Uncertainty-aware Visual Analytics for Spatio-temporal Data Exploration

Dissertation submitted for the degree of
Doctor of Natural Sciences

Presented by

Hansi Vihara Senaratne

at the

Faculty of Sciences

Department of Computer and Information Science

Date of the oral examination: 05. May 2017

First referee: Prof. Dr. Tobias Schreck

Second referee: Prof. Dr. Daniel Keim

Acknowledgements

I would like to thank everyone who supported me during the preparation of this dissertation. At first, I am deeply grateful to Prof. Dr. Tobias Schreck and Prof. Dr. Daniel Keim for their supervision and constant guidance towards this final result. With many constructive discussions, ideas, time, and untiring efforts they have helped me to shape my research.

Furthermore, I greatly appreciate the collaborations with my colleagues at the data analysis and visualisation group at the University of Konstanz, Dominik Sacha, Dominik Jäckle, Sebastian Mittelstädt, Geoffrey Ellis, and Bum Chul Kwon, as well as Amin Mobasheri, Ahmed Loai Ali, Muki Haklay, and Cristina Capineri with whom I had the pleasure of collaborating externally. Furthermore, I want to thank all the other co-authors with whom I worked on publications in context of this research. Not forgetting is the geo-analytics sub research group that I would like to express my gratitude for taking time to give me great many feedback.

Special thanks go to the students who I have supervised during their Master or Bachelor theses: Christina Jacob, Dominic Lehle, and Manuel Mueller. By aligning their works with the conceptual framework of this dissertation, fruitful collaborations could emerge. I thank them for constructive discussions, their implementation work, as well as their contributions to mutual publications.

For their efforts in reviewing this dissertation and providing such helpful feedback, I would like to thank my supervisors. Furthermore, I would like to thank all those reviewers of my publications who have provided me insightful feedback to improve my work.

Without the financial support through the German Research Foundation (DFG) under the grant *GK-1042* and the DFG program *SPP-1335* this research would not have been possible. Thus, I would like to thank my supervisors as well as the professorial staff of the Computer and Information Science department of the University of Konstanz for giving me the opportunity to freely develop research within these

projects. Kindly acknowledged are all the institutions that provided me with the valuable datasets and other resources to carry out my research.

Finally, I want to thank my family for their encouragement and patience, my parents particularly for their direction and undying support. My deep gratitude goes to my husband, Arne, for his love, care, and constant encouragement to tread on.

Summary

Uncertainty in spatio-temporal data is described as the discrepancy between a measured value of an object and the true value of that object. Common causes of uncertainty in data can be identified as errors of precision in the data measurement devices, inadequate domain knowledge of the data collector, absence of gatekeepers etc., known in this dissertation as inherent or source uncertainties. These inherent uncertainties further vary depending on the type of data (e.g., geotagged text or image data), as well as the explicit and implicit nature of the spatial dimension in the data.

Static and dynamic visualisation methods have been used to communicate uncertainties. However, a gap we see in such uncertainty visualisations is that users have little to no leeway of controlling the system outcomes (e.g., by weighing in their domain expertise, control to what extent uncertainty plays a role in the analysis, or reduce uncertainty in the data). Visual analytics help to fill this gap by allowing the user to steer the analysis process through interaction. The challenge of uncertainty analysis with visual analytics is that we not only have to encounter the inherent data uncertainties, but also the uncertainties that keep propagating through every component in a visual analytics system (the data, data models, data visualisations and model-visualisation couplings), and through every interaction from the user.

To address this challenge, this dissertation introduces a framework that defines the role of uncertainty throughout the visual analytics knowledge generation process. At each component of the visual analytics system, guidelines in terms of methods are specified for assessing the uncertainties. Following this framework, four novel visual analytics approaches are introduced that enable a user to explore, assess, and mitigate context-specific uncertainties in heterogeneous data types: image data, text data, location data, and numerical data. By enabling a strong interaction between the user and the system, uncertainties are mitigated and trustworthy knowledge is extracted, thereby bridging the gap identified in static and dynamic uncertainty visualisations. The approaches developed are evaluated against anecdotal evidences and a usability experiment.

Zusammenfassung

Unsicherheit in raum-zeitlichen Daten ist die Diskrepanz zwischen dem wahren Wert und dem gemessenen Wert einer Messgröße eines Objekts. Typische Gründe für das Entstehen von Unsicherheit in Daten sind z.B. Fehler in der Präzision der Messgeräte oder unzureichendes Domänenwissen des Datenerfassenden. Diese Art von Unsicherheiten in raum-zeitlichen Daten werden in dieser Dissertation als inhärente Unsicherheiten bezeichnet. Diese unterscheiden sich sowohl in Abhängigkeit vom Typ der raum-zeitlichen Daten (z.B. georeferenzierte Text oder Bilddaten), als auch in der expliziten bzw. impliziten Natur der Georeferenz dieser Daten.

Bislang wurden statische und dynamische Visualisierungsmethoden benutzt, um Unsicherheiten zu kommunizieren. Allerdings haben Nutzer dabei wenig oder gar keinen Spielraum, das System zu kontrollieren. Visual Analytics hilft diese Lücke zu schließen, indem der Nutzer den Analyseprozess durch Interaktion steuern kann. Die Herausforderung der Analyse von Unsicherheiten mit Visual Analytics ist dabei, dass wir nicht nur den inhärenten Unsicherheiten begegnen, sondern auch den Unsicherheiten, die sich durch jede Komponente des Visual Analytics Systems (d.h. die Daten, Datenmodelle, Visualisierungen, und Modell-Visualisierungs-Kopplungen) und durch jede Interaktion des Nutzers verbreiten.

Um diese Herausforderung zu adressieren, führt diese Dissertation ein Framework ein, dass die Rolle von Unsicherheit im gesamten Wissenserzeugungsprozess der Visual Analytics definiert. Für jede Komponente eines Visual Analytics Systems werden Leitlinien in Form von Methoden zur Bestimmung der Unsicherheiten spezifiziert. Diesem Framework folgend, werden vier neuartige Visual Analytics Ansätze entwickelt, die den Nutzer befähigen, Kontext-spezifische Unsicherheiten in heterogenen Typen von Daten (Bild-, Text-, Geo-, sowie numerische Daten) zu untersuchen, zu bestimmen und zu mildern. Durch das Ermöglichen einer starken Interaktion des Nutzers mit dem System können Unsicherheiten abgeschwächt werden und vertrauenswürdiges Wissen kann extrahiert werden. Die entwickelten Ansätze werden durch anekdotische Evidenz sowie eine Nutzbarkeitsstudie evaluiert.

Contents

1 Introduction **1**

1.1 Need for Visual Analytics in Uncertainty-aware Data Exploration . . 1

1.2 Dissertation Structure 5

1.3 Contributions of this Dissertation 7

1.4 Publications 8

2 The Role of Uncertainty in Spatio-temporal Data Analysis **15**

2.1 Background & Related Work 16

2.1.1 Uncertainty Measurements & Indicators 17

2.1.2 Uncertainty Visualisation 19

2.1.3 Taxonomies for Uncertainty Visualisation 28

2.2 A Framework and Guidelines for Uncertainty Inheritance and Propagation in Visual Analytics 30

2.2.1 Guidelines to Assess Source Uncertainty in Spatio-temporal Data 32

2.2.2 Guidelines to Assess Propagated Uncertainties in Spatio-temporal Data 54

2.2.3 Guidelines to Aggregate Uncertainties 60

2.2.4 Guidelines to Visualise Uncertainty Information 61

2.2.5 Guidelines to Enable Interactive Uncertainty Exploration . . . 62

2.3 Discussion & Future Work 62

2.4 Conclusions 63

3 Uncertainty Analysis of Image-based Volunteered Geographic Information **65**

3.1 Background & Related Work 66

3.2 Reverse-viewshed Analysis for Assessing the Positional Accuracy of Image-based VGI 69

3.2.1 Flickr Metadata Retrieval with the FlickrMetaCrawlr 71

3.2.2 Reverse-viewshed Analysis for POIs 72
3.3 Credibility as an Uncertainty Indicator for Flickr Images 75
3.4 Discussion & Future Work 82
3.5 Conclusions 86

4 Uncertainty-aware Movement Analysis in Text-based Volunteered Geographic Information 87
4.1 Background and Related Work 89
4.2 Movement Detection in Implicitly Referenced Spatial Data 91
4.2.1 Keyword-based and #hashtag-based Data Gathering 92
4.2.2 Hostspot & Cluster Analysis with KDE & DBSCAN 93
4.2.3 Conversation Movement Trajectories 97
4.3 Sructural Characterisation of Movement Trajectories 99
4.3.1 Geospatial Structure-based Characterisation 99
4.3.2 Content Structure-based Characterisation 102
4.4 Feature-based Trajectory Ranking 109
4.4.1 Example Scenario 1 110
4.4.2 Example Scenario 2 111
4.5 Conversation Movement Analysis for Sports Journalism 116
4.6 Discussion & Future Work 118
4.7 Conclusions 123

5 Uncertainty-aware Space-time Exploration from Location-based Mobile Communication Data 125
5.1 Background & Related Work 127
5.2 Location-based Mobile Communication Data 128
5.3 Spatio-temporal Analytics through Movement Patterns 129
5.3.1 Movement Trajectory Extraction from Location-based Mobile Internet Usage Data 129
5.3.2 Spatial and Temporal Movement Similarity of Users 132
5.3.3 Place Classification based on Home and Work Area Detection 135
5.3.4 Regional Partitioning with Origin-Destination Analysis 137
5.3.5 Spatial Area Analysis based on Temporal Usage Patterns . . . 140
5.4 Uncertainty in Movement 141
5.4.1 Space-time Prisms to Analyse Uncertain Movement Path Segments 143
5.4.2 Positional Uncertainty Reduction in Mobile Communication-based Movement Data 146

5.5 Discussion & Future Work 148
5.6 Conclusions 150

6 Uncertainty Analysis of Bi-dimensional Numerical Data 151
6.1 Background & Related Work 152
6.2 Numerical Data from a Smart Grid Network 155
6.3 Applying Monte Carlo Simulation and Sampling Method for Uncertainty Assessment 157
6.4 Glyph Design for Bi-dimensional Uncertainty Analysis in Smart Grid Environments 159
6.5 Performance and Preference of Glyph Designs 162
6.5.1 Design of the Usability Study 164
6.5.2 Study Results 169
6.6 Discussion & Future Work 173
6.7 Conclusions 175

7 Conclusion and Future Work 177
7.1 Summary of Conclusions 177
7.2 Interdisciplinary Visual Analytics Research 180
7.3 Future Work 181

List of Abbreviations 183

List of Figures 185

List of Tables 195

Bibliography 197

Chapter 1

Introduction

Contents

1.1 Need for Visual Analytics in Uncertainty-aware Data Exploration 1

1.2 Dissertation Structure 5

1.3 Contributions of this Dissertation 7

1.4 Publications 8

1.1 Need for Visual Analytics in Uncertainty-aware Data Exploration

Today, a plethora of *spatial* and *spatio-temporal data* exist and come in various types (e.g., numeric, text, image, video, or audio data). They are available to users of many domains for intelligent analysis. Such data can serve as valuable sources to researchers for the investigation of phenomena in their spatial and temporal context (e.g., tracking of moving objects, impacts of hazards, global stock market finances, spread of diseases, or the occurrence of events). Spatio-temporal data, based on how they are collected, can be classified into *authoritative* and *non-authoritative*. Authoritative data are collected by domain experts, in place of gate keepers, following standardised data collection procedures (e.g., official surveyors). Non-authoritative data however are collected by *volunteers*, who do not necessarily have deep domain expertise. They do not follow quality management procedures and may use a variety of methods to collect data. Data collected by volunteers is known as Volunteered Geographic Information (VGI) (Goodchild, 2007).

The spatial dimension is not always straight forward in data. Depending on the kind of data, the spatial reference may be explicit (e.g., the Times Square mapped in

OpenStreetMap[1] with geographic coordinates) or implicit (e.g., in a Twitter[2] tweet: "I'm going to see the Lady Gaga concert in *New York*"). Similarly, the nature of volunteering can be distinguished. I.e., on some platforms such as OpenStreetMap the user explicitly contributes for the sake of data collecting, while on other platforms, the collection of data through the user is only a bi-product of his interactions (e.g., on social media). These categories are illustrated in Figure 1.1.

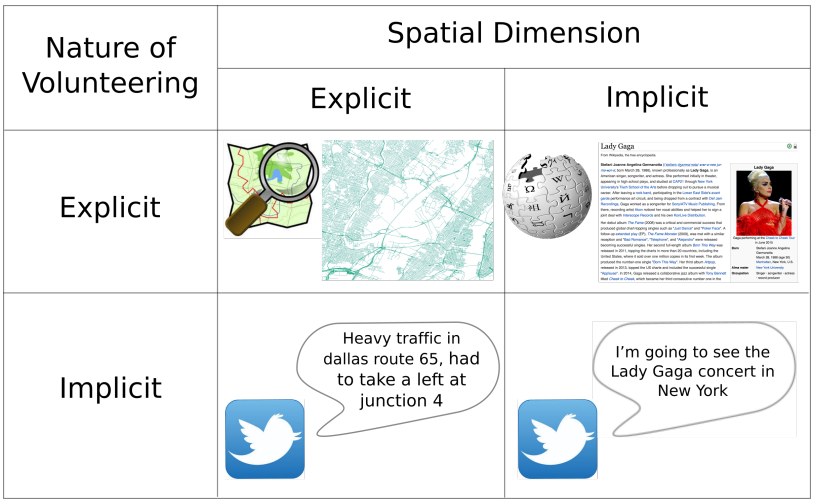


Figure 1.1: Classification of VGI based on the nature of volunteering and the spatial dimension (figure is adapted from M. Craglia (2012)).

Spatial data are also inherently uncertain. According to the Joint Committee for Guides in Metrology (JCGM), *uncertainty* determines the discrepancy between the measured value of an object and the true value of that object (JCGM, 2008). Works such as Beard et al. (1991); Buttenfield and Beard (1994); Thomson et al. (2005) define uncertainty in the context of spatio-temporal data as a collection of elements such as positional accuracy, thematic accuracy, or completeness (described in detail in Section 2.1.1). This uncertainty can be due to various reasons, e.g., imprecision of GPS (Global Positioning Systems) devises, inaccurate geodetic datums, lack of expertise of data collectors, and unwanted omissions or commissions of data. When data is collected by volunteers, elements such as the credibility or the trustworthiness of the volunteer is verified to assess the uncertainty in the data. For the analysis of spatio-temporal data, it is imperative to be aware of uncertainties in data in order to reduce errors in the analysis outcomes. Figure 1.2 shows an example of uncertainty depiction in the weather segment of the German daily news channel, Tagesschau. In

[1] `http://www.openstreetmap.org`
[2] `https://twitter.com`

this figure a forecast for the temperature in Frankfurt is shown through a white time series visualisation, and the uncertainty of the temperature data is shown through the grey colour interval drawn around the time series visualisation. As we can see, uncertainty increases with the forecast.

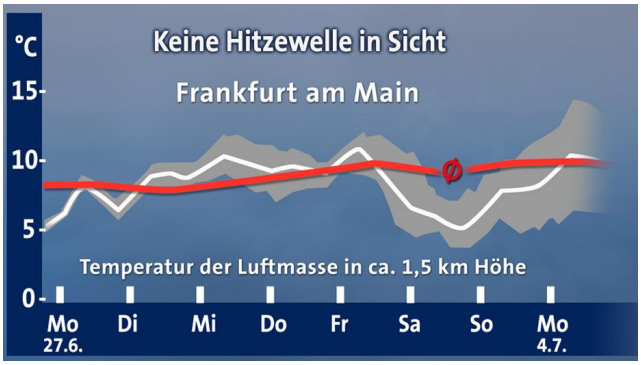


Figure 1.2: An example of uncertainty depiction in the German daily news channel Tagesschau. (Source: `wetter.tagesschau.de`).

In addition to the inherent uncertainties in data, more errors are introduced during the analysis process when the user selects inappropriate methods, data transformations, models, or visualisations. An example for errors caused through inappropriate visualisations is shown in Figure 1.3. This visualisation was used by the National Hurricane Center[3] of the US National Oceanic and Atmospheric Administration to depict the movement of a hurricane using a black trajectory. At first, viewers thought the white conic visualisation around the trajectory represents the *size* of the hurricane-causing more uncertainties. However, the cone actually represents the uncertainty of the hurricane forecast. Soon after this visualisation was released, the National Hurricane Center had to add a note (at the top of the Figure 1.3) stating that the conic visualisation represents the probable path of the storm and not its size.

The examples above demonstrate that the complexity of spatio-temporal data and their inherent uncertainties demand effective means to *explore* these data. It is crucial to sufficiently acknowledge the associated uncertainties. Often, static visualisations of data are not sufficiently drawing attention to uncertainties, as seen e.g. in Figure 1.3. As a result, users cannot grasp the uncertainty in static (or also animated) visualisations. This can be prevented through a stronger involvement of the user in the data exploration task. In the figure above, if the user was more involved, she/he could have chosen a more suitable metaphor for the uncertainty visualisation. This general concept is followed in *Visual Analytics*.

