at a very small scale, e.g. molecular scale, nor very large scale, e.g. movement of celestial objects. In this section we discuss how the information related to movement can be organized, characterized and communicated to users that wish to analyse that information and gain insight from it. We use the same definition as Andrienko *et al.* [16] when discussing movement data. It consists of position records, with an object identifier, a time of recording (possibly future in cases of predictive software) and spatial coordinates. We acknowledge, just like them, that real movement is continuous, and that records of movement are discrete and thus inherently force dealing with incomplete information. We also argue that if the quality of the data is sufficient, i.e. has a level of precision that is high enough, for the task that the user is attempting to perform, it is acceptable to discuss the recorded information as if it is the information of the movement itself. Furthermore, presenting information close to reality can be detrimental to some tasks, and thus considerations about desired level of realism are important in some contexts. An interesting example is the work of Hurter *et al.* [78] where airline trajectories are dynamically bundled or presented in detail according to user's needs.

Before displaying data to the user, information has to be collected, transformed into data, cleaned, and according to the needs transformed and displayed into visualizations. For an effective discussion when considering contributions for these steps, it is thus necessary to consider the origin of data, its transformation and representation in a structure to allow for comparison of contributions provided by researchers. One influential contribution is that of Peuquet [139] - namely the WHAT-WHERE-WHEN conceptual framework, illustrated in Fig. 2.1. This approach is the high-level base for most conceptual models that discuss categorization of information related to movement.

Another influential contribution for our work was the framework developed by Andrienko *et al.* [9], which characterized visualizations according to dimensions, thematic attributes, trajectories, time, space. This approach is interesting as it does not limit the diversity of the designs that can be produced, but can be used to describe any visualization, and convey the intent on how to set up the organization of information to convey. Their framework is a systematic and comprehensive way to indicate the possible types of information that can be extracted from analysis of data related to movement.

Following the same categorization of information, Bogorny *et al.* [29] developed CONSTanT, a conceptual data model for semantic trajectories, considering the event-based approach to movement analysis as a UML structure. This contribution is an interesting example, showing that the framework of Andrienko *et al.* [9] can be directed translated into a computational structure. In practice, while none of these contributions directly produce a graph that links them directly, this contribution is a valuable example of a very low level of abstraction for which each element can be additionally characterized with a higher level of abstraction with ease. It's worth noting that the model of Bogorny *et al.* doesn't discuss aggregations for the information stored in their model, they instead leave the choice to the designers on their approach to that question. Doing so simplifies their model. Particularly, according to interpretation, information can be characterized differently. The most flagrant example being spatial information, which can be characterized as qualitative information, part of the WHAT

component from Peuquet Triad. Information can also be grouped, clustered, changing their status from quantitative to qualitative, e.g. speed records ranging from 0 and 10 kilometers being labeled as slow. The When category, which can correspond to single moments or time ranges, can also be considered in various ways. Time can be considered as a numerical value, if one is to measure the time passed from a certain measurement, e.g. in programming, time is often measured as the number of milliseconds passed since January 01, 1970. It is also characterized according to its repetitive nature, e.g. day of the week, or due to social constructs (albeit built upon physical measurements to justify them), e.g. hour of the day, month of the year. Brehmer *et al.* [31] discuss how the variability of approaches to consider time can impact its scale, through data transformation, resulting in the following potential scales:

- Chronological: approach to time that considers it each record within a global frame, the common calendar based on year, month, day, minute, second, and lower levels of management according to precision levels
- Relative: approach to time that considers events according to their difference to a certain time selected according to relevance to the task. That approach does not require fitting a commonly used attribute for time, i.e. 1 in a relative scale could mean any range of time frames, e.g. 1 could represent 34.52 seconds.
- Logarithmic: approach to time that is similar to the relative one, but distorts with a logarithmic scale the chronology. The main point of this approach is to consider and display appropriately data with a skewed distribution, particularly when the distribution is poorly balanced.
- Sequential: approach to time that is solely based on the information that a recording occurs prior or after another, without any indication as to time separating them.
- Sequential + Interim Duration: approach that is hybrid between chronological and relative, particularly useful when details of chronological information want to be preserved for some section of the records, but allows skips.

Considering visualizations, there also exists a large corpus of literature discussing definitions of the elements that constitute visualizations and characterize them. But while naming the most basic elements that constitute a visualization with geometric vocabulary is relatively simple, misunderstandings can occur due to different interpretations of the same word. Thus, work presenting definitions for the basic geometries that compose visualizations have been a necessary contribution for clarity. Bertin [117] introduced the following visual variables: position, size, shape, value, colour, orientation and texture. These variables constituted the foundation of characterization of visual variables. They underwent amendments over time (e.g. colour got split into attributes to characterize it: hue, saturation). Visualizations can then be defined through basic visual variables for the three components of the Triad framework [115, 116]. While using basic visual variables to characterize graphs generated is possible, we would not recommend it, since the visual details are too specific to

allow for a clear categorization of the visualizations used to represent the data.
We thus consider that categorizations of visualizations should be done, not with the basic visual elements that compose them, but rather their composition. The question comes twofold: what data is presented in the visualization, and how. The approach followed by Andrienko *et al.* [9] in their framework is very interesting as it allows to easily link visual representation to conceptual representations of movement and attributes. Their approach allows using letters to indicate the type of information displayed in a graph, e.g. S for spatial positions, O for objects, A for thematic attributes, ΔS for displacements. Following that approach, each letter links to a visual representation, e.g. S to positions on the map, O to geometries on the map, A to retinal properties. It is then possible to generate pictographs to indicate organization of the graphs. We illustrate that approach with one of their tables indicating transitions from elements to portray to visualization techniques descriptions to pictographs, in Fig. 2.2. Their approach does not allow to directly indicate whether the data is detailed or aggregated, and thus this notion has to be discussed separately.

Following the same characterization of information related to movement, Andrienko *et al.* [12] developed a conceptual model that considers movement as a combination of spatial events of diverse types and extents in space in time.

Brehmer *et al.* [31] also discuss layout and representation of time. Their characterization is thus interesting as it allows considering low level characteristics of the representation of time, when discussing the representation, being linear, radial, in a grid, in a spiral, or arbitrary, as well as how it is set around the visualization, when discussing its layout. They identify 4 layouts:

- Unified: approach to display time with a single timeline.
- Faceted: approach to partition time according to a categorical attribute, for which the objective is to compare information over time over the time frame looked at.
- Segmented:approach to cut the time frame into several smaller ones, according to meaningful choices to compare the information displayed at different time, e.g. information of a year displayed over several months.
- Faceted + Segmented: approach that is a mix of the faceted and segmented, which is relevant to consider evolution of different attributes together, e.g. investigation of the potential impact of one attribute over another.

Space, not unlike time, is also a form of information that can be interpreted in several ways. Space relates to information that is recorded over a geographical area. Areas carry contextual information that can be relevant for certain tasks, and thus allowing its consideration is important for a variety of tasks. Contextual information of space can include multi-layered attributes, e.g. a geographical being part of a street, which is part of a district, which is part of a town, which is part of a department, which is part of the town. That information can thus be used as an approach to aggregate spatial information for various representations, and similarly can be used to compare information over time, e.g. compare evolution of an attribute over time according to location aggregated into an attribute.

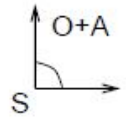
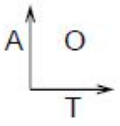
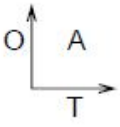
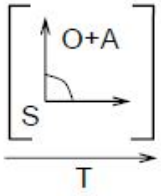
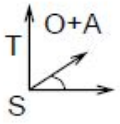
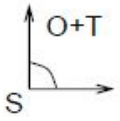
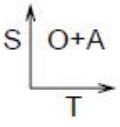
What is portrayed	Visualization technique description	Pictograph
Spatial positions of objects (+ thematic attributes)	Map: spatial positions (S) → positions on the map; objects (O) → geometries on the map; thematic attributes (A) → retinal properties	
Temporal positions of objects (+ thematic attributes)	Time graph: temporal positions (T) → positions on the time axis; thematic attributes (A) → positions on the attribute axis; objects (O) → points or lines	
Temporal positions of objects (+ thematic attributes)	Temporal bar chart: objects (O) → positions on the object axis; temporal positions (T) → positions on the time axis; thematic attributes (A) → retinal properties	
Spatial and temporal positions of objects (+ thematic attributes)	Map sequence: temporal positions (T) → positions in the sequence; spatial positions (S) → positions on the map; objects (O) → geometries on the map; thematic attributes (A) → retinal properties	
Spatial and temporal positions of objects (+ thematic attributes)	Space-time cube: spatio-temporal positions (S+T) → 3D positions in the cube; objects (O) → geometries in the cube (e.g. lines); thematic attributes (A) → retinal properties	
Spatial and temporal positions of objects	Map: spatial positions (S) → positions on the map; objects (O) → geometries on the map; temporal positions (T) → retinal properties **(1)**	
Spatial and temporal positions of objects	Space-time graph: spatial positions (S) → positions on the space axis; temporal positions (T) → positions on the time axis; objects (O) → points or bars; consecutive positions may be connected by lines; thematic attributes (A) → retinal properties **(2) (3)**	

Fig. 2.2 One of the tables of Andrienko *et al.* [9] illustrating the characterization of visual stimuli, considering information to display in the first column, the type of visualization technique that can be used in the second column, illustrated with a pictograph in the third column.

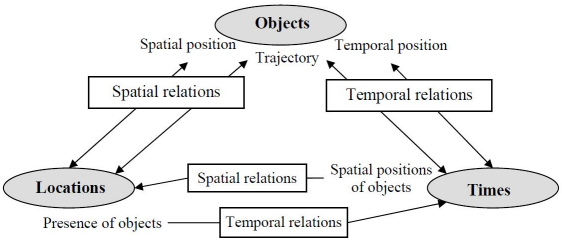


Fig. 2.3 The connections between the basic components of the Triad framework from Peuquet *et al.* [139] as defined by Andrienko *et al.* [9].

Andrienko *et al.* [16] characterize spatial information with the following attributes:

- Spatial resolution: the minimal change of position that can be reported.
- Spatial precision: difference from the position recorded and the real world position that is systematic, e.g. phone position reported by a router recording it.
- Positioning accuracy: error in the measurement reported.
- Spatial coverage: range of potential positions and its uniformity.

These attributes are relevant for information about any entity, whether the entity is able to move or not. According to the scale selected for time, what is considered a moving entity can vary, e.g. some coasts lose part of their land over the water level over time, height of mountains vary over years, glaciers move slowly due to gravity and change shape slowly due to long term changes in climate. The positions recorded in measurements are captured according to the method used to define the precise position recorded, such as latitude and longitude, ECEF also known as earth-centered earth-fixed [185], the reference ellipsoid [72] and then displayed according to map projections [86]. The scope of our thesis is more focused on the analysis of data related to movement and space-time attributes. Andrienko *et al.* [21] define episodic movement data as position measurements for which the positions between records may not be reliably reconstructed due to the large time difference between the measurements. In this thesis, while we acknowledge their importance in certain contexts, we will not discuss those elements of formatting and communicating mapping of that information further than the most common cases. Our discussions about context relating to spatial information is thus independent on measurement approach as long as it is possible to separate an episodic movement from the rest, and with a level of precision that is high enough to consider that the tasks of interest can be performed with their sufficient level of accuracy.

Thus, models that discuss relationships between time and space are more interesting for the scope of our thesis. In their framework, Andrienko *et al.* [9] extend the Triad framework of Peuquet *et al.* [139] defined by the WHAT-WHERE-WHEN components by considering relationships between these elements, illustrated in Fig. 2.3. The WHAT-WHERE-WHEN blocks have been strongly

influential as these are the highest level of abstraction and allow discussing information related to movement and space-time attributes. As discussed previously, information transformation can influence the approach to analyse data.

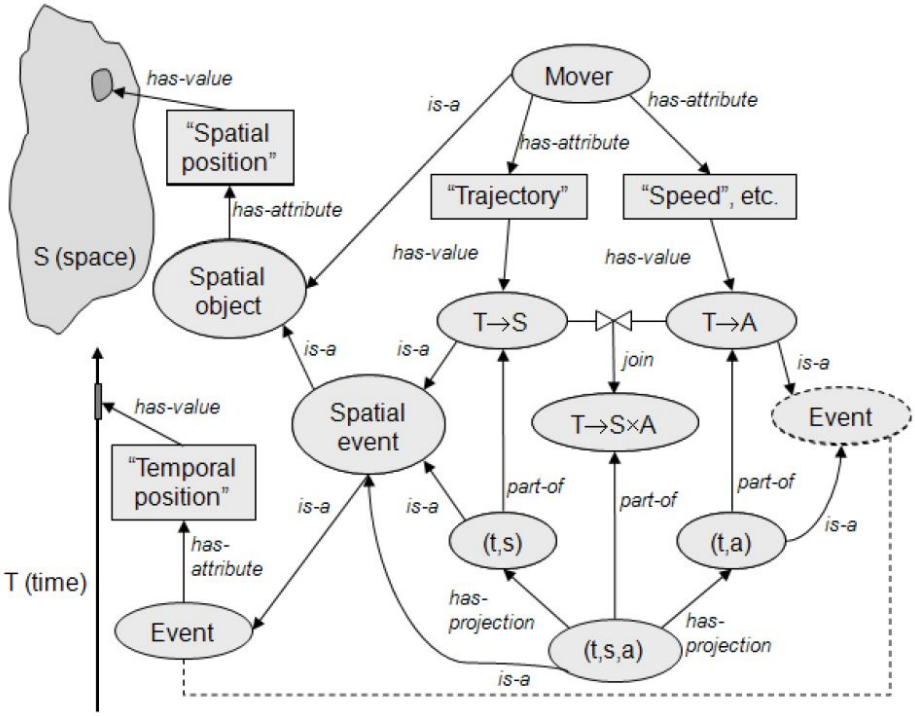


Fig. 2.4 Movement as a composition of spatial events, presented by Andrienko *et al.* [12]. The graph presents how movement is a set of records of "temporal positions" and "spatial positions" for a "mover".

Andrienko *et al.* [12] discuss events, and the importance of models that can link types of objects to superior concepts to ensure communication on lower levels of abstraction can be connected to high level levels of abstractions, e.g. a building is an object, with a spatial position, that is unlikely to vary, as opposed to an animal, which has a position that is likely to vary. They are both spatial objects. But the animal, due to its ability to move, can be the source of an event, whereas a building can only be a subject of an event. They then define movement as a composition of spatial events, as illustrated in Fig. 2.4

Following that high level approach, further work has been included to discuss information related to space, time and movement altogether into lower-level conceptual models.

Santana *et al.* [148] present a Travel History Conceptual Model, which is based on the connections between the following blocks: place, trail, social interaction, traveller, travel history, stay and visit. Their model is presented in an UML graph that allows to numerise expected relations and ownership between elements, and to see what information is shared between these blocks. Santipantakis *et al.* [149] present an ontology for the representation of semantic trajectories at varying levels of

spatio-temporal analysis, for which trajectories can be considered as a sequence of positions of moving objects or aggregated from their raw data, to characterize the activity over the region used for the aggregation. Their approach is interesting for their task of interest, as they wish to analyse the evolution of airspace according to time according to sector configurations defined by flight information regions.
Spatial regions can be transformed into categorical information that can then be displayed in structures that fully disregard their geographical properties, such as lists, or partially, such as in treemaps [16, 159, 160] or as categories for Gantt charts [68].

The work of Zeng et a. [182] is an interesting implementation of a visualization incorporating hybrid aggregation approaches of space-time information. This kind of approach is indicative of the value of considering connections between different levels of information aggregation to result in a contextually rich representation of information.

All of these contributions tend towards a similar understanding and conceptualization of time, space and information attributes. But that understanding has to be nuanced, as there is no consensus within the scientific community as to the appropriate level of abstraction with which to display time-space-attribute information according to any particular task of interest. Furthermore, communications between the level of abstractions should be organized, and there is a clear trend towards developing models that allow multilevelled consideration when making a contribution. As future discussions will continue on how to connect various levels of abstractions, we consider that the safest approach for our framework to remain valid, useful and flexible over time is to set it with a high level of abstraction. If future work establishes clear links between abstraction levels, an updated version of our framework will be possible, but the main blocks presented here will be preserved. We further discuss this notion about the SFNCS in section 3.3.

An influential contribution that has inspired this work is the reference model developed by Card [39] for the usage of visualization to perform tasks using data. Card's model, illustrated in Fig. 2.5, connects the following elements:

- Data: The data that represents the information necessary to perform the task of interest. This element contains two children:
  - Raw Data: data recorded but not adapted for the task performed.
  - Data Tables: data in a format that is fit to perform the task.
- Visual Form: the visual elements displayed, translating information transcribed in the data and presented in a manner that helps in their analysis. It is composed of two children:
  - Visual Structures: spatial substrates, marks and graphical properties set an ensemble of visual elements that are mapped on a display. The visual information is arranged in an effort to ensure a clear and accurate representation of the data.
  - Views: graphical parameters that transform the visual structure according to the physical display used to draw the visualization.

- Human Interaction: modifications of the previous elements made consciously to facilitate performing the task. It is composed of three children:
  - Data Transformations: modifications of the data into a structure more relevant for performing the task of interest.
  - Visual Mappings: selection of the structure of the visual elements to display, i.e. how data is changed to drawn elements.
  - View Transformations: adaptations of the drawings for readability.
- Task: the function for which the visualization is produced.

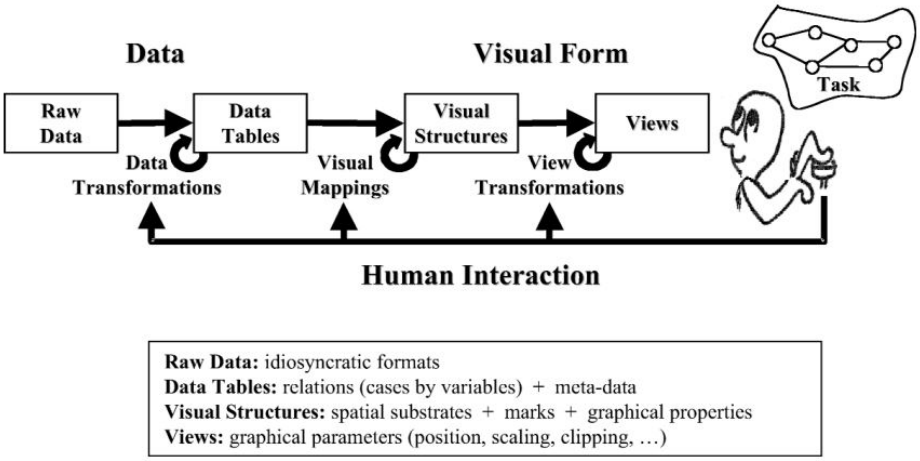


Fig. 2.5 The reference model of Card [39]. In this loop the Raw Data is transformed into Data Tables, which is then mapped to Visual Structures, and transformed into Views presented to the user aiming to perform a Task. Each of these steps can be influenced by Human Interaction.

This model is particularly important as it underlines the necessity to consider the final usage of a visualization as a number of steps for which each decision influences the following one. Note that the elements here do not indicate how each step is done, and elements are not linked together. This approach does not facilitate comparisons between visualizations. Still, this contribution was very influential and helps to consider how novel contributions fit compared to previous work.
This influential work resulted in alternative approaches, such as the *Prefuse* framework of Heer *et al.* [71], illustrated in Fig. 2.6, which presents a first approach to indicate how to populate the elements of the reference model of Card [39]. Using their framework, they developed an user interface for designing interactive visualizations. Their approach, enables the user to interact with every step between data, visual form, and view, is fundamental to various interactive visualization systems including commercial products such as Tableau [173]. Furthermore, commercial software such as Tableau that are developed with a specific focus for flexibility and accessibility offer a variety of approaches to transform data, visual forms and views.

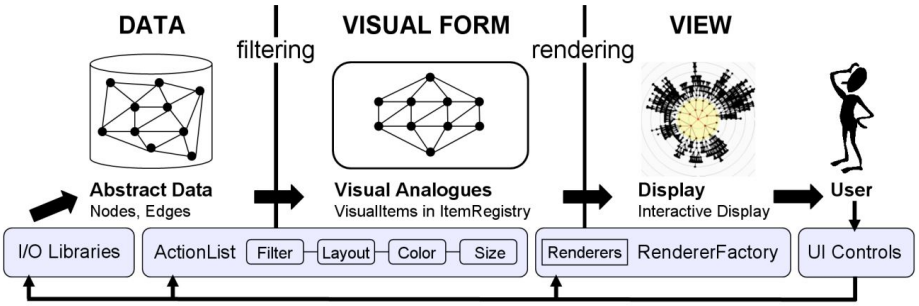


Fig. 2.6 The prefuse visualization framework developed by Heer *et al.* [71]. It presents a list of composable actions to transform the data into a view the participant can use to perform their task of interest. With UI Controls, the user can modify elements from the Data, Visual Form or View.

These contributions show that by linking into one structure the elements ranging from data collected to the visualization presented to the user, the clarity of contributions is reinforced. These contributions illustrate how it is possible to set software systems in which various combinations of the same data can be done. These contributions are valuable, but do not provide an answer as to which visualizations are best suited to particular tasks. In the hypothetical scenario where a software system could produce all variations of visualizations for a certain data set, the user wishing to perform their task of interest would struggle to find which visualization is most fitting. Software designed to be flexible and user-friendly such as Tableau suggests diverse common visualizations according to the data input, and leaves the user to choose among them. This approach has two issues: in the limited number of visualizations presented, there might not be the specific visualization that fits best the task, and the ability of the user to assess which visualization is the fittest is not guaranteed. These facts imply the need to compare efficiency of visualization methods according to tasks. We discuss our approach to formally define elements composing our framework, the SFNCS, derived from the reference model of Card [39], in section 3.3.

2.1.2 The conceptual representations of tasks

There is a wide variety of tasks that are performed using visualizations methods. Illustrative examples for trajectory analysis include : discover peaks of activity within transport network [178], find moving entities' connections, e.g. flocking, moving entities avoiding each other [119], discover attributes of moving entities for a certain area, such as finding if a zone of a town is mainly visited by tourists or working people [131], establish where the future position of a hurricane will be, and with how much certainty [125], find causes of deviations from the optimal or expected path for moving entities, and how likely a moving entity following a future similar trajectory will undergo the same deviations [18].

Tasks can be either very clearly defined and simple to perform, such as finding a value at a certain time, or much more complex, vague and difficult to perform, such as establishing if one object has a meaningful influence over another. We can classify these tasks to consider notions such as scope, object, objective, clarity. Furthermore, there is no method to know with certainty the mental steps

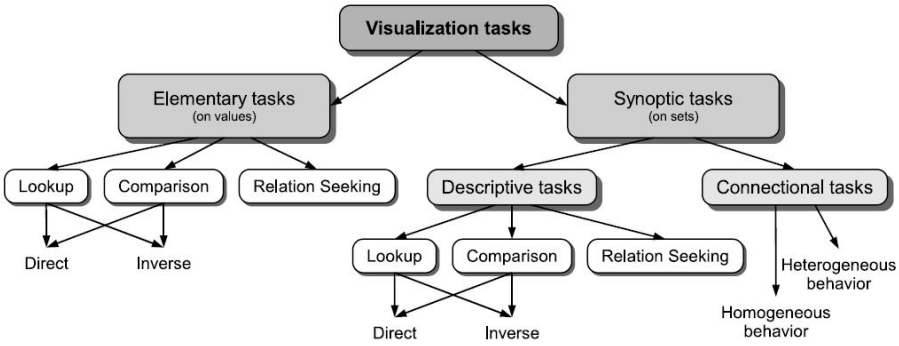


Fig. 2.7 The taxonomy of visualization tasks produced by Andrienko *et al.* [15] as modelled by Aigner *et al.* [3]. The separation of tasks into these categories can be used to compare studies evaluating designs.

taken to solve a task. There are methods to help to understand mental steps to perform tasks, such as records being made of actions when interacting with a visualization system, questions can be asked in studies with a qualitative approach, or automation of computations to reduce necessity to perform simple tasks which are hypothesized to be necessary for the aimed complicated task.

Discussions about tasks are intimately related to notions of levels of abstractions. The more vague the information that's necessary to perform a task, the higher the level of abstraction, and the greater the need to involve humans' abilities to assess information according to overall context that goes beyond the presented data.

Often, tasks are loosely labelled as elementary or high-level. That separation is closer to a spectrum than a strongly separated split. But without clear separations between levels of abstraction of tasks, interpretation is likely to influence how two researchers are going to characterize a task.

Keim *et al.* [89] consider that the following three high-level tasks encompass all the potential objectives that can be aimed for with data visualization:

- Visual Presentation: tasks that aim to communicate certain information derived from the data. The information is known prior to the designing of the visualization and is not designed to generate new insight but to expand the population that possess that knowledge, e.g. teaching material [140].

- Visual Exploration: tasks for which the objective is to generate new insight from the data. The nature of the insight searched for may not be known. These tasks are bound by the notion that recorded information can lead to new discoveries. As the exact connections between recorded and clear data to potential new insights are yet to be discovered, these tasks are most often performed thanks to visualization systems that present a certain flexibility, to allow the exploration of multiple hypotheses.

- Visual Analysis: tasks for which the hypotheses are well-defined and require verification, rather than design to potentially surprise the user. Visualizations produced for such tasks aim to either confirm or reject hypotheses previously defined.

This approach is interesting but does not consider how tasks can connect each other, i.e. if a hypothesis is rejected, others may arise as substitute, necessitating other verifications, but it can also indicate the need to consider exploration of the information available to generate new insight which itself could later on be responsible for new hypotheses to evaluate.

Brehmer *et al.* [33] present a less abstract multi-level typology that connects together the task that a visualization being designed for is supposed to support, how that support is reached, and what is the input and output (if applicable) of the visualization. Their typology is illustrated in Fig. 2.8.

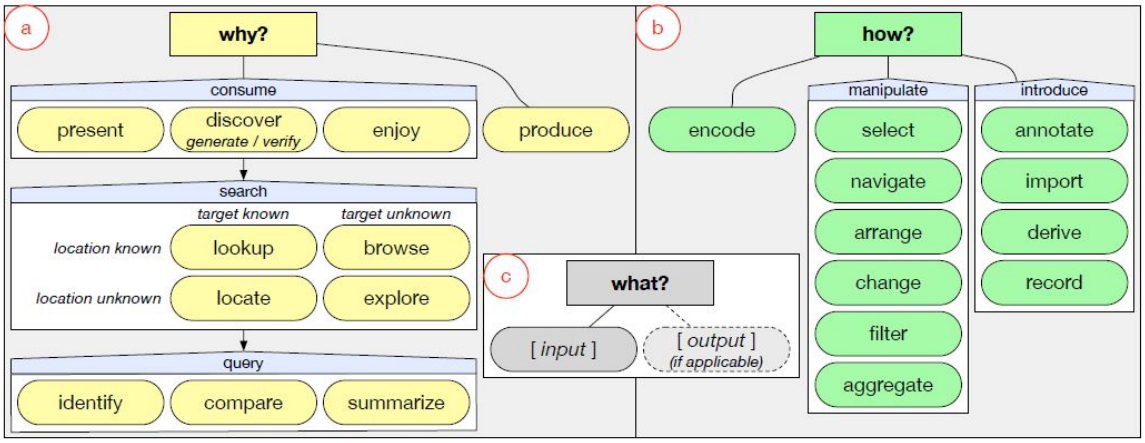


Fig. 2.8 The multi-level typology developed by Brehmer *et al.* [33]. It conveys *why* visualizations are designed, in alignment with the Keim *et al.* [89] model. It also considers *how* the tool supports the task performed, and *what* the task inputs and outputs are. The tasks are cut into three levels of abstraction, with consume (and produce) being high-level, search being mid-level, and query being low-level.

It is not always obvious how to make a distinction between the levels of abstractions. Low-level tasks are specific and precise, in opposition to high-level tasks which are broader. The low-level tasks present little interest on their own, as they could be simply computed, but are considered as necessary in order to complete high-level tasks. A notable difference between tasks with low levels of abstraction and high levels of abstraction is the approach to evaluate participants performing them. The first are often evaluated through quantitative studies and the latter through qualitative studies. Quantitative research is a systematic analysis of quantifiable data to which statistical tests can be applied in ways that allow inferences to be associated with probabilities. A strength of quantitative research is the ease of comparison of numerical output. Oppositely, qualitative research explores problems that can incorporate both quantifiable data and qualitative data prone to interpretation, resulting in different possible conclusion from data analysis. Thus, a strong corpus of quantitative studies have been developed to evaluate low-level tasks [7, 90]. But evaluations of visualizations

for low-level tasks, often done through quantitative studies, do not guarantee the usefulness of these systems for high-level tasks. Thus, contributions for visualizations for high-level tasks are often evaluated differently, through qualitative studies. Kerracher *et al.* [90] summarize methods to ensure tasks validity and threats to each approach, e.g. interview domain experts, which are common for the high-level tasks, with the risk being of skewed by the experiences.

Low-level analysis can be diverse. Amar *et al.* [7] present ten low-level analysis tasks that largely capture people's activities while employing information visualization: Retrieve Value, Filter, Compute Derived Value, Find Extremum, Sort, Determine Range, Characterize Distribution, Find Anomalies, Cluster, Correlate. Even though the connections between these low-level analysis tasks to the high-levels ones can not always be directly linked, such lists are valuable to consider basic functionalities that are likely to help towards performing high-level tasks. They also provide useful structure for controlled quantitative studies of low-level tasks.

Kerracher *et al.* [90] discuss models for tasks classifications, and provide guidance for selection between competing classifications for use in the design and evaluation processes, but do not discuss how information can be transferred from one model to another. But their approach is interesting as it is a step towards ensuring the validity of tasks, and to potentially start considering how performance in low-level tasks can contribute towards the success of the high-level ones. Their solutions to mitigate the threats consist of assessing lower level threats together and evaluation of the output produced with the task generated, e.g. ensure the existence of a wide literature to justify the validity of the task generated as well as evaluation after the output is produced. Following that approach, Mazimpaka *et al.* [119] classify a variety of trajectory mining methods and applications using the framework from Andrienko *et al.* [9] and illustrate the tasks with practical examples selected from academic publications. This work exemplifies the approach to generate a high-level structure that can be populated with relevant matching low-level contributions.

Bertin [117] categorizes tasks with three levels of reading: elementary, intermediate, and overall. That separation indicates whether a task requires considering only a single data element in a group, or all the elements existing in the data set. Considering the task according to the data presents pros and cons. It is valuable to consider tasks that can be performed according to the available information, and these considerations ensure that if a task can't be performed with the available data, it might be that the data selection is wrong. But dependencies between data and tasks mean that to compare tasks, it is necessary to consider data sets that are similar. However, due to how information has to be looked at to find the piece of interest, the need to display a single point or several is not clear with Bertin's framework.

For the framework we discuss in details in section 3, we use the task typology presented by Andrienko *et al.* [15]. We selected their typology for its comprehensiveness and the ability to clearly define the scope and output respective to each task. Instead of considering tasks with three levels of abstraction like Bertin [117], this typology makes the choice of dividing the range of visualization tasks into two broad categories with their own subgroups: elementary tasks and synoptic tasks. The difference between the two types of tasks relates to the number of data items necessary

to consider in order to perform the task; elementary tasks address individual data elements, be it singular or plural, while synoptic tasks involve a general view and require groups of data items to be considered in their entirety. That notion does encompass a certain subtlety when characterizing a task as either elementary or synoptic, as it requires the researcher to establish whether a task requires the data set to be considered as a whole or whether individual elements are enough for the task. Andrienko *et al.* define elementary tasks as "tasks that do not imply dealing with sets of references or characteristics as wholes but, rather, address their elements". This means that elementary tasks deal with data without considering information that can only be extracted by considering the whole data set together as one entity. Any task that does fit within that definition is synoptic. Then, the tasks are divided according to their targets, i.e. the information or insight that is gained in order to perform the task. Elementary tasks contain the sub-categories lookup, comparison, and relation seeking. Both the lookup and comparison subgroup can be either direct or inverse, implying that it can be either the search for a data value or for a space in time and space that matches the data searched. Tasks about relation seeking are similar to comparison tasks, but are not bound by a strong characterization of the nature of the comparison like it is for the other sub-category comparison. Andrienko *et al.* [15] and Aigner *et al.* [3] explain in greater detail. But to ensure sufficient clarity when discussing these notions of tasks characterization, which are complex, we present some illustrative examples:

- Elementary lookup: What was the price of the Apple stock on February 15th?
- Elementary comparison: Is the price of the Samsung TV lower than the price of the Sony TV?
- Synoptic lookup: What is the trend of the world CO2 emissions during the year 2020?
- Synoptic comparison: Compare the behaviour of the stock price of Amazon and the number of deaths due to COVID during the year 2020.

Synoptic tasks require us to consider the whole data set and are divided according to whether they are descriptive or connectional. Descriptive tasks are, like the name implies, about specifying properties in the data available. This sub-category possess subgroups of its own, identical to the elementary tasks, with lookup, comparison, which can be direct or inverse, and relation seeking. Connectional tasks are about establishing connections between two or more sets of data (whether they are displayed in the same graph is not relevant to the task characterization), to investigate relationships between different phenomena. Those can be homogeneous or heterogeneous, i.e. whether the phenomena are different occurrences of the same type of phenomenon or completely distinct in nature.

2.2 Visualizations for data analysis of movement and space-time attributes

In this section, we discuss several visualizations that were interesting for our thesis. As the subject of the thesis was not first clearly defined, we read about numerous diverse visualization methods during

our literature review, some of which present little interest compared to the final subject of the thesis. Although these readings partially influenced us, their relevance to our subject is minor, and thus will not be part of our discussion in this section. In this section, we focus our discussions on visualization designs that relate to our subject.

We noticed a large variety of visualization designs discussed in publications. Some patterns were noticed, such as spatial coordinates being represented over a map, although it is not a necessity. An important element for the analysis of data changing over time is time itself. In their book discussing the visualization of time-oriented data, Aigner *et al.* [3] list various approaches to conceptually represent time, including the aforementioned model from Peuquet *et al.* [139], [15]. That list is illustrated in Fig. 2.9. The representation of time is dependent on how it is fundamentally considered for the needs of the visualization being designed, i.e. which facet of time matters, e.g. its linear characteristic or its repetitive nature.

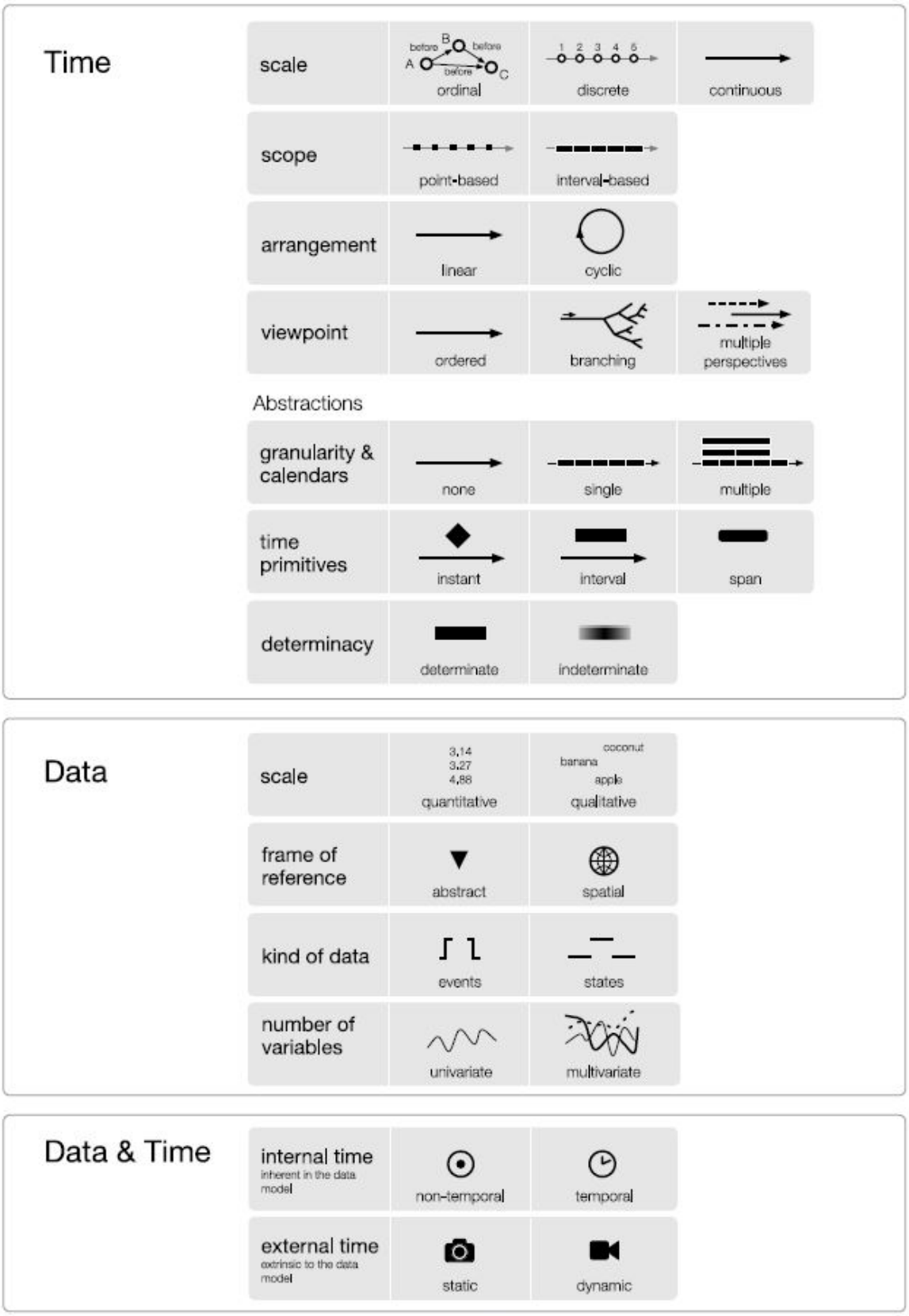


Fig. 2.9 Aigner *et al.* [3] structured and specified the characteristics of time and time-oriented data. Their structure can be translated into other models.

During the beginning of our thesis, no decision was taken over the type of moving entity for which we would later produce new visualization designs for. Thus, there is a variety of types of visualizations for different moving entities and with different contexts designed for various tasks in this section. The approaches to present movement and space-time attributes are diverse, but the concepts used to produce the visualizations are for the majority transferable for different moving entities. A notable case is the analysis of flight records, for which many of the visualizations produced for its analysis use methods transferable to other moving entities, but some attributes specific to aeroplanes push for specific needs. For example, in the graph displayed in Fig. 2.11, Andrienko *et al.* [11] show trajectories that can be chosen by flight planners with each potential path represented by a line over a map with its width indicating the costs of flying over an area. This approach could be used to display trajectories enriched with additional information for other moving entities and/or context, e.g. the display of cars with the predicted likeliness of traffic jams occurrences. The same publication contains another interesting contribution, a clustering algorithm categorizing trajectories with similar paths, with the trajectories coloured according to these categorizations, as illustrated in Fig. 2.12. This approach could be used for other moving entities, such as vehicles entering towns. But analysis of flights can require some specific information, e.g. altitude through representations of trajectories within 3D visualizations, like Buschmann *et al.* [36] who generated visualizations with trajectories enriched with icons and colour to indicate additional information, as illustrated in Fig. 2.13. We found their contribution particularly interesting as it projected the trajectories in the 3D space onto the 2D map to increase readability of paths, which is arduous for trajectories displayed in 3D environments.

Geographical attributes can be characterized and then considered as any qualitative attribute, such as the presentation of regions in a treemap by Slingsby *et al.* [159]. Hybrid approaches do exist, one interesting example being Wood *et al.* [176] who presents OD maps, by tesselating spatial geometries while conserving part of the geographical features. That approach can also be nuanced by considerations of importance of regions due to additional factors, e.g. cities represent a relatively small geographical space within which there is a high amount of activity.

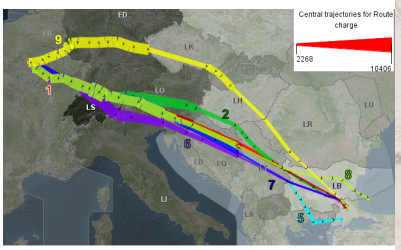


Fig. 2.11 Trajectories represented by lines with the width indicating how important the price of flying over an area is. The colour encoding is used to differentiate several moving entities.

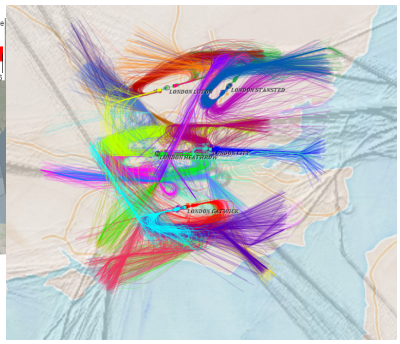


Fig. 2.12 Lines representing approaches of flights arriving at an airport. The similarity of approaches is colour encoded.

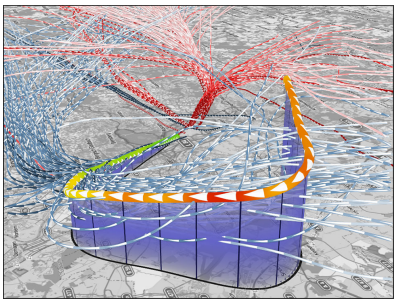


Fig. 2.13 Display in three dimensions of a flight, with colour encoding, for a selected trajectory, the acceleration. The lines and halo going from the points in altitude down to the map help to understand more precisely the movement. While valuable for a single trajectory, this method can not scale.

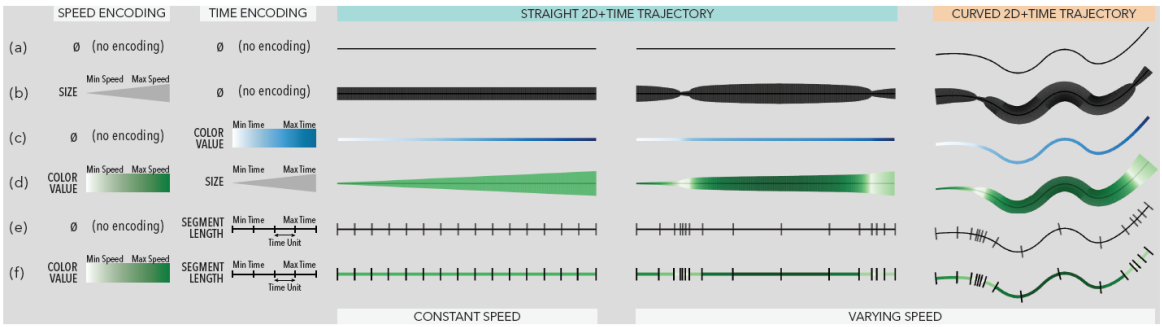


Fig. 2.10 Different encodings of time and speed for straight and curved 2D+time trajectories presented by Perin *et al.* [137]. Both constant speed and varying speed (two slow sections near the start and end, high speed in the middle) are shown.
(a) Neither time nor speed are visually conveyed; (b) size (or stroke width) conveys speed; (c) colour value conveys time elapsed; (d) colour value conveys speed and size conveys time elapsed; (e) segment length (spacing between ticks) conveys time distribution, from which speed can be inferred (the closer two ticks, the slower); and (f) colour value conveys speed on top of segment length. Results from studying nine visual encodings suggest that (e) and (f) are the best choices for conveying both time anspeed,ed and that (d) is the next best.

The amount and quality of the data influences the range of designs that can be made for designing a visualization. A common method to visualise trajectories is to draw a line over a map, to gain insight on points traversed in perspective to their locations. While effective for understanding spatial changes, that method suffers several drawbacks, such as the inability to scale for analysing numerous trajectories, a lack of precise information regarding time when points belonging to the trajectory

were recorded, and the lack of any additional information display that could be relevant, such as attributes of the moving entity, e.g. body temperature of a jogger, heart rate of a cyclist. Additionally, representation of data related to movement is dependent on several factors, e.g. background, level of detail, abstractions, aggregations.

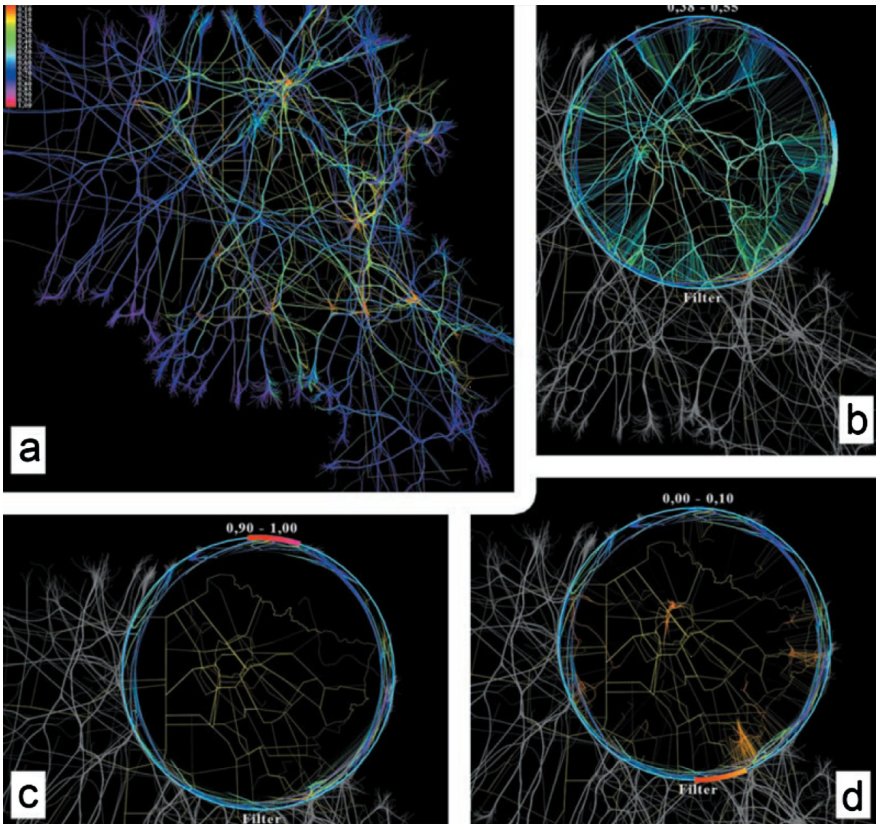


Fig. 2.14 Focus-and-context exploration of bundled airline trajectories in a system introduced by Hurter *et al.* [78].

Perin *et al.* [137] present and evaluate diverse designs for presenting time and speed over trajectories, illustrated in Fig. 2.10. These designs could be reused to represent other quantitative attributes than speed. Additionally, they discuss the elements of the design that characterise the complexity of trajectories. We discuss its influence over our work in section 3.2.3. Enrichment of trajectory display with overlay of information, e.g. with icons, colours, over lines can also scale up to 3D [36]. But that approach is limited when trying to analyse a large amount of movement data over an area. Aggregation can be made in various ways, with diverse levels of abstraction of the trajectories. Willems *et al.* [175] present an overlay of lines and coloured areas to display a summary of movements while keeping the ability to identify outliers. Alternatively, movement within areas can be summarized in a visualization that doesn't display trajectories, but instead information derived from movement, like method of transport [178], or amount of phone calls shared by antennas to indicate density of population over time [10]. Potentially, systems can incorporate different levels

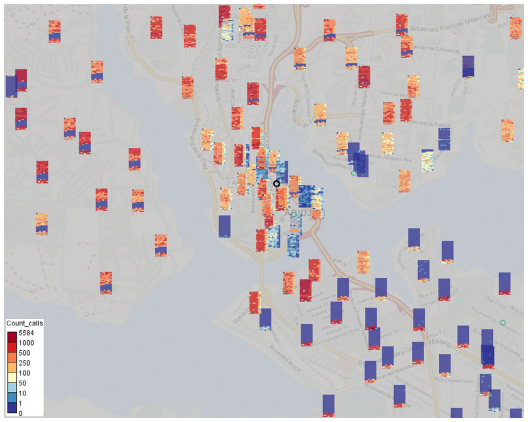


Fig. 2.16 Mosaic diagrams display evolution of an attribute of interest over time. In this case, daily amount of phone calls passing through stations.

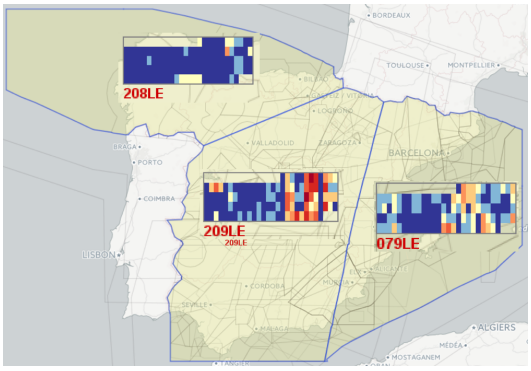


Fig. 2.17 As long as the area delimitation is clear, mosaic diagrams can also be used to display aggregated attributes, such as here with regulations over areas.

of aggregation that communicate through interactions or highlights to convey their connections [16]. Additionally, accuracy of the visualization can be discarded, to display modified data, e.g. distorted trail bundling to display airport connection patterns [95]. The necessity of access to the precise data defines the scope of visualization methods that can help for tasks to complete.

Andrienko *et al.* [18] discuss visualizations in which they compare planned flight trajectories in what they call an 'artificial space', which is about presenting movement in a space where at least one dimension isn't spatial. This method is illustrated in Fig. 2.15, and is original as most visualizations don't display attributes derived from movement, e.g. angles, in relationships to other attributes of interest.

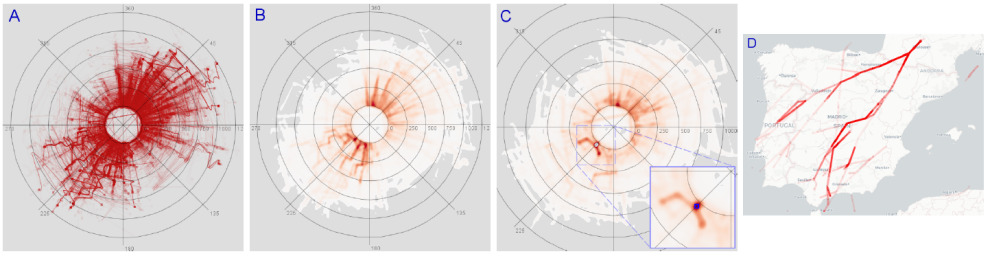


Fig. 2.15 Andrienko *et al.* [18] present visualizations with that are based on the concept of an 'artificial space'. A: Planned flight trajectories are represented in an artificial space with polar coordinates: movement direction (angle) vs. distance from the cruise phase start (radius). B: A density map summarizes the whole trajectories. C: The density map summarizes the segments that were substituted by shorter paths in the real flights. The inset on the bottom right shows a filtering window around a density hot spot. D: The trajectories crossing the hot spot in the artificial space are shown on a geographic map with 5% opacity.

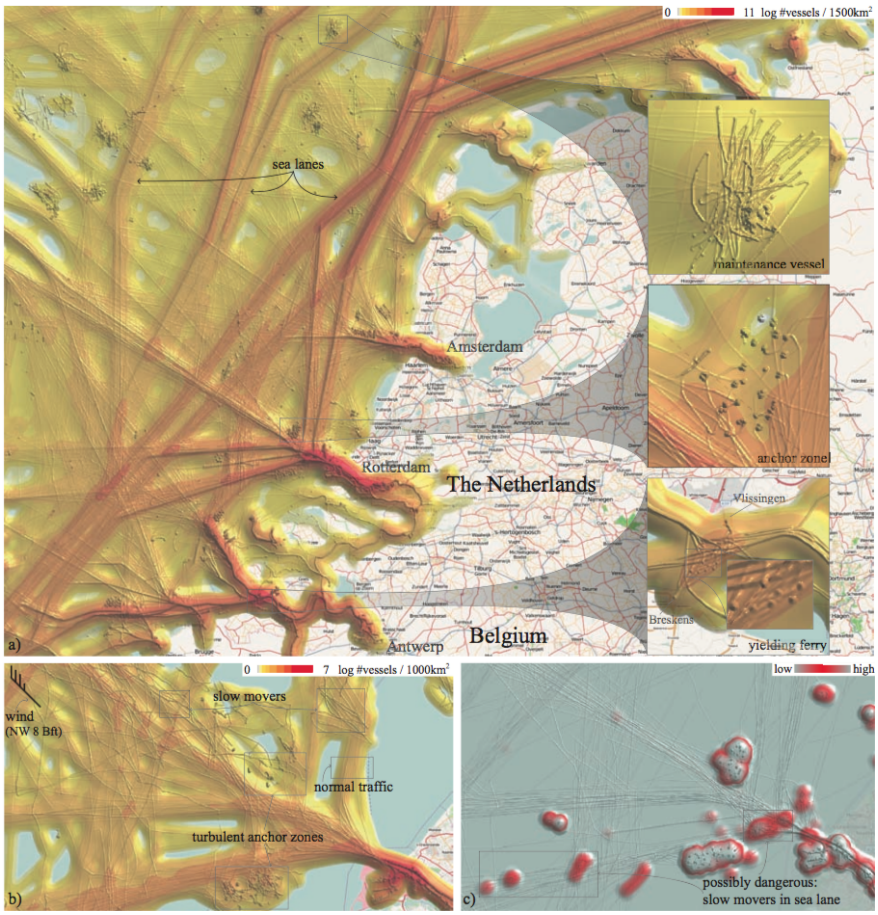


Fig. 2.18 Vessel density of the Dutch coast with a kernel used to calculate the value for each point. This method allows finding outliers that experts can understand thanks to their knowledge of normal vessel behaviour. It is interesting to point out that within the study, the users were capable of finding outliers, but the characterization of the cause was due to their experience with real-life cases.

The visualization that most interested us was the time mask, developed by Andrienko *et al.* [17] and illustrated in Fig. 1.1, which they also used later for the analysis of football games [8]. The time mask is a visual annotation that overlays sections of a time series where conditions, such as those entered in a query, are matched. The overlay is thus indicated visually before filtering the data. This approach allows sections of time sequences to be visualized, and can be applied whether the data displayed is the original recording ('Raw Data' in Card's terms), or derived from the data ('Data Tables') like average positions of footballers over a field. Our interest in the time mask was motivated by the prospect of considering extensions of that concept with additional variations.

2.3 Validation processes for designs and studies

Ensuring the validity and usefulness of contributions is critical for researchers. Many models have been proposed in visualization and beyond to help achieve this, and discussing all of them is not

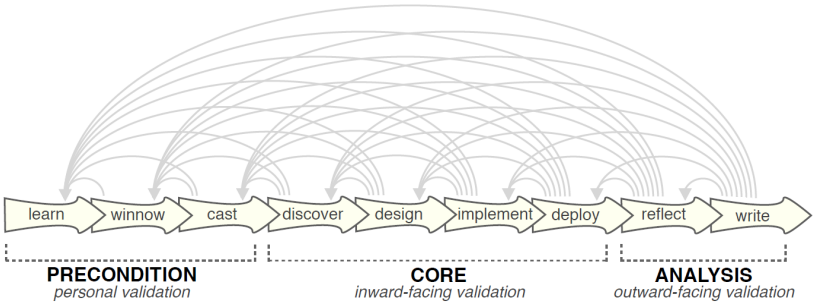


Fig. 2.20 The Design Study Methodology of Seldmair *et al.* [156]. For each steps, pitfalls are to be avoided to ensure a contribution that is valid and valuable.

within the scope of this thesis. In this section, we discuss those that we found the most influential and explain how they helped us shape our approaches to ensure validity and how their strengths and weaknesses drove us to design the SFNCS.

The Nested Model developed by Munzner [126], illustrated in Fig. 2.19, supports visualization designs by considering four levels at which threats to validity may occur that each need consideration, with some issues impacting several layers.

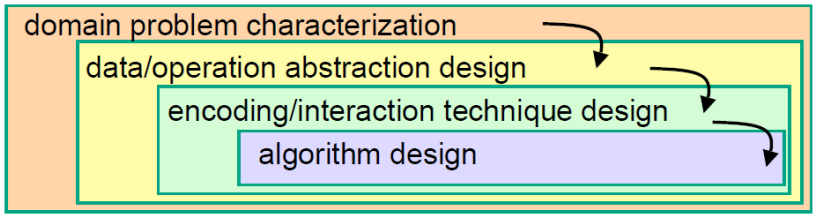


Fig. 2.19 The Nested Model of Munzner [126]. For each step of the design study, threats are to be identified and countered to ensure validity of the contribution.

For design studies, Seldmair *et al.* [156] developed the Design Study Methodology, illustrated in Fig. 2.20, which describes the process of developing a visualization system in an applied context into 9 stages: learn, winnow, cast, discover, design, implement, deploy, reflect and write. These are illustrated in Fig. 2.20. The Design Study Methodology has then been used in many research projects to both ensure and justify the validity of visual designs that are developed. While it is not possible to verify all potential pitfalls of the numerous aspects that entail a design study, indications of reflections over some that were accounted for increase the strength of the claims.

Another approach is the Algebraic Visualization Design (AVD) developed by Kindlmann *et al.* [93], which can be used to justify the appropriate usage of a visualization method, such as McNutt [122] does for table cartograms. Their approach is very different from ours as our approach relies on designers and researchers to evaluate through studies the strengths and weaknesses of designs. But we think such an approach could be valuable to help filter out potential design errors.

An important contribution was made by Munzner [126] with her nested model to indicate a stream of four levels for validation, to identify threats and validations to against them to ensure contribution

validity. While the nested model was critical to help researchers make a claim on the validity of their contribution, it doesn't allow comparing values of contributions according to their context. The problem thus remains, a contribution is made and valid, but is it better than previous work? If so, can we assess by how much? Is it for a specific task or an ensemble?

Assessing the value of a claim concerning a contribution is often addressed by setting up studies in which participants perform tasks that are evaluated, so to ensure that claims made about theoretical concepts apply against particular benchmarks. Pena-Araya *et al.* [134] conducted a study about identifying correlation over space and time, using different representations of the same data. Their approach strongly influenced the studies we set up, specifically regarding the evaluation of the participants' responses. This work was particularly interesting as it illustrated analysis of tasks with different characterization of data. While asking participants to evaluate spatial correlation with the data displayed, they split the data necessary to answer as either a single location, a location in a region, or all locations displayed; time-wise, they split the data as either a single time, a time interval, or all times. Our approach for studies we discuss later in Chapter 5 follows this method to set groups defined by data characteristics.

2.4 The influence of uncertainty for analysis of movement and space-time attributes

An important characteristic for the elements that composes the framework we present is whether they are Measurement or Judgement. This distinction is further discussed in section 3.2.4, but is strongly inspired by our review of literature discussing the subject of uncertainty. The terms Measurement and Judgement relate to the notion of uncertainty that arises from attributes and characteristics that rely upon interpretation. There are many different definitions of uncertainty [53], most are similar to the one presented by Zhang & Goodchild [183]: it is the differences between reality and the knowledge, representation, or understanding of reality. But that definition underlines an issue with discussions about uncertainty: many notions are encompassed into this one term. In this section, we discuss the impact of different types of uncertainties for various tasks, how it can be communicated and how it influenced the conception of the SFNCS.

2.4.1 Communication of uncertainty

Uncertainty, and by extension context, is an important subject that needs to be considered when discussing analysis of data. It is intrinsically related to the structure of data. The transformation of information (be it from a natural phenomenon or human-made entry) can always be subject to error, to various degrees, e.g. the precision of measurements, calibration of instruments or human mistake. Interpretation of data made by a human is strongly dependent on the task performed, and the context associated with it. For example, precision only matters up to a certain point for which

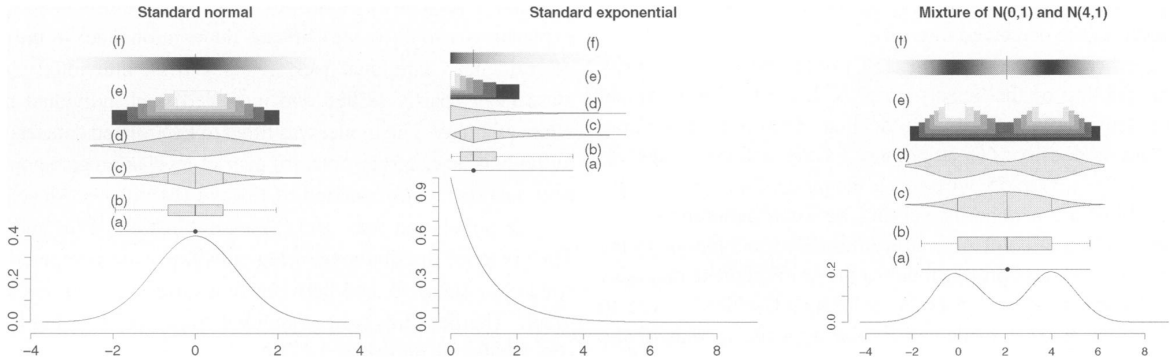


Fig. 2.21 Several methods used to display distribution. Jackson *et al.* [83] created the density strip method, visible at the top. This method is interesting for its efficiency to convey the gradual changes in the distribution while taking little space, making a potential valuable method for connected views within a visualisation.

enough confidence in the data and visualizations are generated to perform the task. Allocating a large screen space to display a graph is often done to ensure readability, but reduces the numbers of graphs that can be presented at once, thus visualizations should aim to allocate enough space for understanding of the information displayed, but no more. What 'enough' means is dependent on the context. Jackson *et al.* [83] discuss how not displaying the entirety of a distribution could result in the loss of nuanced information that is important to trust the decisions made looking at information, but also realize that space to display visualizations is limited and should thus be used in a compact yet clear manner as argued above. They thus developed the density strips method, which uses gradient variations to convey variations of distributions. But the question goes deeper, as considerations of impact of levels of precision, as evaluated by Greis *et al.* [65], or different considerations to include contextual information, as investigated by McCurdy *et al.* [121], have various impacts, and remain open questions for designers of data analysis systems. Furthermore, in some situations where the information presented is either lacking precision for the task at hand, or when the visualizations display predictions with a range of potential values, these have to be presented. The most efficient design to consider how to display that uncertainty is yet to be answered. An interesting set of examples are the visualizations of predictions of hurricane movements, such as Cox *et al.* [50], displaying trajectories predicted by their software, versus Mirzargar *et al.* [125] additionally displaying the estimated area when the hurricane is likely to pass. Due to the shape generated from potential areas, displaying simply the zone was often misunderstood into thinking that could be the area the hurricane would affect, instead of its position. The previous design was technically relevant and understood by experts, but due to context, it was vastly misunderstood by viewers without prior experience to this field. While not all cases of context influencing the perception of a visualization are that radical, it is important to consider how a visualization can be understood in relationship to prior knowledge about the data and contextual expectations.

As part of the effort to display information effectively, several designs can be chosen for the display of the uncertainty around storm (and other) paths. Jackson *et al.* [83] present an interesting design based on shaded monochrome strip whose darkness indicate density. The choice of purposely aggregating the data to ease numerical comparison of attributes or continuous representations of estimations then depends on the desire of the designer to engage trust in the person viewing the visualization. Aggregated representation of the data allows for quicker judgement, but the loss of information can be problematic. Then designers have to consider uncertainty according to two opposite principles: present all the information to induce trust, but modify it to simplify its understanding.

Another interesting work for our research was the methodology developed by Deitrick *et al.* [52], which defines and discusses implicit and explicit uncertainty in visualizations. They define explicit uncertainty visualizations as representations that directly identify errors and unknowns through quantitative values, e.g. error bars, or qualitative values, e.g. confidence reported with a Likert scale. These uncertainty values can only be interpreted in one manner. They define implicit uncertainty, in opposition, as uncertainty that is context dependent and is thus dependent on interpretation. Their approach to deal with implicit uncertainty is to transform it into an outcome space: a range of combinations that are likely to happen according to the values of two uncertain variables, i.e. the risk depending on two uncertain variables is represented with coloured areas, each colour indicating the level of risk.

Deitrick *et al.* [53] further discuss it and the impact of uncertainty over decision-making, again with the philosophy to derive quantitative representation of implicit uncertainty accordingly to the context. As they focus on the impact of communication of implicit uncertainty over decision-making, they present implicit uncertainty visualization in a descriptive model to reflect on how decision makers contend with uncertainty. The resulting matrix displayed in Fig. 2.22 shows combinations of aggregation of data and uncertainty, and how those can be displayed to help reflect upon the decision-making process.

"Strategies to cope with uncertainty fall into three basic groups: reducing, acknowledging, and suppressing (Lipshitz and Strauss 1997)" Lipshitz *et al.* [108]

While this approach is interesting, it is strongly dependent on the ability to transform the implicit uncertainty into a quantitative value. And as exemplified by McCurdy *et al.* [121], it is not always beneficial to transform information to correct its display. Context thus needs to be adapted for information management and communication, both accordingly to the expected end-user of the solution provided.

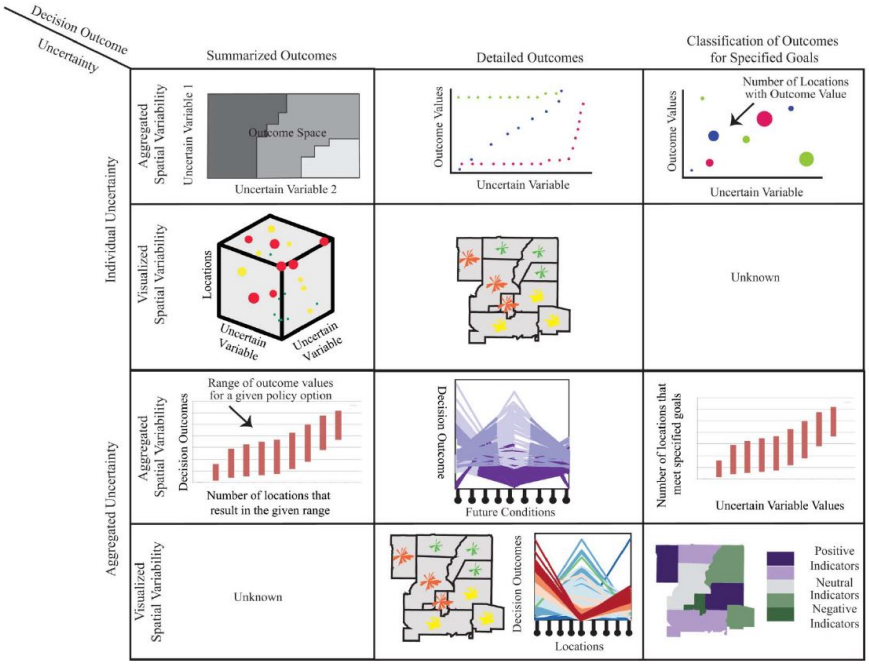


Fig. 2.22 The matrix developed by Deitrick *et al.* [53] depicts the visualization solution space based on the way uncertainty and decision outcomes are conceptualized.

2.4.2 Integrating the communication of uncertainty into systems to refine understanding of movement and space-time attributes.

Uncertainty is dependent on context. The influence of context over uncertainty is notably discussed by McCurdy *et al.* [121] when introducing their framework for externalizing expert knowledge about discrepancies in data. In their prototype, they added to their visualizations the possibility to create annotations, to enrich the data presented with contextual information that can not be captured simply with quantitative data or broad categorization of the information. This approach differs from the ones discussed in section 2.4.1 because the information necessary for the analysis to be carried out is no longer computable. In their work, the necessity of the generation and display of contextual generation originates from different policies between countries regarding their process to report new cases of Zika. The data recorded and displayed using the same scaling over a geographical visualization resulted in data discrepancies. While the experts McCurdy *et al.* were collaborating with were aware of policies requiring them to adapt their mental model of the situation, communication and collaboration were hindered by these data discrepancies.

As the data displayed was technically correct, it still was not presenting a representative depiction of the reality of the situation. McCurdy *et al.* use the term '*implicit error to describe measurement error that is inherent to a given dataset, assumed to be present and prevalent, but not explicitly defined or accounted for in the dataset.*' The implicit error is defined by its:

- Type: whether the error is systematic or random

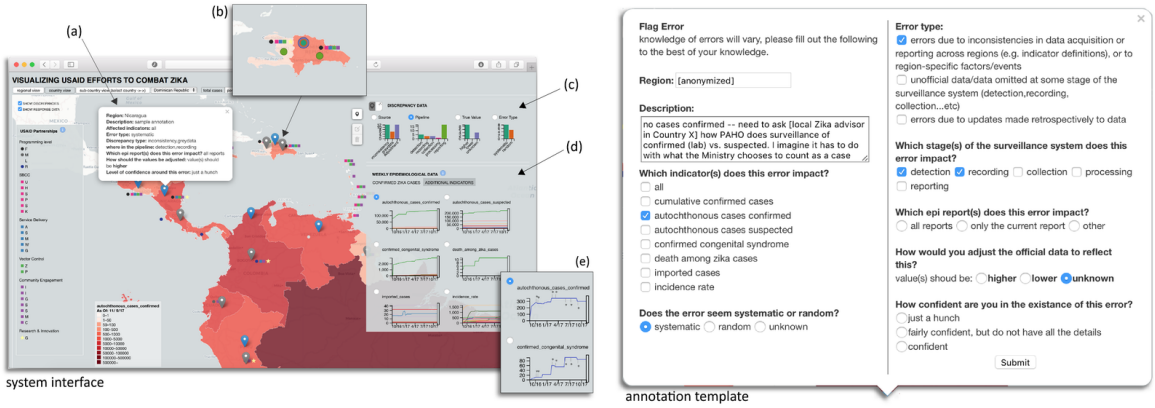


Fig. 2.23 Prototypical instantiation of the framework designed by McCurdy *et al.* [120] for externalizing implicit error. The system allows the inclusion of framework generated by experts when using the prototype. The data is thus not modified to fix the

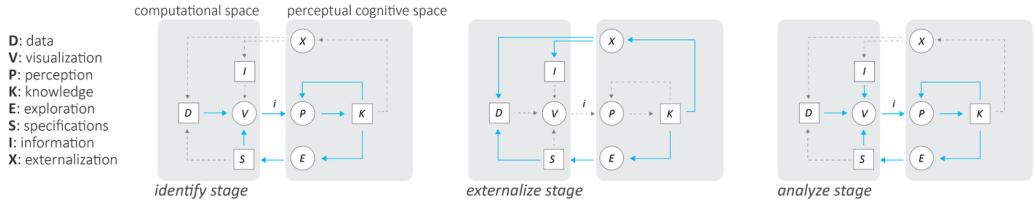


Fig. 2.24 The process model to externalize implicit error, as described by McCurdy *et al.* [120].The purpose of this process is to enrich the computational model with prior knowledge an analyst possesses but isn't integrated to the data.

- Direction: indication of the estimated difference, e.g. negative or positive for quantitative values
- Magnitude: how important is the difference due to the implicit error
- Confidence: how much the person is confident in an implicit error existing
- Extent: how much of the information is impacted by the implicit error, e.g. how many regions are afflicted with the implicit error

This work and informal discussions with the author over extension of the concept were influential over the development of Measurement and Judgement information we define in section 3.2.4. Our framework is defined with high-level concepts but the elements that characterize implicit error could potentially be reused for a lower level of abstraction.

