TU WIEN Informatics

A Study on the Impact of Uncertainties on the Analytical Process

BACHELORARBEIT

zur Erlangung des akademischen Grades

Bachelor of Science

im Rahmen des Studiums

Medieninformatik und Visual Computing

eingereicht von

Sonja Turner
Matrikelnummer 11908540

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Assistant Prof. Dr. Renata Raidou

Wien, 17. November 2023

Sonja Turner	Renata Raidou

TU WIEN Informatics

A Study on the Impact of Uncertainties on the Analytical Process

BACHELOR'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

Media Informatics and Visual Computing

by

Sonja Turner
Registration Number 11908540

to the Faculty of Informatics

at the TU Wien

Advisor: Assistant Prof. Dr. Renata Raidou

Vienna, November 17, 2023

Sonja Turner | Renata Raidou

Erklärung zur Verfassung der Arbeit

Sonja Turner

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 17. November 2023

Sonja Turner

Danksagung

Im Laufe dieser Bachelorarbeit hatte ich das Glück eine Menge Unterstützung erfahren zu dürfen. Daher möchte ich damit beginnen mich bei Assistant Prof. Dr. Renata Raidou zu bedanke, welche meine Arbeit betreut und mich mit viel Geduld und guten Ratschlägen unterstützt hat.
Weiterer Dank gebührt meinen Freunden und Kollegen sowie meiner Familie welche mich nicht nur emotional unterstützt haben, sondern auch die Zeit genommen haben an der Studie teilzunehmen.

Acknowledgements

During the writing of this bachelor's thesis, I was fortunate to receive a lot of support. Therefore, I would like to start by thanking Assistant Prof. Dr. Renata Raidou, who supervised my work and supported me with a lot of patience and good advice.
Further thanks go to my friends and colleagues as well as my family who not only supported me emotionally but also took the time to take part in the study.

Kurzfassung

Unsicherheiten wie fehlende Daten oder Ungenauigkeiten sind in allen Arten von Datensätzen vorhanden. Die Arbeit mit Unsicherheiten in Daten könnte schwerwiegende Auswirkungen auf den entsprechenden Analyseprozess haben.
Es wurden viele Studien und Untersuchungen durchgeführt, um Wege zu finden, Datenunsicherheiten zu verwalten, zu verhindern und sie in Visualisierungen darzustellen. Die Quantifizierung und Visualisierung von Unsicherheiten innerhalb von Analyserahmen ist ein gut erforschtes Thema, aber Informationen über den Einfluss von Unsicherheiten auf den Analyseprozess sind sehr spärlich. Ziel dieser Arbeit ist es, erste Erkenntnisse über die Auswirkungen von Unsicherheiten auf den Analyseprozess zu liefern. Zu diesem Zweck führten wir eine Studie bestehend aus einem Fragebogen und einem interaktiven Dashboard durch. Das für die Studie verwendete Szenario war die Vorhersage von Wetterdaten, wobei Temperatur- und Niederschlagswerte aus dem Bundesstaat Colorado als Daten für die Studie verwendet wurden. Wir haben uns auf bestimmte Arten von Unsicherheit konzentriert, d. h. fehlende Daten, sowohl zufällig fehlende als auch fehlende mit strukturellen Mustern, sowie Daten, deren genaue Position unbekannt ist. Darüber hinaus wurden auch Daten ohne Ungenauigkeiten als Grundlage für die Analyse verwendet, und simulierte Analyseszenarien mit Kombinationen der oben genannten Ungenauigkeiten wurden auch beim Design der Studie verwendet. Die Ergebnisse zeigen einen erheblichen Einfluss der im Szenario simulierten Unsicherheiten auf den Analyseprozess und die Genauigkeit der Prognosevorhersage. Die Stärke des Einflusses unterscheidet sich zwischen den Unsicherheiten.

Abstract

Uncertainties, such as data missingness or inaccuracies, are present in all kinds of data sets. Working with uncertainties in data might have serious implications on the corresponding analytical process. Many studies and investigations have been carried out to find ways to manage and prevent data uncertainties and to display them in visualizations. Quantifying and visualizing uncertainties within analytics frameworks is a well-researched topic, but information on the influence of uncertainties on the analytical process is very sparse. This thesis aims to provide initial insights into the effects of uncertainties on the analytical process. For this purpose, we conducted a study using an interactive dashboard. The scenario used for the study was the prediction of weather data, whereby temperature and precipitation values from the state of Colorado were used as data for the study. We focused on specific types of uncertainty, i.e. missing data, both randomly missing and missing with structural patterns, as well as data for which the exact position is unknown. Furthermore, data without inaccuracies were also used as a baseline of the analysis, and simulated analytical scenarios with combinations of the above-mentioned inaccuracies were also used in the design of the study. The results demonstrate a significant impact of the uncertainties simulated in the scenario on the analytical process and the accuracy of the forecast prediction. Both inaccuracies, positional uncertainty and missingness not at random, reduce the accuracy of the prediction. The data suggests that there is no clear difference in the strength of the influence of these two uncertainties. The accuracy of the prediction decreases as the ratio of missingness at random increases, although the correlation is weak.

Contents

CHAPTER 1

Introduction

Data sets are increasingly becoming larger and more difficult to handle. Therefore, the need for methods and techniques for data handling and analysis is also increasing. These large data sets, also called big data [WB13], can be found in many fields, including social media, medicine, and finances. Automatic analysis is sometimes not sufficient for the exploration and interpretation of the data. Hence, theories and concepts from visual analytics are used to interpret the data [KKP+11]. Visual analytics combine automatic data analysis with visual representation thereof and human perception [Fis05, KKP+11]. However, in all types of data, we encounter uncertainties, which occur, among other things, in the form of missing data, inaccurate values, or positions [unc]. The possible causes of uncertainties in data can be diverse, including inaccurately described real-world entities, data-capture techniques, and abstraction, as well as the processing of data such as the rendering of images [SZZZ18, GSWS21].

Due to the diversity of causes, missing data and inaccurate values are almost impossible to prevent. For this reason, it is also relevant to research what impact they have on the analytical process and to evaluate how they influence the result [SSK+16]. There are currently numerous studies that cover parts of one of these subject areas. During our literature research, we found papers that deal with the individual steps, challenges, and the relevance of visual analytics [Fis05, ALA+18, WV12]. Papers often explain the term inaccuracies in connection with data in more detail and present methods of dealing with them and analyze how these arise [WH16, MSW18, GSWS21]. Furthermore, there is also research about the practical implementation of visualizations that take uncertainties into account like uncertainty-aware applications and frameworks [FG14]. However, existing literature only describes methods to characterize, quantify, and/or display missing and inaccurate data. There is currently no work to specifically address the influence of uncertainties on the analytical process.

To gain further knowledge about the influence of the different uncertainties on dif-

ferent scenarios of the analytical process, further studies are required. The aim of the study presented in this paper is therefore to provide a first impression along this research direction. Since there have already been numerous studies on the details and relevant aspects of the analytical process as well as on the causes of inaccuracies in data, this work will not delve into these topics but will focus instead on how these two areas interact with each other.

The research questions that we will explore within this thesis are:

- *How uncertainties influence the analytical process?*
- *How do different types of uncertainty have a different influence on the analytical process?*

The contribution of this work will comprise an overview of the impact of uncertainties on the analytical process, as well as a preliminary understanding of how this influence differs with different uncertainties. We anticipate that the influence of inaccuracy reduces the accuracy of predictions, depending on the type and extent of the inaccuracy.

In the course of this thesis, we conducted a literature research to review the state of the art. The result of this research is summarized in Chapter 2 together with definitions relevant for this paper. Subsequently, a user study was conducted, forming the thesis's main part. This study consists of a questionnaire and an interactive dashboard. The Details about the design are described in Chapter 3. The results gained from this study are then analyzed and represented in Chapter 4 and, afterward, in Chapter 5 they are discussed to evaluate the possible reasons behind those.

CHAPTER 2

Background

2.1 Definitions

2.1.1 Analytical Process

The term analytical process refers to the combination of a person's analytical thinking and decision-making about the available data [Los13]. The purpose of an analytical process is to achieve new knowledge and insights from these large and sometimes complex data, for which purely automatic analysis methods would not have been sufficient[Kei09, ALA+18]. The data can be processed and presented in different ways to support the activity of analytical thinking. According to [Los13], the process can essentially be divided into two parts, the acquisition of the data up to the representation and the transfering of the data to the user via the representation. If visual techniques and methods or visual data representation are used for the implementation, this is also referred to as visual analytics or visual analytical process [Kei09].

Essentially, information visualizations include two components: the representation of the data and the interaction between data and humans [YKSJ07]. Most research on information visualization focuses on representation , but it is shown by Yi et al. [YKSJ07] that the interaction is equally relevant to enable efficient communication between user and data. According to most definitions, the term analytical process also includes the different techniques used for data representation and analytical reasoning, as well as interaction techniques[Fis05]. For this research, those aspects will not be discussed. Rather, this paper will focus on the result of the analytical process and the final gained knowledge.

2.1.2 Uncertainty

Uncertainty in connection with data occurs when unknown information or information whose accuracy is doubtful is present [WH16, BHP14]. This uncertainty can have several causes, including data collection, variance, and technical limitation [WH16]. Understanding the cause and quantifying uncertainty in data is crucial for making informed decisions, as it provides insights into the reliability, trust, and limitations of the data and the conclusions that can be drawn from it [SSK+16]. Uncertainty can be divided into multiple categories, but for the following study only positional uncertainty and missingness will be considered.