[3] `http://www.nhc.noaa.gov/aboutcone.shtml`

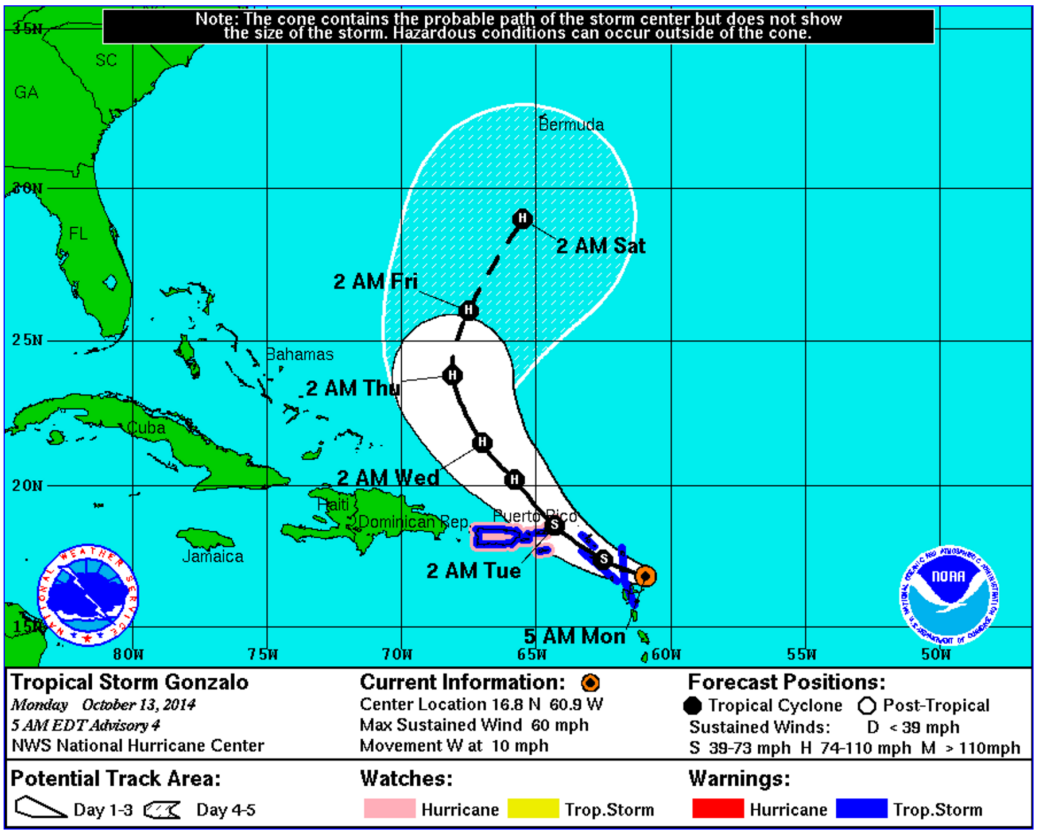


Figure 1.3: An example for errors caused through inappropriate visualisations used for uncertainty depiction. (Source: `http://www.nhc.noaa.gov/aboutcone.shtml`).

Visual analytics methodologies can enable thorough analysis of data by creating a synergy between the user, the automatic analysis methods, and effective visualisations (Keim et al., 2008). By creating an *exploration loop*, a *verification loop*, and a *knowledge generation loop* Sacha et al. (2014a), the user interacts with the system through actions that trigger the system to generate responses or feedback that users can verify though steering the interactions with the system and thereby arrive at insightful knowledge of the data at hand.

Existing visual analytics approaches address workflows to determine the role of uncertainty in data. However, as uncertainty keeps *propagating* throughout an analysis system we need frameworks that extend the current state of the art to explore the role of uncertainty in data model building, model usage, visual mappings, visualisations, as well as model-visualisation couplings, in order to minimise the errors in system outputs. User interaction with data, models, and visualisations promotes awareness of the various inherited and propagated uncertainties. Through such an iterative process of exploring, deriving findings, verifying those findings, and gaining insights into data,

the user has the ability to reduce errors caused by uncertainties in a system (Section 2.2).

Derived from these challenges, the goal of this dissertation is to develop a conceptual framework that incorporates and extends the role of uncertainty in the visual analytics knowledge generation process. Based on this framework, key approaches are developed for visually exploring and analysing the context-based nature of uncertainties in four essential types of data: image, text, location, and numerical data.

These approaches follow the workflow indicated in Figure 1.4.

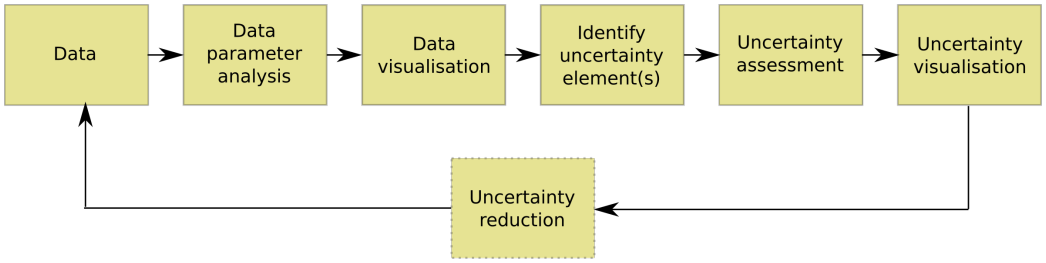


Figure 1.4: Work flow of the approaches followed in the subsequent chapters.

We start by analysing the data first with a specific task in mind (step 2 of Figure 1.4), and visualising the data to aid the analysis process (step 3 of Figure 1.4). This data parameter analysis step leads us to identify the element(s) of uncertainty present in the data, specifically the uncertainty element(s) that influences the analysis process (step 4 of Figure 1.4). This uncertainty element(s) is then assessed following an appropriate method (step 5 of Figure 1.4). The uncertainties are visualised in a final step along with the data (step 6 of Figure 1.5), before the process is reiterated. Throughout the reiteration process uncertainties are reduced (where applicable) following appropriate methods.

The next section describes how achieving the goal of this dissertation is approached through the structuring of this dissertation. Section 1.3 provides a summary of contributions. A list of all publications where parts of this dissertation appeared is given in Section 1.4.

1.2 Dissertation Structure

As outlined in Figure 1.5, this dissertation is organised as follows:

Chapter 2 begins with an introduction to *uncertainty* in spatio-temporal data, and an emphasis on the importance of being aware of various kinds of uncertainties that underlie data. Followed by a review of the existing state of the art on visually

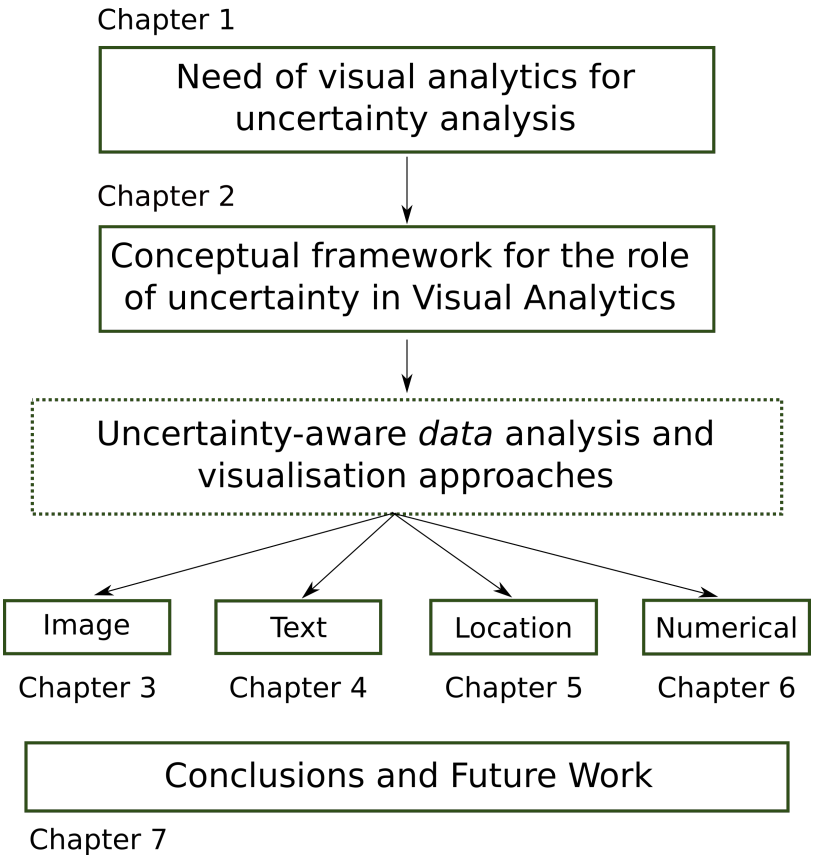


Figure 1.5: Structure of this dissertation.

communicating uncertainties, this chapter identifies key gaps in the existing work that inhibits the users from fully exploring uncertainty in their data. This can be bridged by the use of visual analytics. Therefore, a conceptual framework has been developed that defines the role of uncertainty in data, models, visualisations, and visualisation-model couplings. As a result several guidelines have been formulated that are intended to advise users in exploring uncertainty and thereby reducing errors in there findings throughout the visual analytics knowledge generation process.

This conceptual framework together with the defined guidelines act as the foundation for the following chapters. Chapter 3 - 6 investigate visual analytics approaches of uncertainty analysis for different data types.

Chapter 3 introduces an approach that relies on the local horizon of observers (visibility from an observer point) to assess the positional accuracy and the credibility of *image-based VGI*.

Chapter 4 explores uncertainty analysis approaches for movement trajectories that are extracted from implicitly referenced *text-based VGI*. By identifying several geographic- and content-based characteristics, and interestingness-based filtration

processes, the accuracy of trajectory extraction from text data can be improved.

Chapter 5 investigates space-time prisms to analyse the uncertainty in movement trajectories that are approximated by interpolating *geographical coordinate data* of antennas. Furthermore, this chapter identifies an approach to reduce positional uncertainties caused by antenna jumps in this type of data, and explores how such movement trajectories of humans can be utilised to determine patterns in the surrounding urban environment.

While the chapters above deal with one-dimension of uncertainty at a time in any given data type, Chapter 6 introduces several visual designs for presenting bi-dimensional uncertainties in *numerical data*. Further, a thorough evaluation process for these designs is included to select the most effective candidates for a given application scenario.

Chapter 7 draws conclusions from the research conducted in the chapters above, highlights the interdisciplinary research carried out, the lessons learned, and provides an outlook to the future perspectives of this work.

1.3 Contributions of this Dissertation

- *Framework for identifying the role of uncertainty within the visual analytics knowledge generation process.* Uncertainties are inherited and propagated throughout each component within a system. With the establishment of these distinct forms of uncertainty, causes have been identified for uncertainty inheritance in data, and for uncertainty propagation from thereon to data models, data visualisations, and model-visualisation couplings. Thirteen guidelines have been identified for acknowledging and assessing these uncertainties within visual analytics. Further, suggestions are made on how to reduce errors in the knowledge outcomes by dealing with the inherited and propagated uncertainties in a system. This framework acts as a foundation for exploring spatio-temporal data under the influence of various uncertainties in the forthcoming chapters. (Senaratne et al., 2017a; Sacha et al., 2016, 2014b). (Appeared in Chapter 2).

- *A novel method to assess the uncertainty of image-based VGI.* The standard viewshed analysis method has been modified to introduce the reverse-viewshed analysis to assess the positional accuracy of image-based VGI. This method relies on the line of sight (horizon of visibility) between observers who take photographs and the target that the photographs are labelled as. For each photograph the line of sight is calculated based on the surface elevation data that exists between

the observer and a target. Further based on these assessments of positional accuracy for images, image and user metadata are derived as suggestions for credibility indicators. (Senaratne et al., 2013a,b). (Appeared in Chapter 3).

- *A novel method to derive movement trajectories and improve the accuracy of movement trajectory extraction from text-based VGI.* Episodic sequential hotspots and the various drifts that are identified, e.g., through sentiments or content, are introduced as an approach for context-based movement trajectory extraction from implicitly referenced text-based VGI. A new grouping strategy based on the geographic and content characteristics of Tweets has been identified to improve the accuracy of trajectory extraction from these text-based VGI. To complement this, an interestingness-based trajectory ranking approach is introduced that further improves the accuracy of extracted trajectories for the task at hand. (Senaratne et al., 2014a, 2016, under review). (Appeared in Chapter 4).

- *A novel method for extracting movement trajectories from mobile communication data, a visual analysis of spatio-temporal patterns based on movement data, and a method for uncertainty analysis in mobile communication-based movement trajectories.* A method that relies on the temporal patterns of Internet usage is introduced to extract movement patterns of users from location-based telecommunication data. Uncertain markers are introduced to the space-time prism to volumetrically analyse the positional uncertainties in the extracted movement trajectories. This approach further extends to reduce uncertainties in movement trajectories by detecting antenna jumps. An encompassing visual analytics approach is identified that explores patterns in urban environments based on user movements. (Senaratne et al., 2017b). (Appeared in Chapter 5)

- *The design and evaluation of a glyph visualisation for bi-dimensional uncertainties.* Several glyph visualisations are designed and evaluated to visualise bi-dimensional uncertainties in one view. (Senaratne et al., 2014b). (Appeared in Chapter 6).

1.4 Publications

In the following, publications are listed where parts of this dissertation were published in and what the contribution of the author were:

1. H. Senaratne, A. Mobasheri, A. Loai Ali, C. Capineri and M. Haklay. **A Review of Volunteered Geographic Information Quality Assessment Methods**. *International Journal of Geographical Information Science, Taylor & Francis, 2016.*

   This publication is a result of a collaboration that I initiated at the VGI Vespucci summer school[4]. In this publication, my role as the first author was driving the entire literature review and writing a majority of the text in the paper. My contribution was primarily defining a typology for source uncertainty assessment methods within image-, text-, and map-based VGI. Furthermore, I have extended the classification by Goodchild and Li (2012) by defining *data mining* as an approach for source uncertainty assessment within these three types of VGI. A. Mobasheri and A. Loai Ali contributed with a review of most of the map-based VGI and identified some future improvements for uncertainty analysis in map-based VGI. Since I wrote the paper, the text has been taken into this dissertation with only slight modifications. C. Capineri and M. Haklay advised us throughout the two years of working on this publication.

2. H. Senaratne, A. Bröring and T. Schreck. **Using Reverse Viewshed Analysis to Assess the Location Correctness of Visually Generated VGI**. *Transactions in GIS, Wiley-Blackwell, 17(3), 369-386, 2013.*

   Refer to the contribution statement below.

3. H. Senaratne, A. Bröring and T. Schreck. **Assessing the Credibility of VGI Contributors based on Metadata and Reverse Viewshed Analysis - An experiment with Geotagged Flickr Images.** *Proceedings of the 16th AGILE International Conference on Geographic Information Science, 2013.*

   Both publications above were written by myself, and my contributions as the first author were the definition and implementation of the reverse-viewshed analysis as an uncertainty assessment technique for image-based VGI. A. Bröring and T. Schreck played a significant role in advising, revising, and thereby moulding the paper to its successful publication. Since I wrote the paper, the text has been taken into this dissertation with only slight modifications.

4. H. Senaratne, A. Bröring, T. Schreck and D. Lehle. **Moving on Twitter: Using Episodic Hotspot and Drift Analysis to Detect and Characterise**

[4] `http://www.vespucci.org/history/2014`

Spatial Trajectories. *In Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. (**Best student paper award**) ACM, 2014.*

As the first author in this publication, I took lead in writing the paper. My contributions were defining the systematic framework for detecting movement trajectories based on the episodic sequential hotspots in Twitter data, and for detecting drifts through sentiments and content in the data. D. Lehle led the implementation of the approach, and T. Schreck supervised this research by giving insightful feedback and improving the paper through revisions. A. Bröring reviewed the related research and overlooked the implementation. This paper was awarded the best student paper sponsored by Twitter. Since I wrote the paper, the text has been taken into this dissertation with only slight modifications. D. Lehle fulfilled his Master Thesis based on this work. The reference to his thesis is as follows: *D. Lehle, Trajectory-based Visual Analysis and Ranking of Twitter Data, Master's thesis (2015), University of Konstanz, Konstanz, Germany.*

5. H. Senaratne, D. Lehle and T. Schreck. **MovingOnTwitter2: Detection, Characterisation, and Interest-based Ranking of Conversation Trajectories**. *Under review for ACM Transactions on Spatial Algorithms and Systems (ACM TSAS), 2016*

   This paper was an extension of the above paper. As the first author, I took lead in writing the paper. My contributions were defining approaches together with D. Lehle for identifying geospatial- and content-based characteristics to derive movement trajectories of conversations on text-based VGI, as well as formalisation of the feature-based trajectory ranking approach to improve the accuracy of trajectory extraction. D. Lehle led the implementation of the defined approaches, and T. Schreck supervised this research by giving insightful feedback and improving the paper through revisions. Since I wrote the paper, the text has been taken into this dissertation with only slight modifications. D. Lehle fulfilled his Master Thesis based on this work. The reference to his thesis is as follows: *D. Lehle, Trajectory-based Visual Analysis and Ranking of Twitter Data, Master's thesis (2015), University of Konstanz, Konstanz, Germany.*

6. H. Senaratne, M. Mueller, M. Behrisch, F. Lalanne, J. Bustos, J. Schneidewind, D. Keim, T. Schreck. **Urban Mobility Analysis with Mobile Network Data: A visual Analytics Approach**. *Accepted for publication in IEEE Transactions on Intelligent Transportation Systems, DOI: 10.1109/TITS.2017.2727281, 2017.*

As the first author in this publication, I took lead in writing the paper. My contribution within this publication was defining the approach together with M. Mueller for extracting movement trajectories from location-based mobile communication data, exploring spatio-temporal patterns through the derived movements, and assessing the uncertainty of the movement trajectories derived from location-based mobile communication data. M. Mueller further led the implementation of the defined approaches. M. Behrish and T. Schreck defined the approach for temporal change detection through mobile communications, however, this section in the publication is not included in this dissertation. T. Schreck further supervised the entire process with his feedback and helped in structuring and revising the paper. F. Lalanne and J. Bustos provided the dataset and took part in many discussions with their insights to the Chilliean-based dataset. They also wrote part of the related work section. J. Schneidewind was the expert that we referred the paper to, and he gave us valuable expert feedback to the approach and the implemented tool, which was incorporated into the paper. D. Keim advised us with insightful feedback. Since I wrote most sections of the paper, the text that I have written have been taken into this dissertation with only slight modifications. M. Mueller fulfilled his Master Thesis based on this work. The reference to his thesis is as follows: *M. Mueller, Visual Mobile Network Data Analysis: Spatiotemporal Change and Uncertainty Perspectives, Master's thesis (2015), University of Konstanz, Konstanz, Germany.*

7. H. Senaratne, S. Mittelstaedt, C. Jacob and T. Schreck. **Uncertainty Visualization for Crisis Management in Smart Grid Environments**. *Eighth International Conference on Geographic Information Science (GIScience) Workshop on Visually Supported Reasoning with Uncertainty, 2014.*

   As the first author in this publication, I took lead in writing the paper. My contributions in the paper were guiding the comparison between the uncertainty assessment methods, together with S. Mittelstaedt designing the glyph visualisations, defining the step-by-step approach for the user study, and overseeing the implementations. S. Mittelstaedt provided his expert knowledge on the data environment, wrote on the related work section, and helped to revise the paper. C. Jacob led the implementation and the evaluation of the results. T. Schreck supervised this research. Since I wrote most sections of the paper, the text that I have written have been taken into this dissertation with only slight modifications. C. Jacob fulfilled her Bachelor Thesis based on this work. The reference to her

thesis is as follows: *C. Jacob, Visual Analytics for Handling Uncertainty within Smart Grid Environments, Bachelor's thesis (2014), University of Konstanz, Konstanz, Germany.*

8. D. Sacha, H. Senaratne, B. C. Kwon and D. A. Keim. **Uncertainty Propagation and Trust Building in Visual Analytics**. *IEEE VIS 2014 - Provenance for Sensemaking Workshop (poster paper), 2014.*

   Refer to the contribution statement below.