An approach that could be considered for communication of uncertainty is fuzzy logic. Fuzzy logic, a many-valued logic that considers truth value to potentially be any real number between 0 and 1 instead of only false or true. Zadeh [181] describes fuzzy logic as a logic that adds an in-between between continuous and quantized information, the granulated information, also known as the fuzzy logic gambit. Kosko *et al.* [98] discusses fuzzy logic and points out that most systems rely on rules set by an expert. That notion is illustrated in Fig. 2.25. A notable exception is unsupervised learning,

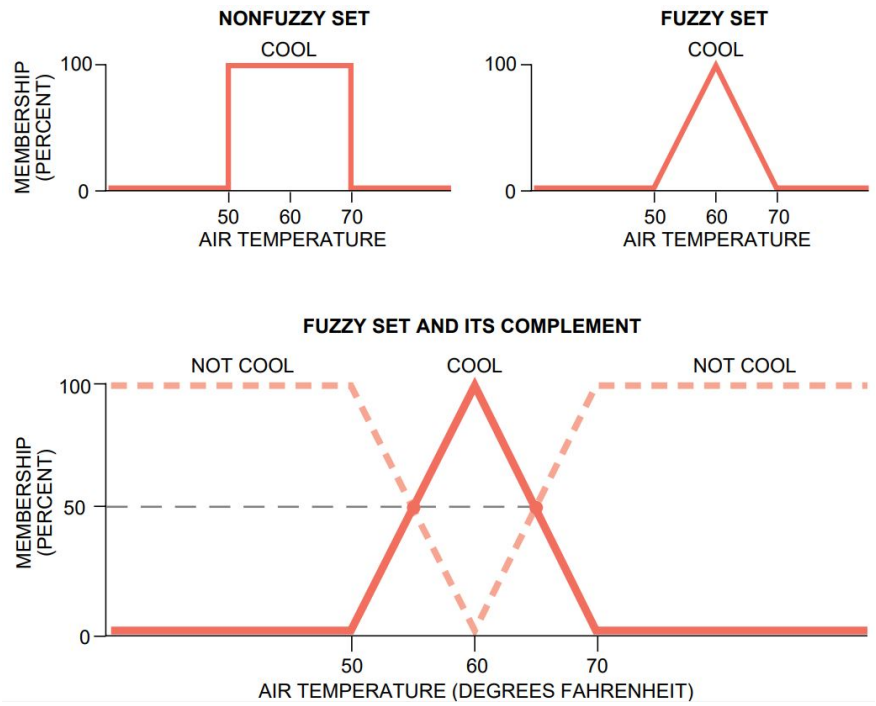


Fig. 2.25 Difference between a fuzzy set annon-fuzzyzy set, and an explanatory graph of a fuzzy set with its complement [98].

but then instead of creating rules, they end up being implicitly created through mimicry. Thus, while fuzzy logic presents an approach to transform and increase the complexity of the concept of truth, we argue that it is still dependent on humans to set its element or at the very least to judge it. It is thus a valuable concept that might be incorporated into our approach in future work, but its focus is not on communication, like us. Similarly, an interesting contribution of Scheepens *et al.* [150] presented a system that generated density maps using a density model that allows users to modify the parameters. That approach is motivated by the objective to allow experts to express their knowledge about the context inherent to the information but not present in the data.

The work of Andrienko *et al.* [16] is also interesting as it discusses the analysis of movement data for extracting and exploring significant places. The notion of significant is here used to define places where the basis of clustering movement according to different attributes results in the creation of a geographical point or area, depending on the level of aggregation. The new places are then overlaid with graphs displaying variations over time of attributes of interest. The notion of significant is here let up to evaluation of the analyst. Uncertainty related to the importance of a location is thus dependent on movement here, as opposed to context-based enriched data, such as Parent *et al.* [131] who overlay indications of positions and names of specific locations likely to be interesting for the moving entity for which records have been gathered, i.e. tourists.

There are thus many facets to uncertainty, relating to either information itself or communication about it. While the focus of our work pivoted from the subject of uncertainty, these readings strongly influenced our approach to define the notions of Measurement and Judgement in section 3.2.4, which acknowledges that interpretation is dependent on context and a clear separation of these notions is necessary to reflect upon the decision-making process, including its evaluations when setting up studies designed to evaluate new contributions.

2.5 Chapter summary

In this chapter, we discussed literature which influenced our thesis. We first reviewed literature which analysed how information was considered and organized. The literature addressed was particularly focused on frameworks discussing connections between different components, and their relations, e.g. the WHAT-WHERE-WHEN of Andrienko *et al.* [12]. The reference model of Card [39] is particularly important as it presents how information goes through a connected ensemble of steps from raw data to the user.

We then discussed papers which formally characterize tasks, and how these relate to previous characterizations of information, identifying how these elements are built thanks to each others and how links can be made explicit by using the same information characterization.

We then reviewed various examples of visualizations in the literature to present interesting approaches to display the information characterized in graphs to help their analysis. Following discussions of visualization methods, we discussed processes to ensure their validity.

These diverse publications indicate a large amount of contributions which discuss information for visualization analysis, how it can be displayed, and how we can ensure the validity of novel contributions. We notice that contributions to ensure validity are not based on the shared components used to characterize information, implying numerous difficulties for comparison of different contributions. We thus developed a framework to define the evaluation process when generating a study to assess a novel contribution in chapter 3.

But first, we discussed the influence of uncertainty over our work. While not the focus of our thesis, uncertainty strongly influenced the development of our framework as considering it drove us to define new important notions: Measurement and Judgement, defined in section sec. 3.2.4. These considerations are important for our framework to support quantitative and qualitative studies.

Chapter 3

The Systematic Framework for N-scales Characterizations of Studies (SFNCS)

The main contribution that results from this thesis is the **Systematic Framework for N-scales Characterizations of Studies (SFNCS)** presented in this section.
In this chapter we first discuss its origin, with a taxonomy for which its development allowed us to reflect on the high level concepts. Then we discuss the concepts that were selected to allow discussing its elements, the process to select the blocks that compose it, and its structure. Finally, we discuss examples of usage of the SFNCS for diverse fictive studies set-up.

3.1 Origin of the framework

During our literature review at the beginning of our thesis, the choice of the data set(s) we would work with was not definitive. To summarize, the potential sources were: pedestrian movement, ship moment, aircraft movement and/or traffic movement. We thus sought and read diverse scientific contributions to understand the current problematics that existed with these moving entities and the approaches whether existing or researched to tackle them. During our readings, we started to notice some similarities of visualizations depending on moving entities. To assess whether the pattern was recurrent or random, we decided to investigate whether there were relationships between types of moving entities, tasks that were performed when analysing data related to them, the type of visualizations and algorithms produced to help tackle them. Additionally, we wished to investigate whether display of contextual data visually encoded was more common for a certain type of moving entity or a certain type of task. Finally, we were inspired by Kitchenham *et al.* [94] and our approach was motivated by the goal to *'assess the frequency or rate of a project development factor such as the adoption of a technology, or the frequency or rate of project success or failure'* with a rigorous method that would allow us to focus future design concepts with the most successful approaches.
To evaluate the existence of patterns, we decided to categorize contributions and to populate a taxonomy to allow for comparisons of their attributes. We read previous literature that were organized using

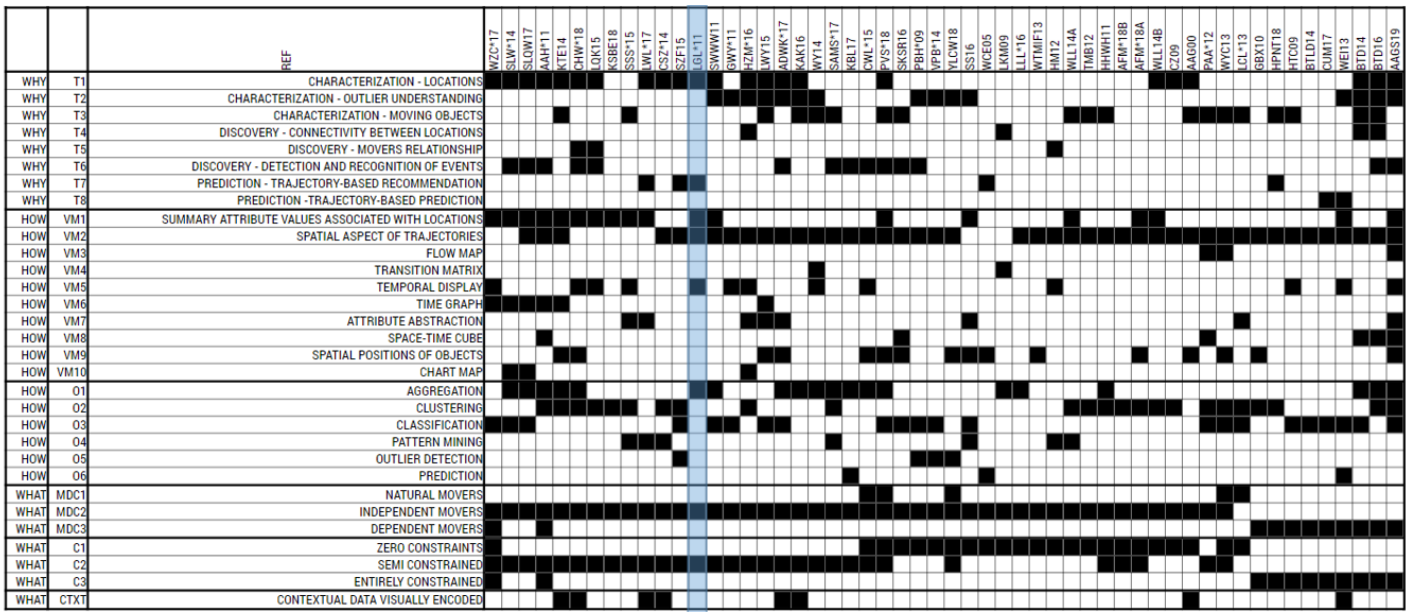


Fig. 3.1 Our WHAT-WHY-HOW taxonomy of trajectories visualization research illustrated using the Bertifier technique [135]. The documents are ordered through Bertifier's visual similarity algorithm that makes patterns easier to discern. Find the data here: https://bit.ly/2vyoSoQ. The blue column indicates a document discussed as a populating example discussed in section 3.1.

the Bertifier technique [135], which allowed to conveniently reorder entries, and also incorporated methods to order entries according to visual similarity to render patterns easier to discern.

Many elements from this section are discussed in the poster we published at EuroVis 2019 [5]. The poster is available in the appendix A. We discuss them in more detail here, as this process was an important step towards the development of the SFNCS.

We decided to develop the taxonomy by linking the high-level of abstraction WHAT-WHY-HOW typology [32] discussed in section 2.3 to identify the data, tasks and idioms being employed. We were inspired by the list of visualization methods of Andrienko *et al.* [16] to connect for every single paper the task to visualization methods used. In their review of methods and applications for trajectory data mining, Mazimpaka *et al.* [119] included a finite categorization that named both visualization methods and algorithms. Their taxonomy interested us as it listed unique names for each visualization method and algorithm, which would frame our search for patterns.

Additionally, as mentioned more in depth in section 2.4.1, our interest to investigate the impact of context over the analysis of movement and space-time attributes drove us to add a category to the taxonomy which was a binary quality to capture the presence or absence of contextual information. Our categories are set up as a lower-level of abstraction approach to populate the WHAT-WHY-HOW typology. The categories populating the HOW group are taken directly from the taxonomy developed by Mazimpaka *et al.* [119] with minor additions, as we encountered examples in the literature that qwere not well captured by the phrasing of these categories.

The WHY block The tasks listed in the taxonomy fit into the WHY category from the typology

of Brehmer *et al.* [32]. Mazimpaka *et al.* [119] define high-level tasks as "application problems" and tasks with a lower level of abstraction "main task". We followed the requirements of Brehmer *et al.* [32] and for each task what is the expected input and output. Our model was derived from the semantic trajectory model of Bogorny *et al.* [29], but made simpler for clarity. We list the resulting tasks selected for our taxonomy:

- **T1 - Characterisation of locations:**
  Input: trajectories
  Output: attributes for locations
  This task aims to label **locations** by using data such as environment or trajectory data
  For example: Willems *et al.* [175] encode details of speed variations of individual vessels within a small kernel to highlight anchoring zones where multiple vessels stop.

- T2 - Outlier understanding: Input: trajectories - Output: attributes to trajectories - This task aims to find trajectories within a set that can either be errors within the data or cases too extraordinary to be considered relevant within the analysis. Mazimpaka *et al.* [119] consider this task as a data mining method, however, there are enough visualisation methods designed directly to explain outliers to warrant an explicit task under our task taxonomy, e.g. Hurter *et al.* [78] use visualisations to detect outliers within flights records or Laxhammar [101] who uses a model to detect anomalies for sea surveillance.

- T3 - Characterisation of moving objects: Input: trajectories - Output: attributes of movers - The aim of this task is to gain additional information about the **mover**, e.g. characteristic features, categorisation, or places frequently visited. For example, Parent *et al.* [131] present how from the stops of a trajectory and the information about elements in the area indicate that the mover is likely a tourist.

- T4 - Discovery and characterisation of the **connectivity** between locations: Input: trajectories and locations - Output: relationships between locations - The aim of this task is to discover and characterise the existence of a connection between locations. For example, Wood *et al.* [176] present an encoding of Origin-Destination (OD) maps and use it to illustrate links in between US counties, using data of migration or commuting between each.

- T5 - Discovery of movers' relationships: Input: trajectories - Output: relationships between movers - This task establishes whether two movers share some connection based on regular trajectories parameters such as a mover following another. Mazimpaka *et al.* [119] named this task "Discovery of social relationship". It was changed in an effort not to be domain-specific. One example is Li *et al.* [103] who introduce data mining functions used to discover animal movement patterns.

- T6 - Detection and recognition of events: Input: trajectories - Output: event - Originally this task aimed to detect social events only, but we renamed it, following the scope from the

framework of Andrienko *et al.* [9]. They have stated that "any pair" of a point and a location "in a trajectory can be treated as a spatial event" but not "any such event is significant with respect to the goals of analysis". Relevant events are extracted in combination to relevant context, such as "meetings of two or more movers" or "Attaining particular values of movement attributes, e.g. cars exceeding speed limits".

- T7 - Trajectory-based recommendation: Input: trajectories and locations - Output: recommendations of trajectories - This task aims to suggest a potential route or sets of places based on a model of the user goal. Data operations can be computed to help, by using semantic data about the mover and the environment, where it is assumed that movers with similar movement records are likely to share other semantic attributes such as preferences. One example is Qian *et al.* [144] who present an algorithm to suggest trajectories for multi-aircraft planning.

- T8 - Trajectory-based prediction:Input: trajectories and locations - Output: predictions of trajectories - The prediction task is defined by the interest of the user to know a possible future position for a mover, its destination, or the route it will take. One example of this task is Cox *et al.* [50] who display a set of potential path predictions for hurricanes.

The HOW block The HOW section of our taxonomy contains **visualization methods**. Here we are concerned with HOW the trajectories are encoded - this is the Visual Structures phase of Card's model. Our examples are derived from the work of Andrienko *et al.* [16] and types of operations performed on the data to produce additional content, inspired from the review of Mazimpaka *et al.* [119] In this taxonomy we name the visualizations after the ones listed by Andrienko *et al.* [16]. That approach can not be exhaustive, as it limits the number of visualizations to a finite number. We use the method names proposed by Mazimpaka *et al.* [119], to which we added the category 'Aggregation' which is used in a number of papers but wasn't included in their work due to their different focus. Do note that we evaluated the visualisations within the paper that were representing movement, i.e. if a paper is composed of graphs representing results of a user study, those would not be used to populate the taxonomy. We complemented the baseline taxonomy with the 'Attribute Abstraction' to distinguish techniques where all the dimensions used are attributes related to the movement but none are spatial. The visualizations methods listed originate from the taxonomy of Andrienko *et al.* [9], with the distinct separation of the operation of O1 'Aggregation'. The resulting list of visualizations is:

- Vm1 - Summary attribute values associated with locations: representation of movement aggregated over a point or a region, and then maps the calculated sum of times with movement occurring over it to a colour scheme. E.g. the work of [152] shown in Fig. 2.18.

- Vm2 - Spatial aspect of trajectories: representation of trajectories drawn by displaying points of the recorded positions of a moving entity and then connecting them with lines. This approach is often done over a display of the geographical context of the movement, e.g. a map representation of the area.

- Vm3 - Flow map: this visualization is a representation that does not represent an accurate depiction of movement, but that aggregates different movement records into fewer lines that represent the set. The more aggregated the movement is, the larger the width of the line representing the movement. This approach trades off spatial precision for readability of movement of numerous moving entities. The lines generated may split, at their origin or destination, into more detailed views, e.g. several lines can merge together from flights originating from different airports in the USA, before merging across the Atlantic and then splitting again to provide detail about destinations in European airports.
- Vm4 - Transition matrix: representation of movement in a matrix where each column and line represent an area. Detailed information of the trajectories is lost, but trends between regions can be easier to spot with this approach.
- Vm5 - Temporal display: display of an aggregated attribute related to movement over time, e.g. speed over time.
- Vm6 - Time graph: similar to the temporal display, but detail for each object is kept in the visualization.
- Vm7 - Attribute abstraction: representation of two non-spatial attributes, with the x and y axis used to indicate the characteristics of movers.
- Vm8 - Space-time cube: representation of movement similar to Vm2, but within a 3D space with one axis used to indicate time. The resulting drawn trajectories thus both move in a 2D space to indicate movement over geographical scales, e.g. latitude and longitude, but the lines drawn continuously move up the view as time passes.
- Vm9 - Spatial positions of objects: indications of positions in which entities have been spotted. Unlike Vm2, there are no lines used to indicate how movement could have happened between the recorded positions.
- Vm10 - Chart map: representation of the space over which the movement is recorded with additional visualizations positioned over certain points to communicate aggregated data at these locations.

The list of operators is the following:

- O1 - Aggregation: Aggregating data involves taking a set of trajectories and reducing them to one that represents some overall way to indicate what matters for all of them. It is a method to retain global information while accepting the loss of detail
- O2 - Clustering: Clustering operations labels groups of trajectories together based on their similarities, but no additional trajectory is created to represent them.

- O3 - Classification: Classification is an operation that adds semantic data to the trajectories, which can later on be used within the visualization of the trajectories directly or ensemble statistics.
- O4 - Pattern mining: Computed operations done to find how to qualify the type of movement for the trajectory, such as going to work, going back home, etc.
- O5 - Outlier detection: Computed operations to detect trajectories that are very likely to be errors, such as cases where the mover moves too far in a very short time frame.
- O6 - Prediction: Computed operation to estimate, based on known previous positions, and possibly environmental factors that influence the mover.

The WHAT block Our reflection on how to define the elements of the WHAT block were originally motivated by the consideration that context influences data analysis [120, 121]. Regarding analysis of movement and space-time attributes, an interesting example is described by Bonham *et al.* [30] where trajectories of vessels appear counterintuitive unless the person looking at the data is aware of specific rules or motivations due to the context, e.g. loop nearby a dock available for a long time in order to land when the varying price of the cargo is at high as possible. Another context is discussed by Andrienko *et al.* [18] who discuss flight variability and draw attention to the importance of context such as weather, or exceptional events such as strikes called by airport workers. Brehmer *et al.* [32] advocate for a "*bring your own WHAT*" approach. We thus developed our own blocks for WHAT, with the objective to convey differences between moving entities according to their unique context. The resulting list is:

- Mover's decision capabilities(MDC): The mover's decision capabilities is a category that indicates whether the mover is the one deciding for the trajectory they follow. The mover's decision capability is useful to assess whether a trajectory is an error, if the mover possessed the ability to take another trajectory, or the existence of potential interactions between several movers.
  - MDC1 - Natural movers: this category describes objects where the movement will not undergo modifications due to the will or action of a sentient being, e.g. storms or glaciers.
  - MDC2 - Independent movers: Independent movers are responsible for deciding their own movement, e.g. a pedestrian or a car being driven by an occupant.
  - MDC3 - Dependent movers: Dependent movers are not making the decisions for the movement they are undergoing, e.g. a plane following direction given by an agent outside.
- Levels of constraints(C): This section presents categories that define how constrained a mover is, i.e. how many rules the mover has to abide to. This notion is a continuum rather than a series of precisely defined ordered categories, and this notion can also change depending on the context, e.g. a car is semi-constrained, limited normally by legal constraints, but has access to a

range of velocities and several directions. It can however be forced into a deviation in various ways: when being towed, or when under instruction by an external person, making it entirely constrained exceptionally. This notion also depends to a certain degree of the precision of movement available to the moving entity, e.g. in normal situations planes are limited to a certain path but for safety are allocated a certain range of movement, due to both safety concerns and difficulty to very precisely keep a specific desired trajectory in problematic context, e.g. high winds.

  - C1 - Zero constraints: whereby the mover is able to go in any direction within its physical capability so to reach its destination.
  - C2 - Semi constrained: whereby the mover has sets of possibilities for trajectories, but is not free to take all of them.
  - C3 - Entirely constrained: whereby movers are unable to move in ways other than those predefined, e.g. trains are forced to move on rail-roads, and are unable to deviate from the planned routes.

- Contextual data (CTXT): This category is used to label documents with visualizations that add contextual data different to movement, e.g. display of metadata of points of interest that can help to understand the reason behind the time a mover stops at a specific location.

The taxonomy was first populated following a convenience sampling [56] approach. The convenience sample was useful to indicate issues with the categories of the taxonomy, but was neither systematic nor reproducible, thus we restarted populating the taxonomy, following this approach:

- **(1)** Search on Scopus for all conference papers and journals articles that discuss “trajector(y/ies)”,“visuali(s/z)ation” and exclude keywords that were representative of notions that fell out of our scope, e.g. "trajectories of eye movement".

- **(2)** Remove posters, short papers and VAST challenges to ensure contributions at a full paper level. Full papers provide a representation of the quality of the state of the art, rather than methods used out of habit or specifically contextual reasons, e.g. news reports often display cones cones to represent predictions of hurricanes positions even though it is commonly misunderstood [125].

- **(3)** Remove the papers outside our scope and use the remaining ones to populate the taxonomy.

This resulted in 54 documents populating the taxonomy [4, 18, 19, 36, 46, 45, 69, 76, 79, 100, 110, 133, 136, 146, 147, 157, 162, 177, 180, 2, 13, 37, 35, 42, 47, 67, 73, 81, 80, 87, 97, 99, 51, 105, 106, 109, 111, 113, 114, 158, 1, 132, 152, 163–168, 170, 169, 172, 178, 161].

For clarity, here’s an example : Liu *et al.* [111] discuss route suggestions for taxi trips and reasons as to why they follow them or take other paths. The paper presents visualisation techniques to display attributes and statistics associated with different routes to analyse their diversity and help choose the

best route. Taxis are instructed on the beginning and end of a trip, but decide of the route to take, and thus fit the category 'Independent movers (MDC2)'. They are following a set of rules but have a range of options to perform their movement, thus fitting in the category 'Semi-constrained (C2)'. The tasks the paper discusses fit the 'Characterisation of location (T1)' category when using techniques to display hot spots in a city, and additionally fit the category 'Trajectory-based recommendation (T7)' when using their system to help the user choose the best route possible. The paper presents several methods to support those tasks. Some visualisations include heat maps to indicate hot spots, achieved using aggregated data, thus fitting categories 'Aggregation (O1)' and 'Summary attribute values associated with locations (Vm1)'. Other visualisations present routes that were taken, fitting the category 'Spatial aspect of trajectories (Vm2)', present activity over roads over a timeline, fitting the category 'Temporal display (Vm5)', or views including both. Finally, the document does not display contextual data. Do note that our taxonomy aggregates elements of documents classified, i.e. separate visualisations, operations, movers, tasks.
The development of the taxonomy and the organization of the contributions allowed us to make a number of observations:

- Tasks 'Characterisation of locations (T1)' and 'Characterisation of mov-ing objects (T3)' are mainly discussed while using the visualisation method 'Spatial aspect of trajectories (Vm2)'
- Most moving entities discussed in contributions are 'Independent movers (MDC2)'
- Contributions discussing 'Context (CTXT)' are in documents discussing (T1), indicating the usefulness of displaying contextual data for providing richer semantic context.

But these observations have to be considered carefully, as the over-representation of '(MDC2)' could indicate a lack of diversity in our population, making the emergence of strong links less likely. Furthermore, all cases of 'Entirely constrained (C3)' are linked to 'Dependent movers (MDC3)', potentially indicating combinations of these WHAT elements that are not separable, and thus flaws in this structure. In retrospect and in light of the data, the taxonomy produced suffered from several issues:

- The categories used to build the taxonomy were set up at a mid-level of abstraction. While there is nothing fundamentally wrong with this approach, fitting some contributions into them was not easy due to this choice.
- This classification aggregated different visualizations within one contribution (scientific paper/article) into one entry. That choice was made for readability of the taxonomy, but valuable patterns might have been lost due to this approach, and either specific entries for each visualization of the contribution or a classification that indicates several visualizations displayed in one dashboard.
- The characterization of the WHAT category showed that most likely some categories were not completely independent. Ability to take decisions is particularly intriguing, as discussions with

experts in different fields or papers underlined the importance of that aspect. But in practice, discussions of that aspect were important to understand the data more clearly, but also were put forward by researchers as these cases indicate two issues; these events could show limitations of their designs, but that assessment is hard to make as it is not clear yet for the researchers how much that contextual information is to be considered as exceptional or systematic.

- Context is an important aspect of data analysis, but as we aggregate the visualization methods together for each contribution, the details of where its additional display was deemed most important is lost. Future work to extend the framework to account for context would be required to specify these details.

It is important to explain the influence of the conference in which the poster was presented over our later work. During the IEEE EuroVis 2019 conference, we discussed with several researchers about the concepts presented in our taxonomy, and about the potential approaches to improve the blocks used to built it. Our poster drew a certain interest, and discussions with fellow researchers were strongly beneficial to identify clearly the strengths and weaknesses of our approach, and even more importantly, what drew researchers' interest towards this contribution. While we are aware that informal conversations during conferences can only be accounted as anecdotal evidence, we can not underline enough how valuable the conversations with fellow researchers bringing fresh and diverse ideas were. The discussions with a range of people working in this area indicated a particular interest for a system to compare contributions about the analysis of trajectories and space-time information. This was not a surprise, as comparison of contributions had already proven to be sometimes complicated at the beginning of the thesis, but such discussions reinforced our confidence that the problem was common. Additionally, these conversations were valuable to reflect upon the level of abstraction that would be most appropriate for the building blocks of a theoretical structure. Particularly, discussions with fellow researchers were valuable concerning connections between the blocks of Card [39]. An interesting suggestion was that all the elements of a theoretical framework would be much simpler to understand, connect and compare if they shared as many elements as possible, i.e. the characterization of information should be based on universal notions between blocks wherever possible.
The taxonomy was thus an important step in our effort to understand the visualization field for analysis of trajectories and space-time attributes. But more importantly, reflections over the taxonomy's strengths and weaknesses and discussions about it indicated the needs for a structure that would allow the comparisons of diverse contributions.

3.2 Data characterization

An important aspect of the SFNCS is the formal characterization of data and its categorization. Data characterization is the enrichment of data with new information that defines it. Data categorization is the process of grouping data with common characteristics or features that are likely to influence

visual design (HOW) and be important to task performance (WHY). This distinction is important as data characterization is necessary to build our framework, and data categorization is used to evaluate visualization contributions. We discuss data categorization for our studies in section 5.1.

The SFNCS is the result of both the development of our taxonomy described in section 3.1 and the following reflections afterwards which pointed us to the model of Card *et al.* [39], which we discuss in section 2.1.1. The model of Card is a mapping of data to visual form to perform visualization tasks. This approach has been successfully used in frameworks [71] within the academic environment, but also for commercial products such as Tableau [173]. Still, we argue that further enriching this approach with a new framework that would incorporate a formal characterization of the elements composing it would prove valuable. In this section, after presenting our reasoning for that use, we describe the elements which will be used to characterize our framework.

3.2.1 Motivation to characterize data

Characterizing data according to the information it originates from is not a novel contribution, and there are several approaches to do so, as discussed in section 2.1.1. As we present a novel approach towards presenting studies for researchers to discuss and report their protocols, it is legitimate to question whether that approach is valid and a positive contribution. In this section we present the reasoning behind our approach, its expected strengths and the limits that may hinder its adoption.
The reference model of Card [39] links Data (Raw Data and Data Tables) to Visual Form (Visual Structures and Views) to perform a Task, with these elements potentially modified through Human Interaction. An important notion that is not discussed when presenting this model is who is the person responsible for the series of choices made to select and modify the data, select and specify details of the visualizations, as well as the interactions executed to perform the task. Is it a single person, e.g. a researcher exploring a data set of interest, or is it the product of the work of several people, e.g. a data analyst prepared the data set, for which a designer developed a software that will present visualizations, from which a field expert can perform the task they are interested in? This interrogation is important because a vast part of research contributions is produced for the second case, for which we need to assess the value of the visualizations contributions and their impact on efficiency, accuracy, confidence, for different tasks. Furthermore, contributions should be comparable to ensure a nuanced understanding of their strengths and weaknesses.
To allow for interdependent comparisons between combinations of elements of the reference model, researchers should use the same characterizations of information amongst the blocks wherever possible. Our framework is thus built upon a consistent usage of data characterization. We selected the geo-temporal framework WHAT-WHERE-WHEN proposed by Peuquet *et al.* [139] to build it. With their framework, we can consider data according to its qualities, its geographical attributes, or its temporal attributes. How their model fits into our framework is detailed in section 3.3. In this section,

we discuss the motivation behind our approach of characterizing and categorizing data, the versatility of the WHAT-WHERE-WHEN framework, and how its elements can be used to categorize data, using different levels of granularity, similarly to [134, 63, 153], at different levels of abstractions.

By characterizing data complexity, our objective is to facilitate the comparison of different studies run with different data sets, and to enable the comparison of visualizations efficiency for similar data sets. The assumption behind that approach is that the more complex the data is, the more complex the resulting visualization will be.

This assumption is true as long as the higher complexity of the data results in more visual complexity for the participants of the studies performing the tasks asked of them. But this statement has to be nuanced with the influence of data operations such as aggregation and clustering methods that imply simpler visual stimuli but loss of detailed information. We thus consider that the assumption holds true if the data discussed is the one directly presented to the user, not before it undergoes transformations. Furthermore, for the assumption to hold true, the researchers must select the right measurement which correlates to perceived difficulty for participants to perform tasks. Within this thesis, choices have been made to select characteristics as measurement units for complexity, discussed in sections 3.2.2 and 3.2.3.

An important point to consider is that our characterization is here based on attributes that can be measured, e.g. 'average sinuosity of a trajectory between 1.25 and 1.5'. It is necessary for the construction of our framework to only consider characterizations that are based on measurable elements to avoid different interpretations of meanings. If researchers try to characterize terms that require human judgement, e.g. 'while the value is high', differences of opinion on what 'high' means can arise. This distinction between type of information is discussed in section 3.2.4.

When setting the characteristics that define the complexity of the data presented to their participants, several choices are made by the researcher:

- The researcher has to consider which characteristic or set of characteristics that define the complexity present in the data and how this characteristic can be captured in a manner that is indicative and usable.

- The researcher can define the complexity of a type of data thanks to combinations of different calculations, or decide that it is necessary and relevant to keep those methods of calculation distinct, e.g. if two measurements derived from the quantitative data are deemed relevant by the researcher, they can decide not to aggregate the resulting categories.

- Once characteristics are set to order data complexity, the researcher has to consider how these are going to be used to organize studies structure and communicate the results. In practice, that choice is important because it has consequences over the structuring of studies: if the researcher splits the abstract space of complexities into a high number of categories, it will result in a high amount of categories to evaluate separately if the researcher intends to evaluate the influence

of each axis individually. Simultaneously, cutting the abstract space of complexities into a large number of categories will result in more precise claims. The researcher thus has to make practical choices adapted to constraints due to real-life, such as funding and time available, to make valid selections within the complexity space that is linked to interesting claims, e.g. evaluations of use-cases similar to real-life cases. We expect that as future studies are developed and run, precision of the ranges selected can be increased, learning from previously published results and their associated data sets.

In the following subsections, we will discuss the choices we have made regarding the data characterization process and the selections of ranges we made.

3.2.2 WHAT: quantitative and qualitative attributes

The WHAT element of Peuquet's framework relates to qualities and quantities that are independent on space. When discussing WHAT elements, we discuss diverse attributes in a wide variety of contexts, be it a constant or a variable. Attributes can be considered constant or variable depending on context, e.g. height of a mountain is considered a constant in most cases, but over very long time frames is variable. The WHAT element from Peuquet's framework should thus be considered according to the context it is analysed in. We referred to diverse attributes when discussing WHAT. Those are qualitative (sometimes called categorical or nominal) be it binary or polynomial, ordinal (where the values are discrete, but order matters), quantitative (aka continuous or real-valued). Card [39] claims that the most important distinctions are those of level of measurement [**?**]: whether data are nominal, ordinal or quantitative. This distinction is critical for data evaluation (WHY), data representation (HOW) and the subtle relationships between the two that we seek to describe and explore. We thus invite fellow researchers to adapt their categorizations of attributes according to whether they are qualitative, quantitative or spatial, as WHAT_Ql, WHAT_Qn and WHERE, with the possibility to enrich the categorization accordingly to more specific details, e.g. WHAT_Ql_Nom, WHAT_Ql_Ord, with Nominal and Ordinal being sub-categories of Qualitative data. We do not consider that the specifications of sub-categories should be communicated if the claims researchers wish to make are valid for all the sub-categories. We considered a distinctive group for binary data as it commonly discussed within this thesis, but decided against as the implications of dealing with binary or ordinal data are minor, with the only distinction being the number of elements characterizing an attribute being 2 or more than 2. Communication and analysis of nominal data is relatively similar to binary data, oppositely to ordinal or quantitative data. In this section, we discuss the characterization of nominal qualitative *WHAT_Ql* and quantitative *WHAT_Qn* data, as it was the focus for the studies we planned to later set up.

An important aspect of our work was to consider a manner to characterize the complexity of data used to generate the visualization displayed to participants of studies. We are not experts in regard to various mathematical formulations that could be used to define complexity, e.g. Kolmogorov-Chaitin Complexity, Stochastic Complexity, Statistical Complexity or Structural Complexity [57]. Our lack of

expertise on these approaches meant attempting to use them in a new visualization context presented risks for the validity of our claims. As we are aware that potential future work might present new usage of such methods in a visualization context, we wish to state again our invitation to make data open source to make it possible to generate new metadata out of previous studies. We intend to later publish the work resulting from this thesis, and will then make our data and our code to analyse it open source.
Thus our search for parameters to define the complexity characteristics was limited to contributions in the visualization research domain. We searched for these characteristics solely in publications related to the visualization field that discussed elements that influenced perceived complexity once the information is displayed.

WHAT_Qn

Many parameters are potentially influential concerning the perceived complexity of a quantitative attribute, but literature rarely discusses it. Rather, we found during our literature review papers discussing complexities of lines in two dimensions representing trajectories, which present some similarities to quantitative attributes. The contribution that fit closest to our aim to characterize the perceived complexity of quantitative attributes in two dimensions was the graphical perception study run by Perin *et al.* [136] in which they define the following parameters to characterize a trajectory complexity:

- The tasks participants are asked to perform (WHY?)
- The 2D path of the moving entity, including its curvature, direction, length, range of angles (abrupt changes of direction), and crossings (trajectories going over positions previously visited)
- The time function (ranges of speeds and time distributions).
- The background, with its colour and texture (e.g., a map) (CONTEXT)

Their characterisation of a trajectory complexity inspired us for our approach to characterize quantitative data, but we deviated from it for a series of reason which we present before discussing our approach to characterize quantitative attributes. While we agree that tasks are likely important on perceived visualization complexity, we argue that tasks should be separated from characterization of the data and visualizations as data sets and visualizations can be produced for various tasks, or one task can require several data sets or visualizations, implying that a framework assessing their relations need to consider variations separately. Task will thus be part of the SFNCS, but separated from characterization of the attributes. Time is not an attribute unique to the moving entities that can be assessed by researchers, which drives us to consider it is an attribute that influences perceived complexity of the quantitative attributes but it is not an appropriate characteristic. The background is an element that can influence perceived complexity of a line, either helping by providing context

or hindering by reducing readability, but it is an attribute that is exterior to the information the quantitative attribute represents in itself, and thus is not a relevant characteristic. The remaining element that can be used to categorize the quantitative attribute is thus the 2D line (path is only a relevant term for a trajectory), which itself is defined by several elements. The approach selected by Perin *et al.* [136] was to draw straight and curved lines, with using B-splines with six control points to produce a smooth curve with randomly assigned y values within a certain range and a fixed x increment between the points. The output from their trajectory generation is thus very similar to a line describing the evolution of a quantitative attribute over time. Due to the characteristics of the quantitative attribute, some expectations can be made about the 2D line since we do not consider set-valued analysis [22], and thus expect a unique value produced for each quantitative data point, i.e. no crossing will be present in the data.

Priority was set for characteristics that seemed to be most important and would have the higher influence as it varies. While other factors can be influential, our reflection was built on two considerations that justify prioritization over attempting to directly bundle together all potential contenders to relevant characterizations. First is that assumptions require real-life verification to be confirmed, and while motivated by experience and readings of previous literature, some choices have to be made based on researchers' expectations. Second, realistic constraints have to be considered when setting up evaluation structures: if too many characteristics are set as variables, evaluating becomes particularly complicated, while studies that have a limited number of characteristics are doable and allow the presence or absence of particular effects to be established across meaningful and differentiable conditions. It is our expectation that the knowledge generated from studies in which researchers use characterizations they deem fit for particular purposes, will allow characterizations of the attributes of the framework to be refined over time, as the impacts of particular attributes on particular tasks are understood. Originally, the motivation to consider the notion of noise corresponded to the observation that lines that were varying a lot '(range of angles, abrupt changes of directions)' were more difficult to be assessed. We thus aimed for a calculation method that could be used to characterize and categorize quantitative data.

We investigated for metrics that would effectively capture this loose notion of spatial complexity. Metrics that were considered as potentially suitable included range of angles, straightness and sinuosity. Li *et al.* [104] list a series of features to characterize movement, which can be used for quantitative data. The list is composed of speed, acceleration, turning angle, and straightness. The straightness appeared as a valid measurement to characterize quantitative data. It is defined as:

$$Straightness = distance(p_{i-1}, p_i) + distance(p_i, p_{i+1}/distance(p_{i-1}, p_{i+1}) \tag{3.1}$$

This was our first approach, as using the mean of that value could describe an overall amount of changes, instead of a range of changes. But at the time of implementing the calculation in our code, we noticed a potential issue: the definition does not clearly consider whether the differences in indexes should be made based on indexes alone or account for distances variations. We ran tests with

straightness calculated based on index or distances and noticed different values, which was expected and tolerable, but also noticed that certain selections of index or distance differences resulted in different straightness values of lines. Such cases were rare and occurred when the points p_{i-1} and p_{i+1} were relatively far from each others. While we can expect researchers to select such differences adaptively to the context of the studies they wish to run, this indicated that straightness would produce different categorizations according to researchers' set up, which would hinder transferability of knowledge generated between future studies using our approach. The choice was thus made to use sinuosity, which is not going to vary according to researchers' input for calculation. Sinuosity is a commonly used calculation to describe rivers [155] and is calculated as such:

$$Sinuosity = \sum_{i=0}^{n-1} distance(p_i, p_{i+1}) / distance(p_0, p_n) \tag{3.2}$$

We first hesitated to use sinuosity as the method to characterize quantitative attributes complexity, due to its origin being made for a specific geographical task and not intended as a common calculation to use for quantitative attributes. But the same considerations that lead us to disregard straightness made us reconsider that decision. Sinuosity is scale independent and appeared as a valid candidate to characterize complexity of quantitative data. That statement has to be nuanced with three points:

- First, sinuosity is a measurement that is scale independent but for which the interpretation varies according to context, e.g. if looking at two lines that are different entities but share the same sinuosity, e.g. one for a river and one for a train line, over the same area,a user would likely consider for the same value to be more or less important due to context, as train lines tend to be less sinuous than many rivers. The same could be said of two quantitative attributes with different contexts. Interpretation of the sinuosity in accordance to its context is thus still present.

- Second, sinuosity is only valid for as a measurement in a Euclidean space.

- Third, sinuosity is to be considered as a valid metric for characterization of quantitative attributes only within a set with diverse variances. This point is motivated by the observation of extreme cases, such a line with a very high sinuosity but a very low variance will be most likely still easier to read than a line with a very low sinuosity but a high variance. This observation has to be nuanced as it is dependent on context and has to be considered as an ensemble of sets of quantitative attributes, not one quantitative attribute on its own. The observation we made about the extreme case is problematic if the two cases are considered using the same axes to scale them. Within the context of a study, it is commonly expected that these don't vary unless indicated and accounted for by researchers.

WHAT_Ql

The same approach drove our selection for the qualitative characterization. Several characteristics could be considered as potentially valid for our framework. As stated previously, we focus on characterizations that are likely to have a strong influence over perceived complexity, considering that future work can refine these elements of the framework. Qualitative attributes are defined by a limited number of potential values present in the data. As we can only hypothesize the influence of data being nominal or ordinal over perceived complexity of qualitative data, we aimed to use a method that could be valid for both subtypes of qualitative data.

The number of measurements that seemed relevant to us to characterize qualitative data was limited. Overall, our expectations of data complexity resulting in perceived complexity were based on the assumption that numerous changes of status in a small time frame were difficult to consider in detail. This loosely defined notion drove us to consider calculations based on frequencies of changes of status of the qualitative data. We first considered how to measure the distributions of the variations as potential complexity measurements. But due to complex variations being only possible with a relatively high number of variations, we concluded that these were consequences of qualitative complexity rather than cause. If you consider over a set number of qualitative data points that the changes are responsible for perceived complexity, two types of situation can occur: changes can be relatively equally distributed or concentrated in a section of those points. It can be argued that the first case is more complex as overall likeliness to consider a potential change is relatively high overall, but oppositely with the concentration of changes in a small section, one large section will be relatively less complex and one small section will be very complex. This implies that these two approaches are but different distributions of the complexity, meaning that the complexity is not the frequency of changes, but the number of changes over the number of points itself. We thus selected the number of status changes over the number of points as the characterization of the qualitative attributes, commonly called frequency.
In section 5.1 we discuss the characterization of the data used for our studies according to frequency.

3.2.3 WHERE: Trajectory complexity

In this section we discuss how we selected the characterization of trajectories. The process to select a method to characterize trajectories took place simultaneously as we worked on characterizing the quantitative and qualitative attributes. We thus invite the reader to consider both sections as part of the same effort to characterize perceived complexity while accounting for context.
Looking at spatial information, potential elements influencing complexity are to consider from both the environment and the moving entities themselves. Due to technical constraints and time, evaluating all the elements that could potentially be interesting as characterizations of trajectories was not feasible. We thus aimed to describe characterizations that were likely to be most influential. Following this decision, within the scope of our thesis we only consider contributions in which the background that conveys geographical contextual information, i.e. the map, is not changing and only

representations of moving entities are evolving. We suppose that the influence of the background over trajectory complexity is rather due to the geographical constraints it contains than the method to archive contextual geographical data. Furthermore, our decision is justified by the consideration that movement is influenced by constraints of real-life, but according to how data is processed that importance might not be accounted for, which reinforces our confidence that while trajectories are dependent on real-life constraints, the data worked with might not be under their influence. Still, future work is required to formally support this supposition. This decision reduces the number of dimensions to consider the characterization of the trajectory.
To understand the selection we made for data characterization, we discuss the elements that also seemed valid and were disregarded following further reflection. Part of our reflection originates from considerations of how the parameters could be evaluated to justify the selection validity.
Taking the list of factors listed by Perin *et al.* [137], we considered several combinations to define the levels of granularity for the categorisation of the spatial data. Following a reasoning similar to the one for the selection of characterization of the WHAT elements, our first approach was to consider the 2D path as the only controlled variable to characterize the trajectories. Straightness, as defined by Li *et al.* [104], seemed a valid trajectory characterization measure, as it fits the assumption, shared by Perin *et al.* [137] that a curved line was more complex than a straight one. It thus seemed like it could be used as the qualifier of complexity for our study. But further reflections drove us to nuance this approach. Our first approach was to consider straightness as a measurement to characterize the trajectories. As stated in section 3.2.2, our selections for characterizations were based on assumptions over the elements that will likely be the most influential to the task. This means that some other characteristics then have to be either ignored, constant or randomly varied as long as it is diverse enough. We thus discuss other parameters and justify our choices on whether incorporating them in the characterization, keeping them constant, or ensuring they are distributed evenly. Particularly, we discuss the influence of direction and length. Our reflections about length drove us to set up an exploratory study to evaluate perceived complexity that pivoted our approach. Elements that we do not discuss in detail are assumed to be randomly varied.

Direction is an element that differentiates a line representing a quantitative attribute from a trajectory. The display of the quantitative attribute is dependent on the order selected to display it, most commonly in Western cultures with time evolving from left to right, resulting in a single direction to display evolution of time no matter the quantitative attribute. Trajectories are influenced by parameters of visualization methods selected, e.g. scaling or rotation of maps, but they are not limited to a single direction. Direction should thus be considered as a potential valuable characterization of the trajectory. Direction can be considered with several levels of detail, i.e. a series of turns or a series of angles changes according to the order of the points in which the moving entity movement was recorded. We did not suppose that directions were likely a strong influence over perceived complexity of trajectories that would differ from sinuosity, but wanted to assess whether spatial directions influenced tasks requiring the temporal aspect of trajectories to be considered, e.g. is the same trajectory going

from left to right is as complex if going from right to left. We thus decided not to use direction to characterize trajectories, but in our studies flipped the stimuli for some answers to establish whether direction could later on be a valid characterization of trajectories. More details about direction are discussed in section 5.4.

Another important element to consider was length. We were aware that it was considered a potentially important factor for the characterization of the complexity of the trajectory, as mentioned by Perin *et al.* [136], but how important was that impact was not immediately clear to us. We considered that the perception of length influencing trajectory complexity had to be nuanced. We wished to understand whether length was influential when differences were relatively minor. To do so, we set up a study to refine our understanding of perceived complexity of trajectories.
At the time of setting up this study, we had already decided to use the IEEE VAST 2014 challenge data for future studies, as the trajectories generated for this set were recognized as realistic but did not present a risk of confounding effect due to participants' personal contextual knowledge that can occur with real-life trajectories data set. We thus used the IEEE VAST 2014 challenge data, which contains trajectories of a series of cars moving in a fictitious city. The trajectories are constrained by a block-based road network and vary in terms of sinuosity and length. This provided a realistic sample of trajectories, that had been derived for and used in spatio-temporal visualization, with a diversity of shapes, directions and lengths. We asked participants to indicate, when presented with a series of couples of trajectories juxtaposed, which was more complex. Participants could say which of the two trajectories seemed more complex, or indicate if they did not notice a difference in complexity. An example of a question asking to compare the complexity of two trajectories is displayed in Fig. 3.2. A convenience sample of 63 participants answered our survey, each generating 30 comparisons, from 14 different trajectories selected to generate a set with diverse sinuosities and lengths, resulting in 91 different combinations of comparisons, for which each was compared the number of answer. An example of stimuli is illustrated in Fig. 3.2. The participants were not expected to have particular skills or competencies in data visualization. The results are available here: https://tinyurl.com/trajscomparisonresults and displayed in Fig. 3.3. This allowed us to order the trajectories based on perceived complexity. Using this new baseline allowed to consider differences between factors and combinations of factors. At the time of running this study we were still considering straightness as a measurement for perceived complexity, before pivoting, as detailed in section 3.2.2 to sinuosity. When analysing the results of the study, we noticed that ordering the trajectories according to the participants' perception of complexity varied from the one made based on the computed trajectory straightness metric. To understand the meaning of the output on the ordering of trajectories according to complexity, we initiated informal discussions with participants after the study and asked them whether they followed a certain logic that they could describe to make their choices. The output was clear: participants, without receiving any instructions, often relied on counting turns and looked a length of the trajectories. Some mentioned that they first considered complexity comparison without using any specific methods, considering whether they could rely on a

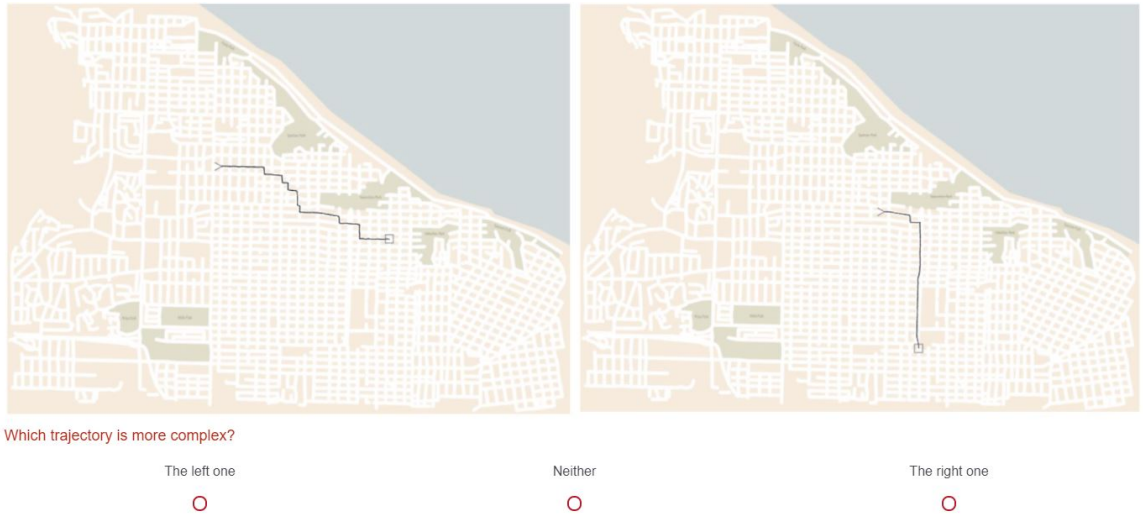


Fig. 3.2 An example of the stimuli presented to participants asking them to compare the complexity between two trajectories. This design, which consists in indicating the beginning of the trajectory with an arrow, and its end with a square, was first explained to participants. The design was later slightly updated, as described in section 4.3, but we used the same trajectories.
No criteria of complexity were communicated to the participants, since our objective was to use their answer as a base to consider characteristics of the trajectories that were influential over perceived complexity.

measurement as complexities differences were not intuitive anymore, while others started immediately by using a method and sticking with it for the duration of the study. Additionally, when asking participants about lengths of trajectories, participants responded that they either disregarded it or use it as an alternative if shapes didn't convey a strong difference between trajectories, in which case they would set the longer trajectory as the more complex one. Following these discussions and accounting for the characteristics of the trajectories, we consider that the most important characteristic of complexities for the trajectories presented are the number of turns, followed by length of trajectories.
We should note that due to the IEEE VAST 2014 challenge data set being a representation of cars inside a city, with sharp turns and mostly rectangular buildings between the roads, it is possible that the common approach is an artefact of the clarity of the number of turns, compared to potentially less constrained trajectories, e.g. birds flying. Our claim of universality for our method to characterize trajectories is thus nuanced. Still, we consider that the method of characterization we selected is valid when displaying positions of moving entities constrained inside a grid-like environment, but note that future studies set up to compare trajectories complexities evaluating more parameters and more types of moving entities would be a valuable contribution.

In section 5.1 we discuss the characterization of the data used for our studies based on the number of turns and lengths of the trajectories.

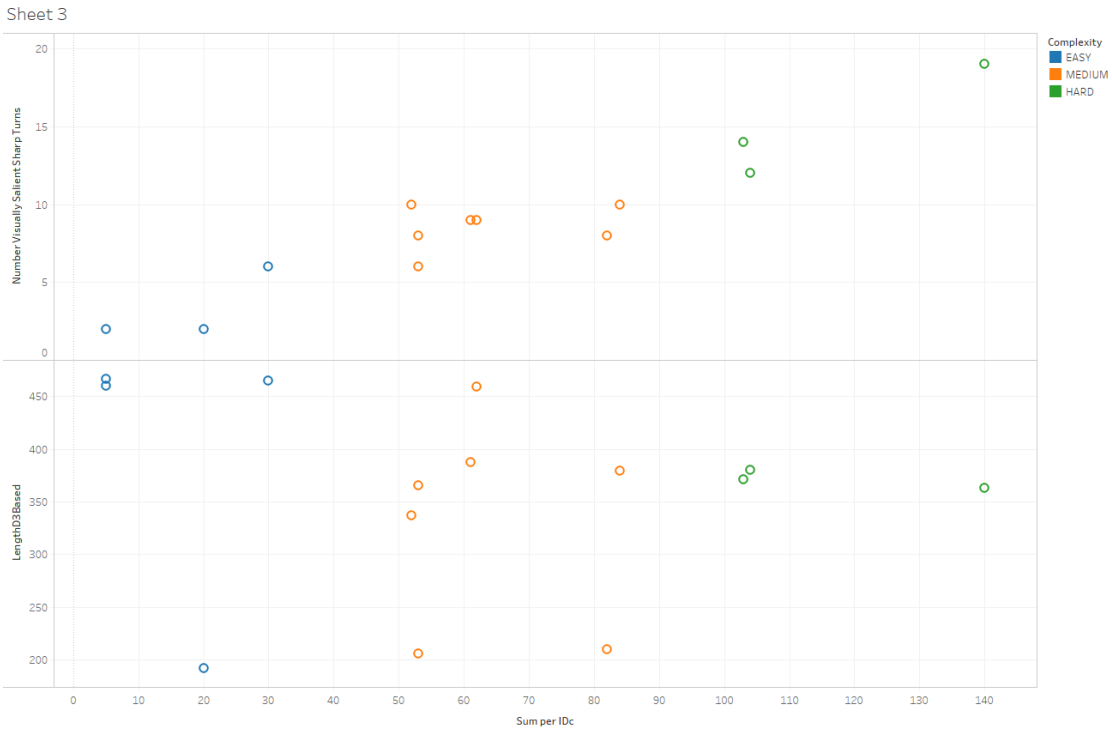


Fig. 3.3 The results of our trajectory complexity comparison study. For each of the 14 trajectories compared, a dot represents them. The dots are horizontally organized according to the number of times they were deemed more complex than their peer (combinations of comparisons were balanced), as illustrated in Fig. 3.2. The dot plot at the top vertically orders the trajectories according to the number of turns of the trajectories and the bottom one orders them vertically according to their length. The results of the study and informal discussions with participants informed us that length of the trajectory were used as comparison method if shapes did not convey a strong enough complexity difference. These results informed us how to categorize trajectory complexity. Here, the dots representing them are coloured according to that categorization. The details of the data categorization are presented in Fig. 5.2.

3.2.4 Measurement and Judgement

As we aimed for the SFNCS to allow the characterization of studies with high or low levels of abstraction, also often quantitative or qualitative studies, with the first ones being potentially influenced by contextual uncertainties and participants' interpretations, we deemed it necessary to characterize the differences between information presented as objective or subjective.

The very terms of 'objectivity' and 'subjectivity' have to be scrutinized. The notions of objectivity and subjectivity are critical for the interpretation of information. The terms objective and subjective are defined in the Cambridge Dictionary as 'based on real facts and not influenced by personal beliefs or feelings' [Press] and 'influenced by or based on personal beliefs or feelings, rather than based on facts' [142].

We consider that objectivity and subjectivity are not notions separated by a unique separator, but instead consider that subjectivity is the result of different steps in the process of interpreting and communicating information. To communicate information clearly and efficiently, indications of human judgement in the process of information generation is important for various reasons, such as trust in the information retrieved or reproducibility of data analysis processes. The transfer from information as a fact from how the world around us exists to data can be done in various manners. Measurements can be produced using tools, for which the measurements are independent of each other, e.g. thermometers, rulers, or assessed as information that is context dependent. That difference is important as the uncertainty, and by extension the trust that can be associated with information, will vary according to that process of transferring information. That difference has to be nuanced, as communication about the generated information is important. If a measurement is done at a time T by a thermometer that gives a value of 23.12 degrees Celsius, it is an objective statement to declare that the measurement made at the recorded time T was 23.12 degrees Celsius, but it would not be true to claim that the actual temperature is 23.12 degrees Celsius, as the measurement process is subject to precision and accuracy. Language is often shortened for comfort of communication, with the notion of how much information retrieved can be shortened being selected according to the necessary need, i.e. if it is 'good enough'. That notion is subjective and context dependent, albeit originating from an objective measurement. Subjectivity thus brings its own type of uncertainty, but as stated by Meyer *et al.* [123], its presence is not be considered as a lack of rigour. For complex problems that are usually tackled by experts in a specific field, it is common to present information within a context deemed appropriate or respecting the regular conventions they are accustomed to. Furthermore, sentences and questions are complicated structures that can incorporate various levels of subjectivity into a single phrase [66]. Any part of a sentence that would deviate from the most possible objective statement would by definition be subjective, e.g. the choice to communicate a measurement with a lower level of precision than the record for clarity changes an information from objective to subjective. As communication over information is thus likely to alter the status of its objectivity, establishing what absolute objectivity is can become surprisingly complicated. If we keep our example of a machine measuring temperature at a certain point in time, many information

points are likely not to be communicated due to expectation of lack of usefulness, e.g. how many digits of precision can the machine produce, what is it accuracy levels, how these were calculated and tested, what year was the machine created. These issues can add up as we consider more dimensions to that information. Spatial or temporal attributes carry their own information that are likely to be changed from objective to subjective, e.g. the level of precision of for spatial records or time records are subject to the same issue. Expectations due to context imply that absolute objectivity is hard to define and categorizing information as objective is in itself a subjective process, as the evaluation of what 'all the information present that can be reported' is a human judgement in itself.

Thus, when we discuss the notions of objectivity and subjectivity, we want to underline that we are using such terms in a manner that implies reasonable assumptions due to the context. To ensure clarity when discussing these notions as we characterize the elements of SFNCS, we thus characterize these as Measurement and Judgement.

The separation between the two categories depends on whether the information from the question can only be defined with terms that require human judgement, e.g. the term "hot" is only true if a human considers that categorisation to be true within the context. This concept, while easy to qualify when scientists think of one single element of the study they plan, can become confusing when some conventions are widely used and thus expectations blur the judgement of scientists while they set up their studies.

Coining the terms was not an easy process, as we could not find a word that was perfectly matching our intention for what we ended up naming Measurement. Luckily, the term Measurement is already used within the scientific community, and its meaning is close to what we mean to express. The main difference when discussing our definition of Measurement and the common definition of measurement is the necessity for the measured information to be quantitative, e.g. width with a ruler, while our definition of Measurement also includes the recovery of qualitative information as long as it is not subject to interpretation, e.g. in a football match, the number assigned to a player to identify them does not change and is not to be interpreted in various ways, unless mistakes are made during the recovery of that information.

Following upon our previous example of a machine making a measurement: absolute objectivity is hard to achieve, as there is much information to convey for a statement about a measurement to be entirely objective, if that is even achievable. We thus consider that Measurement is a characterization of information which did not imply subjective judgement from the person gathering or communicating the information, while some information can be disregarded if we consider it presents no interest according to the context (e.g. rulers haven't changed for years and thus not communicating year of production of the ruler used for a measurement is acceptable for a Measurement).

Do note that this notion is potentially context-dependent in some cases, depending on the information set considered. An interesting example is colour: 'red' can be considered as a qualitative attribute

that is going to be interpreted by all participants without vision impairment as unique information for which distinction is easy, given the values separating it from other colours displayed being large enough, i.e. compared to clearly distinct colours, e.g. 'blue' or 'green'. But that information can also be considered based on quantitative attributes bundled together, e.g. a colour encoded in a computer program as rgb(255,0,0,1) would be interpreted by virtually all the people without vision impairment as 'red', but within that approach to consider information and data, it becomes one 'red' out of a series of potential values that could be labelled as 'red'. This one example value is surely always going to be considered as a 'red', but the same approach would likely result in different colour names if displayed a colour encoded as rgb(255,64,25), called Orange-red, and thus becomes subject to subjectivity.

We thus define the two terms as thus:

- *Measurement*: the characterization of information that is independent on human contextual enrichment.
- *Judgement*: the characterization of information that is dependent on a human enriching the information, due to their contextual knowledge.

Note that we use a capital letter when using our definition of Measurement and Judgement.

Our motivation behind the settings for our definitions originates principally from the importance of various interpretations of data according to whether it is a human that potentially was influenced by context when making an estimation of information versus an information that would result in a similar value if performed in different contexts, i.e. a measurement. As stated previously, we are aware that the interpretation of different measurements is still possible, particularly when measurement methods differ [120, 121]. But the differences between the interpretations are dependent on the users including contextual information (or lack thereof) in their analysis.

The consequences of the separation of that notion spread through several elements of the SFNCS. Information that can be characterized as a measurement can be used to generate baselines for assessing performance in studies. That approach is common in quantitative studies [134, 137, 73]. Many studies evaluating visualizations ignore the evaluations of judgements. But a critical notion to consider is that in cases where a baseline can be computed, visualization may not be a necessity. If a baseline can be generated, the value of the visualization is then limited to verification of the information presented, potentially reinforcing trust in the computed output. We argue that visualizations are more relevant for tasks that are based on judgements. While the quantitative evaluations of visualizations allow for strong claims about ability to perform low-level tasks, the assumption is often made that the low-level tasks are intermediary steps necessary to perform high-level tasks, albeit the steps and reasoning occurring to perform high-level tasks are unknown.

We argue that the incorporation of characterization as either Measurement or Judgement in future studies will allow for a less opaque separation between qualitative and quantitative studies and nuance the validity of claims over contributions for tasks with various degrees of levels of abstractions. Our

framework is thus designed to incorporate the characterization of questions and answers that are generated during studies for the evaluation of various tasks, for which the final objective can be for the participants to find either a Measurement, e.g. maximum value among others, or make a Judgement, e.g. estimate whether the implications of an event justify actions, such as policy changes.

3.3 The structure of the framework

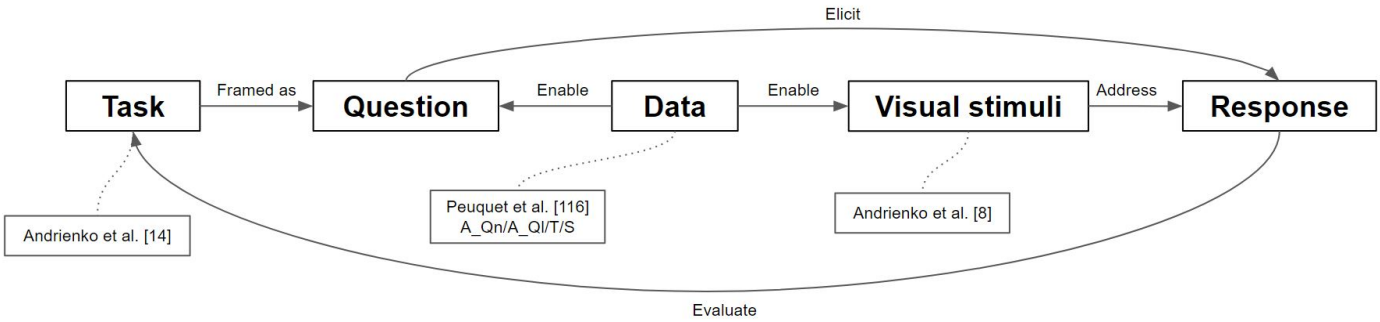


Fig. 3.4 The blocks of the Systematic Framework for N-scales Characterizations of Studies (SFNCS): connecting Task, Question, Data, Visual Stimuli, and Response into one structure from which characteristics transfer. The arrows indicate how selections within a certain category at its source will directly impact the element at its end. The SFNCS is built by connecting tasks as defined by Andrienko *et al.* [15], the framework for characterization of visual stimuli of Andrienko *et al.* [9], and data characterization as defined by Peuquet *et al.* [139]. We developed the blocks Question and Response to characterize the entire flow of studies set by researchers.

The framework we have developed is linking together five blocks – concepts that are fundamental to visualization, its design and evaluation. It is illustrated in picture 3.4. It is composed of the following blocks:

- **Data**: The digital recordings of a real life-event, transcribed into either position records, quantitative attributes and qualitative attributes, as well as times of their creations. Our categorisation of data is based on the model developed by Peuquet *et al.* [139], where they are defined as part of one of the following group: WHAT,WHERE,WHEN. This high level categorisation can be expanded into subcategories, for further details when that separation is important.
  Our thesis exemplifies the expansion into subcategories as we evaluate a novel design using our framework, for which we need to add the following sub-categories to WHAT: WHAT_Qn, the quantitative attributes, and WHAT_Ql, the qualitative attributes, which itself WHAT_Ql encompasses lower categories: WHAT_Ql_Nom for nominal data and WHAT_Ql_Ord for ordinal data.
  The design in our studies are not specific to the WHAT_Ql_Ord category and are thus not evaluated, but we underline the mention of the WHAT_Ql_Ord category, as we deemed its differences with the WHAT_Ql_Nom category strong enough to require a separation instead of

keeping the aggregated category WHAT_Ql. Such selections of data characterization are to be done by researchers accordingly to the contributions they wish to evaluate.

- **Visual stimuli**: characterization of the visualizations presented to the participant, using the framework of Andrienko *et al.* [9]. The characterization of the visualization is defined by the attributes displayed and how they are organized together, and can be described using pictographs 2.2. The pictographs are built using letters to indicate the type of information displayed, e.g. a T for time, O for object, A for a thematic attribute or S for space. We discuss their taxonomy in detail in section 2.2.
- **Task**: The tasks that users perform in a real situation. We categorize the tasks using the taxonomy produced by Andrienko *et al.* [15]. The same taxonomy is used as part of the Question block categorization. The details of this structure are discussed in section 2.1.2 and illustrated in Fig. 2.7.
- **Question**: The questions that are asked to participants are categorized with several parameters that we describe and define in this section. The Question block is made up of sub-blocks that indicate how questions are asked, how they relate to the task evaluated through them, and the information defining it. We further discuss the Question block and its sub-blocks in sec. 3.3.
- **Response**: the methods available to answer the question asked, once again, defined in this section.

In this section, we discuss the two novel contributions: the Question and Response blocks. According to Bertin [27], types of tasks are distinguished based on the type of information sought. Following that philosophy, we considered that our framework required a specific focus on the way the researchers ask the question in order to evaluate participants' abilities to perform the tasks of interest.

The Question block

When considering the evaluation of a contribution for a task, the questions are characterized to allow comparison between studies. Note that while we propose a systematic approach to characterize questions that supports transparency and reproducibility, we recommend (and indeed provide the flexibility for) adapting sentences so that they make sense for the purpose of the study that researchers wish to run, rather than using our own specific terms if others fit better. The categories we describe are meant as indicators for type of information first, and sentence structure only if it does not hinder sentence comprehension.

The categories of the questions are illustrated in Fig. 3.5, and examples of these categories being Measurement or Judgement are illustrated in Fig. 3.6. Question are split in three blocks:

- **Question form**: The type of questions depends on the task the researcher wishes the participant to partake in. It is defined by four criteria that define the task:

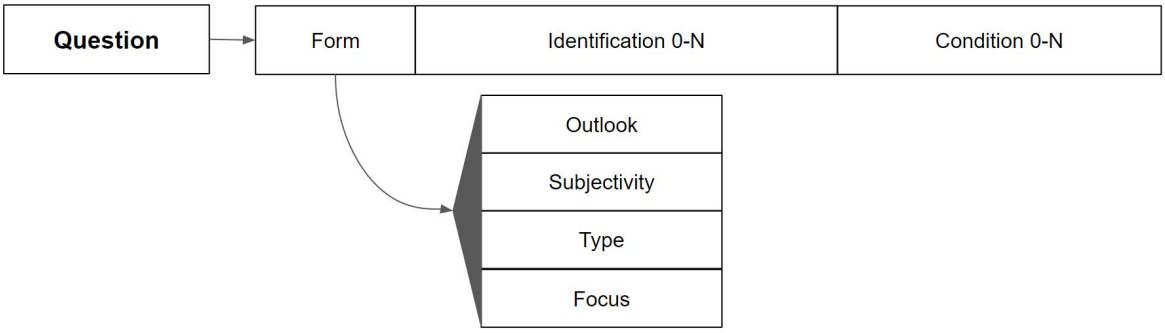


Fig. 3.5 The sub-blocks of the Question block. A Question is made up of a Form, 0, 1 or several (N) Identifications, and 0, 1, or several (N) Conditions.
The Form is defined by its Outlook, its Subjectivity, its Type and its Focus. Respectively, this depends on whether: the task is elementary or synoptic; the information requested is a measurement or a judgement; the task involves lookup, comparison, relations, or connections; and the information requested is quantitative, qualitative, or spatial.

- **Task outlook**: Whether the task is elementary or synoptic. This separation is defined by Andrienko *et al.* [15] as elementary tasks dealing with "elements of data, i.e. individual references and characteristics", and synoptic tasks dealing with "sets of references and the corresponding configurations of characteristics, both being considered as unified wholes."
- **Task subjectivity**: Whether the answer asked of the participant resulting from performing the task is a measurement or a judgement. The notion of measurement and judgement is discussed at length in section 3.2.4. The separation is clearly obtained when the researcher considers if they wish to ask for something where a baseline that's computable exists, e.g. comparison of two numerical values is a Measurement, while assessing whether an attribute is influenced by another is a Judgement.
- **Task type**: the categorization of the task that must be performed to answer the question. Our reference for these are the taxonomy of visualization tasks developed by Andrienko *et al.* [15], also used and illustrated by Aigner *et al.* [3]. It does not have to be the exact same as the Task block of the SFNCS, as it can be considered more valid for researchers to perform a global task by performing several intermediate ones. The list of tasks, discussed in detail in section 2.1.2, and illustrated in Fig. 2.7, is the following:
  - Lookup: tasks about searching for data values and searching for specific points in space and time.
  - Comparison: tasks about assessing the direction and size of a difference between several elements.
  - Relation seeking: tasks about search of occurrences between data elements with specific characteristics.
  - Connectional tasks: in the case of synoptic tasks, connectional tasks are about establishing connections between at least two sets with at least two variables

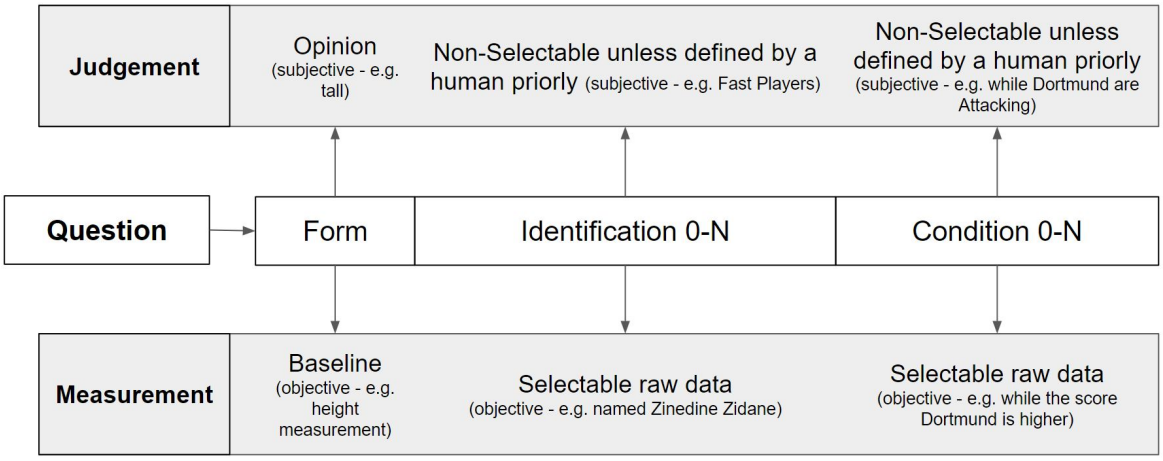


Fig. 3.6 The sub-blocks to characterize the Question block, where every information piece is either a measurement or a judgement. They are respectively used for identification as part of the question can relate to 0 to n conditions. A question can contain several identifications, e.g. comparing the number of passengers of two trains. If there is no condition, it implicitly means over the whole time displayed. The condition can be separated from the element of identifications, e.g. comparing the numbers of passengers in a train from 14:30 to 16:30, or have their own related conditions, e.g. comparing the number of passengers in train A from 10:15 to 10:45 to the number of passengers in train B from 18:45 to 19:15.