Positional Uncertainty

For the term positional uncertainty, multiple different definitions exist depending on the context in which the term is used [ROM, TG11]. Because no unified definition could be found in literature, for the scenario followed in this thesis, we define it as the uncertainty of data values at a specific geographic location within the border of a larger region for which the data value is known.

Missingness

Missingness describes the absence of values from the data set. A distinction can be made between the general categories of data missing at random, missing completely at random, and missing not at random [MSW18]. Missingness at random and missingness completely at random describe not systematically missing data. The difference is that with missing at random, the distribution of the observed data is likely to be similar to the distribution of the missing data, while for the distribution of missingness completely at random, no assumption can be made [BS14]. Missingness not at random occurs because certain factors or events have not been measured [MSW18]. This includes, among other things, the lack of measurement data at certain locations or time points. Due to time limitations, we only considered missingness at random and missingness not at random for this thesis.

2.2 Related Work

This chapter provides an overview of the state of the art in the area covered in this thesis and related topics.

Many research directions deal with the connection between uncertainties and big data. A fairly general overview of this topic is given in the paper of Wang et al. [WH16], which gives an overview of how uncertainties in big data arise, what types there are, and how they can affect the management of big data. In their work, the authors also show techniques that, depending on the present type of uncertainty, should help to manage and minimize the influence of the uncertainty.

The paper of Mack et al. [MSW18] follows a similar approach and deals with missing data in the field of medicine, more precisely with missingness that appears in patient registries. In addition to possible techniques for managing such inaccuracies, the paper also addresses possible effects on reports from studies with were conducted with missing data. However, this work only mentions problems with the statistical analysis, not with the analytical process.

Gillman et al. [GSWS21] focus on uncertainties in the visualization pipeline. The survey shows in detail which uncertainties exist in which steps of the pipeline and, similar to other works mentioned above, it also looks at the impact of uncertainties on the statistical analysis techniques that use medical images to gain knowledge. All three works provide a basis for developing technologies and applications that can cope as well as possible with different types of uncertainties.

The book "Illuminating the Path"[Fis05] offers a broad overview of this topic. In addition to a comprehensive explanation of the meaning of this term, it also includes what it entails. The book describes how factors such as representation, data type, and transformation can affect the analytical process and what they look like in detail. Although this book offers detailed information about what the term analytical process and visual analytics means and entails, it only helps to a limited extent in answering the question of this thesis because uncertainties are not taken into account here.

Andrienko et al. [ALA+18] offer a conceptual framework and definitions that are intended to facilitate carrying out studies in the field of visual analytics. The focus of the work is on the workflow and its components as well as the possible classifications, categorizations, and definitions that it contains. The work does not offer any concrete information about how external factors affect the process but does provide a structural basis for future research that deals with this.

The scientific work of Schiewe [Sch09] deals with the development of an approach using visual analytics which takes uncertainties into account in the change analysis process. The focus here is on the question of how inaccurate and missing information can be integrated into the analysis so that the result is as reliable as possible. To achieve this, they present techniques that can be used to visualize uncertainties and analyze the efficiencies of these techniques in different scenarios. The same question is also addressed in the work of Correa et al. [CCM09], who present the answer to a draft of a framework that is intended to simplify the visualization of uncertainties.

In their work, Wong and Varga [WV12] discussed why the visible representation of missing data in visual analytics is a very relevant topic for achieving a reliable result. To address this type of challenge, Fernstad and Glen [FG14] examine how visual techniques can be used to represent missingness. Furthermore, they describe techniques for analyzing the missing data to recognize patterns and trends, which is important because there is

also knowledge and insight that can be gained by analyzing the data that is not there.

In summary, there is currently a large amount of studies and research on both the topic of analytical processes and uncertainty. These not only provide precise definitions of the individual terms but also explain how uncertainties arise and why they represent a problem. However, the area covered by the papers found is in all cases limited to one of the three areas: visual analytics, management of uncertainty, or presentation of uncertainty. Work addressing specifically the impact of uncertainty on the visual analytics workflow and the corresponding analytical process is not available in the literature.

CHAPTER 3

Methodology

To get a better understanding of the impact of uncertainties on the analytical process a user study based on an analytical scenario was designed. The following chapter describes the details of the study design as well as the underlying hypotheses on which the study was built.

3.1 Hypotheses

The study aims to analyze the influence of inaccuracies on the analytical process. The main focus of the analysis of the data is therefore to look for patterns and discuss the possible explanations behind those. The following three hypotheses were put forward:

H1: *With increasing ratio of missingness at random, the accuracy of the predicted data compared to the prediction based on data without missingness at random decreases* As the proportion of missing data values increases, the absolute deviation between the expected value and the value predicted by the same participant for the same location, without any data missing at random, also increases. As in the work of Ayilara et al. [AZS+19], missing data can distort and lead to incorrect results. Although the paper only examines the influence using statistical methods and does not include analytical procedures, it is to be expected that the result will be comparable. Depending on how much and what data is missing, we assume that it will be more difficult for the analysts to make a precise decision. As a result, the expected value based on data with increasing missingness also increasingly deviates from the expected value based on complete data. However, in this context, it is difficult to argue whether the value is generally estimated too high or too low or whether the distribution is evenly distributed.

H2: *The impact of data missingness at random is unaffected by other uncertainties*
The absolute deviation of the expected values at a place affected by missingness at random compared to the expected values at the same place but with no missingness at random hardly differs, regardless of whether these places are affected by other uncertainties or not. We assume that when the data at the requested location is compromised by missingness at random and another type of uncertainty, the forecasting will be based on the remaining data. If the data missingness at random is distributed evenly between all data, then the remaining data, on which the prediction will be based, is also compromised by missingness at random. Therefore, we assume that the magnitude of the impact of missingness at random is consistent regardless of whether other uncertainties are present or not.

H3: *Data missingness not at random and positional uncertainty reduce the data value accuracy of the prediction*
For places affected by missingness not at random or positional uncertainty, we assume that the absolute deviation between ground truth and prediction is greater than compared to places unaffected by uncertainty. This hypothesis is based on the same reasoning as **H1**. Since missingness not at random also reduces the number of available data, this leads to the same problem that was already addressed in **H1**. Thus, the influence of the two types of missingness will be the same. The fact that positional uncertainty can lead to misinterpretations is also shown, among other things, by the work of Beconyte et al. [BB22]. This paper shows that the way data is aggregated in larger geographical units compared to smaller units can influence the perception of regional patterns and differences. It is therefore to be expected that the accuracy is reduced in the presence of missingness not at random and positional uncertainty, but it is difficult to estimate which type of uncertainty has more influence than others. Even from the information found in the related works, no evidence was found that could substantiate this hypothesis.

3.2 Study Design

To find an answer to the research question we designed and conducted a qualitative study. The following chapter first describes what scenario and corresponding data were used for the study and how the different types of uncertainty were simulated using those data. Subsequently, the study, which used a weather forecast scenario as a basis, was depicted. In this scenario design, we include the uncertainty types missingness and positional uncertainty. The study consisted of a digital questionnaire and an interactive dashboard to display the data. At the end of the chapter, an explanation of how the study was distributed and an analysis of the demography of the participants is included.

3.2.1 Data

The research objective addressed in this study does not require a specific scenario, so we decided to choose a scenario that is familiar to most people. The choice fell on predicting the weather because, on the one hand, it can be assumed that this is something that the majority of participants have already done at some point in their lives and, on the other hand, this is a scenario for which there is sufficient publicly available data. To analyze the influence of uncertainty on two different parameters, both temperature values and precipitation values were used for the study. Of the different weather parameters, these were the ones for which the most data was available. Since the distribution of weather stations is not the same everywhere in the world, when selecting the area from which the data comes, care was taken to ensure that the distribution was as even and as complete as possible. The National Oceanic and Atmospheric Administration [OoC] provides daily weather data from weather stations all around the world, so for this study, the data about Colorado from this website was used. The period covered by this data was exactly one year starting from the end of May 2022 to the end of May 2023. Because the units for precipitation and temperature usually used in Vienna differ from the units used by the National Oceanic and Atmospheric Administration, the data was converted from Kelvin and inches to degrees Celsius and millimeters.

The types of uncertainty that we decided to simulate within this scenario were positional uncertainty and two types of missingness, missingness at random and missingness not at random. The following paragraphs describe the procedure used to simulate these uncertainties.

Missingness at Random

In the case of missingness at random, the distribution roughly equals the distribution of the rest of the data [BS14]. To simulate this, a certain proportion of the data was randomly deleted from the data. To ensure that this happens randomly without bias, we used the Java function *Random*. The individual values were considered independently when deleted, which means that there are days in the data set for which either only the temperature value or precipitation value was deleted.

Since one of the goals of the study is to analyze whether the influence of missingness at random depends on the proportion of missing data, we created four different data sets using the procedure just described. The proportion of missing data is 2%, 5%, 10%, and 20%. For simplicity, these five, the original and the four new, data sets are referred to as charts 1 to 5, where chart 1 is the original dataset with 0% missingness and chart 5 is the one with 20% missingness.