9. D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis and D. A. Keim. **The Role of Uncertainty, Awareness, and Trust in Visual Analytics**. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of the Visual Analytics Science and Technology), 22(01), 240-249, 2016.*

   In both publications above, my contribution as the second author was defining the role of uncertainty within the visual analytics knowledge generation process. In doing so, I have derived 13 guidelines for uncertainty analysis within visual analytics. D. Sacha, derived guidelines for assessing the trust on the human side of the visual analytics knowledge generation process. G. Ellis worked mainly on deriving awareness perspectives, and how this affects the knowledge generation process. B.C. Kwon and D. Keim advised us throughout the paper with insightful feedback and supervised this research. Since I wrote the uncertainty sections in the paper, this text has been taken into this dissertation with only slight modifications.

Publications I have contributed to but not included in this dissertation are as follows:

1. H. Senaratne and L. Gerharz. **An Assessment and Categorisation of Quantitative Uncertainty Visualisation Methods for Geospatial Data**. *Proceedings of the 14th AGILE International Conference on Geographic Information Science (Poster Paper.* ***Best poster award****), 2011.*

2. H. Senaratne, L. Gerharz, E. Pebesma and A. Schwering. **Usability of Spatio-temporal Uncertainty Visualisation Methods**. *Proceedings of the 15th AGILE International Conference on Geographic Information Science - Bridging the Geographic Information Sciences, Springer, pages 3-23, 2012.*

3. D. Reusser, H. Senaratne, L. Gerharz, T. Sterzel, T. Nocke and M. Wrobel. **User Preferences for the Presentation of Uncertainties on Web Platforms for Climate Change Information.** *EGU2012. In proceeding of Geophysical Research Abstracts, 2012.*

4. H. Senaratne, D. Reusser and T. Schreck. **Usability of Uncertainty Visualisation Methods: A Comparison between Different User Groups**. *GeoViz Hamburg 2013: Workshop on Interactive Maps that Help People Think, 2013.*

5. D. Jäckle, H. Senaratne, J. Buchmüller and D. A. Keim. **Integrated Spatial Uncertainty Visualization using Off-screen Aggregation.** *EuroVis Workshop on Visual Analytics (EuroVA), The Eurographics Association, 2015.*

6. J. Buchmüller, M. Stein, A. Jäger, S. Schmidt, H. Senaratne and H. Janetzko. **Fusing Events, Tasks and Spatial Awareness in an Ambient-Enabled Work Environment**. *IEEE Conference on Visual Analytics Science and Technology (VAST Challenge 2016 MC1), to appear, 2016.*

Chapter 2

The Role of Uncertainty in Spatio-temporal Data Analysis

Contents

2.1 Background & Related Work 16

2.1.1 Uncertainty Measurements & Indicators 17

2.1.2 Uncertainty Visualisation 19

2.1.3 Taxonomies for Uncertainty Visualisation 28

2.2 A Framework and Guidelines for Uncertainty Inheritance and Propagation in Visual Analytics 30

2.2.1 Guidelines to Assess Source Uncertainty in Spatio-temporal Data 32

2.2.2 Guidelines to Assess Propagated Uncertainties in Spatio-temporal Data 54

2.2.3 Guidelines to Aggregate Uncertainties 60

2.2.4 Guidelines to Visualise Uncertainty Information 61

2.2.5 Guidelines to Enable Interactive Uncertainty Exploration . 62

2.3 Discussion & Future Work 62

2.4 Conclusions 63

In this chapter a conceptual framework is introduced that defines the *role of uncertainty* in the visual analytics knowledge generation process. All kinds of data are *inherent* with uncertainty. We review the *definitions* of uncertainty within the context of spatio-temporal data, the *elements* of uncertainty that vary based on the *type of data*, as well as the existing visual metaphors that have been introduced to *visually communicate* the uncertain elements in data. Cognitive limitations, and the inability to contextually analyse static, dynamic, or animated uncertainty visualisations create a gap in the state of the art. Furthermore, many existing works have assumed that

uncertainty is a constant in the data, and therefore have neglected the fact that uncertainty keeps *propagating* through data, models, visualisations, and model-visualisation couplings. A lack of approaches to acknowledge these propagated uncertainties further increase the gap. This chapter aims at filling these gaps by introducing *guidelines* for acknowledging and assessing the inherited source uncertainties and the propagated uncertainties within a visual analytics knowledge generation process, thereby completing the circle of uncertainty analysis in visual analytics.

This chapter unfolds as follows: In Section 2.1 we review the basic concepts of uncertainty and uncertainty visualisation for the analysis of spatio-temporal data. In doing so, Section 2.1.1 describes the fundamental elements of uncertainty that are used for assessing the discrepancies in data. Section 2.1.2 describes the three main dichotomous categories of visualisation techniques for uncertainty, followed by taxonomies of uncertainty visualisations from the literature. In Section 2.2 we present a framework for defining the role of uncertainty within a visual analytics knowledge generation process. As such we present guidelines for assessing source uncertainties and propagated uncertainties in Sections 2.2.1 and 2.2.2. Guidelines to aggregate uncertainties, visualise uncertainties, and interactively explore uncertainties are presented in Sections 2.2.3, 2.2.4, and 2.2.5.

The contents of this chapter are based on the publications Senaratne et al. (2017a)[1], and Sacha et al. (2014b, 2016)[2].

2.1 Background & Related Work

Uncertainty, generally known as the *state of not knowing*, is attributed to the discrepancy between a measured value of an object and the true value of that object (JCGM, 2008). According to Griethe and Schumann (2006), elements of uncertainty are *errors*, *imprecision*, *accuracy*, *lineage*, *subjectivity*, *non-specificity*, and *noise*. Similarly, many works have classified uncertainty into different measurements and indicators.

[1]Appears in Sections 2.1.1 and 2.2.1. This work is a result of a collaboration with A. Mobasheri from the University of Heidelberg, A. Ali from the University of Bremen, C. Capineri from the University of Sienna, and M. Haklay from the University College London. My contribution as the first author within this collaboration was primarily defining a typology for source uncertainty assessment methods within image- and text-based volunteered geographic information (VGI), as well as reviewing methods for map-based VGI. Furthermore, I extended the classification of Goodchild and Li (2012) by defining *data mining* as an approach for source uncertainty assessment within these three types of VGI.

[2]Appears in Section 2.2.2. This work is a result of an intense collaboration with D. Sacha, B. Kwon, G. Ellis, and D.A. Keim. My contribution as the second author within this collaboration was defining the role of uncertainty within the visual analytics knowledge generation process, thereby deriving guidelines for uncertainty analysis within visual analytics

Uncertainties vary mainly depending on the *data type* and the *context* of application domain. For example, *topological consistency* as an uncertainty measurement for street network data. Throughout this dissertation we will refer to any one of these elements as uncertainty.

2.1.1 Uncertainty Measurements & Indicators

Uncertainty of spatio-temporal data can be described by uncertainty *measures* and uncertainty *indicators* (Antoniou and Skopeliti, 2015). Uncertainty measures, mainly adhering to the ISO principles and guidelines refer to those elements that can be used to ascertain the discrepancy between the collected spatio-temporal data and the ground truth (e.g., completeness of data), mainly by comparing to authoritative data. When authoritative data is no longer usable for comparisons, and the established measures become no longer adequate to assess the uncertainty of data, researchers have explored more intrinsic ways to assess the uncertainty by looking into other proxies for uncertainty measures (in most cases, this is relevant for volunteered geographic information - VGI). These are called uncertainty indicators, that rely on various participation biases, data contributor expertise or the lack of it, background, etc., that influence the uncertainty of spatio-temporal data, but cannot be directly measured (Antoniou and Skopeliti, 2015). In the following these uncertainty measures and indicators are described in detail.

Uncertainty Measures

ISO[3] (International Standardisation Organisation) defined geographic information quality as *the totality of characteristics of a product that bear on its ability to satisfy stated and implied needs.* A lack of it therefore creates uncertainty. ISO/TC 211[4] (Technical Committee) developed a set of international standards that define the measures of geographic information uncertainty (standard 19138, as part of the meata-data standard 19115). These quantitative uncertainty measures are: *completeness*, *consistency*, *positional accuracy*, *temporal accuracy*, and *thematic accuracy.*

Completeness describes the relationship between the represented objects and their conceptualisations. This can be measured as the absence of data (errors of omission) and presence of excess data (errors of commission). Consistency is the coherence in the data structures of the digitised spatial data. The errors resulting from the lack of it are indicated by (i) conceptual consistency, (ii) domain consistency, (iii) format consistency,

[3] `http://www.iso.org/iso/home/standards.htm`
[4] `http://www.isotc211.org/`

and (iv) topological consistency. Accuracy refers to the degree of closeness between a measurement of a quantity and the accepted true value of that quantity, and it is in the form of positional accuracy, temporal accuracy and thematic accuracy. Positional accuracy is indicated by (i) absolute or external accuracy, (ii) relative or internal accuracy, (iii) gridded data position accuracy. Thematic accuracy is indicated by (i) classification correctness, (ii) non-quantitative attribute correctness, (iii) quantitative attribute accuracy. In both cases, the discrepancies can be numerically estimated. Temporal accuracy is indicated by (i) accuracy of a time measurement: correctness of the temporal references of an item, (ii) temporal consistency: correctness of ordered events or sequences, (iii) temporal validity: validity of data with regard to time.

Uncertainty Indicators

As part of the ISO standards, geographic information uncertainty can be further assessed through qualitative uncertainty indicators such as the *purpose*, *usage*, and *lineage*. These indicators are mainly used to express the uncertainty overview for the data. Purpose describes the intended usage of the dataset. Usage describes the application(s) in which the dataset has been utilised. Lineage describes the history of a dataset from collection, acquisition to compilation and derivation to its form at the time of use (Van Oort and Bregt, 2005; Hoyle, 2001; Guinée, 2002).

In addition, where ISO standardised measures and indicators are not applicable, we have found in the literature more abstract uncertainty indicators to imply the uncertainty of mainly non-authoritative data -VGI. These are: *trustworthiness*, *credibility*, *text content quality*, *vagueness*, *local knowledge*, *experience*, *recognition*, *reputation*.

Trustworthiness is a receiver judgment based on subjective characteristics such as reliability or trust (good ratings on the creations, and the higher frequency of usage of these creations indicate this trustworthiness) (Flanagin and Metzger, 2008).

In assessing the credibility of VGI, the source of information plays a crucial role, as it is what credibility is primarily based upon. However, this is not straight forward. Due to the non-authoritative nature of VGI, the source may be unavailable, concealed, or missing (this is avoided by gatekeepers in authoritative data). Credibility was defined by Hovland et al. (1953) as the *believability of a source or message, which comprises primarily of two dimensions, the trustworthiness* (as explained above), *and expertise*. Expertise contains objective characteristics such as accuracy, authority, competence, or source credentials (Flanagin and Metzger, 2008). Therefore, in assessing the credibility of data as an uncertainty indicator one needs to consider factors that attribute to

the trustworthiness and expertise. Metadata about the origin of VGI can provide a foundation for the source credentials of VGI (Frew, 2007).

Text content quality (mostly applicable for text-based VGI) describes the uncertainty of text data by the use of text features such as the text length, structure, style, readability, revision history, topical similarity, the use of technical terminology etc.

Vagueness is the ambiguity with which the data is captured (e.g., vagueness caused by low resolutions) (De Longueville et al., 2010).

Local knowledge is the contributors' familiarity to the geographic surroundings that she/he is implicitly or explicitly mapping. Experience is the involvement of a contributor with the VGI platform that she/he contributes to. This can be expressed by the time that the contributor has been registered with the VGI portal, number of GPS tracks contributed (for example in OSM) or the number of features added and edited, or the amount of participation in online forums to discuss the data (Van Exel et al., 2010).

Recognition is the acknowledgement given to a contributor based on tokens achieved (for example in gamified VGI platforms), and the reviewing of their contributions among their peers (Van Exel et al., 2010).

Maué (2007) described reputation as a tool to ensure the validity of VGI. Reputation is assessed by, for example the history of past interactions that are happening between collaborators. Resnick et al. (2000) described contributors' abilities and dispositions as features where this reputation can be based upon. Maué (2007) further argue that similar to the eBay rating system[5], the created geographic features on various VGI platforms can be rated, tagged, discussed, and annotated, which affect the data contributor's reputation value.

Many methods have been developed in the state of the art to visually communicate these quantitative and qualitative uncertainty measures and indicators to the end users.

2.1.2 Uncertainty Visualisation

Visualisation of uncertainty in spatio-temporal data is a complex procedure, due to reasons such as the heterogeneity of data, spatial and temporal variation of data, different elements of uncertainty in data, and also the various abstract definitions of uncertainty that are adapted to suit the context of data usage (Gerharz et al., 2012). Various taxonomies have been built to support the user in identifying suitable visual

[5] `http://ebay.about.com/od/gettingstarted/a/gs_feed.htm`

variables to visualise the uncertainty elements under investigation. Many visualisation methods that vary from static, dynamic, to interactive in nature have been developed over the years in various spatio-temporal settings, to visually communicate the different elements of uncertainty.

MacEachren et al. (1998) asserted that importance should not only be given to the visual syntactic with which uncertainty measures and indicators are matched with visual variables, but also to the way data and uncertainties are linked and represented. As such, the works of MacEachren (1992) and Howard and MacEachren (1996) have identified three prominent dichotomous categories for uncertainty visualisation: *intrinsic/extrinsic* (w.r.t. situating data and uncertainty), *coincident/adjacent* (w.r.t. view organisation), and *static/dynamic* (w.r.t. to the interactive nature of the display). Most existing uncertainty visualisations focus on intrinsic (visual variable of the data is manipulated to represent uncertainty, such as colour transparency), coincident (data and uncertainty are integrated in one view), and static (no interaction with the display) techniques, while extrinsic (additional glyphs are used to visualise uncertainty) and adjacent (data and uncertainty on two adjacent views) techniques are being used in seldom (Kinkeldey et al., 2014). Dynamic techniques that involve user interaction in most cases, are sparse, as it is shown through studies such as by Senaratne et al. (2012). Such techniques require advance experience in spatial uncertainty analysis.

Vsualisation should primarily accomplish *detection*, *notice*, *identification*, and *quantification* of spatial data uncertainty, taking into account also legal and psychological connotations of these terms (Beard et al., 1991). Furthermore, Beard et al. (1991) emphasised the advantages of visualising spatial data uncertainty as, *speed of pattern recognition*, *motion detection*, *change detection*, *seeing the intangible* (e.g., in remote sensed data, molecular structures) etc. They also indicated that visualising meta data surpasses the *barriers of language*, *reduce cultural bias*, and *recover hidden structures in the patterns of data uncertainty.*

In the following sections, examples of the different dichotomous categories of uncertainty visualisation are presented.

Intrinsic Visualisations

The intrinsic visualisation technique has been adapted in two approaches in the state of the art and incorporates manipulating the visual variable to represent uncertainty. These two approaches are the use of colour models and focus metaphors.

Colour models are utilised in different methods to visualise the uncertainty in spatio-temporal data. Hengl et al. (2002); Cliburn et al. (2002); MacEachren et al. (2005)

showed how data is mapped to the colour hue and its uncertainty represented through the saturation of the colour hue, i.e. higher saturation indicates lower uncertainty. Cliburn et al. (2002) evaluated the use of colour saturation to depict uncertainty in a usability study against two other uncertainty visualisation methods that use colour transparency and glyphs to visualise uncertainties in their data (surplus and deficits in water). The participants were domain experts, usability experts, and decision makers. The results indicated that participants with less experience with sciences preferred the colour saturation method to visualise uncertainty, as it gave them a better impression at a first look without any complexities. Participants with a scientific background and decision makers however preferred to use colour transparency and glyphs to depict uncertainty respectively.