  - **Task focus**: the task focus is defined by the information type that the study participant should find, be it WHAT-WHERE-WHEN, or a sub-category e.g. WHAT_Qn.

- **Identification**: the Identification sub-block describes the data type(s) used to query the element(s) that need to be identified in order to answer the question, e.g. asking about maximum speed of a football player: the question needs to establish who/what the question is about, in this case a football player identified with their name, a qualitative attribute (WHAT_Ql). Similarly to the Condition sub-block, it is defined by some operants and can be either a Measurement or a Judgement.
  Some questions don't require identification (e.g. questions about global weather for an area don't require identification). The result from that identification can be a single element, e.g. if we need to identify a moving entity based on its identifier, then this question identification is WHAT_Ql resulting in a single selection, or if we select all moving entities within a certain range of latitudes, the question identification is WHERE resulting in several selections. Thus the type of query for the identification can be WHAT,WHERE, WHEN and resulting in 0,1 or several elements selected.
  Note that the number of identifications to generate a question can range from 0 to any integer, and due to N being commonly used as a mathematical annotation to indicate an integer, the same letter is used to indicate the range of potential identifications in a question, i.e. N-scale of the Systematic Framework for N-scales Characterizations of Studies (SFNCS).

An example of a question which requires at least two identifications is for comparisons between two different moving entities, e.g: asking about which player ran for the longest between Ronaldo and Messi during a game requires two identifications, both based on the name, a WHAT_Ql attribute.
Note that the type of information for Identification does not have to be the same as the Focus of the Form, e.g. for a question asking about the position of a football player based on their name, the Form Focus is WHERE and the Identification is WHAT_Ql.

- **Condition**: a condition is a filter over the data, defined by some operands. It can be either a Measurement or a Judgement, following the same logic as Identification and Task. E.g. 'precipitation levels greater than 1 inch' is a condition from a Measurement; 'going fast' is a condition from a Judgement.

The Response block

There are various ways to deliver answers according to the questions set, and these answer methods are options managed by the researcher. The Response block is dependent on how the questions are structured, and should be set up according to the approach the researcher thinks is the most interesting. As they define the combination of questions and answer tools for their studies, researchers should aim for a balance between allowing participants to express themselves and communicate their understanding or misunderstanding of the task, while adapting to their study structure, i.e. whether it follows a quantitative or qualitative approach. That selection is thus made by the researcher according to context specific to their study, but does not have to be unique for a combination of Task, Question, Data and Visual stimuli. At of the time of writing this paper, a complete characterization of answer methods is, to the best of our knowledge, yet to be generated.
As we set to develop a structure to characterize answer methods, we considered two approaches: search for answer methods from our literature review and specific searches, or find lists already established, albeit incomplete.
The issue with limiting the list to existing published literature were the lack of generalization and difficulty to consider whether the approach is specific and rarely used. This issue could be mitigated with the development of a taxonomy of answering methods from the literature, but technical constraints and time meant this option was not available. In parallel, using a list already created and used in a commercial software implied that the answering methods listed were sought after and thus relevant for the set-ups of studies, but such structures set with a focus on ease of implementation can separate answering methods that are conceptually similar. We thus decided to settle on a hybrid approach to list answering methods, albeit we acknowledge the development of a future taxonomy could prove valuable to redefine it.
The list of answering methods we set up is derived from the options provided by leading survey software system Qualtrics [28]. Some of the options offered by Qualtrics are too similar to consider them as separate entries for answering methods, and are thus aggregated in our list. Some of

these methods were encountered in literature we reviewed and are thus listed . Additionally during our literature review discussing spatio-temporal attributes analysis, some answering methods were encountered in literature [171, 77, 41, 17, 8, 44] but not Qualtrics. We thus added these Response methods to our list if they were significantly different from the others.
As we introduce the elements composing the list of answer methods, we discuss their strengths and weaknesses, considering their value as methods to gather data from participants in our context. The distinct answering methods are the following:

- **Multiple choice**: several options for a question for which the user can enter one or many entries. That approach is valuable for a fast answer, but that approach can force the participant to select a value by default rather than an actual match between the participant's opinion and the option. This answering method can include answer of various focuses, e.g. nominal, ordinal or quantitative, measurement or judgement, but is limited by the number of answers set by the researcher.

- **Text entry**: set a field for the participant to enter a detailed answer. This approach is valuable to get rich information from participants, but is strongly dependent on their ability to express themselves. This can be a long process that participants wish to avoid. Similarly to the *Multiple choice* option, the data generated can be of different focuses.
  Furthermore, once the answer is generated, if it relates to a judgement, the difficulties arising from interpretation is increased, as it is dependent on both the ability of the participant to clearly enunciate their thoughts, and of the interpretation from the researchers analysing the data. The analysis of text can also be problematic due to the difficulty to interpret a large amount of it, be it time to read a large amount of text or difficulties to summarize the information communicated in it.

- **Matrix table**: several questions and answers arrange in a common structure. The answers have to share the same sets of potential answers for the participant to select. The participant has to select at least one answer per question. It is similar to the 'Multiple choice' option, but the need to set the same range of potential answers for all the questions can make the organization of the study more complicated.

- **Slider**: a bar that the participant can drag to a position to indicate an ordered value, i.e. ordinal or numerical, that represents the answer. It is set by the researcher to have limits over the precision with which the participants can answer.

- **Form field**: similar to a 'Text entry' but constraints can be entered to ensure the participant is communicating information in the format desired by the researcher. Forms can limit the size of the answers, force a specific format, e.g. email address or phone number. Restrictions imply that forms have to consider the potential answers participants will detail to enter.

- **Rank order**: set of entries that are to be ordered according to answer the question, e.g. order according to size, monetary value, sentimental importance.
- **Pick, group and rank**: Similar approach to 'Rank order', but the participant can enter a justification behind their choice into text boxes.
- **Heat map**: records of one or several clicks entered by participants over a 2D stimulus, commonly an image. That approach can be interesting for tasks in which a precise input is necessary to answer, e.g. if the participant is asked to communicate about a geographical point, selecting it with the mouse is an easy approach to indicate the position. The input made with the mouse is bound to lack a certain level of precision, but alternatives such as textual or numerical inputs are likely to suffer from the same issue in a stronger degree. The participants can also provide textual feedback on the positions they selected. The heatmap is a composite showing the density of all responses collected from such questions.
- **Hot spot**: similarly to a heat map, this is a set of pre-made selections of parts of an image for which the participant can provide information about.
- **Highlight**: option to select pieces of texts from a larger corpus and indicate additional information about those selections, e.g. evaluations of their meaning, such as approvals or rejections.
- **Airbrush**: approach similar to the heat map and the hot spot, with dynamic identification of areas on a picture without requiring boundaries, with the possibility to spread magnitude. This approach is not listed by Qualtrics, but has been used to record vague or fuzzy areas of interests from participants [171, 77]. While an interesting approach, it's been discussed that difficulties to set up and analyse results for this approach may have limited its adoption within academia and in practice [41].
- **Discussion**: observations made by participants as they perform the task asked of them. This Response method requires the researcher to record what participants say during the study. Commonly, researchers filter and report discussions participants made according to what they estimate insightful in the context of their study. This approach is more common for qualitative studies in which synoptic tasks are evaluated [17, 8, 44].

Note that this list is limited to methods to answer through interactions with elements displayed in 2D (including buttons). Interactions unique to 3D models (e.g. as listed by Bach *et al.* for a space-time cube [23]) are out of the scope of this thesis. While not incorporating them in our list of methods, we wish to suggest that by offering interactions specific to 3D environments, the methods to answer previously listed are likely to still fit, albeit with some adaptations in forms of input from participants. We discuss interaction further in our section 6.2.3.

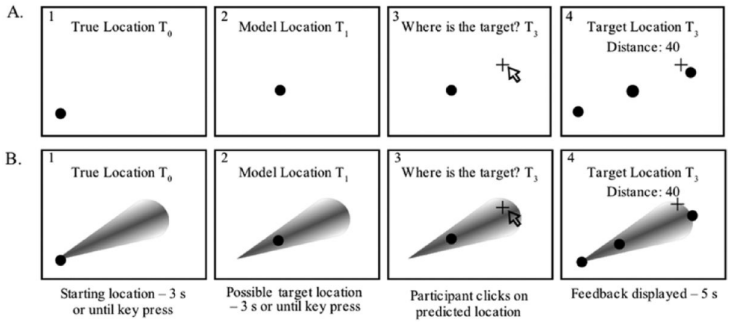
3.4 Examples of usage with the SFNCS

In this section, we illustrate how the SFNCS can be used to characterize studies. We select two papers discussing different tasks and different visual stimuli to show diversity amongst the examples. Note that we don't discuss every detail of the studies, but rather mention elements which allow characterizing the studies.
Our selection is not made to be exhaustive as the SFNCS incorporates many different notions and finding examples of all variations would be too considerable. Instead, we wish to illustrate different examples to discuss the richness offered by the scope of the characterizations incorporated in the blocks of the SFNCS. The first example we selected is a contribution from Pugh *et al.* [143] assessing the impact of a visualization method, the cone of uncertainty, for prediction of future trajectory position. The second example we selected is a contribution from Chen *et al.* [44] where they present an analytical workflow for movement-related event contextualization for pattern detection of spatio-temporal distribution of patterns.
We do not discuss all the details about the selected literature, but rather focus our discussion on how the SFNCS can be used to characterize different studies.

For each example, we introduce its context quickly and then discuss how we can characterize it using the blocks of the SFNCS. Note that we only characterize Data and don't categorize data complexity, as it is not communicated in these studies.

Example 1: Effect of Visualization Training on Uncertain Spatial Trajectory Predictions

Pugh *et al.* [143] presented the result of two experiments in which they assess the differences between participants answers and actual trajectories paths using a cone of uncertainty. Their first experiments consist in comparing difference of performance between the participants of the control group and those who are presented with the uncertainty cone as a visual helper. The process is illustrated in Fig. 3.1a Their second experiment consists in predicting the future position of curved series of points while being helped with the uncertainty cone with various distributions of points. An example of the stimuli is illustrated in Fig. 3.1b.
The data presented to participants is solely spatial, we thus consider that in this case, the Data block is solely populated by WHERE data.
Participants are asked about a question where a baseline exists, and there is no subjectivity to consider if participants are correct or not, we can thus characterize the Task they are asked to perform as a Measurement. The task asked of participants requires them to consider the mean directions and speeds of moving entities according to displayed points, indicating the participants required to analyse the data presented to them a whole, meaning the Task asked of them was synoptic.
Furthermore, participants are asked to identify a specific point, meaning the Task asked is a Lookup with a WHERE focus. It is not indicated in the paper whether Pugh *et al.* asked participants in order to perform the Task. We thus consider that if they were, it would follow the same characterization as the Task. Participants are asked to click to indicate their answer, thus indicating the Response block is



(a) Structure for the first study of Pugh *et al.* [143] with participants predicting next position of a trajectory according to prior points. According to their group, they are asked to perform the task with or without the visual help of the uncertainty cone.

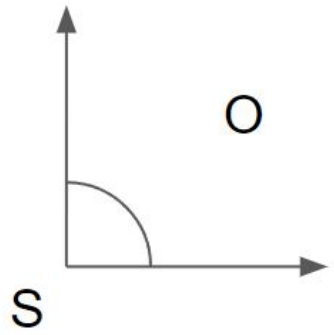
(b) An example of stimuli for the second study of Pugh *et al.* [143] in which they ask participants to perform the same task, with the help of the uncertainty cone, but with a higher variety of curvatures and changes of speeds.

to be characterized as a Heat map. We illustrate the resulting set of blocks in Fig. 3.7. We illustrate

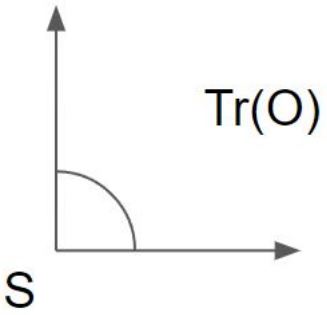
Task	Question	Response
Synoptic Measurement Lookup WHERE	Synoptic Measurement Lookup WHERE	Heat map

Fig. 3.7 The characterization of the blocks Task, Question and Response in the SFNCS for the study of Pugh *et al.* [143].

the characterization of the Visual Stimuli block for graphs presented to participants using pictographs, following the method of Andrienko *et al.* [9] in figures 3.2a and 3.2b.



(a) The control group of the first experiment of Pugh *et al.* only sees a series of points. As the information presented is a trajectory, we consider the background to be spatial.



(b) The test group of the first experiment, and all the participants of the second experiment, of Pugh *et al.* , were presented cones of uncertainty. The cone of uncertainty displays a clear connection between the records of positions, and thus it is labelled as a trajectory.

Example 2: Contextualized Analysis of Movement Events

Chen *et al.* [44] presented an analytical workflow including event contextualization for pattern detection and exploration of spatio-temporal distribution of patterns. To evaluate the efficiency of their workflow, they collaborated with domain experts and verified whether they understood the data displayed to them and whether the observed patterns. The workflow consists of extracting events as segments of trajectories preceding and following events and display additional attributes in the time frames of these events characterizing or describing external circumstances.

This visualization allows participants to interact with the software, and thus several variations of composite graphs can be produced. We analyse the composite graph presented in the contribution of Chen *et al.* for the characterization of the Data and Visual Stimuli block. Chen *et al.* present several variations of their composite graphs that participants can modify through interaction. The database discussed is composed of movement of trucks in Greece combined with attributes of the moving objects, e.g. fuel level, attributes of their movement, e.g. speed, and parameters of the environment, e.g. temperature. They extract from the movements events they label, e.g. harsh brakes. While their approach considers events as a quality occurring over a single moment, we still argue that that type of information is still to be considered qualitative, albeit with a very short duration. The Data block of this contribution is thus WHAT_Qn, WHAT_Ql and WHERE.

We notice that all their variations of composite graphs are composed of a limited number of variations of the same graphs. We thus list all the different graphs presented concurrently in their contribution and use them to characterize the Visual stimuli block. The composite graph used as a reference is illustrated in Fig. 3.8 and the resulting pictograph is illustrated in Fig. 3.9.

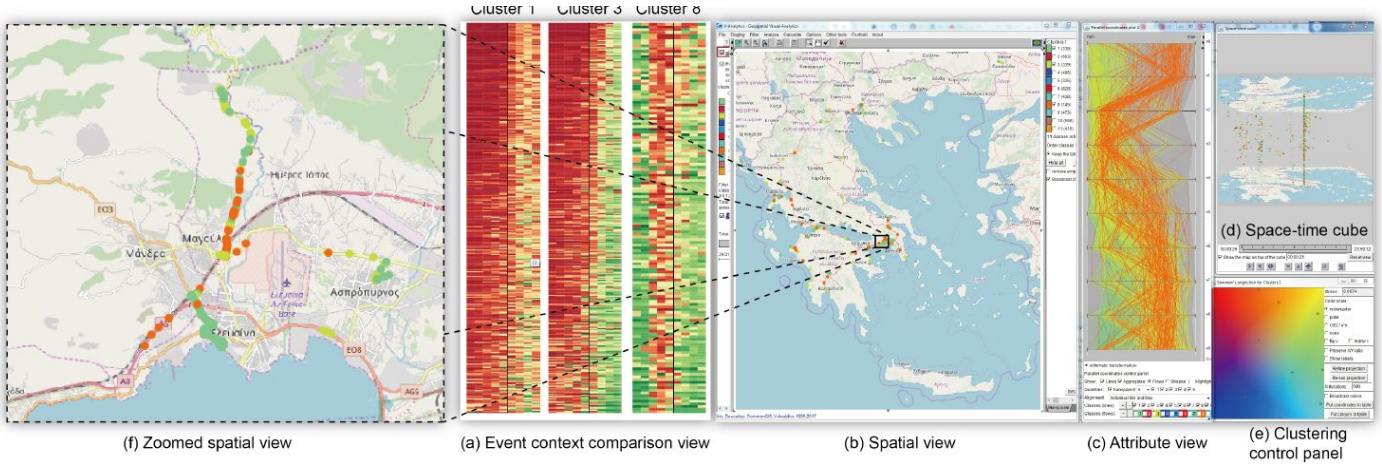


Fig. 3.8 The example of composite graph presented by Chen *et al.* [44]. This composite graph includes: (a) Event Context Comparison View, (b) and (f) Spatial View, (c) Attribute Parallel Coordinates View, (d) Space Time Cube, (e) K-Means Clustering Control Panel.

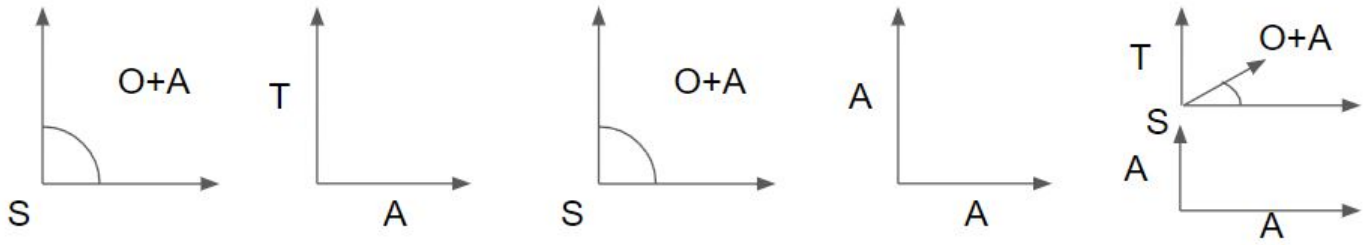


Fig. 3.9 The pictographs illustrating the Visual stimuli block displayed to participants in the studies ran by Chen *et al.* [44].

The analytical workflow developed by Chen *et al.* [44] was used to perform two different tasks: context pattern discovery, and exploration of the spatial and temporal distribution of events. The two tasks are motivated by an overall objective to analyse the context around records of harsh brake events.

What constitutes a pattern is to be decided by the domain experts, indicating that the Task is a Judgement. This task requires considering the whole of the data presented together to judge relevancy of observations, meaning the task is synoptic. Additionally, this first task consists only in finding occurrences of such patterns to observe y their positions. We thus consider this task to be a lookup. Participants were not directly asked questions, but rather domain experts already had tasks they aimed to perform. We thus characterize the Question block the same as the Task block. The resulting characterization of the blocks Task, Question and Response are illustrated in Fig. 3.10.

Once the patterns were found, the following task was to analyse the distribution of spatio-temporal attributes of the events. The composite graphs Chen *et al.* used to illustrate exploration of the clusters contextualized for this task are composed of the same previous graphs but with different sections. Participants search to understand the relationships between all the data types listed previously, indicating that the Task type is Relation seeking, and that the focus of the Task is WHAT_Qn, WHAT_Ql and WHERE. Similarly to the previous task, once more the Task outlook is Synoptic and the Task subjectivity is Judgement. We illustrate the Task, Question and Response blocks in Fig. 3.11.

Task	Question	Response
Synoptic Judgement Lookup WHERE	Synoptic Judgement Lookup WHERE	Discussion

Fig. 3.10 The Task, Question, and Response blocks using the SFNCS for the *'Context Pattern'* Discovery task of the Chen *et al.* [44] study.

Task	Question	Response
Synoptic Judgement Relation seeking WHAT_Qn/WHAT_QI/WHERE	Synoptic Judgement Relation seeking WHAT_Qn/WHAT_QI/WHERE	Discussion

Fig. 3.11 The Task, Question, and Response blocks using the SFNCS for the *'Exploration of the Spatial and Temporal Distribution'* task of the Chen *et al.* [44] study.

3.5 Chapter summary

In this chapter, we discussed our main contribution, the Systematic Framework for N-Scales Characterizations of Studies (SFNCS).

This chapter begun with the presentation of a taxonomy, for which we developed some novel categorizations. We later pivoted from this approach, but this work helped us to consider how to later build our framework, the SFNCS.

This chapter continues as we discuss data characterization. This discussion was important as it allowed to claim that the Card model could be considered with the same elements to characterize different aspects of visual analysis: Task, Data, Visual stimuli. We then define complexity measurements for the data, which is an important step for later considerations of study setup and communication of results: as we make claims of ranges of complexities from the data, we initiate a conversation about comparison of visualization analysis studies using different data sets. This section ends with the introduction of the notions of Measurement and Judgement, which is a necessary step to separate cases where human subjective judgement is used to collect or communicate information.

The SFNCS is then finally presented, extending the Card model with the Question and Response blocks. These new blocks are set to characterize questions asked to participants during studies, and what tools are provided for them to answer, and thus perform the task asked of them.

The two following chapters discuss a novel visualization method, the ATS-ATS Mask, and technology related to displaying it and controlling data for studies set up. In chapter 5 we then discuss how the SFNCS is used to analyse results of studies set to analyse the ATS-ATS Mask.

Chapter 4

The ATS-ATS Mask

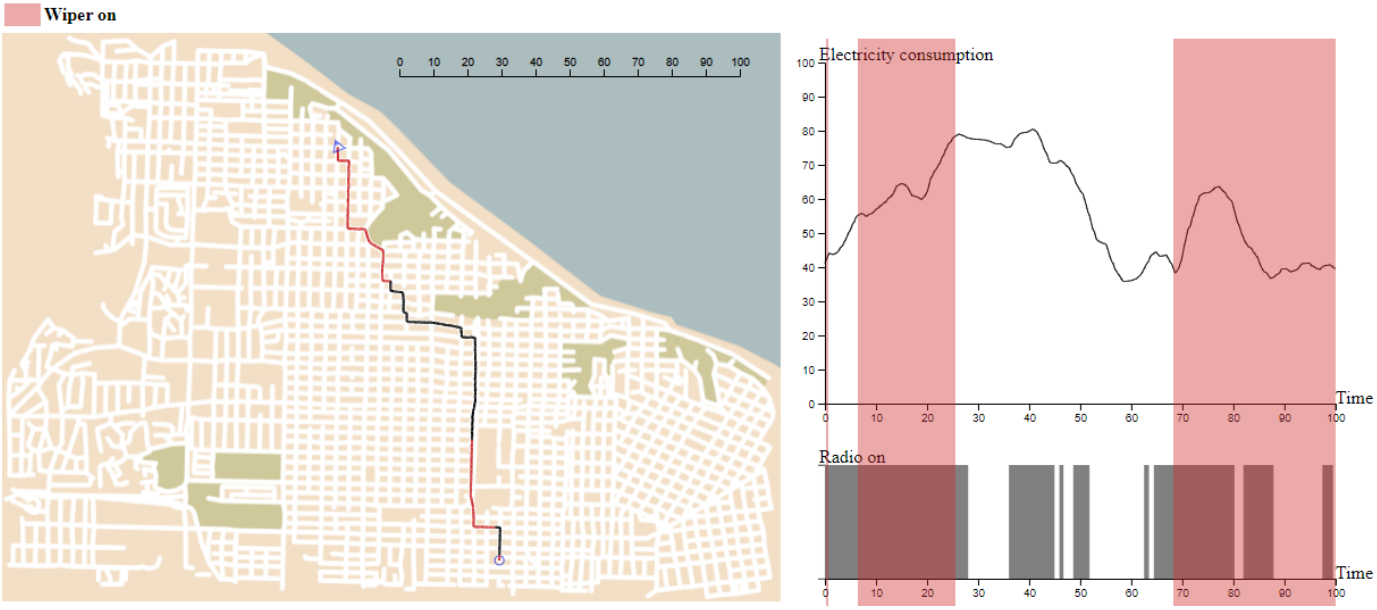


Fig. 4.1 The concept of the ATS-ATS Mask is to indicate through overlays on time frames the data that matches one or more conditions. Conditions may be described through queries.
For each use of the ATS-ATS Mask, its naming is adapted to the type of information queried, and the type of information displayed before the application of the overlay: condition-graphic
In this example, the query is about an *attribute* matching a certain condition, meaning the first section of this Mask name is **A**. Additionally, the Mask is overlaying a visual form displaying information about *attributes*, *time* and *spatial information*, meaning the second section of this Mask name is **ATS**. This is thus an example of an A-ATS Mask.
In this design, the beginning of the trajectory is indicated with a coloured circle, and its end with a triangle, oriented and with one edge coloured to indicate the overall direction of the trajectory.

The time mask, developed and used by Andrienko *et al.* [17, 8], deals with the temporal concept of querying and filtering spatio-temporal data based on attributes. A time mask filter allows an analyst to see when certain conditions are fulfilled and what else happened during those times. It can be

considered a visual annotation that supports comparison of times when conditions are met, and those when they are not. In Card's terms [39] (Fig. 2.5), it is a visual mapping that adds a condition to an existing visual structure. Andrienko *et al.* [17, 8] claim that this approach *may be very useful in joint analysis of several time-referenced datasets for finding relationships between different phenomena*. As they mention, the novelty of their approach is that previous work used to do queries based on time, either with a linear view of time, or a cyclic one. Their approach adds a temporal filtering in which time intervals are selected based on satisfaction of query conditions formulated in terms of time-variant attributes. Such selections, and the resulting time frames, are displayed before the filtering is applied, over one or more time series displaying attributes. An example in which a time mask is applied to time graphs showing variation in quantitative and qualitative attributes is shown in Fig. 1.1.
The concept seemed promising, and thus we aimed to further study it, either by generating new designs using that concept, or extending the concept further. As discussed in section 3, no matter the approach, our resulting contribution would require formalization for its subsequent evaluation. First, a formalization of the time mask would prove valuable as a step to justify the direction of our efforts to extend it. The formalization of the time mask using the framework of Andrienko *et al.* [9] lead us to the concept of the ATS-ATS Mask, discussed in this chapter. In this chapter we discuss how we extended the concept of the time mask, resulting in the ATS-ATS Mask and how that concept can generate a variety of designs. We then discuss the designs selected for the studies described in detail in section 5, and the purpose of prioritizing that design. We also discuss the importance to consider dependencies between designs and masks overlays.

4.1 Extending the time mask

Similarly to how the time mask of Andrienko *et al.* [17] extends the range of queries, we initiate a step further by considering the possibility to make queries based on geography. Queries based on geography can then be defined either for a location itself, e.g. a frame of latitudes and longitudes, or location of an identified object, moving or not. Note that the status of an object as a moving one is a relative statement, e.g. the position of a car is likely to move at some stage, whereas the position of a supermarket is unlikely to move relative to the coordinate system in which it is positioned. Characterizing entities as moving or motionless is dependent on the scale of the data being recorded, e.g. depending on spatial scale selected to analyse data, all objects on Earth are moving. Similarly, according to the time scale considered, the nature of an object can vary, e.g. a mountain is immobile in most modern navigation systems, but over a long time does move.

While there are a variety of potential visual mappings that can support an overlay that indicates conditions being met, the time mask as defined by Andrienko *et al.* [17] is only considered as an overlay upon time series displaying quantitative or qualitative attributes, aligned and sharing the same time axis. We consider that since the concept of the time mask can be defined with various data

types for the input, and result in different visualizations, specifying these in a more precise manner is important in ensuring validity when evaluating any of these variations.
Thus, we build on the time mask by defining the ATS-ATS Mask, which encompasses the potential variations. As with the Peuquet *et al.* [139] framework, information is considered to be of three types, A for Attributes (WHAT), T for Time (WHEN), and S for Space (WHERE). The first section of the name is used to indicate the type of information queried with the mask and the second section is used to indicate the information displayed in the visualizations onto which the condition is visually mapped. Following our approach, we can characterize the time mask presented by Andrienko *et al.* [17] as a *A-AT Mask*.
The approach we introduce is not sufficient to precisely detail a unique visualization, but indicates the purpose of the type of mask: what it's displaying, and what overlays are applied over it. Additionally, it is important to consider that if several visualizations are connected, i.e. the information is shared between them (e.g. two views showing information at the same time), we consider these composite graphs as one mask overlay.

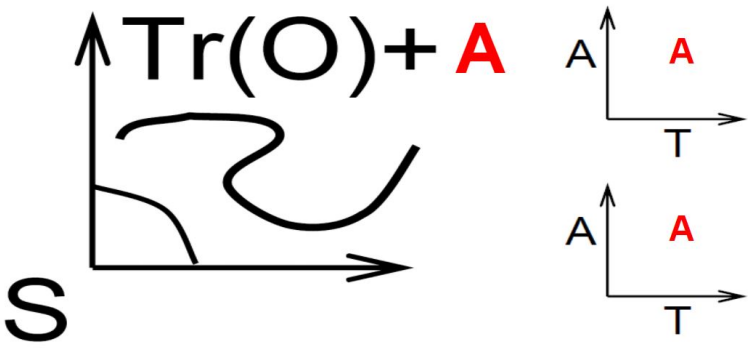


Fig. 4.2 The A-ATS design used in our studies, characterized using the framework of Andrienko *et al.* [9]. The letters encode the following information: spatial positions (S), temporal positions (T), trajectories (Tr), attributes (A), for detailed representation of the data (not aggregated). This design presents detailed information (not aggregated). For clarity, we coloured in red the information overlaid by the A-ATS Mask.

This approach does not indicate for a set of visualization how the layouts of information is set up, but indicates the intention of the selections made by the designer. We still recommend other approaches to characterize the visualization more specifically, such as pictographs as presented by Andrienko *et al.* [9] to facilitate comparison between visualization methods evaluated in different studies.

The selection of the scale of abstraction that is most appropriate to describe and compare visualizations is an open question, which has been a point of important reflection when discussing how could we describe our extension of the time mask concept into the ATS-ATS Mask. Such selection has to be done by researchers as they consider the level of specificity unique to their contribution, and it seems probable that as the number of research contributions increase, alternative theoretical models

arise. We fully support the modifications of the blocks of the SFNCS, including the Visual stimuli one, as long as new contributions discuss how their new approaches fit in regards to previous ones.
To efficiently discuss the ATS-ATS Mask and evaluations of its variations, it is necessary to consider the porous definitions of qualitative, quantitative, and spatial attributes. Spatial attributes are first dependent on the recording method used to collect them. If a spatial record is made by saving its longitude and latitude, it is effectively a combination of two quantitative attributes, while if the record is reduced to the region in which it is produced, e.g. a town, it is effectively a quality with implicit spatial and contextual knowledge embedded. Note that a place does not have to be real to allow for spatial data to be composed of quantitative attributes combinations or contextually rich qualitative data, e.g. a town in a video game [112]. The consideration of whether data is spatial is thus linked to context. What this implies is that while spatial data can be displayed with visualization designed to display quantitative or qualitative attributes, they are most of the time unique to spatial data.
These considerations are important as they influence our choices considering how the ATS-ATS Mask can be used and how it overlays elements displayed in the composite graphs.

First, the spatial attributes can be considered as specific combinations of quantitative or qualitative attributes, we consider that its contextual uniqueness requires it to be considered as a specific distinct element, i.e. the S of the ATS-ATS Mask.

Second, the ATS-ATS Mask is to consider data as it is presented in priority to the source data, e.g. if quantitative data is transformed (e.g. aggregation) into qualitative data it should then be considered that qualitative data is presented.

Third, the ATS-ATS Mask is a concept that does not modify the scaling of the display of the composite graph it overlays, but is expected to be applied with values fitting the composite graph, i.e. the ATS-ATS Mask is expected to generate overlays with comparable orders of magnitude over time, space, and values of attributes. Furthermore, the ATS-ATS Mask is not expected to imply changes over the composite graph in order to force the display of the overlay, e.g. the ATS-ATS Mask is not expected to extend the size of a map to display an area that would fit the condition entered.

Fourth, similarly to the time mask, the ATS-ATS Mask is a concept that can result in diverse approaches to display overlays indicating its conditions being met. We reiterate our previous statement that we recommend the use of pictographs to indicate how researchers display the overlay they evaluate, but we do not claim that specifying ATS-ATS Masks combinations and pictographs is bound to be sufficient to compare contributions. Elements such as scaling or levels of aggregation have to be considered by researchers to ensure comparisons between contributions to be valid.

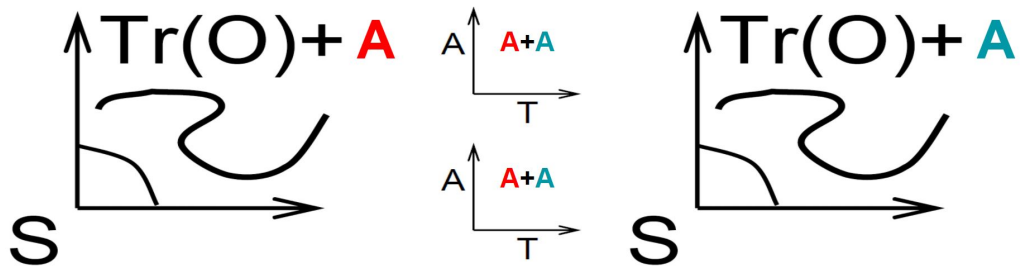


Fig. 4.3 Several ATS-ATS Masks can be overlaid over composite graphs. This pictograph exemplifies how several Masks can be applied over composite graphs, with each hue indicating a different A-ATS Mask. This example displays four different visualizations, with two bearing only one overlay and two bearing two overlays. This fictious example could for example represent analysis of trajectories over different geographical areas over a shared time frame.

4.2 Assessing the range of potential designs: set up of a workshop and reflections

As discussed in section 2.2, there are a variety of potential graphs representing movement and attributes related to space-time attributes. As part of our process to select the visualization designs that would suit our needs best, we categorized the visualizations of time-oriented data listed by Aigner *et al.* [3]. This process was useful to grasp the scope of potential visualizations that could be selected, but was not sufficient to claim with confidence a single design selection.
One factor that influenced our selection of designs over which to apply the ATS-ATS Mask was our consideration of future evaluation processes. The novelty of the ATS-ATS Mask implied a lack of knowledge regarding its effectiveness for performing tasks with a low level of abstraction. Acknowledging this implied that we should first focus on evaluating the understanding of the ATS-ATS Mask concept and ability to perform low-level tasks. We discuss the studies we ran further in section 5, but this decision impacted the designs selected for studies, i.e. the studies that would be run to evaluate the design would be quantitative, and thus require a fairly high number of participants. Considering a high number of participants for our studies implied that we could not afford to select designs that would only be understood by experts. Additionally, this choice was justified by our interest in evaluating the impact of the Masks over a common visualization, rather than the visualizations on their own. Setting up a complex design as a base on top of which to overlay the ATS-ATS Mask meant increasing the risk of confounding effect of results indicating a certain level of understanding or ability to perform for the base design rather than the combination of the base design and the ATS-ATS Mask.

To assess the types of visualizations that would be most commonly used to display the information we would use for the studies, we ran a workshop with a group of visualization experts, with various levels of expertise, ranging from PhD students to professor. Workshops have been successfully used in many previous contributions to either generate or evaluate visualization designs [91, 64, 120, 160].

During the workshop, the 5 participants were first presented with the concept of the time mask as defined by Andrienko *et al.* [17] and their query system. They were then asked, for a series of types of information, to produce drawings of as many types of plausible designs to display the information as possible. As the workshop progressed, the variations of design grew more diverse and rich in terms of the number of elements displayed as well as the tasks that could potentially be performed using the visualizations.
At the time we ran the workshop, the number of moving elements and attributed to display to participants of our future studies was not set yet. We had planned already not to consider interactions in our studies. Thus, for our workshop we decided to aim for a number of elements that was high enough to consider design methods to account for it without requiring aggregation of data, or without requiring interaction. Participants were asked to generate visualization designs to display information following these specifications:

- 3 moving entities that vary their position over 30 points of time.
  In a space that contains areas with different static characteristics:
  - 2 categorical variables
  - 1 numeric variable
- Each object having two *constant* attributes:
  - 1 categorical attribute
  - 1 numeric attribute
- And Each object having two *varying* attributes
  - 1 categorical attribute
  - 2 numeric attributes

To do so, they were provided with paper and colouring pencils, or marker pens and a whiteboard. This setup made it difficult to capture interactions, but as this was not our focus and still enabled participants to convey design ideas quickly and effectively. The participants were not specifically instructed use Masks in their drafts. Once each participant generated multiple design approaches, they were asked to increase the number of elements to display, either keeping the same design principles they used to draw their drafts, or modifying them accordingly.

As a wide variety of designs were generated, trends appeared:

- All the participants, when asked to generate visualization designs for trajectories, started by making drawings of those over the map background.
- The usage of colour was used to either:

  - Indicate different moving entities (when we increased the complexity of the information to be represented)
  - Indicate areas of interest
  - Indicate time frames, sections of trajectories, or regions that would match a query

- All participants, when asked to display the evolution of some attributes over time, drew at least one visualization with time as the x axis.

- Designs grew in complexity as we increased the number of elements to draw (moving entities, number of attributes) but discussions during the workshop indicated that aggregation would not be considered by most until a significantly higher number of elements to draw would have to be considered.

Most of the variations in the design happened when we asked the participants to display several moving entities or several attributes in their drawings. Especially, aggregation of the data was only considered for either several moving entities or several time dependent attributes. These supported our decision to display precise levels of information (as opposite to aggregated) if we were to evaluate a single moving entity or a relatively small number of attributes. The drafts produced during the workshop are available in the appendix B.

The ATS-ATS design is thus inspired by the time mask of Andrienko *et al.* [17] and our interpretations of the data collected during the workshop to develop a design that would have a base of visualization methods commonly used to display the information of interest and support overlays to display masks for queries being met.

4.3 Prioritizing the A-ATS Mask

The richness of the ATS-ATS Mask means many sets of evaluations should be developed in future work to fully understand its impact for diverse tasks. We did not have the resources to evaluate all the variations possible, and thus a selection had to be made. We had to decide what would be the most important combination for a first set of studies we would run. As stated in section 4.1, the ATS-ATS Mask is the result of two additions compared to the time mask: addition of the possibility to make queries based on spatial information, and consideration of information displayed by the visualizations the Mask is being overlaid on top of. These two considerations were the motivators towards the selection we made later among the various combinations that can be generated with the ATS-ATS Mask. Our final decision considering our selection of visualization designs, tasks, and evaluation methods is therefore an aggregate of our interests and estimations that would yield the most indicative contribution, as well as an artefact from the necessity to prove that the SFNCS and the designs we selected are efficient in a situation with a simple set up before future research could expand upon those claims.

As we discuss in section 2.2, the biggest inspiration for our ATS-ATS mask is the time mask of

Andrienko *et al.* [17] which was only displaying the time mask over visualizations depicting space-time attributes, be it quantitative or qualitative, at a precise level, i.e. the data was not aggregated. Using the characterization we defined, the time mask is a A-AT Mask.
Various combinations could prove valuable to assess strengths and weaknesses of the ATS-ATS Mask while increasing the scientific literature corpus. We considered spatial information a priority as it highlights the addition our contribution brought to the time mask. We thus considered that we needed to evaluate tasks that required a focus on quantitative attributes, qualitative attributes, like in a A-AT Mask, but also a focus on spatial information, resulting in our choice to select a design for a A-ATS Mask.
As we considered how to design and implement the A-ATS Mask, we aimed for solutions that could support a variety of tasks. The characterization of tasks by Andrienko *et al.* [9] indicated us a certain range of parameters necessary so that tasks could be performed, e.g. synoptic comparative tasks required to indicate the chronology of time records. While many implementations can be achieved for a A-ATS Mask, as that of Andrienko *et al.* [17] indicated that it was understood and effective for experts for a specific use-case. We infer that a design similar to theirs would likely be understood by participants for future studies, albeit modifications would be necessary to extend the range of tasks that could be supported with our design implementation.

The design we selected for our studies is illustrated and categorized in Fig. 4.2 using the framework of Andrienko *et al.* [9]. Our design presents similarities and differences compared to the one of Andrienko *et al.* [17]. An example of the same approach with two masks is illustrated in Fig. 4.3. Similarities with our design include:

- The visualizations depicting the evolution of quantitative and qualitative attributes are aligned, and information is recorded and depicted at the same time.
- The design highlights time frames that match the conditions with visual annotations: rectangles that cover the substrate are orthogonal to the temporal axes with low opacity so as not to obscure other marks and their visual encoding.
- Quantitative information is displayed with a time series visualization that connects each recorded point.
- Qualitative information is displayed with coloured blocks to indicate its value for each record. The visualizations set up for the studies only display binary attributes, but colouring could be used to indicate various categorical values, if those were limited to a finite number, for which each they would be easily discernible.

But while our inspiration for the ATS-ATS Mask originated from the time mask, we aimed to assess different designs based on that same concept. The differences in our designs are either due to an interest in variations in the design, or adaptations for the studies. The differences include:

- Unlike the implementation of the time mask of Andrienko *et al.* [17], within our A-ATS Mask, we do not repeat the data from which the time mask is derived. The time mask highlights with

coloured horizontal bars the conditions in the visualizations that display the queried attributes. The dependencies between data present in the visualization before the application of the A-ATS Mask and the information it depicts are illustrated in Fig. 4.4.

- Our dashboard includes a visualization displaying a geographical area, over which a trajectory is drawn. Since we aimed to indicate chronological order of the spatial records, our designs needed to indicate that information. We considered several approaches, e.g. text and/or numbers to indicate starts and ends of trajectories, arrows indicating direction at the starts and ends of trajectories, series of arrows over the entire trajectories, icons. We decided to indicate the beginning of the trajectory with a coloured circle with no fill, and its end with a coloured triangle, with a section of it coloured, to indicate the global direction from its beginning to its end. This approach was selected as it establishes movement order trajectories that end close to the position at which they begin. The colours of the circles and triangles are unique to each trajectory, to ensure future work could include multiple moving entities and allow to differentiate trajectories including in cases of overlap.

- We extend the encoding of the A-ATS Mask to the trajectory that matches the query entered. We decided to communicate this information by colouring the trajectory using the same colour used to indicate the condition is met, with the mask over the visualizations for the qualitative and quantitative attributes. Colour has associative properties, and we can assume that the semantic association is made given the visual association [27].

		Mask		
		WHAT_Qn	WHAT_Ql	WHERE
Focus	WHAT_Qn	Independent	Independent	Independent
		Dependent	Dependent	Dependent
	WHAT_Ql	Independent	Independent	Independent
		Dependent	Dependent	Dependent
	WHERE	Independent	Independent	Independent
		Dependent	Dependent	Dependent

Fig. 4.4 This table indicates combinations of focus and mask type. For our studies, we set our mask dependencies differently to how they were set in the work of Andrienko *et al.* [17, 8]. Unlike them, we set a variety of masks, reduced to binary attributes, that are defined by values not presented in the original dashboard of visualizations. The combinations that we use in our studies are coloured in blue. This graph indicates thus that if the focus and the Mask are displaying the same type of information, they are displaying different attributes, e.g. if focus is WHAT_Qn and Mask is WHAT_Qn, the quantitative attribute shown could for example be the engine temperature while the Mask represents a condition on an unrelated attribute, such as the suspension force being above some level.

The design choices made for our implementation of the A-ATS Mask are related to our aim to evaluate them. An important choice we made when we set up our studies was to display the movement and space-time attributes of a single moving entity, a single quantitative attribute and a single qualitative attribute. This decision originates from:

- Most quantitative studies present a single visualization that enables participants to perform one evaluated task. Qualitative studies often require multiple visualizations for participants to perform the task asked of them. As we aimed for an implementation of the A-ATS Mask that could be used for quantitative and qualitative studies, our design is displaying a dashboard of visualizations, including some which are not necessary to perform the tasks asked of them. This approach is fairly unorthodox, which justifies the necessity to first evaluate a small number of elements to display to the participants of our studies.

- As in their follow-up work with the time mask, Andrienko *et al.* [8] develop additional visualizations to analyse movement during football games and display either several moving entities at the same time or calculate the means of several moving entities and display the resulting one. A particularly interesting approach they took was to measure for each player his position in the team space (an artificial space) and then aggregate those for the whole team, resulting in an average of one position for each time recorded. As that method proved useful for analysts [8] it reinforced our confidence in the validity of our approach.

The design we set up can easily support additional attributes to be displayed. We actually made prototypes that supported the display of more moving entities and more quantitative and qualitative attributes. We intend for future work to increase the number of attributes and moving entities displayed in future evaluations. We discuss the potential future work in section 6.2.

4.4 From theoretical framework to data to evaluate it

We introduced the SFNCS that allows to characterize studies and the ATS-ATS Mask. We aim to evaluate the A-ATS Mask using the SFNCS. Evaluating the A-ATS Mask introduced in section 4 originates from an interest into expanding the time mask, used to help data analysis of time-space attributes for moving entities in diverse contexts, e.g. football matches, air-traffic-control and shipping movement [17, 8].
Before evaluating the A-ATS Mask, we discuss the process through which we selected a data set which fits our requirements. The requirements were the following:

- Access to the data: we required a data set which we could have access to and with which communication of our work was possible.

- A level of detail sufficient to perform the tasks we wished to assess: as the A-ATS Mask is built to expand from the time mask, we prioritized the evaluation of the A-ATS Mask with a similar level detail as discussed in the paper which introduced the time mask. Tasks are discussed in detail in section 5.1.1. The implication of this requirement is that we disregarded potential data sets containing only aggregated data.

- Ensuring privacy of moving entities: sharing of data is important to allow scrutiny from other researchers and transparent communication of results, but doing so implies ensuring that

anonymity of the people recorded as moving entities is preserved. If researchers aim to use geographical movement, they can either use methods to anonymize trajectories [75, 127] or use fictitious data.

- In the case where a fictitious one was chosen, to be realist enough to avoid its misinterpretation.

Data selection

We considered several potential data sources with their own associated context and how their potential tasks of interest could fit our research interest:

- **Ship movement** using the Automatic Identification System (AIS) in the vicinity of harbours, where risks increase while moving entities compete for landing ports access at specific times depending on cargo price changing dynamically [30]. Assessing causes of occurrences of ships navigating too closely is a complex problem that is likely to require visualizations to analyse time-space attributes. We contacted several groups (Office for National Statistics, Maritime & Coastguard Agency, Peel Ports Group) who initially indicated interest for collaborations but then pulled away from the project. Assessing causes of occurrences of ships being dangerously close is a synoptic task, for which experts knowledge is necessary to assess whether observations are significant. With access to neither experts nor data, we thus disregarded this context.

- **Aircraft movement** for planning assistance, for which optimisations could reduce the fuel consumption, up to 16.7% according to Qian *et al.* [144]. Colleagues had previously worked and published with a data set that they even shared with us. But due to policy changes from the group that generated and owned the data, our authorisation to work and publish with that data set was removed.

- **Traffic movement** for analysis of cities clutter was also considered. The analysis of that data is important to consider limitations of roads in various situations, such as traffic jams or evacuations [20]. Detailed real data records of cars positions around cities are not simple to get access to unless partnerships with other groups are set, due to privacy concerns. Detailed real-life records were thus not sought after.

- **The IEEE VAST 2014 Challenge data** [49, 174, 59, 38] set was our final choice for the data set used for our studies. It is a set of trajectories set in a fictitious town, called Kronos, in which movement of certain cars with unique identifiers is recorded. The fictitious town is represented with a drawing of the map, over which can be overlaid trajectories of moving entities. The IEEE VAST 2014 Challenge data set is composed of realistic trajectories, with a detailed level of precision. This set requires no anonymization, as the data is fictitious. Additionally, its fictitious nature meant there could not be a confounding effect due to pre-existing knowledge of the displayed area. The IEEE VAST 2014 Challenge data set thus met our needs, since it was composed of realistic trajectories in a fictitious area.

The selection of the IEEE VAST 2014 Challenge data set meant we possessed a set that met our requirement for the TS part of the A-ATS Mask we wished to evaluate. The following step is thus to consider what attributes should be used for our studies.

Data enrichment

Thus, we generated quantitative and qualitative data to complement the trajectories, doing so through several iterations and with controlled constraints. As discussed in section 3.2.2, a diverse set of values for the quantitative and qualitative attributes was necessary to avoid confounding factors.

For our studies, we required quantitative data that seemed realistic and rich, e.g. presented diversity for means, ranges, different levels of straightness. To do so, we selected the Perlin noise, as it's been used for decades to generate data with realistic noise characteristics [138]. Specific values selected to characterize the data we generated are discussed in section 5.1.2.
We used a Processing sketch that draws upon the Perlin library [60] to rapidly generate sequences of bounded quantitative values with diverse characteristics in terms of mean, variance and autocorrelation. We show a screenshot of the software we developed and used to verify distributions of characteristics of attributes generated in Fig. 4.5. The sequences of quantitative values are represented with lines, similarly to how they would be displayed in a time series visualization. Their characteristics are shown by the overlain grid, with cells representing different combinations. Complexity increases from left to right, mean (arithmetic mean for the quantitative data, proportion 'on' for the binary data) from bottom to top. Red cells indicate that a sequence of quantities has been generated with the combination of complexity and mean represented by the grid position. Blue cells show that a binary sequence of the appropriate complexity and mean has been generated.
Having generated many such sequences, a subset was selected as representative of the range of characteristics through structured sampling. Others were categorised into binary sequences, and again a selection were chosen through structured sampling as representative of the range of possible variations. The generated attributes were then assigned to trajectories. We considered that each record of position should be accompanied by a series of quantitative and qualitative attributes that would reflect potentially interesting measurements done in real-life analysis of car traffic, i.e. information specific to the car. Information external to the car, e.g. temperature, precipitations, were considered but disregarded due to concern over potential interpretation from participants that external attributes could influence the internal ones. Thus, for each record of position, a car is assigned a value of the following attributes:

- Quantitative attributes:
  - Engine temperature
  - Fuel consumption
  - Suspension force

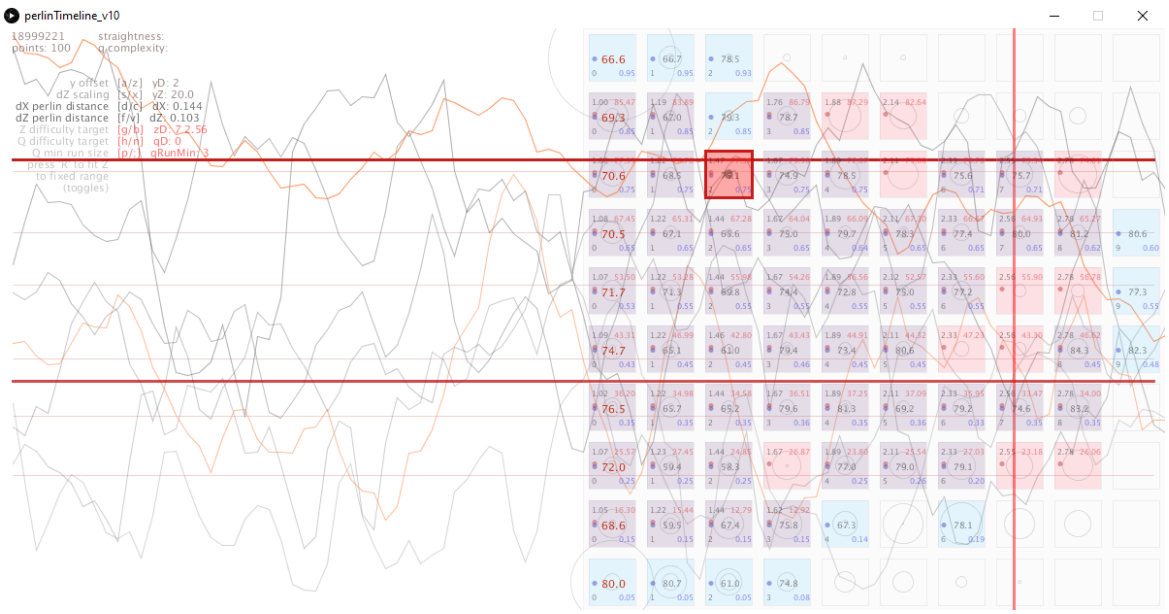


Fig. 4.5 A screenshot from our software that generates quantitative and qualitative values using Perlin noise to control noise variations. The lines drawn are matching the category highlighted when the mouse cursor is over them. The x axis represents the complexity measured, and the y axis represents the range of potential means in that group. The squares are red if the algorithm has only generated quantitative values for that set of criteria, blue if only qualitative values have been generated for that set of criteria, and purple for both.

  - Electricity consumption

- Qualitative attributes:
  - Gps on
  - Heating on
  - Wiper on
  - Radio on

Our first approach was trying to generate values with ranges that reflected real-life measurements. This approach was potentially interesting as we wished to make the study as realistic as possible, but since we aimed for claims to be generalizable, we decided not to follow it, as the different attributes could have different ranges, with different minimum or maximum values that might affect task performance. As we evaluate the A-ATS Mask, we want to make it possible to consider the impact of the A-ATS Mask over different composite graphs displaying either WHAT_Qn, WHAT_Ql. We thus ensure the validity of the comparisons by standardizing ranges for all quantitative values ranging from 0 to 100 and set all the time frames of the composite graph ranging from 0 to 100.

4.5 Chapter summary

In this chapter, we introduced the ATS-ATS Mask, a method to indicate time frames which match conditions. This method is an extension of the time mask developed by Andrienko *et al.* [17]. The initials A,T and S stand for Attribute, Time and Space, following a set of characterizations of information defined by Peuquet [139]. The three first letters of the ATS-ATS Mask indicate the type of information being queried, and the three last letters indicate the information displayed by the graphs over which the Mask is displayed.

In this chapter we first explained motivation to extend the time mask and then presented our reasoning resulting in the ATS-ATS Mask. We wanted to evaluate the ATS-ATS Mask. Since it is a method to overlay visualizations with an indication of when conditions are met, we first had to consider over which visualization methods can the ATS-ATS Mask be displayed. We discussed how we assessed the range of visual mappings which allow displaying spatio-temporal attributes through a workshop ran with visualization experts.

Following reflection on the range of designs produced, we discuss and justify our choices for variations of the ATS-ATS Mask to evaluate first. Our interest focused on the A-ATS Mask, as the time mask was an A-AT Mask, and by adding the overlay of the Mask over spatial information, we could then consider tasks with diverse focuses.

In the next chapter, we discuss technology to display spatio-temporal attributes and overlay Masks over it, and the selection or generation of data to allow evaluation of the A-ATS Mask.

Chapter 5

Studies to evaluate the A-ATS Mask

In chapter 3 we present the SFNCS used to characterize studies, and in chapter 4 we present the ATS-ATS Mask, and discuss the reasons why we prioritized studies designed to analyse the strengths and weaknesses of the A-ATS Mask. In this chapter, we detail the studies we ran, present the points that motivated them and their structures, and discuss the conclusions we drew after analysing the results. To simplify discussion, we named the three studies we ran based on elements that are important to them. In the following sections we will thus discuss the Distractor, Scaling, and Measurement studies.

There were two objectives of the Measurement study: to set up a study to evaluate the strengths and weaknesses of the A-ATS Mask, and to illustrate using the SFNCS to characterize studies. The studies termed Distractor and Scaling were set up as we identified research questions that arose during reflective discussions over the set-up of the Measurement study, predominantly to ensure that our A-ATS stimuli were ecologically valid. In section 5.5 we show some breadth of the SFNCS by outlining a complimentary A-ATS study that highlights the benefit of the consideration of information being a Measurement or a Judgement, particularly the Form Subjectivity. This Judgement study was not prioritized and thus not run, but is differentiated as its questions involve terms and interpretation that are judgement-based. We discuss how we would attempt to structure the study and our expectations, if it were to be run later. Through this section, we will discuss how the SFNCS can be used for diverse studies.

To ease communication throughout this chapter, we indicate in maroon letters terms which refer to studies parameters, e.g. *FOCUS* indicating the focus of the Question participants are asked to perform.

5.1 Commonalities

The specific motivations for each study are detailed in their appropriate sections, but can be all summarized by our research questions, previously introduced in section 1.2, and discussed in detail in sections 5.2, 5.3 and 5.4. These research questions are similar as they all aim to investigate how some factors influence the ability to correctly conduct synoptic comparative tasks within multivariate

spatio-temporal data analysis, and their impact over trust while performing said tasks.
We set the three studies for different purposes: the Measurement study is set to help us understand a rich combination of factors, based on their categorization; the Distractor study is set to help us understand if the presence of graphs not necessary to perform tasks evaluated, named distractors, affect the ability to perform the tasks evaluated; the Scaling study is set to help us understand the impact of visual space allocated for the display of graphs.
Due to their common origin, these studies share many similarities. We thus first start this chapter by discussing in this section these common traits over the studies structures, the tasks asked of participants and the questions generated to ask them to perform the tasks, as well as similarities on data displayed and visual stimuli.

5.1.1 Study Structure, Tasks and Questions

The studies share a common structure, where each visualization presented to the participants is explained, followed by a series of elementary questions to ensure their understanding. The questions set to ensure participants' understanding of the visualizations are elementary identification tasks. Following the introduction questions, participants are asked to answer questions about data displayed to them in a composite A-ATS graphic for a series of stimuli. The resulting studies workflows are illustrated in Fig. 5.10,Fig. 5.24, and Fig. 5.35.

An important element for the design of our studies was to make claims about various types of tasks using the same composite graph. We thus set the studies with *QUESTIONS* with various *FOCI* . The structure to characterize the questions is defined in section 3.3. The questions of the different studies have thus been set to share the same structures but with different *FOCI* . For each *FOCUS* , various characteristics of the information displayed to the participants could potentially be evaluated and have participants tasked with retrieving such characteristics. The selection of the information participants are asked to retrieve is flexible and dependent on researchers' interests, but should still be justified. Our studies and our selection of information characteristic are strongly influenced by the work discussing the time mask [17, 8], but also by previous literature that shows the information that we ask participants to retrieve to be useful. The characteristics we ask participants to retrieve for each focus (*FOCUS*) are the following:

- WHAT_Qn: For questions with a *FOCUS* on the quantitative attribute, the information evaluated by the participants are the quantities associated with the moving object as they vary over time.
  It is important to evaluate participants' ability to retrieve this type of information, as it is necessary to establish relationships between different attributes. For our studies, participants were asked to consider proportions of time and variations of the quantitative attribute, according to the condition encoded by the A-ATS Mask.

- WHAT_Ql: For questions with a *FOCUS* on the qualitative attribute, the information evaluated by the participants are the time frames within which the status of the attribute is positive or

negative. Assessing participants' understanding of the information when displayed qualitative attributes is important, as such attributes are likely to be important for performing high-level tasks such as understanding relationships between attributes. In our studies, the qualitative attributes are binary. We decided to do the same as Andrienko *et al.* [17, 8] when discussing the time mask as multivalued qualitative attributes would increase the complexity of the data presented, and we had yet to evaluate the simpler case where the qualitative attribute is binary. For our questions, participants are asked to consider means and variations of the time frames in which the attribute is positive, according to the condition encoded by the A-ATS Mask.

- WHERE: For questions with a *FOCUS* on the spatial attribute, the information evaluated by the participants are the distances between the points of the trajectory and a point of interest (POI). Points of interest are important for inferring human movement and activities [61] and thus ensuring which tasks participants can perform using them is relevant. The inclusion of a point of interest also presented an advantage for the study setup: a point of interest with a set longitude and latitude meant a reduced risk of different interpretations, e.g. distance to a park, represented by an area, could be interpreted as distance to the closest point of the park to the trajectory, the park centre, or an arbitrary corner of the park. The questions ask participants to consider means and variations of the distances between the trajectory and the point of interest, according to the condition encoded by the A-ATS Mask.

Our interest was set on evaluating synoptic comparative *TASKS* that varied according to *FOCUS* . We aimed to evaluate participants' performance and assess their confidence in their performance. The records of trust are not meant to possibly invalidate the performances of the participants, but rather to nuance direct applicability for designers that would later wish to implement the design for their own work. If a visualization is effective but participants are not confident in the validity of their estimations, it could mean that additional efforts have to be made towards a clear communication of the design.
Thus, to assess participants' accuracy, we first planned to ask them whether they agreed with statements about the data displayed to them. For each stimulus, participants would be asked, for the *FOCUS* attribute, whether the value was on average higher while the condition was met, and whether the value was on average more stable while the condition was met. This answering method is a case of *Multiple choice* within our Response block. As further reflections and discussions occurred, we considered that this approach was suboptimal, as this would only indicate ability from participants to consider whether the difference between the means of the attributes during times in which the condition was met and the means of the attributes during times in which the condition was not met was greater than zero.

We thus considered that by asking participants to numerically assess this type of difference, the generated answers would be more nuanced and precise information about participants' ability to

perform the tasks asked of them.

We thus generated some prototype of survey generator that would ask such questions. The sentences asking to calculate the differences between means seemed fairly verbose, and the questions asking about calculations of variations were, despite our best efforts, impossible to generate as we could not find a reasonable manner to ask participants to provide a number that could describe the differences of variations.
We thus redesigned the questions structure for the studies. The first modification was to reconsider the change of the questions about variations. Due to the lack of commonly known and easily calculable method to quantify strengths of variations differences, we considered that this question was most relevant as a binary comparison of variations. Regarding the means assessments, we decided to split the one question asking to compare means directly into easier questions which were asking steps necessary to compare means. This approach is unorthodox, but is justified by the necessity for direct means comparisons to remember means values to compare them, which was not what we were trying to assess. Our aim was to evaluate participants' capability to assess data according to focus and status of Mask. We thus established their ability to do so by asking them to evaluate mean values when the mask is on and overall. The question of means comparisons was thus cut into three simpler questions:

- Assessment of the mean of the attribute while the condition is met.
- Assessment of the mean of the attribute over the whole time frame displayed.
- Percentage of the time in which the condition is met.

These three questions require participants to enter numerical responses. *Multiple choice* was considered as a Response option, but we considered the process would be distracting and typos likely to occur, albeit restrictions over them, could reduce them, which would thus be a *Form field* Response. We considered that maximal ease would result in the least distracting effort for participants to answer. Thus, we considered that providing the ability to answer everything with the mouse would be the most comfortable, intuitive and least distracting approach to provide answers, resulting in the three questions about numerical means being set to be answered using sliders, thus characterizing these Responses as *Slider* in the SFNCS. Additionally, we set the resolution level at 1, as a high level of precision in the input was likely to result in participants getting distracted.

The questions of the surveys are generated automatically in our JavaScript code that also generates the visualization. The questions thus adapt accordingly to the *FOCUS* required. The three first questions require the participant to use a slider to provide a numerical answer, and the fourth one asks the participant to claim whether they agree, disagree, or neither agree or disagree with the statement describing the data presented to them.
These changes effectively resulted in our studies SFNCS Response block being a *Slider*. Questions set to evaluate the ability to compare variations of the attributes remain *Multiple choice* in the SFNCS Response block.

The questions asked to participants are the implementations of the questions listed in the Question block illustrated in Fig. 5.1.
We list the structures of the sentences, with the italicized terms adapting to the displayed attributes to generate sentences that are comparable across *FOCI* and linguistically sound:

- WHAT_Qn:
  - Mean with Mask MwM
    What is the average value of the *attrDrawnQn* while the *sentenceCondition*?
  - Mean Overall MO
    What is the average value of the *attrDrawnQn* over the whole time displayed?
  - Mask Proportion MP
    For what percentage of the time displayed is the *sentenceCondition*?
  - Stability Comparison SC
    The *attrDrawnQn* is on average more stable while the *sentenceCondition* than at other times.
- WHAT_Ql:
  - Mean with Mask MwM
    For how long are both the *attrDrawnQl* and the *sentenceCondition*?
  - Mean Overall MO
    For how long is the *attrDrawnQl*?
  - Mask Proportion (MP)
    For what percentage of the time displayed is the *sentenceCondition*?
  - Stability Comparison SC
    The *attrDrawnQl* switches less frequently while the *sentenceCondition* than at other times.
- WHERE:
  - Mean with Mask MwM
    What is the average distance between the trajectory and the point of interest while the *sentenceCondition*?
  - Mean Overall MO
    What is the average distance between the trajectory and the point of interest?
  - Mask Proportion MP
    For what percentage of the time displayed is the *sentenceCondition*?
  - Stability Comparison SC
    The distance between the trajectory and the point of interest is more stable while the *sentenceCondition* than at other times.

The variations are calculated using the following equations:

- WHAT_Qn: $[abs(x_{(i)} - x_{(i+1)}) ... + abs(x_{(n-1)} - x_{(n)})]/n$
  With x being the quantitative values displayed, i representing indices, and n the number of data points used for the calculation, depending on Mask status selection.

- WHAT_Ql: cql/n
  With cql being the number of changes of status, and n being the number of data points used for the calculations, depending on Mask status selection.

- WHERE: $[dist(p_{(i)} - poi) ... + dist(p_{(n)} - poi)]/n$
  With p being the positions of the points selected and poi being the position of the point of interest, with the selection depending on Mask status selection.

The stability of the groups with the Mask status being positive and the Mask status being negative are compared using these calculations. Baselines for SC are set as calculations of variations of the time frames with the Mask status being positive being inferior to the variations of the time frames with the Mask status being negative, i.e. $variation_{(Mask)} <= variation_{(NoMask)}$.
The baselines are set independently of how big the difference between the variations are. To our knowledge, there are no recommendations concerning minimum differences of variation that can be visually perceived, but are aware that these vary due to our data generation set up. We thus decided to consider participants responding "neither agree nor disagree" to the statements provided to them as being neither correct nor incorrect, and filter them out from our analysis process.

As stated previously, we also considered it was valuable to understand participants' confidence in their ability to perform the tasks asked of them. It is important to evaluate the self-reported confidence of participants to perform tasks, as it is expected to relate to effort and persistence to perform tasks [118]. Thus, for each question, the participants were asked to evaluate their confidence in their answer, ranging from 0 to 5 included, with 5 meaning complete confidence and 0 meaning no confidence at all in their response. We loosely use the term trust to indicate the same notion of confidence, as these two terms relate to the same notions and are commonly used alternatively in informal conversations as well as literature.

Using the blocks of the SFNCS, we summarize the elements of the study in Fig. 5.1.

Task	Question		Response
Synoptic Measurement Comparison WHAT_Qn/WHAT_Ql/Where	Mean with Mask (MwM) Mean Overall (MO) Mask Proportion (MP)	Synoptic Measurement ***Lookup*** WHAT_Qn/WHAT_Ql/Where	Slider
	Stability Comparison (SC)	Synoptic Measurement ***Comparison*** WHAT_Qn/WHAT_Ql/Where	Multiple choices

Fig. 5.1 The Task, Question, and Response blocks from the SFNCS as they are used within our study. For each stimulus the task is the same, it is a Synoptic Measurement Comparison task, with a *FOCUS* on either a quantitative (WHAT_Qn), qualitative (WHAT_Ql) or spatial (WHERE) attribute. Characteristics of the questions are detailed in the Question block, in the right column. The left column indicates specific objects of the questions participants are asked to return. Note that the questions are not all the same as the task. While the questions are designed towards helping perform the task, the approach is not limited to questions matching the same characterization. The Response block indicates for each question the tool provided to participants their answers.

The number of responses we aimed to achieved was inspired by the study of Pena-Araya [134]. The number of participants we set for each study was calculated prior to the studies by considering for each the number of factors we aimed to use to assess ability to perform MwM , MO , MP and SC . Pilot studies using a convenience sample of 5 participants helped us to assess the length participants would take to answer each question, which informed our assessment of the total number of questions it would be reasonable to ask of a participant without them losing their attention.

We thus obtained the desired number of participants by dividing a certain number of stimuli by the number of questions per participants. The combinations of parameters are indicated in table 5.1.

	Measurement study	Scaling study	Distractor study
DATA COMPLEXITY	27	5	5
FOCUS	3	2	3
MASK COMPLEXITY	3	3	3
Distractor variations	1	1	2
Scaling variations	1	3	1
Number of stimuli	486	90	90
Number of responses in total (HITs)	1440	270	270
Number of participants	160	30	30
Number of responses per stimulus	3	3	3

Table 5.1 Distributions of participants and their answers according to the studies.

In the following sections, we discuss specifics of each study. We do not discuss every detail of the graphs due to the information produced, but discuss what we learned from the results. Our approach to analyse the results is strongly influenced by the studies ran by Heer *et al.* [70] and Pena-Araya *et al.* [134], but presents some differences. For the questions for which the answer provided is a number, we do not expect participants to answer with precision the value of the baseline generated. We

thus consider that a measurement more relevant for the ability of participants to answer the question asked of them is the difference between their answers and the baselines. We use BCa bootstrapping to generate 95% confidence intervals [48]. If the confidence intervals of differences do not cross 0, it provides evidence of differences being significant, with the further away from 0 and the smaller the CI the stronger the evidence.
Additionally, responses for questions in which participants indicate whether they agree, disagree, or neither agree nor disagree with a statement describing the data stability (SC) are evaluated by generating error rates. While both approaches would be potentially valid, we decided not to consider answers where participants claim they neither agree nor disagree with the statement as incorrect. We considered that if participants assess that they are not able to perform the task, then they are unlikely to make wrongful conclusions about the data following performing that task.

5.1.2 Data Displayed and Visual Stimuli

All the stimuli of the studies are produced using selections from the IEEE VAST 2014 Challenge data set, which has been enriched with quantitative and qualitative attributes. This process is detailed in section 4.4. The resulting data is categorized into ordered levels of complexity which we vary during the studies. The resulting characteristics for each category are illustrated in Fig. 5.2.

Data					
WHAT_Qn		**WHAT_QI**		**WHERE**	
Easy	Sinuosity min: 1.0 Sinuosity max: 1.5	Easy	# of qualities status variations: 1 # of qualities status variations: 3	Easy	# turns min: 2 # turns max: 8 Length min: 163.5 Length max: 245.8
Medium	Sinuosity min: 1.6 Sinuosity max: 2.1	Medium	# of qualities status variations: 4 # of qualities status variations: 6	Medium	# turns min: 6 # turns max: 12 Length min: 327.7 Length max: 464.7
Hard	Sinuosity min: 2.3 Sinuosity max: 2.8	Hard	# of qualities status variations: 7 # of qualities status variations: 9	Hard	# turns min: 14 # turns max: 23 Length min: 286.0 Length max: 499.7

Fig. 5.2 The Data block characterization for our studies. *Sinuosity* and *number of* (indicated by with the '#' symbol) are scale-independent, but that is not the case of the *length* used to characterize the trajectories that compose our studies. The lengths are defined by the pixels used to draw the trajectories. The number of turns that are part of the characterization of the trajectories were manually counted, which was possible in the context of the study data that consisted of trajectories of cars travelling along constrained routes in a block-based road system.

For each question, the participant is presented with a composite graph of visualizations overlain with an A-ATS Mask. We introduced the A-ATS Mask in section 4.3, and provided an example of the

resulting composite graph with an A-ATS Mask overlaid in Fig. 4.1. The different studies present variations that allow us to use SFNCS to structure stimuli and observations regarding the importance of different aspects of the graphs we produced. We detail our motivations for these changes in sections 5.2, 5.3 and 5.4.

The categorization of the attributes displayed in our study sets our expectations over participants. We overall expect higher data complexity to result in lessened performances for MwM , MO , MP and SC . Our overall expectations for the results were thus the following:

- Measurement study:
  - The higher the complexity of the attribute focused on, the higher the difference between answers and baselines.
  - The higher the complexity of the masks displayed, the higher the difference between answers and baselines.
  - Differences between answers and baselines will be smallest for the questions with a qualitative focus and highest for those with a spatial focus.
  - The higher the complexity of the attribute focused on, the lower the self-reported confidence.
  - The higher the complexity of the masks displayed, the lower the self-reported confidence.
  - Self-reported confidence will be smallest for the questions with a qualitative focus and highest for those with a spatial focus.