Missingness not at Random

The other missingness that will be discussed in this paper is missingness not at random, which will be simulated as the missingness of the data of a region [BS14]. This region is a state in Colorado which, to avoid bias, we selected with the same Java function that was also used for the other missingness uncertainty simulation. To ensure that a ground truth is available for the analysis afterward, the states that could selected by the function was limited to state where at least the data from one weather station is included in the data collection. All data from the selected region, which turns out to be Boulder, was deleted from the dataset. Additionally, we selected a weather station in Boulder as a second place to simulate missingness not at random. This was done to analyze if there is a difference in the result between a geographic location and a region. Because in the selected state of Boulder, only one weather station with available data was present, it was not necessary to use a random function to select one. This station will be referred to as Mark 1 in the remainder of the thesis.

Positional Uncertainty

Within this research, positional uncertainty was defined as the uncertainty of the exact values at a geographical location within a region whose values are known. Since a ground truth is necessary for the following analysis, we selected a weather station whose values are known in the period in question as the geographical position. The values of the states are calculated from the average values of the weather stations present within the borders. This means that for those states for which there is only a single weather station whose values are known, the calculated values for the state are identical to those of the station and thus also to those of the location where positional uncertainty would be simulated. To prevent this case, the selection of possible locations was limited to those states that had at least two different data sets available. The final selection of the location was carried out using Java's pseudorandom function *Random* to avoid bias. Since deleting or changing the data is not necessary to simulate positional uncertainty, no new data sets were created. Because the selected weather station was labeled using an ID consisting of numbers and letters and this is unintuitive to people, we will be referring to the selected station as Mark 2 in the remainder of the work.

3.2.2 Dashboard Design

To visually display the data based on which the question of the questionnaire should be answered, a dashboard was created using the program Tableau [NSS16]. The dashboard, which can be seen in Figure 3.3, can be divided into two parts, a map to provide an overview of the precipitation and temperature in Colorado in the last month of the period covered by the data and timelines to display the details for a single state over the whole period. As mentioned in Section 2.1.1, interactivity is a relevant aspect of visual analytics, therefore the dashboard also provides interaction which combines the two parts of the

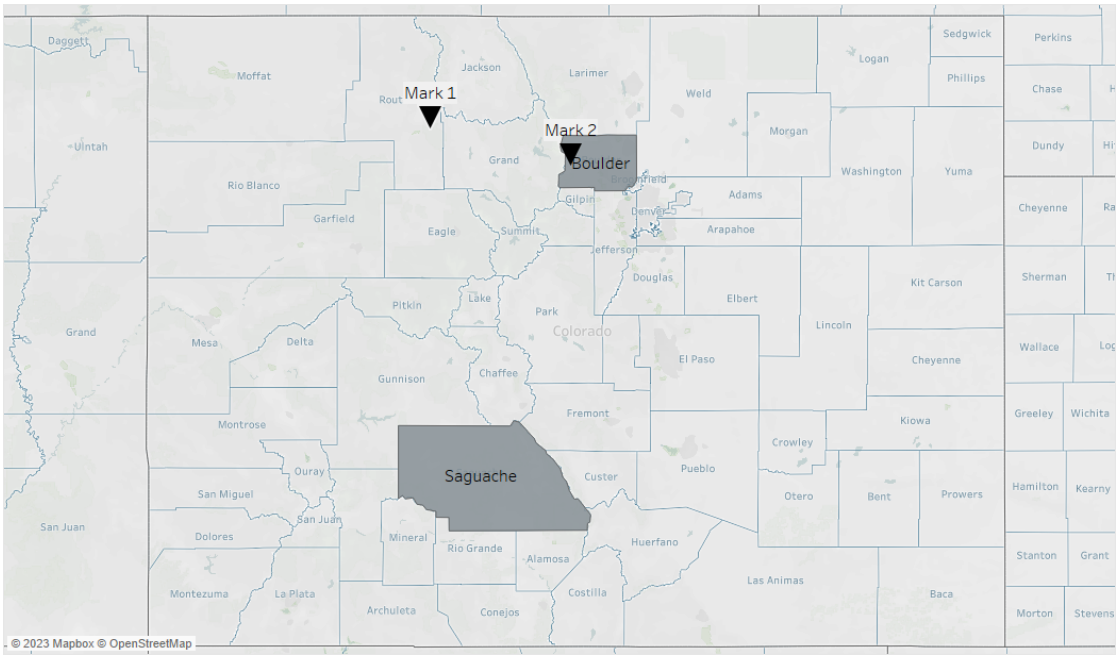


Figure 3.1: The positions and the labels of the places in Colorado that were used in the user study. The position of the two black markers equals the position of weather stations.

Place	Geographic Region Type	Uncertainty
Mark 1	Weather station	Missingness not at random
Mark 2	Weather station	Positional uncertainty
Boulder	State	Missingness not at random
Sanguache	State	none

Figure 3.2: Places used in the question of the questionary with the associated uncertainty and the type of geographic region.

dashboard to access details about the data.

Map

To display the precipitation and the temperature, we used a combination of two different map types, a chloropleth map that displays the average rainfall in the state and a bubble map to display the temperature. This map is shown in Figure 3.3 on the upper half of the dashboard. The difference in the precipitation was represented by the color saturation of the shade. To be able to display both, the average minimum and the average maximum temperature, the bubbles from the bubble map were divided in the middle. This was achieved by using a pie cart function where the proportion where fixed at 50% for both used variables. Originally it was planned to divide the bubble horizontally instead of

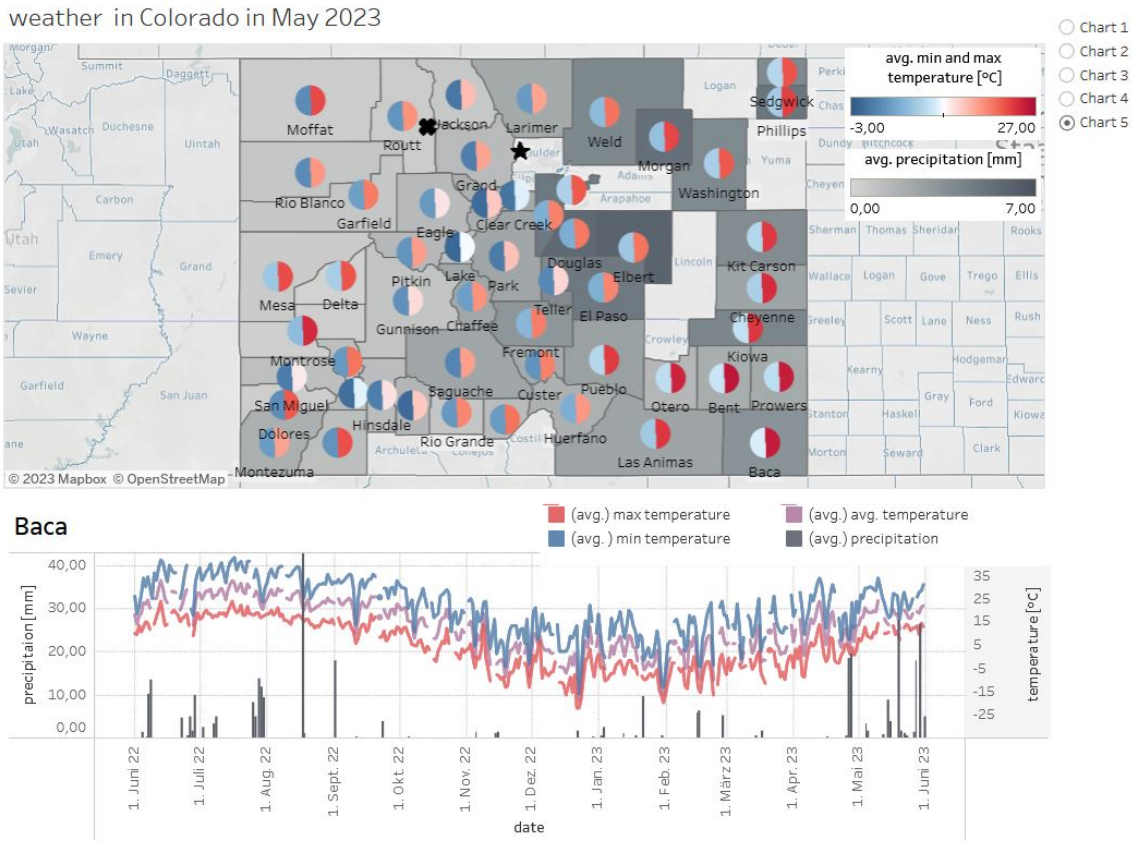


Figure 3.3: Dashboard created with Tableau where chart 5 and the state Baca are selected. The upper half of the dashboard shows the map with the marks, black star and cross, and the lower half shows the timelines. The radio button on the right side allows switching between the data of the charts.

vertically to increase the intuitiveness but because pie carts normally start with an angle of 0° [TS23], this was not possible to achieve with the integrated functions of Tableau. States in Colorado where no data was available were displayed as blank and the integrated background map of Tableau was used to display the name of these states.

To define the position of Mark 1 and 2 in this map we used black markers. To prevent confusion between the labels of the states and those for the two marks, the labels for Mark 1 and 2 were only visible when hovered over the marker. This interaction is shown in Figure 3.4. Two different Symbols, a star, and a cross, were used to differentiate between these two places. Originally, we planned to use a second different representation of the map, shown in Figure 3.5, in which no choropleth map would be used. Instead, the temperature would have been displayed with the same representation as described

above, but individually for each existing measuring station. The representation of the precipitation would have been displayed by changing the size of the bubble according to the value. The purpose of this would have been to examine the influence of different types of representation. However, this would have significantly increased the number of questions required in the survey. Therefore, for time-related reasons, we only implemented one type of representation.