Hengl (2003) introduced a colour model which he called, whitening, where the colour hue is used to represent the data, and the saturation-intensity (whiteness) is used to represent the associated uncertainty. The amount of white colour, proportional to the uncertainty is mixed in with the hue which represents the prediction. A similar result is obtained by the technique of pixel mixing in which uncertainty is represented by adding amounts of white pixels proportional to the normalised prediction error (Hengl and Toomanian, 2006). An example is shown in Figure 2.1 where the data is represented by the colour hue and the uncertainty is represented by the saturation-intensity of colour (whiteness). Gerharz and Pebesma (2009) evaluated this method by assessing the usability in terms of preference of the method. The participants who were all from the Geoinformatics domain found it difficult to quantitatively interpret the associated uncertainty in the data as opposed to two other uncertainty visualisation methods: the adjacent maps (Figure 2.5) and exceedance probability mapping method (Figure 2.6).

Focus metaphors are based on the human perception of focused and non-focused (blurred) views. Uncertain data is depicted out of focus, making it less precisely visible, e.g. foggy. More certain data is depicted in focus, e.g. crisp boundaries. Another metaphor of the focus method is the opacity method where less uncertain data is seen less opaque and more uncertain data is more opaque (MacEachren et al., 2005; Prassni et al., 2010). This can also be used in reverse where uncertain data is shown more transparently (Drecki, 2002). Vullings et al. (2013) were one of the very few to visualise the incompleteness in geospatial data, using mainly fuzziness to depict the incompleteness in land-use data. Vullings et al. (2013) also evaluated these methods, and found the participants (who were stakeholders) preferred fuzzy boundaries to represent the incompleteness in the land-use data, over a noise lines method.

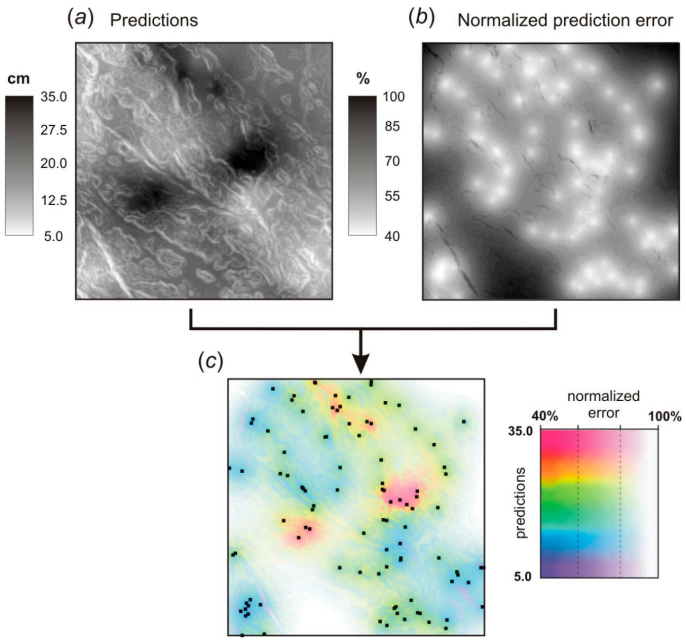


Figure 2.1: Whitening. Colour hue represents top soil thickness data and the saturated intensity represents the uncertainty of top soil thickness. Figure is taken from Hengl and Toomanian (2006).

Extrinsic Visualisations

In extrinsic approaches glyphs are used to visualise uncertainties in geospatial data, following the size, shape, orientation, colour visual variables. Pang (2001) demonstrated this in a use case where angular uncertainty of wind direction was depicted using glyphs in the form of arrow plots (Figure 2.2). The width of the arrow head represents the angular uncertainty and the magnitude of the vector field is depicted through the length of the arrow plot.

Cliburn et al. (2002) and Slocum et al. (2003) used line glyphs to visualise the uncertainty in surplus and deficits of water in different locations over Asia. The length of the line glyph is used to indicate the high and low uncertainty. Cliburn et al. (2002) evaluated the usability of glyphs in terms of user preference, in a study where users came from decision support, domain expertise and usability engineering. The results indicate that the participants who were from a decision support background preferred to use glyphs over colour saturation and colour transparency to visualise uncertainty.

MacEachren et al. (2012) carried out an experiment with thirty participants with a GIS background, on symbolisation of uncertainty. They first assessed the symbol intuitiveness, and secondly they assessed the task performance when multiple symbols appear on the display. They carried out experiments in two different settings, to

assess first, the comprehensibility (intuitiveness) of the methods, and second the performance of the users within the uncertainty visualisation methods depicting data and uncertainty using more than one symbol. They found in terms of intuitiveness that the visual variables fuzziness, location, value, and arrangement worked well for representing uncertainty, while size and transparency were found to be potentially usable. However, colour saturation was ranked low in usability, similar to some other findings that are reported in this work. MacEachren et al. (2012) further concluded that dominant perceptual symbols are more effective for pre-attentive tasks such as visual search, symbol comparison, visual aggregation, or region comparison.

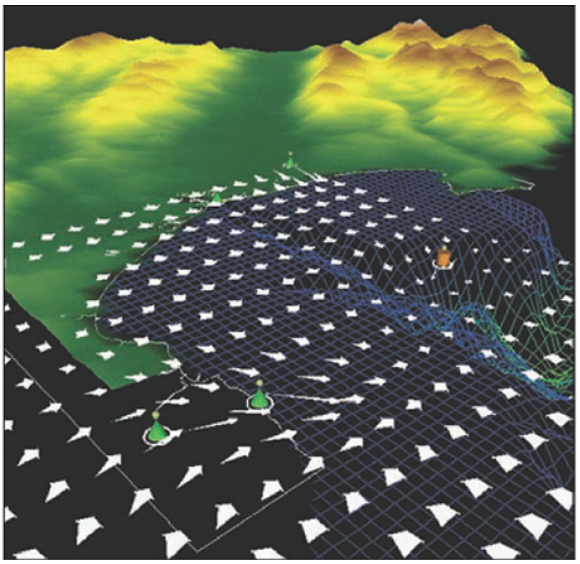


Figure 2.2: Uncertainty visualisation with glyphs. Wind velocity data is depicted through the length of the arrow plot and its angular uncertainty through width of the arrow head. Figure is taken from Pang (2001).

Coincident Visualisations

Coincident approaches are where the data and the uncertainty are integrated into one view. Two examples of this approach are presented in this section: contouring method and the error bars method.

In general, contour lines connect locations where a chosen attribute has a constant value (Longley, 2005). This technique is adapted to visualise uncertainty. In a multivariate mapping environment, contour lines of different colours are used to distinguish between different variables and their uncertainties with the intensity of colour. Similarly, contours of varying thicknesses are used to represent uncertainty in a coincident depiction where the size visual variable (thickness) is used in proportion

to the uncertainty in the data. Using this method to visualise the uncertainty in concentration of PM10 (particulate matter with a diameter of 10 µm or less) is shown in Figure 2.3. Likewise, positional uncertainty can be depicted through gap widths in the dots of dotted contour lines where higher uncertainty leads to wider gaps, as seen in the works of Dutton (1992); Allendes Osorio and Brodlie (2008).

The concept of contouring is also used in an animated environment as seen in the work of Fauerbach et al. (1996). Senaratne et al. (2012) evaluated the usability of this uncertainty visualisation technique with participants coming from different domains – GIS, map visualisation, statistics, decision support, urban planning. In terms of user performance and user preference this method ranked first in usability among a majority of the users. These users found it easy and convenient to distinguish the similar areas of high and low uncertainties by following the contour lines. Data and uncertainty being given in one coincident representation was also found to be easy to comprehend as compared to adjacent maps (Section 2.1.2), where the users need to shift their gaze between two adjacent maps to distinguish the associated uncertainties in different locations on the map. Similar results were also found in Fauerbach et al. (1996).

Error bars are a means to indicate uncertainties of data represented in charts and graphs, in proportion to the confidence intervals taken into consideration (Figure 2.4). Works of Tukey (1977); Cleveland and McGill (1986); Tufte and Graves-Morris (1983); Olston and Mackinlay (2002); Senaratne et al. (2012), discuss the use of error bars to communicate statistical uncertainty in the data. As seen in Figure 2.4, the ground level Ozone concentration on a given day for the city of Osnabrük in Germany is depicted in a time series visualisation, and the associated uncertainty of Ozone concentration at each hour is visualised through error bars. The height of the error bars represents the high and low uncertainties in the data. For each data hour, the associated Probability Distribution Function (PDF) graph is presented adjacent to it. Senaratne et al. (2012) evaluated the usability of this uncertainty visualisation method against adjacent maps (with colour saturation to show the uncertainty), contouring (with size of the contour lines to show the uncertainty), and exceedance probability mapping (with colour value to show the uncertainty) methods. Error bars method ranked third with higher preference from users but very low in user performance. Users suggested that this uncertainty visualisation method is most suitable for expert users who had a deeper knowledge in statistical data analysis.

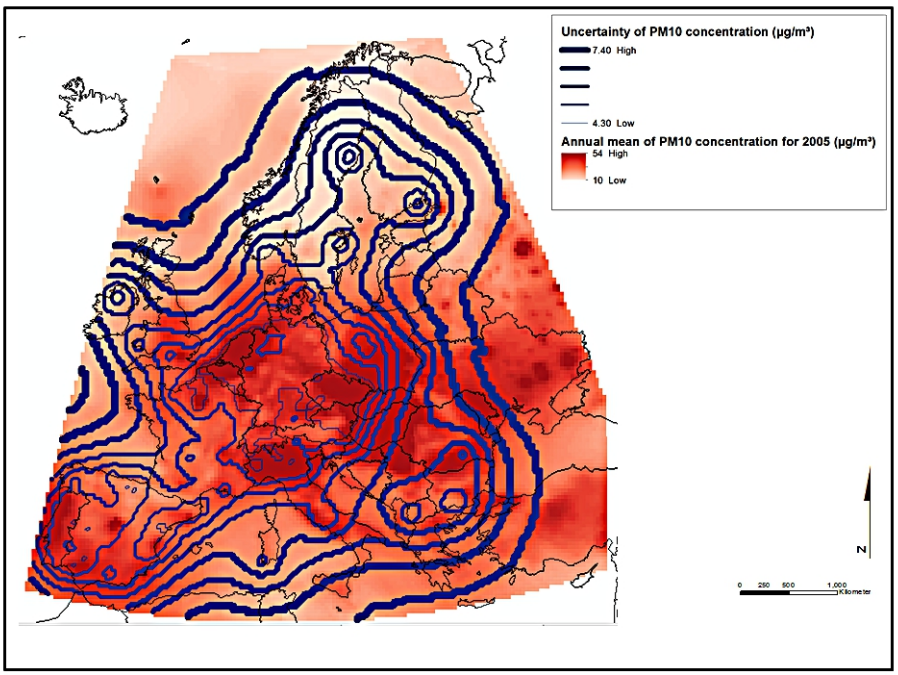


Figure 2.3: Uncertainty visualised through contouring method. PM10 concentration data is shown in the background with higher saturation of red corresponding to higher concentration and lower saturation corresponding to lower concentration. The uncertainty of PM10 concentration is shown in the foreground with thickening contour lines. Thicker contours correspond to higher uncertainty and thinner contours to lower uncertainty in PM10 concentration. Figure is taken from Senaratne et al. (2012).

Adjacent Visualisations

This technique presents data and data uncertainty on two separate maps adjacent to each other. Through comparing the two maps, the degree of uncertainty at different points can be comprehended. MacEachren et al. (1998) first introduced this technique by using colour saturation to represent the uncertainties, however, any other visual variable that supports the given data and their uncertainties can also be used following this technique. In the given example in Figure 2.5 high and low colour saturation is used to represent the high and low concentration of PM10 data for Europe (left), and high and low uncertainties of these concentration data (right). MacEachren et al. (1998) evaluated this technique against a visually separable coincident technique (the user is able to toggle between the data layer and the uncertainty layer) where the uncertainty is represented through a texture, and a second intrinsic method where the colour value (light to dark grey) is used to present uncertainty in health statistics data. A majority out of the 84 participants, preferred the coincident method over the rest, and the reasons being the flexibility and easiness to analyse data and uncertainty

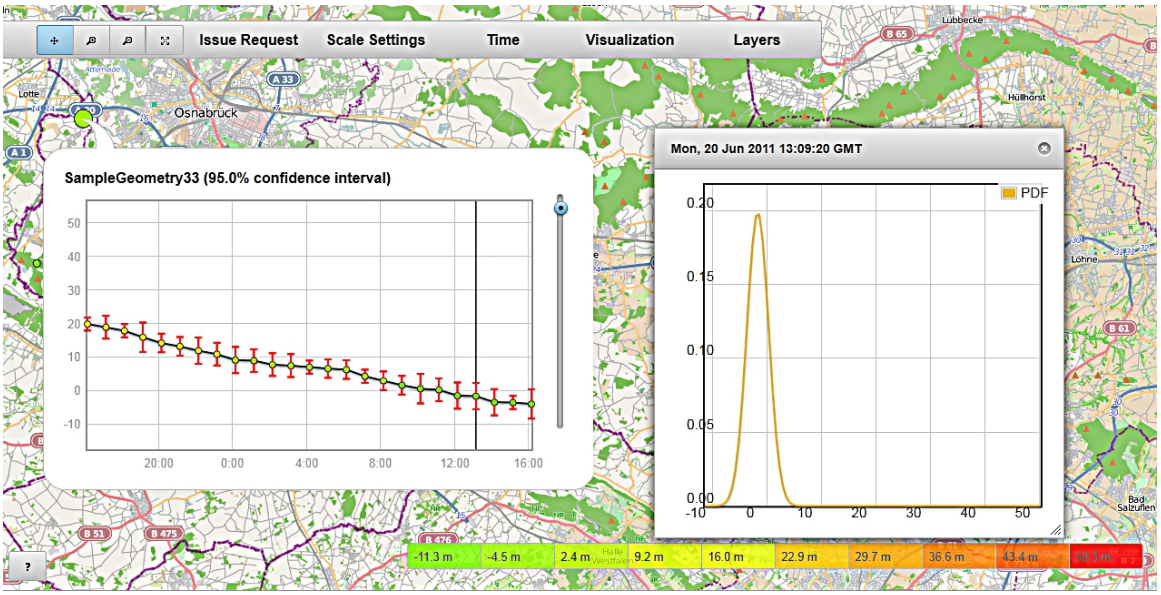


Figure 2.4: Uncertainty of ground-level Ozone concentration data for each hour is visualised through the error bars. The height of each error bar represents the amount of uncertainty associated for the given hour. Figure is taken from Senaratne et al. (2012).

separately by disregarding the textual overlay when needed, thus was said to be most useful for exploratory causes.

In another study by Aerts et al. (2003), 60% of the expert users and 83% of the novice users that took part preferred this adjacent maps method where uncertainty was represented through colour saturation, while the toggling method where data and uncertainty were presented in two different layers through colour saturation was least preferred due to the irritation that is caused by constant alternation of the layers.

Senaratne et al. (2012) found in their study, that the adjacent maps methods with colour saturation to depict data and uncertainty (Figure 2.5) was ranked higher in usability among novice users ,and was only the second choice among expert users. The participants of this study suggested that this method is comparatively easier to assess the uncertainty by comparing with the data adjacent to it, unlike other coincident visualisation methods such as the contouring method (Figure 2.3), that some participants found difficult to determine the uncertainty in comparison to the underlying data.

Static and Dynamic Visualisations

Visualisation methods such as those described above have been implemented in static and dynamic environments. Animation, which was a popular dynamic approach for uncertainty depiction, is a presentation technique which can be used for different

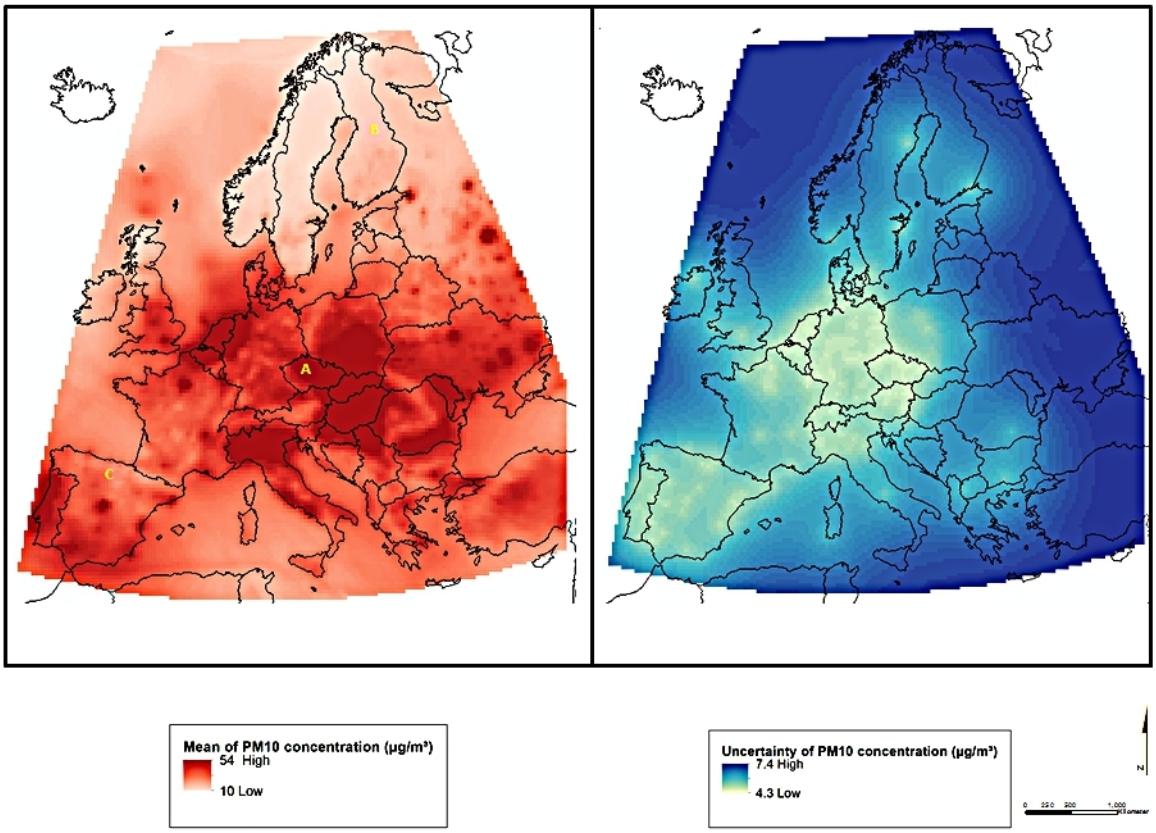


Figure 2.5: Adjacent Maps. Colour saturation is used to depict high/low PM10 values and high/low uncertainties of PM10. Figure is taken from Senaratne et al. (2012).

methods, and is the integral part of three particular visualisation methods in the state of the art.