  These expectations are also shared for the Distractor and Scaling studies.

- Distractor study:
  - Differences between answers and baselines will be higher with all elements displayed.
  - Self-reported confidence will be lower with all elements displayed.

- Scaling study:
  - The higher the scaling of the attributes displayed, the lower the differences between answers and baselines.
  - The higher the scaling of the attributes displayed, the higher the self-reported confidence.

5.1.3 Control of data for the studies

We have previously discussed in section 4.4 the process that justified the selection of the IEEE VAST 2014 Challenge data set and why and how we enriched it. In this section, we discuss how we controlled the selection of trajectories and quantitative and qualitative attributes to possess a data set

categorized that would fit our studies needs.

Qualitative and qualitative values were generated and attributed to the trajectories according to expectations concerning complexity. The process generation was done using Perlin noise [138] to produce realistic-looking data. The data was then uploaded to the MongoDB database, so it could be queried according to the categories desired for each stimulus.
As part of our process to verify participants were not likely to often face questions for which the baselines would be the same values, which could have resulted in a confounding factor with participants repeating the same answers, we analysed the distributions of the baselines for the questions asked of them. We illustrate the distribution of baselines in figures 5.3, 5.4 and 5.5. We iterated through several generations of the attributes while modifying parameters to ensure a diverse distribution of baselines.

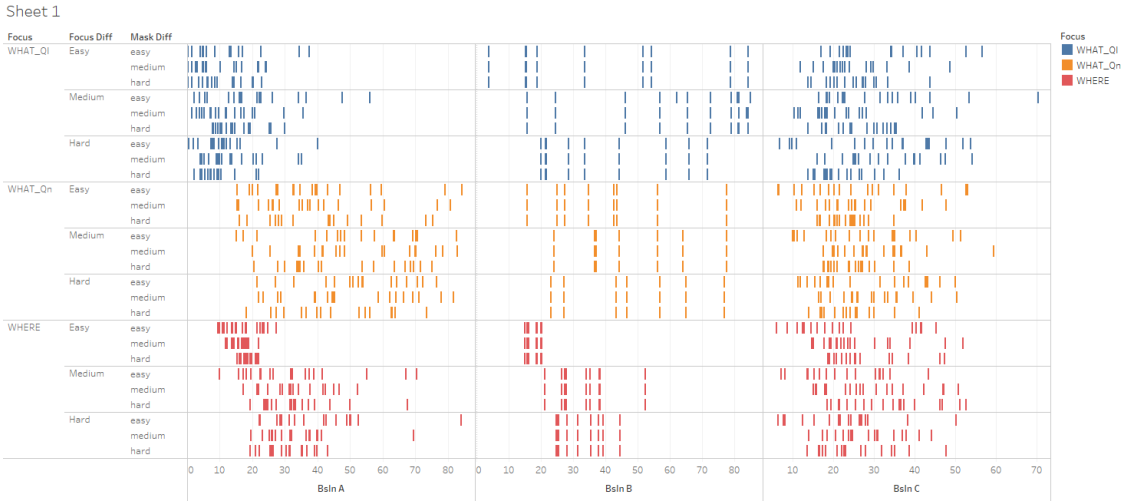


Fig. 5.3 The distribution of baselines for questions asking participants to provide numerical answers (MwM , MO , MP), according to categories.

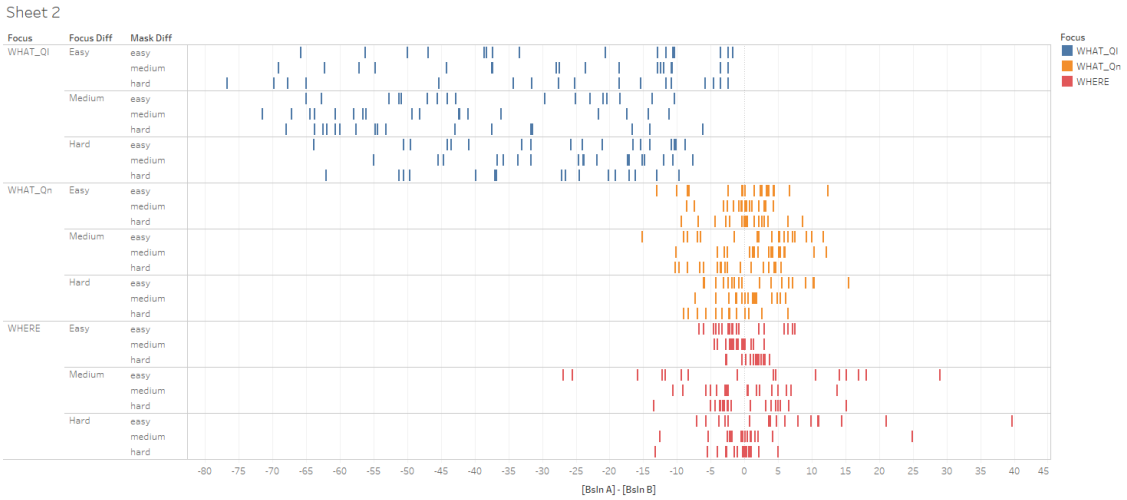


Fig. 5.4 The differences between the distributions of baselines asking about the means with either the condition being met or overall (MwM against MO). The details of the differences between the baselines present little interest to us, but ensuring diversity within them was important to avoid participants potentially encountering questions with exactly the same answer several times, resulting in a confounding effect.

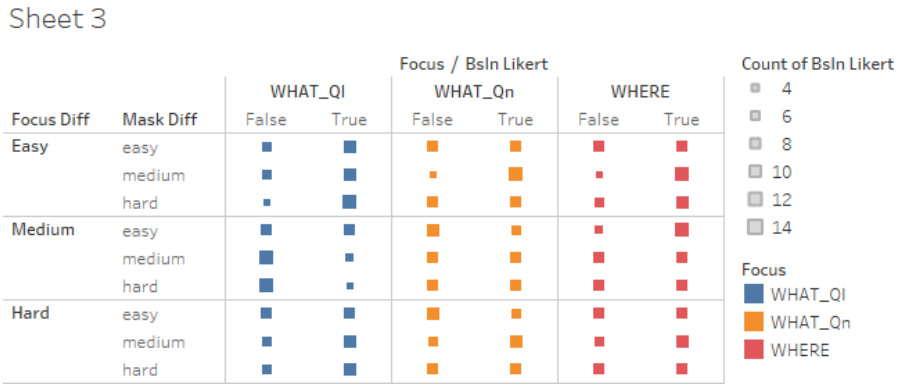
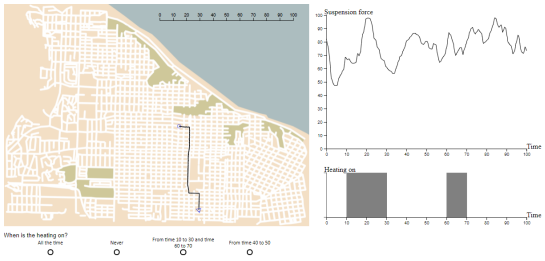


Fig. 5.5 This graph displays the distribution of baselines for questions asking participants whether they agree with statements describing variations of the data (SC). Details about these questions are discussed in section 5.1.1. The Likert questions can be either True or False; each block represents for a certain combination of studies factors the number of baselines which are either True or False. This graph indicates that while there is a diversity of baselines distributions, there is no combination of factors for which all the baselines are a single value, which could have created a potential confounding effect.

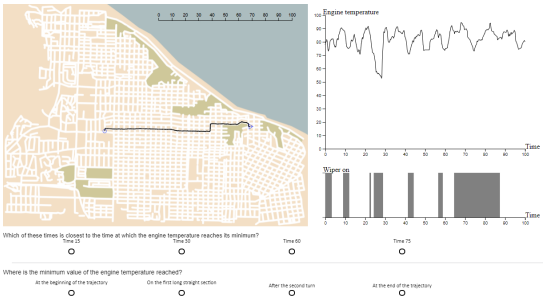
5.1.4 Considerations of answers quality

To ensure that participants understood the visualizations presented to them, we set up elementary questions following the visualizations' introduction. The questions of the introduction are displayed in figures 5.2a, 5.2b, 5.2c and 5.2d. We name this set of questions the introduction test (**TI**).

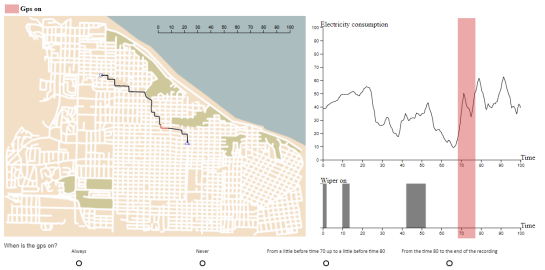
(a) The first question asked to check participants' understanding of the representation of the trajectory and its direction. We name this introduction question A.



(b) The second question to check participants' understanding of the representation of the qualitative attribute evolution over time. We name this introduction question B.



(c) The third and fourth questions to check participants' understanding of the representation of the quantitative attribute evolution over time. Answering this question requires the participant to connect the explicit depiction of time of the quantitative attribute's axis to its implicit representation in the display of the trajectory. We name the first question C1 and the second C2.



(d) The last question to check participants' understanding of the A-ATS Mask.

Table 5.2 The questions asked to participants after they were introduced to the visualizations. Except for the first question, all the composite graphs present three types of information. The left graph is set to communicate WHERE information: Trajectory position in space varies over time. The top right graph is set to communicate WHAT_Qn information: a numeric quantity varying over time. The bottom right graph is set to communicate WHAT_Ql information: a binary quality varying over time. We name this introduction question D.

Prior to our recruitment of paid participants, we first ran a series of tests with a small convenience sample [56]. All participants answered all questions without errors in the introduction test. Our convenience sample was not composed of data visualization experts, giving us some evidence to suggest that our introduction questions were understandable without prior expertise in the field.

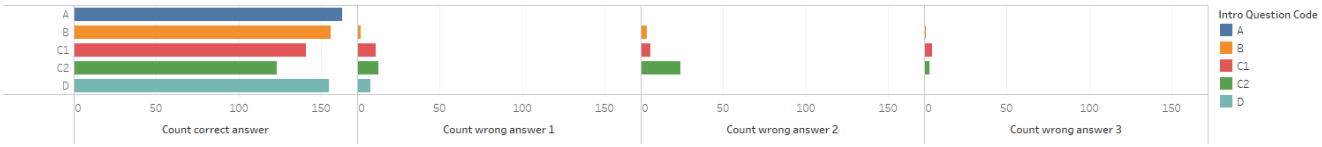


Fig. 5.6 The distribution of answers for the introduction test (TI). The questions are listed in the order they are presented to participants, as illustrated in table 5.2. The number of participants answering correctly are displayed on the left and the number of responses for alternative potential answers are displayed in the three other bar graphs.

We thus assumed that Prolific participants with a maximal approval rate, who claimed to be fluent in English (**FiE**) would perform well, at least for the introduction test, but still they performed poorly at the introduction tests. Some participants who answered poorly were contacted through the Prolific messaging platform. These discussions indicated that some participants may have been less fluent in English than they claimed, as their answers were poorly written, e.g. typos, wrong order of words, or that others participants directly admitted either not understanding English or the tasks asked of them. Additionally, we were alerted that a high surge of registrations on Prolific followed a viral post on TikTok that may have polluted the Prolific panel and resulted in low performance, on the 24th of July 2021 [43].
We thus decided to run the studies again with a narrower set of participant selection criteria, limiting participation to our study only to participants who claimed English as a First Language (**EFL**) and who registered to use the Prolific platform before the influx of July 24th 2021.

To assess whether the ability to successfully answer the introduction test (**TI**) was significantly different between these two groups of participants, we ran a CHI-Square analysis. The result is illustrated in Fig. 5.7.

Results						
	post-influx FiE	pre-influx EFL				***Row Totals***
TI : pass	123 (123.23) [0.00]	110 (109.77) [0.00]				233
TI : fail	60 (59.77) [0.00]	53 (53.23) [0.00]				113
Column Totals	183	163				**346 (Grand Total)**

The chi-square statistic is 0.0029. The *p*-value is .957124. The result is *not* significant at $p < .05$.

Fig. 5.7 The CHI Square test to compare introduction test (**TI**) performance, according to whether participants registered to Prolific after the TikTok influx and are Fluent in English (**FiE**) or registered prior to the influx and clam English as a First Language (**EFL**).

Our experiment showed no difference in performance for the introduction test between the pre-influx EFL and the post-influx FiE. This surprised us, but this test output allows us to make the following claims:

- *Post-influx FiE did not perform differently to pre-influx FiE participants.*

- *Prolific participants with a maximal approval rate found the elementary questions that required them to read a graphic challenging - much more so than our convenience sample.*

These introduction test results hindered our trust in the quality of the answers we collected. We considered that further evaluating the quality of the answers could prove valuable to reinforce said trust in the answers provided. We thus investigated for additional tests that could reinforce our confidence in the data quality. We looked at detailed answers provided by participants, searching for characteristics in the answers that were likely to indicate answers of low quality. The first characteristic we noticed was that some participants answered the studies particularly quickly. Time was not part of our interest, but could indicate low quality of answers. This statement has to be nuanced, as many factors can influence time necessary for a participant to perform a study, e.g. distractions can force participants to perform another activity in the middle of the study, or experienced participants might not spend as much time on the introduction sections of the studies. According to their experience in answering surveys, participants will perform differently [34], and thus we considered that time was an interesting approach to filter out answer sets to investigate in detail, but not to characterize answers as invalid.
We thus aimed to set an additional data quality checking over the answers provided, which was informed by observations of participants' answers that seemed to be of low quality due to short study completion times (rapid responses). The detailed analysis of these answers indicated two additional characteristics that we could use to filter out low performance:

- Some participants constantly self-reported a confidence of 0 out of 5 or 5 out of 5 for all the questions asked to them. If participants constantly self-reported a confidence of 0 or 5, for the total of 45 questions (4 questions for each of the 9 stimuli), we consider that these answers are of low quality.

- One specific error participants can make is particularly interesting: for questions with a WHAT_Ql focus, participants are asked to estimate for how long are both the qualitative attribute on and the condition met (WHAT_Ql MwM). We noticed several participants who provided higher values for WHAT_Ql MwM than for WHAT_Ql MO , which is impossible, as WHAT_Ql MwM is a subset of WHAT_Ql MO . Responses in which WHAT_Ql MwM is higher than WHAT_Ql MO indicate low answer quality, either due to lack of engagement or understanding of either the visualization or the question formulation.

We thus set an additional test, termed the Rigorous Test (**TR**), which verifies whether the participants did not constantly self-report a confidence of 0 or 5 (out of 5), and did not provide answers with higher values for WHAT_Ql MwM compared to WHAT_Ql MO . We verify whether the differences in distributions of participants failing the TR is significant according to post-influx FiE or pre-influx EFL and illustrate the result in Fig. 5.8.

Our results indicate no difference in ability to pass the Rigorous Test, between the pre-influx EFL participants and the post-influx FiE participants. This once again surprised us, but our results allow the following claim to be made:

Results						
	post-influx FiE	pre-influx EFL				***Row Totals***
TR : pass	85 (79.34) [0.40]	65 (70.66) [0.45]				150
TR : fail	98 (103.66) [0.31]	98 (92.34) [0.35]				196
Column Totals	183	163				**346 (Grand Total)**

The chi-square statistic is 1.5157. The *p*-value is .218277. The result is *not* significant at $p < .05$.

Fig. 5.8 The CHI Square test to compare ability to pass the Rigorous Test (**TR**) according to whether participants registered to Prolific after the TikTok influx and are Fluent in English (**FiE**) or registered prior to the influx and clam English as a First Language (**EFL**).

Post-influx FiE participants did no perform differently to pre-influx EFL participants in terms of the Rigorous Test.

The statistical tests we ran over the introduction questions indicated that participants from our online panel performed less well than we expected given our experience with the convenience sample during piloting, even with an explanation of the composite graphs and examples discussing them.

Orientation of the spatial stimuli for the Measurement study

A minor study parameter not mentioned previously is that half the participants of the Measurement study were presented to a stimulus of a map and its trajectory and if present point of interest flipped horizontally and vertically. This approach does not modify in any way the questions nor answers for participants, but was set up to try to reduce directionality of trajectory as a confounding factor, as we noticed a majority of trajectories records were going from left to right, potentially easing the implicit connection between time frames where attributes of the moving entities match the condition displayed by the A-ATS Mask. We run a CHI Square test to assess whether the difference in ability to pass TI is influenced by the orientation of the spatial stimuli. We display the result in Fig. 5.9

Results						
	no flip	flip				*Row Totals*
TI: pass	59 (53.99) [0.47]	51 (56.01) [0.45]				110
TI: fail	21 (26.01) [0.97]	32 (26.99) [0.93]				53
Column Totals	80	83				163 (Grand Total)

The chi-square statistic is 2.8106. The *p*-value is .093645. The result is *not* significant at $p < .05$.

Fig. 5.9 The CHI Square test to compare ability to pass the Introduction Test (**TI**) according to whether participants saw a spatial stimulus (map, trajectory, point of interest) that was flipped horizontally and vertically. The results are not significant, and we thus do not make further separation according to this criterion when discussing answers of the Measurement study.

The difference according to whether the spatial stimuli is flipped is not significant, and thus discussions of results of those answers do not make any distinction between these two groups.

5.1.5 Analysis of results

The studies are built using the SFNCS with the same method to characterize the data and its resulting stimuli, and participants are asked to perform the same tasks for each study. We thus construct our argumentation for results analysis using the same approach for each study.
For each task MwM , MO , MP and SC we assess whether performance is dependent on the factors used to set up the studies: i.e. *FOCUS* and its *DATA COMPLEXITY* , *MASK COMPLEXITY* , and scaling and presence of distractors for the relevant studies.
We use the same approach to evaluate characteristics of the answers.

Throughout our analysis of studies results, we use data characteristics of responses by either using them for comparisons, or to factorize the results. The characteristics selected to view results and make comparisons are called ***variants*** *and for readability are indicated by icons on the left of graphs displaying the responses. The* ***factors****, which split the responses into groups according to data characteristics are, if they are present, indicated by icons on the right of the graphs displaying the responses.*

We structure our analysis of the results by searching whether the following statements match participants performances:

- MwM
  - MwM 0 - is reasonable / acceptable
  - MwM 1 - is independent on variant
  - MwM 2 - is consistent in direction according to the variant

- MO
  - MO 0 - is reasonable / acceptable
  - MO 1 - is independent on variant
  - MO 2 - is consistent in direction according to the variant
- MP
  - MP 0 - is reasonable / acceptable
  - MP 1 - is independent on variant
  - MP 2 - is consistent in direction according to the variant
- SC
  - SC 0 - is reasonable / acceptable
  - SC 1 - is independent on variant
  - SC 2 - is consistent in direction according to the variant

Both considerations of performances being acceptable and performances varying consistently with the variant rely on subjective judgements that we, as researchers, make when considering that the strengths of the observed effects are worth reporting and draw conclusions from. Similarly, observations about performances being independent on variant are dependent on our interpretation of results in their global context. Considering results in their overall context is done to reduce likeliness of claims over false positive. We thus consider that factored results can be used as a method to identify potential false positive for global trends.
The following sections use graphical methods to present the results and as a basis for discussing them. All the studies will present the same type of graph.
The numerical questions (MwM , MO and MP) are analysed using the log absolute error as measured by Cleveland and McGill [48] and subsequently employed by Heer and Bostock [70]: $log_2(|baseline - answer| + 1/8)$.
We also directly refer to values of Responses when communicating the importance of an effect we note.
Additionally, we discuss the relationships between the factors of interest and self-reported confidence in ability to perform the tasks. Significance of differences between self-reported confidence is evaluated with first a singular Kruskal-Wallis test, followed by Dunn's Multiple Comparisons [55].

Graphs discussing numerical questions display information about the questions (MwM , MO and MP) and are composed of 6 columns, presenting the following information, from left to right:

- Absolute error for MwM .

- Differences between absolute errors for MwM *.
- Absolute error for MO .
- Differences between absolute errors for MO *.
- Absolute error for MP .
- Differences between absolute errors for MP *.
  **In the difference graphics, those differences that are statistically significant are indicated with a red circle to the left of the confidence interval of the differences. We selected a p-value of $p < 0.05$, which is commonly used [82].*

Our graphs indicate the number of responses collected and used for their generation. This is particularly important as artefacts due to low number of responses occurred, i.e. no drawing of violin plot generated with very low number of responses.

Graphs discussing Likert questions display error rates according to factors, with two columns to display the following calculations, from left to right:

- Error rate for SC .
- Differences between error rates for SC . Statistically significant differences between factors are indicated with a red circle left to the confidence interval of the differences. Similarly, that for numerical questions, we selected p-value of $p < 0.05$ to label differences as statistically significant.

These graphs also indicate the number of responses used for their generation. It is interesting to note that violin plots are not drawn, and confidence intervals have a width length of 0 when all participants were either correct nor incorrect. These occurrences have to be nuanced, as it is probable that over a larger sample of responses some variations are still likely to occur.

Due to the high number of variations of arrangements of factors, we do not discuss them in detail. Instead, our analysis focuses on the display and analysis of variations that are the most relevant to our interests as expressed in our research questions 1.2.
The resulting observations for each study are listed in table 5.3. We detail our conclusions drawn from the data in the appropriate sections for each study. Our approach to assess strengths of claims is to first analyse the data without factoring it according to the data characteristic we wish to evaluate, and then assessing whether potentially significant observations are likely false positives by analysing answers with factored groups.

Self-reported confidence will be presented as stacked bar charts, with each coloured stack element presenting proportions of answers according to factors as well as the differences of confidences according to factors evaluated with the Dunn Test. The calculations of significance are listed in tables

5.4, 5.5 and 5.6.
Additionally, we assess the relationships between self-reported confidence and ability to perform MwM , MO , MP and SC using violin plots.

Study	Index	Variant	Factor	MwM 0	MwM 1	MwM 2	MO 0	MO 1	MO 2	MP 0	MP 1	MP 2	SC 0	SC 1	SC 2	img CI	img ER
D	0	Distractor	None	X	X	X	X	X	?	X	X	X	X	X	X	5.12	5.17
D	1	Distractor	*FOCUS*	X	X	X	?	?	X	X	X	X	X	?	X	5.13	5.18
D	2	Distractor	*MASK COMPLEXITY*	X	X	X	X	?	X	X	?	X	?	X	?	5.14	5.19
D	3	Distractor	*FOCUS _ DATA COMPLEXITY*	X	X	X	?	?	?	X	?	?	X	X	X	5.16	5.21
D	4	Distractor	*FOCUS _ MASK COMPLEXITY*	X	X	X	Y	?	Y	X	Y	?	?	?	?	5.15	5.20
S	5	Scaling	None	X	Y	?	X	X	X	X	?	X	X	X	X	5.25	5.30
S	6	Scaling	*MASK COMPLEXITY*	?	Y	?	X	X	X	X	?	?	X	X	X	5.27	5.9b
S	7	Scaling	*FOCUS*	Y	Y	Y	X	X	X	X	?	X	X	X	X	5.26	5.9a
S	8	Scaling	*FOCUS _ DATA COMPLEXITY*	Y	Y	Y	Y	Y	?	X	?	?	X	X	?	5.29	5.32
S	9	Scaling	*FOCUS _ MASK COMPLEXITY*	?	?	?	Y	?	?	?	?	?	X	?	?	5.28	5.31
M	10	*FOCUS*	None	?	Y	Y	?	Y	Y	X	X	X	X	?	X	5.36	5.43
M	11	*MASK COMPLEXITY*	None	X	?	?	X	X	X	X	?	?	X	X	X	5.38	5.45
M	12	*DATA COMPLEXITY*	*MASK COMPLEXITY*	?	X	?	?	?	X	X	X	X	X	?	?	5.41	5.48
M	13	*DATA COMPLEXITY*	*FOCUS*	X	Y	?	?	Y	?	X	X	X	X	X	X	5.40	5.47
M	14	*DATA COMPLEXITY*	*FOCUS _ MASK COMPLEXITY*	X	?	Y	?	Y	?	X	X	X	X	X	X	5.42	5.49

Table 5.3 Notes about the results gathered by our studies.Observations are organized according to the study: D for Distractor study, S for Scaling study and M for Measurement study. The Index column is used to discuss the conclusions drawn from these observations. The Variant indicates which answers are used to generate the confidence intervals. The Factor column indicates the characteristic of the study used to factor the answers. The tasks are listed according to their shortened titles as presented at the beginning of the section 5.1.5. 'X' indicate no evidence, 'Y' indicate evidence supported and '?' indicates evidence limited in the context which requires detailed discussion.

Study	Input	Question type	p value	p<0.05
Measurement	trust according to question type		0	TRUE
Measurement	trustA1 according to *FOCUS*	MwM	2.20E-16	TRUE
Measurement	trustA2 according to *FOCUS*	MO	2.20E-16	TRUE
Measurement	trustA3 according to *FOCUS*	MP	0.7654	FALSE
Measurement	trustB according to *FOCUS*	SC	0.006992	TRUE
Measurement	trustA1 according to *MASK COMPLEXITY*	MwM	0.194	FALSE
Measurement	trustA2 according to *MASK COMPLEXITY*	MO	0.523	FALSE
Measurement	trustA3 according to *MASK COMPLEXITY*	MP	0	TRUE
Measurement	trustB according to *MASK COMPLEXITY*	SC	0.43	FALSE
Measurement	trustA1 according to *DATA COMPLEXITY*	MwM	0.172	FALSE
Measurement	trustA2 according to *DATA COMPLEXITY*	MO	0.039	TRUE
Measurement	trustA3 according to *DATA COMPLEXITY*	MP	0.808	FALSE
Measurement	trustB according to *DATA COMPLEXITY*	SC	0.56	FALSE
Distractor	trustA1 according to Distractor	MwM	0.34	FALSE
Distractor	trustA2 according to Distractor	MO	0.711	FALSE
Distractor	trustA3 according to Distractor	MP	0.006	TRUE
Distractor	trustB according to Distractor	SC	0.015	TRUE
Scaling	trustA1 according to Scaling	MwM	0.113	FALSE
Scaling	trustA2 according to Scaling	MO	0.042	TRUE
Scaling	trustA3 according to Scaling	MP	0.18	FALSE
Scaling	trustB according to Scaling	SC	0.997	FALSE

Table 5.4 The results of the Kruskal-Wallis tests to assess significance of factors over self-reported confidence.

Study	Question	Characteristic	Comparison	Z	P unadjusted	P adjusted	Significant
Measurement	MwM	*FOCUS*	WHAT_Ql - WHAT_Qn	4.44	8.91E-06	2.67E-05	TRUE
Measurement	MwM	*FOCUS*	WHAT_Ql - WHERE	9.04	1.54E-19	4.62E-19	TRUE
Measurement	MwM	*FOCUS*	WHAT_Qn - WHERE	4.59	4.23E-06	1.27E-05	TRUE
Measurement	MO	*FOCUS*	WHAT_Ql - WHAT_Qn	6.48	9.04E-11	2.71E-10	TRUE
Measurement	MO	*FOCUS*	WHAT_Ql - WHERE	11.53	8.73E-31	2.62E-30	TRUE
Measurement	MO	*FOCUS*	WHAT_Qn - WHERE	5.05	4.34E-07	1.30E-06	TRUE
Measurement	MP	*FOCUS*	WHAT_Ql - WHAT_Qn	0.72	0.47	1	FALSE
Measurement	MP	*FOCUS*	WHAT_Ql - WHERE	0.47	0.64	1	FALSE
Measurement	MP	*FOCUS*	WHAT_Qn - WHERE	-0.25	0.80	1	FALSE
Measurement	SC	*FOCUS*	WHAT_Ql - WHAT_Qn	0.77	0.44	1	FALSE
Measurement	SC	*FOCUS*	WHAT_Ql - WHERE	3.03	0.0024	0.0073	TRUE
Measurement	SC	*FOCUS*	WHAT_Qn - WHERE	2.26	0.024	0.072	FALSE
Measurement	MwM	*MASK COMPLEXITY*	easy - hard	1.61	0.11	0.32	FALSE
Measurement	MwM	*MASK COMPLEXITY*	easy - medium	1.53	0.13	0.38	FALSE
Measurement	MwM	*MASK COMPLEXITY*	hard - medium	-0.05	0.96	1	FALSE
Measurement	MO	*MASK COMPLEXITY*	easy - hard	0.48	0.63	1	FALSE
Measurement	MO	*MASK COMPLEXITY*	easy - medium	-0.65	0.52	1	FALSE
Measurement	MO	*MASK COMPLEXITY*	hard - medium	-1.14	0.26	0.77	FALSE
Measurement	MP	*MASK COMPLEXITY*	easy - hard	3.66	0.00025	0.00075	TRUE
Measurement	MP	*MASK COMPLEXITY*	easy - medium	3.06	0.0022	0.0066	TRUE
Measurement	MP	*MASK COMPLEXITY*	hard - medium	-0.56	0.58	1	FALSE
Measurement	SC	*MASK COMPLEXITY*	easy - hard	1.29	0.20	0.59	FALSE
Measurement	SC	*MASK COMPLEXITY*	easy - medium	0.56	0.58	1	FALSE
Measurement	SC	*MASK COMPLEXITY*	hard - medium	-0.73	0.47	1	FALSE

Table 5.5 The results of the Dunn's Test with Bonferroni correction to assess whether differences of self-reported confidence are significant according to *FOCUS* and *MASK COMPLEXITY* in the Measurement study.

Study	Question	Characteristic	Comparison	Z	P unadjusted	P adjusted	Significant
Scaling	MwM	Scaling	0 - 1	1.71	0.088	0.26	FALSE
Scaling	MwM	Scaling	0 - 2	-0.43	0.66	1	FALSE
Scaling	MwM	Scaling	1 - 2	-1.84	0.065	0.19	FALSE
Scaling	MO	Scaling	0 - 1	1.93	0.054	0.16	FALSE
Scaling	MO	Scaling	0 - 2	-0.74	0.46	1	FALSE
Scaling	MO	Scaling	1 - 2	-2.31	0.021	0.062	FALSE
Scaling	MP	Scaling	0 - 1	2.36	0.018	0.055	FALSE
Scaling	MP	Scaling	0 - 2	-1.29	0.20	0.59	FALSE
Scaling	MP	Scaling	1 - 2	-3.19	0.072	0.22	FALSE
Scaling	SC	Scaling	0 - 1	0.14	0.89	1	FALSE
Scaling	SC	Scaling	0 - 2	-0.84	0.40	1	FALSE
Scaling	SC	Scaling	1 - 2	-0.90	0.37	1	FALSE
Distractor	MwM	Distractor	hidden - normal	-3.55	0.34	0.34	FALSE
Distractor	MO	Distractor	hidden - normal	-2.34	0.71	0.71	FALSE
Distractor	MP	Distractor	hidden - normal	-5.22	1.81E-07	1.81E-07	TRUE
Distractor	SC	Distractor	hidden - normal	-3.26	0.0011	0.0011	TRUE

Table 5.6 The results of the Dunn's Test with Bonferonni correction to assess whether differences of self-reported confidences are significant according to scaling and presence of elements not necessary to perform the task in the Scaling and Distractor studies.

Data transformation for results analysis

In this section, we discuss the process to transform the data resulting from the studies and generate visualizations that help researchers for their analysis. This section is specific to our studies structure, but constitutes an example for potential future researchers considering the best approach to evaluate study results of their own. Our approach is strongly influenced by the works of Heer *et al.* [70] and Pena-Araya *et al.* [134]. Our approach is based on the display of 95% confidence intervals to analyse distributions of answers. Two calculations are used to produce the confidence intervals. According to whether the question requests participants to provide a numerical answer or a binary answer, the confidence intervals calculate either differences between answers from participants to baselines priorly calculated or their error rate [130]. The choice to calculate differences between numerical answers and baselines was due to the fact we could not expect participants to provide the actual correct value precisely, and thus our approach differs on this calculation compared to error rate. This approach is often used for evaluating measurements, e.g. answering time [134, 70], blood pressure [6]. The results are returned as CSV files from the Qualtrics website. Each variation of the study is its own Qualtrics study, e.g. the studies with and without the distractors result in two different studies and thus two different CSV files. For each file, we filter the answers from participants who failed to answer correctly the questions we set up in the introduction to assess their understanding of the visualizations displayed. The Puppeteer code that was run to generate the stimuli also generates a JSON file in which, for each image, important information about it is stored, e.g. the focus of the question, the complexity of the elements displayed or the baselines of the questions asked. Using that JSON file, we use a Node.js code we wrote in which the questions are transformed into a table in which each set of answers to a stimulus are associated together with the information of the data presented in the stimuli, i.e. the answers to the questions A1, A2, A3 and B, the self-reported trust levels, and information such as complexity of the data displayed or baselines and differences between answers and baselines. The data are then aggregated accordingly, with a new column indicating their group, e.g. for the Scaling study, a "scaling" column is added to indicate for each stimulus the scaling of the visualizations.

It's important to note that we implemented an additional filter for the results. First analysis of results indicated some occurrences of answers which seemed unlikely if participants were focusing and understanding what was asked of them, e.g. participants finishing the studies drastically faster than the average, always reporting confidence in their answers as either 0 or 5 over every single of their answers, or providing numerical answers which are not possible. We used the Prolific messaging system to discuss with participants who provided these odd answers. The ones who answered often wrote messages with multiple typos, or even admitted a poor understanding of what was asked of them, or simply of the English language. We discuss the implications of these observations in section 5.1.4.

We thus generated an R code that filters out answers from participants who constantly reported a confidence of 0 or 5, and those who provided answers which indicated misunderstanding of the

sentences or visualizations presented to them. The R code then generates, for the filtered data, bootstraps to produce results graphs to help us analyse the answers. The code is set with a series of functions that split the results by factoring according to the categories. The code allows for any organization of factoring, but we noticed that splitting the data too many times resulted in difficulties to compare confidence intervals. To simplify the analysis of the results, we calculate differences of confidence intervals and indicate with red dots juxtaposed the cases for which the differences are statistically significant.
Additionally, to ensure nuanced analysis of results, we use Violin plots [74] to display the distribution of the differences between provided responses and baselines.

5.2 The Distractor Study

5.2.1 Study motivation and structure

Within the scope of the studies we ran, each task requires only one of the visualizations shown in the composite graph to be consulted. The implications being that instead of considering the A-ATS Mask as an aggregated entity, it could be valid to consider evaluations made with A-AT Masks and A-S Masks. Our design choice is unorthodox and originates from the need to use a common visual structure, such a visualization composite graph, for tasks of varying complexity to achieve ecological validly through realistic stimuli and to ensure that results are comparable across a range of studies. The elements of the composite graph that are not required for the participants to perform elementary tasks may act as distractors. We communicate their presence or not according to the distractor status: hidden (h) or normal (n).

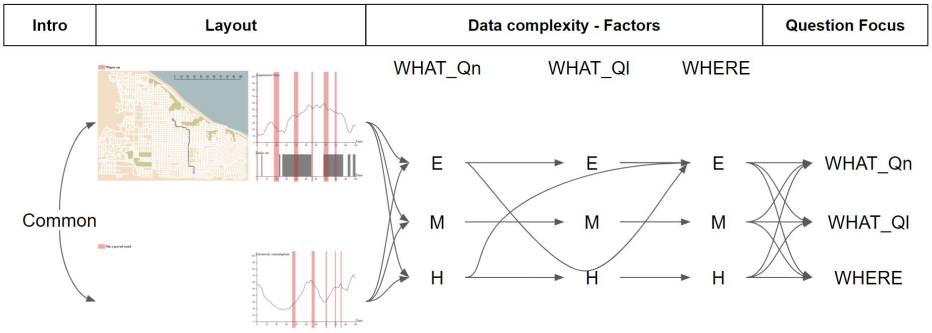


Fig. 5.10 The workflow of the distractor study. Participants are split into two groups. After following the same introduction, they are then presented with a series of nine questions with data complexity varying as defined in Fig. 5.11, but with one group being shown the full composite graph, i.e. with the distractor status being normal (n), and one group being shown only the necessary elements to perform the tasks, i.e. with the distractor status being hidden (h).

Qualitative studies set to assess performances in complex tasks very often display composite graphs of visualizations and additionally some studies are set by having participants interact with

prototype software that allows them to vary the information presented to them [184, 58, 78, 16]. Another reason for the Distractor study was our intent to make a composite graph similar to what would be encountered in professional software, i.e. realistic. To achieve this, the stimuli presented to participants in our studies has to look as if it were records from real-life events, and that the composite graph looked similar to some encountered in the real world. We designed the SFNCS so that it could be used for the set up of complex studies, with tasks that are complex and require interactions with prototype software that researchers intend to evaluate. While interactions are not within the scope of this thesis, we expect that future work using the SFNCS will evaluate complex tasks that require them.

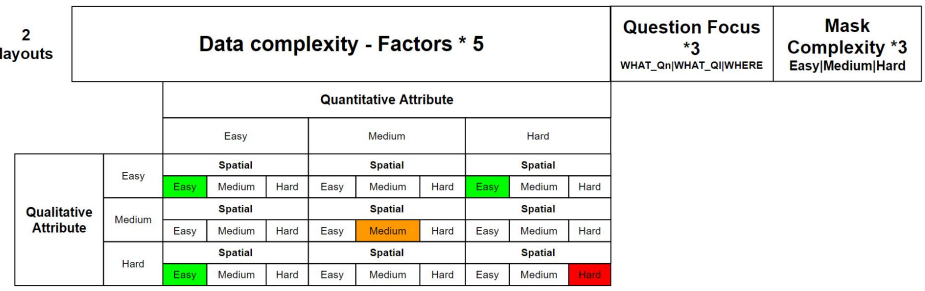


Fig. 5.11 The structure of the Distractor study. The categories vary according to data complexity, focus of the questions, and complexity of mask applied. We colour the data complexities used to generate stimuli presented to the participants of the Distractor study. The colours loosely indicate *DATA COMPLEXITY* used to generate the stimuli, with green indicating relatively Easy complexity, orange indicating Medium complexity, and red Hard complexity.

Thus, we set the Distractor study as a means of assessing the differences in performance between participants with or without the additional (irrelevant) information in the composite graph when performing synoptic comparative tasks. The results of the Distractor study will help us to answer the two following research questions that are intended to validate the results of our more comprehensive assessment of synoptic comparative tasks :

- **Research question 3.5**: How does the display of additional information that does not contribute to task completion impact the ability to correctly conduct synoptic comparative tasks within multivariate spatio-temporal data analysis?
- **Research question 3.6**: How does the display of additional information that does not contribute to task completion impact self-reported trust for conducting synoptic comparative tasks within multivariate spatio-temporal data analysis?

We illustrate the study structure in figures 5.11 and 5.10. The studies were set to compare between the group with information that does not contribute to task completion and the one composed solely of the necessary information. The objective of the study was not to go into details of each combination of categories, but rather to discuss whether observed differences in performance between the two groups were significant. Did the distracting information, that provides a realistic stimulus for ecological

validity, have an effect on elementary synoptic comparative tasks?

5.2.2 Results analysis

The Distractor study focuses on the impact of additional information that does not contribute to task completion. In line with the processes outlined in section 5.1.4, we rejected 40% of responses as participants failed the Introduction Test (TI). This large proportion of rejection for TI forced us not to filter out participants who failed the rigorous test (RT) to allow us to analyse the results.

The means of results according to the factors are illustrated in table 5.7.

study	factor1	factor2	factor3	responses	MwM	MO	MP	SC (mean correctB)	SC (error rate)	responses (neither)	SC (neither mean correctB)	SC (neither error rate)	str
d				237.00	9.72	9.67	5.56	0.39	0.61	175.00	0.53	0.47	
d	distractor			117.00	8.53	9.91	6.28	0.36	0.64	87.00	0.48	0.52	h
d	distractor			120.00	10.89	9.43	4.86	0.42	0.58	88.00	0.57	0.43	n
d	distractor	*FOCUS*		39.00	8.68	7.11	5.26	0.36	0.64	35.00	0.40	0.60	h-WHAT_QI
d	distractor	*FOCUS*		39.00	7.78	7.54	6.28	0.44	0.56	30.00	0.57	0.43	h-WHAT_Qn
d	distractor	*FOCUS*		39.00	9.13	15.09	7.31	0.28	0.72	22.00	0.50	0.50	h-WHERE
d	distractor	*FOCUS*		42.00	12.90	6.41	4.17	0.38	0.62	32.00	0.50	0.50	n-WHAT_QI
d	distractor	*FOCUS*		42.00	9.86	8.57	5.16	0.52	0.48	33.00	0.67	0.33	n-WHAT_Qn
d	distractor	*FOCUS*		36.00	9.75	13.97	5.30	0.33	0.67	23.00	0.52	0.48	n-WHERE
d	distractor	*MASK COMPLEXITY*		38.00	8.80	7.73	5.84	0.26	0.74	24.00	0.42	0.58	h-easy
d	distractor	*MASK COMPLEXITY*		38.00	7.74	11.12	8.25	0.32	0.68	32.00	0.38	0.63	h-medium
d	distractor	*MASK COMPLEXITY*		41.00	9.01	10.81	4.87	0.49	0.51	31.00	0.65	0.35	h-hard
d	distractor	*MASK COMPLEXITY*		40.00	11.03	11.71	3.38	0.50	0.50	27.00	0.74	0.26	n-easy
d	distractor	*MASK COMPLEXITY*		40.00	11.00	6.53	6.23	0.35	0.65	29.00	0.48	0.52	n-medium
d	distractor	*MASK COMPLEXITY*		40.00	10.64	10.06	4.97	0.40	0.60	32.00	0.50	0.50	n-hard
d	distractor	*FOCUS*	*MASK COMPLEXITY*	12.00	11.55	4.66	6.31	0.17	0.83	8.00	0.25	0.75	h-WHAT_QI-easy
d	distractor	*FOCUS*	*MASK COMPLEXITY*	13.00	3.87	6.37	5.37	0.31	0.69	13.00	0.31	0.69	h-WHAT_QI-medium
d	distractor	*FOCUS*	*MASK COMPLEXITY*	14.00	10.68	9.89	4.25	0.57	0.43	14.00	0.57	0.43	h-WHAT_QI-hard
d	distractor	*FOCUS*	*MASK COMPLEXITY*	14.00	5.78	4.85	4.43	0.36	0.64	10.00	0.50	0.50	h-WHAT_Qn-easy
d	distractor	*FOCUS*	*MASK COMPLEXITY*	12.00	11.20	11.85	10.77	0.50	0.50	11.00	0.55	0.45	h-WHAT_Qn-medium
d	distractor	*FOCUS*	*MASK COMPLEXITY*	13.00	6.78	6.46	4.12	0.46	0.54	9.00	0.67	0.33	h-WHAT_Qn-hard
d	distractor	*FOCUS*	*MASK COMPLEXITY*	12.00	9.56	14.16	7.00	0.25	0.75	6.00	0.50	0.50	h-WHERE-easy
d	distractor	*FOCUS*	*MASK COMPLEXITY*	13.00	8.42	15.20	8.80	0.15	0.85	8.00	0.25	0.75	h-WHERE-medium
d	distractor	*FOCUS*	*MASK COMPLEXITY*	14.00	9.41	15.77	6.19	0.43	0.57	8.00	0.75	0.25	h-WHERE-hard
d	distractor	*FOCUS*	*MASK COMPLEXITY*	13.00	10.72	7.71	2.86	0.62	0.38	8.00	1.00	0.00	n-WHAT_QI-easy
d	distractor	*FOCUS*	*MASK COMPLEXITY*	15.00	14.63	2.97	6.10	0.20	0.80	13.00	0.23	0.77	n-WHAT_QI-medium
d	distractor	*FOCUS*	*MASK COMPLEXITY*	14.00	13.07	8.88	3.32	0.36	0.64	11.00	0.45	0.55	n-WHAT_QI-hard
d	distractor	*FOCUS*	*MASK COMPLEXITY*	14.00	12.81	10.65	2.67	0.50	0.50	11.00	0.64	0.36	n-WHAT_Qn-easy
d	distractor	*FOCUS*	*MASK COMPLEXITY*	13.00	6.66	5.35	6.10	0.62	0.38	10.00	0.80	0.20	n-WHAT_Qn-medium
d	distractor	*FOCUS*	*MASK COMPLEXITY*	15.00	9.88	9.42	6.67	0.47	0.53	12.00	0.58	0.42	n-WHAT_Qn-hard
d	distractor	*FOCUS*	*MASK COMPLEXITY*	13.00	9.43	16.85	4.65	0.38	0.62	8.00	0.63	0.38	n-WHERE-easy
d	distractor	*FOCUS*	*MASK COMPLEXITY*	12.00	11.17	12.25	6.52	0.25	0.75	6.00	0.50	0.50	n-WHERE-medium
d	distractor	*FOCUS*	*MASK COMPLEXITY*	11.00	8.58	12.45	4.75	0.36	0.64	9.00	0.44	0.56	n-WHERE-hard
d	distractor	*FOCUS*	*DATA COMPLEXITY*	13.00	9.28	4.06	5.38	0.31	0.69	12.00	0.33	0.67	h-WHAT_QI-E
d	distractor	*FOCUS*	*DATA COMPLEXITY*	15.00	6.48	3.90	5.02	0.27	0.73	12.00	0.33	0.67	h-WHAT_QI-M
d	distractor	*FOCUS*	*DATA COMPLEXITY*	11.00	10.96	15.09	5.45	0.55	0.45	11.00	0.55	0.45	h-WHAT_QI-H
d	distractor	*FOCUS*	*DATA COMPLEXITY*	17.00	7.54	7.56	9.64	0.53	0.47	14.00	0.64	0.36	h-WHAT_Qn-E
d	distractor	*FOCUS*	*DATA COMPLEXITY*	7.00	12.64	13.86	2.69	0.29	0.71	5.00	0.40	0.60	h-WHAT_Qn-M
d	distractor	*FOCUS*	*DATA COMPLEXITY*	15.00	5.79	4.57	4.14	0.40	0.60	11.00	0.55	0.45	h-WHAT_Qn-H
d	distractor	*FOCUS*	*DATA COMPLEXITY*	22.00	6.11	13.88	6.59	0.45	0.55	13.00	0.77	0.23	h-WHERE-E
d	distractor	*FOCUS*	*DATA COMPLEXITY*	8.00	18.61	20.36	7.52	0.13	0.88	6.00	0.17	0.83	h-WHERE-M
d	distractor	*FOCUS*	*DATA COMPLEXITY*	9.00	8.06	13.35	8.89	0.00	1.00	3.00	0.00	1.00	h-WHERE-H
d	distractor	*FOCUS*	*DATA COMPLEXITY*	14.00	19.50	5.68	3.54	0.43	0.57	12.00	0.50	0.50	n-WHAT_QI-E
d	distractor	*FOCUS*	*DATA COMPLEXITY*	15.00	5.94	3.70	4.00	0.40	0.60	10.00	0.60	0.40	n-WHAT_QI-M
d	distractor	*FOCUS*	*DATA COMPLEXITY*	13.00	13.83	10.32	5.05	0.31	0.69	10.00	0.40	0.60	n-WHAT_QI-H
d	distractor	*FOCUS*	*DATA COMPLEXITY*	19.00	6.49	7.20	6.63	0.58	0.42	16.00	0.69	0.31	n-WHAT_Qn-E
d	distractor	*FOCUS*	*DATA COMPLEXITY*	6.00	20.35	17.16	3.51	0.17	0.83	4.00	0.25	0.75	n-WHAT_Qn-M
d	distractor	*FOCUS*	*DATA COMPLEXITY*	17.00	9.93	7.07	4.10	0.59	0.41	13.00	0.77	0.23	n-WHAT_Qn-H
d	distractor	*FOCUS*	*DATA COMPLEXITY*	20.00	5.95	12.71	3.39	0.45	0.55	13.00	0.69	0.31	n-WHERE-E
d	distractor	*FOCUS*	*DATA COMPLEXITY*	6.00	13.14	15.05	8.94	0.00	1.00	4.00	0.00	1.00	n-WHERE-M
d	distractor	*FOCUS*	*DATA COMPLEXITY*	10.00	15.32	15.84	6.94	0.30	0.70	6.00	0.50	0.50	n-WHERE-H

Table 5.7 Results summary for the Distractor study. The table indicates the absolute difference between responses from participants and the baselines, according to factors we varied. The columns on the left indicate the factors and the ones on the right indicate the detail of the factored selections, e.g. "h-WHERE-medium" indicates responses where distractors were hidden, the *FOCUS* was WHERE, and the *MASK COMPLEXITY* was medium. The columns with "neither" in their titles discuss the results once responses for which participants answered "Neither agree nor disagree" are filtered.

We list notes in the table 5.3 and use its indices to refer to them, but focus our discussion on our interpretation of these notes.

Numerical questions: Mean with Mask (MwM), Overall Mean (MO), Mask Proportion (MP)

The overall results for the Distractor study are displayed in Fig. 5.12, and results factored according to focus and Mask are displayed in figures 5.13 and 5.14.

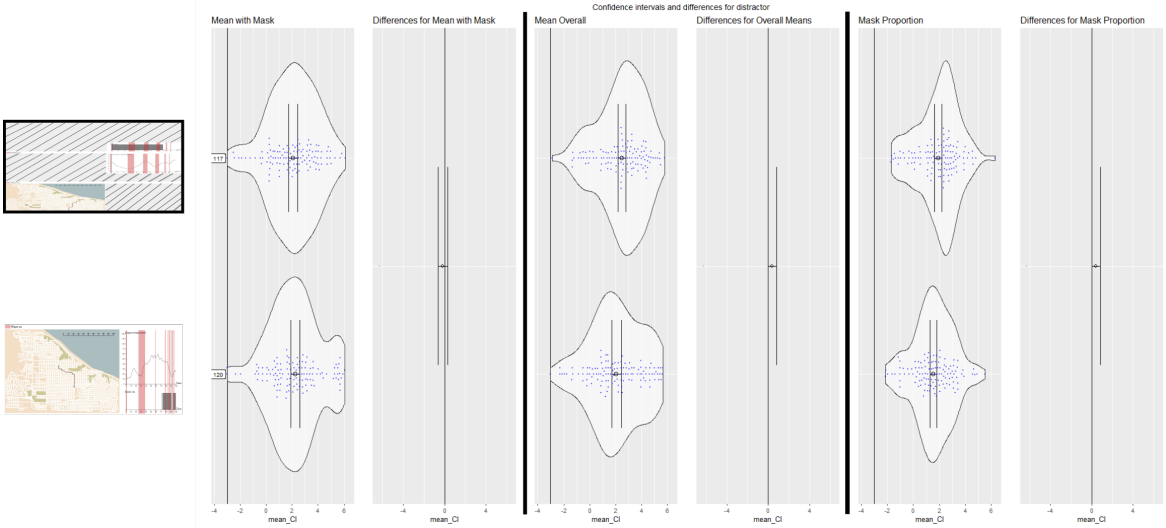


Fig. 5.12 The overall results (all *FOCI* , all *DATA COMPLEXITY* and all *MASK COMPLEXITY*) for MwM , MO and MP the Distractor study. The left icons indicate the variant: the top left icon indicate that graphical elements not required to perform the tasks are hidden; and the bottom left icon indicates all graphical elements are displayed - some are distractors in this full composite graph.

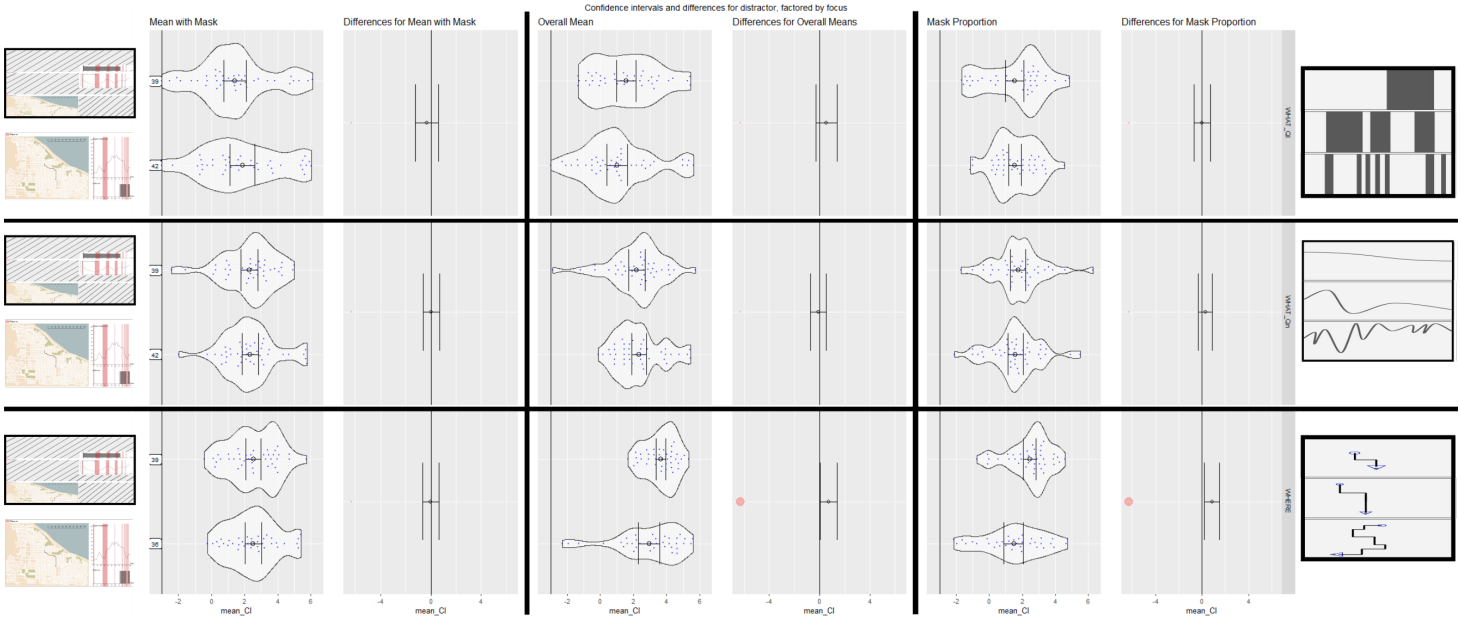


Fig. 5.13 The MwM , MO and MP answers factored according to *FOCUS* . We note occurrences of significant differences between display or not of distractor for MO and MP when the *FOCUS* is WHERE.

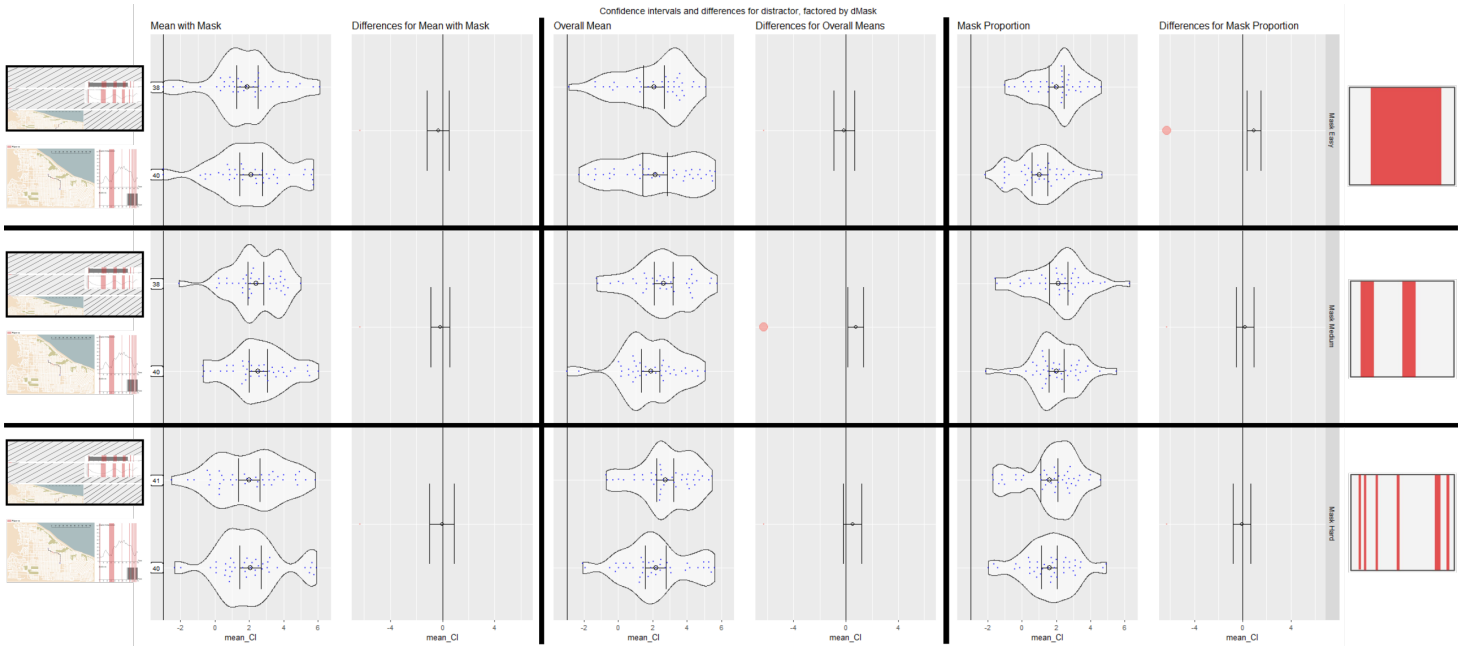


Fig. 5.14 The MwM , MO and MP answers factored according to *MASK COMPLEXITY* .

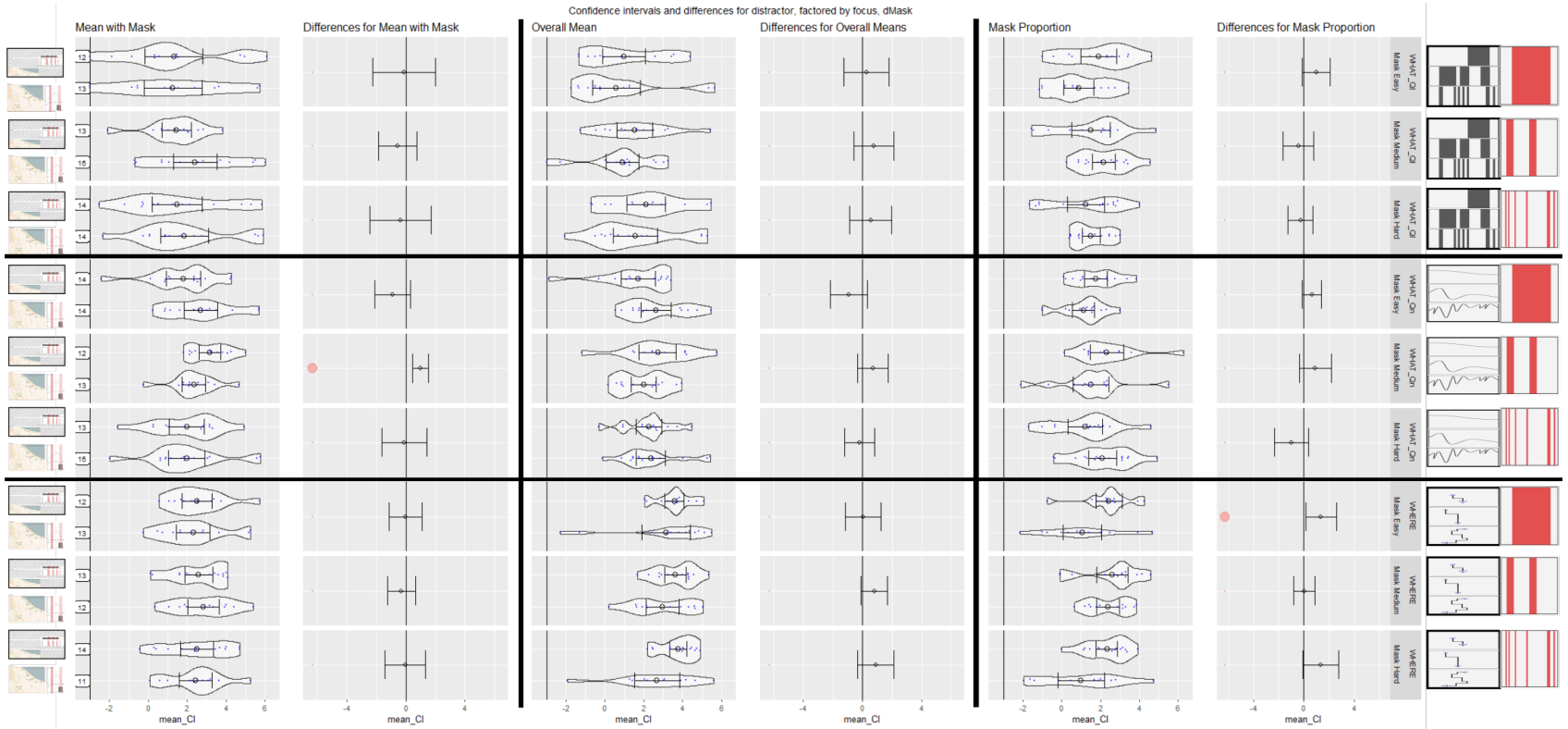


Fig. 5.15 The MwM , MO and MP answers factored according to *FOCUS* and *MASK COMPLEXITY* .

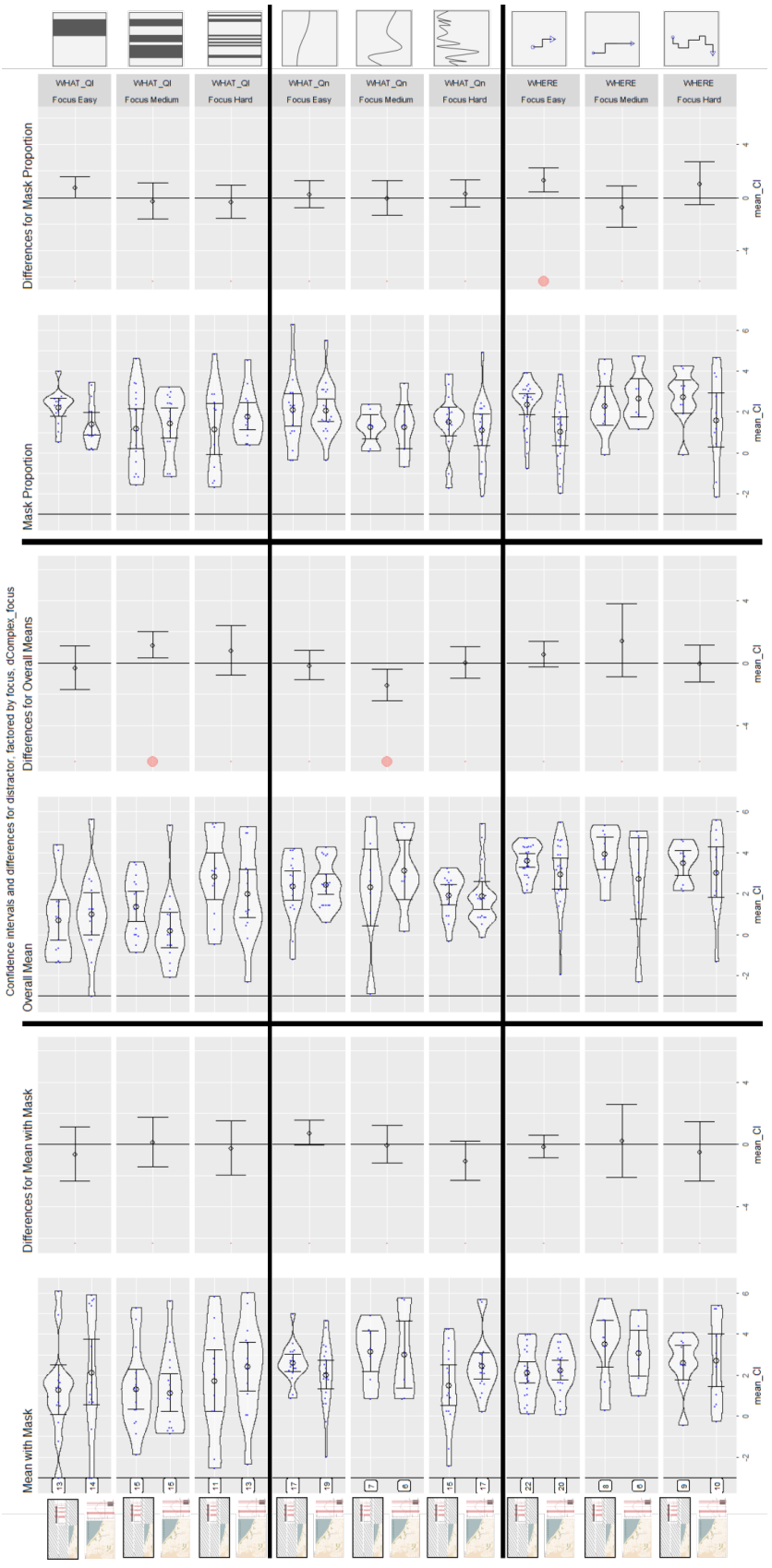


Fig. 5.16 The MwM , MO and MP answers factored according to *FOCUS* and *DATA COMPLEXITY* .

We report in index 0 of the table 5.3, a doubt about whether MO can be influenced by the display of elements that are not necessary for the completion of the task. The difference is significant in Fig. 5.14 when Mask is set to Medium. We notice in Fig. 5.15 that as MO is factored by both focus and Mask, the difference of performance for MO is not significant in this case, indicating the effect doesn't show a consistent pattern across factors. This indicates to us that the previous observations were likely false positive. This leads us to conclude that Mask has no significant effect over the ability to perform MwM , MO and MP when results are organized according to presence of elements not necessary to perform the tasks.
Fig. 5.13 indicates that both MO and MP are significantly poorer in the distractor condition for WHERE tasks. Once the results are factored by both *FOCUS* and the *DATA COMPLEXITY* , the claim is no longer true for MO , and only true for the factor WHERE Easy for MP . This leads us to conclude that previous observations were likely false positive, and that we do not have strong evidence that the focus has significant effects on the ability to perform MwM , MO or MP when results are organized according to the presence of elements not necessary to perform the tasks - the composite graph distractors.
We see no strong evidence that the distractors are having widespread or systematic effects that we can explain, and therefore speculate that the significant differences between the two conditions are spurious false positives.

Likert question: Stability Comparison (SC)

We observe in Fig. 5.17 that performance for Stability Comparison (SC) is fairly poor, independent on distractor condition. As reported in index 2 of table 5.3 and illustrated in Fig. 5.19, performances for SC are significantly poorer when the elements which are not necessary for the completion of the task are displayed and the complexity of the Mask is set to Easy. We notice that while not displaying significant differences, performances for SC are not following the same order according to whether the Mask is set to Medium or Hard. We notice that both when the Mask is Easy and Medium, participants with elements not necessary performed less well, but the opposite occurred with a Hard Mask. This statement is supported by the analysis of the SC answers as displayed in Fig. 5.18.

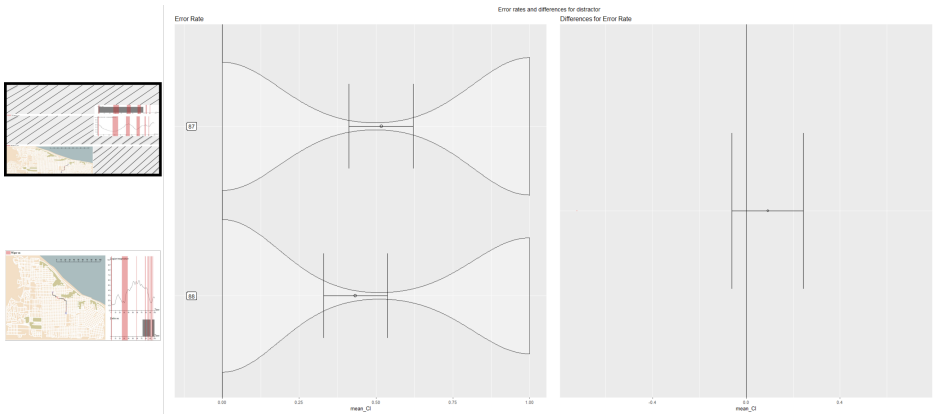


Fig. 5.17 The error rate for SC of the Distractor study with presence of elements not necessary to perform the task (distractors) as the variant.

Analysis of the SC when results are factored by both focus and the *MASK COMPLEXITY* , as displayed in Fig. 5.20. indicate that this trend is not global for each case where the Mask is set to Easy, but rather a unique case where the focus is WHAT_Ql and the Mask is Easy. We thus conclude that this difference is a false positive.

Note that Fig. 5.21, which displays error rates for SC , factored by both focus and *DATA COMPLEXITY* , does not contain information that strongly changes our conclusions. It does show very narrow confidence intervals for both cases of focus WHERE Medium and WHERE Hard. This is an artefact of all our answers being incorrect for these groups. Due to the low number of answers we have gathered for these groups, we do not consider these observations as valid for the interpretations of the results.

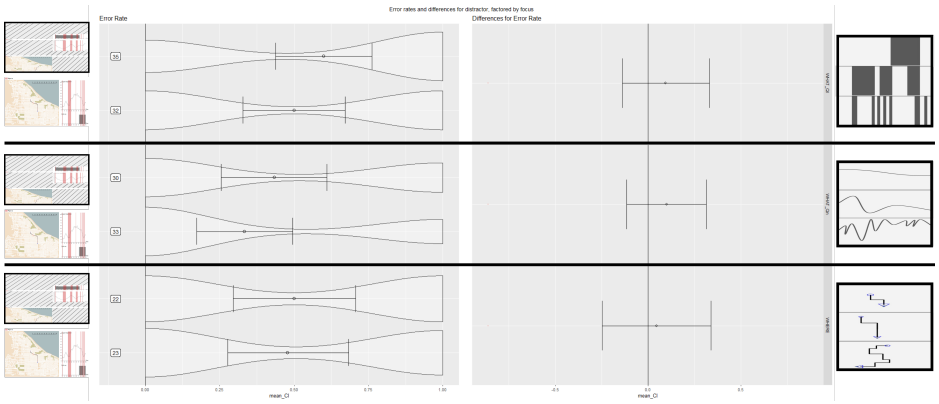


Fig. 5.18 The error rate for SC with presence of elements not necessary to perform the task as the variant and *FOCUS* as the factor.

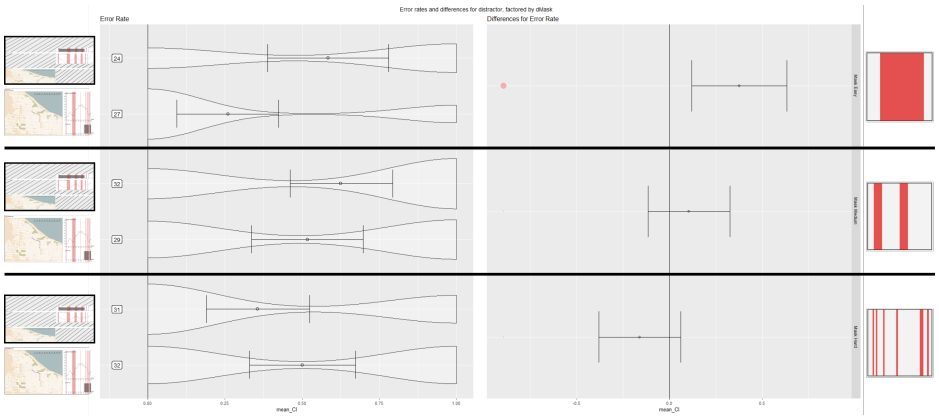


Fig. 5.19 The error rate for SC with presence of elements not necessary to perform the task as the variant and *MASK COMPLEXITY* as the factor.