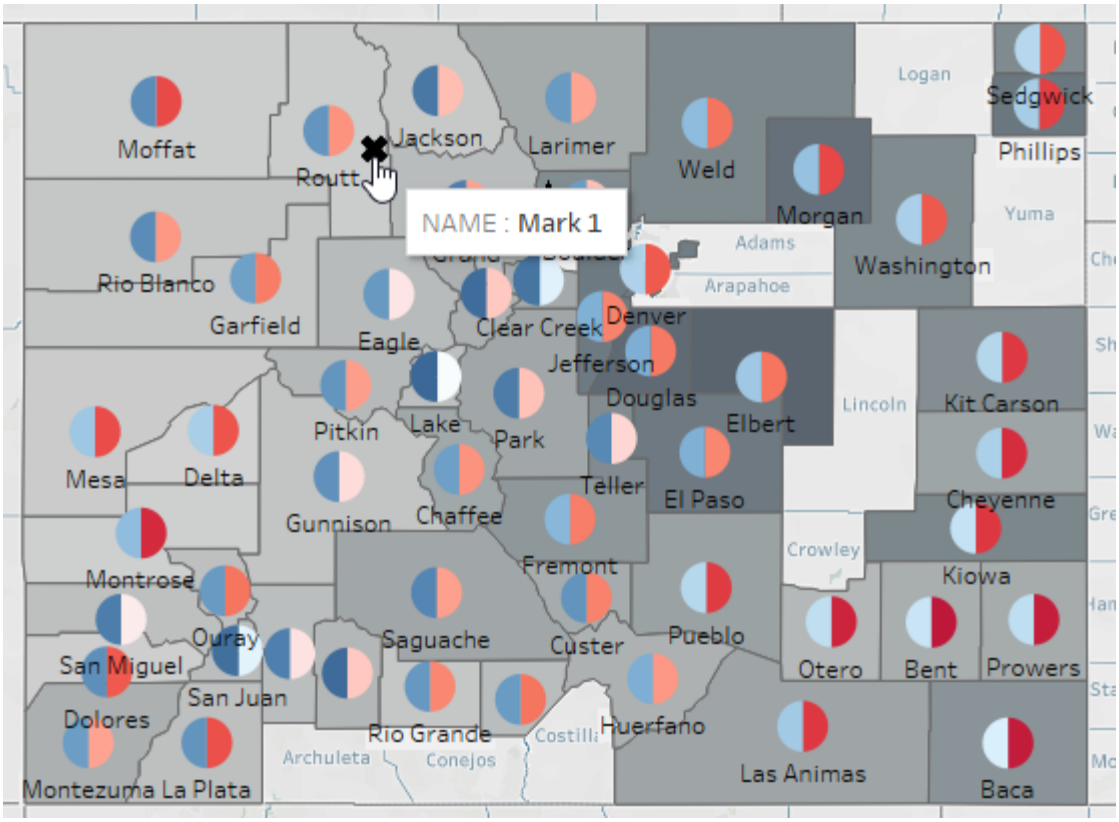


Figure 3.4: A representation of the map used for the study. The mouse hovers over the location marker which is called Mark 1, which is why the label is displayed.

Timelines

The second part of the dashboard are timelines to provide detailed information for a single state. To display precipitation and temperature in one chart, we used a combination of a line and a bar chart. The line chart was used for the temperature, where a separate line was used for the minimum and the maximum temperature. Additional to this the average temperature for each day was calculated and displayed as a third line in the chart. The average temperature was calculated after die data collection was edited to simulate uncertainty, therefor if for a day either the minimum or the maximum temperature is missing, the average temperature is also missing. We positioned these timelines in the bottom half of the dashboard shown in Figure 3.3. The precipitation was presented by a

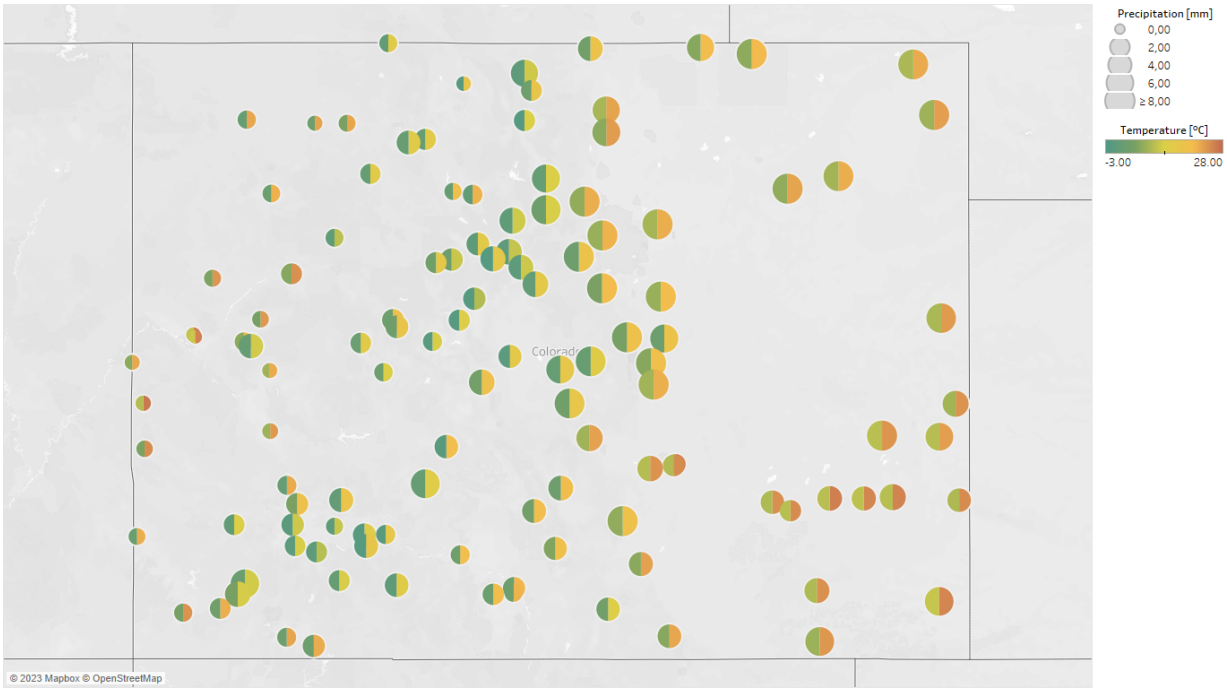


Figure 3.5: Different representation of the map which was not used in the study. The temperature of the weather stations is represented by color and the precipitation with the size of the bubbles.

bar chart. The axis for the bars was, concerning the axis of the line chart, set further down to prevent an overlap of the lines and bars. If no state is selected, as explained in the next section, the time chart will display the details for Colorado.

Interactivity

As previously mentioned, interactivity is an important aspect of the analytical process. The design of the dashboard we decided on contains two options for interactivity. The first interaction was necessary due to the chosen method to simulate the uncertainties. As described in Section 3.2.1, several data sets were created to simulate missingness at random. To display these data sets separately without overloading the user with too much information at once, we implemented the possibility to select the dataset that will be displayed. To switch between these in the dashboard, a radio button was added to the right side of the map. Changing the selected chart updates the data used to create the map and the timelines. To make it easier to switch between the charts, the scaling of the axes and the value range used were fixed. This means that the participant is not forced to adapt to a new scaling every time they switch between the charts.

The second interaction we implemented in the dashboard is to enable the user to access details about the data. The timelines, which by default show the data from Colorado,

can be limited to a state by clicking on it on the map. To cancel the selection, the user can click on it again or select any other location.

3.2.3 Questionnaire

To carry out this study, we created a digital survey that could be accessed with a link. There was no time limit for answering the questions, nor were there restriction about the place where they worked on the study or the device they used to access it. The main decision for this was that the chosen scenario, weather prediction, is neither time-critical nor restricted to a certain place. Rather, it is something people probably do in different locations on different devices in their everyday lives. Furthermore, even if the participants were asked to perform the analytical process in a certain way, there would be no easy way to ensure that they followed the request.

In terms of structure, the survey can be divided into three main parts. First, we provided our user study participants with the consent form at the beginning of the survey. Besides the information required by law, like the exact data that will be collected by the study, the form also contains contact information and a description of the anonymization procedure. If this is not agreed to, the following parts will be skipped and the survey will be ended immediately. In this case, an empty questionnaire will be sent. If accepted, the second part which contains general questions will follow regarding approximate age and experience with the topic being discussed. The answers to these questions will be used to determine the demography of the people who participated in the study

The main part of the survey is a series of questions that participants should answer based on the information they receive through the dashboard. An example of these questions would be: *What temperature (°C) would you expect on June 10 2023 afternoon in Boulder according to chart 4?*. The questions follow the format of asking about the expected value of either precipitation or temperature at a particular location at a particular time depending on a particular data set. The point in time was the same for all questions, namely a day slightly later than the period covered by the data. The participants could therefore not read the value they were looking for directly from the dashboard but had to speculate about it using the provided data in the dashboard. An example of such a question and the dashboard display is shown in Figure 3.6.