Fisher (1993a) was one of the first to use animation to depict uncertainty through blinking pixels. The data in each grid cell of a map is represented through colour hue. This colour remains stable for pixels with less uncertain classifications in soil data, and changes continuously proportional to the uncertainty in the data creating a flickering environment for data with higher uncertainties. A study that was carried out by Evans (1997) showed that a majority of the 66 participants who took part in the study found the blinking pixels animation method to be comprehensible and useful. Ehlschlaeger et al. (1997) presented uncertainty through animating its different realisations to emphasise the underlying spatial uncertainties. Similarly, Kardos et al. (2006) introduced blinking Regions method, where census data is visualised though colour hue on one layer and its uncertainty is visualised through colour saturation on another layer and these two layers are then overlaid on top of one another and alternatively displayed.

The blinking effect is said to inform the user of the data and the underlying uncertainty alternatively. Kardos et al. (2006) further evaluated this method against eight other uncertainty visualisation methods with 34 participants who had an understanding in GIS. These other methods were: adjacent maps, texture overlay, focus metaphors (e.g., fuzziness, image sharpness), pixel mixture, sound, colour saturation and another animation technique. The results which were primarily based on the visual appeal, speed of comprehension, and overall effectiveness of the uncertainty visualisation methods indicated that participants performed better in blinking regions method in terms of speed of comprehension and overall effectiveness. However, this method was undesired by the participants in terms of visual appeal, as the constantly flickering display was found to be irritating. Another study that was carried out by Aerts et al. (2003) with expert and novice participants showed that method of alternating between different layers to highlight data uncertainty was least preferred by participants, and instead adjacent maps with colour saturation to depict uncertainty was highly preferred by all participants.

Also, within a dynamic environment Pebesma et al. (2007); Senaratne et al. (2012); Verstegen et al. (2012) used the open source *Aguila*[6] software as an interactive approach for exploratively analysing data uncertainty. The tool presents uncertainty of the data through the cumulative probability functions for each pixel or vector object. An example of visualising the uncertainty in PM10 concentration data with this tool is shown in Figure 2.6. Depending on the chosen quantile or threshold value, the map colour scale shows the associated value or probability (Figure 2.6, left). PDF (Probability Distribution Function) graphs are generated to show the exceedance probability of threshold values of PM10 concentrations (Figure 2.6, right). By analysing the PDF and the probability maps together, associated uncertainties of PM10 concentration at different locations over Europe can be estimated.

2.1.3 Taxonomies for Uncertainty Visualisation

Many of the challenges in uncertainty visualisation have been met with taxonomies that were developed to ease the user of having to go through multitudes of uncertainty visualisation methods to choose the appropriate methods according to their data and user requirements. Amongst many early discussions on uncertainty visualisation, MacEachren (1992); Beard et al. (1991); Fisher (1994); Van der Wel et al. (1994) are prominent examples in grounding the visualisation of uncertainty indicators and

[6] `http://pcraster.geo.uu.nl/projects/developments/aguila/`

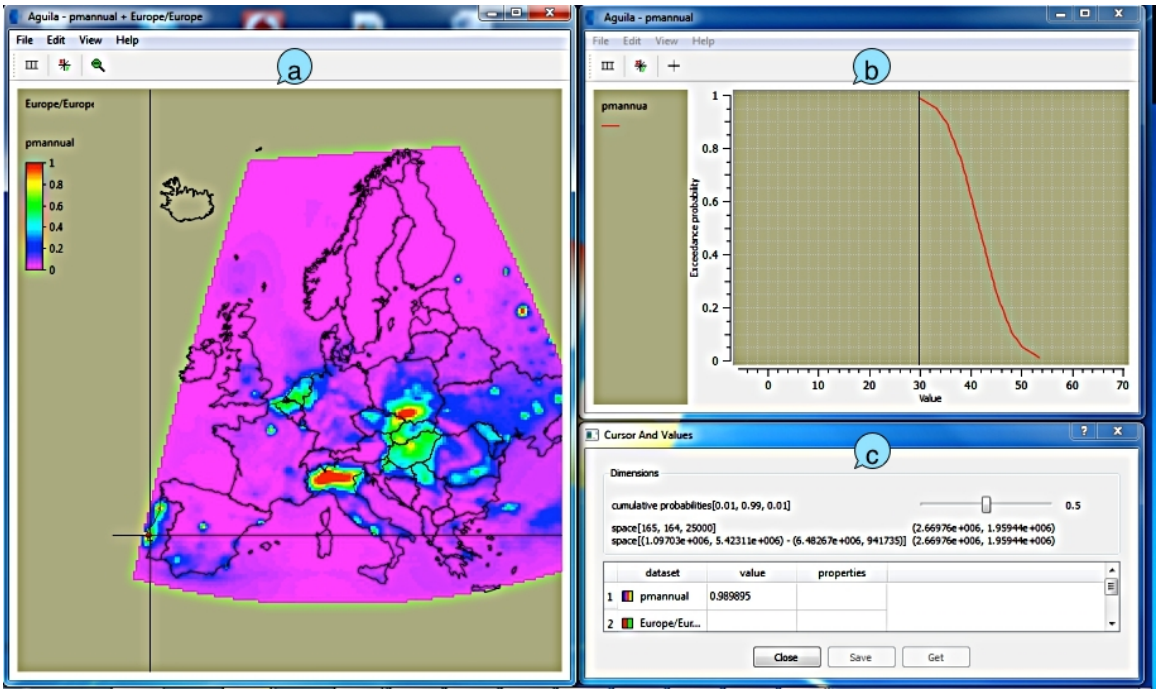


Figure 2.6: Uncertainty in PM10 concentration data for Europe is visualised through (a) a probability map with a rainbow colour scale, (b) PDF graph, and (c) the corresponding probability values. Figure is taken from Senaratne et al. (2012).

measurements. Buttenfield and Weibel (1988) were among the first to present a framework for categorising the different elements of data uncertainty with a focus on how they can be cartographically presented with respect to the measurement scale of the data. Their approach focused on five categories of data uncertainty and three types of data; *categorical* (area features represented by categories or attributes represented by classes), *discrete* (point and line features), and *continuous* (surfaces and volume). The resulting matrix helps one to realise which visual variables are most appropriate to be used in representing a respective category. Buttenfield and Beard (1994) extended this framework by including also *location*, *attribute*, *time*, and *resolution* components. Pang et al. (1997) emphasised the importance for users to select appropriate methods for a given visualisation task. Therefore they classified existing visualisation methods with respect to the *value* (scalar, vector or multivariate), the *dimensionality* (1D, 2D, 3D, and time as another dimension), *data extent* (continuous and discrete) and *visualisation extent* (surface, volume, point etc.). Furthermore, Senaratne and Gerharz (2011) classified existing most popular uncertainty visualisation methods according to the supported data type, data format, uncertainty type, and interaction type. They further evaluated these methods with various domain experts, and extended this taxonomy in Senaratne et al. (2012) to include the user domain, thereby helping the

user to choose uncertainty visualisation methods according to their data as well as user requirements.

With this body of work for uncertainty visualisation, Beard et al. (1991) further discussed various aspects of cognitive limitations, such as the amount of image complexity that humans can handle, limited detail in mental images, limited resolution with which users can perform mental overlay of two maps especially in the case of intrinsic visualisations. These cognitive limitations restrict the design space for data uncertainty visualisation. Furthermore, with static or dynamic animation methods the users have very little control to decide what role uncertainty should play in the decision making process. Uncertainty visualisation approaches should adapt to the context of analysis at hand. Semi automatic and more *visual-interactive approaches* help users to overcome some of these pitfalls.

2.2 A Framework and Guidelines for Uncertainty Inheritance and Propagation in Visual Analytics

In this section, guidelines for assessing source uncertainties in data, as well as the propagated uncertainties in the system are presented. These guidelines help the users to be aware of the role of uncertainty in a data analysis system. This awareness together with the right tools help users to reduce propagated uncertainties where necessary.

The goal of visual analytics is to make the processing of data as transparent as possible for constructive knowledge generation (Keim et al., 2008). This can be achieved by enabling an effective collaboration between the machine and the human, thereby involving the human in the exploration and verification stages of data processing (Sacha et al., 2014a).

To succeed in the process of *constructive knowledge generation*, the user needs to be aware of the various uncertainties inherited in the data (also known as source uncertainty; Section 2.2.1), and those uncertainties that are propagated throughout the various stages of data processing on the machine (also known as propagated uncertainties; Section 2.2.2). Ignorance or unawareness of these uncertainties lead to inaccurate derivations of information, misinformed decision making, and incomplete analyses.

Several works have previously tackled data uncertainties in the visualisation pipeline. E.g., Correa et al. (2009) suggest to propagate and communicate the uncertainties that

arise inherently in the data and its transformations in the information visualisation pipeline. Zuk and Carpendale (2007) extend the data uncertainty visualisation pipeline of Pang et al. (1997) to include these uncertainties. These workflows facilitate the analyst in identifying the inherent and propagated uncertainties in their data. While these works focus on considering uncertainties in *data*, they do not specifically consider uncertainties propagated in the phases of *data visualisation*, *data modelling* and *data model-visualisation coupling*.

Thus, a novel conceptual framework is introduced here, taking into account the previous works above, to include uncertainties in all stages of a system in visual analytics. This framework includes defining the role of uncertainty in the visual analytics knowledge generation process, which is illustrated in Figure 2.7.

This framework in Figure 2.7 builds up on the visual analytics knowledge generation process of Sacha et al. (2014a), where it depicts how the human interacts (right side of the figure) with a data analysis system (left side of the figure) that consists of the components- data, model, and visualisation. The addition of uncertainty inheritance and propagation throughout the system as well as the human decision making process is indicated by the red flow of arrows through each component.

Through this framework recommendations are drawn on how to assess uncertainties at every component of a system of the visual analytics knowledge generation process. Therefore each component is examined in the following sections. As a result guidelines are derived to assess the uncertainties and examples of uncertainty elements for each component are provided.

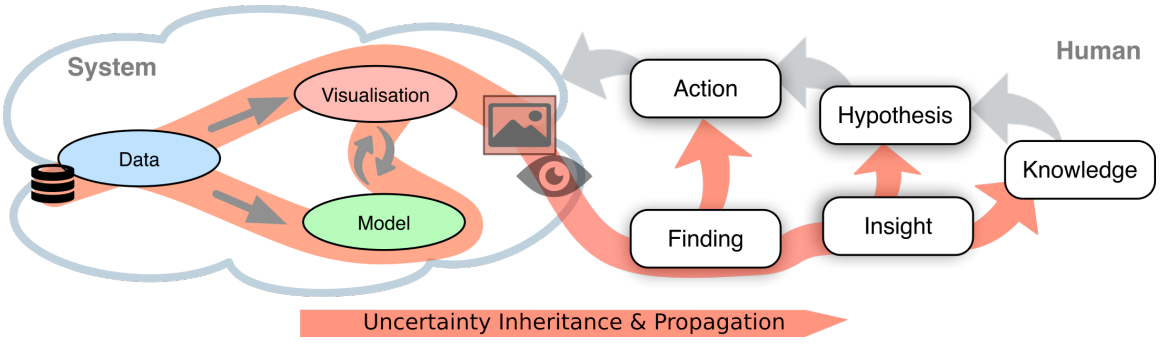


Figure 2.7: A framework for uncertainty inheritance and propagation within the visual analytics knowledge generation process. This figure appeared in Sacha et al. (2016).

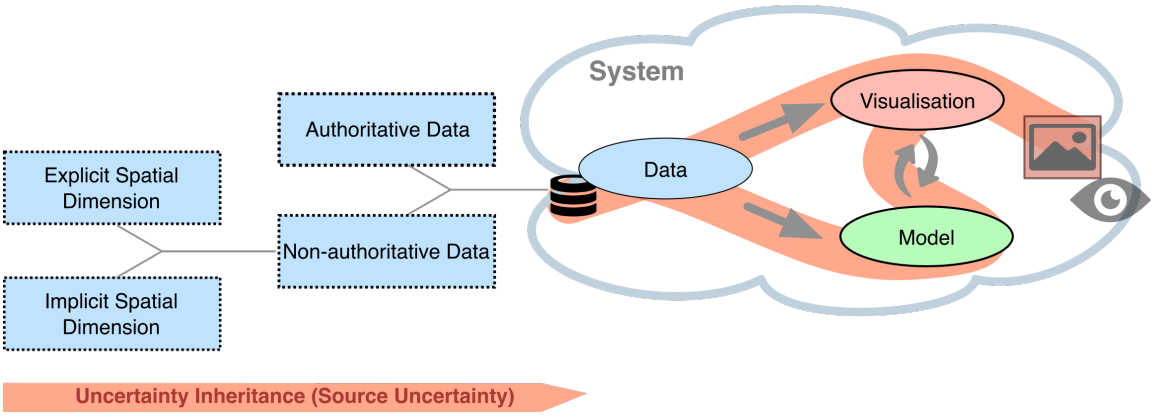


Figure 2.8: Source uncertainty is inherent to the data. It further varies depending on the authoritative and non-authoritative nature, as well as the implicit and explicit geography that is captured in the data.

2.2.1 Guidelines to Assess Source Uncertainty in Spatio-temporal Data

Source uncertainty is inherent to the data and largely depends on the way in which data is collected. This is shown in Figure 2.8. Thereby, spatio-temporal data can be classified as *authoritative data* and *non-authoritative data*: Authoritative data is data that is collected by professional domain experts in place of professional gatekeepers to moderate the quality in the data (lower uncertainty in the data means higher quality data). This minimises the inherent uncertainties in the data. Non-authoritative data, such as Volunteered Geographic Information (VGI) (Goodchild, 2007), are collected by citizens. Those users are often untrained, and regardless of their expertise and background, create geographic information on dedicated Web platforms, e.g., OpenStreetMap (OSM), Wikimapia[7], Google MyMaps [8], Map Insight [9] or Flickr[10].

In a typology of VGI, the works of Antoniou et al. (2010) and Craglia et al. (2012) classified VGI based on the type of explicit/implicit spatial dimension being captured and the type of explicit/implicit volunteering (see also Figure 1.1). In explicit VGI, contributors are mainly focused on mapping activities. Thus, the contributor explicitly annotates the data with geographic contents (e.g., geometries in OSM, Wikimapia, or Google). Data that is implicitly associated with a geographic location could be any kind of media: text, image, or video referring to or associated with a specific geographic location. For example, geotagged microblogs (e.g., Tweets), geotagged

[7] http://www.wikimapia.org
[8] https://www.google.com/maps/mm
[9] http://www.mapsharetool.com/external-iframe/external.jsp
[10] http://www.flickr.com

images from Flickr, or Wikipedia articles that refer to geographic locations. Source uncertainties in such citizen collected data is introduced due to the fact that humans perceive and express geographic regions and spatial relations imprecisely and in terms of vague concepts (Montello et al., 2003). This vagueness in human conceptualisation of location is due not only to the fact that geographic entities are continuous in nature, but also due to the quality and limitations of spatial knowledge (Hollenstein and Purves, 2010).

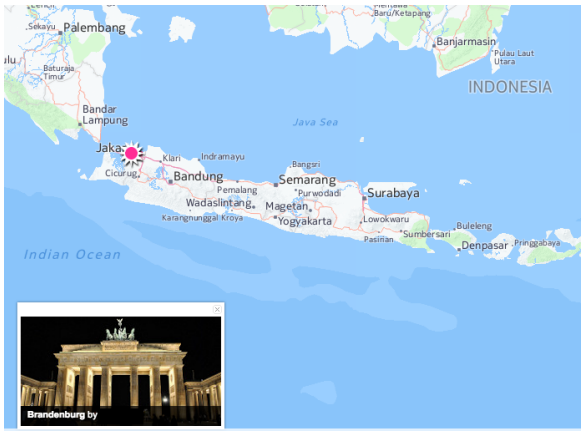


Figure 2.9: Example of an incorrectly geotagged photo on Flickr (Brandenburg Gate in Berlin is tagged in Jakarta). This figure appeared in Senaratne et al. (2017a).

Figure 2.9 shows an example of a photo of the famous tourist site, the Brandenburg Gate in Berlin, as incorrectly geotagged in Jakarta (Indonesia). Providing reliable services or extraction of useful information requires data with a fitness-for-use quality standard. Uncertainty stemming due to incorrect or malicious geographic annotations could be assessed in place of appropriate uncertainty assessment methods.

Goodchild and Li (2012) have discussed three approaches for assessing the source uncertainty of VGI: (1) *crowd-sourcing* (the involvement of a group to validate and correct errors that have been made by an individual contributor), (2) *social approaches* (trusted individuals who have made themselves a good reputation with their contributions to VGI can for example act as gatekeepers to maintain and control the uncertainty of other VGI contributions), and (3) *geographic approaches* (use of laws and knowledge from geography, such as Tobler's first law (Tobler, 1970) to assess the uncertainty).

Many works have developed methods[11] to asses the uncertainty of VGI following

[11]A method is considered to be a systematic procedure that is followed to assess the uncertainty measures and indicators. For example, comparing with satellite imagery is a method to assess the positional accuracy of maps.

these approaches. Based on an extensive literature review of those methods (Section 2.2.1), guidelines are presented in Sections 2.2.1 to 2.2.1 for source uncertainty analysis of three types of VGI: (1) image data, (2) text data, (3) map data. These three types of VGI are chosen based on the ways that are used to capture the data (maps: as GPS points and traces, image: as photos, text: as plain text), and because they are the most popular forms of VGI currently used.