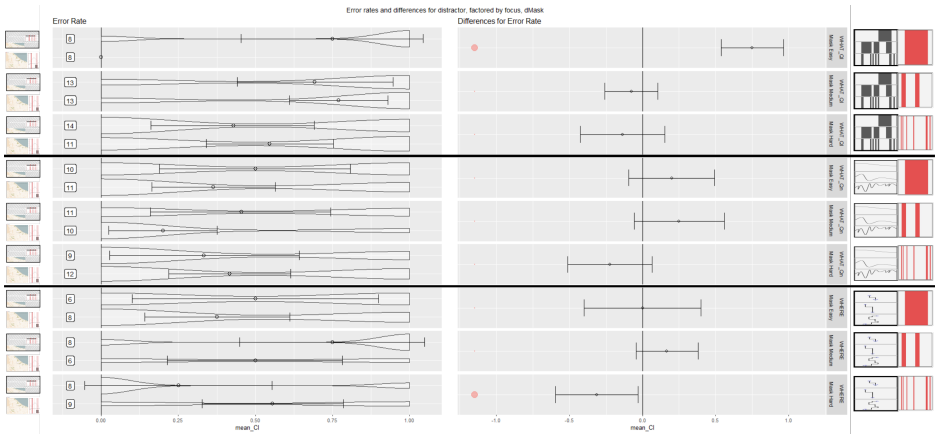


Fig. 5.20 The error rate for SC with presence of elements not necessary to perform the task as the variant and both *FOCUS* and *MASK COMPLEXITY* as the factor.

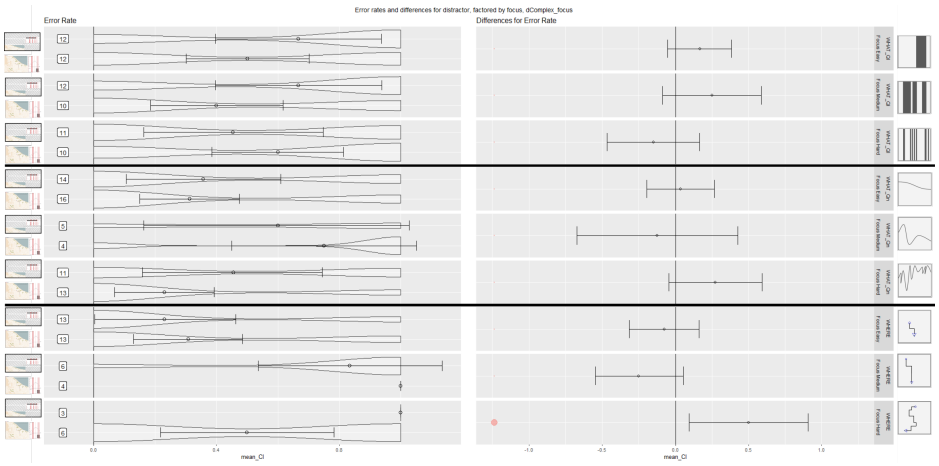


Fig. 5.21 The error rate for SC with presence of elements not necessary to perform the task as the variant and both focus and *DATA COMPLEXITY* as the factor.

Self-reported confidence

The bar charts in Fig. 5.22 indicate, for questions MwM , MO , MP and SC , self-reported confidence, factored by the distractor status: hidden (h) or normal (n).
Differences in self-reported confidence between the two groups for questions MwM and MO are not significant, indicating the presence of elements not necessary does not affect perceived ability to perform the task. For the question MP the Dunn test indicates that the difference is significant between the two groups. The mean self-reported confidence for graphs without elements not necessary to perform the task is 3.02 against 3.47 when those elements are displayed. We can thus claim that participants are more confident for MP when elements that are not necessary are displayed. In the composite graphs presented to participants, the Masks are displayed both over the WHAT_Ql and WHAT_Qn views, as well as indicated by the trajectory being coloured accordingly. We thus consider that this result could potentially indicate that the repetition of the Mask information reinforces participants' trust in their answer.
Additionally, the Dunn test between the two groups is significant for the SC question. The means of self-reported confidence are 2.84 for D(without) and 3.33 for D(within). As this effect is light and we are unable to consider an explanation for it, we consider that this is observation is likely a false positive.

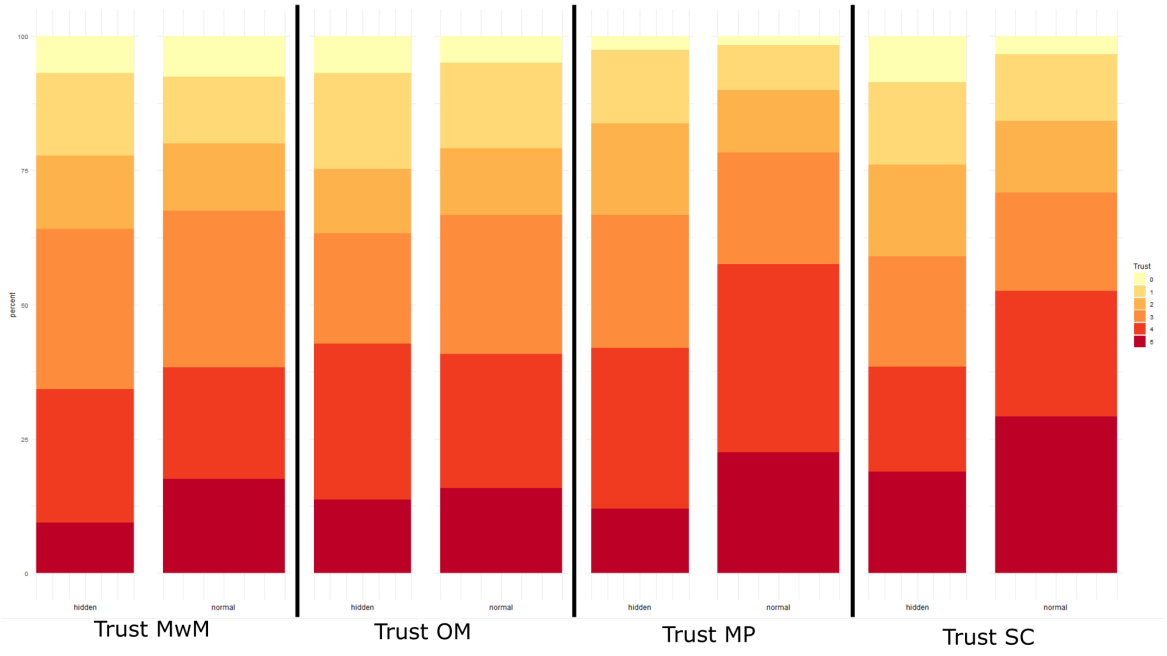


Fig. 5.22 The bar charts labelled as hidden indicate the reported self-reported trust according to question when there is no display of attributes which are not necessary to perform the task. The bar charts labelled normal indicate the self-reported trust according to question when there are distractors drawn. Bars are coloured from yellow to red, with 0 indicating no confidence at all in their answer and 5 indicating absolute confidence in their answer.

Conclusion

We verified whether characteristics of potential significance without factoring results presented the same significance once factored, for MwM , MO , MP and SC . No case of potentially significant characteristic was present when comparing different numbers of factors. This suggests that the characteristics with a significant impact over performances were likely false positives, the chances of our highly factored analysis returning these being high given the 5% significance levels used in our testing and the lack of any post-hoc correction to account for family effects. From this we conclude that there is no evidence that the presence of graphs which are not necessary to answer the questions hinders performances.

Our analysis of the self-reported confidences in participants answer did not indicate significant differences according to whether participants were presented elements not necessary to perform the task or not.

This is an interesting result, as it reinforces the validity to set up one visualization system with composite graphs to assess different tasks. Furthermore, with the SFNCS, composite graphs can be characterized, and we suspect that this approach can facilitate comparisons of composition of graphs for tasks which require to make implicit connections between elements displayed in different visualizations.

The analysis of the results thus gave us strong evidence to claim:

- Presence of the distractors displayed does not modify significantly ability to perform MwM , MO , MP , or SC .
- Presence of the distractors increases self-reported confidence for MP .

5.3 The Scaling study

5.3.1 Study motivation and structure

As we designed the visualization to assess the A-ATS Mask, specific efforts were made to balance out external validity, i.e. make the composite graph look real, and internal validity, i.e. allow tasks with different foci to be compared without confounds. Comparing ability to perform synoptic comparative tasks within multivariate spatio-temporal data analysis according to different foci requires stimuli to be presented for each task with enough space allocated to do so. According to the focus, the visual space necessary may differ, however.

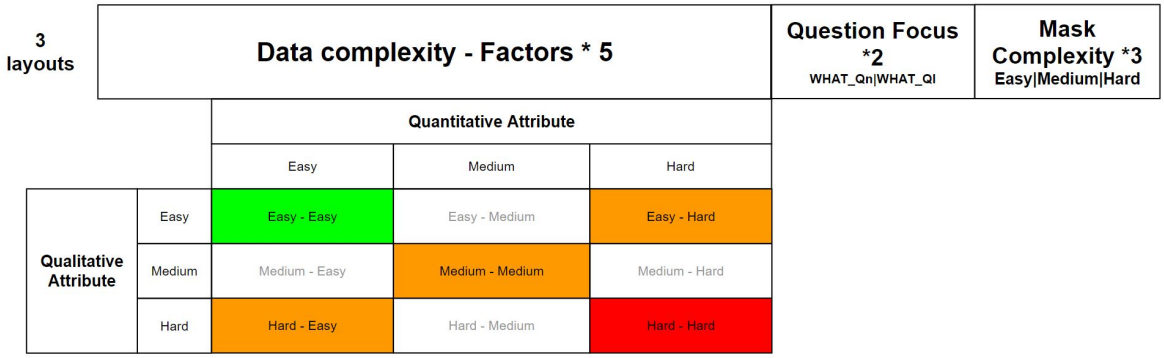


Fig. 5.23 The structure of the Scaling study. The categories vary according to *DATA COMPLEXITY* , *FOCUS* , and *MASK COMPLEXITY* . We colour the data complexities used to generate stimuli presented to the participants of the Scaling study.

There are still unanswered questions concerning the impact of screen size and interaction methods over performances of visualization tasks and understanding of information communicated [129, 92]. Our intent for the Scaling study is to evaluate for a certain range of potential scaling combinations whether they meaningfully impact ability to perform synoptic comparative tasks. The set of studies presented in this thesis are all run on computers, i.e. no smartphones and no pads. This sets certain expectations concerning screen size available that can be used for the set up of the studies [96].
The choices made for the set up of the composite graph presented to participants are motivated by a will to present a design similar to the Time Mask [17, 8] and the subjective impression that the view is 'good enough' to allow participants to understand the information presented to them. Once again, our confidence in this was reinforced by small informal studies with convenience samples in which participants answered elementary questions, e.g. ability to detect the number of turns for trajectories or identification of direction of trajectories, finding the maximal value of a quantitative attribute displayed, or the number of variations of status of a qualitative attribute.
Designing the composite graph we would present to participants, we had to consider whether some constraints had to be respected. One constraint for the studies we set up was the usage of the map used as a background from the data generated for the IEEE VAST 2014 challenge. We justified the selection of this data set, and by extension this background map in section 4.4. Selecting the

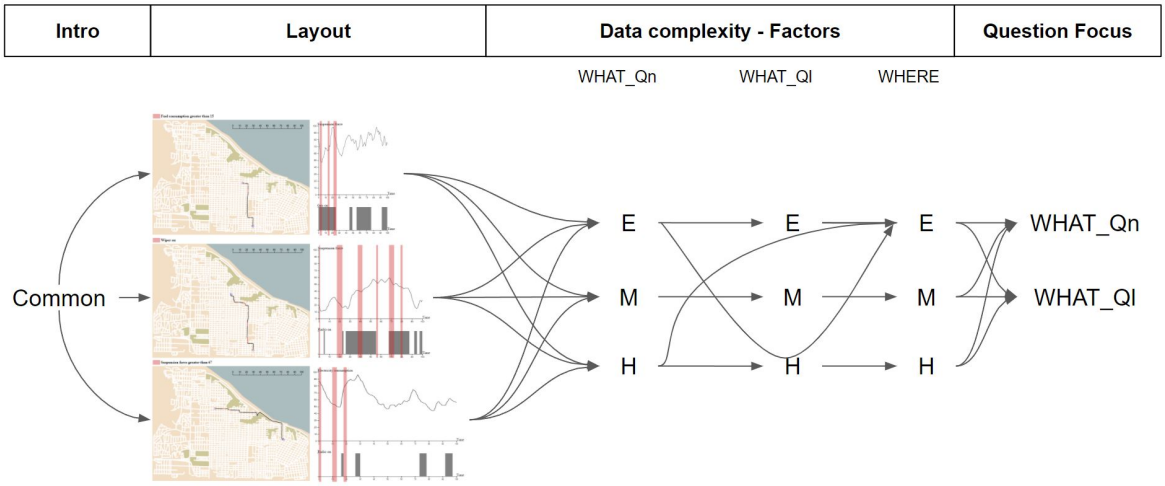


Fig. 5.24 The workflow of the Scaling study. Participants are split into three groups. After following the same introduction, they are then presented with a series of nine questions with data complexity varying as defined in Fig. 5.23. The size allocated for the display of the quantitative and qualitative attributes vary according to the group, being either 300, 450 or 600 pixels wide. Note that there are no questions with a WHERE focus, as this aspect is evaluated in other studies and not influenced by the scaling of the temporal axis in the quantitative and qualitative attributes displays.

visual space allocated to display the map background was based on consideration about screen size participants were likely to have when participating in our studies [124] and how to set up a composite graph with elements sharing the allocated space with each element being allocated enough space to perform the task according to the focus. We selected a space of 500 pixels of height, resulting in 686 pixels of width. We did not modify the data set of trajectories generated for the IEEE VAST 2014 challenge, to ensure we did not alter its validity.

To guarantee internal validity, participants were provided the same tools to answer questions no matter the focus, i.e. three sliders ranging from 0 to 100 for the MwM , MO and MP questions, as well as a set of answers to express their opinion for the SC question. This required questions with a WHERE focus to be presented on a map background at a scale that would result in our baselines ranging from 0 to 100. Considerations of how long this scale could be with our set of trajectories set constraints for our design. To ensure the baselines for questions with a WHERE focus would range from 0 to 100, we thus added a scale ranging from 0 to 100 with a width of 300 pixels.

The length of this axis thus sets potential values that could be used for the scaling of the axes of the quantitative and qualitative attributes displayed. The first approach considered was for all axes of the composite graph to share the same length, i.e. 300 pixels here. This approach would thus result in the graphs presenting quantitative and qualitative attributes to also have a width of 300 pixels. While consistent, this approach does not use the screen width as one would expect from a realistic software, i.e. the screen area is likely to have a relatively large space that isn't used to help the participant perform the task, while readability of the attributes might be hindered by that relatively small width. Alternatively, we considered setting the axes for the quantitative and qualitative attributes to be double

the size of the one of the geographical one, i.e. 600 pixels here. This approach produces the largest space available for the display of the quantitative and qualitative attributes. It is possible that this approach would result in the highest readability, but we argue might be 'too much' space, i.e. take more space than necessary to perform the tasks.
The third approach we considered was motivated by the aim to set up the visualizations with 'enough' space, i.e. wide enough to perform the task. We are aware that this notion is subjective and requires tests to ensure such a claim to be made confidently. We adopted a space that is in line with these expectations with a ratio of 1.5 times the axis of the geographical graph, i.e. 450 pixels. This resulted in a ratio is not uncommon in composite graphics that involve time series [**?**].
Accounting for the effects of the width allocated to perform the tasks with a WHAT_Qn and WHAT_Ql focus allows us to assess the validity of the other studies and reinforce our trust in the reusability of the system for future work. We thus set up the Scaling study, which compares ability to perform synoptic comparative tasks within multivariate temporal data analysis and self-reported confidence over ability to perform such tasks, according to the width set to display the graphs.

The set up of this study is made to help us answer the following research questions:

- **Research question 3.3**: How does the scaling of the axis displaying time variations impact the ability to correctly conduct synoptic comparative tasks within multivariate temporal data analysis?
- **Research question 3.4**: How does the scaling of the axis displaying time variations impact self-reported trust for conducting synoptic comparative tasks within multivariate temporal data analysis?

The structure of the scaling study is illustrated in figures 5.23 and 5.24.

5.3.2 Results analysis

The Scaling study focuses on the impact of varying the width allocated to the display of the WHAT_Ql and WHAT_Qn attributes. As mentioned in section 5.1.4, data quality implied a high amount of answers is filtered. The analysis of the answers produced for the Scaling study is done using data of participants who passed the introduction test (TI). 40% of responses were rejected as a result.
We list observations in the table 5.3 and use its indices to refer to them, but focus our discussion on our interpretation, which follows these observations. We discuss the different variants of width by considering their ratio, i.e. 300 px is a ratio of 1, 450 pixels is a ratio of 1.5 and 600 pixels is a ratio of 2, e.g. we refer to 'scaling 0' instead of 'cases with a width of 300 pixels assigned for width' for readability.

The means of the results according to the factors are illustrated in table 5.8.

study	factor1	factor2	factor3	responses	MwM	MO	MP	SC (mean correctB)	SC (error rate)	responses (neither)	SC (neither mean correctB)	SC (neither error rate)	str
s				171.00	8.08	6.00	6.02	0.34	0.66	134.00	0.43	0.57	
s	scaling			54.00	5.38	4.86	6.01	0.33	0.67	44.00	0.41	0.59	0.00
s	scaling			63.00	5.15	5.05	4.60	0.38	0.62	55.00	0.44	0.56	1.00
s	scaling			54.00	14.21	8.24	7.70	0.30	0.70	35.00	0.46	0.54	2.00
s	scaling	*FOCUS*		29.00	3.47	5.02	4.54	0.38	0.62	24.00	0.46	0.54	0-WHAT_QI
s	scaling	*FOCUS*		25.00	7.59	4.66	7.71	0.28	0.72	20.00	0.35	0.65	0-WHAT_Qn
s	scaling	*FOCUS*		31.00	5.65	5.73	5.28	0.39	0.61	27.00	0.44	0.56	1-WHAT_QI
s	scaling	*FOCUS*		32.00	4.67	4.40	3.93	0.38	0.63	28.00	0.43	0.57	1-WHAT_Qn
s	scaling	*FOCUS*		29.00	18.15	8.23	5.32	0.31	0.69	21.00	0.43	0.57	2-WHAT_QI
s	scaling	*FOCUS*		25.00	9.64	8.25	10.46	0.28	0.72	14.00	0.50	0.50	2-WHAT_Qn
s	scaling	*MASK COMPLEXITY*		18.00	5.95	2.81	3.39	0.33	0.67	15.00	0.40	0.60	0-easy
s	scaling	*MASK COMPLEXITY*		21.00	4.88	6.44	5.79	0.38	0.62	16.00	0.50	0.50	0-medium
s	scaling	*MASK COMPLEXITY*		15.00	5.38	5.10	9.45	0.27	0.73	13.00	0.31	0.69	0-hard
s	scaling	*MASK COMPLEXITY*		26.00	3.43	4.63	5.27	0.42	0.58	24.00	0.46	0.54	1-easy
s	scaling	*MASK COMPLEXITY*		18.00	5.25	7.08	2.56	0.50	0.50	17.00	0.53	0.47	1-medium
s	scaling	*MASK COMPLEXITY*		19.00	7.42	3.72	5.60	0.21	0.79	14.00	0.29	0.71	1-hard
s	scaling	*MASK COMPLEXITY*		19.00	16.13	8.36	8.14	0.32	0.68	13.00	0.46	0.54	2-easy
s	scaling	*MASK COMPLEXITY*		17.00	10.21	11.36	6.20	0.35	0.65	12.00	0.50	0.50	2-medium
s	scaling	*MASK COMPLEXITY*		18.00	15.96	5.17	8.64	0.22	0.78	10.00	0.40	0.60	2-hard
s	scaling	*FOCUS*	*MASK COMPLEXITY*	9.00	1.57	1.63	2.32	0.56	0.44	8.00	0.63	0.38	0-WHAT_QI-easy
s	scaling	*FOCUS*	*MASK COMPLEXITY*	11.00	4.27	6.53	4.67	0.45	0.55	8.00	0.63	0.38	0-WHAT_QI-medium
s	scaling	*FOCUS*	*MASK COMPLEXITY*	9.00	4.39	6.58	6.60	0.11	0.89	8.00	0.13	0.88	0-WHAT_QI-hard
s	scaling	*FOCUS*	*MASK COMPLEXITY*	9.00	10.33	3.99	4.45	0.11	0.89	7.00	0.14	0.86	0-WHAT_Qn-easy
s	scaling	*FOCUS*	*MASK COMPLEXITY*	10.00	5.55	6.34	7.02	0.30	0.70	8.00	0.38	0.63	0-WHAT_Qn-medium
s	scaling	*FOCUS*	*MASK COMPLEXITY*	6.00	6.87	2.88	13.73	0.50	0.50	5.00	0.60	0.40	0-WHAT_Qn-hard
s	scaling	*FOCUS*	*MASK COMPLEXITY*	16.00	3.25	5.04	6.93	0.50	0.50	16.00	0.50	0.50	1-WHAT_QI-easy
s	scaling	*FOCUS*	*MASK COMPLEXITY*	7.00	7.92	10.89	3.82	0.43	0.57	6.00	0.50	0.50	1-WHAT_QI-medium
s	scaling	*FOCUS*	*MASK COMPLEXITY*	8.00	8.47	2.58	3.27	0.13	0.88	5.00	0.20	0.80	1-WHAT_QI-hard
s	scaling	*FOCUS*	*MASK COMPLEXITY*	10.00	3.72	3.96	2.63	0.30	0.70	8.00	0.38	0.63	1-WHAT_Qn-easy
s	scaling	*FOCUS*	*MASK COMPLEXITY*	11.00	3.55	4.65	1.76	0.55	0.45	11.00	0.55	0.45	1-WHAT_Qn-medium
s	scaling	*FOCUS*	*MASK COMPLEXITY*	11.00	6.66	4.55	7.29	0.27	0.73	9.00	0.33	0.67	1-WHAT_Qn-hard
s	scaling	*FOCUS*	*MASK COMPLEXITY*	10.00	22.74	7.85	4.89	0.30	0.70	6.00	0.50	0.50	2-WHAT_QI-easy
s	scaling	*FOCUS*	*MASK COMPLEXITY*	12.00	12.13	12.62	6.09	0.33	0.67	10.00	0.40	0.60	2-WHAT_QI-medium
s	scaling	*FOCUS*	*MASK COMPLEXITY*	7.00	21.91	1.26	4.60	0.29	0.71	5.00	0.40	0.60	2-WHAT_QI-hard
s	scaling	*FOCUS*	*MASK COMPLEXITY*	9.00	8.80	8.93	11.75	0.33	0.67	7.00	0.43	0.57	2-WHAT_Qn-easy
s	scaling	*FOCUS*	*MASK COMPLEXITY*	5.00	5.60	8.33	6.45	0.40	0.60	2.00	1.00	0.00	2-WHAT_Qn-medium
s	scaling	*FOCUS*	*MASK COMPLEXITY*	11.00	12.17	7.66	11.22	0.18	0.82	5.00	0.40	0.60	2-WHAT_Qn-hard
s	scaling	*FOCUS*	*DATA COMPLEXITY*	9.00	4.98	6.66	7.15	0.33	0.67	7.00	0.43	0.57	0-WHAT_QI-E
s	scaling	*FOCUS*	*DATA COMPLEXITY*	14.00	3.04	2.60	2.63	0.50	0.50	12.00	0.58	0.42	0-WHAT_QI-M
s	scaling	*FOCUS*	*DATA COMPLEXITY*	6.00	2.22	8.22	5.08	0.17	0.83	5.00	0.20	0.80	0-WHAT_QI-H
s	scaling	*FOCUS*	*DATA COMPLEXITY*	12.00	4.52	5.18	10.17	0.33	0.67	10.00	0.40	0.60	0-WHAT_Qn-E
s	scaling	*FOCUS*	*DATA COMPLEXITY*	6.00	15.31	4.50	4.18	0.17	0.83	5.00	0.20	0.80	0-WHAT_Qn-M
s	scaling	*FOCUS*	*DATA COMPLEXITY*	7.00	6.22	3.92	6.50	0.29	0.71	5.00	0.40	0.60	0-WHAT_Qn-H
s	scaling	*FOCUS*	*DATA COMPLEXITY*	15.00	6.87	4.56	7.86	0.47	0.53	13.00	0.54	0.46	1-WHAT_QI-E
s	scaling	*FOCUS*	*DATA COMPLEXITY*	12.00	4.58	3.36	3.04	0.33	0.67	11.00	0.36	0.64	1-WHAT_QI-M
s	scaling	*FOCUS*	*DATA COMPLEXITY*	4.00	4.29	17.19	2.36	0.25	0.75	3.00	0.33	0.67	1-WHAT_QI-H
s	scaling	*FOCUS*	*DATA COMPLEXITY*	14.00	3.30	4.43	3.07	0.50	0.50	11.00	0.64	0.36	1-WHAT_Qn-E
s	scaling	*FOCUS*	*DATA COMPLEXITY*	6.00	8.05	3.83	2.46	0.33	0.67	6.00	0.33	0.67	1-WHAT_Qn-M
s	scaling	*FOCUS*	*DATA COMPLEXITY*	12.00	4.59	4.66	5.67	0.25	0.75	11.00	0.27	0.73	1-WHAT_Qn-H
s	scaling	*FOCUS*	*DATA COMPLEXITY*	9.00	25.04	0.82	4.43	0.22	0.78	6.00	0.33	0.67	2-WHAT_QI-E
s	scaling	*FOCUS*	*DATA COMPLEXITY*	13.00	11.32	11.36	6.43	0.23	0.77	9.00	0.33	0.67	2-WHAT_QI-M
s	scaling	*FOCUS*	*DATA COMPLEXITY*	7.00	21.98	11.95	4.41	0.57	0.43	6.00	0.67	0.33	2-WHAT_QI-H
s	scaling	*FOCUS*	*DATA COMPLEXITY*	10.00	6.01	4.96	9.52	0.30	0.70	7.00	0.43	0.57	2-WHAT_Qn-E
s	scaling	*FOCUS*	*DATA COMPLEXITY*	7.00	19.43	12.15	13.27	0.29	0.71	4.00	0.50	0.50	2-WHAT_Qn-M
s	scaling	*FOCUS*	*DATA COMPLEXITY*	8.00	5.60	8.97	9.15	0.25	0.75	3.00	0.67	0.33	2-WHAT_Qn-H

Table 5.8 Results summary for the Scaling study. The table indicates the absolute difference between responses from participants and the baselines, according to factors we varied. The columns on the left indicate the factors and the ones on the right indicate the detail of the factored selections, e.g. "0-WHERE-medium" indicates responses where the scaling is indexed 0, i.e. 300 pixels, the *FOCUS* was WHERE, and the *MASK COMPLEXITY* was medium. The columns with "neither" in their titles discuss the results once responses for which participants answered "Neither agree nor disagree" are filtered.

Numerical questions: Mean with Mask (MwM), Overall Mean (MO), Mask Proportion (MP)

We discuss the answers of questions MwM , MO and MP , which asked participants to provide numbers as an answer in this section. We split the sets of observations according to characteristics of the stimuli.

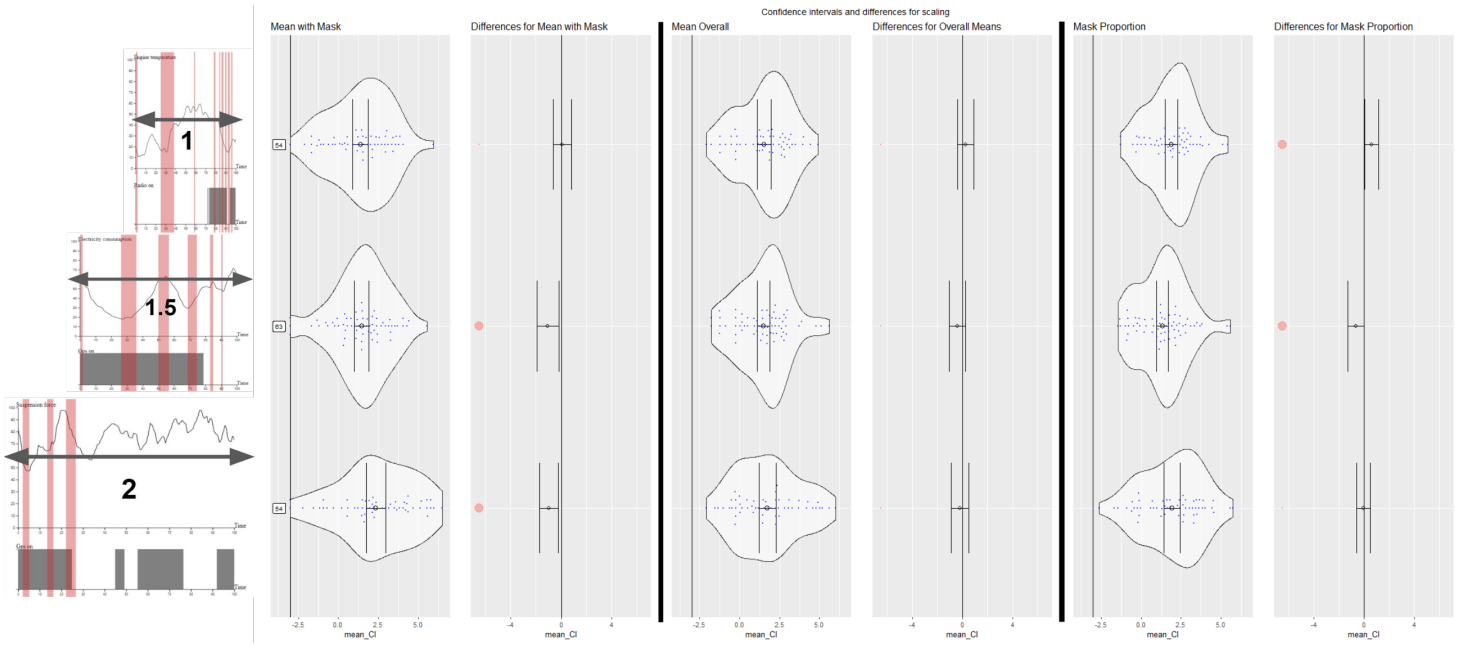


Fig. 5.25 The absolute errors for MwM , MO and MP with the variant being Scaling.

The overall results for the Scaling study are displayed in Fig. 5.25. We observe in Fig. 5.25 significant differences for Mean with Mask (MwM) according to scaling. Scaling 2 results in significantly poorer performances for MwM , while scaling 0 and 1 are strongly similar. We also observe that performances for Mask Proportion (MP) are significant better when scaling is 1.5 (MP 1).

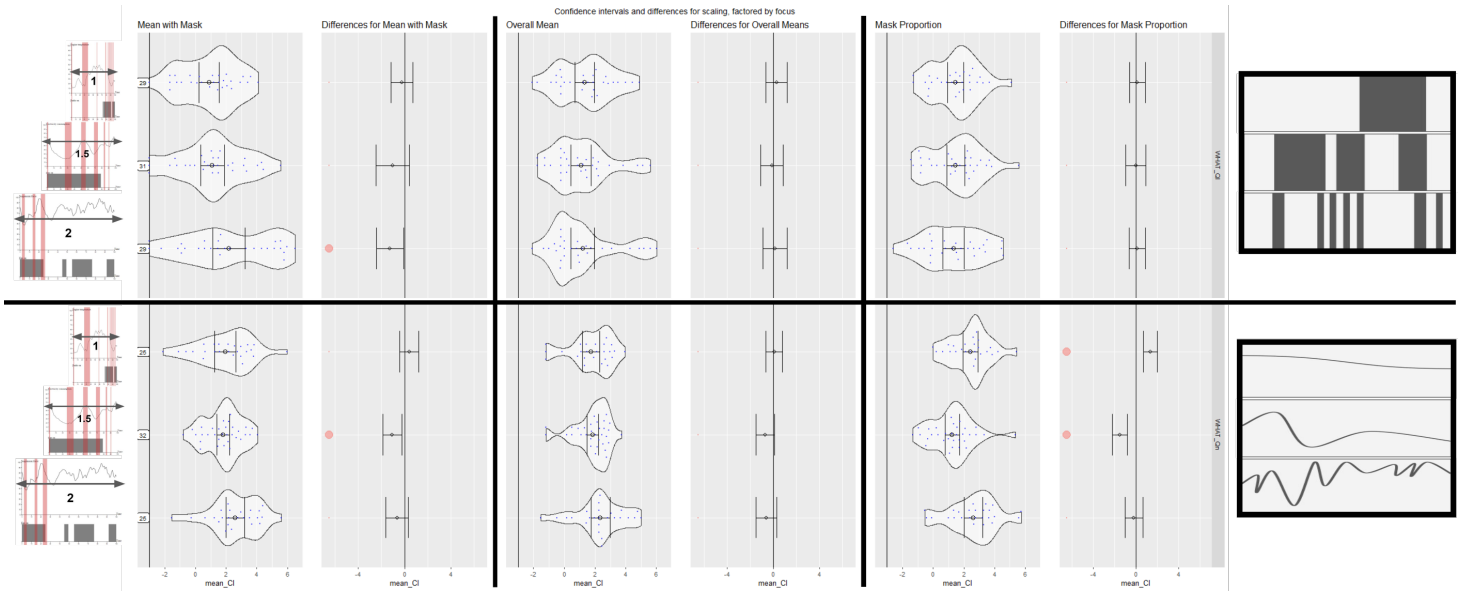


Fig. 5.26 The absolute errors for MwM , MO and MP with the variant being Scaling and the factor being focus.

The results factored according to focus are illustrated in Fig. 5.26. Factoring according to focus, the same trend is observed, with performance deteriorating as scaling increases (MwM 2). Those differences of performances are significant between scaling 1 and 2 for WHAT_Ql and 1.5 and 2 for WHAT_Qn.

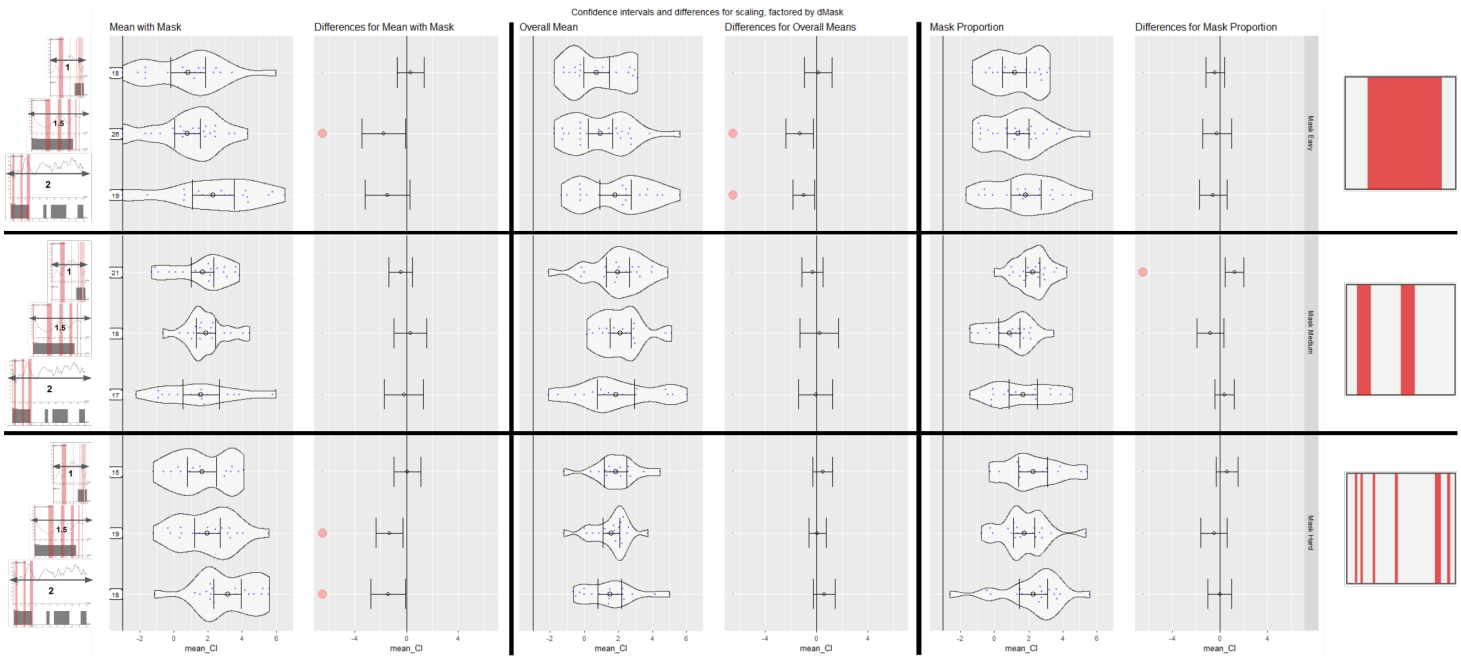


Fig. 5.27 The absolute errors for MwM , MO and MP with the variant being Scaling and the factor being *MASK COMPLEXITY* .

The ordering of performances for MwM is also noticeable when variants are factored according to masks, as illustrated in Fig. 5.27. The trend is consistent and provides some evidence that performance deteriorates as scaling increases. This observation should be nuanced, as differences are only statistically significant (5%) between scaling 1 and 2 and 1.5 and 2, when the Mask is Hard. The previous observation MP 1 about performances being significantly higher for scaling 1.5 is no longer true except in the case where Mask is Medium.

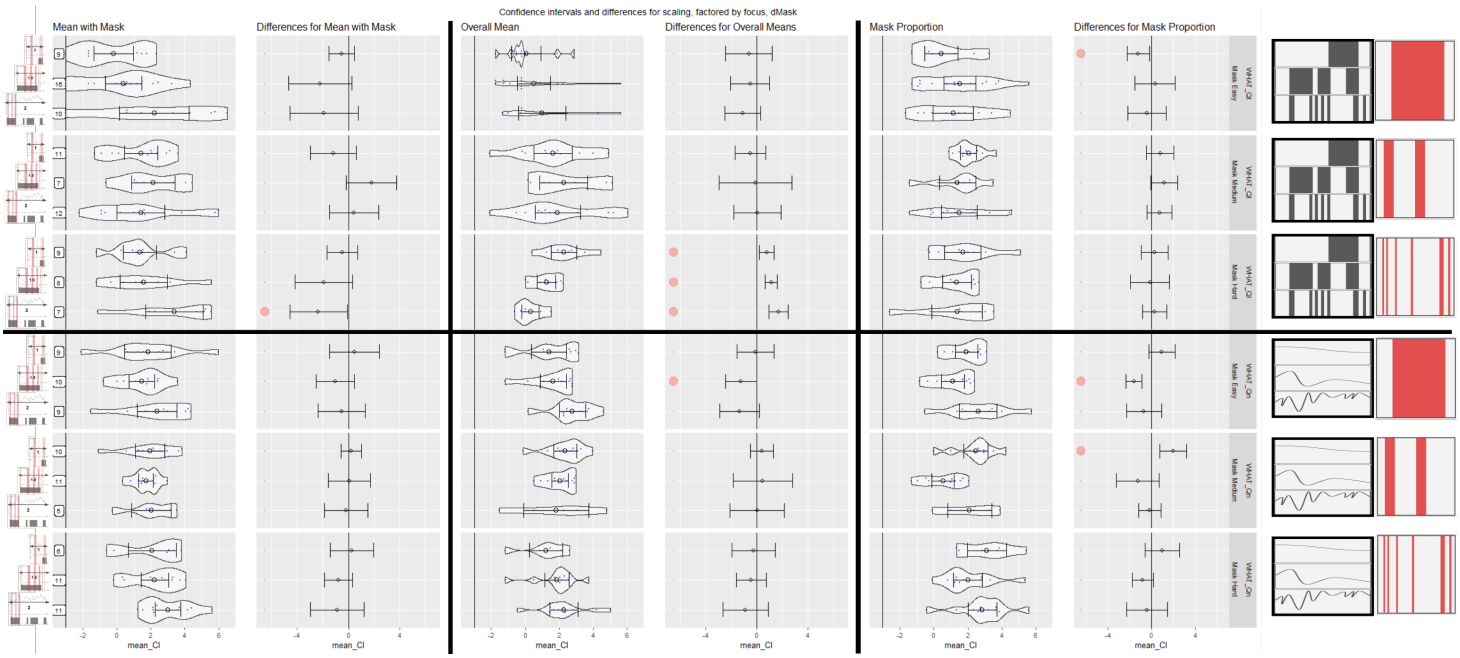


Fig. 5.28 The absolute errors for MwM , MO and MP with the variant being Scaling and the factors being *FOCUS* and *MASK COMPLEXITY* .

Fig. 5.28 shows that MwM 2 is true when results are factored according to Focus and *MASK COMPLEXITY* . We notice that significant differences of performance between each combination of scaling with a WHAT_Ql focus and Mask Hard (MO 1). This combination is unique due to the previous statement and the observation that it presents a trend of performance inverse to most MO results, them being either poorer as MO as scaling increases, or without strong trend. This leads us to conclude that an unpredicted element of our study setup may result in a confounding factor. The previous observation about MP 1 is once more contradicted, which leads us to conclude it was a false positive. It is interesting to note that MO here indicates a strong trend, with accuracy being higher, the lower the scaling, with the particular exception of WHAT_Qn and Mask Hard as a factor. This leads us to think that Overall Mean varies accordingly to Scaling (MO 2).

Fig. 5.29 The absolute errors for MwM , MO and MP with the variant being Scaling and the factors being focus and its complexity.

Fig. 5.29 shows that MwM 2 is true when results are factored according to focus and its complexity. The previous observation about MO 2 is here nuanced with the factor WHAT_Ql set to Easy. We thus conclude that MO 2 is significant within a certain range of difficulties parameters.

Likert question: Stability Comparison (SC)

We discuss the answers of questions SC , which asked participants to indicate whether they agreed with the description of stability in the data. We split the sets of observations according to characteristics of the stimuli.

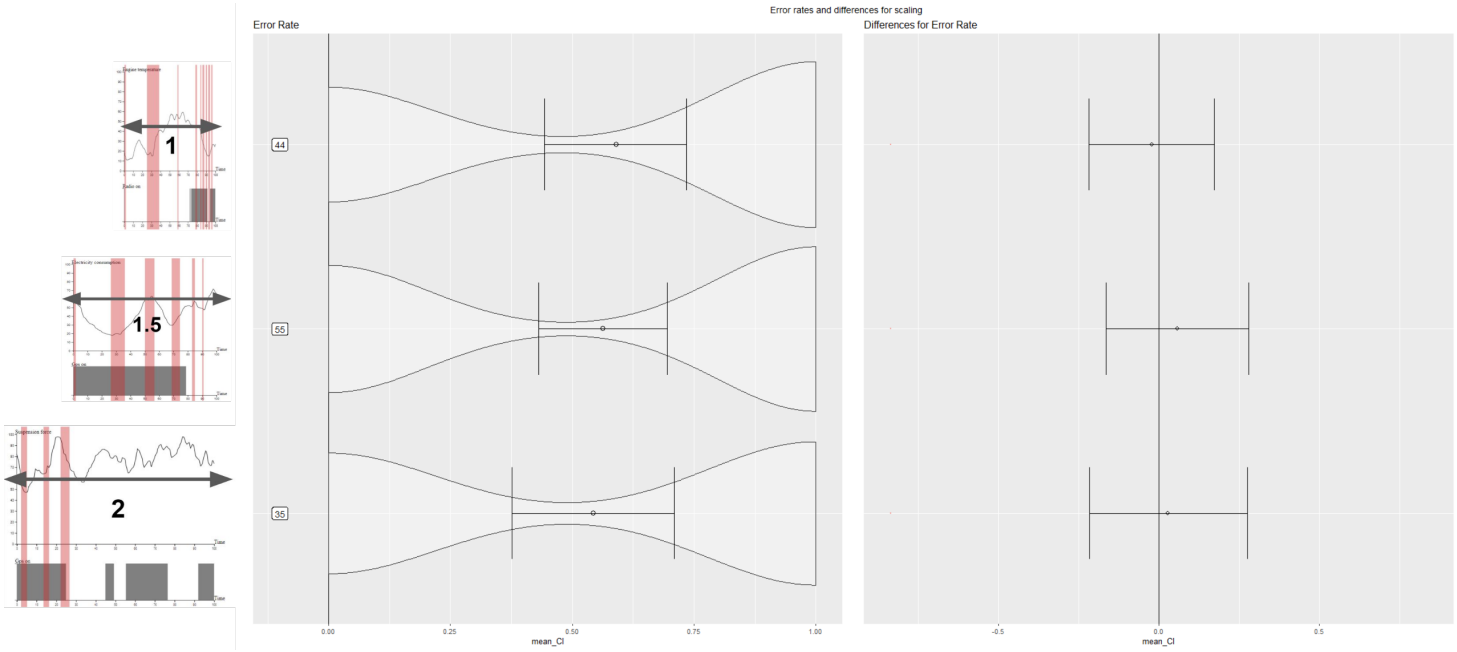
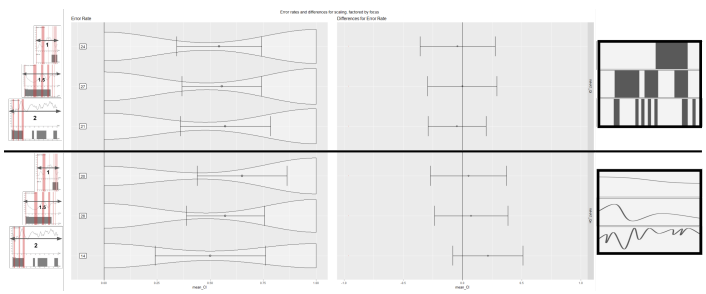
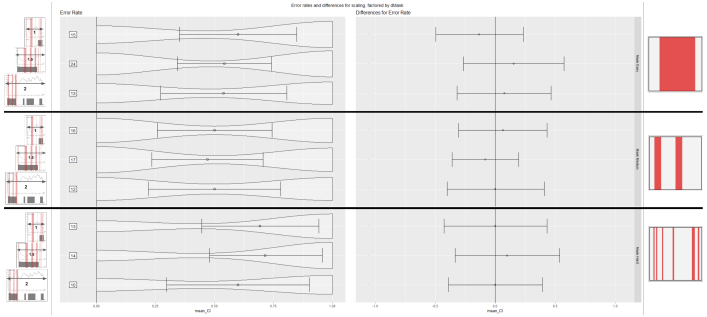


Fig. 5.30 The error rates for the SC of the scaling study. We note no significant difference of performance according to the Scaling.



(a) The error rates for the SC of the scaling study factored by *FOCUS* .



(b) The error rates for the SC of the scaling study factored by *MASK COMPLEXITY* .

Table 5.9 We see no evidence that Scaling has an effect on the error rates when factored by *FOCUS* (Qn|Ql) or *MASK COMPLEXITY* (Easy|Medium|Hard).

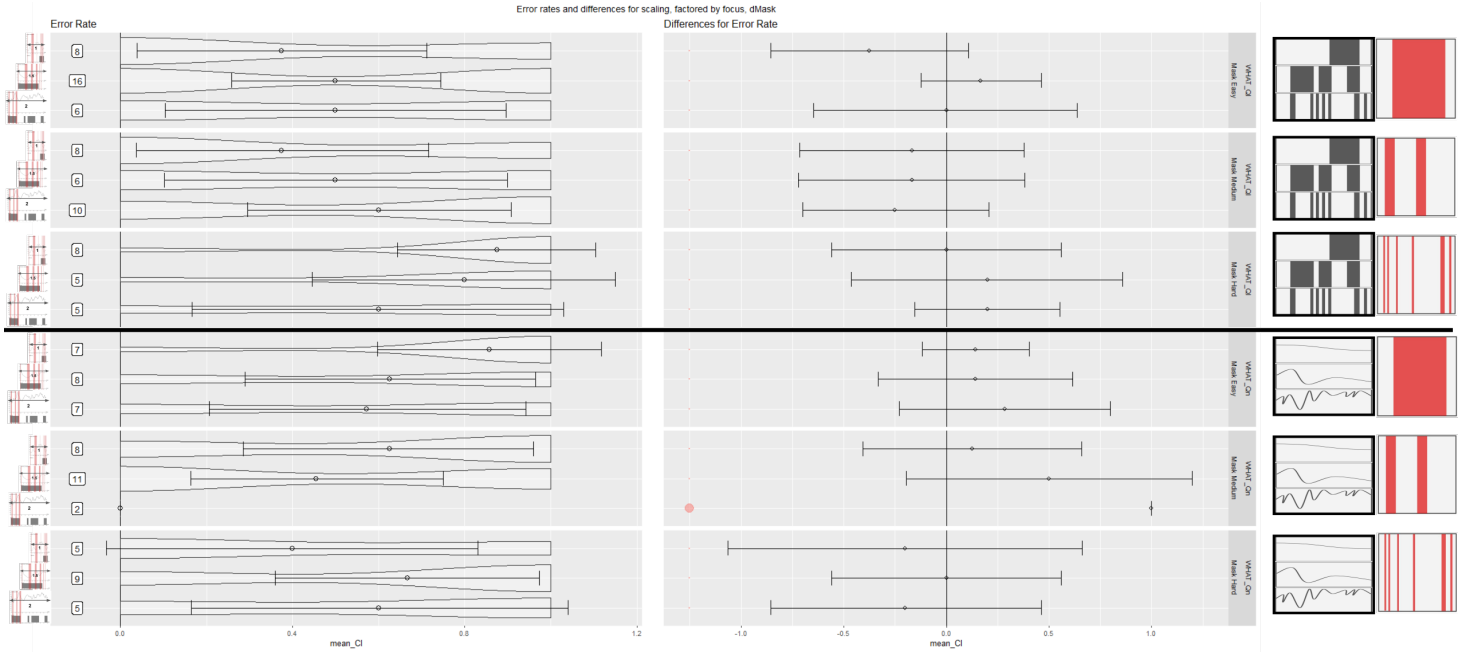


Fig. 5.31 The error rates for the SC of the scaling study, with the scaling being the variant, factored by both *FOCUS* and *MASK COMPLEXITY* .

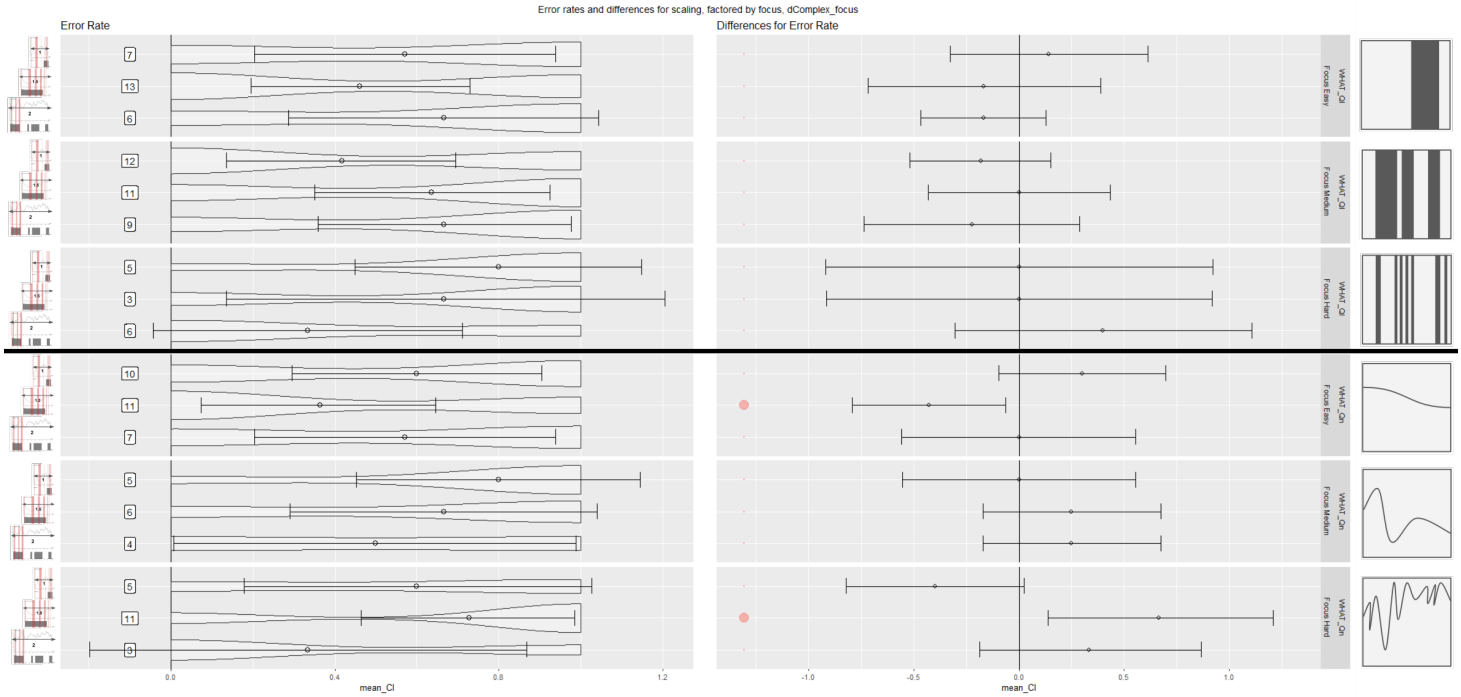


Fig. 5.32 The error rates for the SC of the scaling study, with the scaling being the variant, factored by both *FOCUS* and *DATA COMPLEXITY* .

Note that we have very wide confidence intervals for the representation of the error rates, indicating answers are low on both accuracy and precision, although this has to be nuanced with the relatively low number of participants who passed the introduction test (TI). We noticed that performances for SC seem to get poorer with scaling getting higher when the focus is WHAT_Ql and focus is Easy and Medium, but answers with WHAT_Ql and focus is Hard follow an inverse relationship. We thus consider that these observations are not strong enough to draw conclusions upon, at least for SC .

Self-reported confidence

We display the distribution of self-reported confidence in Fig. 5.33. For each question, we assess whether different scaling results in significantly different levels of confidence using a Kruskal-Wallace test for significance and a pot-hoc Dunn test to account for family-wise error and identify the differences. The results of these tests are listed in table 5.6.
We did not have strong expectations concerning the impact of the horizontal scaling of the WHAT_Ql and WHAT_Qn graphs. As discussed previously, our understanding of the impact of the space allocated for a graph is dependent on the subjective notion that there is 'enough' space allocated for participants to confidently perform the tasks asked of them. We consider that this notion is likely shared for both accuracy and confidence, although they may differ in values which define the subjective thresholds, e.g. participants could theoretically be overconfident and claim a high confidence with a certain graph size, while accuracy evaluations could indicate a higher graph size is necessary to perform said task efficiently.
The visual analysis of thee bar charts seem to indicate differences between groups, particularly for the group where the stimuli is 450 pixels wide. The Dunn tests indicate that the differences are not statistically significant, except for the single case of question MP and scaling being either 450 pixels wide and 600 pixels wide. This finding is not in line with others and is not easily explained, which leads us to report it, but consider it likely that this is a false positive.
We thus conclude that there are no significant differences in self-reported confidence between groups according to scaling for questions MwM , MO , MP and SC .

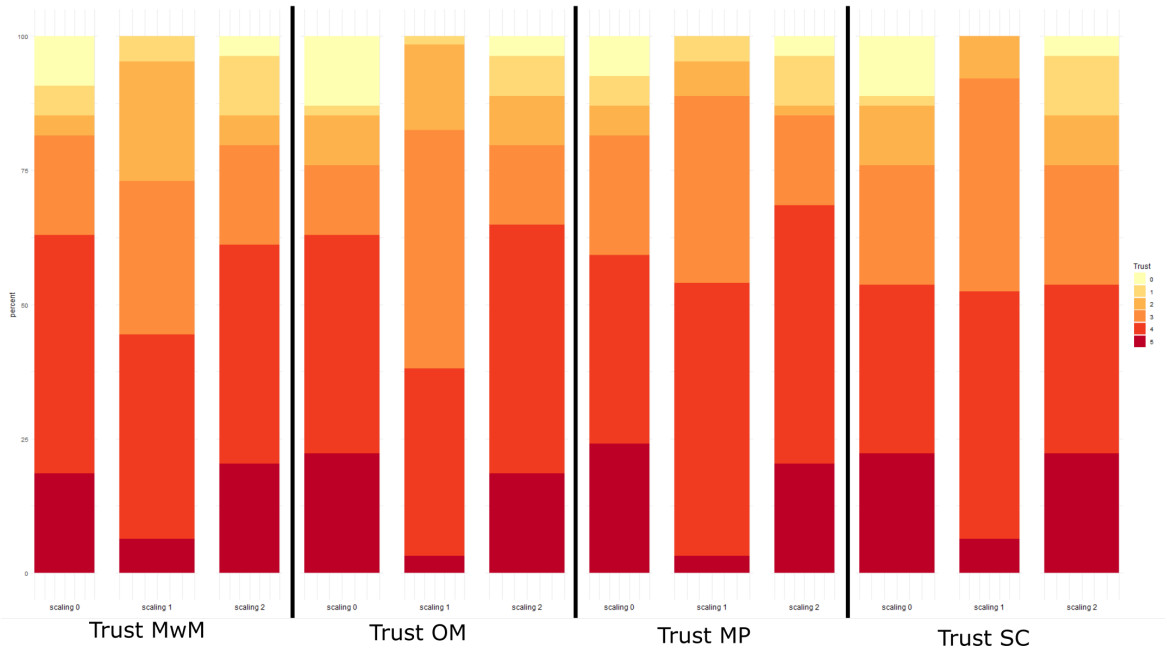


Fig. 5.33 The figures are composed of bar charts indicating the reported self-reported trust according to question when the stimuli are 300 pixels wide in figure when index of scaling is 0, 450 pixels in figure when index of scaling is 1 and 600 pixels when index of scaling is 2. Each set of bar charts indicates the self-reported trust in the following order: MwM , MO , MP and SC . Bars are coloured from yellow to red, with 0 indicating no confidence at all in their answer and 5 indicating absolute confidence in their answer.

Conclusion

We verified the impact of the width of the graphs set to display the WHAT_Qn and WHAT_Ql information, through their scaling, over the performances and self-reported confidence of participants for Mean with Mask (MwM), Mean Overall (MO), Mask Proportion (MP) and Stability Comparison (SC). As stated previously, the relatively low number of participants who passed the introduction test (TI) requires some caution in interpreting our observations.

We noticed that performances for MwM were poorest with the widest space allocated for the graphs, while it had little effect over performances for MO , MP and SC . Additionally, the variations of performances noticed for MwM were not consistent as we factored the results according to the focus of the questions or the *MASK COMPLEXITY* . We thus consider that the difference in performance observed for MwM when the width is at its highest is likely a false positive.

We also conclude that performances for MO , MP and SC are not impacted by the scaling of the WHAT_Ql and WHAT_Qn graphs, although the fairly low results of SC might simply indicate an inability to perform the task as we set it for the studies, independently of scaling. This warrants further investigation.

Our analysis of reported self-confidence indicates no significant difference according to the scaling assigned to the graphs. In summary, Scaling (1.5) time series graphics performed no worse than

Scaling(1.0) or Scaling (2.0) and that in some circumstances they may have been associated with marginally improved performance. This supports the use of Scaling(1.5) for the time series graphics used in the realistic composite graphs for the Measurement tests that follow.

The analysis of the impact of the Scaling factor over results let us claim tentatively that, for our three selected Scaling values of 300, 450 and 600 pixels width allocated for the graphs of WHAT_Ql and WHAT_Qn:

- MwM , MO , MP and SC are not impacted by the Scaling.
- Self-reported confidence is not significantly impacted by the Scaling selected.

5.4 The Measurement Study

5.4.1 Study motivation and structure

The Measurement study is the main study that we ran during this thesis. As discussed in section 4, the novelty of the A-ATS Mask means that we lack knowledge about the strengths, weaknesses and limitations that arise with this design. Informally, we could simply state that the motivation to run the Measurement study originates from our impression that the time mask of Andrienko *et al.* [17, 8] was promising, and that further design extensions seemed likely to be valuable.

1 layout	Data complexity - Factors * 27										Question Focus *3 WHAT_Qn|WHAT_QI|WHERE	Mask Complexity *3 Easy|Medium|Hard
		Quantitative Attribute										
		Easy			Medium			Hard				
Qualitative Attribute	Easy	Spatial			Spatial			Spatial				
		Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard		
	Medium	Spatial			Spatial			Spatial				
		Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard		
	Hard	Spatial			Spatial			Spatial				
		Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard		

Fig. 5.34 The structure of the study termed Measurement. The categories vary according to data complexity, focus of the questions, and complexity of mask applied. We colour the data complexities used to generate stimuli presented to the participants of the Measurement study.

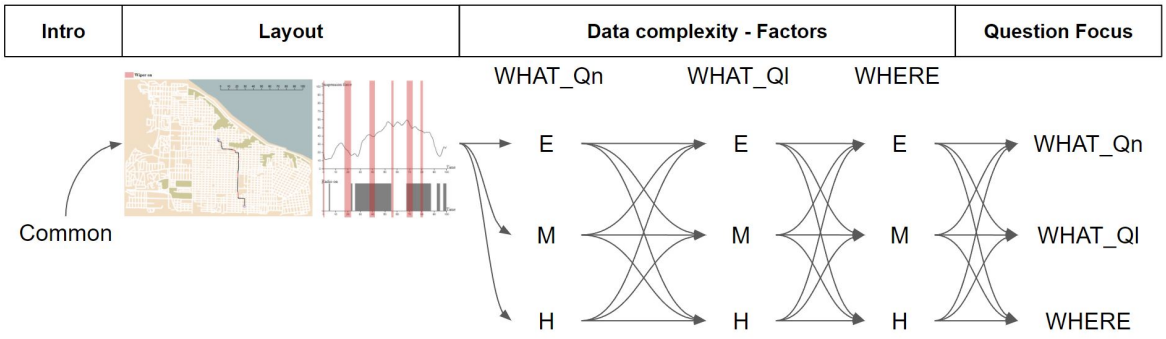


Fig. 5.35 The workflow of the Measurement study. This study details more *DATA COMPLEXITY* combinations and *FOCI* on their impact on effectiveness for performing synoptic comparative tasks. The detail of the data complexities are illustrated in Fig. 5.34.

Due to the novelty of the design presented, the first studies to run would have to be limited concerning the number of elements to show to the participants of the studies, i.e. we display a single moving entity, a single quantitative attribute, a single qualitative attribute, and a single mask resulting from a query.

The study is configured to help us answer the following research questions:

- **Research question 3.1**: How does the visualization of conditions over time-space-attributes support synoptic comparative tasks within multivariate spatio-temporal data analysis?
- **Research question 3.2**: How does the visualization of conditions over time-space-attributes impact self-reported trust for conducting synoptic comparative tasks within multivariate spatio-temporal data analysis?

The structure of the Measurement study is illustrated in figures 5.34 and 5.35.

5.4.2 Results analysis

The Measurement study was set to evaluate the impact of diverse factors over the performances and self-reported confidence while performing Mean with Mask (MwM), Overall Mean (MO), Mask Proportion (MP) and Stability Comparison (SC). As in the case of the Distractor and Judgment studies (see section 5.1.4) a high proportion of respondents were omitted through the quality filter. The analysis of the answers produced for the Measurement study is done using data from participants who passed the introduction test (TI) and the rigorous test (RT).
The Measurement study was designed to assess the impact of the different Mask difficulties, the different *FOCI* of the questions, and the variations of difficulties for each focus. Thus, unlike the Scaling and Distractor studies, the Measurement studies assess different variants. We thus have graphs to analyse absolute errors and error rates according to different *FOCI* , *MASK COMPLEXITY* , and difficulties for each focus.

The list of of results according to factors can be found in table 5.10. For MwM , MO , MP , the values are the means of the absolute differences between responses and baselines, and for SC the values are, as listed in the table, either the mean or of responses being correct or the error rate. The table lists both performances for SC with or without considering responses "Neither agree nor disagree" as false. The ones who filter out these responses are indicated by the keyword "neither".

The following subsections split the results according to their answer method.

Numerical questions: Mean with Mask (MwM), Overall Mean (MO), Mask Proportion (MP)

The results for the three tasks (MwM , MO and MP , columns) factored by *FOCUS* are displayed in Fig. 5.36.

- For *Mean with Mask* *MwM* (left), performance differs significantly according to *FOCUS* : performance in WHERE tasks is inferior to both WHAT_Qn and WHAT_Ql.
- *Mean Overall* *MO* shows significant differences between all *FOCI* : WHAT_Ql performance is significantly better than WHAT_Qn and WHERE, with WHAT_Qn performance signifi-

study	variant	factor1	factor2	# responses	MwM	MO	MP	mean SC	error rate SC	log MwM	log MO	log MP	#responses noNeither	mean noNeither SC	error rate noNeither SC	str
m				585	10.93	9.25	15.15	0.37	0.63	2.82	2.49	3.42	432.00	0.50	0.50	
m	*FOCUS*			195	9.44	6.33	16.24	0.33	0.67	2.62	1.85	3.54	145.00	0.45	0.55	WHAT_Ql
m	*FOCUS*			195	9.38	7.16	14.72	0.43	0.57	2.67	2.33	3.30	147.00	0.56	0.44	WHAT_Qn
m	*FOCUS*			195	13.97	14.26	14.48	0.35	0.65	3.17	3.29	3.41	140.00	0.49	0.51	WHERE
m	*MASK COMPLEXITY*			192	12.50	8.64	18.04	0.40	0.60	3.00	2.43	3.68	148.00	0.52	0.48	easy
m	*MASK COMPLEXITY*			191	8.58	9.08	15.02	0.32	0.68	2.52	2.41	3.46	136.00	0.46	0.54	medium
m	*MASK COMPLEXITY*			202	11.67	10.00	12.51	0.38	0.62	2.93	2.62	3.12	148.00	0.52	0.48	hard
m	*DATA COMPLEXITY*			195	8.79	7.83	15.73	0.35	0.65	2.48	2.26	3.46	135.00	0.51	0.49	E
m	*DATA COMPLEXITY*			195	13.10	10.93	15.03	0.37	0.63	3.16	2.65	3.36	147.00	0.49	0.51	M
m	*DATA COMPLEXITY*			195	10.90	8.99	14.68	0.38	0.62	2.82	2.56	3.43	150.00	0.50	0.50	H
m	*FOCUS*	*MASK COMPLEXITY*		60	9.90	4.82	18.05	0.37	0.63	2.61	1.63	3.73	41.00	0.54	0.46	WHAT_Ql-easy
m	*FOCUS*	*MASK COMPLEXITY*		70	7.92	6.75	17.34	0.29	0.71	2.44	1.72	3.74	51.00	0.39	0.61	WHAT_Ql-medium
m	*FOCUS*	*MASK COMPLEXITY*		65	10.66	7.27	13.37	0.35	0.65	2.83	2.18	3.15	53.00	0.43	0.57	WHAT_Ql-hard
m	*FOCUS*	*MASK COMPLEXITY*		58	10.58	6.20	22.25	0.45	0.55	2.88	2.19	4.04	48.00	0.54	0.46	WHAT_Qn-easy
m	*FOCUS*	*MASK COMPLEXITY*		64	6.38	7.08	12.69	0.39	0.61	2.25	2.35	3.04	50.00	0.50	0.50	WHAT_Qn-medium
m	*FOCUS*	*MASK COMPLEXITY*		73	11.07	8.01	10.51	0.44	0.56	2.88	2.42	2.94	49.00	0.65	0.35	WHAT_Qn-hard
m	*FOCUS*	*MASK COMPLEXITY*		74	16.10	13.64	14.75	0.39	0.61	3.42	3.26	3.37	59.00	0.49	0.51	WHERE-easy
m	*FOCUS*	*MASK COMPLEXITY*		57	11.86	14.19	14.78	0.30	0.70	2.92	3.32	3.59	35.00	0.49	0.51	WHERE-medium
m	*FOCUS*	*MASK COMPLEXITY*		64	13.38	15.03	13.91	0.34	0.66	3.10	3.30	3.30	46.00	0.48	0.52	WHERE-hard
m	*FOCUS*	*DATA COMPLEXITY*		65	6.93	4.80	16.80	0.28	0.72	2.20	1.42	3.51	45.00	0.40	0.60	WHAT_Ql-E
m	*FOCUS*	*DATA COMPLEXITY*		65	10.60	8.22	15.99	0.38	0.62	2.85	2.08	3.51	48.00	0.52	0.48	WHAT_Ql-M
m	*FOCUS*	*DATA COMPLEXITY*		65	10.79	5.98	15.92	0.34	0.66	2.81	2.05	3.60	52.00	0.42	0.58	WHAT_Ql-H
m	*FOCUS*	*DATA COMPLEXITY*		65	8.89	5.98	15.25	0.38	0.62	2.53	2.13	3.43	45.00	0.56	0.44	WHAT_Qn-E
m	*FOCUS*	*DATA COMPLEXITY*		65	10.49	8.32	15.23	0.46	0.54	2.91	2.47	3.21	50.00	0.60	0.40	WHAT_Qn-M
m	*FOCUS*	*DATA COMPLEXITY*		65	8.77	7.20	13.68	0.43	0.57	2.58	2.38	3.26	52.00	0.54	0.46	WHAT_Qn-H
m	*FOCUS*	*DATA COMPLEXITY*		65	10.56	12.72	15.16	0.40	0.60	2.71	3.23	3.44	45.00	0.58	0.42	WHERE-E
m	*FOCUS*	*DATA COMPLEXITY*		65	18.21	16.26	13.87	0.26	0.74	3.73	3.39	3.37	49.00	0.35	0.65	WHERE-M
m	*FOCUS*	*DATA COMPLEXITY*		65	13.14	13.79	14.42	0.38	0.62	3.07	3.24	3.42	46.00	0.54	0.46	WHERE-H
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	21	7.46	6.90	23.92	0.33	0.67	2.15	1.58	4.13	14.00	0.50	0.50	WHAT_Ql-easy-E
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	16	12.55	2.30	13.56	0.44	0.56	2.83	1.25	3.28	11.00	0.64	0.36	WHAT_Ql-easy-M
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	23	10.29	4.69	15.81	0.35	0.65	2.88	1.95	3.68	16.00	0.50	0.50	WHAT_Ql-easy-H
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	24	4.72	1.50	14.74	0.33	0.67	1.88	1.11	3.48	18.00	0.44	0.56	WHAT_Ql-medium-E
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	26	9.49	13.69	19.43	0.31	0.69	2.73	2.43	3.78	17.00	0.47	0.53	WHAT_Ql-medium-M
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	20	9.72	4.02	17.76	0.20	0.80	2.73	1.53	3.99	16.00	0.25	0.75	WHAT_Ql-medium-H
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	20	9.03	6.54	11.79	0.15	0.85	2.63	1.61	2.89	13.00	0.23	0.77	WHAT_Ql-hard-E
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	23	10.51	6.15	13.79	0.43	0.57	3.00	2.26	3.36	20.00	0.50	0.50	WHAT_Ql-hard-M
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	22	12.28	9.12	14.37	0.45	0.55	2.82	2.62	3.17	20.00	0.50	0.50	WHAT_Ql-hard-H
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	22	8.63	6.04	22.67	0.45	0.55	2.62	2.19	4.08	18.00	0.56	0.44	WHAT_Qn-easy-E
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	19	9.37	5.63	24.46	0.47	0.53	2.83	2.19	4.19	15.00	0.60	0.40	WHAT_Qn-easy-M
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	17	14.44	7.04	19.25	0.41	0.59	3.26	2.18	3.82	15.00	0.47	0.53	WHAT_Qn-easy-H
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	21	4.68	4.21	13.64	0.38	0.62	2.05	1.82	3.30	15.00	0.53	0.47	WHAT_Qn-medium-E
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	16	10.46	10.13	10.98	0.56	0.44	2.88	2.78	2.56	14.00	0.64	0.36	WHAT_Qn-medium-M
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	27	5.29	7.49	12.96	0.30	0.70	2.03	2.51	3.12	21.00	0.38	0.62	WHAT_Qn-medium-H
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	22	13.16	7.60	9.36	0.32	0.68	2.90	2.38	2.91	12.00	0.58	0.42	WHAT_Qn-hard-E
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	30	11.22	9.06	11.65	0.40	0.60	2.97	2.48	2.94	21.00	0.57	0.43	WHAT_Qn-hard-M
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	21	8.65	6.94	10.10	0.62	0.38	2.73	2.39	2.99	16.00	0.81	0.19	WHAT_Qn-hard-H
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	23	12.03	13.98	18.32	0.39	0.61	3.11	3.35	3.71	16.00	0.56	0.44	WHERE-easy-E
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	30	20.06	13.35	12.47	0.37	0.63	3.76	3.23	3.13	25.00	0.44	0.56	WHERE-easy-M
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	21	14.90	13.68	14.07	0.43	0.57	3.29	3.20	3.32	18.00	0.50	0.50	WHERE-easy-H
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	20	8.86	13.67	15.37	0.45	0.55	2.44	3.26	3.54	13.00	0.69	0.31	WHERE-medium-E
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	18	13.64	13.56	14.88	0.17	0.83	3.43	3.19	3.59	12.00	0.25	0.75	WHERE-medium-M
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	19	13.33	15.35	14.05	0.26	0.74	2.94	3.51	3.65	10.00	0.50	0.50	WHERE-medium-H
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	22	10.57	10.54	11.65	0.36	0.64	2.55	3.07	3.06	16.00	0.50	0.50	WHERE-hard-E
m	*FOCUS*	*MASK COMPLEXITY*	*DATA COMPLEXITY*	17	19.78	24.27	15.25	0.18	0.82	3.98	3.90	3.56	12.00	0.25	0.75	WHERE-hard-M

Table 5.10 Results summary for the Measurement study. The table indicates the absolute difference between responses from participants and the baselines, according to factors we varied. In this table, the factor most to the right is the variant. The columns on the left indicate the factors and the ones on the right indicate the detail of the factored selections, e.g. "WHERE-medium-E" indicates responses where the *FOCUS* was WHERE, the *MASK COMPLEXITY* was medium and the *DATA COMPLEXITY* was easy. The columns with "neither" in their titles discuss the results once responses for which participants answered "Neither agree nor disagree" are filtered.

cantly better than WHERE. These observations reveal a particularly strong order concerning performance and focus for MO (MO 2).