The locations and charts that appeared in the questions were those described in Section 3.2.1. The questions were divided into two sections, the questions about temperature and those about precipitation. The order within a section was chosen randomly for all participants. The answers to the questions were in numerical form without a unit and, if necessary, separated by commas or dots. Both separators were accepted and standardized to commas during the analysis. Before the questions relevant to the analysis, we added two simplified training tasks and a visual explanation of the dashboard and its interactions. The answers to this part of the user study were not taken into account in the analysis, as they were meant just for user familiarization.

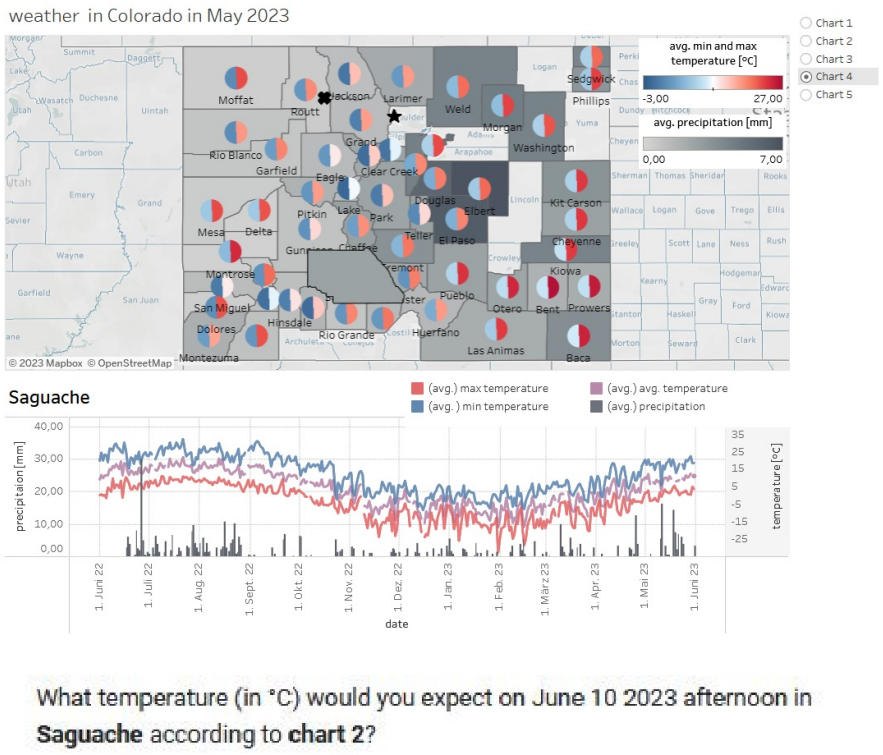


Figure 3.6: An example of a study question and the dashboard displaying the data that should be used by the participants to answer the question. The place in question is selected on the map and displayed in the timelines. The chart was selected with the radio button on the right side.

3.3 User Study Participation

For the study, 15 people volunteered to participate. The recruitment took place via instant messaging chatrooms and by word of mouth. Furthermore, we asked the participants to pass on the information and access to the study to other people to gain more participants. Therefore it is not possible to determine who exactly participated in the study. Since these chat rooms are mainly visited by people who, in the broadest sense, work or study in the field of computer science, we expect that this also applies to most participants of this study.

In addition to access to the study, participants also received an estimate of the time required to reduce the probability that people abandon the study because of faulty time management or demotivation. To guarantee the anonymization of the participants, we neither asked for their name nor their email address nor was this found out about during their participation in the study. Instead, we used an individually chosen ID by the participant to enable the possibility to identify the associated data to a participant in

Demographic	**size**
Age	
18-25	9
26-35	1
36-45	0
>45	5
Experience with visualizations	
expert	0
good	7
a little bit	3
none	5
Experience with meteorology	
expert	0
good	0
a little bit	9
none	6

Table 3.1: Demographic statistic of the participants.

the case that a participant wishes to withdraw their consent. However, this means that it is not possible to evaluate whether multiple answers were given.

CHAPTER 4

User Study and Results

4.1 Pilot Study

We conducted a pilot study with three participants to check whether the study design contained unclear parts. One of these participants was a peer researcher, and the other two people had no prior knowledge in this or similar areas of knowledge.

During the pilot study, we were able to observe that the map and its interactions were perceived as intuitive and therefore easy to understand. The exception to this was the state of Boulder where no data exists in the charts, which can be seen in Figure 3.3. One of the three participants was unsure at first if the interaction for this place was not working correctly. A note for explanation was added before the questions to avoid confusion like this for following participations.

The task description on the other hand confused all three participants. They expected to be able to read the value asked for in the question directly from the chart and were unsure how to answer when they realized that there was no definite answer. To prevent this problem in the finalised study further explanations were added to the already existing explanation mentioned above. The hints that were added are shown in Figure 4.1.

One noticeable thing when conducting the pilot study was the fact that how the three participants arrived at their answers differed greatly. This not only applies to the order in which they looked at the individual components of the dashboard but also which factors they considered to be more important than others. It cannot be ruled out that this may influence the results of the study, but it did not lead to a change in the design as the analysis of the individual steps of the analytical process is not part of the study.

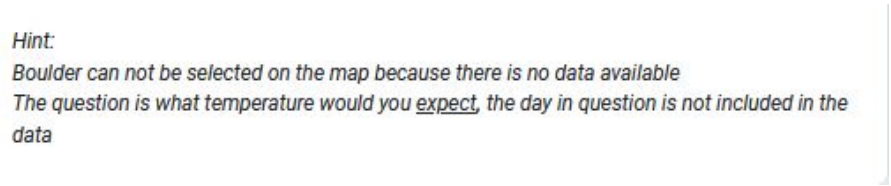


Figure 4.1: Hint that was added in the questionnaire before each question to clarify the interactions with the dashboard and what data are included. The hint was added to prevent the confusion which was evident in the pilot study

4.2 Tasks

The questionnaire consisted of three different blocks of questions, the questions to determine the demography of the participants, the training tasks, and the tasks for the analysis to answer the study's question.

The questions of the first block were multiple choice questions to determine the study-relevant characteristics of the participation group. The possible answers are shown in Table 3.1. The training tasks included two questions designed to introduce the participants to the interactive dashboard before the actual questions that were relevant to the analysis. These two tasks were similar to the questions in the last block, they only differed in the day that was asked about. The answers to the training tasks were not included in the results of the study.

The biggest part of the questionnaire was the third part. This part was divided into two, the questions about the temperature and those about the precipitation. Both parts consist of 20 questions, every combination of the four places and the five charts. The date that was asked about was the 10th of June 2023, which was 10 days after the last day that was included in the data displayed by the dashboard. The order of the questions in these two parts, separated according to temperature and precipitation, was randomized. This was done to prevent bias introduced by previous questions or concentration difficulties.

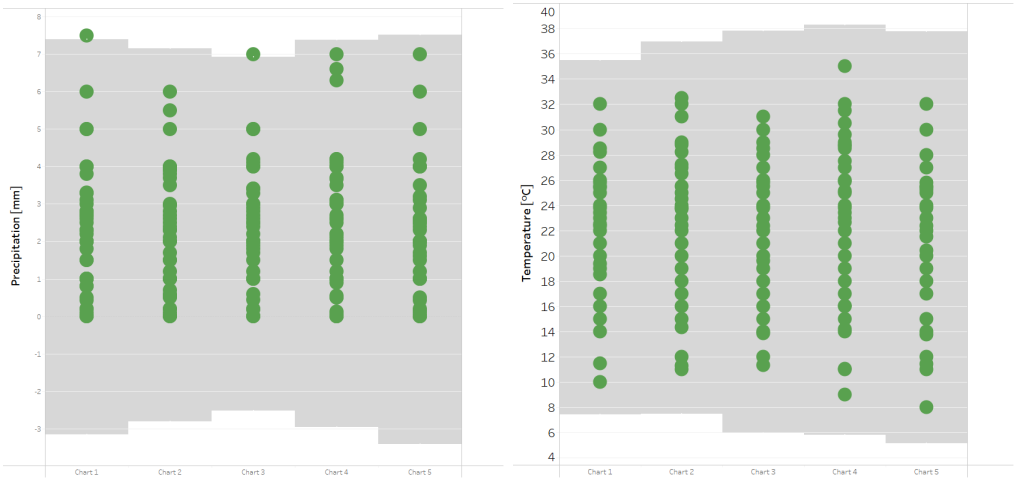
4.3 Results

4.3.1 Outliers

Since outliers have a major influence on the evaluation and can distort the results, we removed them from the data. For this purpose, the temperature and precipitation values were compared individually and the standard deviation was determined for each chart. The values that were more than three times the standard deviation away from the mean of the associated charts were removed from the data. This was true for two precipitation values, which can be seen in Figure 4.2b. No temperature values were outside the range.

After this step, the participants' values were then examined individually to check whether

there were participants who should be suspected of not taking the study seriously. The data from one participant has very large differences between the expected values for the individual charts at all four locations. The differences were in both directions without recognizable patterns. We assume that this person gave arbitrary answers when answering or made a mistake when interacting with the dashboard. To ensure that the results are not distorted, this participant's data were completely removed from the data set. Furthermore, a participant reported that a technical error occurred while answering the study questions. Since the error only affected the display of the data in chart 1, we decided to remove the answers that could be affected by this error, to avoid distortion.