As an outcome of the review of methods, *data mining* has been identified as another approach to assess VGI source uncertainty, which extends the approaches of Goodchild and Li (2012) here. Thereby, data mining utilises computational processes for discovering patterns and learning purely from data, irrespective of the laws and knowledge from geography, and independent from social or crowd-sourced approaches. Extending the spectrum of approaches will sprout more uncertainty assessment methods in the future, especially for VGI types that have not been extensively researched so far.

The Literature Review Methodology

The conducted review provides an overview of the state of the art methods to assess the source uncertainty of selected types of VGI. To achieve this goal, the review breaks down into three categories. Firstly, it is shown how the topic of uncertainty assessment within map, image, and text VGI has evolved over the years since the beginnings of VGI in 2007 until mid of 2015. Secondly, the reviewed papers are classified according to the type of uncertainty measure or indicator that is assessed within each of the papers. Thirdly, all the uncertainty measures and indicators that are addressed within each of the reviewed papers are classified with the different methods utilised to assess them.

The following strategy was used to select the literature for the review. Google Scholar was used to search for papers that include the following terms in their title or abstract: *data quality assessment*, *methods and techniques*, *uncertainty*, *volunteered geographic information*, *map*, *microblog*, *photo*. This query resulted in 425 research papers. The search results were sorted according to the Google Scholar relevance ranking [12]. This relevance ranking follows a combined ranking algorithm that contains a weighting for the full text of each article, author of article, publisher, and how often the article has been cited in other scholarly articles. The collection of papers was refined by filtering out the papers according to the following criteria: (1) papers were published from 2007, (2) papers should describe uncertainty assessment methods, or techniques, or tools, (3) a latest paper was selected when multiple versions of similar

[12] `https://scholar.google.com/scholar/about.html`

methods were available from the same research group. As such, 55 papers were selected in total.

Out of the 55 papers, 40 papers discuss methods for assessing the uncertainty of map-based VGI, in most cases taking OSM data as the VGI source. 18 papers introduce methods for text-based VGI taking mainly Twitter, Wikipedia, and Yahoo Answers[13] as the VGI source. 12 papers introduce methods for image-based VGI taking Flickr and Panoramio[14] as their VGI source. In reference to Craglia et al. (2012)'s typology of VGI with the reviewed papers, most uncertainty assessment work is done on explicit VGI and lesser amount of work is done on implicit VGI, although implicit VGI due to its very nature has more concerns regarding their uncertainty. Figure 2.10 shows the distribution of the reviewed papers for VGI uncertainty assessment methods.

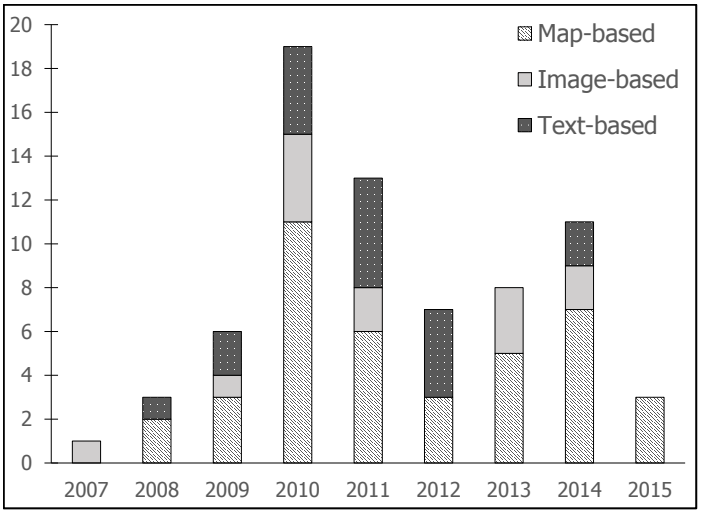


Figure 2.10: Distribution of the surveyed papers. This figure appeared in Senaratne et al. (2017a).

Guideline 1: Choosing Uncertainty Measures and Indicators

We have found 17 uncertainty measures and indicators (7 measures and 10 indicators) that are addressed within the 55 papers surveyed. We found that papers particularly focusing on map-based VGI are clearly using only ISO standardised measures for

[13] `https://answers.yahoo.com`
[14] `http://www.panoramio.com`

uncertainty assessment, whereas text-based VGI have been assessed only on the credibility, text content quality, and vagueness. Image-based VGI have been assessed in several papers on the positional/thematic accuracy, credibility, vagueness, experience, recognition, and reputation. Guidelines for choosing the uncertainty measures and indicators according to the type of VGI are summarised in Table 2.1. The table further provides a guideline for users to learn about specific uncertainty measures and indicators by following the referenced literature.

These uncertainty measures/indicators gather previously discussed spatial data uncertainty elements in the literature, but also extend the previous categorisations such as Thomson et al. (2005), to include further spatial data uncertainty indicators such as reputation, text content quality, or experience.

Table 2.1: Classification of the reviewed papers according to the uncertainty measures and indicators. ⋆ = map-based, • = image-based, ◇ = text-based, and ⋈= all types of VGI.

Papers	Uncertainty measures and indicators																
	Positional accuracy	Thematic accuracy	Topological consistency	Completeness	Temporal accuracy	Geometric accuracy	Semantic accuracy	Lineage	Usage	Credibility	Trustworthiness	Content quality	Vagueness	Local knowledge	Experience	Recognition	Reputation
Jacobs et al.(2007)	•																
Agichtein et al.(2008)												◇					
Schmitz et al.(2008)			⋆														
Mummidi&Krumm (2008)		⋆															
Hasan et al.(2009)												◇					
Kounadi (2009)	⋆																
Ather (2009)	⋆			⋆													
De Longueville et al.(2010)													⋈				
Bishr&Janowicz(2010)											⋈						
Mendoza et al.(2010)										◇							
Haklay(2010)	⋆			⋆													
Ciepluch(2010)	⋆			⋆													
Corcoran et al.(2010)			⋆														
Girres&Touya (2010)		⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆								
Haklay et al. (2010)	⋆																
Poser&Dransch (2010)		⋈															
Brando&Bucher (2010)	⋈	⋈	⋈	⋈	⋈												

Papers	Positional accuracy	Thematic accuracy	Topological consistency	Completeness	Temporal accuracy	Geometric accuracy	Semantic accuracy	Lineage	Usage	Credibility	Trustworthiness	Content quality	Vagueness	Local knowledge	Experience	Recognition	Reputation
Huang et al. (2010)											⋈						
De Tré et al. (2010)	⋆	⋆															
Al Bakri&Fairbairn (2010)	⋆																
van Exel et al. (2010)															⋈	⋈	⋈
Ciepluch et al. (2011)			⋆														
Neis et al. (2011)			⋆														
Codescu(2011)		⋆															
Castillo et al.(2011)										◇							
Becker et al (2011)												◇					
Canini et al.(2011)										◇							
Ostermann&Spinsanti (2011)										⋈							
Kessler et al. (2011)								⋆									
O'Donovan et al.(2012)										◇							
Kang et al.(2012)										◇							
Gupta et al.(2012)										◇							
Morris et al. (2012)										◇							
Helbich et al.(2012)	⋆																
Mooney&Corcoran(2012)		⋆															
Koukoletsos et al. (2012)				⋆													
Kessler&deGroot(2013)		⋆	⋆	⋆													

Papers	Positional accuracy	Thematic accuracy	Topological consistency	Completeness	Temporal accuracy	Geometric accuracy	Semantic accuracy	Lineage	Usage	Credibility	Trustworthiness	Content quality	Vagueness	Local knowledge	Experience	Recognition	Reputation
Zielstra&Hochmair(2013)	●																
Canavosio-Zuzelski et al.(2013)	⋆																
Hecht et al.(2013)				⋆													
Vandecasteele&Devillers(2013)		⋆															
Jackson et al. (2013)	⋆			⋆													
Foody et al. (2014)		●															
Barron et al.(2014)			⋆	⋆													
Siebritz(2014)			⋆														
Wang et al.(2014)			⋆														
Fan et al.(2014)	⋆			⋆													
Tenney(2014)	⋆			⋆													
Ali et al.(2014)		⋆															
Bordogna et al. (2014)													●◇		●	●	●
Forghani&Delavarl(2014)			⋈														
Hollenstein&Purves(2014)	●																
Arsanjani (2015)		⋆															
Vandecasteele&Devillers (2015)							⋆										
Hashemi&Abbaspour (2015)			⋆														

The following sections describe the selected types of VGI: 1) *map*, 2) *image*, and 3)*text*, their uses, how data source uncertainty arises, and the guidelines for assessing the source uncertainties. A summary of all the following guidelines are presented in Table 2.2.

Guideline 2: Source Uncertainty Assessment in Image-based VGI

Image-based VGI is mostly produced implicitly within portals such as Flickr, Panoramio, Instagram etc., where contributors take pictures of a particular geographic object or surrounding with cameras, smart phones, or any hand held device, and attach a geospatial reference to it. These objects/surroundings can be spatially referenced either by giving geographic coordinates and/or by giving user-assigned geospatial descriptions of these photographs in the form of textual labels. These photo sharing websites have several uses such as environmental monitoring (Fuchs et al., 2013), pedestrian navigation (Robinson et al., 2012), event and human trajectory analysis (Andrienko et al., 2009), creation of geographical gazetteers (Popescu et al., 2008), or even to complement institutional data sources in your locality (Milholland and Pultar, 2013).

Tagging an image is a means of adding metadata to the content in the form of specific keywords that describe the content (Golder and Huberman, 2006), or in the form of geographic coordinates (geotagging) to identify the location linked to the image content (Valli and Hannay, 2010). There exist several approaches to geotag an image: record the geographic location with the use of an external GPS device, with an in-built GPS (in many of the modern digital cameras, smart phones), or manually positioning the photo on a map interface.

Not only the GPS precision and accuracy errors resulting from various devices, but also other factors influence the uncertainty of image-based VGI. For example, instead of stating the position from where the photo was taken (photographer position) some contributors tend to geotag the photo with the position of the photo content, which could be several kilometers away from where the photo originated causing positional accuracy issues (as also discussed in Keßler et al. (2009)). This is a problem when we want to utilise these photos for example in human trajectory analysis. Furthermore, due to the lack of sufficient spatial knowledge, contributors sometimes incorrectly geotag their photographs (Figure 2.9), also in lower geographic resolutions (in case of Flickr, some contributors do not zoom enough to the street level, instead they zoom up to country or city level to geotag their photos). Or some contributors geotag and textually label random irrelevant photos for actual events, causing the users to doubt

the trustworthiness of the content. Such content are not fit for use for tasks such as disaster management, environmental monitoring, or pedestrian navigation. Citizen Science Projects such as GeoTag-X[15] combine machine learning and crowd-sourcing methods to discover unauthentic material and clean them.

The guidelines to assess these uncertainties of image-based VGI are given below in the form of uncertainty assessment methods. These methods particularly support one or more of the uncertainty elements that can be found in image-based VGI. In addition to the reviewed methods below, in Chapter 3 we developed a novel method that assesses the positional accuracy and credibility of image-based VGI.

Guideline 2.1: Positional Accuracy

This guideline specifies methods to assess the positional accuracy of image-based VGI. The first method is by Jacobs et al. (2007), where they explored the varying positional accuracy of photos by matching photos with ancillary satellite imagery. They localise cameras based on satellite imagery that correlates with the camera images taken at a known time. Their approach helps where it is important to know the accurate location of the photographer instead of the target object. The method by Zielstra and Hochmair (2013) on the other hand compared the geotagged positions of photos to the manually corrected camera position based on the image content. Their results indicate better positional accuracy for Panoramio photos compared to Flickr photos. The method by Hollenstein and Purves (2010) assessed the positional accuracy of such photos by manually inspecting these photos for their correspondence between the tagged geographic label and geotagged position.

Guideline 2.2: Thematic Accuracy

This guideline specifies methods to assess the thematic accuracy of image-based VGI. Foody et al. (2014)'s method used Geowiki as the data source, where it contains a series of satellite imagery. Volunteered contributors were given the task to label the land use categories in these satellite imagery from a pre-defined set of labels. The accuracy of the labeling was assessed through conducting a latent class analysis (LCA). LCA allows the analyst to derive an accuracy measurement of the classification when there are no reference datasets available to compare with. The authors further emphasise that this method can be applied to image-based VGI. Further, their approach characterises the volunteers based on the accuracy of their labels of land use classes. This helps to ultimately determine the volunteer quality. The method of Zhang and

[15] `http://geotagx.org/`

Kosecka (2006) used feature-based geometric matching using the image recognition software SIFT (Lindeberg, 2012) to localise sample photos in urban environments. Although their work was not based on VGI, this is a potential method to solve uncertainty related issues within image-based VGI.

Guideline 3: Source Uncertainty Assessment in Text-based VGI

Text-based VGI (typically microblogs) is mostly produced implicitly on portals such as Twitter, Reddit or various Blogs, where people contribute geographic information in the form of text by using smart phones, PCs, or any hand held devices. Twitter for example is used as an information foraging source (MacEachren et al., 2011), in journalism to disseminate data to the public in near real-time basis (O'Connor, 2009; Castillo et al., 2011), detect disease spreading (Chunara et al., 2012), event detection (Bosch et al., 2013), and for gaining insights on social interaction behavior (Huberman et al., 2008) or trajectories of people (Andrienko et al., 2013a).

In text-based VGI, the spatial reference can be either in the text, where the contributor refers to a place-name (e.g., 'Lady Gaga is performing in New York today'), or the spatial reference can be the geotag where the tweet is originating from. While some people contribute meaningful information most others use these mediums to express personal opinions, moods, or for malicious aims such as bullying or trolling to harass other users. Gupta et al. (2012) conducted a study to investigate how much information is credible and therefore useful, and how much information is spam, on Twitter. They found that 14% of Tweets collected for event analysis were spam, while 30% of the Tweets contained situational awareness information, out of which only 17% of the total tweets contained credible situational awareness information. Such spam makes it difficult to derive useful information that could be of interest for the above named use cases. Therefore uncertainty analysis of these data is important to filter out the useful information, and disregard the rest.

Other than the inherent GPS errors in devices, a bigger role for uncertainty issues is played by the contributor herself/himself based on the information she/he provides. Also due to the lack of spatial knowledge of some contributors the location is incorrectly specified, and at times at a low resolution (in the Twitter interface on PCs the contributor can specify the location not only at the city level, but also at a more coarse state level). Sometimes if the contributor is writing about an event that takes place a few hundred kilometers away from a contributor's position, she/he would geotag her content with the location of the event rather than her position. Or the other way around.

The guidelines to assess these uncertainties of text-based VGI are given below in the form of uncertainty assessment methods. These methods for text-based VGI particularly support to assess the credibility and text content quality based on contributor, text, and content features. In addition to the reviewed methods below, in Chapter 4 we developed a novel method to visually analyse and explore the text content quality and credibility as uncertainty indicators for text-based VGI. Further, our developed method helps to reduce the uncertainties in the derived results.

Guideline 3.1: Credibility

This guideline specifies methods to assess the credibility of text-based VGI. These methods are as follows.

Relating to a social approach of uncertainty analysis, Mendoza et al. (2010) found out that rumors on Twitter tend to be more questioned by the Twitter community during an emergency situation. They further indicate that the Twitter community acts as a collaborative filter of information.

Castillo et al. (2011) introduced a method that employed users on mechanical turk[16] to classify pre-classified 'news-worthy events' and 'informal discussions' on Twitter according to several classes of credibility (i. almost certainly true, ii. likely to be false, ..). This is then used in a supervised classification to evaluate which Tweets belong to these different classes of credibility. This helped the authors to derive credibility indicators. The user features such as average status count or the number of followers among others were found to be the top ranked user-based credibility features.

The method of Gupta et al. (2012) is similar to Castillo et al. (2011), and followed a supervised feature classification PageRank like method to propagate the credibility on a network of Twitter events. They use event graph-based optimization to enhance the trust analysis at each iteration that updates the credibility scores. A credible entity (node) links with a higher weight to more credible entities than to non-credible ones. Their approach is similar to that of Castillo et al. (2011), but the authors proposed a new technique to re-rank the Tweets based on a Pseudo Relevance Feedback.

Canini et al. (2011)'s method divided credibility into implicit and explicit credibility. Implicit credibility is the perceived credibility of Twitter contributors, and is assessed by Twitter users by evaluating an external data source together with the Tweeters content topicality and its relevance to the context, and social status (follower/ status counts). Explicit credibility is evaluated by ranking Tweeters (Twitter contributors)

[16] `https://www.mturk.com`

on a scale from 1 to 5 based on their trustworthiness. End result is a ranking recommendation system on whom to follow on Twitter regarding a particular topic.

O'Donovan et al. (2012) provided an analysis of the distribution of credibility features in four different contexts in the Twitter network: diversity of topics, credibility, chain length and dyadic pairs. The results of their analysis indicate that the usefulness of credibility features depends on the context in question. Thus the presence of a credibility feature alone is not good enough to evaluate the credibility of the context, but rather a particular combination of different credibility features that are 'suitable' for the context in question.

Morris et al. (2012) designed a pilot study with participants (with no technical background) to extract a list of features that are useful to make their credibility judgments. Finally to run the survey, the authors sent the survey to a sample of Twitter users in which they were asked to assess how each feature impacts their credibility judgment on a five-point scale. Their findings indicate that features such as verified author expertise, re-tweets from someone you trust, or author is someone you follow have higher credibility impact. These features differ somewhat to the features extracted through the supervised classification of Castillo et al. (2011). These features were further ranked according to the amount of attention received by Twitter users.