- Performance for *Mask Proportion MP* is consistent for each *FOCUS* , indicating, as we would expect, that the condition can be read independent on the *FOCUS* .

The results for the Measurement study of MwM , MO and MP with the variant being the *MASK COMPLEXITY* are displayed in Fig. 5.38.

- Responses to MwM (left) indicate that performance is relatively weak , but significantly less so when *MASK COMPLEXITY* is Medium. We also note that the performances are similar between *MASK COMPLEXITY* Easy and Hard. This goes against our expectations as instead of a trend of performances getting worse with *MASK COMPLEXITY* increasing, we find no global trend.

- MO (center) shows little difference in performance according to *MASK COMPLEXITY* , which fits our expectations.

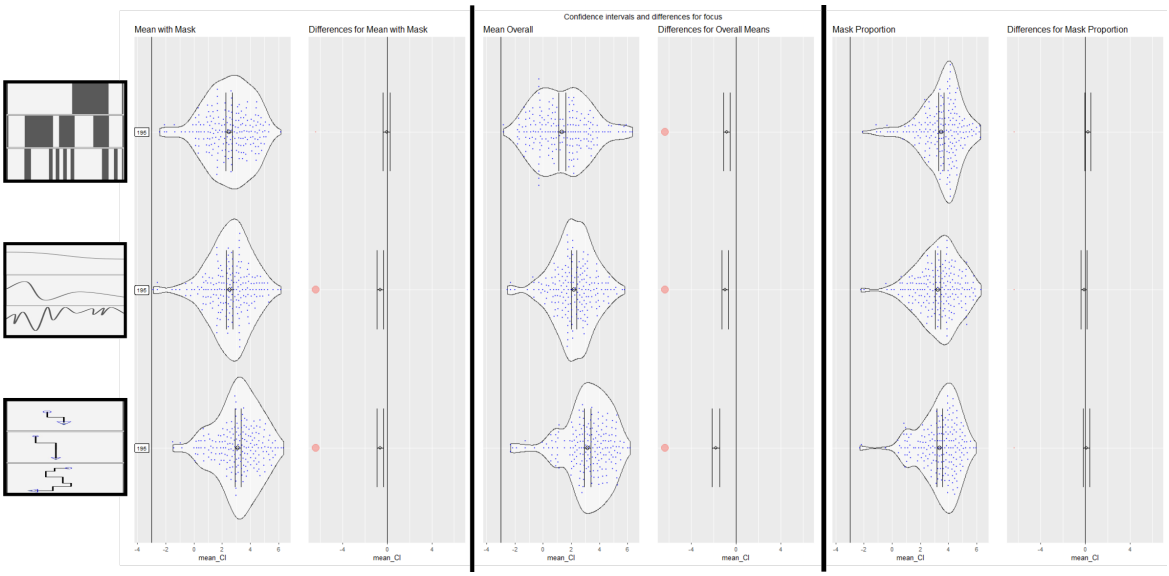


Fig. 5.36 The overall results for the Measurement study according to *FOCUS* .
FOCUS has an effect on performance in MwM and MO tasks. Specifically, we see that both MwM and MO are estimated less well in spatial (WHERE) than binary attribute (WHAT_Ql) *FOCUS* conditions.

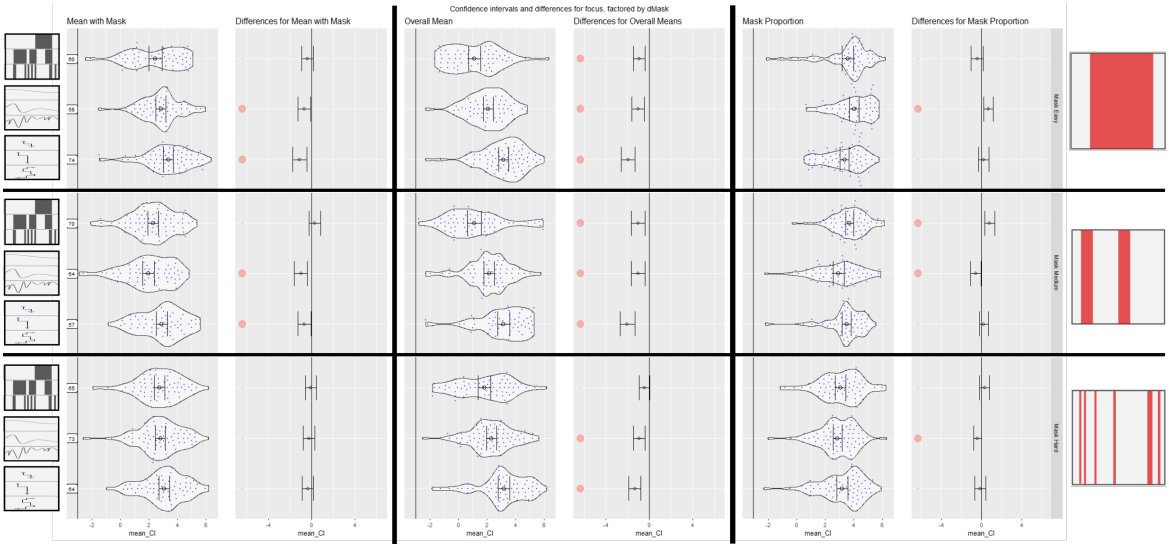


Fig. 5.37 The overall results for the Measurement study according to *FOCUS* factored by *MASK COMPLEXITY* . We note a significant trend for which performances of MO are ordered as such: WHAT_Ql, WHAT_Qn, WHERE. This trend is always not statistically significant between WHAT_Ql and WHAT_Qn when the *MASK COMPLEXITY* is Hard, but remains true, thus indicating that the trend is globally important.
For the task MwM , the *FOCUS* WHAT_Ql results in significantly better performances than WHAT_Qn and WHERE with *MASK COMPLEXITY* being Easy and Medium, while the differences are minor while the *MASK COMPLEXITY* is Hard.

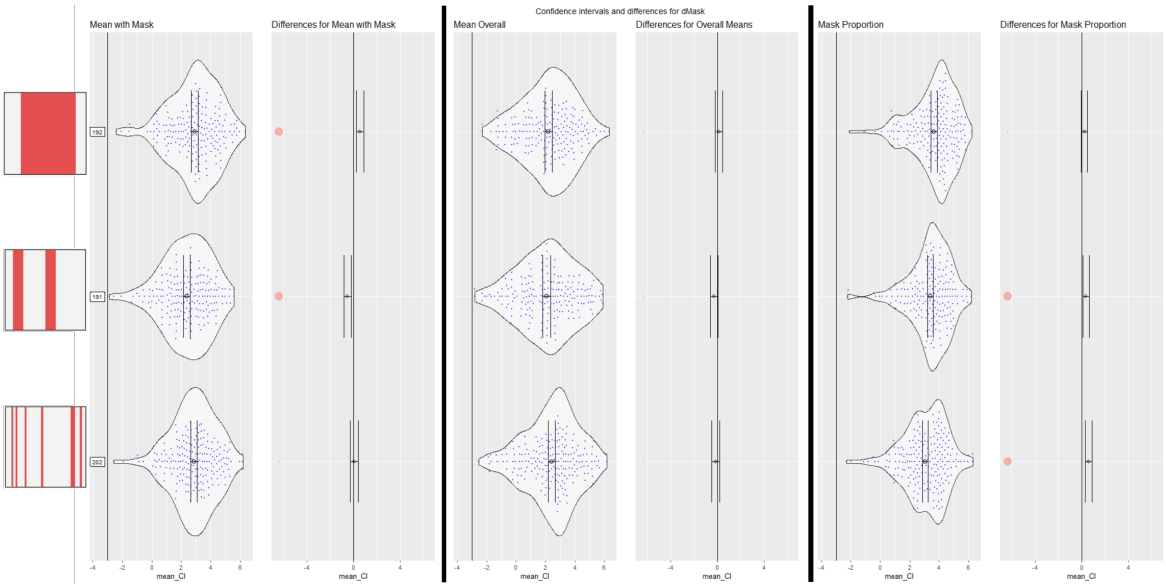


Fig. 5.38 The overall results for the Measurement study according to *MASK COMPLEXITY* . *MASK COMPLEXITY* has an effect on task (MP) with performance *improving* as mask complexity increases. Performance in Mean with Mask tasks (MwM) was better when *MASK COMPLEXITY* was *Medium* than in the Easy or Hard conditions.

- MP performance (right) is significantly different between *MASK COMPLEXITY* Medium and Hard, and *MASK COMPLEXITY* Easy and Hard, with performance deteriorating as the *MASK COMPLEXITY* increase. This is counter-intuitive, potentially indicating that other parameters than the ones we selected are impacting ability to perform MP .

The results for the Measurement study of MwM , MO and MP with the variant being the *DATA COMPLEXITY* factored by *FOCUS* are displayed in Fig. 5.40.

- MwM (left) presents a single case of significant difference when the focus is WHAT_Qn, which leads us to think it is a false positive.

  There are two significant differences when the focus is WHERE. Each group with the complexity of WHERE is Easy is significantly better than groups with WHERE is set to a Medium complexity. We notice some occurrences where performances for MwM follow the order of the complexity of the focus, but Medium complexity results several times in poorer performances than Hard, which drives us not to consider these as valid for our analysis. WHERE Medium ask performance is significantly worse than the Easy or Hard conditions for MwM . This is unexpected and not something we can explain for certain. We discuss visual analysis of stimuli with the poorest performances in section 5.6.2.

- MO (center) The 95% confidence intervals of performance for Overall Mean for focus being WHAT_Ql at an Easy complexity contain the baseline, indicating fairly good performances

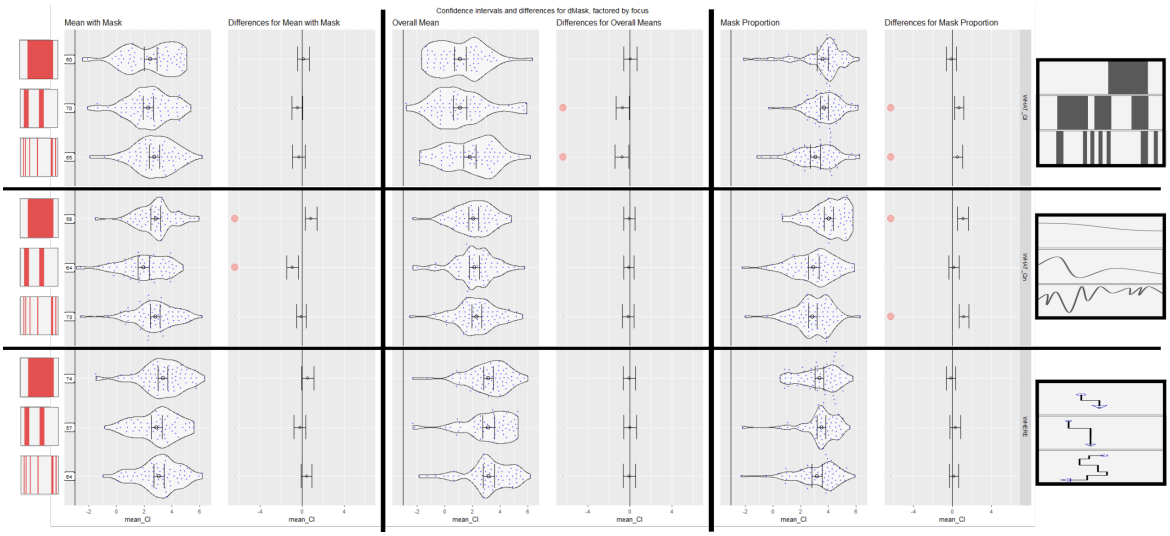


Fig. 5.39 The overall results for the Measurement study according to *MASK COMPLEXITY* factored by *FOCUS* .
For *FOCUS* WHAT_Ql we note some significant differences between *MASK COMPLEXITY* Medium and Hard and Easy and Hard for MO and MP , both cases due to strong differences in performances for *MASK COMPLEXITY* Hard, but worse performances for MO and better performances for MP . We also note unexpected differences of performances for *FOCUS* WHAT_Qn, i.e. for MwM the performances are significantly better for *MASK COMPLEXITY* Medium and MP are significantly worse for *MASK COMPLEXITY* Easy.

from participants in this case (MwM 0). But absolute errors increase as *FOCUS* complexity increases. Performance in the WHAT_Ql condition with Easy complexity is better than the harder conditions, adding to the evidence that tasks can be undertaken with some success at this focus and *DATA COMPLEXITY* level (MwM 1).

- MP (right)Performances for Mask Proportion indicate little variations according to the factors we set.

Measurement study data for the MwM (left), MO (centre) and MP (right) tasks with **variation in *DATA COMPLEXITY* , factored by *MASK COMPLEXITY*** are displayed in Fig. 5.41.

- MwM (left) Performance is relatively weak but significantly stronger where *FOCUS* complexity is Easy over Medium for all *FOCI* .
- MO (center) performance is not affected by *MASK COMPLEXITY* across all *FOCI* . The tendency for MO performance to be better when *FOCUS* is Easy reported above is reflected in the lower error rates for Easy WHAT_Qn and WHERE, one of which is significant in the former case.
- MP (right) Performances for Mask Proportion is not dependent on *FOCUS* complexity. This could have been a distractor when interpreting Time Masks, but we see no evidence of this in

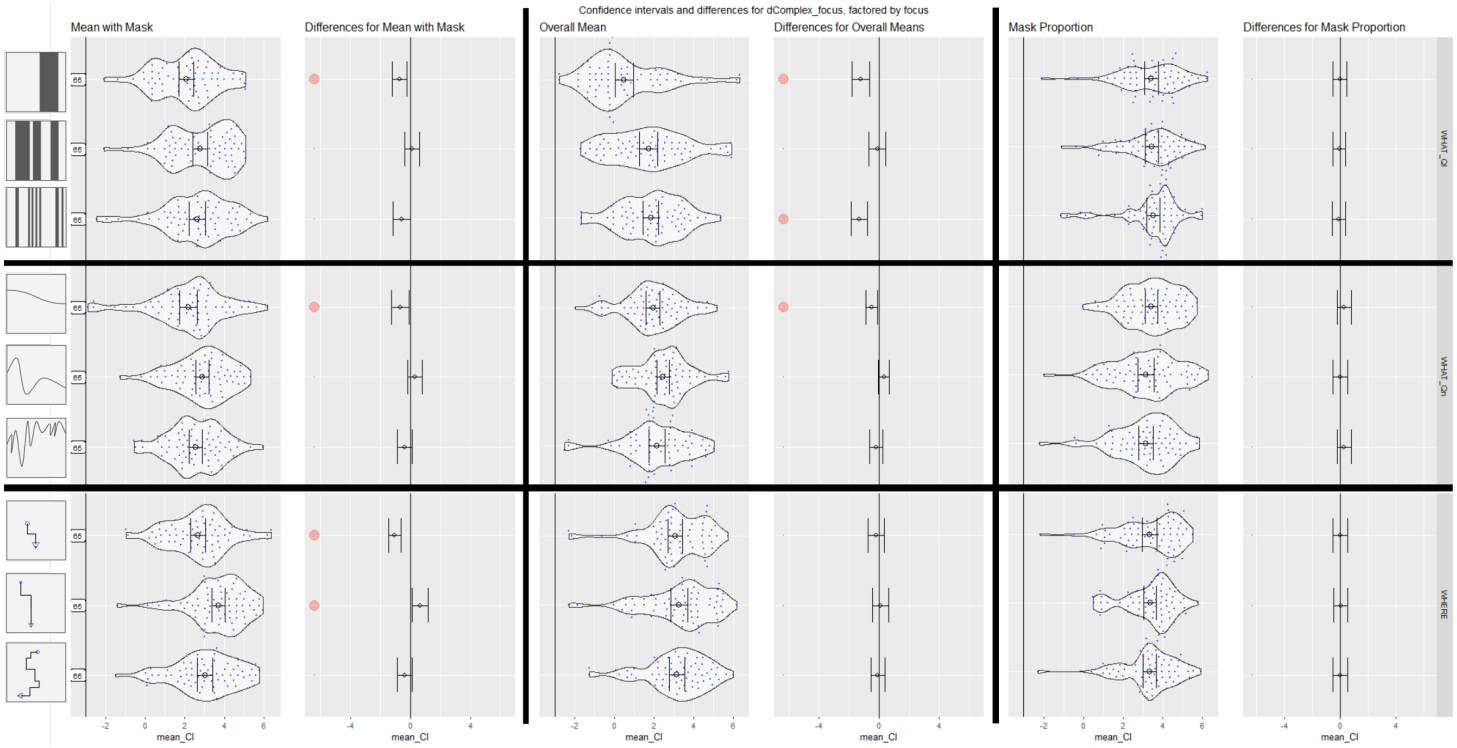


Fig. 5.40 Results for the Measurement Study, with the variant being the *DATA COMPLEXITY* , factored by *FOCUS* .
We note some evidence that MO performance is better when *FOCUS* complexity is lower, but not in the case of WHERE, where performance is consistently poor.
MwM performance is low (add absolute numbers - means) but with some evidence that this is less true when *FOCUS* complexity is low as MwM performance is significantly better for Easy level of complexity than at least one of the more challenging complexity levels in all three *FOCI* .
As expected, mask proportion (MP) performance does not vary with *FOCUS* .

our results, suggesting that masks and data can be interpreted independently.

Measurement study data for the MwM (left), MO (center) and MP (right) tasks with variation in *DATA COMPLEXITY* factored by both *MASK COMPLEXITY* and *FOCUS* are displayed in Fig. 5.42.

Considering the high amount of categories from this double factorisation, only three observations stand out: performances for MwM are significantly better when the focus is WHERE Easy, that WHERE Medium, independently of the *MASK COMPLEXITY* ; performances for Overall Mean are significantly good for WHAT_Ql with an Easy complexity, independently of the Mask; and performances for WHAT_Ql with an Easy complexity are significantly better than WHAT_Ql Medium and WHAT_Ql Hard independently of *MASK COMPLEXITY* .

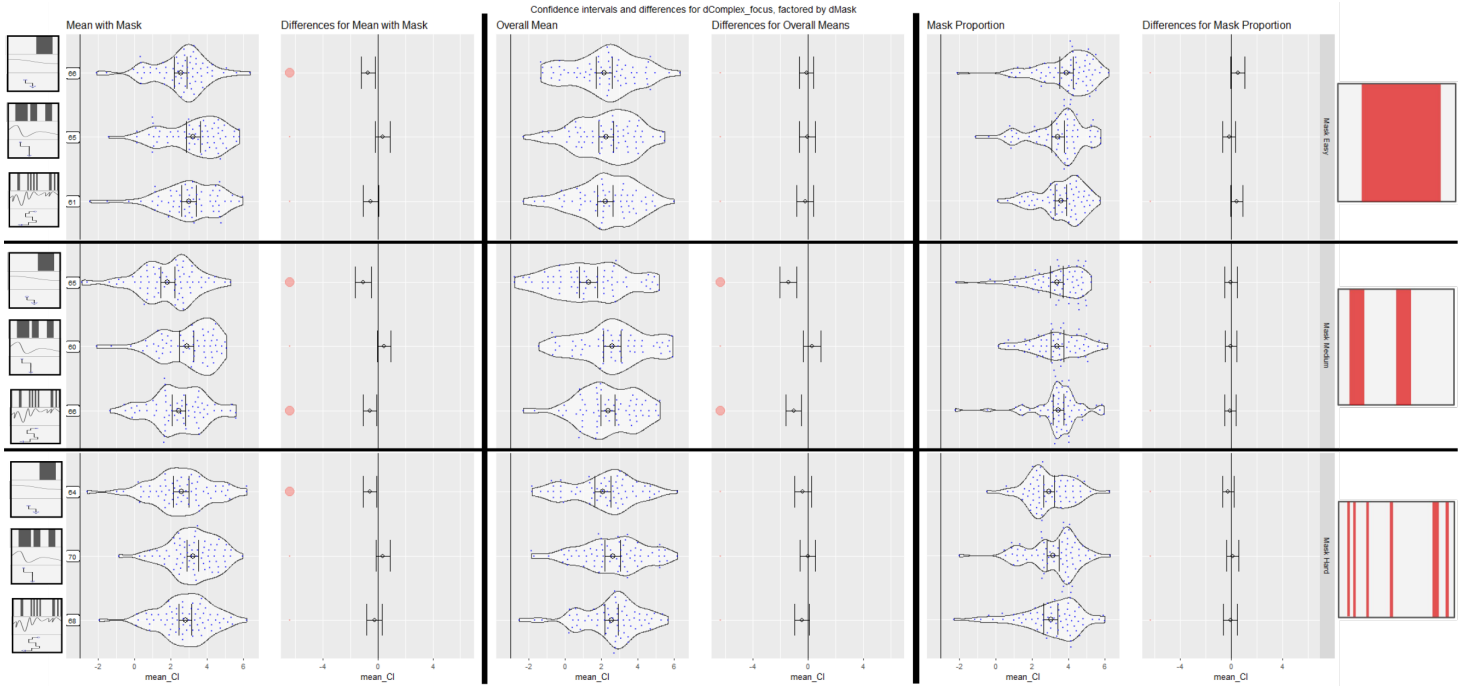


Fig. 5.41 Results for the Measurement Study, with the variant being the *DATA COMPLEXITY* , factored by the *MASK COMPLEXITY* .
MwM performance is low , but significantly better for Easy/Simple *FOCUS* than higher levels of *FOCUS* complexity across all types of *FOCUS* and all levels of mask complexity. This is important as it suggests that our abilities to interpret Masks are relatively limited in terms of the complexity of underlying data.
MO performance is weak, but there is a suggestion that this is better when *FOCI* are less complex, as shown by the single significant difference (*MASK COMPLEXITY* Medium) and supporting trends in other mask complexities. This effect is shown more clearly in Fig. 5.36 and Fig. 5.40and, as anticipated, is not particularly dependent upon *MASK COMPLEXITY* .
As expected, mask proportion (MP) performance does not vary with *FOCUS* , when factored by *MASK COMPLEXITY* .

- MwM task performance is reasonable for WHAT_Ql (binary) settings (add average error in absolute terms), with no systematic bias. Errors are lower where *FOCUS* complexity is low (Easy) irrespective of *MASK COMPLEXITY* in with WHAT conditions, but this effect is weakest in the complex (Hard) Mask case. In the WHERE *FOCUS* , performance is slightly less strong (add average error in absolute terms and see below) but there is a consistent difference between the Easy and Medium *FOCUS* complexity levels, with the less complex WHERE *FOCUS* resulting in significantly smaller errors for all levels of *MASK COMPLEXITY* . Surprisingly, this pattern is not apparent in the Most complex (hard) *FOCI* , with MwM performance in tasks with complex (Hard) trajectories being indistinguishable from the with simple (Easy) trajectories for all levels of *MASK COMPLEXITY* . (But then there is of course the question

about whether these errors are equivalent across *FOCUS* given that they use different numbers of pixels, etc!)

- MO There is a relatively consistent pattern here that shows MO performance to be better when *FOCUS* complexity is low in the WHAT_Qn and particularly the WHAT_Ql *FOCI* . This is supported by general trends and seven significant differences between conditions that all point in this direction. We do not see strong evidence that this is dependent upon *MASK COMPLEXITY* and nor would we expect it to be. It is not the case in WHERE tasks, where performance is low irrespective of *FOCUS* complexity. So we can claim with some confidence that Overall Mean estimation tasks are achievable where *FOCI* are simple, but performance deteriorates for WHAT, but not WHERE, as *FOCUS* complexity increases.

- MP task performance is not influenced by *FOCUS* complexity for any *FOCUS* for any level of *MASK COMPLEXITY* .

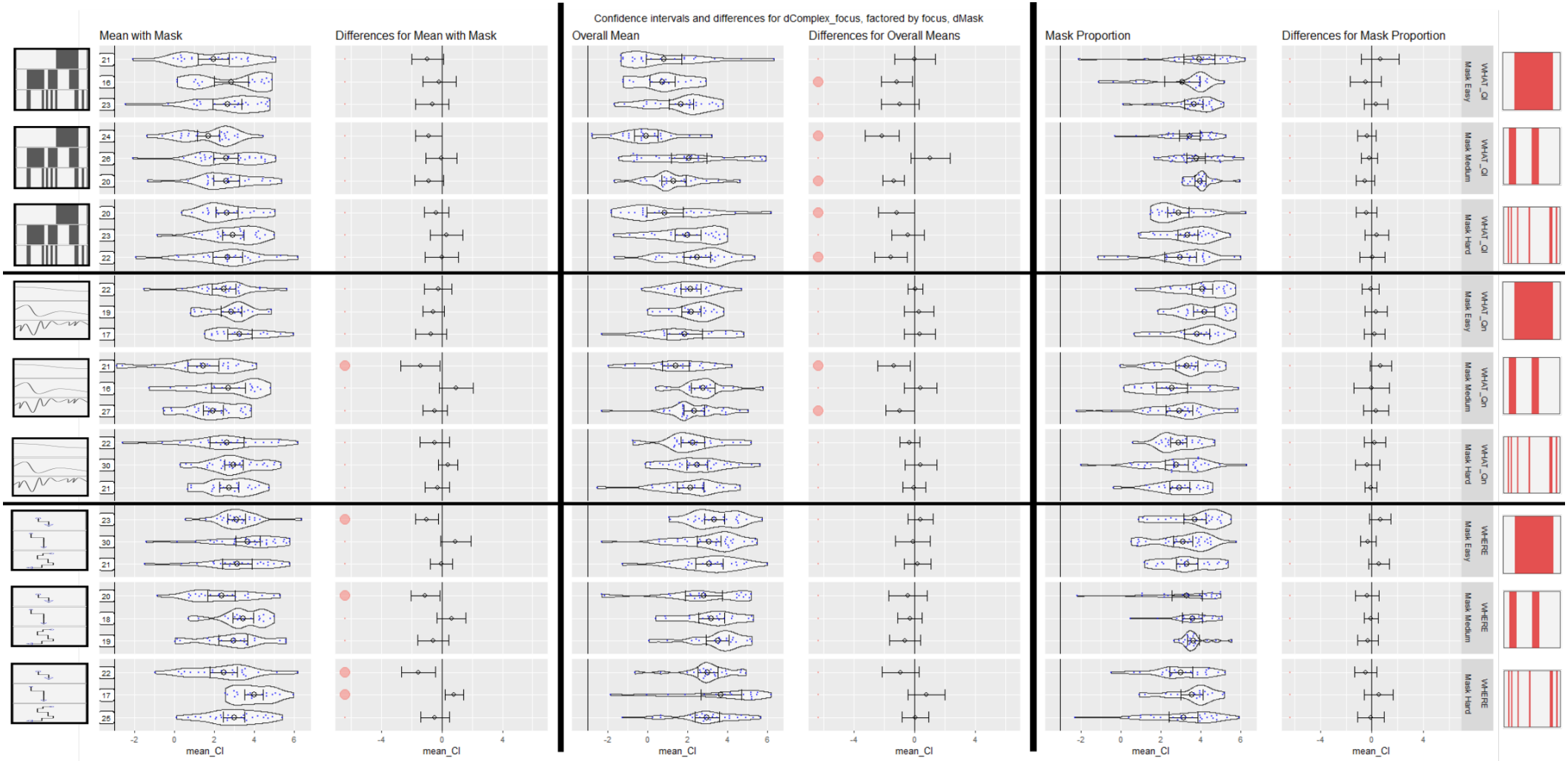


Fig. 5.42 The overall performances for the Measurement study for MwM , MO and MP with the variant being *DATA COMPLEXITY* , factored by *FOCUS* and *MASK COMPLEXITY* . We note that distribution of responses for *FOCUS* WHAT_Ql is not always normal for the MwM . We consider this could be explained by different understandings of either the data presented, the task asked of participants, or the phrasing somehow misleading. We discuss potential explanations in section 5.6. For *FOCUS* WHAT_Qn MwM and MO resulted in better performances when the *DATA COMPLEXITY* was Easy, but the effect was moderate. For MO and *FOCUS* WHAT_Ql, *DATA COMPLEXITY* differences are often significant, following the trend we expect of higher *DATA COMPLEXITY* resulting in higher error rate, except for the case where the *MASK COMPLEXITY* is medium, for which *DATA COMPLEXITY* medium has significantly higher error rate.
Performances for MP are low overall, with relatively low differences according to factors.

Likert question: Stability Comparison (SC)

The graphs to discuss the Likert questions follow the same structure as the graphs to discuss the Numerical questions.

Data showing performance for Stability Comparison with the variant set to focus are displayed in Fig. 5.43. Performance in this task was poor. As showed in table 5.10 when considering answers where participants answered "neither agree nor disagree" as wrong, the mean of error rate was higher than 50%, i.e. worse than random. It is equal to 50% when considering responses "neither agree nor disagree" as neither correct nor incorrect but filtering them out. As mentioned previously, we consider that labelling responses "neither agree nor disagree" as correct nor incorrect is fallacious, and thus filter these out from our analysis. This leads us to ask questions about abilities to interpret levels of variation within and beyond the conditions displayed visually in A-ATS Masks. Given this, we do see that performance for Stability Comparison was significantly better for WHAT_Qn than the other *FOCI* .

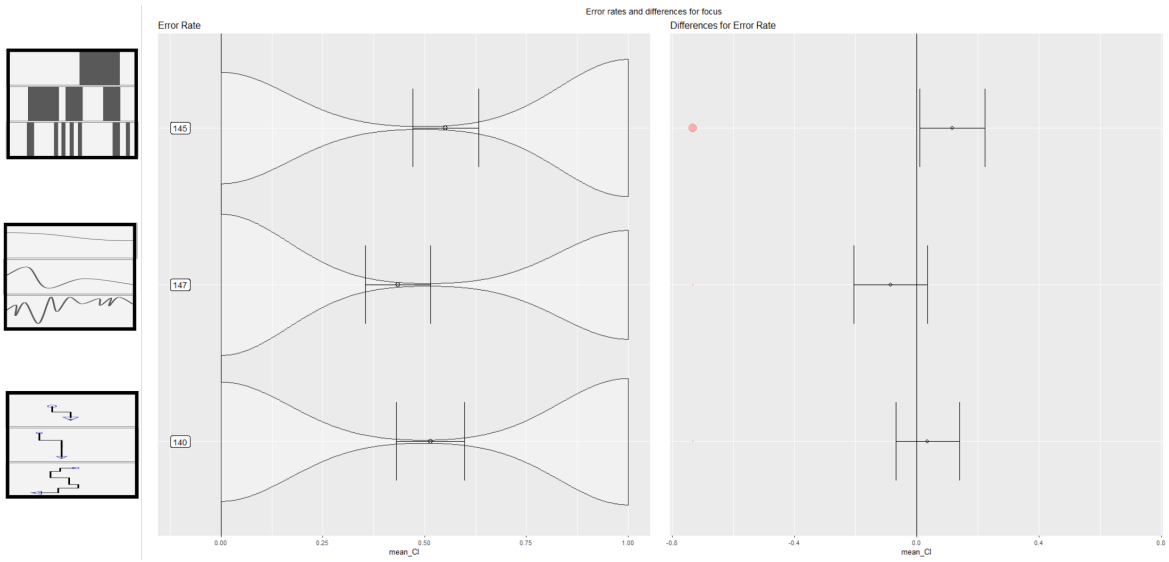


Fig. 5.43 The SC answers with the variant being *FOCUS* . The performances are significantly better when the *FOCUS* is WHAT_Qn.

Data showing performance in the Stability Comparison task, varying by *MASK COMPLEXITY* are displayed in figure Fig. 5.45. Stability Comparison does not vary according to *MASK COMPLEXITY* , and there is no trend to suggest that Mask complexity has an effect on performance.

We can dig deeper with graphics that set the variant to complexity of the *FOCUS* and factor by *FOCUS* or Mask.

The results for the SC with the variant being *FOCUS* factored by *MASK COMPLEXITY* are displayed in Fig. 5.44 and show the strongest differences between *FOCI* in the results where the *MASK COMPLEXITY* is Hard. Particularly, *FOCUS* WHAT_Qn responses are significantly better. This observation is true with all *MASK COMPLEXITY* but only statistically significant in the

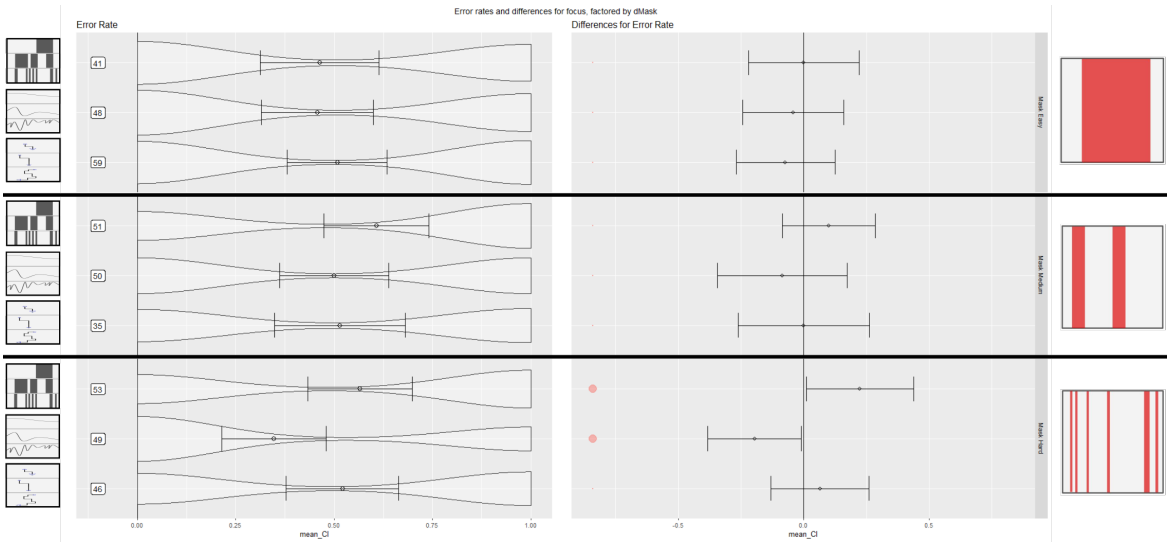


Fig. 5.44 The SC answers with the variant being *FOCUS* , factored by *MASK COMPLEXITY* . We note that differences are relatively small between *FOCI* , except in the case of *MASK COMPLEXITY* being Hard where the *FOCUS* WHAT_Qn results in significantly better performances.

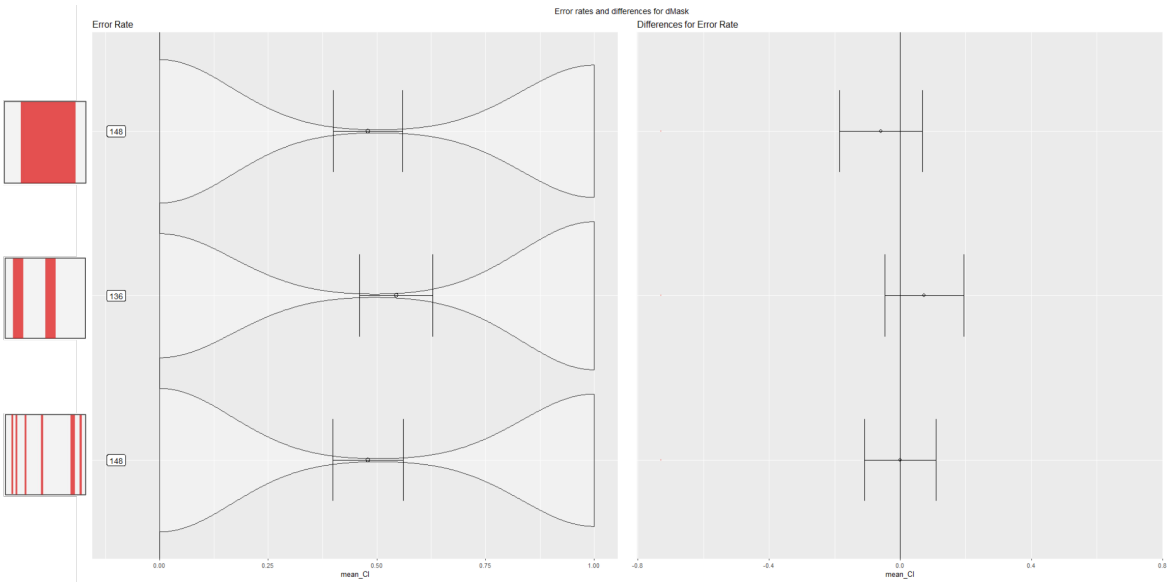


Fig. 5.45 The SC answers with the variant being *MASK COMPLEXITY* . The differences between variants are not statistically significant.

previously mentioned case, which leads us to consider that the higher *MASK COMPLEXITY* results in a greater ease to perform SC . This goes against our expectations.

Performance data for the Stability Comparison task, with varying *DATA COMPLEXITY* , factored by *MASK COMPLEXITY* are displayed in Fig. 5.48. We notice that performances get poorer for SC as *DATA COMPLEXITY* increases for both *MASK COMPLEXITY* Easy and Medium, but the opposite

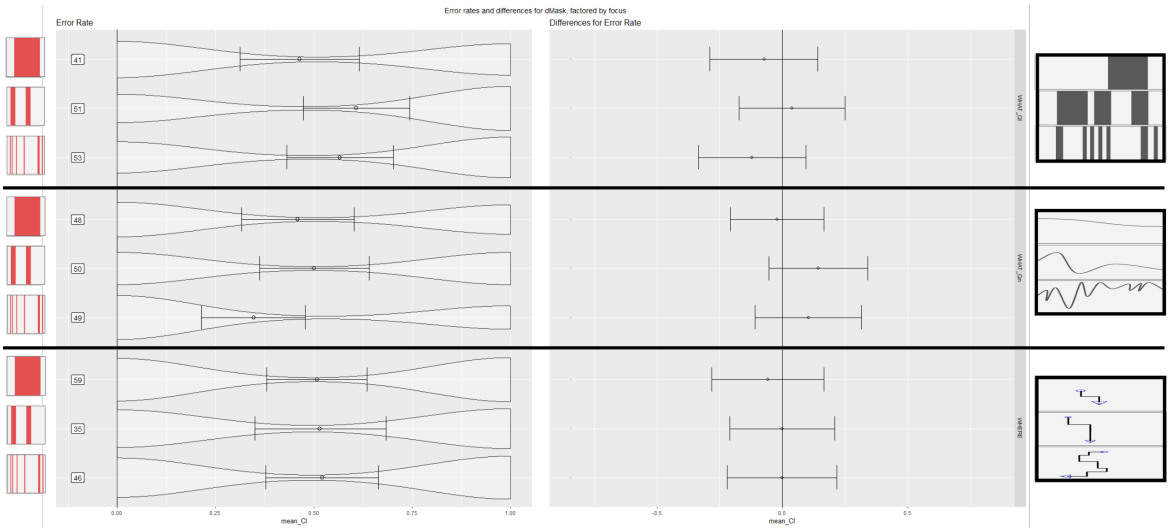


Fig. 5.46 The SC answers with the variant being *MASK COMPLEXITY* , factored by *FOCUS* . The differences between variants are not statistically significant.

occurs for *MASK COMPLEXITY* Hard. Additionally, performances with *MASK COMPLEXITY* Hard and *DATA COMPLEXITY* Hard are better than the others, which goes against our expectations.

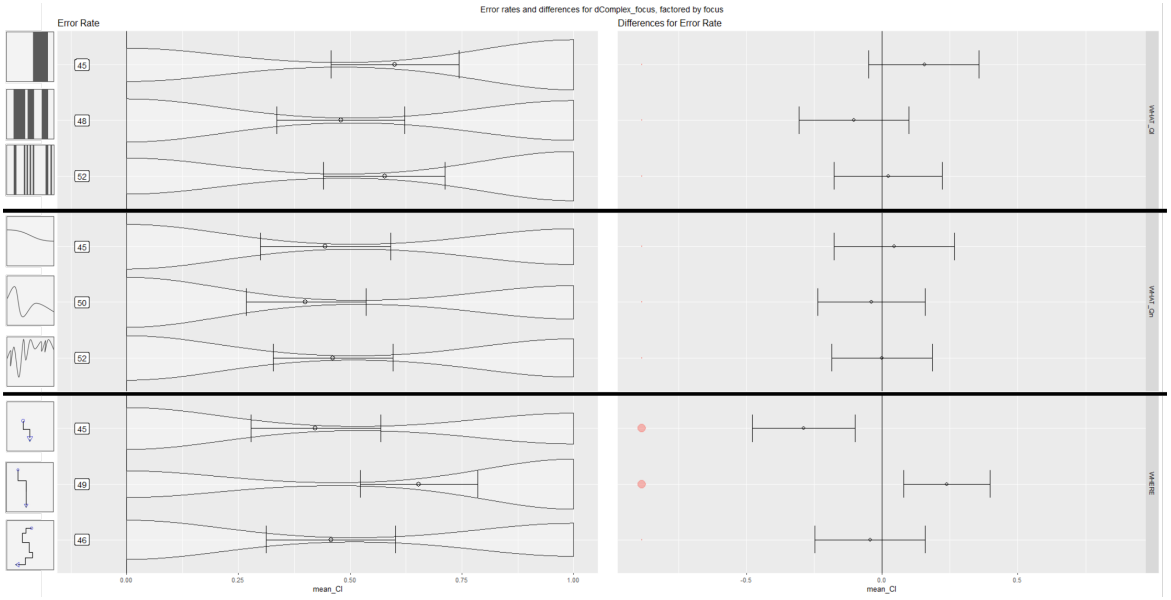


Fig. 5.47 SC according to *DATA COMPLEXITY* factored by *FOCUS* . Variations of *DATA COMPLEXITY* are not significant for *FOCUS* WHAT_Qn and WHAT_Ql, but are significant for *FOCUS* WHERE in two variations (Easy-Medium, Medium-Hard) as the performances for the SC are much worse when the *MASK COMPLEXITY* is set to Medium.

The Stability Comparison performance data are presented with the variant set to *DATA COMPLEXITY* , factored by *FOCUS* in Fig. 5.47. Performance is low and there is no difference in error

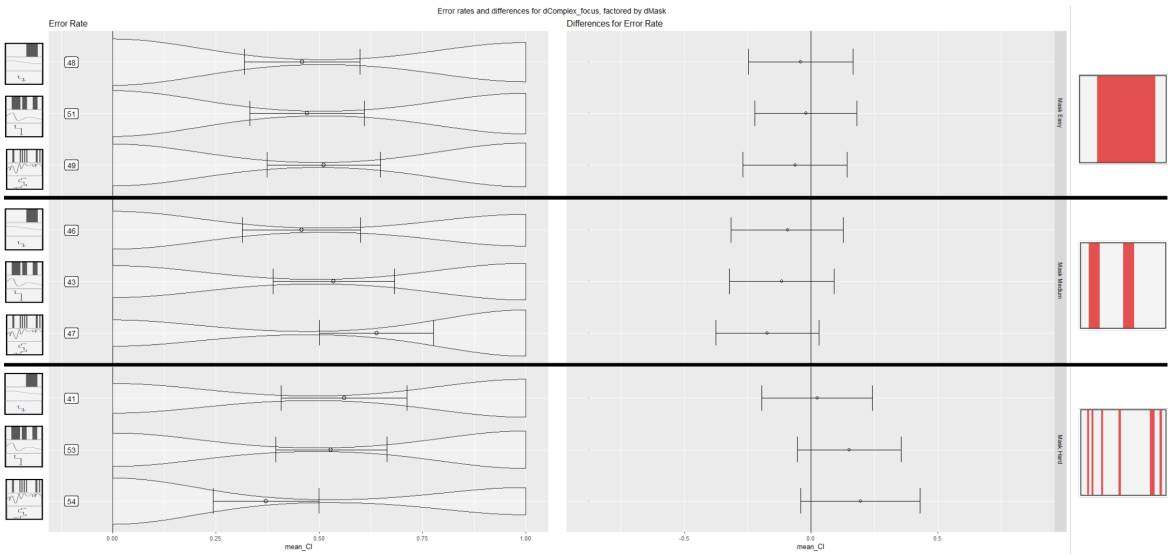


Fig. 5.48 Performances for SC with *DATA COMPLEXITY* as the variant and factored by *MASK COMPLEXITY* . We note that while light, there is a trend of *DATA COMPLEXITY* increasing resulting in poorest performances for *MASK COMPLEXITY* Easy and Medium, but the trend is inverted for *MASK COMPLEXITY* Hard.

rates within or between factors, meaning that we see no evidence that *DATA COMPLEXITY* has an effect for any *FOCUS* .

The performance data for the Stability Comparison task are presented with the variant set to *DATA COMPLEXITY* , factored by both *FOCUS* and *MASK COMPLEXITY* in Fig. 5.49. This reveals low performance across the board, and given that we are looking at 27 relationships, finding two that are statistically significant at the 95% confidence level should not surprise us. We consider them likely to be false positives as they do not relate to any other evidence and are not easily explained. The double factoring of the data results in small numbers of observations being used in our bootstrapping and thus wide confidence intervals, which give us reason to be cautious concerning claims about light trends observed. As such, we conclude that task performance is surprisingly and consistently low for Stability Comparison SC and that *FOCUS* complexity has no effect for any mask complexity for any *FOCUS* .

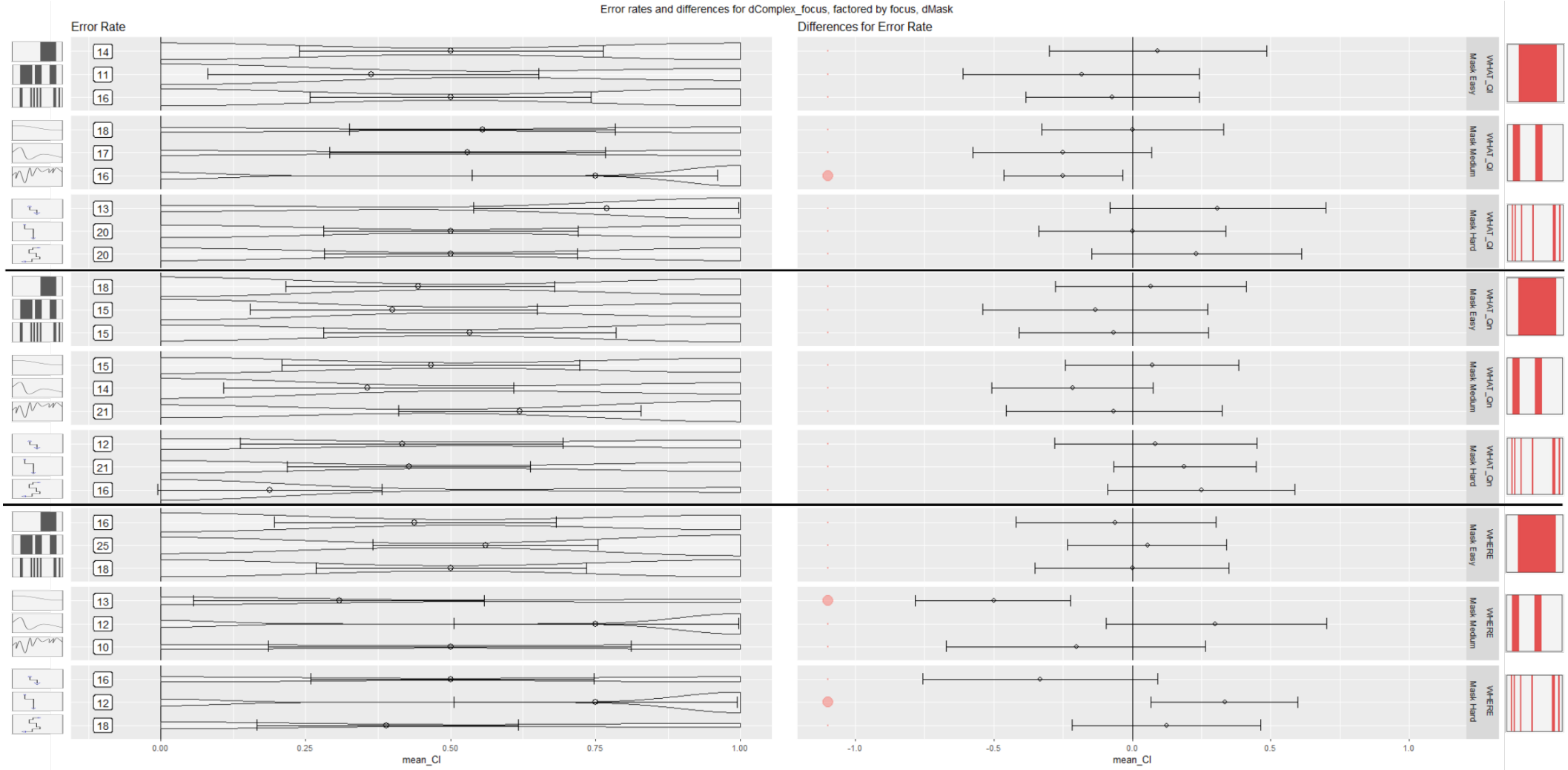


Fig. 5.49 The overall performances for the Measurement study for SC with variant *DATA COMPLEXITY* , factored by *FOCUS* and *MASK COMPLEXITY* . With the *FOCUS* being WHAT_Ql, the only significant difference is with *MASK COMPLEXITY* Medium and between the *DATA COMPLEXITY* Easy and Hard. There is no significant difference between results when the *FOCUS* is WHAT_Qn. We note two significant differences when the *FOCUS* is WHERE: one with *MASK COMPLEXITY* Medium, between *DATA COMPLEXITY* Easy and Medium, and *MASK COMPLEXITY* Hard between *DATA COMPLEXITY* Medium and Hard. Both cases indicate much poorer performances when the *DATA COMPLEXITY* is Medium. Trend of difficulty do not follow our expectations of performances getting poorer as *DATA COMPLEXITY* increases.

Self-reported confidence

In this section, we discuss the impact of the focus, *MASK COMPLEXITY* and their combination over the self-reported confidence for questions MwM , MO , MP and SC .

We display the distribution of self-reported confidence in Fig. 5.50 and Fig. 5.51. For each question, we assess whether different scaling results in significantly different levels of confidence using the Dunn test. The results of these tests are listed in table 5.5.

The Dunn tests comparing self-reported confidence according to focus indicate significant differences for MwM and MO . Visual analysis of the graphs in Fig. 5.50 suggest a global trend: for both MwM and MO , self-reported confidence is highest for WHAT_Ql, followed by WHAT_Qn, and lowest for WHERE. The average of self-reported confidence is respectively 3.65, 3.11 and 2.52.

The differences between *FOCI* for MP are not statistically significant and visually look quite similar Fig. 5.50. Visually, the bar charts indicate similar trends according to the different focus. With average means of self-reported confidence of 3.55, 3.43 and 3.47 for WHAT_Ql, WHAT_Qn and WHERE, this seems to indicate a relatively high confidence in the ability to perform MP , independently of the focus.

The differences between focus for SC is only significant between WHAT_Ql and WHERE. With respective means of self-reported confidence of 3.39, 3.32 and 3.05 for WHAT_Ql, WHAT_Qn and WHERE, and a relatively strong visual difference of trends in the graphs of Fig. 5.50. These results show that SC is perceived to be most straightforward for WHAT_Ql, perhaps slightly more challenging for WHAT_Qn, and significantly harder for WHERE.

In combination, these results provide strong evidence to support a global trend for tasks with WHERE to result in lower self-reported confidence than the two WHAT tasks.

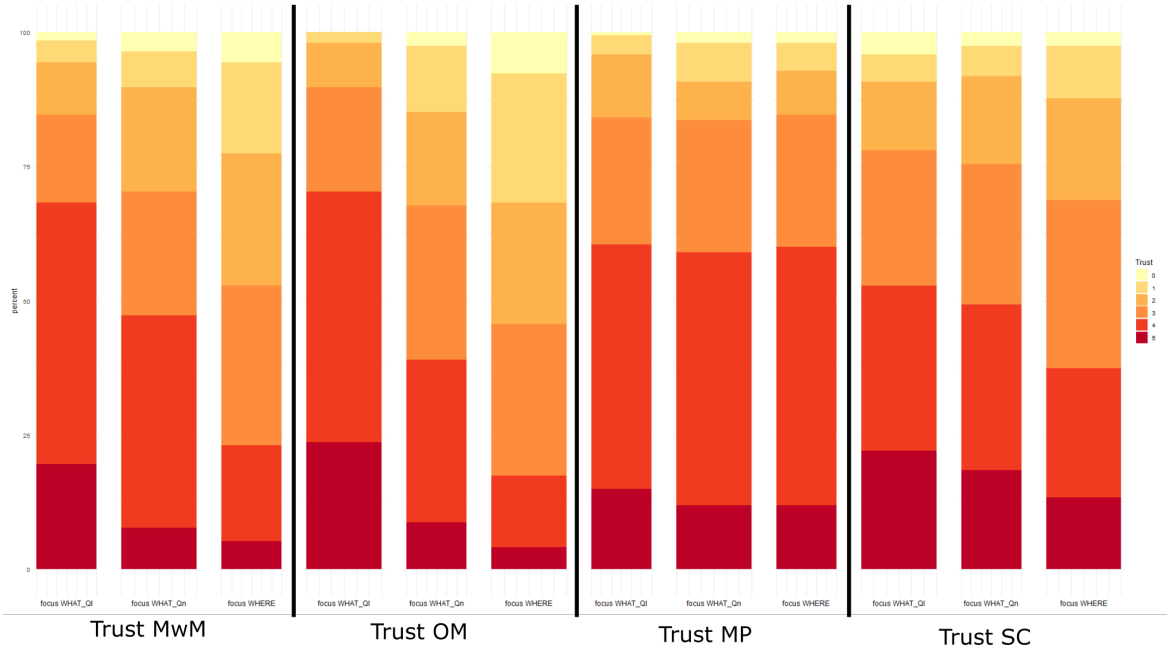


Fig. 5.50 The figures are composed of bar charts indicating self-reported trust according to *FOCUS* . Each set of bar charts shows trust in the four tasks in the following order: MwM , MO , MP and SC . Bars are coloured from yellow to red, with 0 indicating no confidence at all in an answer and 5 indicating absolute confidence in an answer.

We follow the same approach to assess the impact of *Mask complexity* as the one we just used for *FOCUS* . According to the Dunn tests listed in table 5.5, *MASK COMPLEXITY* does not result in significant differences in self-reported confidence for MwM , MO and SC . This was expected for MO as this questions does not rely in any way on considerations of the Mask, but was a surprise for MwM and SC , both of which involve consideration of the mask and its form.

MP indicated significant differences between self-reported confidence according to Mask for Mask Easy against Mask Medium and Mask Easy against Mask Hard. Visual analysis of the graphs show a large portion (almost 75%) of self-reported trusts are 4 or 4 out of a maximum 5. We thus conclude that participants are more confident in their ability to perform the tasks asked of them when the Mask is Easy.

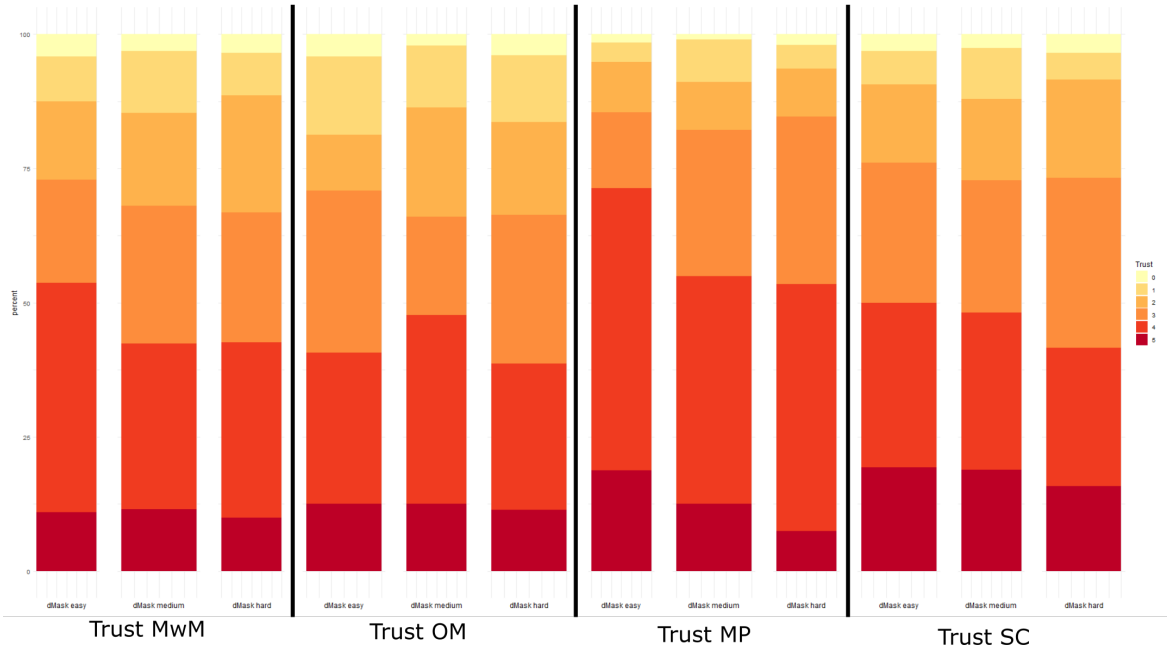


Fig. 5.51 The figures are composed of bar charts indicating self-reported trust according to the *MASK COMPLEXITY* . Each set of bar charts shows trust in the four tasks in the following order: MwM , MO , MP and SC . Bars are coloured from yellow to red, with 0 indicating no confidence at all in an answer and 5 indicating absolute confidence in an answer.

The relationships between performances and self-reported confidence

It is important to assess the relationship between performance and self-reported confidence, as two scenarios are to be assessed and avoided:

- Participants under evaluating their ability to perform the tasks asked of them using the A-ATS Mask implies that they would likely not make decisions based on graphs using this visualization method.
- Participants over evaluating their ability to perform the tasks asked of them using the A-ATS Mask implies that they would likely make decisions based on wrongfully interpreted information using this visualization method.

We thus analyse the distribution of answers according to self-reported confidence and verify differences of trends according to them. Similarly to our approach to analyse results according to previous factors, we analyse the data with different levels of factorization to reinforce or reject interpretation of results.

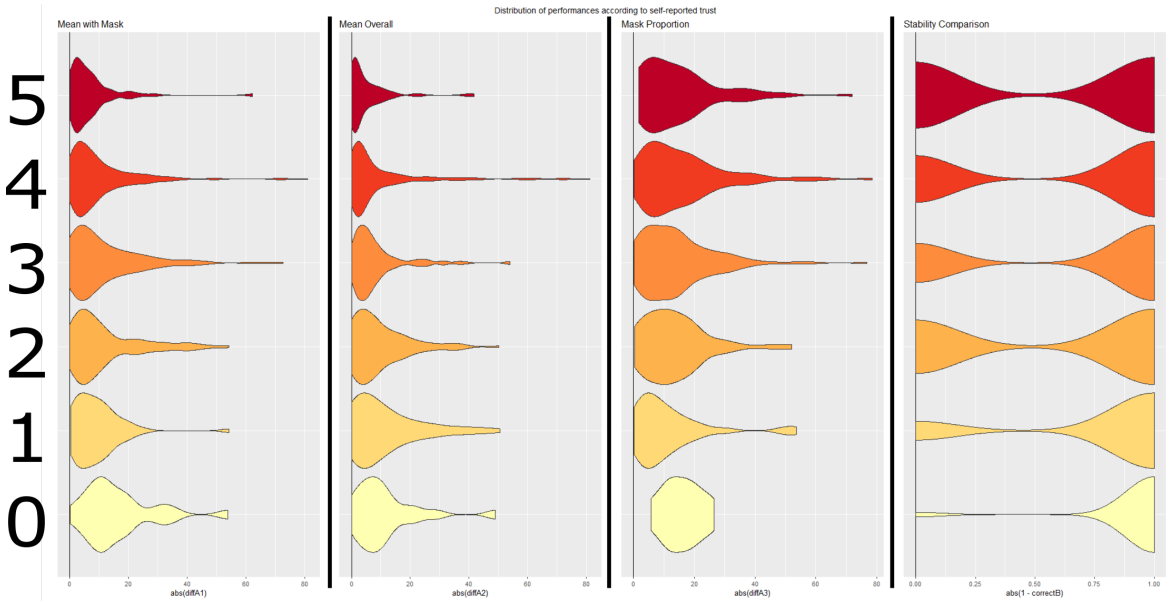


Fig. 5.52 The overall performances for the Measurement study according to the self-reported confidence. The plots are ordered with the highest trust (5 out of 5) at the top, and with the lowest trust (0 out of 5) at the bottom. The performances for MwM , MO and MP are set with the absolute value of the participants answers minus the baseline, thus the lower the difference, the better the performances. The display of the SC distribution is set with the correct answers being on the left and the wrong answers on the right.
Distributions seem to display a minor positive correlation between levels of self-reported confidence for MwM . MO seems to indicate a minor correlation relationship between self-reported confidence and performance. MP seems to indicate a minor negative correlation between self-reported confidence and performance. SC shows a minor positive correlation between performance and self-reported trust.
These results indicate that participants appear fairly effective at assessing their own ability to perform the tasks asked of them, except for MP .

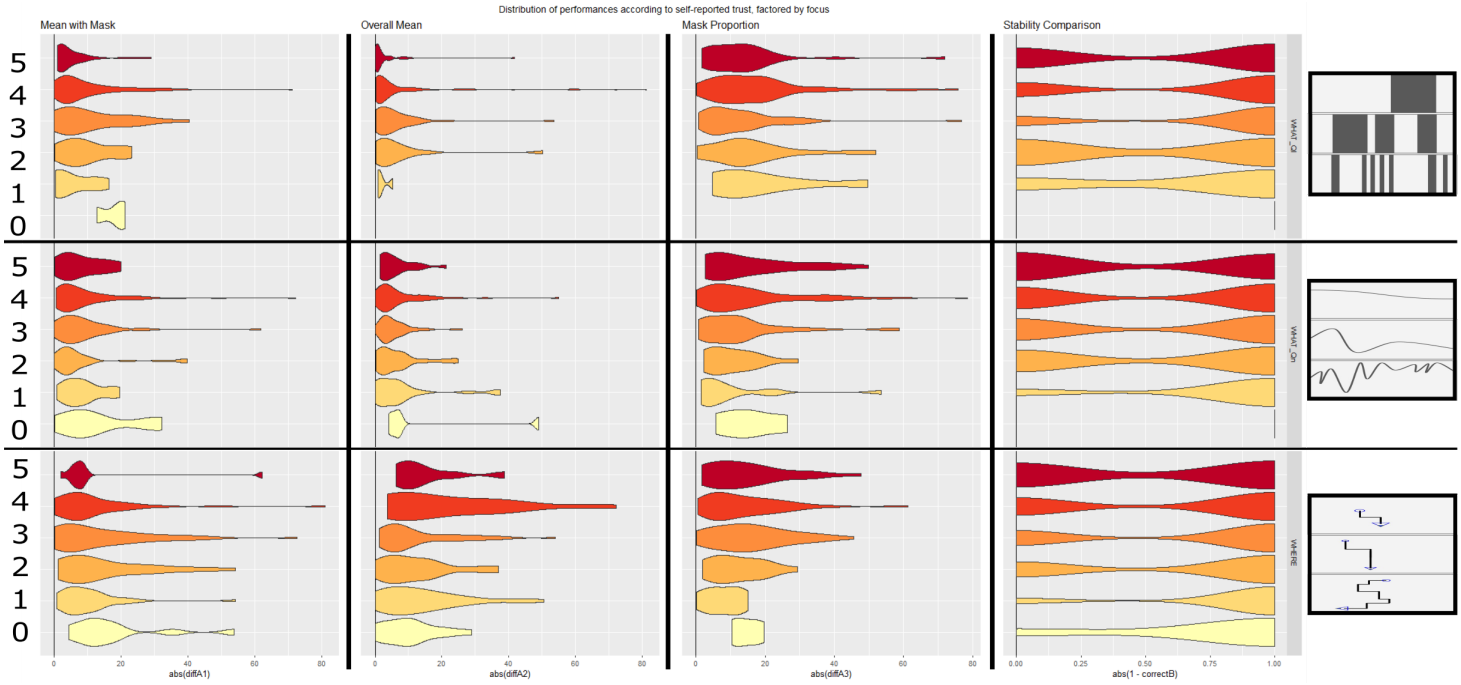


Fig. 5.53 The performances for the Measurement study according to the self-reported confidence factored by *FOCUS* . The plots are organized the same way as for Fig. 5.52.
We note that for MwM and MO , responses for WHAT_Ql show a stronger tendency to have self-reported confidence match performance. WHAT_Qn shows a similar tendency but is less strong, particularly for MwM .
Some factors result in limited ranges of performances, particularly WHAT_Ql with the lowest confidence for MO and MP . This is odd, as these are not particularly worse performances. These results may be the results of confounding factors. We discuss visual analysis of stimuli to assess potential confounding factors is 5.6.2.
Performance is consistent across all levels of self-reported confidence for WHAT_Ql, as opposed to WHAT_Qn and WHERE which display fairly strong positive correlation.
Note that absence of violin plot for a category indicates that the number of responses is inferior or equal to two, as the R program can not generate underpopulated violin plots.

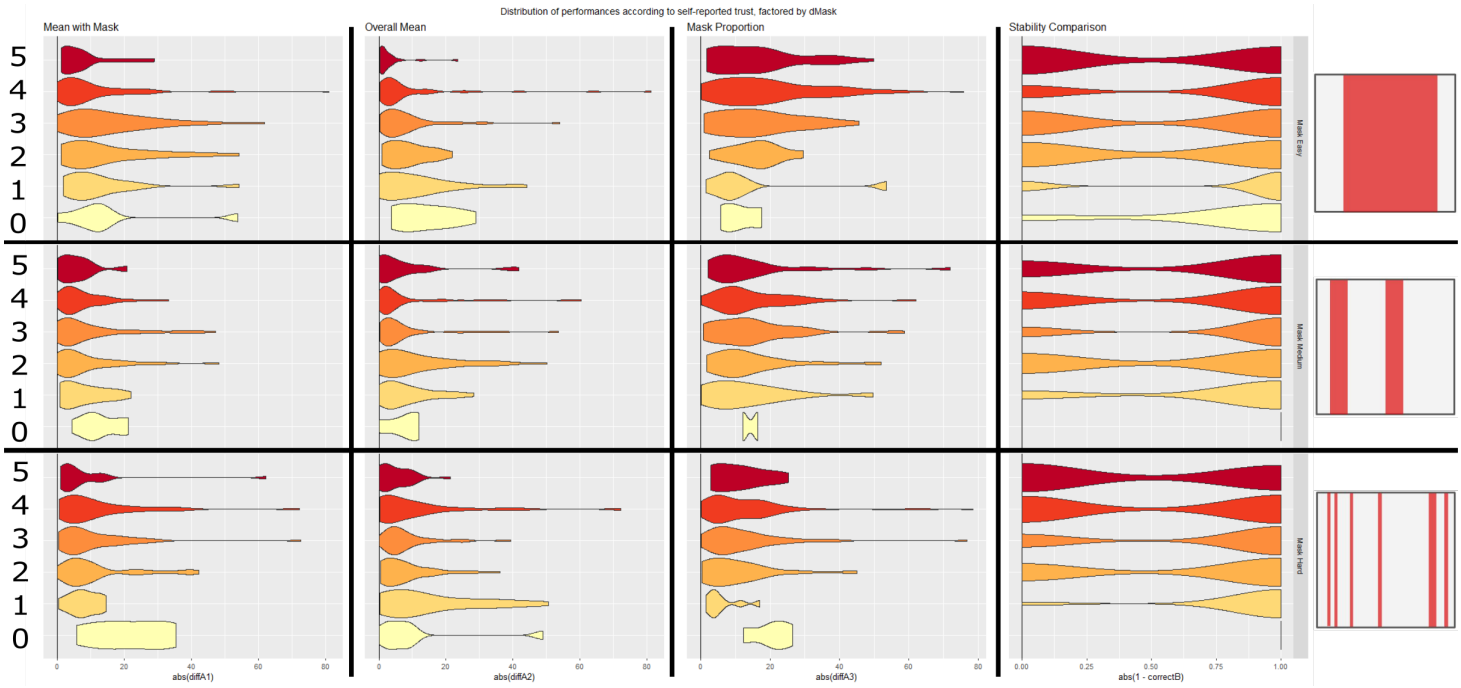


Fig. 5.54 The performances for the Measurement study according to the self-reported confidence factored by *MASK COMPLEXITY* . The plots are organized the same way as for Fig. 5.52.
Questions for MwM do not seem to show very strong trend between performance and self-reported confidence.
Questions for MO show a fairly strong trend of performances being better according to self-reported confidence. This could indicate that *MASK COMPLEXITY* has no effect on ability of participants to evaluate their ability to perform tasks not relying upon analysing them.
MP does not seem to present strong patterns for *MASK COMPLEXITY* Easy and *MASK COMPLEXITY* Hard, but we note that for MP with *MASK COMPLEXITY* Medium, the trend seems to be inverted. This contrasts with Fig. 5.38 where significant differences of performance for MP according to *MASK COMPLEXITY* followed a negative trend. These may indicate a poor understanding of the task MP .
SC shows a positive correlation between the performances and self-reported confidence for *MASK COMPLEXITY* Easy and *MASK COMPLEXITY* Hard, but not for *MASK COMPLEXITY* Medium. These responses seem to indicate that distribution of *MASK COMPLEXITY* Medium result in perceived complexity being skewed.