(a) Precipitation (b) Temperature

Figure 4.2: The precipitation and temperature values obtained in the user study separated by chart. The grey area marks the three standard deviation range. Data points outside this area were treated as outliers.

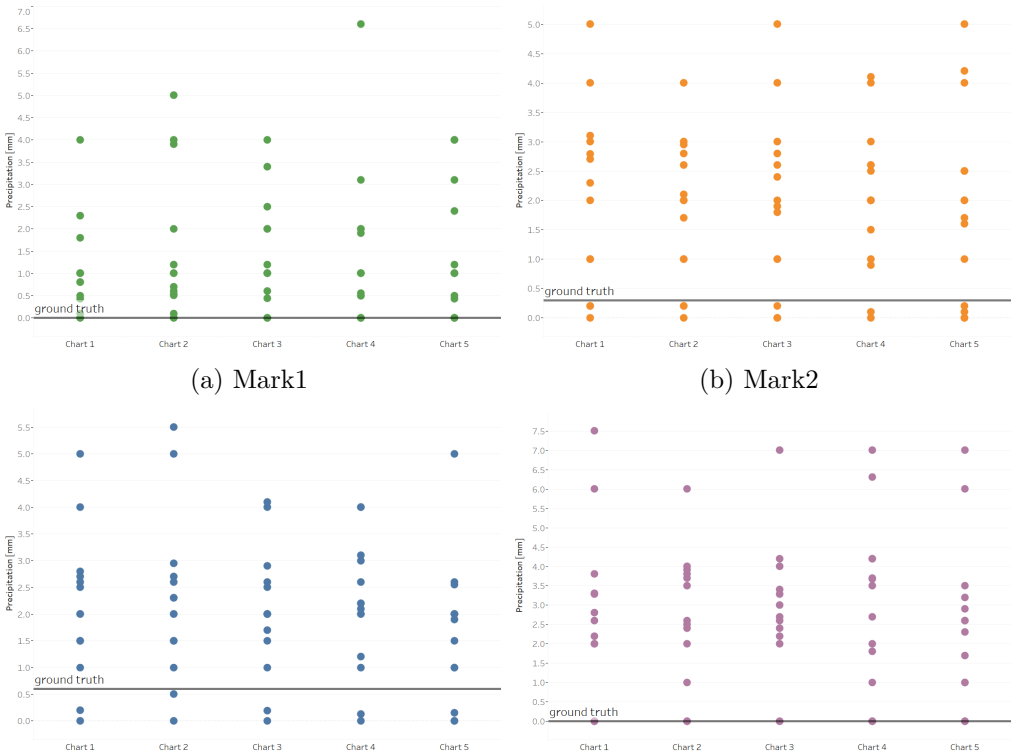
4.3.2 Data Distribution

The actual temperature and precipitation values at the location and day were used as ground truth. We obtained this data, like the previous data used for the dashboard, from the National Oceanic and Atmospheric Administration website [OoC]. It should be noted that ground truth was very similar to the data from the same day a year earlier, which was available in the dashboard. The exception to this is the precipitation of Saguache, which will be discussed in more detail in further analysis.

In Figures 4.3 and 4.4 it is possible to see that the data is approximately normally distributed with a similar scattering for all charts individually. Furthermore, the data in all charts shows individual values that are significantly above or below the other data. These minimum and maximum values, some of which are just below the limit at which you would have been considered an outlier, are the answers of different people. So it is not the case that individual participants generally gave higher answers, but rather that

the answers of participants in individual charts differ significantly from the values they gave in the other charts.

In terms of distribution, the values for the four locations do not differ noticeably in terms of precipitation. When it comes to temperature, it should be noted that the number of outliers at Saguache, i.e. when there is no uncertainty, is significantly lower than at the other three locations where missingness is not at random and positional uncertainty is present.



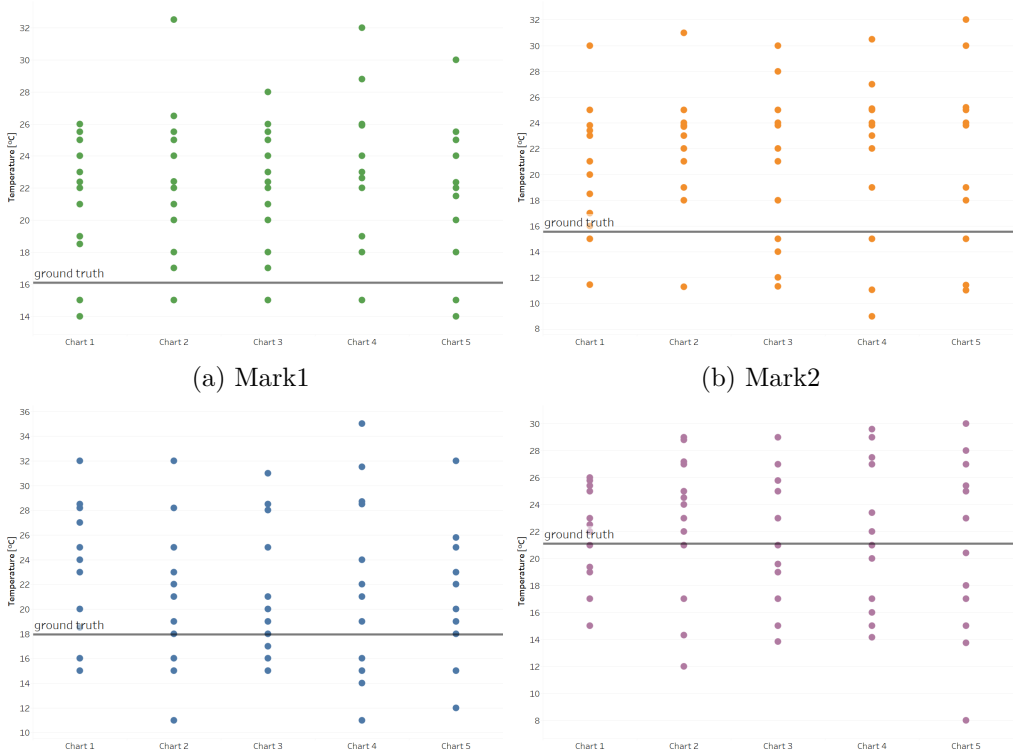
(a) Mark1 (b) Mark2

(c) Boulder (d) Sanguache

Figure 4.3: The answered precipitation values sorted by chart and compared to the ground truth. The ground truth corresponds to the values measured in reality at the corresponding weather stations and is indicated with a black line

4.3.3 Hypothesis H1

To show the influence of missingness at random more precisely, the mean absolute deviation was calculated between the data predicted for chart 1, the data without missingness at random, and the data for the other charts. In the resulting charts, Figures 4.6 and 4.7, we can see that outliers are present for all four places. Regarding **H1**, the hypothesis that the deviation increases when the ratio of the missingness at random increases, Figures



(a) Mark1

(b) Mark2

(c) Boulder

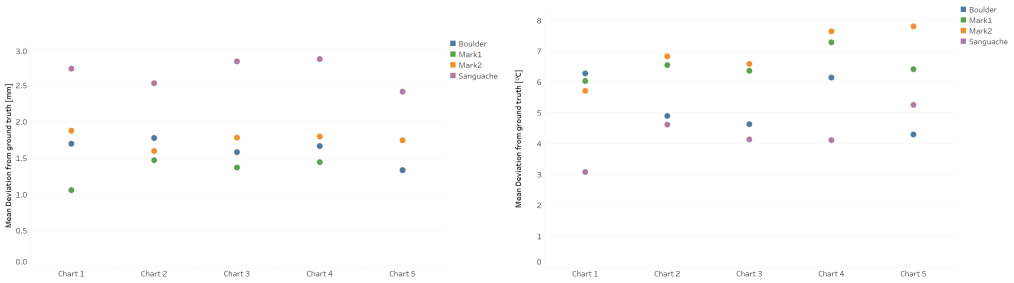
(d) Sanguache

Figure 4.4: The answered temperature values sorted by chart and compared to the ground truth. The ground truth corresponds to the values measured in reality at the corresponding weather stations and is indicated with a black line

4.8a and 4.8b indicate that there is a small positive correlation between the charts and the predicted values. To evaluate whether the hypothesis is statistically correct, we calculated Spearman's rank correlation [Dai] for the precipitation values ($r_s = 0.04, \rho = 0.55$), and for the temperature values ($r_s = 0.05, \rho = 0.38$). The Spearman's Rho r_s indicates how strong the correlation is between the variables to be examined. A value of +1 would correspond to a perfect positive correlation, and -1 to a perfect negative correlation. With a value of 0, no correlation could be determined. Therefore, we can say that there is a very small but significant positive correlation between the ratio of missingness at random and the deviation of the forecasted value.