Kang et al. (2012) defined three different credibility prediction models and studied how each model performs in terms of credibility classification of Twitter messages. These are: 1. social model, 2. content-based model, and 3. hybrid model (based on different combinations of the two previous models). The social model relies on a weighted combination of credibility indicators from the underlying social network (e.g., re-tweets, no. of followers). The content-based model identifies patterns and tweet properties that lead to positive reactions such as re-tweeting or positive user ratings, by using a probabilistic language-based approach. Most of these content-based features are taken from Castillo et al. (2011). The main results from the paper indicate that the social model outperformed all other models in terms of predication accuracy, and that including more features in the predication task doesn't mean a better predication accuracy.

Guideline 3.2: Text Content Quality

This guideline specifies methods to assess the uncertainty of text-based VGI in terms of the text content quality. These methods are as follows.

Agichtein et al. (2008) described a generic method for all text-based social media data. They use three inputs for a feature classifier to determine the content quality:

1. textual features (e.g., word n-grams up to length 5 that appears in the text more than 3 times, semantic features such as punctuations, typos, readability measures, avg. no. of syllables per word, entropy of word lengths, grammarticality), 2. user relationships (between users and items, user intuition such as good answers are given by good answerers, and vote for other good answerers), 3. usage statistics (no. of clicks on an item, dwell time on content).

Becker et al. (2011) used a two tier approach for the uncertainty analysis of text-based Twitter data in an event analysis context. To identify the events, they first cluster tweets using an online clustering framework. Subsequently, they use three centrality-based approaches to identify messages in the clusters that have high textual quality, strong relevance, and are useful. These approaches are: 1. centroid similarity approach that calculates the cosine similarity of the tf-idf statistic of words, 2. degree centrality method which represents each cluster message as a node in a graph, and two nodes are connected with an edge when their cosine similarity exceeds a predetermined threshold, 3. LexRank approach distributes the centrality value of nodes to its neighbors, and top messages in a cluster are chosen according to their LexRank value.

Hasan Dalip et al. (2009) on the other hand used text length, structure, style readability, revision history, and social network as indicators of text content quality in Wikipedia articles. They further used regression analysis to combine various such weighed quality values into a single quality value, that represents an overall aggregated quality metric for text content quality.

Bordogna et al. (2014) measured the validity of text data by measuring the number of words, proportion of correctly spelled words, language intelligibility, diffusion of words, and the presence of technical terms as indicators of text content quality. They further explored indicators such as experience, recognition and reputation to determine the uncertainty of VGI.

Guideline 4: Source Uncertainty Assessment in Map-based VGI

Map-based VGI concerns all VGI sources that include geometries as points, lines and polygons, the basic elements to design a map. Among others, OSM, Wikimapia[17], and Google Map Maker[18] are examples of map-based VGI projects. However, OSM is the most prominent project due to the following reasons: (i) It aims to develop a free map of the world accessible and obtainable for everyone; (ii) It has millions of

[17] `http://wikimapia.org`
[18] `https://mapmaker.google.com`

registered contributors; (iii) It has active mapper communities in many locations; and (iv) It provides free and flexible contribution mechanisms for data (useful for map provision, routing, planning, geo-visualization, point of interests (POI) search etc.).

As in most VGI projects, the spatial dimension of OSM data is annotated in the form of nodes, lines, or polygons with latitude/longitude referencing, and attributes are annotated by tags – in the form of key-value pairs. Each tag describes a specific geographic entity from different perspectives. There are no restrictions to the usage of these tags: endless combinations are possible, and the contributors are free to choose the tags they deem appropriate. Nevertheless, OSM provides a set of recommendations of accepted key-value pairs, and if the contributors want their contributions to become a part of the map, they need to follow the agreed-upon standards. This open classification scheme can lead to misclassification and reduction in data quality. Map-based VGI is commonly used for purposes like navigation and POI search. For these purposes the positional accuracy and the topological consistency of the entities are as important as their abstract locations.

The other dimension is the attribute accuracy, where the annotations associated with an entity should reflect its characteristics without conflicts (e.g., for road tags, "oneway"="true" and "two-way"="true"). In OSM, the loose contribution mechanisms result in problematic classifications that influence the attribute accuracy. In addition to accuracy, providing reliable services is affected by data completeness; features, attribute, and model completeness. Whether a map includes all the required features, whether a feature is annotated with a complete set of attributes, and if the model is able to answer all possible queries, all these points are related to the completeness uncertainty measure. Especially due to the lack of ground-truth data for comparison, assessing VGI completeness still raises some challenges.

The guidelines to assess these uncertainties of map-based VGI are given below in the form of uncertainty assessment methods. These methods for map-based VGI particularly support to assess the positional accuracy, topological consistency, thematic accuracy, semantic accuracy, completeness, temporal accuracy, lineage, usage, and purpose.

Guideline 4.1: Positional Accuracy

This guideline specifies methods to assess the positional accuracy as an uncertainty element of map-based VGI. These methods are as follows.

In the works of Kounadi (2009); Ather (2009); Haklay (2010); Ciepłuch et al. (2010); Al-Bakri and Fairbairn (2010); Zandbergen et al. (2011); Helbich et al. (2012); Jackson

et al. (2013); Fan et al. (2014); Tenney (2014); Brando and Bucher (2010); Al-Bakri and Fairbairn (2010), authors employed officially gathered reference datasets to assess the positional accuracy of map-based VGI (mostly OSM data) by comparison. The comparison with reference data method has been further employed for the assessment of thematic accuracy (Girres and Touya, 2010; Poser and Dransch, 2010; Kounadi, 2009; Brando and Bucher, 2010; Arsanjani et al., 2015), completeness (Haklay, 2010; Ciepłuch et al., 2010; Kounadi, 2009; Ather, 2009; Ciepłuch et al., 2011; Hecht et al., 2013; Jackson et al., 2013; Fan et al., 2014; Tenney, 2014; Brando and Bucher, 2010), geometric accuracy (Girres and Touya, 2010). For geometric accuracy OSM objects of same structure were manually matched. This manual approach was preferred over an automated approach to avoid any processing errors.

Haklay (2010) applied the Linus Law and found out that higher the number of contributors on a given spatial unit on OSM, lower the uncertainty, and therefore higher the quality. This study shows that comparison to reference datasets isn't the only way to assess the uncertainty of OSM data as done in many use cases.

De Tré et al. (2010) used a Possibilistic Truth Value (PTV) as a normalized possibility distribution to determine the uncertainty of the POIs being co-located. The uncertainty regarding the positioning of a POI is primarily caused by the imprecision with which the POI are positioned on the map interface. The proposed technique further semantically checks and compares the closely located POIs. Their method helps to identify redundant VGI, and fuse the redundancies together. Furthermore, this approach has been applied to also assess the thematic accuracy of map-based VGI.

In a rather different approach, Canavosio-Zuzelski et al. (2013) performed a photogrammetric approach for assessing the positional accuracy of OSM road features using stereo imagery and a vector adjustment model. Their method applies analytical measurement principles to compute accurate real world geo-locations of OSM road vectors. The proposed approach was tested on several urban gridded city streets from the OSM database with the results showing that the post adjusted shape points improved positional accuracy by 86%. Furthermore, the vector adjustment was able to recover 95% of the actual positional displacement present in the database.

Brando and Bucher (2010) presented a generic framework to manage the uncertainty of ISO standardised uncertainty measures by using formal specifications and reference datasets. Formal specifications facilitate the assurance of high quality in three manners with means of integrity constraints: i) support on-the-fly consistency checking, ii) comparison to external reference data, iii) reconcile concurrent editions

of data. However, due to a lack of proof of concept the practical applicability of this approach is difficult to conceive.

Guideline 4.2: Topological Consistency

This guideline specifies methods to assess the topological consistency as an uncertainty element of map-based VGI. These methods are as follows.

The topological consistency in OSM data is assessed mainly using intrinsic data checks to detect and alleviate problems occurring through for example overlapping features or overshoots and undershoots in the data (also known as dangles where start and end point of two different lines should meet but do not, due to bad practices in digitisation). The authors Schmitz et al. (2008); Neis et al. (2011); Barron et al. (2014); Siebritz (2014) have demonstrated that for each of these measures a separate topology integrity rule can be designed and applied. Further, based on the definition of planar and non-planar topological properties Corcoran et al. (2010) and Da Silva and Wu (2007) have used geometrical analysis methods to assess the topological consistency of the OSM data.

In another work, the concept of spatial similarity in multi-representations have been employed in order to perform both extrinsic and intrinsic uncerainty analysis (Hashemi and Abbaspour, 2015). The authors discuss that their method could be efficiently applied to VGI data for the purpose of vandalism detection. Other studies have also focused on evaluating the topological consistency of OSM data with a focus on road network infrastructures (Will, 2014). In Wang et al. (2014) and Girres and Touya (2010) the authors have used the Dimensional Extended nine-Intersection Model (DE-9IM) in order to compute the qualitative spatial relation between road objects in OSM. This method and model allow them to check for topological inconsistencies and be able to locate the junctions of roads in order to, for example generate expected road signs.

Guideline 4.3: Thematic Accuracy and Semantic Accuracy

This guideline specifies methods to assess the thematic and semantic accuracy as uncertainty elements of map-based VGI. These methods are as follows.

Mooney and Corcoran (2012) points out that most errors in OSM are caused by manual annotation by contributors who sometimes misspell the feature values. Addressing this issue, Codescu et al. (2011); Vandecasteele and Devillers (2013); Ali et al. (2014) have developed semantic similarity matching methods, which automatically assess the contributor annotation of features in OSM according to the semantic meaning

of such features. In the work of Girres and Touya (2010), they found semantic errors were mainly due to the mis-specification of roads. For example: roads that were classified as 'secondary' in the reference dataset were classified as 'residential', or 'tertiary' by contributors in OSM data. The reasons for these inaccuracies as seen by authors are the lack of a standardised classification, looseness for contributors to enter tags and values that are not present in the OSM specification, lack of naming regulations w.r.t. for example capitalisation or prefixes. The authors emphasise the need for standardised specifications to improve semantic and thematic accuracy of OSM data.

Furthermore, in regard to semantic accuracy of map-based VGI, Vandecasteele and Devillers (2015) introduced a tag recommender system for OSM data which aims to decrease the semantic uncertainty of tags. OSMantic is a plugin for the Java OpenStreetMap editor which automatically suggests relevant tags to contributors during the editing process. Mummidi and Krumm (2008) used clustering methods on Microsoft's Live Search Maps[19] to group user contributed pushpins of POIs that are annotated with text. Frequent text phrases that appear in one cluster but infrequently in other clusters help to increase the confidence that the particular text phrase describes a POI.

Guideline 4.4: Completeness

This guideline specifies methods to assess the completeness as an uncertainty element of map-based VGI. These methods are as follows.

Koukoletsos et al. (2012) proposed to use a feature-based automated matching method for linear data using reference datasets. Barron et al. (2014) and Girres and Touya (2010) used intrinsic data checks to record the statistics of the number of objects, attributes, and values, thereby keeping track of all omissions and commissions to the database.

Guideline 4.5: Temporal Accuracy

This guideline specifies methods to assess the temporal accuracy as an uncertainty element of map-based VGI. These methods are as follows.

Very few works exist to assess the temporal accuracy. Among the few, Girres and Touya (2010) used statistics to observe the correlations of the number of contributors to the mean capture date, and to the mean version of the capture object in order to assess how many objects are updated. Their results show a linear increase of the

[19] `http://maps.live.com`

mean date, and the mean version of captured object in relation to the number of contributors in the chosen geographic area. Concluding results show higher the number of contributors, more recent the objects were, and the more up-to-date the objects were.

Guideline 4.6: Lineage, Usage, Purpose

This guideline specifies methods to assess the lineage, usage, and purpose as uncertainty elements of map-based VGI. These methods are as follows.

In Keßler et al. (2011), following a data oriented approach with a focus on the origins of specific data items, their provenance vocabulary explicitly showed the lineage of data features of any online data. They base their provenance approach on Hartig (2009) on 'provenance information in the web of data'. Their approach allows them to classify OSM features according to recurring editing and co-editing patterns. To keep track of the data lineage, Girres and Touya (2010) urge the need for moderators who have control over screening the contributions (as in Wikipedia) for necessary source information. They further analyse the usage of data by comparing the limitations that were observed in previous evaluations of map-based VGI.

As a generic approach to assess ISO standardised uncertainty indicators, Keßler and de Groot (2013) propose Trust as a proxy to measure the topological consistency, thematic accuracy, and completeness in these map data based on data provenance, a method which relies on trust indicators as opposed to ground truth data.

Guideline 5: Generic Approaches for Source Uncertainty Assessment

This guideline specifies generic methods to assess the source uncertainty in all types of VGI. These methods are as follows.

As a generic method for all VGI, Forghani and Delavar (2014) proposed a new uncertainty metric for the assessment of topological consistency by employing heuristic metrics such as minimum bounding geometry area and directional distribution (Standard Deviational Ellipse). Van Exel et al. (2010) proposed to use contributor related uncertainty indicators such as local knowledge (e.g., spatial familiarity), experience (e.g., amount of contributions), and recognition (e.g., tokens achieved). A conceptual workflow for automatically assessing the uncertainty of VGI in crisis management scenarios was proposed by Ostermann and Spinsanti (2011). VGI is cross-referenced with other VGI types, and institutional ancillary data that are spatially and temporally close. However, in a realistic implementation this combination of different VGI data types for cross referencing is a challenging task due to their heterogeneity.

Bishr and Janowicz (2010) proposed to use trust together with reputation as a proxy measure for VGI uncertainty, and established the spatial and temporal dimensions of trust. They assert that shorter geographic proximity of VGI observations provide more accurate information as opposed to higher geographic proximity VGI observations (implying that *locals know better*, *the proximate spectator sees more*). On a temporal perspective of trust, they further claim that trust in some VGI develop and decay over time, and that the observation time of an event has an affect on the trust we endow in one's observation. Furthermore, to assess the trust of VGI Huang et al. (2010) developed a method to detect outliers in the contributed data. De Longueville et al. (2010) proposed two methods to assess the vagueness in VGI. 1. contributor encodes the vagueness of their contributed spatial data in a 0 - 5 scale (e.g., 5 = it's exactly there, 0 = I don't know where it is. 2. the second type is system created vagueness that is assessed through automatically capturing the scale at which VGI is produced. VGI produced in lower scales is classified as more vague.

Summary of Guidelines for Source Uncertainty Assessment

Table 2.2 shows a summary matrix of all uncertainty measures and indicators observed in the literature review, with various methods that can be applied to assess these uncertainty measures/indicators. The sparse cells in the table indicate the uncertainty measures/indicators that have not been explored excessively. Following the classification by Goodchild and Li (2012), the methods have been categorised into (i) geographic, (ii) social, and (iii) crowdsourcing. However, additionally to their categorisation, new methods have been found here under the category (iv) *data mining*. Table 2.2 can be used as a guideline for users who want to solve various uncertainty issues within map, text, and image-based VGI. Nevertheless, this should be followed with caution, as this literature review can only reflect what has been discovered at time of writing, and the presented methods could be applied beyond this discovery, and therefore need to be further explored.

Table 2.2: Uncertainty measures/indicators classified according to type of method to assess them. Methods are grouped in geographic, social, crowdsourced (abbrv. 'C.'), and the newly found data mining approaches. Type of VGI indicated as: ⋆ = map-based, ● = image-based, and ◇ = text-based.

	Geographic													Social							C.	Data mining									
	Compare w. reference data	Line of sight	Formal specifications	Semantic consistency check	Geometrical analysis	Intrinsic data check	Integrity constraints	Automatic tag recommend.	Geographic proximity	Time between observations	Automatic scale capturing	Geographic familiarity	Manual inspection	Manual inspect./annotat.	Manual annotation	Comparing limitation	Linguistic decision making	Meta-data analysis	Tokens achieved	Applying Linus law	Possibilistic truth value	Cluster analysis	Latent class analysis	Correlation statistics	Automatic outlier detection	Regression analysis	Supervised classification	Feature classification	Provenance vocabulary	Heuristic metrics
Positional accuracy	⋆●◇	●	⋆●◇										●	●						⋆	⋆									
Thematic accuracy	⋆●◇		⋆●◇	⋆											⋆							⋆	●						⋆	
Topological consistency	⋆●◇		⋆●◇	⋆	⋆	⋆	⋆																						⋆	⋆●◇
Complete-ness	⋆●◇		⋆●◇			⋆		⋆																					⋆	
Temporal accuracy																								⋆						
Geometric accuracy	⋆																													
Semantic accuracy	⋆							⋆																						
Lineage													⋆																⋆	

		Credibility	Usage	Trust	Content quality	Vagueness	Local knowledge	Experience	Recognition	Reputation
Data mining	Heuristic metrics									
	Provenance vocabulary									
	Feature classification	◇			◇					
	Supervised classification	◇								
	Regression analysis				◇					
	Automatic outlier detection			✶●◇						
	Correlation statistics									
	Latent class analysis									
	Cluster analysis									
	Possibilistic truth value									
C.	Applying Linus law									
Social	Tokens achieved								✶●◇	
	Meta-data analysis							✶●◇		
	Linguistic decision making				●◇			●◇	●◇	●◇
	Comparing limitation		✶							
	Manual annotation					✶●◇				
	Manual inspect./annotat.									
	Manual inspection									
Geographic	Geographic familiarity						✶●◇			
	Automatic scale capturing					✶●◇				
	Time between observations			✶●◇						
	Geographic proximity			✶●◇						
	Automatic tag recommend.									
	Integrity constraints									
	Intrinsic data check									
	Geometrical analysis									
	Semantic consistency check									
	Formal specifications									
	Line of sight	●								
	Compare w. reference data	✶●◇								

2.2.2 Guidelines to Assess Propagated Uncertainties in Spatio-temporal Data

Uncertainty is created and passed on from the source to the model and subsequently to the visualisation. Haber and McNabb (1990) introduced uncertainty propagation to their visualisation reference model, where the visualisation of uncertainty focuses on the uncertainties that are in the measurement and simulation data (seen in data source uncertainty). They discuss how uncertainty propagates from the filtering stage, mapping stage, to the rendering stage of a traditional pipeline model. They call this *uncertainty of visualisation*.