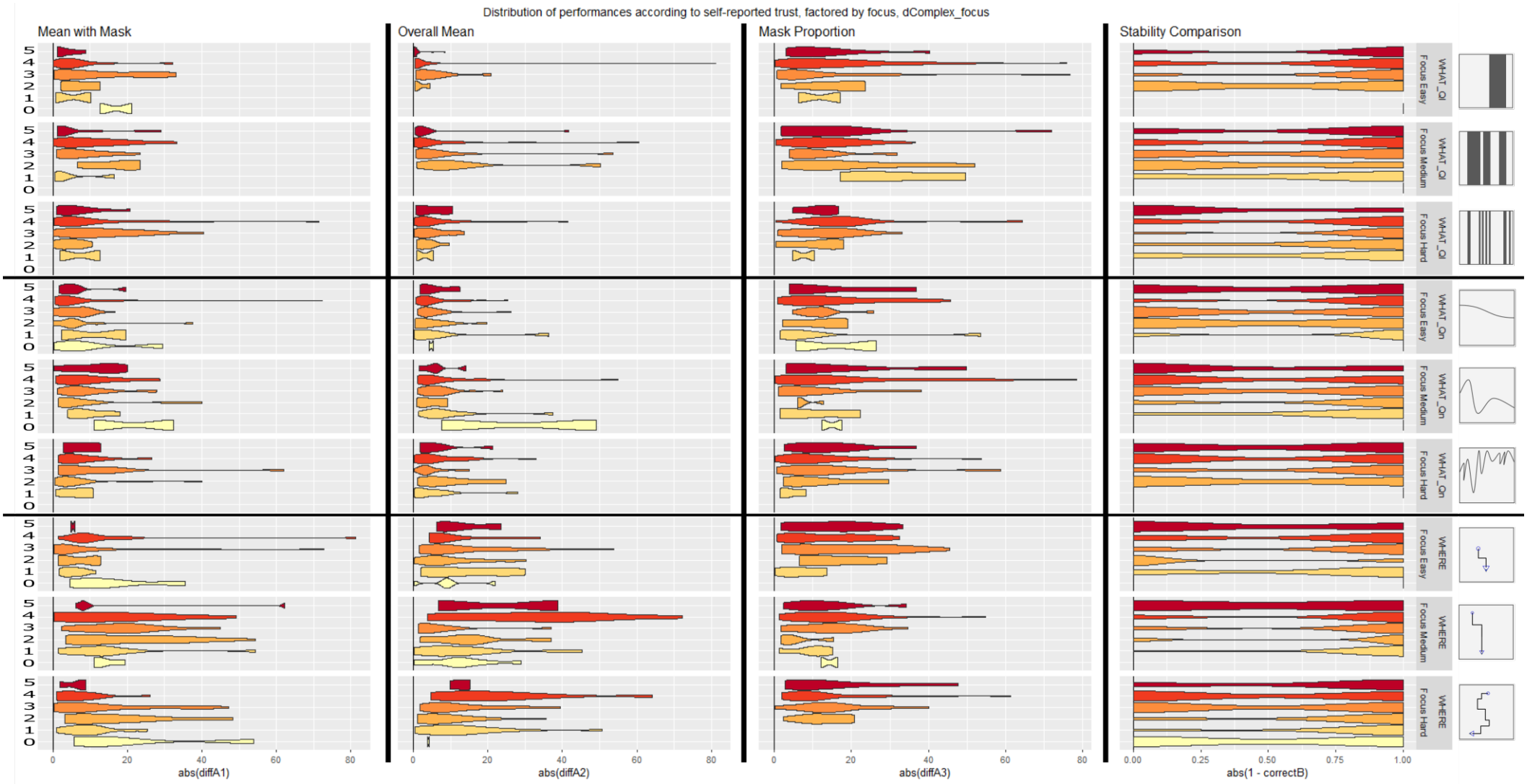


Fig. 5.55 The performances for the Measurement study according to the self-reported confidence factored by *FOCUS* and *DATA COMPLEXITY* . The plots are organized the same way as for Fig. 5.52.Note the lines instead of violin plots for self-reported trust 0 for *FOCUS* WHAT_Ql and WHERE when the *DATA COMPLEXITY* is Easy: these cases were all only composed of incorrect responses. Considering the high number of combinations, we focus our discussion on elements which we judge could lead to claims concerning the participants responses.
We note no violin plots for the trust being 0 for the questions with a *FOCUS* WHAT_Ql, and overall fewer cases of violin plots with low confidence compared to other *FOCI* . This reinforces the observations made in Fig. 5.53. We also note that participants are fairly bad at assessing their ability to answer for WHAT_Ql when the level of trust is high and the *FOCUS* complexity is Easy, which could indicate a misunderstanding of this specific combination of parameters.

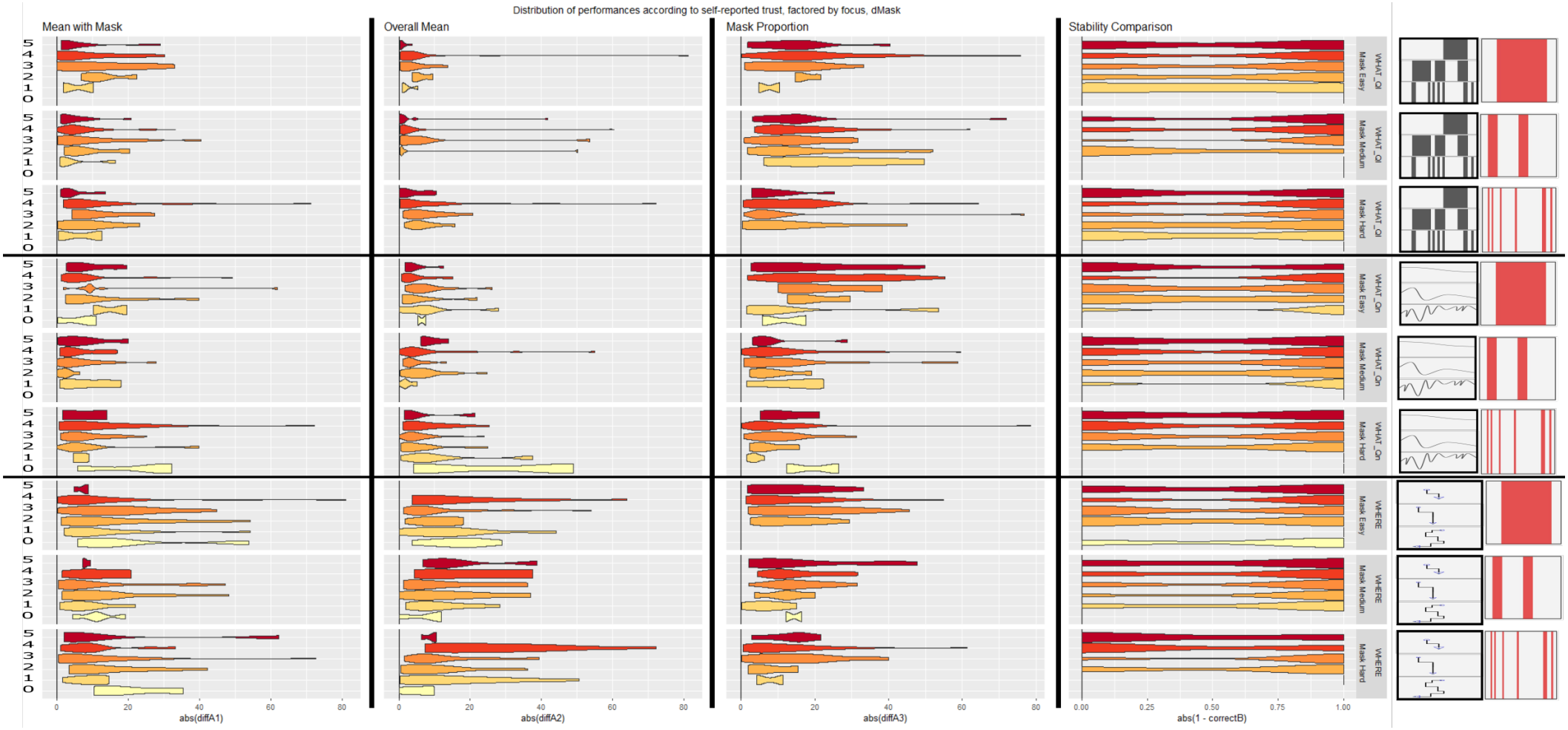


Fig. 5.56 The performances for the Measurement study according to the self-reported confidence factored by *FOCUS* and *MASK COMPLEXITY* . The plots are organized the same way as for Fig. 5.52. Note the single lines on the right of the SC graphs instead of violin plots for *FOCUS* WHAT_Ql, *MASK COMPLEXITY* Medium and Hard, *FOCUS* WHAT_Qn, *MASK COMPLEXITY* Hard and *FOCUS* WHERE, *MASK COMPLEXITY* Hard, which indicate that these factors combinations were composed only of incorrect responses.
The combination of factors allows us to see more nuances in the results, but overall we note no combination which strongly contradicts previous observations made with less factored figures 5.53 and 5.54.

The responses from participants seem to indicate that participants are fairly good at assessing their ability to perform the tasks asked of them, except MP . We estimate it is possible a confounding factor is responsible for both poorer performances when the *MASK COMPLEXITY* is Easy compared to when it is Hard, and the *MASK COMPLEXITY* being Medium resulting in the poorest ability to assess ability to perform MP . We discuss visual analysis of some stimuli to reflect upon this in section 5.6.2.

The analysis of the relationships according to the self-reported confidence allows us to make the following claims:

- Participants can relatively effectively evaluate their own ability to perform MwM and MO .
- Participants can relatively effectively evaluate their own ability to perform SC but that statement has to be nuanced, as some combinations of factors, such as WHAT_Ql Easy or *MASK COMPLEXITY* Medium, indicate the opposite trend.

Conclusion

We structured a study using SFNCS to assess the impact of *FOCUS* , *MASK COMPLEXITY* , and *DATA COMPLEXITY* on performance and self-reported confidence in four tasks that relate to values (MwM , MO , MP) and their variation (SC) established through visual interpretation of A-ATS Masks.

We noticed a strong difference in performance when the focus is WHERE for MwM , and a strong indication that participants perform best for MO when the focus is WHAT_Ql, followed by WHAT_Qn, and finally WHERE. Performances were poor, particularly for MP and SC .

As the studies were set to understand strengths and weaknesses of the A-ATS Mask in a realistic context, the questions participants were asked to answer were relatively complex, but results suggest that some participants likely misunderstood the concept of the Mask, as indicated by the inconsistency between *MASK COMPLEXITY* and performances. We tentatively discuss potential explanations behind these surprising results in section 5.6.1.

Following the analysis of the results of the study, we detail a set of claims we draw from them.

- We claim that categorization of *DATA COMPLEXITY* according to *FOCUS* is correlated to performances while performing MO , and to a lesser extent MwM for *FOCUS* WHAT_Ql and WHAT_Qn. The correlation is likely light though.
- For MwM and MO , categorizing *FOCUS* WHERE *DATA COMPLEXITY* according to the parameters we set for our studies does not match performance, as the mid-group Medium resulted in the worst performances. Considering the categorization was based on perceived complexity of trajectory, we claim tentatively that the perceived complexity of the trajectory does not correlate to performance for MwM and MO .

- Performance for MP is independent on the *FOCUS* or the *DATA COMPLEXITY* .
- MP is not strongly influenced by the *MASK COMPLEXITY* applied in the A-ATS Mask.
- SC has an error rate higher than 50% for WHAT_Ql *DATA COMPLEXITY* being Easy or Hard. This leads us to speculate that participants did not understand the stability comparison qualitative attribute according to whether time frames overlap with the status of the *MASK* .
- The error rate for SC is significantly higher when the *DATA COMPLEXITY* of the *FOCUS* WHERE is Medium.

The previous points highlighted direct insight from the result analysis. We further synthesize the information gathered from the previous claims and following reflections upon their implications, discuss what we learned through the Measurement study in these statements:

- Performances for MP not being significantly influenced by *MASK COMPLEXITY* while MO varies significantly according to *DATA COMPLEXITY* for *FOCUS* WHAT_Ql indicates different levels of understandings from participants for tasks which are fundamentally the same, i.e. assess proportion of either grey or red rectangles over a timeline. Stronger communication and training are thus necessary to consider usability of the A-ATS Mask to ensure its proper use. As stated in our literature review, similar visualization methods to the A-ATS Mask have been used efficiently, but noteworthily by experts in scientific domains, indicating scientific education could be a critical factor.
- The distributions of the responses are relatively wide for all numerical tasks, indicating diversity in participants' ability which may be due to characteristics yet unknown, e.g. experience of interacting with visualizations.
- Participants are fairly efficient at assessing their ability to correctly perform MwM , MO , MP and SC . We thus estimate that alternative approaches of categorizing visualization attributes could be assessed effectively by asking participants self-reported confidence in performing these tasks by varying attributes.
- Performances not varying according to our categorization of *MASK COMPLEXITY* indicates that other factors of the Masks are responsible for difficulty to perform MwM , MO , MP and SC .

Following these reflections, we can summarize what we have learned in order to answer our research questions:

- *Research Question 1- How does the visualization of conditions over time-space-attributes support synoptic comparative tasks within multivariate spatio-temporal data analysis?*
  Evaluating how the visualization of conditions over time-space attributes during synoptic comparative tasks within multivariate spatio-temporal data analysis allowed us to understand

more about abilities of participants to do such tasks. Variations of the parameters of the visualization method we developed to display conditions over time-space-attributes, the ATS-ATS Mask, resulted in differences of performances, but varied differently to our expectations. We previously discussed nuances behind these results, but can summarize the following answer: the display of conditions impacts performances, but further studies to evaluate additional parameters for condition displays could reinforce our understanding of which parameters are the most influential. Additionally, variations of data parameters over which the conditions are displayed impact more significantly the ability to perform the tasks.

- *Research Question 2 - How does the visualization of conditions over time-space-attributes impact self-reported trust for conducting synoptic comparative tasks within multivariate spatio-temporal data analysis?*
  Through our studies, we have learned that participants were fairly effective at evaluating their own ability to correctly perform synoptic comparative tasks within multivariate spatio-temporal data analysis with conditions of time-space-attributes overlaid. This contrasts with the relatively low performances overall. As mentioned previously, we suspect that some participants did not fully understand the ATS-ATS Mask displaying conditions. Our interpretation from these results is thus that participants were able to fairly accurately evaluate whether their understanding of the visualization was correct.

5.5 About a Judgement study

In this section, we briefly discuss plans for the Judgement study. The Judgement study is designed as a complementary study to the Measurement study, and is similar to the Measurement study, but with some of the elements defining its questions based on Judgements rather than Measurements.
An important element of the SFNCS is that it allows for the characterization of diverse elements to generate a variety of potential studies, with variations over elements which are to be evaluated, be it elements from the visual stimuli, data, question or answer. Each of these elements can be Measurements or Judgements.

The characterizations that the SFNCS allows has consequences over the possible approaches to evaluate participants' responses.
If any element of the question asked to the participant is a Judgement, it implies that the question subjectivity is overall a Judgement and thus no baseline exists.
The implication due to a lack of baselines is that the methods that can be employed to analyse the responses provided thus have to differ from quantitative studies. Questions formed with judgements are thus more appropriate to evaluate notions such as differences of perceptions between participants, self-reported trust, comfort, or impact of variations of factors on speed.
The Judgement study is designed to be strongly similar to the Measurement study, with the major difference being questions will be generated with judgement terms.
For the design of our studies, we aim to modify the Task subjectivity of the question: instead of asking for measurements, we are interested in asking for judgements. This approach can help researchers evaluate if there are shared perceptions over terms which are judgements, e.g. when is a temperature hot. We discussed this notion in detail in section 3.2.4. Solely gathering information about agreement on transformation from quantitative data to qualitative, e.g. most people in the United Kingdom agree that a temperature of 30 degrees Celsius is "hot", presents a limited interest. Potentially, fuzzy logic could be used to consider characterisations of the participants' responses to do so. But we consider that using the SFNCS to precisely characterize diverse elements (belonging to the Task, Question, Data, Visual Stimuli and Response block) could be used to assess responsible factors of participants' interpretation of the information presented to them. This enables us to, for example, assess understanding of complex relationships between attributes present in the information presented to participants.
Considering that the Judgement study could be valuable to search for elements that are responsible for participants' perception and interpretation of the information first entails two questions: which elements drive participants to evaluate a certain information relationship exists, and which elements drive participants to evaluate that a relationship is significant.
This approach implies that by varying some factors from Measurements to Judgements, the Judgement study can help researchers assess both performances of participants performing tasks, but also whether they judge the evaluated effect significant.

Further refinement is needed regarding question automation, but questions asking about either a relationship existing or if the relationship, pre-defined, is significant, could be generated using a structure similar to the following questions:

- The *attribute 1* is related to *attribute 2*.
- The variations of the *attribute 1* are much stronger while *condition* is met.

The first question structure is an example where the existence of a relationship is not declared and participants would have to determine whether it is strong enough to report it, while the second question structure is an example where the relationship is defined, but using a Judgement term to qualify it, participants have to assess whether their interpretation of the qualification matches their perception. We consider that a study following this approach could help us further understand the relationships between self-reported confidence, data displayed, and interpretations of what participants consider to be significant observations when using visualizations.

5.6 Global studies reflections

The analysis of the results have indicated that our categorization does not consistently result in a match between performances and complexity, be it for *MASK COMPLEXITY* or *DATA COMPLEXITY* . In this section, we thus discuss elements which may have influenced participants ability to perform the tasks asked of them. To do so, we visually analysed stimuli that resulted in very low performances according to different factors. We first consider whether visual elements could present confounding effects we missed during our study set up, and then discuss alternative elements which may explain differences between categorization of Data and performances of Responses.

5.6.1 Discussion

The claims made in section 5.4.2 are direct reflections from the responses. As the study is complex, multiple factors have to be considered to draw further insight from the participants performances. Within this subsection, we discuss potential explanations behind data observations:

- A high number of participants failed the introduction test (TI). As stated previously, our convenience sample when we prepared the study was composed of five participants who were not visualization experts, and none of them failed the introduction test. As we personally know the academic background of our convenience sample, we are certain that the introduction test can be passed independently of it. But an element we did not factor was that our convenience sample being composed of friends and family motivated only to help and not by financial compensation meant they fully focused on the study as they participated in the test. This is an important notion to consider because due to our efforts to present a set of stimuli that looked

realistic and similar to what could be found in a professional software, our sentences were particularly set to try to avoid different interpretations of information presented and tasks asked of participants. If we are to consider that participants on crowdsourcing platforms like Prolific try to perform the tasks as quickly as possible to perform as many studies as possible, resulting in the highest remuneration, they may consider that the time and effort to read attentively the information presented to them is not worth the effort.
The set up of a study, similar to the Measurement study, but set with sentences asking directly about visual stimuli rather than attributes, could help to indicate whether realistic studies result in poorer performances.

- As the previous point states, we ask participants complex tasks, and they may consider that they are not paid accordingly to the effort asked of them.
  Assessing the appropriate monetary incentive considering the complexity of the tasks asked of participants is not simple. Potentially, work should be set to ensure perceived fairness in pay [179], which would require to understand the perceived amount of effort necessary to perform the tasks. Furthermore, we should consider the impact of intrinsic motivations, such as enjoyment, as indicated by Kaufmann *et al.* [88] are likely more important than extrinsic motivational categories (immediate payoffs, delayed payoffs, social motivation). The tasks we set up for our studies are similar to activities done by professionals but might not be what participants are used to, implying an additional difficulty to assess correctly the appropriate amount to reward participants with.

- Our efforts to set up realistically looking stimuli might have contained confounds aimed at ecological validity, e.g. details about cars setting up expectations for participants, or background map leading participants to misinterpret the information presented to them.

- Participants also often failed the Rigorous Test (RT) (40% out of the participants who passed TI), most of them by providing higher values for MwM than MO when the focus is WHAT_Ql, which is a result that can not occur, as the answer MwM is necessarily a subsection of MO .
  This likely indicates that some participants misunderstand the 'and' of questions such as *'For how long are both the gps on and the wiper on?'* as the sum of the time frames, instead of their overlap.
  Running some studies with only focus on WHAT_Ql and different phrasing to ask MwM , MO and MP could help us to understand which sentences variations are likely to be misunderstood.

- Participants were significantly more confident when performing tasks with a Mask set to Easy, which is contradictory with their ability to perform MP according to MP . This dissonance between the performance and self-reported trust could indicate that some elements of the study were misunderstood by participants.
  It is possible that we poorly explained either the A-ATS Mask or the tasks to perform using it.

One approach to evaluate if the explanations are poor would be to re-run some studies with the same type of stimuli, factors configurations, but with minimal explanations. If performance is not poorer with minimal explanations, it would indicate that our explanations were not clear enough.

- Overall performances were fairly poor, but it is unclear whether more experience with this type of tasks would result in higher performances.
  We could assess the impact of experience for these types of tasks by running the study again with visualization experts, and compare whether the differences are significant according to the two groups.

5.6.2 Visual analysis of stimuli

Since our studies had several *FOCI* and participants performed different tasks, we made our selection of worse performances accordingly. We thus decided to select visual stimuli using the following protocol: we factor the responses by *FOCUS* , and then for questions about MwM , MO and MP , we selected for each task the three responses with the responses being the furthest from their respective baselines; for questions about SC , with the same factoring, we selected the three stimuli which resulted in the most incorrect answers, and if the number of stimuli matching that criteria was higher than three then the selection was made randomly. Following this process we visually analysed 32 stimuli (the selection was set for 36 stimuli but four of them resulted from our selections twice).

To investigate what could be the potential causes for the poorest performances, we first analysed the visual stimuli displayed when these poorest performances occurred, considering only the visual stimuli and the task asked of participants for it. The next step of our analysis process was to consider if alternative measurements could have been performed, indicating a potential misunderstanding of the task. The final consideration was to investigate whether the answer could have been provided for another question, potentially indicating the participant entering their answers without paying attention on Qualtrics. We also investigated whether some participants provided consistently the poorest responses, or if these answers were likely low effort.
In this section, we do not discuss each stimulus that resulted in low performances but rather use some of these stimuli to illustrate our reflection about potentially impactful factors to investigate in future studies or alternative design approaches.

We organize this discussion by considering notes for each *FOCUS* and reflect upon the most interesting ones.

WHAT_QI

We note for several stimuli that resulted in low performances some occurrences of Mask that either overlap or have edges in close proximity to the qualitative attribute displayed.

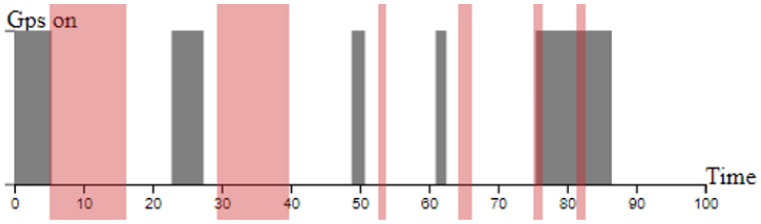


Fig. 5.57 An example of the stimuli where the distinction whether the edges of the WHAT_Ql belong to the time frames of the Mask is particularly difficult, due to the proximity of their edges.

This phenomenon occurred for 8 out of the 11 stimuli we analysed for WHAT_Ql (one of them repeatedly resulted in low performances), with the number of occurrences ranging from a single time up to eight times. While we estimate it is unlikely to influence strongly assessing the duration of either the WHAT_Ql status being positive or the Mask status being positive, it seems possible that it can affect, at least slightly, the ability to assess time frames with the WHAT_Ql status being positive and the Mask status being positive. Analyzing the stimuli, we considered that assessing status stability for WHAT_Ql could be fairly compromised by occurrences of overlap or close proximity between time frames with the status WHAT_Ql or Masks being positive. If it is unclear whether the edge of a time frame belongs or not to the time frame of the other status, assessing their proper chronology may become arduous.

Still, one out of the three stimuli had no occurrence of overlap or close proximity, and the two other stimuli having either one or two occurrences of overlap or close proximity. We thus consider that investigating this effect in future work could be valuable. One approach to evaluate this factor would be to generate an alternative design, e.g. one group with these stimuli and another with the Mask juxtaposed to the display of the WHAT_Ql without overlapping it, and compare performances of the two groups.

WHAT_Qn

Out of the 11 stimuli we analysed for WHAT_Qn, only three of them are *DATA COMPLEXITY* Hard (one of them repeatedly resulted in low performances). The display of the Mask over the WHAT_Qn does not seem to hinder in any way its perception. We thus consider that the Mask is unlikely to strongly influence ability to perform MO for WHAT_Qn.

Yet, visual analysis of the stimuli shows that some of them are overlaid with relatively narrow and closely spaced Masks. While we do not expect this to be related to difficulty to perform MwM , MO or MP , we tentatively claim that the narrow width of the Masks may result in a higher difficulty to consider how the stability of the WHAT_Qn while the narrow Masks overlay them weigh compared to the other values.

We also noticed one response where the answer for MO was actually fairly close to the range of the WHAT_Qn displayed. But since the range of WHAT_Qn was an output of the random generation over which range was only partially controlled, and not logged as a factor, it is unsafe to claim this observation to be more than anecdotal.

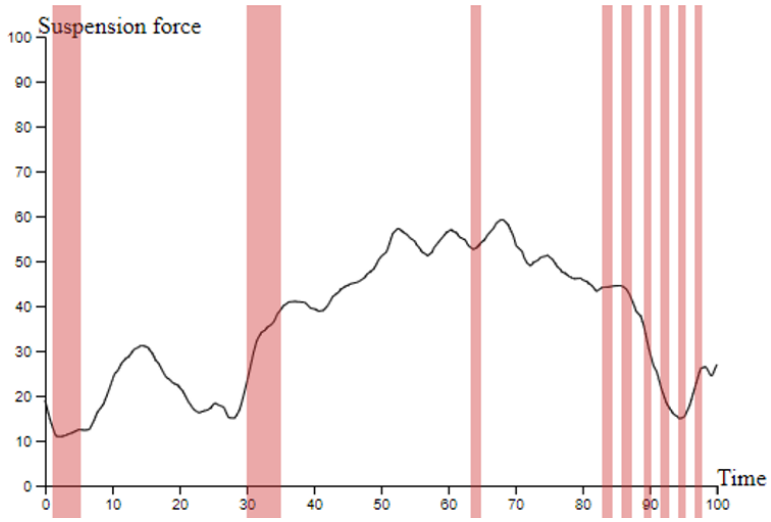


Fig. 5.58 An example of the stimuli where some of the Masks are narrow and closely spaced. This type of stimuli may result in higher difficulty to consider proportions of time ranges affected by the status of the Mask.

WHERE

The diversity in trajectories displayed in the stimuli resulting in the worst performances indicate that it is likely other factors can influence the Tasks where the *FOCUS* is WHERE. We notice that the Masks with a narrow time frame displayed are likely easier to spot when overlaid over the WHAT_Qn and WHAT_Ql attributes. We considered that this approach was likely to simplify the process of spotting sections of the trajectory recorded over time frames in which the status of the Mask was positive. Relatively bad results from participants, including for the introduction tests (TI), may indicate that this assumption was wrong, despite our best efforts to communicate that the A-ATS Mask was the display of the same condition projected over different graphs.

Additionally, the position of the point of interest (POI) was set with constraints due to the study, i.e. it had to be set in a manner that resulted in baselines for Tasks of *FOCUS* WHERE to belong in a range between 0 and 100. This resulted in POI being often fairly close to the trajectory, while that was not specifically set as our standards (POI was supposed to be within the range of latitude and longitude of the edges of the trajectories). It is possible that position of the POI for the Tasks relating to *FOCUS* WHERE influences participants' performances. We suggest that a future study evaluating this parameter as a factor could refine our understanding of its potential impact.

5.6.3 The study set up

The studies we set up resulted in a rich amount of information to analyse for which we followed approaches previously exploited [134, 70] but as the results went against our expectations, they indicated that our categorization of Data complexity did not consistently reflect on performance.

Fig. 5.59 An example of the stimuli where the POI is quite close to the trajectory, potentially resulting in a difficulty to evaluate differences between distances of points where the Mask is positive or negative.

It is interesting to reflect upon the relatively high number of participants who failed the introduction test (TI). The high number of errors indicate that we overestimated abilities of participants to correctly answer these relatively simple questions. But beyond participants' low performances, it is an opportunity to reflect upon the way to prepare participants to studies set to evaluate visualization methods. The distribution of the errors for each question made in the introduction test indicates that despite our best efforts to clearly communicate about the visualization methods, some aspects of them were perceived as more complex. The introduction questions assessing whether participants can easily spot the minimum WHAT_Qn value, and link the time of this occurrence to space resulted in the poorest responses, but neither these questions were indicative of comprehension of the graph displaying the WHAT_Ql attribute. Our approach concerning participants filtering was conservative, and thus we rejected any false set of responses provided by participants who failed any of the TI. This choice was motivated by the aim to produce valid results, but as we reflect on the study set up, we consider that broader discussions about validity of approaches to assess response quality could enrich the scientific literature.

Since the studies we set up had the objectives of both assessing the A-ATS Mask and the SFNCS in practice, being conservative was likely necessary to ensure the validity of claims would not be questioned beyond reasonable doubt. But studies are not limited to conservative approaches, e.g. Nobre *et al.* [128] who set up online studies in which participants providing wrong responses are indicated of their mistake and their system only allows participants to continue the study once they provide the correct response. Nobre *et al.* mention that participants had to correctly answer two tasks to proceed with their online studies, but did not mention participants could provide random answers until they have them right. We think that such choices should be discussed, both in publications but also between researchers concerning their validity, and the impact of such approaches on claims validity.

Concurrently, our conservative approach has resulted in rejecting a fairly high number of participants, and it is hard to assess whether the mistakes made were due to participants input mistakes, misunderstandings of the visualization presented to them, or of the phrasing of the questions. Striking a proper balance is a complicated matter, for which researchers also have to consider extra-academical factors, such as funding or ease to access participants quickly.

Additionally, during our literature review we noted that communication of studies set up in the visualization field are often concise in publications, as studies are tools to generate insight about contributions, and not themselves the focus of publications. We argue that either detailed communication of the studies set up or online access are necessary to fully understand the strength of the claims made concerning results analysis.

5.7 Chapter summary

In this chapter, we discussed the studies we ran to evaluate different aspects of the A-ATS Mask. To simplify the communication of the results, we begin by describing the commonalities between studies: across studies, participants are asked the same questions, with the same data, with variations of stimuli according to each study. We list the tasks and their corresponding questions according to focus, following the WHAT-WHERE-WHEN model of Peuquet [139].
We then detail the motivation and specific elements for the Distractor, Scaling and Measurement studies. Using the same methodology for each study, as done by Heer *et al.* [71], we analysed results using confidence interval for answers about means with mask, overall mean, or mask proportion, and confidence intervals for error rates for questions about stability comparison.
The results allowed to make certain claims on the ability of participants to perform synoptic comparative tasks within multivariate spatio-temporal data analysis within multivariate spatio-temporal data analysis. Performances were relatively poor, potentially indicating that the A-ATS Mask requires further training to ensure comprehension of the visualization method is necessary to perform complex tasks. This hypothesis is reinforced by a high proportion of participants failing a relatively simple test to verify their understanding of the visualization method used for the studies.
Still, the results of the studies allow making some interesting claims; we can claim that the display of elements not necessary to perform synoptic comparative tasks does not impact significantly performance, and that considering a minimum width (300 pixels in our set up), synoptic comparative tasks is not significantly affected by scaling.
The results of the Measurement study allows us to claim that the categorization of the quantitative and qualitative attributes correlates to the performance to synoptic comparative tasks. Additionally, the Measurement study allow us to claim that estimation of mean overall of either a quantitative, qualitative, or spatial attribute is not affected by the overlay of a Mask. We discussed how the notion of Judgement, introduced in sec. 3.2.4 could be used to generate a qualitative study generated in a similar approach to the Measurement study.

Chapter 6

Conclusion

In this thesis we developed the SFNCS that characterizes the elements that define studies that evaluate the efficacy of multivariate spatio-temporal data analysis, presented a novel method for encoding conditions over existing visualizations that are realistic in nature and used the SFNCS to conduct a systematic assessment of the performance of the proposed encoding for a range of spatio-temporal tasks with different foci and levels of complexity.

This final chapter summarizes the benefits and limitations of our contributions, discusses options for future work, and concludes with overall thoughts about the thesis.

6.1 Benefits and limitations

Our first research question was the following: **Research question 1:**
How can the information visualization reference model be expanded to define the evaluation process when generating a study to assess a novel contribution?
We address this question with the Systematic Framework for N-scales Characterizations of Studies (SFNCS): connecting Task, Question, Data, Visual Stimuli, and Response into one structure from which characteristics transfer.
Our second research question was: **Research question 2:**
How can the time mask be extended into a theoretical framework that enables filtering according to time, attributes, and space?
We address this question with the ATS-ATS Mask. The ATS-ATS Mask is a visualization method which indicates through overlays on time frames the data that matches one or more conditions.
Our third research question was the following: **Research question 3:**
How does the visualization of conditions over time-space-attributes affect people's capabilities in conducting synoptic comparative tasks within multivariate spatio-temporal data analysis?
We addressed this question with a series of studies from which we drew claims.
Our discussion of benefits and limitations reflect upon the output from our contributions.

Benefits

Characterizing the Task, Question, Data, Visual Stimuli, and Response blocks of a study The SFNCS allows characterizing blocks which are necessary to set a study. This model derived from the Card model [39] can be populated with other frameworks to fill its blocks, and thus presents an advantage to previous literature discussed in chapter 2 by enabling connections between blocks with the same characterization of the information.

Extending the time mask to allow indications through overlays of time frames for which data matches one or more conditions over visualizations displaying space, time or attributes The ATS-ATS Mask extends the time mask of Andrienko *et al.* [17, 8] by allowing to represent and communicate about queries related to space, time, or qualities over existing visualizations. We discussed queries based on time, space, and qualities, and presented one implementation of the ATS-ATS Mask with the A-ATS Mask.

Enrichment of knowledge concerning the ability to perform synoptic comparative tasks within multivariate spatio-temporal data analysis Through the evaluation of the A-ATS Mask, we contributed to further understanding about ability and confidence to perform synoptic comparative tasks within multivariate spatio-temporal data analysis. Synoptic comparative tasks within multivariate spatio-temporal data analysis were evaluated by participants assessing the following values: Mean with Mask (MwM), Mean Overall (MO), Mask Proportion (MP), and performing Stability Comparison (SC).

These studies allow us to claim that ability to perform MwM , MO , MP and SC are not affected by the presence of elements unnecessary to perform said tasks nor by the scaling of the graph displaying the quantitative attributes; participants performed relatively poorly for MwM , MO , MP and SC , but were fairly efficient at evaluating their ability to perform successfully those tasks. Performances for synoptic multivariate spatio-temporal data analysis was impacted by the complexity of the output of the query of the A-ATS Mask, but no strong systematic trend regarding their impact was noted.

Limitations

Interaction and results analysis not integrated in the SFNCS

The SFNCS does not characterize interactions allowing to modify selected Data or displayed Visual Stimuli, nor does it characterize the approach selected by researchers to draw insight from the Response. We discuss potential approaches to include Interaction and Results Analysis into the SFNCS in sections 6.2.3 and 6.2.4.

No categorization in the SFNCS

The SFNCS allows characterizing information throughout the steps to set up a study for researchers to evaluate their contributions, but does not offer standardized approaches to categorize the elements used in the studies. We presented one method to categorize Data in our studies, in section 3.2.1, but make no claim about our approach being valid for alternative studies set up. We further discuss this in section 6.2.2.

Display of the ATS-ATS Mask is not universal The ATS-ATS Mask can potentially be applied over any visualizations which displays information that relate to the query, but due to constraints created by each specific visual stimuli, we do not claim that the approach to indicate the output of the ATS-ATS Mask can always follow the same output to generate an overlay of Visual Stimuli, e.g. in our composite graphs made with the A-ATS Mask, the Mask is represented by colouring the time frames for which the condition is met, while it is represented with colouring trajectories sections for which the condition is met.

Limitations inherent to the frameworks populating the blocks The framework developed by Andrienko *et al.* [9] was envisioned with a priority over the analysis of spatio-temporal attributes. While this framework allows to characterize components of visual elements used to encode visualizations, it does lack the ability to consider potentially important notions, such as connections, particularly useful if we were to attempt to characterize node-link diagrams.
Additionally, our set of characterizations for Response was designed while considering most commonly used hardware, i.e. a keyboard and a mouse, and thus our characterization of Response does not allow incorporating Responses produced using VR [26], AR [24] technologies, or tactile interaction, e.g. on mobile devices [84].

6.2 Future work

In this section, we discuss potential future work and potential approaches to perform it. Even though these ideas are not all yet as refined as we would like them to be, we wish to discuss the potential value they could add to our current contributions.

6.2.1 Enriching the blocks of the SFNCS

As mentioned in section 6.1, the claims about the SFNCS are limited by the blocks used to build them. Replacing the blocks that compose the SFNCS is possible, and we consider it to be a valid approach, particularly when the researcher wishes to run studies for tasks with a low-level of abstraction. But we wish for the researcher to ensure that the blocks share some characterizations between blocks. Were the researchers to replace each block with the level of abstraction they wish would be valid, but would make comparison with other contributions complicated. We thus advocate for connecting elements with low levels of abstraction to their respective block with a high level of abstraction, whenever possible.
Potentially, the data block could be refined, as alternative measurements are later considered to be more fit as complexity measurements. It is likely that as research evolves, alternative complexity measurement methods are created, perhaps on the basis of improved (empirical) understanding of the relationships between measured complexity, perceived complexity and performance. The design of the SFNCS is set with blocks that can be replaced, and thus we hope that this issue will also be a fertile ground for refinement of knowledge.

Alternatively, further research could indicate that new frameworks could be better fits into the SFNCS. Whichever scenario occurs, an element we wish to be refined is the consideration of time. With the current version of the SFNCS, considering all the nuances related to time is complicated. It is particularly important as its characteristics, such as cyclic proprieties, increases the range of tasks that can be incorporated into the model, with different levels of abstraction.
The Visual stimuli block could be enriched to allow for the incorporation of connections between the elements displayed, as its current version, with the Visual Stimuli block populated by the framework developed by Andrienko *et al.* [9], presents limitations e.g. the impossibility to characterize node-link diagrams.

6.2.2 Data characterization to data categorization

For our studies we have categorized data selections according to their quantitative, qualitative, spatial attributes, and the displayed A-ATS Mask, either as "Easy", "Medium" or "Hard" based on our expectations of the impact of number of Masks over perceived complexity. This approach is helpful to simplify organizations of studies and discussions of results, but even with their best efforts and checks of their expectations, researchers can set up the categories based on false expectations. This statement has to be nuanced, as the names chosen for the categories indicate prior expectations over the influence of characteristic values used for categorisation, they fundamentally do not change the methodology to analyse the results, since these elements are not communicated to the users and analysis methodology does not vary according to these labels.
We worked towards ensuring that our categorizations are relevant, but still advocate for sharing of the data and results for any study run, so that if later research indicates alternative selections of categorizations or structures to be more relevant, new results can be produced from the previously recorded answers. Furthermore, the reusability of studies reports with different methods to characterize the data complexity presents the pragmatic advantage of reducing costs to run studies, i.e. a diverse study results population might be easier to reach if new studies re-use previously published data for results comparison.
A potential alternative would be to consider only the characteristics of the data and not categorize it. This approach implies further reflections considering evaluations of influences of characteristics over tasks, e.g. this approach is incompatible with reporting confidence intervals of differences between answers and baselines factored by categories like we do in section 5.4.

6.2.3 Characterization of interactions

An important element of the visualizations we did not discuss is the possibility to include consideration of interaction. Future work should include interactions as part of the SFNCS. Our first approach would be to consider interaction as an approach to either filter or aggregate information presented, or to vary the information display. The incorporation of these concepts in the SFNCS would likely

result in the transformation of the Visual stimuli block into an array of possible Visual stimuli the participant can access, with each Visual stimuli block being attributed a series of potential interaction, from 0 to n according to what suits it. We illustrate a potential update of the SFNCS in Fig. 6.1.
Another point to consider for future work to include interaction into the SFNCS would be how to characterize the interactions available to the users, as different sets of visualizations can potentially have their own sets of interactions.
Furthermore, we suspect that communication of interactions records are likely to make valuable contributions, but complex to communicate with potentially different sets of visual stimuli.

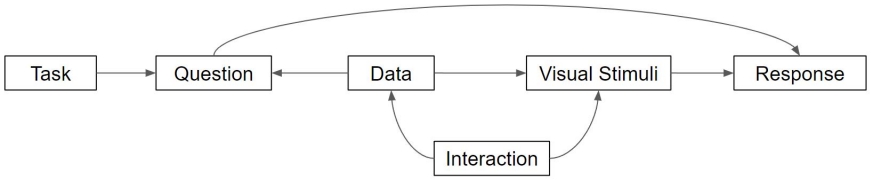


Fig. 6.1 Potential approach to consider the inclusion of characterizations of interactions for future iterations of the SFNCS. The Interaction block has arrows directed towards the Data and Visual Stimuli blocks, indicating the range of elements participants could potentially modify while participating in studies.

6.2.4 Characterization of results analysis

Selecting the methods to analyse results from studies is a step that is to be taken prior to running them. The range of potentially valid analytical methods is dependent of the type of information collected from participants - the details of the Responses block e.g. whether these are nominal, ordinal, numerical, spatial, text-based, or whether researchers are interested in their precision, accuracy while performing tasks with baselines, participants' perception, trust or understanding of the visualizations, data, and tasks asked of them. Relations between these elements and their characteristics are necessary to select analytical approaches that can be considered valid.

We consider that the SFNCS could be enhanced by a formal characterization of methods to analyse results from studies. Similarly to other blocks of the SFNCS, we consider an Analysis block, connected to the Responses one, and as other blocks of the SFNCS, it is composed of a hierarchy of blocks.

Our approach is the result of the reflections about the separation between measurement and judgement to characterize information, as discussed previously in section 3.2.4. While we consider that the separation of information is not binary but composed of nuances dependent of sum of elements that have to be considered, discussions about results analysis is often classified as quantitative or qualitative. We thus consider these two elements as characterization blocks of the highest level of abstraction to categorize approaches to analyse results.

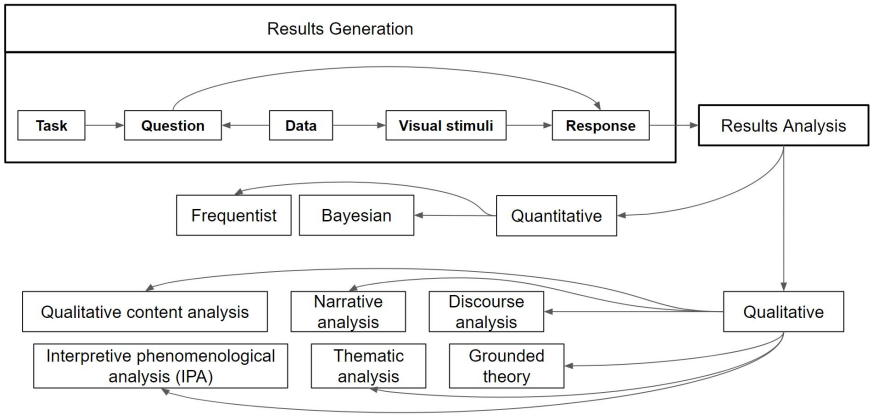


Fig. 6.2 Potential approach to characterize results analysis by extending the framework.

For quantitative studies, we considered that their characterization should be whether the result process follows a Frequentist or a Bayesian methodology, as they are the most used for results analysis [25] of quantitative results analysis. Other approaches could be considered as blocks, such as likelihood approaches [145], but as they are not as commonly used [25], we consider statisticians with a higher level of competence should be the ones to consider whether such addition would be beneficial.

There are numerous qualitative data analysis methods. Justifying a selection of the nuances of each method is beyond the scope of this thesis, but as we aimed to consider a potentially valuable approach to the enriched SFNCS, we focused on methods that are commonly used. Carrera *et al.* [40] presented a comprehensive list of Qualitative methods of data analysis in psychology. While the subject differs from ours, the methods can still be relevant in our field and in showing the flexibility of SFNCS, we list the 6 methods most used in publications as reported in their work [40] :

- Qualitative content analysis.
- Narrative analysis.
- Discourse analysis.
- Thematic analysis.
- Grounded theory.
- Interpretive phenomenological analysis (IPA).

Following the characterizations previously discussed, we illustrate a potential representation of the extended SFNCS in Fig. 6.2.

6.2.5 Future usage of the SFNCS

We hope that the SFNCS can be used in the future for the development of further studies. The details of the Judgement study are discussed in section 5.5, but within this section we wish to consider more variations on combinations of measurements and judgements. Our studies were set to only consider one moving entity, displaying one quantitative attribute, one qualitative attribute and its trajectory, with one condition over time. By increasing the number of elements presented to the user, the range of tasks that can be evaluated becomes richer. The potential combinations of different categories with different levels of subjectivity also opens up the door to interesting discussions about interpretation of measurements impact over judgement according to various sets of controlled combinations of stimuli, data, or tasks.

More categories

The taxonomy of tasks we use for the characterization of tasks, designed by Andrienko *et al.* [15], is valuable, but precise definitions of relations that can be searched for, such as patterns, could help to clarify comparison between different studies evaluating visualization designs. An interesting contribution that could be used to enrich the taxonomy is the review of research carried by Dodge *et al.* [54] which discusses classifications of movement patterns.

More entities displayed

As emphasized by the 'N' in SFNCS, the framework can be used to characterize numerous elements. The Question block is set with the Form, Identifications, and Conditions, with the latter two being potentially non-existent or a potentially infinite number of elements, as the 'N' notation commonly indicates in equations writing.
While it must be nuanced that the numbers N that can be used for the Identifications and Conditions blocks are theoretically infinite, we do expect that there will be a certain threshold beyond which usefulness will be lost, and before that a certain threshold beyond which neither clear communication nor understanding of the question can be achieved. Future studies that vary the numbers of identifications and conditions will be interesting to enrich the corpus of understanding concerning strengths and weaknesses of visualizations for the tasks assessed, but they also present an opportunity to consider formation of sentences with many elements to communicate while aiming for clarity of communication between researchers and participants doing the studies they set up.

6.2.6 Alternative designs to evaluate

The Gradual ATS-ATS Mask

As discussed in section 3.2.4,the SFNCS allows for the characterization of information as being a measurement or a judgement. We first ran studies set to ask questions about judgement over visualization depicting measurements, discussed in section 5. One limitation of the ATS-ATS Mask is that it

can only account for queries based on information which can be directly set into a query similar to a database-system, be it quantitative, qualitative, or spatial. By incorporating selections like a database query would, the ATS-ATS Mask presents a limitation: participants who interact and enter a query value wrong can hinder their analysis by disregarding data very close to matching their interest, but is rejected by the binary filter of the query. This limitation was discussed informally by the authors of the Time Mask [17]. Our studies discussed in chapter 5 were set to first evaluate the A-ATS using a binary filter and displaying the output as returned by the query.
When exploring a data set, it is likely that the precise values which have to be queried to discover new insights are unknown. We thus consider that extending queries output beyond data points fitting binary queries would greatly increase the use of the ATS-ATS Mask.
Thus, one design approach we wish to develop and assess in future work is a variation of the ATS-ATS Mask where the masks overlay vary according to a degree of query match, as opposite to a binary query match as it is currently implemented, i.e. a Mask which indicates a data point is close or far from the value queried rather than a binary match. Indicating the importance of the difference could be done with Masks with visual parameters, e.g. opacity with its value related to the difference to the query entered. Such visualization design could be named a Gradual ATS-ATS Mask. The concept of the Gradual ATS-ATS Mask would be to use adapt colour parameters according to data points matching the query, but take into account the difference between the value of the query and the one of the data point.

The Distinctive ATS-ATS Mask

Another design variation that could be interesting to consider would be alternative designs to consider the display of separate queries and integrate them into the ATS-ATS Mask. As the system currently is, it aggregates several queries into one mask. Providing unique colours for different queries was implemented in the time mask by Andrienko *et al.* [8], but that approach will not scale up beyond a number for which differentiating the unique hues would be simple. One approach we considered, during the premises of the studies set ups, was to convey time ranges belonging to spaces by using a structure similar to UpSet [102]. Icons whose design would be inspired by UpSet could be juxtaposed to time ranges accordingly to each combination of query matches. Such design approach could be named a Distinctive ATS-ATS Mask.

6.2.7 Population of results comparison and extensions

The SFNCS is designed as a tool that allows rich characterization of many elements composing the blocks Task, Question, Data, Visual Stimuli, and Response. We are aware that the relative complexity of the framework, with numerous elements that can be categorized, induces the risk to be unappealing to researchers that wish for a fast approach to consider their design ideas in respect to previous

contributions. One approach we believe could help to simplify the adoption of the SFNCS would be to generate an online software with its embedded database that could be used to query previous contributions according to selected criteria. We found inspirations in websites such as the survey from Schottler *et al.* [154] present online at the following address: https://geonetworks.github.io. One element that requires further reflection is whether the scope of the contributions should be limited, and if so, how. Limitations of the SFNCS are discussed in section 6.1 and indicate that due to the high level of abstraction of the SFNCS, the scope of contributions that can be incorporated is large, albeit modifications of the components of its block might have to be updated accordingly. We consider that, if a survey of previous contributions were to be generated and uploaded into an online queryable database, its first elements should discuss movement and event data, as the framework to characterize tasks of [9] is focused on analysing movement.
A modification that would be necessary to extend the scope of the SFNCS would be to provide a richer method to characterize visualization techniques, as the one from Andrienko *et al.* [9] is solely designed for movement and event data. Visualizations that are currently difficult to characterize include, for example, node-link diagrams and Sankey diagrams. The model from Andrienko *et al.* [9] would have to encode more information about angles of drawn elements, and be more precise on inclusion of drawn elements within others.

6.3 Conclusion

Extending Card's model [39], we developed the SFNCS and used it to characterize blocks of studies set to evaluate efficacy of multivariate spatio-temporal data analysis, populating these blocks with either blocks of previous frameworks [9, 139, 15] or our own characterizations for the novel blocks Question and Response.

The studies were set with a novel approach consisting in characterizing the information presented to the user according to its potential resulting perceived complexity. Part of the information characterization included novel distinctions of Judgement and Measurement, as part of an approach to allow different levels of abstraction to coexist in classifications and thus compare contributions with different approaches.
The work presented is discussing only a part of the scope that our contributions characterize. Our approach addresses an ensemble of gaps to simplify characterization and comparisons of contributions for studies about analysis of multivariate spatio-temporal data.

We hope that the framework presented here will contribute towards ongoing discussion of the quality, nature, effectiveness and scope for developing and using Spatio-Temporal Visualization methods.

References

[1] Agarwal, P. K., Fox, K., Munagala, K., Nath, A., Pan, J., and Taylor, E. (2018a). Subtrajectory clustering. *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - SIGMOD/PODS 18*.

[2] Agarwal, P. K., Fox, K., Munagala, K., Nath, A., Pan, J., and Taylor, E. (2018b). Subtrajectory clustering: Models and algorithms. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. ACM.

[3] Aigner, W., Miksch, S., Schumann, H., and Tominski, C. (2011). *Visualization of time-oriented data*. Springer Science & Business Media.

[4] Al-Dohuki, S., Wu, Y., Kamw, F., Yang, J., Li, X., Zhao, Y., Ye, X., Chen, W., Ma, C., and Wang, F. (2017). Semantictraj: A new approach to interacting with massive taxi trajectories. *IEEE transactions on visualization and computer graphics*, 23(1).

[5] Allain, K., Turkay, C., and Dykes, J. (2019). Towards a what-why-how taxonomy of trajectories in visualization research.

[6] Altman, D. G. and Bland, J. M. (2011). How to obtain the confidence interval from a p value. *Bmj*, 343.

[7] Amar, R., Eagan, J., and Stasko, J. (2005). Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 111–117. IEEE.

[8] Andrienko, G., Andrienko, N., Anzer, G., Bauer, P., Budziak, G., Fuchs, G., Hecker, D., Weber, H., and Wrobel, S. (2019). Constructing spaces and times for tactical analysis in football. *IEEE transactions on visualization and computer graphics*.

[9] Andrienko, G., Andrienko, N., Bak, P., Keim, D., Kisilevich, S., and Wrobel, S. (2011a). A conceptual framework and taxonomy of techniques for analyzing movement. *Journal of Visual Languages & Computing*, 22(3):213–232.

[10] Andrienko, G., Andrienko, N., and Fuchs, G. (2016a). Understanding movement data quality. *Journal of location Based services*, 10(1):31–46.

[11] Andrienko, G., Andrienko, N., Fuchs, G., and Garcia, J. M. C. (2017a). Clustering trajectories by relevant parts for air traffic analysis. *IEEE transactions on visualization and computer graphics*, 24(1):34–44.

[12] Andrienko, G., Andrienko, N., and Heurich, M. (2011b). An event-based conceptual model for context-aware movement analysis. *International Journal of Geographical Information Science*, 25(9):1347–1370.

[13] Andrienko, G., Andrienko, N., Hurter, C., Rinzivillo, S., and Wrobel, S. (2011c). From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE.

[14] Andrienko, G., Andrienko, N., Hurter, C., Rinzivillo, S., and Wrobel, S. (2012a). Scalable analysis of movement data for extracting and exploring significant places. *IEEE transactions on visualization and computer graphics*, 19(7):1078–1094.

[15] Andrienko, N. and Andrienko, G. (2006). *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media.

[16] Andrienko, N. and Andrienko, G. (2013). Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, 12(1):3–24.

[17] Andrienko, N., Andrienko, G., Camossi, E., Claramunt, C., Garcia, J. M. C., Fuchs, G., Hadzagic, M., Jousselme, A.-L., Ray, C., Scarlatti, D., et al. (2017b). Visual exploration of movement and event data with interactive time masks. *Visual Informatics*, 1(1):25–39.

[18] Andrienko, N., Andrienko, G., Garcia, J. M. C., and Scarlatti, D. (2018). Analysis of flight variability: a systematic approach. *IEEE transactions on visualization and computer graphics*, 25(1):54–64.

[19] Andrienko, N., Andrienko, G., and Gatalsky, P. (2000). Supporting visual exploration of object movement. In *Proceedings of the working conference on Advanced visual interfaces*. ACM.

[20] Andrienko, N., Andrienko, G., and Rinzivillo, S. (2016b). Leveraging spatial abstraction in traffic analysis and forecasting with visual analytics. *Information Systems*, 57:172–194.

[21] Andrienko, N., Andrienko, G., Stange, H., Liebig, T., and Hecker, D. (2012b). Visual analytics for understanding spatial situations from episodic movement data. *KI-Künstliche Intelligenz*, 26(3):241–251.

[22] Aubin, J.-P. and Frankowska, H. (2009). *Set-valued analysis*. Springer Science & Business Media.

[23] Bach, B., Dragicevic, P., Archambault, D., Hurter, C., and Carpendale, S. (2017a). A descriptive framework for temporal data visualizations based on generalized space-time cubes. In *Computer Graphics Forum*, volume 36, pages 36–61. Wiley Online Library.

[24] Bach, B., Sicat, R., Beyer, J., Cordeil, M., and Pfister, H. (2017b). The hologram in my hand: How effective is interactive exploration of 3d visualizations in immersive tangible augmented reality? *IEEE transactions on visualization and computer graphics*, 24(1):457–467.

[25] Bayarri, M. J. and Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80.

[26] Bergmann, T., Balzer, M., Hopp, T., van de Kamp, T., Kopmann, A., Jerome, N. T., and Zapf, M. (2017). Inspiration from vr gaming technology: Deep immersion and realistic interaction for scientific visualization. In *VISIGRAPP (3: IVAPP)*, pages 330–334.

[27] Bertin, J. (1983). Semiology of graphics; diagrams networks maps. Technical report.

[28] Boas, T. C., Christenson, D. P., and Glick, D. M. (2020). Recruiting large online samples in the united states and india: Facebook, mechanical turk, and qualtrics. *Political Science Research and Methods*, 8(2):232–250.

[29] Bogorny, V., Renso, C., Aquino, A. R. D., Siqueira, F. D. L., and Alvares, L. O. (2013). Constant - a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, 18(1).

[30] Bonham, C., Noyvirt, A., Tsalamanis, I., and Williams, S. (2018). Analysing port and shipping operations using big data.

[31] Brehmer, M., Lee, B., Bach, B., Riche, N. H., and Munzner, T. (2016). Timelines revisited: A design space and considerations for expressive storytelling. *IEEE transactions on visualization and computer graphics*, 23(9):2151–2164.

[32] Brehmer, M. and Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12).

[33] Brehmer, M. M. (2016). *Why visualization?: task abstraction for analysis and design*. PhD thesis, University of British Columbia.

[34] Bri Cho, J. L. (2021). Data quality at prolific - part 2: Naivety and engagement. https://blog.prolific.co/data-quality-at-prolific-part-2-naivety-and-engagement/. [Online; accessed 24-Nov-2021].

[35] Buschmann, S., Trapp, M., and Dollner, J. (2014a). Real-time animated visualization of massive air-traffic trajectories. *2014 International Conference on Cyberworlds*.

[36] Buschmann, S., Trapp, M., and Döllner, J. (2016). Animated visualization of spatial–temporal trajectory data for air-traffic analysis. *The Visual Computer*, 32(3).

[37] Buschmann, S., Trapp, M., Lühne, P., and Döllner, J. (2014b). Hardware-accelerated attribute mapping for interactive visualization of complex 3d trajectories. In *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*. IEEE.

[38] Cai, Z., Chen, M., Zhao, H., Zhao, Y., Zhou, F., and Zhang, K. (2014). Vast 2014 mini-challenge 1: Meat—multiview event analysis tool of diverse data sources. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 315–317. IEEE.

[39] Card, M. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

[40] Carrera-Fernandez, M. J., Guardia-Olmos, J., and Peró-Cebollero, M. (2014). Qualitative methods of data analysis in psychology: An analysis of the literature. *Qualitative Research*, 14(1):20–36.

[41] Carver, S., Watson, A., Waters, T., Matt, R., Gunderson, K., and Davis, B. (2009). Developing computer-based participatory approaches to mapping landscape values for landscape and resource management. In *Planning support systems best practice and new methods*, pages 431–448. Springer.

[42] Chang, and Zhou, B. (2009). Multi-granularity visualization of trajectory clusters using sub-trajectory clustering. *2009 IEEE International Conference on Data Mining Workshops*.

[43] Charalambides, N. (2021). We recently went viral on tiktok - here's what we learned. https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned/. [Online; accessed 24-Nov-2021].

[44] Chen, S., Andrienko, G. L., Andrienko, N. V., Doulkeridis, C., and Koumparos, A. (2019). Contextualized analysis of movement events. In *EuroVA@ EuroVis*, pages 49–53.

[45] Chen, W., Huang, Z., Wu, F., Zhu, M., Guan, H., and Maciejewski, R. (2018). Vaud: A visual analysis approach for exploring spatio-temporal urban data. *IEEE transactions on visualization and computer graphics*, 24(9).

[46] Chen, Y.-C., Wang, Y.-S., Lin, W.-C., Huang, W.-X., and Lin, I.-C. (2015). Interactive visual analysis for vehicle detector data. In *Computer Graphics Forum*, volume 34, pages 171–180. Wiley Online Library.

[47] Chu, D., Sheets, D. A., Zhao, Y., Wu, Y., Yang, J., Zheng, M., and Chen, G. (2014). Visualizing hidden themes of taxi movement with semantic transformation. *2014 IEEE Pacific Visualization Symposium*.

[48] Cleveland, W. S. and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554.

[49] Cook, K., Grinstein, G., and Whiting, M. (2014). The vast challenge: History, scope, and outcomes: An introduction to the special issue.

[50] Cox, J., House, D., and Lindell, M. (2013). Visualizing uncertainty in predicted hurricane tracks. *International Journal for Uncertainty Quantification*, 3(2).

[51] Cummings, M. (2017). Automation bias in intelligent time critical decision support systems. *Decision Making in Aviation*.

[52] Deitrick, S. (2013). Uncertain decisions and continuous spaces: Outcomes spaces and uncertainty visualization. In *Understanding Different Geographies*, pages 117–134. Springer.

[53] Deitrick, S. and Wentz, E. A. (2015). Developing implicit uncertainty visualization methods motivated by theories in decision science. *Annals of the Association of American Geographers*, 105(3):531–551.

[54] Dodge, S., Weibel, R., and Lautenschütz, A.-K. (2008). Towards a taxonomy of movement patterns. *Information visualization*, 7(3-4):240–252.

[55] Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252.

[56] Etikan, I., Musa, S. A., and Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1):1–4.

[57] Feldman, D. P. (1998). *Computational mechanics of classical spin systems*. PhD thesis, University of California, Davis.

[58] Ferstl, F., Kanzler, M., Rautenhaus, M., and Westermann, R. (2016). Time-hierarchical clustering and visualization of weather forecast ensembles. *IEEE transactions on visualization and computer graphics*, 23(1):831–840.

[59] Fischer, F., Stoffel, F., Mittelstädt, S., Schreck, T., and Keim, D. A. (2014). Using visual analytics to support decision making to solve the kronos incident (vast challenge 2014). In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 301–302. IEEE.

[60] Foundation, P. (2021). Perlin noise processing.

[61] Furletti, B., Cintia, P., Renso, C., and Spinsanti, L. (2013). Inferring human activities from gps tracks. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, pages 1–8.

[62] García-Constantino, M., Atkinson, K., Bollegala, D., Chapman, K., Coenen, F., Roberts, C., and Robson, K. (2017). Cliel: context-based information extraction from commercial law documents. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 79–87.

[63] Goodwin, S., Dykes, J., Slingsby, A., and Turkay, C. (2015). Visualizing multiple variables across scale and geography. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):599–608.

[64] Greis, M., Hullman, J., Correll, M., Kay, M., and Shaer, O. (2017). Designing for uncertainty in hci: When does uncertainty help? In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 593–600.

[65] Greis, M., Joshi, A., Singer, K., Schmidt, A., and Machulla, T. (2018). Uncertainty visualization influences how humans aggregate discrepant information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

[66] Groenendijk, J. A. G. and Stokhof, M. J. B. (1984). *Studies on the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, Univ. Amsterdam.

[67] Guo, H., Wang, Z., Yu, B., Zhao, H., and Yuan, X. (2011). Tripvista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection. *2011 IEEE Pacific Visualization Symposium*.

[68] Gupta, S., Dumas, M., McGuffin, M. J., and Kapler, T. (2016). Movementslicer: Better gantt charts for visualizing behaviors and meetings in movement data. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pages 168–175. IEEE.

[69] Güting, R. H., Behr, T., and Xu, J. (2010). Efficient k-nearest neighbor search on moving object trajectories. *The VLDB Journal—The International Journal on Very Large Data Bases*, 19(5).

[70] Heer, J. and Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 203–212.

[71] Heer, J., Card, S. K., and Landay, J. A. (2005). Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430.

[72] Heine, G. (2013). Euler and the flattening of the earth. *Math Horizons*, 21(1):25–29.

[73] Heuer, B., R. H. Z. J. and Maucher, J. (2012). Empirical analysis of passenger trajectories within an urban transport hub.

[74] Hintze, J. L. and Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184.

[75] Hoh, B., Gruteser, M., Xiong, H., and Alrabady, A. (2007). Preserving privacy in gps traces via uncertainty-aware path cloaking. In *Proceedings of the 14th ACM conference on Computer and communications security*, pages 161–171.

[76] Huang, X., Zhao, Y., Ma, C., Yang, J., Ye, X., and Zhang, C. (2016). Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. *IEEE transactions on visualization and computer graphics*, 22(1).

[77] Huck, J., Whyatt, J., and Coulton, P. (2014). Spraycan: A ppgis for capturing imprecise notions of place. *Applied Geography*, 55:229–237.

[78] Hurter, C., Conversy, S., Gianazza, D., and Telea, A. C. (2014). Interactive image-based information visualization for aircraft trajectory analysis. *Transportation Research Part C: Emerging Technologies*, 47:207–227.

[79] Hurter, C., Puechmorel, S., Nicol, F., and Telea, A. (2018). Functional decomposition for bundled simplification of trail sets. *IEEE transactions on visualization and computer graphics*, 24(1).

[80] Hurter, C., Tissoires, B., and Conversy, S. (2009). Fromdady: Spreading aircraft trajectories across views to support iterative queries. *IEEE Transactions on Visualization and Computer Graphics*, 15(6).

[81] Höferlin, M., Höferlin, B., Weiskopf, D., and Heidemann, G. (2011). Interactive schematic summaries for exploration of surveillance video. *Proceedings of the 1st ACM International Conference on Multimedia Retrieval - ICMR 11*.

[82] Ioannidis, J. P. (2018). The proposal to lower p value thresholds to. 005. *Jama*, 319(14):1429–1430.

[83] Jackson, C. H. (2008). Displaying uncertainty with shading. *The American Statistician*, 62(4):340–347.

[84] Jang, S., Kim, L. H., Tanner, K., Ishii, H., and Follmer, S. (2016). Haptic edge display for mobile tactile interaction. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3706–3716.

[85] Jayaram, K. and Sangeeta, K. (2017). A review: Information extraction techniques from research papers. In *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pages 56–59. IEEE.

[86] Jenny, B., Šavrič, B., and Liem, J. (2016). Real-time raster projection for web maps. *International Journal of Digital Earth*, 9(3):215–229.

[87] Karim, L., Boulmakoul, A., and Lbath, A. (2017). Real time analytics of urban congestion trajectories on hadoop-mongodb cloud ecosystem. *Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing - ICC 17*.

[88] Kaufmann, N., Schulze, T., and Veit, D. (2011). More than fun and money: Worker motivation in crowdsourcing-a study on mechanical turk.

[89] Keim, D. A., Mansmann, F., Schneidewind, J., and Ziegler, H. (2006). Challenges in visual data analysis. In *Tenth International Conference on Information Visualisation (IV'06)*, pages 9–16. IEEE.

[90] Kerracher, N. and Kennedy, J. (2017). Constructing and evaluating visualisation task classifications: Process and considerations. In *Computer graphics forum*, volume 36, pages 47–59. Wiley Online Library.

[91] Kerzner, E., Goodwin, S., Dykes, J., Jones, S., and Meyer, M. (2018). A framework for creative visualization-opportunities workshops. *IEEE transactions on visualization and computer graphics*, 25(1):748–758.

[92] Kim, D. and Kim, D.-J. (2012). Effect of screen size on multimedia vocabulary learning. *British Journal of Educational Technology*, 43(1):62–70.

[93] Kindlmann, G. and Scheidegger, C. (2014). An algebraic process for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2181–2190.

[94] Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

[95] Klein, T., Van Der Zwan, M., and Telea, A. (2014). Dynamic multiscale visualization of flight data. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 1, pages 104–114. IEEE.

[96] Knerl, L. (2021). What are typical monitor sizes and which is best?

[97] Komamizu, T., Amagasa, T., and Kitagawa, H. (2016). Visual spatial-olap for vehicle recorder data on micro-sized electric vehicles. *Proceedings of the 20th International Database Engineering Applications Symposium on - IDEAS 16*.

[98] Kosko, B. and Isaka, S. (1993). Fuzzy logic. *Scientific American*, 269(1):76–81.

[99] Krueger, R., Thom, D., and Ertl, T. (2014). Visual analysis of movement behavior using web data for context enrichment. *2014 IEEE Pacific Visualization Symposium*.

[100] Krüger, R., Simeonov, G., Beck, F., and Ertl, T. (2018). Visual interactive map matching. *IEEE transactions on visualization and computer graphics*, 24(6).

[101] Laxhammar, R. (2008). Anomaly detection for sea surveillance. In *2008 11th international conference on information fusion*. IEEE.

[102] Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). Upset: visualization of intersecting sets. *IEEE transactions on visualization and computer graphics*, 20(12):1983–1992.

[103] Li, B., Hoi, S. C., and Gopalkrishnan, V. (2011). Corn: Correlation-driven nonparametric learning approach for portfolio selection. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–29.

[104] Li, X. (2014). Using complexity measures of movement for automatically detecting movement types of unknown gps trajectories. *American Journal of Geographic Information System*, 3(2):63–74.

[105] Li, Y., Liu, R. W., Liu, J., Huang, Y., Hu, B., and Wang, K. (2016). Trajectory compression-guided visualization of spatio-temporal ais vessel density. *2016 8th International Conference on Wireless Communications Signal Processing (WCSP)*.

[106] Liebig, T., Körner, C., and May, M. (2009). Fast visual trajectory analysis using spatial bayesian networks. *2009 IEEE International Conference on Data Mining Workshops*.

[107] Liem, J., Goudarouli, E., Hirschorn, S., Wood, J., and Perin, C. (2018). Conveying uncertainty in archived war diaries with geoblobs. In *IEEE VIS 2018 Electronic Conference*.

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

[109] Liu, C., Qin, K., and Kang, C. (2015). Exploring time-dependent traffic congestion patterns from taxi trajectory data. *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*.

[110] Liu, D., Weng, D., Li, Y., Bao, J., Zheng, Y., Qu, H., and Wu, Y. (2017). Smartadp: Visual analytics of large-scale taxi trajectories for selecting billboard locations. *IEEE transactions on visualization and computer graphics*, 23(1).

[111] Liu, H., Gao, Y., Lu, L., Liu, S., Qu, H., and Ni, L. M. (2011). Visual analysis of route diversity. *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*.

[112] Lofgren, E. T. and Fefferman, N. H. (2007). The untapped potential of virtual game worlds to shed light on real world epidemics. *The Lancet infectious diseases*, 7(9):625–629.

[113] Lu, K., Chaudhuri, A., Lee, T.-Y., Shen, H.-W., and Wong, P. C. (2013). Exploring vector fields with distribution-based streamline analysis. *2013 IEEE Pacific Visualization Symposium (PacificVis)*.

[114] Lu, M., Wang, Z., and Yuan, X. (2015). Trajrank: Exploring travel behaviour on a route by trajectory ranking. *2015 IEEE Pacific Visualization Symposium (PacificVis)*.

[115] MacEachren, A. M. (1994). *Some truth with maps: A primer on symbolization and design*. Assn of Amer Geographers.

[116] MacEachren, A. M. (2004). *How maps work: representation, visualization, and design*. Guilford Press.