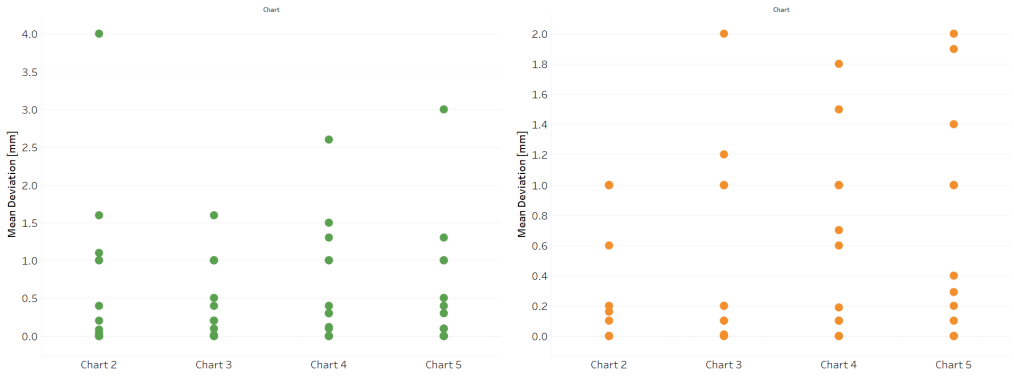
4.3.4 Hypothesis H2

The other hypothesis we put forward about missingness at random states that the influence of missingness remains the same regardless of what other inaccuracies were



(a) Precipitation (b) Temperature

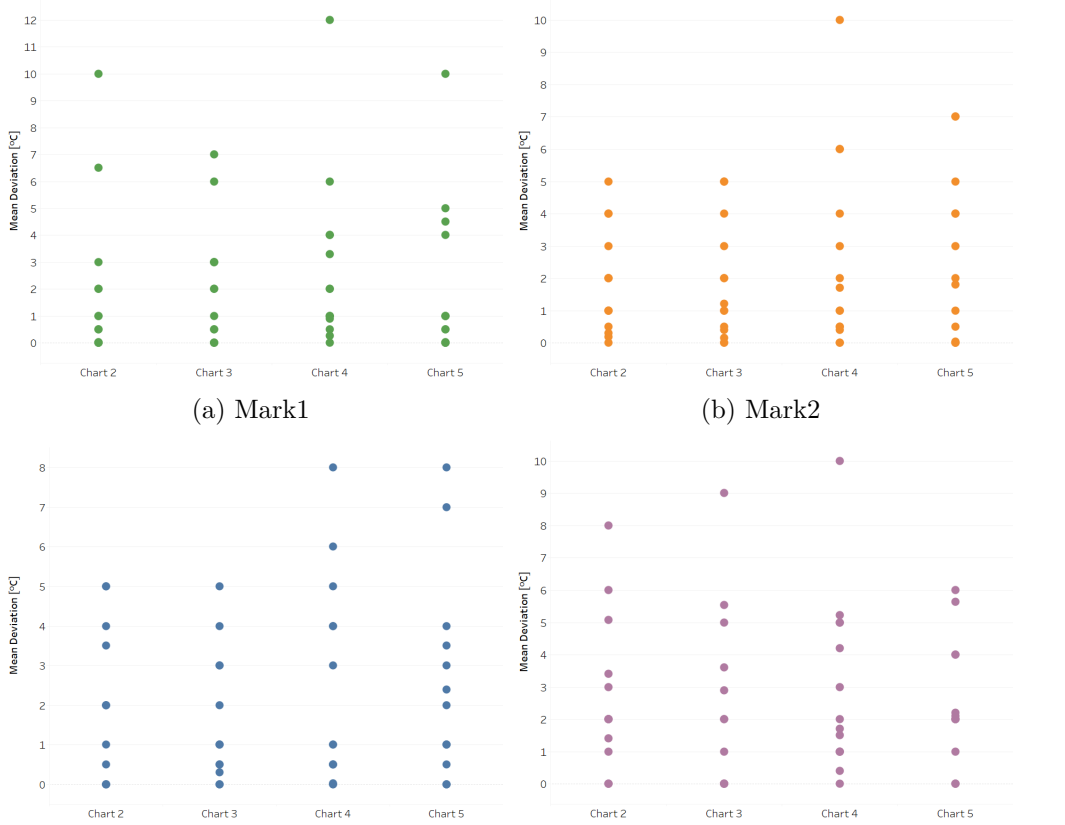
Figure 4.5: The means of the absolute deviations between the answered temperature and precipitation values and the ground truth. The computation was carried out separately for each chart and location.



(a) Mark1 (b) Mark2

(c) Boulder (d) Sanguache

Figure 4.6: Absolute deviation of the answered precipitation values of charts 2 to 5, which contain missingness at random, compared to the answered values of chart 1, which does not contain missing data. The comparison was carried out between the data values coming from the same participant



(a) Mark1

(b) Mark2

(c) Boulder

(d) Sanguache

Figure 4.7: Absolute deviation of the answered temperature values of charts 2 to 5, which contain missingness at random, compared to the answered values of chart 1, which does not contain missing data. The comparison was carried out between the data values coming from the same participant

present (**H2**). In Figure 4.8b we can see that the deviation change between charts is almost the same for all places, which contradicts our hypothesis. We performed a two-way ANOVA [Lis] to analyze if there is a significant interaction between the ratio of the missingness at random and the place. Carrying out the test with the temperature ($F(12, 256) = 0.68, p = 0.999$) and the precipitation ($F(12, 256) = 0.68, p = 0.768$) resulted in a non-significant interaction in both cases. This means that the results do not refute our hypothesis **H2**. This does not prove that the presence of other uncertainties than missingness at random, has no impact on the strength of the influence of missingness at random, but we can assume that **H2** holds.

(a) Precipitation (b) Temperature

Figure 4.8: The mean absolute deviation of the answered precipitation values of charts 2 to 5, which contain missingness at random, compared to the answered values of chart 1, which does not contain missing data. The absolute deviation was calculated between the data values that were obtained from the same participant. The mean was determined separately for each location and chart

4.3.5 Hypothesis H3

As described in Section 3.1, it was hypothesized that the positional uncertainty and missingness not at random decreases the accuracy of the forecast. To analyze whether this assumption is correct, we calculated the absolute deviation of temperature and precipitation from ground truth was calculated. The values used as ground truth were those which, according to the [OoC], were measured at the respective location. The calculated values can be seen in Figure 4.3 and 4.4, where the ground truth is indicated with a black line. In Figure 4.5b and 4.4d, it can be seen that the deviation of Saguache from the ground truth, regardless of the ratio of missingness at random, is greater than the deviation of the other locations. As already noted, the values in this location for ground truth differ greatly between years, as the period used in the survey was unusually dry compared to the period displayed on the dashboard. Saguache was the place where no uncertainty was simulated, as seen in Table 3.2. A meaningful comparison with the data from Saguache, which would be necessary for a verifiable judgment of the hypothesis, is not possible.

To evaluate whether the hypothesis is statistically correct, we performed an ANOVA [Lis] with the data from chart 1. Even though we can not use the precipitation to analyze the hypothesis, we decided to perform an ANOVA between the remaining three places to check whether there are significant differences depending on what inaccuracy is present. Regarding the temperature, the result from the ANOVA ($F = 2.56, p = 0.066$) shows no significant difference. To examine the differences in more detail, we carried out t-tests [Dai] between the data of the four places. The results of these t-tests, which can be seen in Table 4.1, show that there is a significant difference between Saguache and the other three places. The difference between the deviation of Mark 1, Mark 2, and Boulder is not significant. Regarding **H3** we can say that our hypothesis is accepted, which means that

places	t	p
Mark 1 - Mark 2	0.23	0.821
Mark 1 - Boulder	0.17	0.863
Mark 1 - Saguache	2.94	*0.007*
Mark 2 - Boulder	0.36	0.72
Mark 2 - Saguache	2.16	*0.041*
Boulder - Saguache	2.65	*0.014*

Table 4.1: The resulting p-values and t-values of separate t-tests computed with the temperature data. The t-tests were performed for the comparisons of all combinations of the places Mark 1, Mark 2, Boulder, and Saguache. The p-values were tested at the significance level of 0.05. Italics indicate a statistically significant result.

places	t	p
Mark 1 - Mark 2	-1.62	0.118
Mark 1 - Boulder	1.42	0.17
Mark 2 - Boulder	-0.35	0.726

Table 4.2: The resulting p-values and t-values of separate t-tests computed with the precipitation data. The t-tests were performed for the comparisons of all combinations of the places Mark 1, Mark 2, and Boulder. The p-values were tested at the significance level of 0.05. A negative t-test value indicates that the sample mean is lower than the population mean.

our results show a significant decrease in the accuracy of the forecast for places impacted by positional uncertainty and missingness not at random.

Regarding the precipitation, the ANOVA shows a significant difference ($F = 5.21, p = 0.013$). According to the separate t-test, which results can be seen in Table 4.2, this difference can be attributed to the difference between Mark 1 and the other two locations. On the other hand, the t-test regarding the temperature shows no noticeable difference between Mark 1 and the places Boulder and Mark 2, as shown in Table 4.1. This could indicate that the strength of the influence of missingness not at random is different for precipitation data than for temperature data.

4.4 Discussion

4.4.1 Hypotheses

The objective of the study was to gain insight into the impact of uncertainty, precisely missingness at random, missingness not at random, and positional uncertainty, on the analytical process. Since we could not find work that conducted a comparable study, it is not possible to compare the results obtained with other sources.

The hypothesis **H1** that was put forward was that with more data missing at random, the deviation of the values also increases. As far as this assumption is concerned, the data here also suggests that the assumption is correct. The influence of missingness at random is clear at 2% missingness, as the absolute difference between chart 1 and chart 2 in Figure 4.8b shows, the difference in the absolute difference between charts 2 and 5 is far smaller. This would suggest that either the increase of the difference equals a logarithmic function or another function until it reaches a point, between 2% and 5% missingness, where the increase suddenly stops. Between these two possibilities, the first one is far more plausible. To evaluate how the increase behaves, another study for conclusive results would be necessary.