[117] Maître, J. (1968). Bertin (jacques) sémiologie graphique. les diagrammes. les réseaux. les cartes. *Archives de Sciences Sociales des Religions*, 26(1):176–177.

[118] Maurer, T. J. and Andrews, K. D. (2000). Traditional, likert, and simplified measures of self-efficacy. *Educational and Psychological Measurement*, 60(6):965–973.

[119] Mazimpaka, J. D. and Timpf, S. (2016). Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, (13).

[120] McCurdy, N., Gerdes, J., and Meyer, M. (2018). A framework for externalizing implicit error using visualization. *IEEE transactions on visualization and computer graphics*, 25(1):925–935.

[121] McCurdy, N., Gerdes, J., and Meyer, M. (2019). A framework for externalizing implicit error using visualization. *IEEE transactions on visualization and computer graphics*, 25(1).

[122] McNutt, A. (2021). What are table cartograms good for anyway? an algebraic analysis. *arXiv preprint arXiv:2104.04042*.

[123] Meyer, M. and Dykes, J. (2019). Criteria for rigor in visualization design study. *IEEE transactions on visualization and computer graphics*, 26(1):87–97.

[124] Microsoft (2021). Screen sizes and breakpoints. https://docs.microsoft.com/en-us/windows/apps/design/layout/screen-sizes-and-breakpoints-for-responsive-design. [Online; accessed 24-Nov-2021].

[125] Mirzargar, M., Whitaker, R. T., and Kirby, R. M. (2014). Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE transactions on visualization and computer graphics*, 20(12):2654–2663.

[126] Munzner, T. (2009). A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–928.

[127] Nergiz, M. E., Atzori, M., and Saygin, Y. (2008). Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*, pages 52–61.

[128] Nobre, C., Wootton, D., Harrison, L., and Lex, A. (2020). Evaluating multivariate network visualization techniques using a validated design and crowdsourcing approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

[129] North, C., Dwyer, T., Lee, B., Fisher, D., Isenberg, P., Robertson, G., and Inkpen, K. (2009). Understanding multi-touch manipulation for surface computing. In *IFIP Conference on Human-Computer Interaction*, pages 236–249. Springer.

[130] O'Brien, S. F. and Yi, Q. L. (2016). How do i interpret a confidence interval? *Transfusion*, 56(7):1680–1683.

[131] Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., et al. (2013). Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45(4):42.

[132] Patel, D., Bhatt, C., Hsu, W., Lee, M. L., and Kankanhalli, M. (2009). Analyzing abnormal events from spatio-temporal trajectories. *2009 IEEE International Conference on Data Mining Workshops*.

[133] Pelekis, N., Andrienko, G., Andrienko, N., Kopanakis, I., Marketos, G., and Theodoridis, Y. (2012). Visually exploring movement data via similarity-based analysis. *Journal of Intelligent Information Systems*, 38(2).

[134] Peña-Araya, V., Pietriga, E., and Bezerianos, A. (2019). A comparison of visualizations for identifying correlation over space and time. *IEEE transactions on visualization and computer graphics*, 26(1):375–385.

[135] Perin, C., Dragicevic, P., and Fekete, J.-D. (2014). Bertifier: New interactions for crafting tabular visualizations. In *IHM'14, 26e conférence francophone sur l'Interaction Homme-Machine*.

[136] Perin, C., Vuillemot, R., Stolper, C., Stasko, J., Wood, J., and Carpendale, S. (2018). State of the art of sports data visualization. In *Computer Graphics Forum*, volume 37. Wiley Online Library.

[137] Perin, C., Wun, T., Pusch, R., and Carpendale, S. (2017). Assessing the graphical perception of time and speed on 2d+ time trajectories. *IEEE transactions on visualization and computer graphics*, 24(1):698–708.

[138] Perlin, K. (1985). An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296.

[139] Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of american Geographers*, 84(3):441–461.

[140] Preim, B. and Bartz, D. (2007). *Visualization in medicine: theory, algorithms, and applications*. Elsevier.

[Press] Press, C. U. objective definition. https://dictionary.cambridge.org/dictionary/english/objective.

[142] Press, C. U. (2021). subjective definition. https://dictionary.cambridge.org/dictionary/english/subjective.

[143] Pugh, A. J., Wickens, C. D., Herdener, N., Clegg, B. A., and Smith, C. (2018). Effect of visualization training on uncertain spatial trajectory predictions. *Human Factors*, 60(3):324–339.

[144] Qian, X., Mao, J., Chen, C.-H., Chen, S., and Yang, C. (2017). Coordinated multi-aircraft 4d trajectories planning considering buffer safety distance and fuel consumption optimization via pure-strategy game. *Transportation Research Part C: Emerging Technologies*, 81:18–35.

[145] Reid, N. (2000). Likelihood. *Journal of the American Statistical Association*, 95(452):1335–1340.

[146] Sacha, D., Al-Masoudi, F., Stein, M., Schreck, T., Keim, D. A., Andrienko, G., and Janetzko, H. (2017). Dynamic visual abstraction of soccer movement. In *Computer Graphics Forum*, volume 36. Wiley Online Library.

[147] Sailer, C., Kiefer, P., Schito, J., and Raubal, M. (2016). Map-based visual analytics of moving learners. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 8(4):1–28.

[148] Santana, A. V. and Campos, J. (2016). Travel history: reconstructing semantic trajectories based on heterogeneous social tracks sources. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 311–318.

[149] Santipantakis, G. M., Vouros, G. A., Doulkeridis, C., Vlachou, A., Andrienko, G., Andrienko, N., Fuchs, G., Garcia, J. M. C., and Martinez, M. G. (2017). Specification of semantic trajectories supporting data transformations for analytics: the datacron ontology. In *Proceedings of the 13th International Conference on Semantic Systems*, pages 17–24.

[150] Scheepens, R., Willems, N., Van de Wetering, H., Andrienko, G., Andrienko, N., and Van Wijk, J. J. (2011a). Composite density maps for multivariate trajectories. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2518–2527.

[151] Scheepens, R., Willems, N., van de Wetering, H., and van Wijk, J. (2011b). Interactive density maps for moving objects. *IEEE Computer Graphics and Applications*, 32(1):56–66.

[152] Scheepens, R., Willems, N., Wetering, H. V. D., and Wijk, J. J. V. (2011c). Interactive visualization of multivariate trajectory data with density maps. *2011 IEEE Pacific Visualization Symposium*.

[153] Schiewe, J. (2018). Task-oriented visualization approaches for landscape and urban change analysis. *ISPRS International Journal of Geo-Information*, 7(8):288.

[154] Schöttler, S., Yang, Y., Pfister, H., and Bach, B. (2021). Visualizing and interacting with geospatial networks: A survey and design space. In *Computer Graphics Forum*. Wiley Online Library.

[155] Schumm, S. A. (1963). Sinuosity of alluvial rivers on the great plains. *Geological Society of America Bulletin*, 74(9):1089–1100.

[156] Sedlmair, M., Meyer, M., and Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics*, 18(12):2431–2440.

[157] Shen, Y., Zhao, L., and Fan, J. (2015). Analysis and visualization for hot spot based route recommendation using short-dated taxi gps traces. *Information*, 6(2).

[158] Sitaram, D. and Subramaniam, K. V. (2016). Complex event processing in big data systems. *Big Data Analytics*.

[159] Slingsby, A., Dykes, J., and Wood, J. (2011). Exploring uncertainty in geodemographics with interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2545–2554.

[160] Slingsby, A. and van Loon, E. (2016). Exploratory visual analysis for animal movement ecology. In *Computer Graphics Forum*, volume 35, pages 471–480. Wiley Online Library.

[161] Spretke, D., Stein, M., Sharalieva, L., Warta, A., Licht, V., Schreck, T., and Keim, D. A. (2015). Visual analysis of car fleet trajectories to find representative routes for automotive research. *2015 19th International Conference on Information Visualisation*.

[162] Sun, G., Liang, R., Qu, H., and Wu, Y. (2017). Embedding spatio-temporal information into maps by route-zooming. *IEEE transactions on visualization and computer graphics*, 23(5).

[163] Sun, G., Liu, Y., Wu, W., Liang, R., and Qu, H. (2014). Embedding temporal display into maps for occlusion-free visualization of spatio-temporal data. *2014 IEEE Pacific Visualization Symposium*.

[164] Tahir, A., Mcardle, G., and Bertolotto, M. (2012). Identifying specific spatial tasks through clustering and geovisual analysis. *2012 20th International Conference on Geoinformatics*.

[165] Vasenev, A., Pradhananga, N., Bijleveld, F., Ionita, D., Hartmann, T., Teizer, J., and Dorée, A. (2014). An information fusion approach for filtering gnss data sets collected during construction operations. *Advanced Engineering Informatics*, 28(4).

[166] Waga, K., Tabarcea, A., Mariescu-Istodor, R., and Fränti, P. (2013). Real time access to multiple gps tracks. In *WEBIST*.

[167] Wan, T. R., Chen, T., and Earnshaw, R. A. (2005). A motion constrained dynamic path planning algorithm for multi-agent simulations.

[168] Wang, Y., Lee, K., and Lee, I. (2014a). Directional higher order information for spatio-temporal trajectory dataset. *2014 IEEE International Conference on Data Mining Workshop*.

[169] Wang, Y., Lee, K., and Lee, I. (2014b). Visual analytics of topological higher order information for emergency management based on tourism trajectory datasets. *Procedia Computer Science*, 29.

[170] Wang, Z. and Yuan, X. (2014). Urban trajectory timeline visualization. *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*.

[171] Waters, T. and Evans, A. (2003). Tools for web-based gis mapping of a "fuzzy" vernacular geography. In *Proceedings of the 7th International Conference on GeoComputation*. Citeseer.

[172] Weitz, P. (2013). Determination and visualization of uncertainties in 4d-trajectory prediction. *2013 Integrated Communications, Navigation and Surveillance Conference (ICNS)*.

[173] Wesley, R., Eldridge, M., and Terlecki, P. T. (2011). An analytic data engine for visualization in tableau. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 1185–1194.

[174] Whiting, M., Cook, K., Grinstein, G., Liggett, K., Cooper, M., Fallon, J., and Morin, M. (2014). Vast challenge 2014: The kronos incident. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 295–300. IEEE.

[175] Willems, N., Van De Wetering, H., and Van Wijk, J. J. (2009). Visualization of vessel movements. In *Computer Graphics Forum*, volume 28, pages 959–966. Wiley Online Library.

[176] Wood, J., Dykes, J., and Slingsby, A. (2010). Visualisation of origins, destinations and flows with od maps. *The Cartographic Journal*, 47(2):117–129.

[177] Wu, H.-R., Yeh, M.-Y., and Chen, M.-S. (2013). Profiling moving objects by dividing and clustering trajectories spatiotemporally. *IEEE Transactions on Knowledge and Data Engineering*, 25(11).

[178] Wu, W., Zheng, Y., Cao, N., Zeng, H., Ni, B., Qu, H., and Ni, L. M. (2017). Mobiseg: Interactive region segmentation using heterogeneous mobility data. *2017 IEEE Pacific Visualization Symposium (PacificVis)*.

[179] Ye, T., You, S., and Robert Jr, L. (2017). When does more money work? examining the role of perceived fairness in pay on the performance quality of crowdworkers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

[180] Yu, Q., Luo, Y., Chen, C., and Wang, X. (2018). Trajectory outlier detection approach based on common slices sub-sequence. *Applied Intelligence*, 48(9).

[181] Zadeh, L. A. (2008). Is there a need for fuzzy logic? *Information sciences*, 178(13):2751–2779.

[182] Zeng, W., Fu, C.-W., Arisona, S. M., Schubiger, S., Burkhard, R., and Ma, K.-L. (2017). Visualizing the relationship between human mobility and points of interest. *IEEE Transactions on Intelligent Transportation Systems*, 18(8):2271–2284.

[183] Zhang, J. and Goodchild, M. F. (2002). *Uncertainty in geographical information*. CRC press.

[184] Zheng, Y., Wu, W., Chen, Y., Qu, H., and Ni, L. M. (2016). Visual analytics in urban computing: An overview. *IEEE Transactions on Big Data*, 2(3):276–296.

[185] Zhou, Y., Leung, H., and Blanchette, M. (1999). Sensor alignment with earth-centered earth-fixed (ecef) coordinate system. *IEEE Transactions on Aerospace and Electronic systems*, 35(2):410–418.

[186] Zhuang, C., Zhou, Y., Ge, J., Li, Z., Li, C., Zhou, X., and Luo, B. (2017). Information extraction from chinese judgment documents. In *2017 14th Web Information Systems and Applications Conference (WISA)*, pages 240–244. IEEE.

List of figures

List of tables

Appendix A

EuroVis 2019 Poster: Towards a WHAT-WHY-HOW Taxonomy of Trajectories inVisualization Research

Towards a WHAT-WHY-HOW Taxonomy of Trajectories in Visualization Research

K. Allain [1], C. Turkay[1] and J.Dykes[1]

[1]City, University London

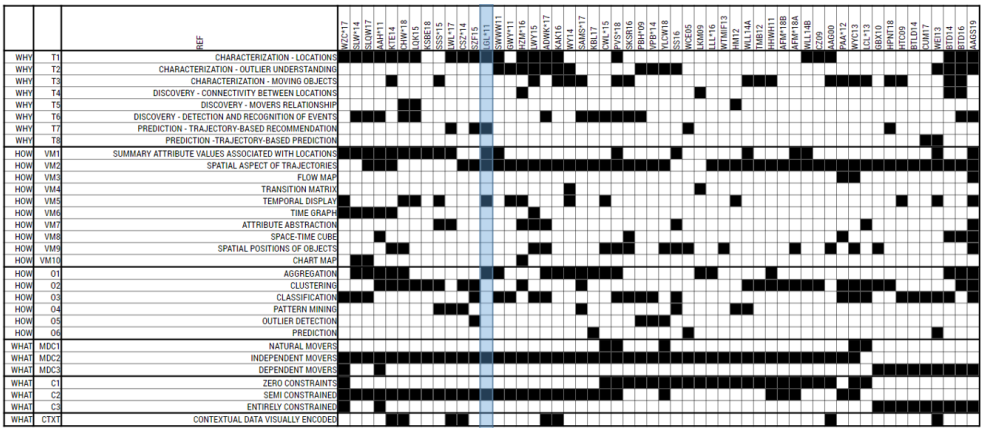


Figure 1: *Our WHAT-WHY-HOW taxonomy of trajectories visualization research illustrated using the Bertifier technique [PDF14]. The documents are ordered through Bertifier's visual similarity algorithm that makes patterns easier to discern. Find the data here:* https://bit.ly/2vyoSoQ. *The blue column indicates a document discussed as a populating example here:*https://bit.ly/3047x5l

Abstract

Effective analysis of movement often requires a comprehensive approach where computational and visual methods are combined to address a wide variety of tasks involving movers with diverse characteristics. In order to help the process of designing effective methods for a wide range of movement analysis cases, we develop a provisional taxonomy that links what Brehmer et al. [BM13] term statements of WHY-WHAT-HOW with tasks, types of movers, context and methods used to compute or visualize data. Within this document we present the origin of this taxonomy, the process we followed to populate it, discuss the novel categories within it, and finally use it to explore relationships between elements of trajectory analysis. Our main contribution is to provide a new means of connecting elements of WHY-WHAT-HOW when analysing trajectories.

1. Introduction

As capacity to collect data records of movements has increased, many methods and tools have been developed in an effort to analyse movement [AA13]. Brehmer et al. [BM13] introduced a typology that links WHY-WHAT-HOW for the analysis of a field. This model can be used to identify the data, tasks and idioms being employed. It provides "*a scaffold to think systematically about design space*" [BM13] and enables us to develop a taxonomy that may help us learn about current practice, guide analysts and designers and identify gaps for research and design. This document uses the model developed by Brehmer et al. [BM13] to link WHAT-WHY-HOW for the analysis of movement. The result is a taxonomy presented in Figure 1, with 54 documents populating it. Two existing works are relevant to our proposed taxonomy: the conceptual framework by Andrienko et al. [AAB*11] that presents a series of approaches to analyse spatio-temporal data by linking low-level analysis tasks and visualisation methods (WHY-HOW); and the taxonomy developed by Mazimpaka et al. [MT16] that details the types of operations that can be applied to support certain higher-level tasks, provides suggestions for mining methods, and to a lesser extent, visualisation methods (WHY-HOW). Both those papers link tasks and methods to analyse movement and discuss different movers

through examples but none attempt to identify mappings between tasks, methods and movers. Their work provides the baseline taxonomy that we build upon. Our scope for movers is the same as theirs, i.e. a mover with a single position, sized at a "human scale", e.g. natural phenomena, animal, urban, naval and aerial mover. In an effort to understand how types of movers and context impact trajectory analysis, we enlarged the scope of the baseline taxonomy by adding attributes of elements belonging to the WHAT aspect of our taxonomy. We first produced a taxonomy populated by papers selected following a convenience sampling [EMA16] approach in order to find categories which needed modifications. The convenience sample was useful to indicate issues with the categories of the taxonomy, but was neither systematic nor reproducible, thus we restarted populating the taxonomy, following this approach: **(1)** Search on Scopus for all conference papers and journals articles that discuss "trajector(y/ies)","visuali(s/z)ation" and exclude keywords that were representative of notions that fell out of our scope, e.g. "trajectories of eye movement". **(2)** Remove posters, short papers and VAST challenges to ensure contributions at a full paper level. **(3)** Remove the papers outside of our scope and use the remaining ones to populate the taxonomy. The categories WHY and HOW within this taxonomy are not described in detail within this work due to their lack of novelty, but further information can be found here: https://bit.ly/2V7iUWt. In this document we discuss the elements composing WHAT, reflect upon the resulting taxonomy, and conclude and discuss potential future work.

2. WHAT-WHY-HOW Taxonomy & Reflections

Analysis of visualisation is influenced by knowledge the user possesses about the elements being part of it [MGM19]. Our motivation to develop the attributes composing the elements of WHAT was based on documents mentioning their importance during the analysis of trajectories. One case is Bonham et al. [BNTW18] discussing explanations of vessels trajectories that appear counter-intuitive unless rules they have to abide to and whether they are the ones deciding the trajectories undertaken are known. Another case is Andrienko et el. [AAGS19a] who discuss flight variability and underlines the importance of context such as weather. Brehmer et al. [BM13] acknowledge that "WHAT" comprises a visualisation isn't agreed upon within the literature and advocate for a "*bring your own what*" mentality. Building upon our baseline taxonomy, we designed the following categories of WHAT:

Mover's decision capabilities (MDC): The mover's decision capabilities is a category that indicates whether the mover is the one deciding for the trajectory they follow. The mover's decision capability is useful to assess if a trajectory is an error, if the mover possessed the ability to take another trajectory, or the existence of potential interactions in between several movers.

MDC1 - Natural movers: this category describes objects where the movement will not undergo modifications due to the will or action of a sentient being, e.g. storms or glaciers. - *MDC2 - Independent movers*: Independent movers are responsible for deciding their own movement, e.g. a pedestrian or a car being driven by an occupant. - *MDC3 - Dependent movers*: Dependent movers are not making the decisions for the movement they are undergoing, e.g. a plane following direction given by an agent outside.

Levels of constraints (C): This section presents categories that define how constrained a mover is, i.e. how many rules the mover has to abide to. This notion is a continuum rather than a series of precisely defined ordered categories, and this notion can also change depending on the context, e.g. a car is semi-constrained, limited normally by legal constraints, but has access to a range of velocities and several directions. It can however be forced into a deviation in various ways: when being towed, or when under instruction by an external person, making it entirely constrained exceptionally.

C1 - Zero constraints: whereby the mover is able to go in any direction within its physical capability so to reach its destination. - *C2 - Semi constrained*: whereby the mover has sets of possibilities for trajectories, but is not free to take all of them. - *C3 - Entirely constrained*: whereby movers are unable to move in ways other than those predefined, e.g. trains are forced to move on rail roads, and are unable to deviate from the planned routes.

Contextual data (CTXT): This category is used to label documents which display contextual data different to movement, e.g. display of metadata of points of interest that can help to understand the reason behind the time a mover stops at a specific location.

Reflecting on the results: Through the development of the taxonomy we made a number of observations, some of which can be seen in the taxonomy table as well, and we reflect on some here. Tasks '*Characterisation of locations (T1)*' and '*Characterisation of moving objects (T3)*' are mainly discussed while using the visualisation method '*Spatial aspect of trajectories (Vm2)*', and most movers are '*Independent movers (MDC2)*', but there is more variety on level of constraints with the task '*Characterisation of moving objects (T3)*' which might indicate this task being researched within more domains, implying more types of movers. Additionally, most cases of '*Context (CTXT)*' are in documents discussing *(T1)*, indicating the usefulness of displaying contextual data for providing richer semantic context. Still, the over-representation of '*(MDC2)*' could indicate a lack of diversity in our population, making the emergence of strong links less likely. Furthermore, all cases of '*Entirely constrained (C3)*' are linked to '*Dependent movers (MDC3)*', potentially indicating combinations of our WHAT elements which are not separable, and thus flaws in our structure.

3. Conclusion

In this poster we introduce a taxonomy that links particular components of WHAT to a WHY-HOW structure of trajectory analysis. The taxonomy is populated with academic papers that involve the visualization of trajectories, and draws links to help navigate the design space in trajectory analysis. In the current form of the taxonomy these links are limited, likely due to particular elements of WHAT being inseparable. Our taxonomy is an in-progress framework that is open for changes to incorporate alternative WHAT structures. Potential modifications could be merging *(MDC)* and *(C)* into one dimension. This dimension would indicate how likely a mover is to follow a trajectory that appears illogical or difficult to interpret, unless provided with information about its intentions or rules it's abiding to. Additional categories could be physical characteristics of trajectories, e.g. sinuosity, granularity, and attributes of locations, and whether those are constant, e.g. elevation of area, or time-dependent, e.g. precipitation.

References

[AA13] ANDRIENKO N., ANDRIENKO G.: Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization 12*, 1 (2013). 1

[AAB*11] ANDRIENKO G., ANDRIENKO N., BAK P., KEIM D., KISILEVICH S., WROBEL S.: A conceptual framework and taxonomy of techniques for analyzing movement. *Journal of Visual Languages Computing 22*, 3 (2011). doi:10.1016/j.jvlc.2011.02.003. 1

[AAG00] ANDRIENKO N., ANDRIENKO G., GATALSKY P.: Supporting visual exploration of object movement. In *Proceedings of the working conference on Advanced visual interfaces* (2000), ACM.

[AAGS19a] ANDRIENKO N., ANDRIENKO G., GARCIA J. M. C., SCARLATTI D.: Analysis of flight variability: a systematic approach. *IEEE transactions on visualization and computer graphics 25*, 1 (2019). 2

[AAGS19b] ANDRIENKO N., ANDRIENKO G., GARCIA J. M. C., SCARLATTI D.: Analysis of flight variability: a systematic approach. *IEEE transactions on visualization and computer graphics 25*, 1 (2019).

[AAH*11] ANDRIENKO G., ANDRIENKO N., HURTER C., RINZIVILLO S., WROBEL S.: From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2011), IEEE.

[ADWK*17] AL-DOHUKI S., WU Y., KAMW F., YANG J., LI X., ZHAO Y., YE X., CHEN W., MA C., WANG F.: Semantictraj: A new approach to interacting with massive taxi trajectories. *IEEE transactions on visualization and computer graphics 23*, 1 (2017).

[AFM*18a] AGARWAL P. K., FOX K., MUNAGALA K., NATH A., PAN J., TAYLOR E.: Subtrajectory clustering. *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - SIGMOD/PODS 18* (2018). doi:10.1145/3196959.3196972.

[AFM*18b] AGARWAL P. K., FOX K., MUNAGALA K., NATH A., PAN J., TAYLOR E.: Subtrajectory clustering: Models and algorithms. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (2018), ACM.

[BM13] BREHMER M., MUNZNER T.: A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013). doi:10.1109/tvcg.2013.124. 1, 2

[BNTW18] BONHAM C., NOYVIRT A., TSALAMANIS I., WILLIAMS S.: Analysing port and shipping operations using big data. 2

[BRA*13] BOGORNY V., RENSO C., AQUINO A. R. D., SIQUEIRA F. D. L., ALVARES L. O.: Constant - a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS 18*, 1 (2013). doi:10.1111/tgis.12011.

[BRSW06] BOMBERGER N., RHODES B., SEIBERT M., WAXMAN A.: Associative learning of vessel motion patterns for maritime situation awareness. *2006 9th International Conference on Information Fusion* (2006). doi:10.1109/icif.2006.301661.

[BTD14] BUSCHMANN S., TRAPP M., DOLLNER J.: Real-time animated visualization of massive air-traffic trajectories. *2014 International Conference on Cyberworlds* (2014). doi:10.1109/cw.2014.32.

[BTD16] BUSCHMANN S., TRAPP M., DÖLLNER J.: Animated visualization of spatial–temporal trajectory data for air-traffic analysis. *The Visual Computer 32*, 3 (2016).

[BTLD14] BUSCHMANN S., TRAPP M., LÜHNE P., DÖLLNER J.: Hardware-accelerated attribute mapping for interactive visualization of complex 3d trajectories. In *2014 International Conference on Information Visualization Theory and Applications (IVAPP)* (2014), IEEE.

[CHW*18] CHEN W., HUANG Z., WU F., ZHU M., GUAN H., MACIEJEWSKI R.: Vaud: A visual analysis approach for exploring spatio-temporal urban data. *IEEE transactions on visualization and computer graphics 24*, 9 (2018).

[CSZ*14] CHU D., SHEETS D. A., ZHAO Y., WU Y., YANG J., ZHENG M., CHEN G.: Visualizing hidden themes of taxi movement with semantic transformation. *2014 IEEE Pacific Visualization Symposium* (2014). doi:10.1109/pacificvis.2014.50.

[Cum17] CUMMINGS M.: Automation bias in intelligent time critical decision support systems. *Decision Making in Aviation* (2017). doi:10.4324/9781315095080-17.

[CWL*15] CHEN Y.-C., WANG Y.-S., LIN W.-C., HUANG W.-X., LIN I.-C.: Interactive visual analysis for vehicle detector data. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library.

[CZ09] CHANG Â., ZHOU B.: Multi-granularity visualization of trajectory clusters using sub-trajectory clustering. *2009 IEEE International Conference on Data Mining Workshops* (2009). doi:10.1109/icdmw.2009.24.

[EMA16] ETIKAN I., MUSA S. A., ALKASSIM R. S.: Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics 5*, 1 (2016), 1–4. 2

[GBX10] GÜTING R. H., BEHR T., XU J.: Efficient k-nearest neighbor search on moving object trajectories. *The VLDB JournalâĂŤThe International Journal on Very Large Data Bases 19*, 5 (2010).

[GWY*11] GUO H., WANG Z., YU B., ZHAO H., YUAN X.: Tripvista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection. *2011 IEEE Pacific Visualization Symposium* (2011). doi:10.1109/pacificvis.2011.5742386.

[HHBR15] HORNAUER S., HAHN A., BLAICH M., REUTER J.: Trajectory planning with negotiation for maritime collision avoidance. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation 9*, 3 (2015). doi:10.12716/1001.09.03.05.

[HHWH11] HÃŰFERLIN M., HÃŰFERLIN B., WEISKOPF D., HEIDEMANN G.: Interactive schematic summaries for exploration of surveillance video. *Proceedings of the 1st ACM International Conference on Multimedia Retrieval - ICMR 11* (2011). doi:10.1145/1991996.1992005.

[HM12] HEUER B. R. H. Z. J., MAUCHER J.: Empirical analysis of passenger trajectories within an urban transport hub.

[HPNT18] HURTER C., PUECHMOREL S., NICOL F., TELEA A.: Functional decomposition for bundled simplification of trail sets. *IEEE transactions on visualization and computer graphics 24*, 1 (2018).

[HTC09] HURTER C., TISSOIRES B., CONVERSY S.: Fromdady: Spreading aircraft trajectories across views to support iterative queries. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009). doi:10.1109/tvcg.2009.145.

[HZM*16] HUANG X., ZHAO Y., MA C., YANG J., YE X., ZHANG C.: Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. *IEEE transactions on visualization and computer graphics 22*, 1 (2016).

[KAK16] KOMAMIZU T., AMAGASA T., KITAGAWA H.: Visual spatial-olap for vehicle recorder data on micro-sized electric vehicles. *Proceedings of the 20th International Database Engineering Applications Symposium on - IDEAS 16* (2016). doi:10.1145/2938503.2938532.

[KBL17] KARIM L., BOULMAKOUL A., LBATH A.: Real time analytics of urban congestion trajectories on hadoop-mongodb cloud ecosystem. *Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing - ICC 17* (2017). doi:10.1145/3018896.3018923.

[KSBE18] KRÜGER R., SIMEONOV G., BECK F., ERTL T.: Visual interactive map matching. *IEEE transactions on visualization and computer graphics 24*, 6 (2018).

[KTE14] KRUEGER R., THOM D., ERTL T.: Visual analysis of movement behavior using web data for context enrichment. *2014 IEEE Pacific Visualization Symposium* (2014). doi:10.1109/pacificvis.2014.57.

[Lax08] LAXHAMMAR R.: Anomaly detection for sea surveillance. In *2008 11th international conference on information fusion* (2008), IEEE.

[LCL*13] LU K., CHAUDHURI A., LEE T.-Y., SHEN H.-W., WONG P. C.: Exploring vector fields with distribution-based streamline analysis. *2013 IEEE Pacific Visualization Symposium (PacificVis)* (2013). doi:10.1109/pacificvis.2013.6596153.

[LGL*11] LIU H., GAO Y., LU L., LIU S., QU H., NI L. M.: Visual analysis of route diversity. *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2011). doi:10.1109/vast.2011.6102455.

[LKM09] LIEBIG T., KÃŰRNER C., MAY M.: Fast visual trajectory analysis using spatial bayesian networks. *2009 IEEE International Conference on Data Mining Workshops* (2009). doi:10.1109/icdmw.2009.44.

[LLL*16] LI Y., LIU R. W., LIU J., HUANG Y., HU B., WANG K.: Trajectory compression-guided visualization of spatio-temporal ais vessel density. *2016 8th International Conference on Wireless Communications Signal Processing (WCSP)* (2016). doi:10.1109/wcsp.2016.7752733.

[LQK15] LIU C., QIN K., KANG C.: Exploring time-dependent traffic congestion patterns from taxi trajectory data. *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)* (2015). doi:10.1109/icsdm.2015.7298022.

[LWL*17] LIU D., WENG D., LI Y., BAO J., ZHENG Y., QU H., WU Y.: Smartadp: Visual analytics of large-scale taxi trajectories for selecting billboard locations. *IEEE transactions on visualization and computer graphics 23*, 1 (2017).

[LWY15] LU M., WANG Z., YUAN X.: Trajrank: Exploring travel behaviour on a route by trajectory ranking. *2015 IEEE Pacific Visualization Symposium (PacificVis)* (2015). doi:10.1109/pacificvis.2015.7156392.

[MGM19] MCCURDY N., GERDES J., MEYER M.: A framework for externalizing implicit error using visualization. *IEEE transactions on visualization and computer graphics 25*, 1 (2019). 2

[MT16] MAZIMPAKA J. D., TIMPF S.: Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, 13 (2016). doi:10.5311/josis.2016.13.263. 1

[PAA*12] PELEKIS N., ANDRIENKO G., ANDRIENKO N., KOPANAKIS I., MARKETOS G., THEODORIDIS Y.: Visually exploring movement data via similarity-based analysis. *Journal of Intelligent Information Systems 38*, 2 (2012).

[PBH*09] PATEL D., BHATT C., HSU W., LEE M. L., KANKANHALLI M.: Analyzing abnormal events from spatio-temporal trajectories. *2009 IEEE International Conference on Data Mining Workshops* (2009). doi:10.1109/icdmw.2009.45.

[PDF14] PERIN C., DRAGICEVIC P., FEKETE J.-D.: Bertifier: New interactions for crafting tabular visualizations. In *IHM'14, 26e conférence francophone sur l'Interaction Homme-Machine* (2014). 1

[PVS*18] PERIN C., VUILLEMOT R., STOLPER C., STASKO J., WOOD J., CARPENDALE S.: State of the art of sports data visualization. In *Computer Graphics Forum* (2018), vol. 37, Wiley Online Library.

[SAMS*17] SACHA D., AL-MASOUDI F., STEIN M., SCHRECK T., KEIM D. A., ANDRIENKO G., JANETZKO H.: Dynamic visual abstraction of soccer movement. In *Computer Graphics Forum* (2017), vol. 36, Wiley Online Library.

[SKSR16] SAILER C., KIEFER P., SCHITO J., RAUBAL M.: Map-based visual analytics of moving learners. *International Journal of Mobile Human Computer Interaction (IJMHCI) 8*, 4 (2016).

[SLQW17] SUN G., LIANG R., QU H., WU Y.: Embedding spatio-temporal information into maps by route-zooming. *IEEE transactions on visualization and computer graphics 23*, 5 (2017).

[SLW*14] SUN G., LIU Y., WU W., LIANG R., QU H.: Embedding temporal display into maps for occlusion-free visualization of spatio-temporal data. *2014 IEEE Pacific Visualization Symposium* (2014). doi:10.1109/pacificvis.2014.56.

[SS16] SITARAM D., SUBRAMANIAM K. V.: Complex event processing in big data systems. *Big Data Analytics* (2016). doi:10.1007/978-81-322-3628-3_8.

[SSS*15] SPRETKE D., STEIN M., SHARALIEVA L., WARTA A., LICHT V., SCHRECK T., KEIM D. A.: Visual analysis of car fleet trajectories to find representative routes for automotive research. *2015 19th International Conference on Information Visualisation* (2015). doi:10.1109/iv.2015.63.

[SWWW11] SCHEEPENS R., WILLEMS N., WETERING H. V. D., WIJK J. J. V.: Interactive visualization of multivariate trajectory data with density maps. *2011 IEEE Pacific Visualization Symposium* (2011). doi:10.1109/pacificvis.2011.5742384.

[SZF15] SHEN Y., ZHAO L., FAN J.: Analysis and visualization for hot spot based route recommendation using short-dated taxi gps traces. *Information 6*, 2 (2015).

[TMB12] TAHIR A., MCARDLE G., BERTOLOTTO M.: Identifying specific spatial tasks through clustering and geovisual analysis. *2012 20th International Conference on Geoinformatics* (2012). doi:10.1109/geoinformatics.2012.6270301.

[VPB*14] VASENEV A., PRADHANANGA N., BIJLEVELD F., IONITA D., HARTMANN T., TEIZER J., DORÃL'E A.: An information fusion approach for filtering gnss data sets collected during construction operations. *Advanced Engineering Informatics 28*, 4 (2014). doi:10.1016/j.aei.2014.07.001.

[WCE05] WAN T. R., CHEN T., EARNSHAW R. A.: A motion constrained dynamic path planning algorithm for multi-agent simulations.

[Wei13] WEITZ P.: Determination and visualization of uncertainties in 4d-trajectory prediction. *2013 Integrated Communications, Navigation and Surveillance Conference (ICNS)* (2013). doi:10.1109/icnsurv.2013.6548525.

[WLL14a] WANG Y., LEE K., LEE I.: Directional higher order information for spatio-temporal trajectory dataset. *2014 IEEE International Conference on Data Mining Workshop* (2014). doi:10.1109/icdmw.2014.48.

[WLL14b] WANG Y., LEE K., LEE I.: Visual analytics of topological higher order information for emergency management based on tourism trajectory datasets. *Procedia Computer Science 29* (2014).

[WTMIF13] WAGA K., TABARCEA A., MARIESCU-ISTODOR R., FRÄNTI P.: Real time access to multiple gps tracks. In *WEBIST* (2013).

[WY14] WANG Z., YUAN X.: Urban trajectory timeline visualization. *2014 International Conference on Big Data and Smart Computing (BIGCOMP)* (2014). doi:10.1109/bigcomp.2014.6741397.

[WYC13] WU H.-R., YEH M.-Y., CHEN M.-S.: Profiling moving objects by dividing and clustering trajectories spatiotemporally. *IEEE Transactions on Knowledge and Data Engineering 25*, 11 (2013).

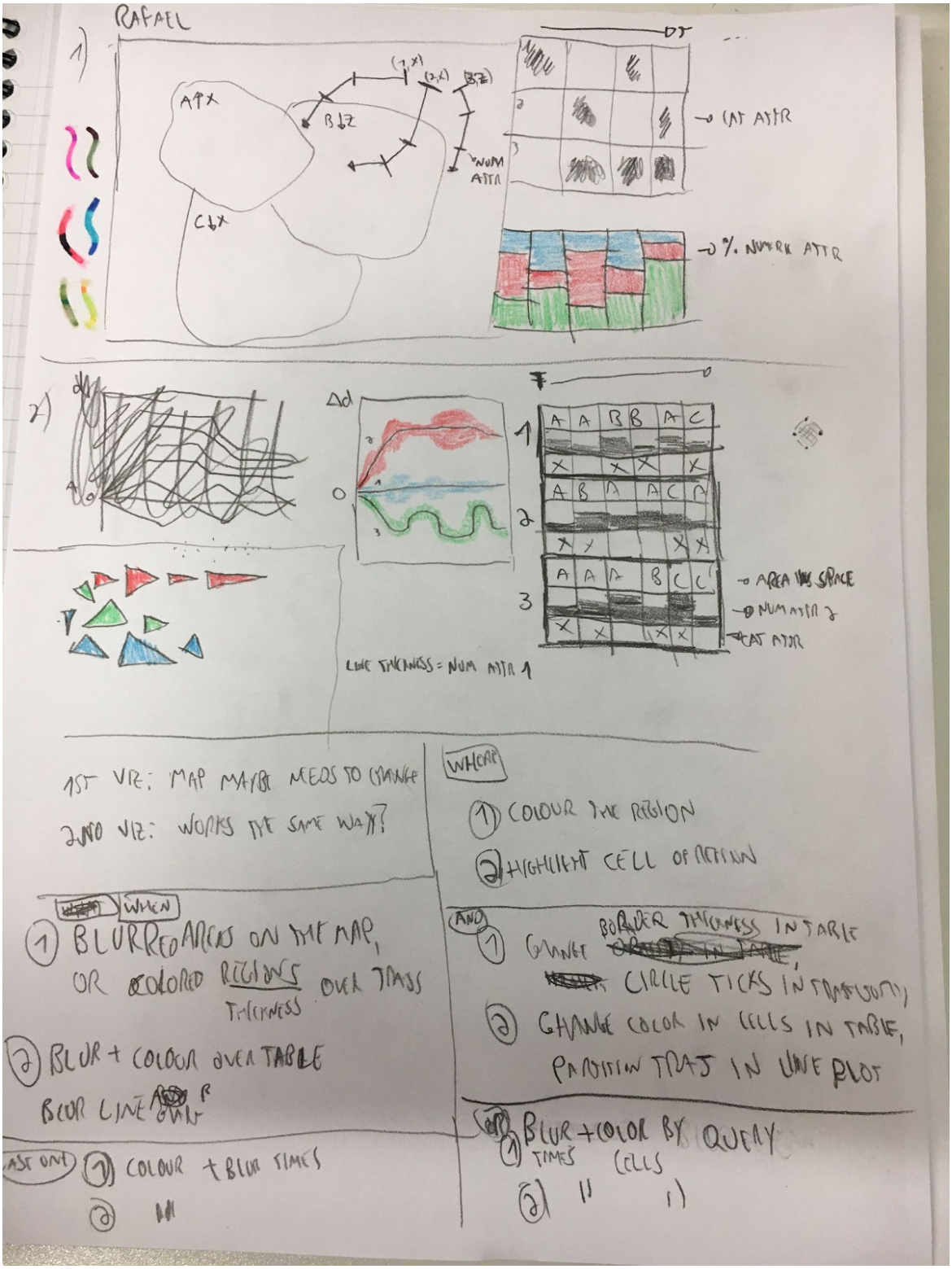
[WZC*17] WU W., ZHENG Y., CAO N., ZENG H., NI B., QU H., NI L. M.: Mobiseg: Interactive region segmentation using heterogeneous mobility data. *2017 IEEE Pacific Visualization Symposium (PacificVis)* (2017). doi:10.1109/pacificvis.2017.8031583.

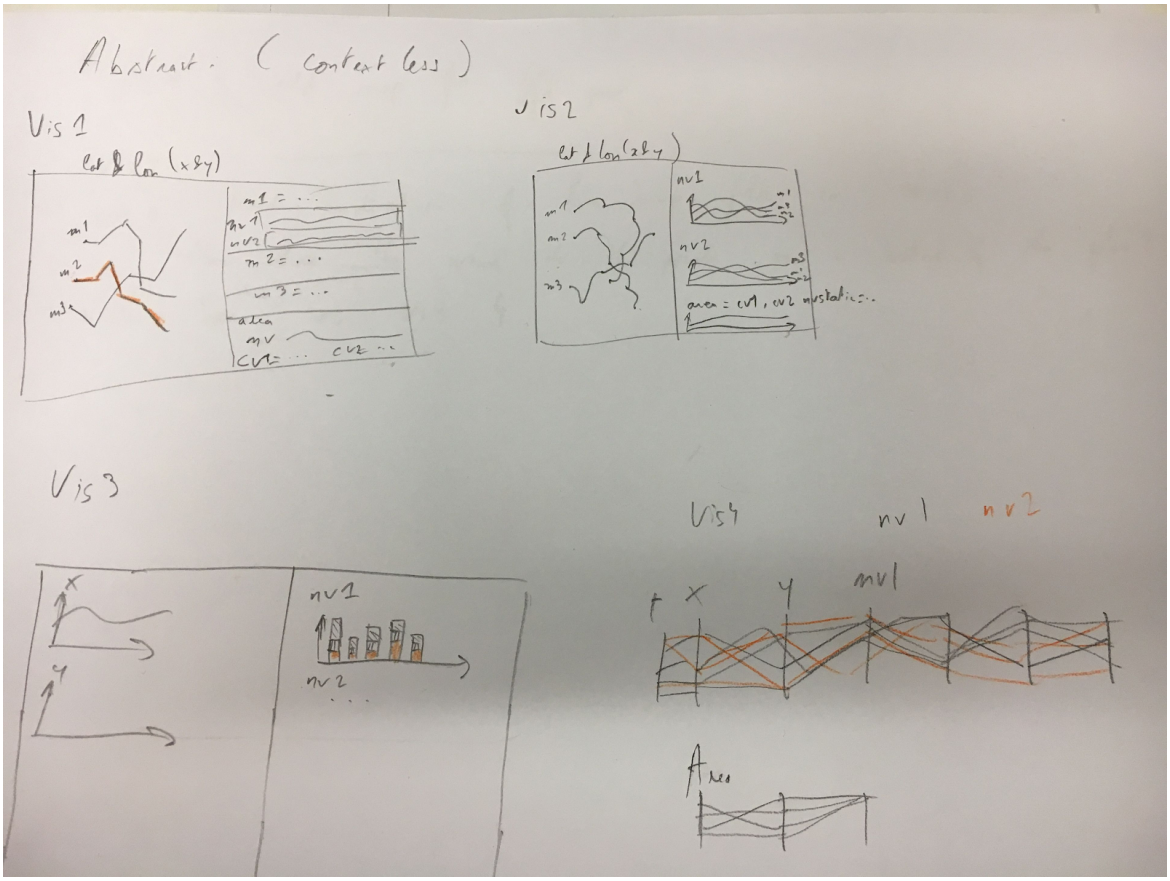
[YLCW18] YU Q., LUO Y., CHEN C., WANG X.: Trajectory outlier detection approach based on common slices sub-sequence. *Applied Intelligence 48*, 9 (2018).

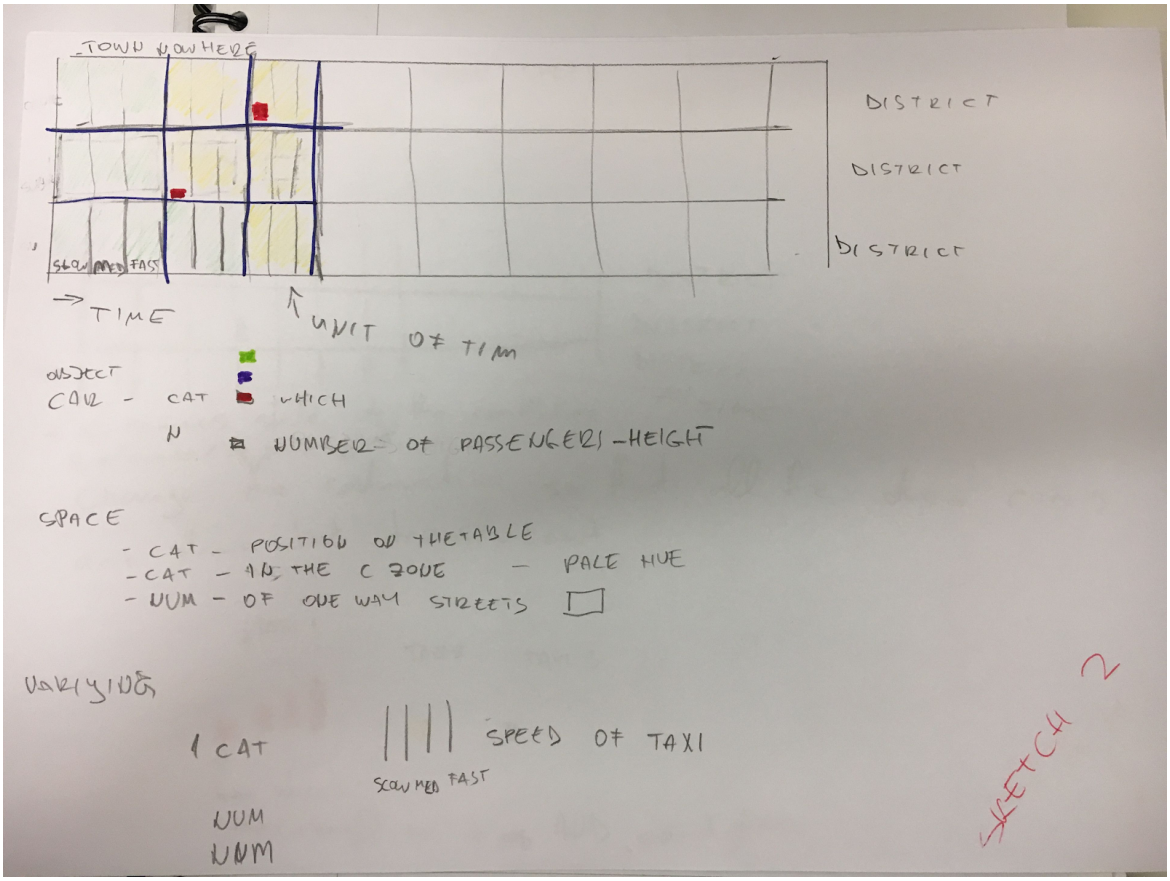
Appendix B

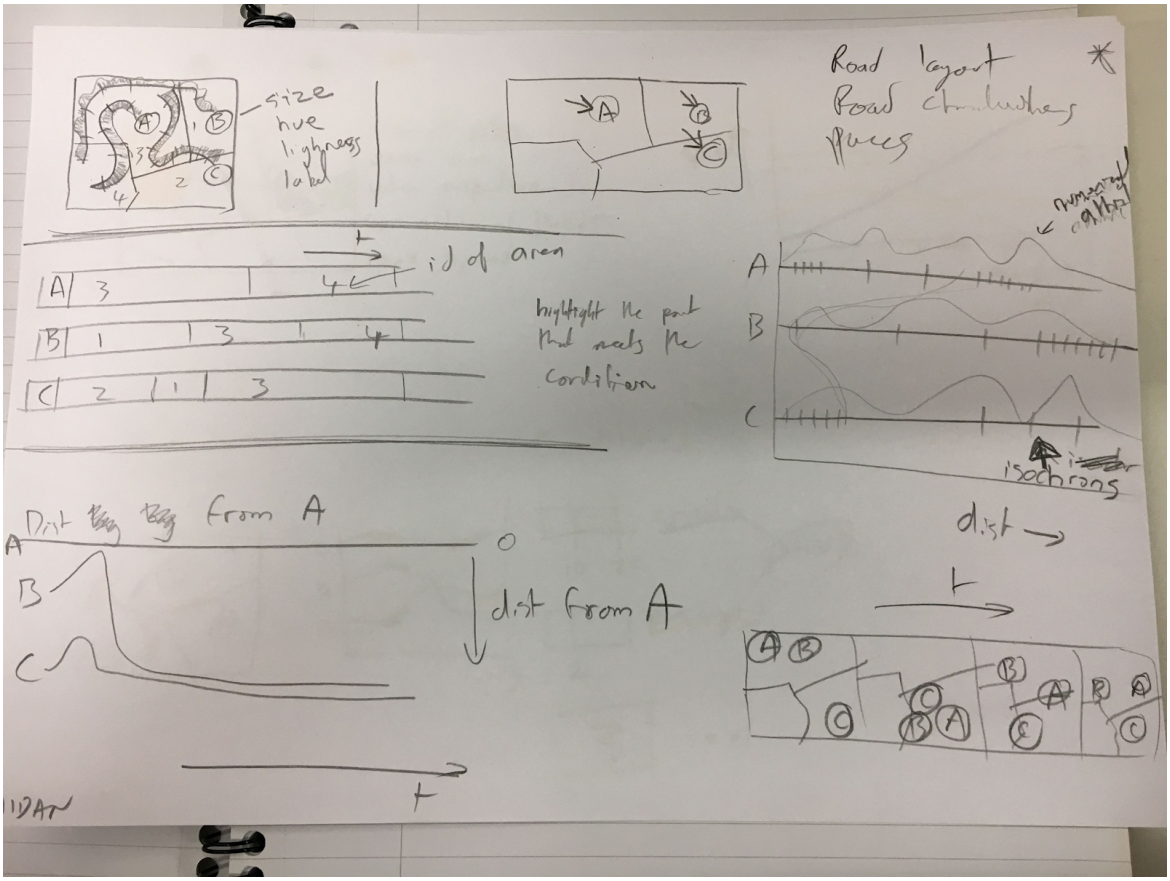
The design workshop

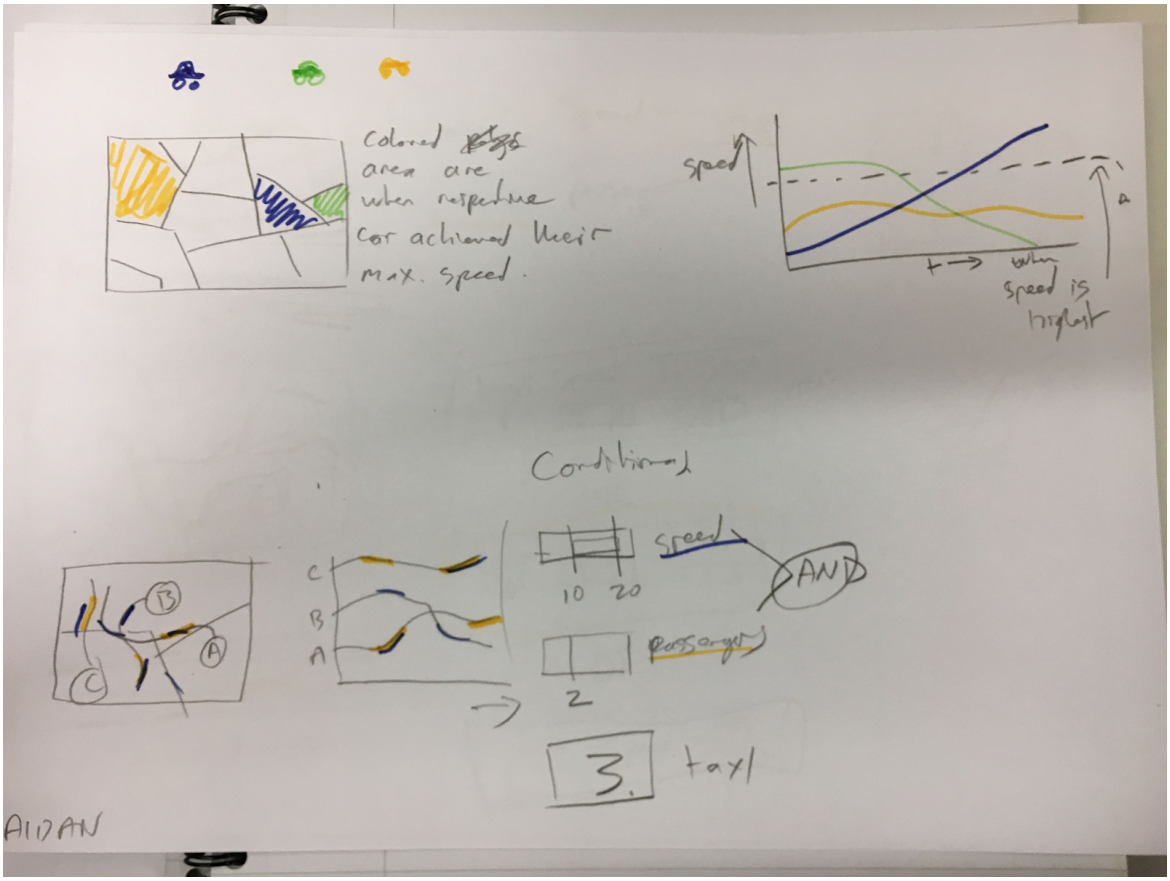
This appendix is set to illustrate the drafts and notes taken by participants of our workshop to investigate visualization designs appropriate for the ATS-ATS Mask. Discussions about our interpretation of those drafts is present in section 4.

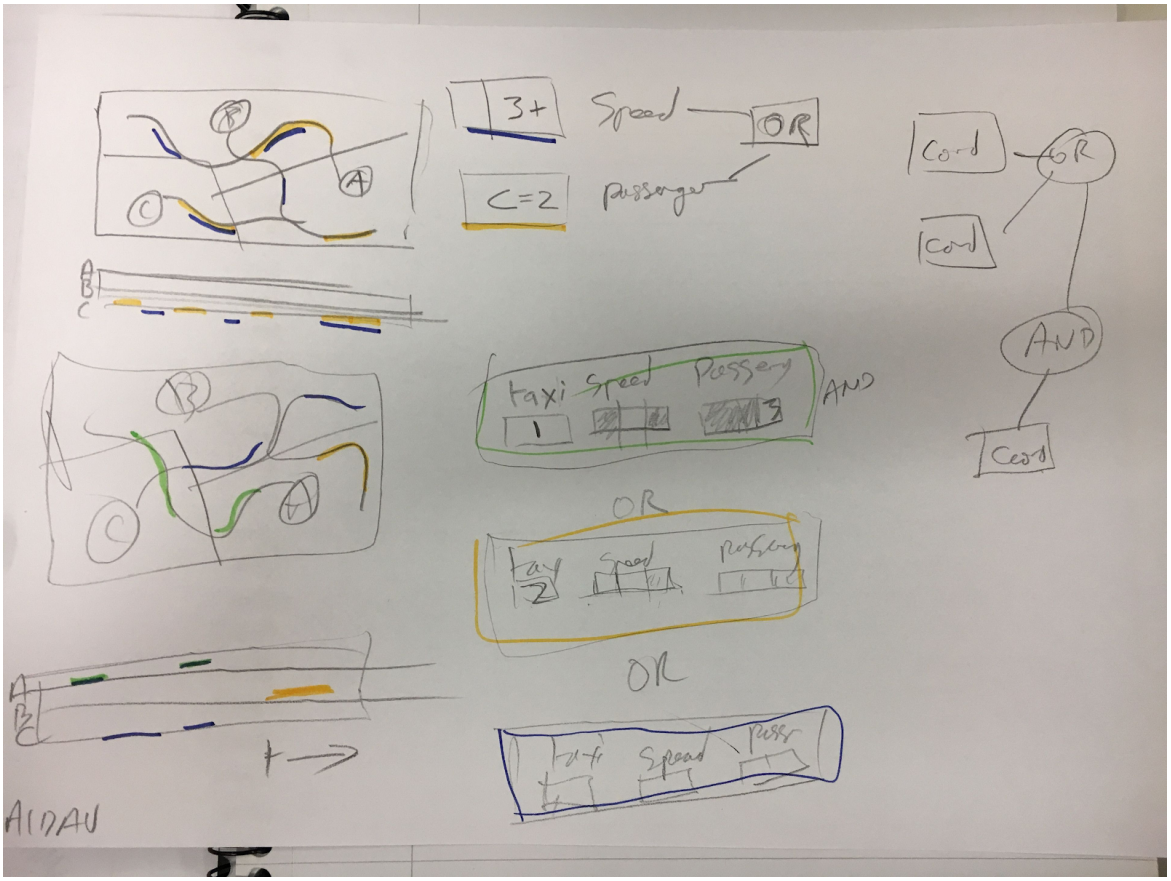

RAFAEL
1)
APX
B↓Z
C↓X
NUM ATTR
CAT ATTR
% NUMER ATTR
2)
Δd
AREA VIS SPACE
NUM ATTR 2
CAT ATTR
LINE THICKNESS = NUM ATTR 1
1ST VIZ: MAP MAYBE NEEDS TO CHANGE
2ND VIZ: WORKS THE SAME WAY?
WHERE
WHEN
1) BLURRED AREAS ON THE MAP, OR COLORED REGIONS OVER TIME
THICKNESS
2) BLUR + COLOUR OVER TABLE
BLUR LINE
WHERE
1) COLOUR THE REGION
2) HIGHLIGHT CELL OF REGION
AND
1) CHANGE BORDER THICKNESS IN TABLE
CIRCLE TICKS IN TRANSLUCENT
2) CHANGE COLOR IN CELLS IN TABLE, POSITION THAT IN LINE PLOT
1) BLUR + COLOR BY QUERY
TIMES
CELLS
2) " "
ASE ONE
1) COLOUR + BLUR TIMES
2)

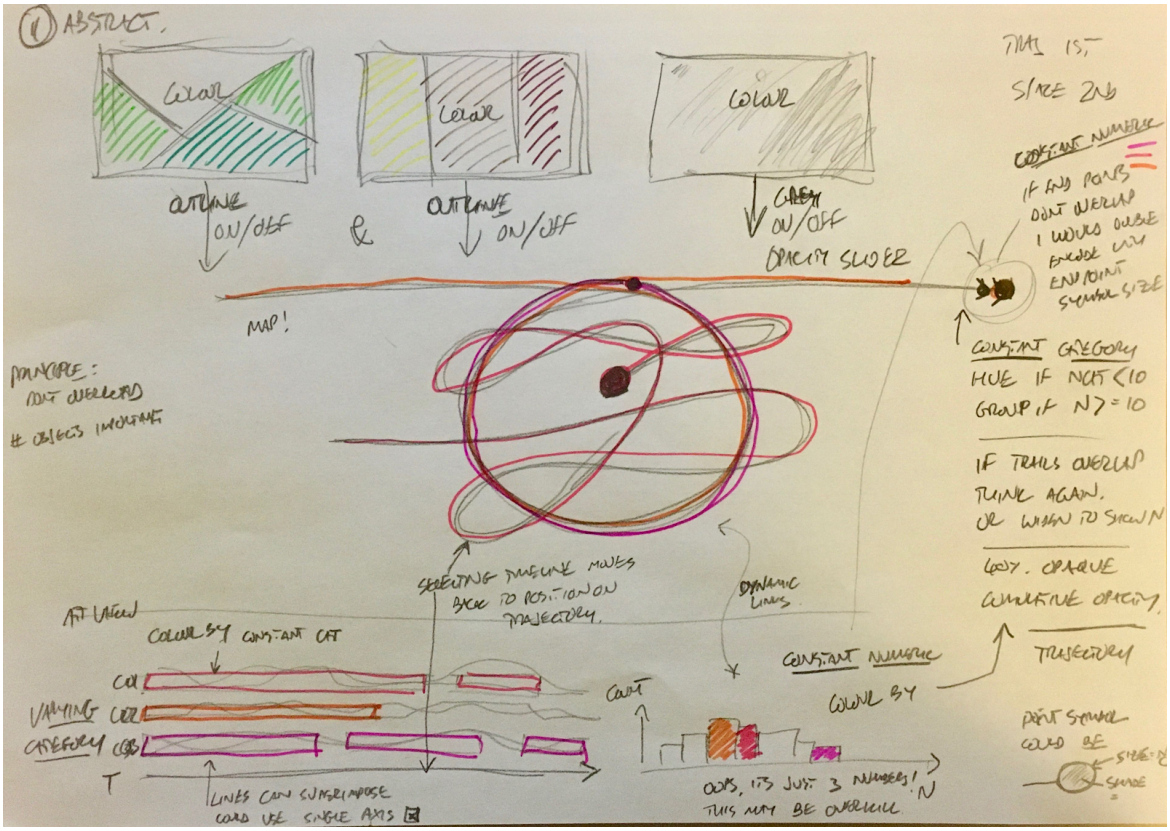

Abstract. (context less)
Vis 1
Vis 2
Vis 3
Vis 4
nv1
nv2

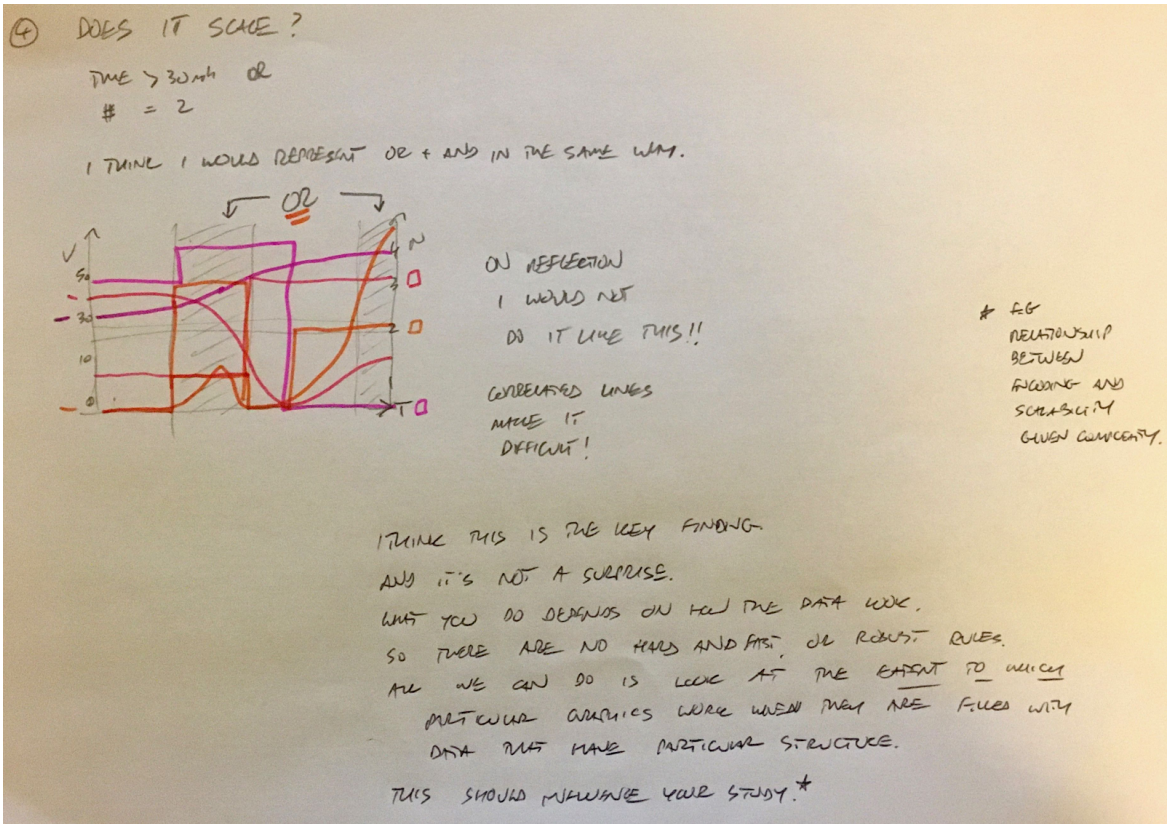

TOWN NOWHERE
DISTRICT
DISTRICT
DISTRICT
SLOW MED FAST
TIME
UNIT OF TIM
OBJECT
CAR - CAT - WHICH
N NUMBER OF PASSENGERS - HEIGHT
SPACE
- CAT - POSITION ON THE TABLE
- CAT - IN THE C ZONE - PALE HUE
- NUM - OF ONE WAY STREETS
VARYING
1 CAT
SPEED OF TAXI
SLOW MED FAST
NUM
NUM
SKETCH 2

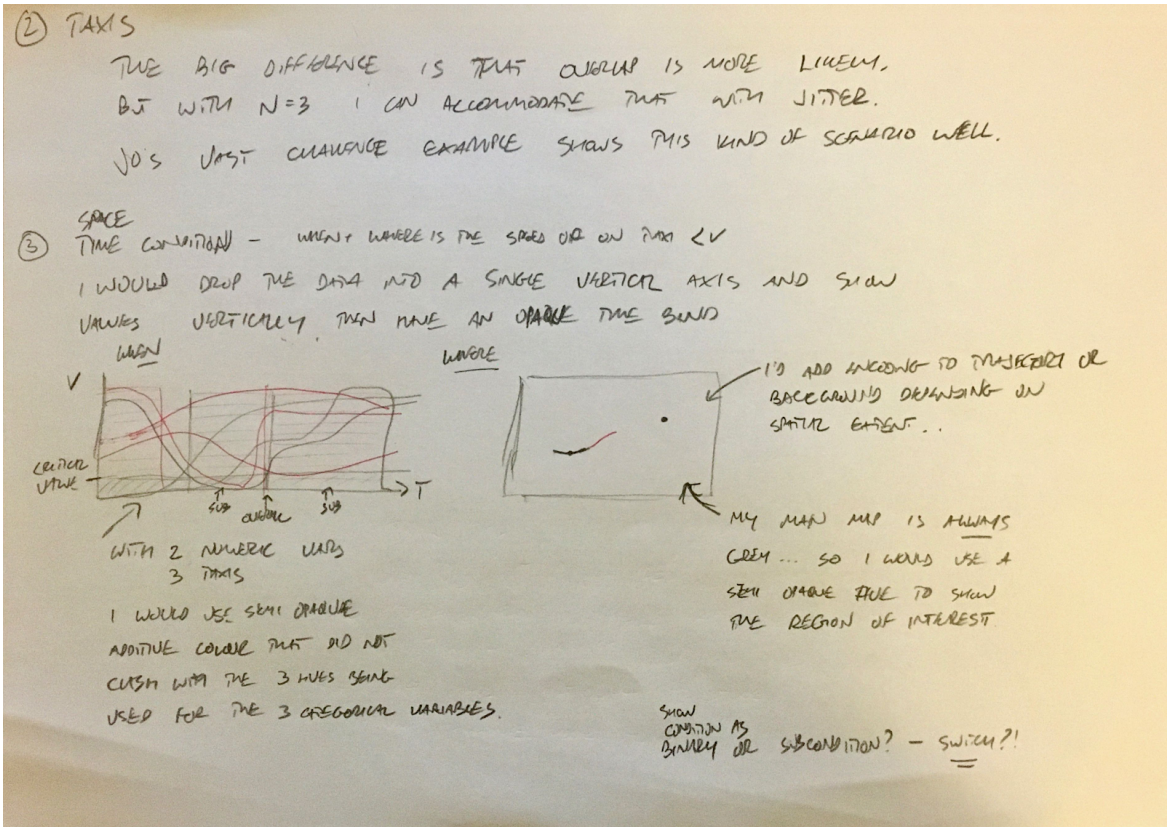

Road layout
size
hue
lightness
label
t
id of area
highlight the part
that meets the
condition
A
B
C
isochrons
dist
t
dist From A
t


Colored area are when respective car achieved their max. speed.
speed
t →
when speed is highest
A
Conditional
speed
AND
10 20
passenger
2
C
B
A
3.
taxi
AIDAN


3+
Speed
OR
C=2
passenger
Card
OR
Card
AND
Card
taxi
Speed
Passeng
AND
1
3
OR
OR
AIDAN


1 ABSTRACT.
COLOUR
COLOUR
COLOUR
OUTLINE ON/OFF
&
OUTLINE ON/OFF
ON/OFF
OPACITY SLIDER
MAP!
PRINCIPLE:
DOT OVERLOAD
OBJECTS IMPORTANT
SCRUBBING TIMELINE MOVES BACK TO POSITION ON TRAJECTORY.
DYNAMIC LINES.
ATT LAYER
COLOUR BY CONSTANT CAT
CAT
VARYING CAT
CATEGORY CAT
T
LINES CAN SUPERIMPOSE
COULD USE SINGLE AXIS
COUNT
CONSTANT NUMERIC
COLOUR BY
ODDS, IT'S JUST 3 NUMBERS! N
THIS MAY BE OVERKILL
THIS IS,
SIZE 2N
CONSTANT NUMERIC
IF END POINTS DON'T OVERLAP I WOULD QUITE ENJOY USE END POINT SYMBOL SIZE
CONSTANT CATEGORY
HUE IF NCAT ≤ 10
GROUP IF N ≥ 10
IF TRAILS OVERLAP TRY AGAIN. OR WARN TO SHOW N
LOW. OPAQUE
CUMULATIVE OPACITY
TRAJECTORY
POINT SYMBOL COULD BE
SIZE
SHAPE


④ DOES IT SCALE?
TIME > 30 mth OR
= 2
I THINK I WOULD REPRESENT OR + AND IN THE SAME WAY.
OR
50
30
10
0
N
T
ON REFLECTION
I WOULD NOT
DO IT LIKE THIS!!
CORRELATED LINES
MAKE IT
DIFFICULT!
* EG
RELATIONSHIP
BETWEEN
ENCODING AND
SCALABILITY
GIVEN COMPLEXITY.
I THINK THIS IS THE KEY FINDING
AND IT'S NOT A SURPRISE.
WHAT YOU DO DEPENDS ON HOW THE DATA LOOK.
SO THERE ARE NO HARD AND FAST OR ROBUST RULES.
ALL WE CAN DO IS LOOK AT THE EXTENT TO WHICH
PARTICULAR GRAPHICS WORK WHEN THEY ARE FILLED WITH
DATA THAT HAVE PARTICULAR STRUCTURE.
THIS SHOULD INFORMING YOUR STUDY.*


(2) TAXIS
THE BIG DIFFERENCE IS THAT OVERLAP IS MORE LIKELY.
BUT WITH N=3 I CAN ACCOMMODATE THAT WITH JITTER.
JO'S VAST CHALLENGE EXAMPLE SHOWS THIS KIND OF SCENARIO WELL.
(3) SPACE TIME CONDITION – WHEN + WHERE IS THE SPACE OUR OWN DATA <V
I WOULD DROP THE DATA INTO A SINGLE VERTICAL AXIS AND SHOW VALUES VERTICALLY. THEN HAVE AN OPAQUE TIME BAND
WHEN
V
CRITICAL VALUE
T
SUB
OVERLAP
SUB
WITH 2 NUMERIC VARS 3 TAXIS
I WOULD USE SEMI OPAQUE ADDITIVE COLOUR THAT DID NOT CLASH WITH THE 3 HUES BEING USED FOR THE 3 CATEGORICAL VARIABLES.
WHERE
I'D ADD ENCODING TO TRAJECTORY OR BACKGROUND DEPENDING ON SPATIAL EXTENT..
MY MAIN MAP IS ALWAYS GREY... SO I WOULD USE A SEMI OPAQUE HUE TO SHOW THE REGION OF INTEREST
SHOW CONDITION AS BINARY OR SUBCONDITION? – SWITCH?!