Regarding hypothesis **H2**, which states that the influence of missingness at random remains the same regardless of whether other uncertainties are also present, it can be seen in Figure 4.8b that it is indeed possible to observe a pattern. The data, at least the temperature data, suggests that the impact of positional uncertainty and missingness not at random on the impact of missingness at random is nearly non-existent. Even though the pattern is not so clear in the precipitation data and the fluctuations in the data are sometimes very extreme in both directions, the statistical test indicates that our assumption also applies to the precipitation. We can assume that these large deviations come from the fact that this type of data is harder to predict [Sem].

The influence of positional uncertainty and missingness not at random on the accuracy of the prediction **H3** can be seen in Figure 4.5. The Saguache values for precipitation differ greatly from those of the others, but as noted, there was an unusually dry period in this location at this time. We therefore assume that this distorted the result and that this deviation therefore says nothing about the influence of the inaccuracies. The temperature and precipitation values at the other locations were in the average range at the time in question, but a slight distortion cannot be ruled out. Nevertheless, the data obtained suggests that positional uncertainty and missingness not at random certainly influence the accuracy of the forecast.

Furthermore, the temperature data shows no significant difference between the accuracy influenced by positional uncertainty and the accuracy influenced by missingness not at random. In the precipitation data a significant difference between the accuracy of Mark 1, a weather station for which missingness not at random was simulated, and the other two positions Mark 2 and Boulder. This could indicate that for missingness at random, it makes a difference whether a single weather station or a larger area is affected by this uncertainty.

In summary, we can conclude the following from the results of this study:

- The accuracy of the forecast decreases slightly with increasing ratio of missingness at random.

- The strength of the impact of missingness at random is not affected by the presence of other inaccuracies.
- Positional uncertainty and missingness reduce the accuracy of forecasting.
- The results indicate that there could be a difference in the impact strength of uncertainties on precipitation data and temperature data.

4.4.2 Limitation

After evaluating the data, some errors in the design of the study emerged that were not foreseen during the planning phase. The design has provided relatively clear results for temperature, but the collected data for precipitation are difficult to interpret. Therefore, it should be considered that the design used may not be suitable for all types of data. Furthermore, it can be problematic with measured values from nature if the data used as ground truth are outliers or are affected by unforeseeable events, as was the case with the Saguache precipitation values. Even if the difference is small, the results suggest that the difference caused by the uncertainties is also rather small, therefore it could still make a relevant difference.

Since the aim of this study was to provide a general overview, the number of participants was sufficient. However, it should be mentioned that the small number also means that individual values have a greater influence on the result, including outliers that may not have been recognized during the analysis.

CHAPTER 5

Conclusion and Future Work

This paper proposes a qualitative study to analyze the impact of missingness at random, missingness not at random, and positional uncertainty on the analytical process. The results have shown that all three different types discussed in this paper have a significant impact on the outcome of the process. Even though the strength of the influence varies, the effect of the impact is in all cases a decrease in the accuracy of the forecast. The study certainly proves that uncertainty is a relevant topic to consider within the analytical process which can have an impact on the result and should be further investigated in the future.

Regarding the results of the study, it should be said that the number of participants was very small, which means that outliers are particularly important for the evaluation. For more reliable results, it would be necessary to carry out this study again with a larger number of participants. The study was also quite limited in terms of the number of questions and locations at which inaccuracies were simulated, which reduces the trustworthiness of the results.

The next step in this research direction could include, on one hand, further investigation of the uncertainties discussed in this paper regarding the influence of different scenarios and visualization techniques and, on the other hand, research about the other kinds of uncertainties.

List of Figures

List of Tables

Bibliography

[ALA+18] N. Andrienko, T. Lammarsch, G. Andrienko, G. Fuchs, D. Keim, S. Miksch, and A. Rind. Viewing Visual Analytics as Model Building. *Computer Graphics Forum*, 37(6):275–299, September 2018.

[AZS+19] Olawale F. Ayilara, Lisa Zhang, Tolulope T. Sajobi, Richard Sawatzky, Eric Bohm, and Lisa M. Lix. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes*, 17(1):106, June 2019.

[BB22] Giedrė Beconytė, Andrius Balčiūnas, Aurelija Šturaitė, and Rita Viliuvienė. Where Maps Lie: Visualization of Perceptual Fallacy in Choropleth Maps at Different Levels of Aggregation. *ISPRS International Journal of Geo-Information*, 11(1):64, January 2022. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[BHP14] Marcel Boumans, Giora Hon, and Arthur C. Petersen, editors. *Error and uncertainty in scientific practice*. Number number 1 in History and philosophy of technoscience. Pickering & Chatto, London, 2014. OCLC: ocn870892164.

[BS14] Krishnan Bhaskaran and Liam Smeeth. What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology*, 43(4):1336–1339, August 2014.

[CCM09] Carlos D. Correa, Yu-Hsuan Chan, and Kwan-Liu Ma. A framework for uncertainty-aware visual analytics. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 51–58, Atlantic City, NJ, USA, 2009. IEEE.

[Dai] Ramon Daines. Academic Success Center Statistics Resources. `https://resources.nu.edu/c.php?g=901567&p=6487593`. Accessed: 2023-12-21.

[FG14] Sara Johansson Fernstad and Robert C Glen. Visual analysis of missing data — To see what isn't there. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 249–250, Paris, France, October 2014. IEEE.

[Fis05] Brian Fisher. Illuminating the Path: An R&D Agenda for Visual Analytics. pages 69–104. January 2005.

[GSWS21] Christina Gillmann, Dorothee Saur, Thomas Wischgoll, and Gerik Scheuermann. Uncertainty-aware Visualization in Medical Imaging - A Survey. *Computer Graphics Forum*, 40(3):665–689, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14333.

[Kei09] Daniel A Keim. Visual Analytics. *University of Konstanz*, 2009.

[KKP+11] Jörn Kohlhammer, Daniel Keim, Margit Pohl, Giuseppe Santucci, and Gennady Andrienko. Solving problems with visual analytics. *Procedia Computer Science*, 7:117–120, 2011.

[Lis] Lisa Sullivan. Hypothesis Testing - Analysis of Variance (ANOVA). `https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_hypothesistesting-anova/bs704_hypothesistesting-anova_print.html`. Accessed: 2023-12-22.

[Los13] David Loshin. Chapter 6 - Business Processes and Information Flow. In David Loshin, editor, *Business Intelligence (Second Edition)*, MK Series on Business Intelligence, pages 77–90. Morgan Kaufmann, January 2013.

[MSW18] Christina Mack, Zhaohui Su, and Daniel Weistreich. Managing Missing Data in Patient Registries. Technical report, Agency for Healthcare Research and Quality (AHRQ), February 2018.

[NSS16] Lekha R Nair, Sujala D Shetty, and Siddhanth D Shetty. Interactive visual analytics on the big data: Tableau VS D3.JS. 12(4), 2016.

[OoC] National Oceanic and Atmospheric Administration U.S. Department of Commerce. National Oceanic and Atmospheric Administration. `https://www.noaa.gov/`. Accessed: 2023-11-15.

[ROM] Craig Roberts, Selin Ozdemir, and Simon McElroy. Where is positional uncertainty? *School of Surveying and Spatial Information Systems, University of New South Wales.*

[Sch09] J Schiewe. Visual analytics approach for cinsidering uncertainty information in change analysis process. *HafenCity University Hamburg, Lab for Geoinformatics and Geovisualization*, 2009.

[Sem] Steven Seman. College of earth and mineral science, department of meteorology and atmospheric science meteo3. `https://www.e-education.psu.edu/meteo3/node/2285`. Accessed: 2023-11-15.

[SSK+16] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249, January 2016.

[SZZZ18] Wenzhong Shi, Anshu Zhang, Xiaolin Zhou, and Min Zhang. Challenges and prospects of uncertainties in spatial big data analytics. *Annals of the American Association of Geographers*, 108(6):1513–1520, 2018.

[TG11] Michele Tucci and Alberto Giordano. Positional accuracy, positional uncertainty, and feature change detection in historical maps: Results of an experiment. *Computers, Environment and Urban Systems*, 35(6):452–463, November 2011.

[TS23] Salesforce Company Tableau Software, LLC. Pie Charts. `https://www.tableau.com/data-insights/reference-library/visual-analytics/charts/pie-charts`, 2023. Accessed: 2023-12-20.

[unc] Types of Uncertainty: Overview Uncertainty Quantification.

[WB13] Jonathan Stuart Ward and Adam Barker. Undefined By Data: A Survey of Big Data Definitions, September 2013. arXiv:1309.5821 [cs].

[WH16] Xizhao Wang and Yulin He. Learning from Uncertainty for Big Data: Future Analytical Challenges and Strategies. *IEEE Systems, Man, and Cybernetics Magazine*, 2(2):26–31, April 2016. Conference Name: IEEE Systems, Man, and Cybernetics Magazine.

[WV12] B.L. Wong and Margaret Varga. Black Holes, Keyholes And Brown Worms: Challenges In Sense Making. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56:287–291, October 2012.

[YKSJ07] Ji Soo Yi, Youn Ah Kang, John Stasko, and J.A. Jacko. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, November 2007.