In the visual analytics knowledge generation process uncertainty is propagated during the data processing (where data undergo transformations such as interpolation, sampling, quantisation, or normalisation), modeling, visualisation, and the model-visualisation coupling stages where these propagated uncertainties keep aggregating as data travel through these stages in the system side of the visual analytics knowledge generation process as shown in Figure 2.7.

The guidelines to assess propagated uncertainties are given below in the form of uncertainty assessment methods.

Guideline 6: Uncertainty Propagation in Data Processing

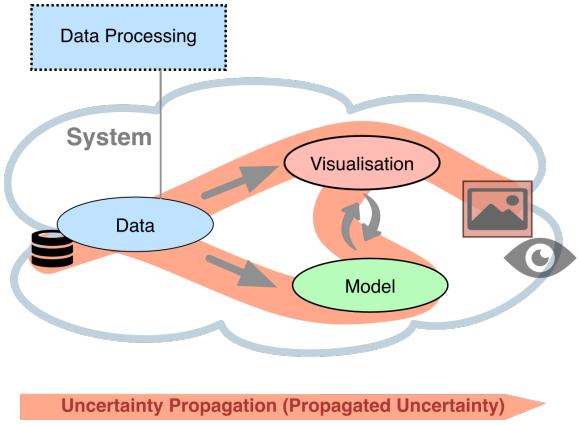


Figure 2.11: Uncertainty propagation through data processing.

System inputs that go through transformations such as interpolation, extrapolation, normalisation etc., propagate uncertainty as seen in Figure 2.11. Choosing a suitable uncertainty propagation method depends on the confidence level, the extent to which you need uncertainty quantification, and the computational cost that one can endure

(Lee and Chen, 2009). Probabilistic approaches (e.g., Monte Carlo Simulation methods) are known to be most robust in quantifying such uncertainties.

This guideline specifies such methods to assess the propagated uncertainty at the data processing stage. These methods are as follows.

Lee and Chen (2009) described in detail five types of probabilistic approaches in their comparative study of uncertainty propagation methods. First, and the most popular type is *sampling based methods* that rely on simulations. Examples include, Monte Carlo method (Pross et al., 2012), importance sampling (Melchers, 1989), and adaptive sampling (Bucher, 1988). Second type is the *local expansion methods*. Examples include Taylor series method and perturbation method (Madsen et al., 2006). The third method is *the most probable point (MPP) based method*. Examples are first-order and second-order reliability methods (Fiessler et al., 1979). The fourth category is *the functional expansion based method*. Examples include Neumann expansion and polynomial chaos expansion methods (Xiu and Karniadakis, 2003). The fifth category is *the numerical integration based methods*. Examples include full factorial numerical integration and dimension reduction methods (Lee and Kwak, 2006). Statistics such as standard deviation, variance and range are further used to propagate data processing uncertainties. Additionally, Cedilnik and Rheingans (2000) used distance based functions to measure the similarity of values, and further point out that interpolated values can also be used.

Guideline 7: Uncertainty Propagation in Data Models

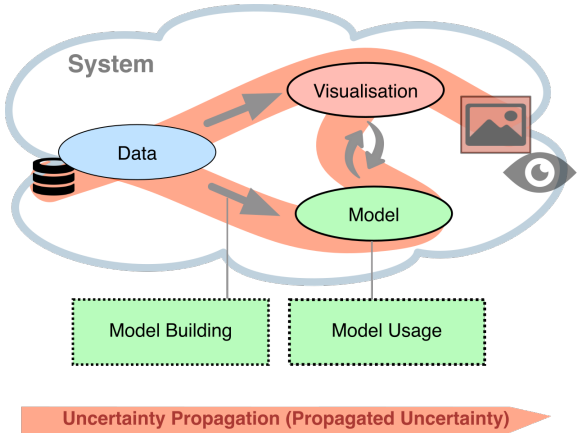


Figure 2.12: Uncertainty propagation through model building and model usage.

Model uncertainty corresponds to the structure of the model and the parametrisation of the model as seen in Figure 2.12. Chatfield (2006) described how uncertainty

is fundamentally propagated in data models that represent real-world phenomena. He also describes the main sources of uncertainty in models. Cullen and Frey (1999) comprehensively address the methods for variability and uncertainty in models, while Lee and Chen (2009) comparatively analyse the various uncertainty propagation methods in terms of their performance.

This guideline specifies methods to identify, alleviate, and assess the propagated uncertainty at the data modelling stage. These methods are as follows.

During the *model building* phase if users have previous knowledge of the model, they achieve a best approximation by typically fitting a parametrised form of the model to the data. Issues of uncertainty arise due to the complexity of the parametrisation (e.g., how many parameters are suitable?) or the appropriateness of the parameters (are the parameters perfect/ good/ bad?), or even the random variation of the model variables. At this stage model calibration introduces a lot of uncertainties by the process of estimating values of unknown parameters in the model. Other uncertainties arise if the distance functions (e.g., euclidean distance or weightings within the similarity function) do not fit data and tasks. Chatfield (2006) classify this type of uncertainties as arising from *model misspecification*.

Such uncertainties arising during the *model building* phase can be lessened by expert background knowledge (e.g., to know which variables to include), and previous experience/information from previous similar datasets (Chatfield, 2006) . However, such expert knowledge may not prevent the user from mistakenly excluding an important variable or adding excess variables. The author points out that one way of avoiding model building uncertainty is to use *nonparametric procedures* that are based on fewer assumptions. One approach to quantify these uncertainties is to use distance functions to measure the distance of parametrisation from the true value.

During the *model usage phase* a lack of previous knowledge of the underlying phenomenon causes inadequacies of the model, which gives rise to structural uncertainties (Chatfield, 2006). He further introduced reasons that give rise to uncertainties during model usage, and they are: (i) specifying a general class of models, where the true model is a special, unknown case, and ii) choosing between several models of different structures.

Additionally, the model carries uncertainties in terms of its suitability to the task at hand. Numerical errors and approximations that occur during the implementation of a model gives rise to algorithmic uncertainty (Kennedy and O'Hagan, 2001).

To propagate the uncertainties in the model selection bias, he further suggested to use the Bayesian averaging approach, and points out the non-triviality of biases. He

recommended replicating the study to check if the new data fits the model, although he makes the point that replicating studies is not all that simple to conduct. Works of Fernandez et al. (2001), and Kennedy and O'Hagan (2001) demonstrate the use of Bayesian approaches to dealing with model uncertainty.

Chapter 4 demonstrates how uncertainties in the outcome are reduced by an iterative specification of the data model.

Guideline 8: Uncertainty Propagation in Visual Mapping

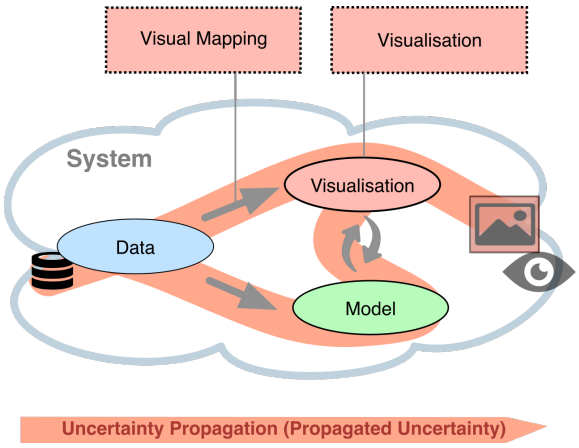


Figure 2.13: Uncertainty propagation through visual mapping and the visualisation.

As seen in Figure 2.13, during the mapping process, the computation of the geometric model (typically done in the mapping process) may be prone to errors due to approximations. Furthermore, the mapping itself causes errors, if the mapping does not fit the underlying data, e.g., when the chosen visual variables do not correspond to the underlying data types. These issues cause uncertainties in this process, which may hinder the comprehensibility of the underlying data. In general, data should be mapped to proper visualisation techniques using the right visual variables (e.g., glyph vs. colour). Uncertainties that occur at the visual mapping stage are mainly due to the use of inappropriate visual variables that do not adhere to the data and task at hand.

Therefore, this guideline specifies how to alleviate propagated uncertainty at the visual mapping stage. This is as follows.

The most sensible approach to alleviate these uncertainties is through analysing the chosen visual variables and metaphors against existing systematic taxonomies. In his *task by data type* taxonomy, Shneiderman (1996) categorised existing information visualisation techniques according to the type of data (e.g., temporal data) and

the task (e.g., zoom or filter). In the case of uncertainty visualisation, we need to consider the added uncertainty dimension to the underlying data. In addition to MacEachren (1992)'s work on manipulating several visual metaphors to represent uncertainty, Buttenfield and Weibel (1988) presented a framework for categorising different cartographic visualisation methods according to the uncertainty elements (e.g., positional accuracy or the lineage of the data) and the measurement scale of the data (e.g. discrete or categorical data). Furthermore, Senaratne and Gerharz (2011) categorised popular uncertainty visualisation methods according to the measurement scale of the data (e.g., continuous or categorical), supported data format (e.g., raster or vector), and the type of uncertainty element in the data (e.g., positional or thematic uncertainty). A taxonomy of uncertainty visualisation is discussed in Section 2.1.3.

Guideline 9: Uncertainty Propagation in Visualisation

The visualisation itself may contain uncertainties as see in Figure 2.13. This is mainly due to the resolution, clutter, and contrast effects of the output visualisation which may hinder the user in gaining insights of the underlying data. Such effects in visualisations that cause uncertainty in the reasoning process are discussed by Zuk and Carpendale (2007) and MacEachren and Ganter (1990).

This guideline specifies how to alleviate propagated uncertainty at the data visualisation stage. This is as follows.

The works of Howard and MacEachren (1996); MacEachren (1992) have developed visual metaphors for representing uncertainty, that fits well with the human cognitive model. Examples are the use of blurring effects, transparency, or coarsely structured surfaces to represent uncertainty. Their impact on decision making under uncertainty has been explored in several studies (e.g., Senaratne et al. (2012)). MacEachren and Ganter (1990) classified visualisation of uncertainties as being developed through two types of errors. *Type 1: seeing what is not really there and Type 2: over-seeing what is really there.* The authors emphasised the need for tools to aid the users in seeing through these type 1 and type 2 errors in visualisations. Relating to the type 2 errors in particular, Brodlie et al. (2012) pointed out to the uncertainties caused by the lower resolution of the visualisation in contrast to the resolution of the data. To alleviate this problem, they suggest that we can use focus-plus-context visualisations to enable the user in viewing data points of interest in full detail, whilst getting an overview of the data at the same time. The off-screen aggregation tool presented by Jäckle et al. (2015) is a tool to solve such resolution bound uncertainties of visualisations.

Guideline 10: Uncertainty Propagation in Model-Visualisation Coupling

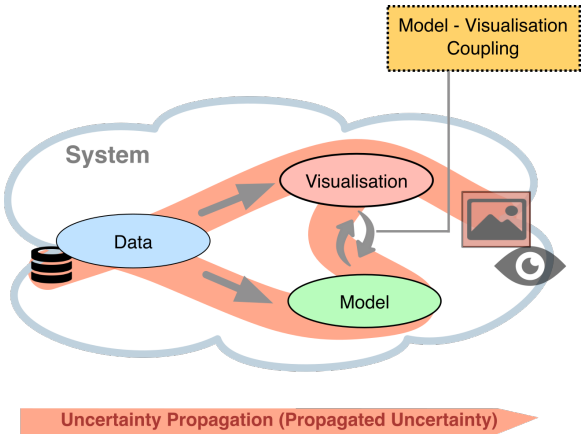


Figure 2.14: Uncertainty propagation through the model-visualisation coupling.

One other aspect that is identified for uncertainty propagation in the system, is the uncertainties caused while coupling the model and the visualisation. This is shown in Figure 2.14. These uncertainties mainly impact the users' interaction with the system and the model steering that is coupled to the visualisation interactions.

This guideline specifies methods to assess the propagated uncertainty at the model-visualisation coupling stage. These methods are as follows.

Endert et al. (2014) proposed an approach where direct interactions on visualisations are directly translated to model steering interactions (e.g., highlighting an item will increase weighting of the models' distance function). If these mappings are not well-designed, these model interactions are translated to model steering interactions that do not fit to the users' intent. Furthermore, the visualisation of the model can be realised in different ways. For example, it is possible to visualise incremental model changes during the training phase (e.g., Fisher et al. (2012)). However, many visual analytics application just visualise the model result.

In the literature review, no methods were found that quantify the uncertainties arising due to the coupling between visualisation and models. One possibility to quantify differences between model and visualisations is to compare measures of the different spaces (e.g., 2D compared to high dimensional spaces in Tatu et al. (2011)) in order to compare model and visualisation characteristics. For example, groups and distances between data items in model space (e.g., between cluster centroids) can be compared to their distances in visual space (e.g., projected distances between cluster centroids). Another approach is to measure how model changes (e.g., via human interaction or data streaming) are propagated to the visualisation. Most of

the visualisations take the final model result but there are several cases, and models that deliver incremental results that can be visualised (e.g., Fisher et al. (2012)).

All these uncertainties are propagated to the final system output which will be observed and used by the human for knowledge generation. As important as it is to account for the uncertainties in a system mentioned above, *which* uncertainties to account for is highly dependent on the application scenario of the data.

2.2.3 Guidelines to Aggregate Uncertainties

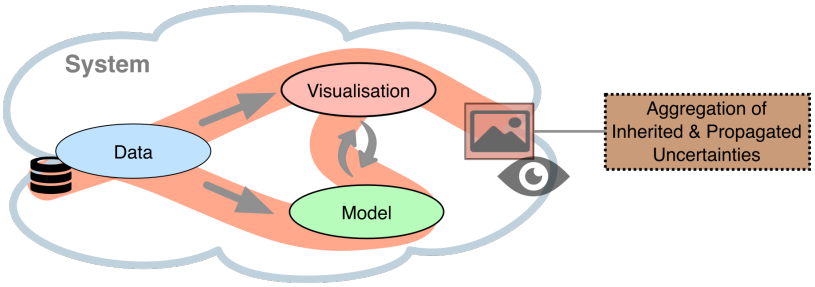


Figure 2.15: Aggregation of inherited and propagated uncertainties.

Systems require powerful and sophisticated techniques to support exploration of large data. Adding different kinds of uncertainty to this data requires an increase in the level of sophistication of the system. In the work of Correa et al. (2009) the data source uncertainty and the propagated uncertainty is estimated through transformations, via sensitivity analysis and error modeling. To simplify the computations, we require intelligent methods to *aggregate* these propagated uncertainties (Figure 2.15).

This guideline specifies methods to assess the aggregated uncertainty at the system outcome. These methods are as follows.

Klir and Wierman (1999) described methods to aggregate source uncertainties and propagated uncertainties in the visualisation pipeline. Also, through a remote sensing classification application, Van der Wel et al. (1998) described the use of an entropy measure to build a weighted uncertainty aggregation measure. They map the different kinds of uncertainty to one measure based on a weighted criteria. Learning from this, an alternative would be for the user of the system to weigh each kind of uncertainty stemming from the system, based on its importance to the use case at hand.

Furthermore, semantic fusion methods can also be utilised for uncertainty aggregation. Based on the work of Castanedo (2013), in a first step the inherited and propagated uncertainty can be incorporated into semantic information. This can be achieved by for e.g., involving the user to utilise formal specifications to annotate the

uncertainties in the system, through annotation tools. In a second stage a pattern aggregation can be performed to provide a semantic interpretation of the aggregated uncertainties.

2.2.4 Guidelines to Visualise Uncertainty Information

Uncertainty visualisation is known to be a most effective medium to communicate such source and propagated uncertainties.

This guideline specifies how to visualise uncertainties in a system. This is as follows.

Griethe and Schumann (2006) presented a pipeline to show the process of uncertainty visualisation. In their pipeline, they differentiated between four different kinds of data flows. (1)*the basic data transformation process through the visualisation pipeline*: is separated into data components and their corresponding uncertainties, such that the user sees the underlying uncertainty; (2) *in/output of the acquisition of uncertainty data*: data at every stage of the visualisation will carry uncertainty, and needs to be considered; (3) *dependencies between the visualisation of the raw data and it's uncertainty*: while the data is explored, it's uncertainty is considered as an integral part of the data. However, decisions in processing the uncertainty is dependent on what raw data is focused on, which rendering techniques and geometric forms for models are chosen. (4) *parametrisation of the pipeline*: uncertainty is not visible by itself at every data component as in (1). Instead, uncertainty will be used to parametrise visualisation of the other data (as done by Schmidt et al. (2004)).

Furthermore, in visualising the uncertainties in the different stages of the data, one needs to carefully consider the different design principles. Works such as of Pang (2001) focused on visualising multi-dimensional uncertainties in data. These can be used as guidelines on how to design visualisations to incorporate different uncertainties propagated through an analysis system. Griethe and Schumann (2006) further emphasised that the decisions on the amount of user interaction on such an uncertainty visualisation process depends on the user's experience and the principles of the visualisation system. Finally, a system should report the uncertainties as cognitive cues about its self-confidence as suggested by Cai and Lin (2010). As a result, users are more comfortable in adjusting their *trust* on the system outputs appropriately.

Chapter 6 introduces several glyph designs for visualising bi-dimensional uncertainties in numerical data, and sets forth a step-by-step evaluation process for these visualisations.

2.2.5 Guidelines to Enable Interactive Uncertainty Exploration

This guideline specifies how uncertainties can be lessened by appropriately incorporating interactive exploration for uncertainty analysis. This is as follows.

Within a visual analytics environment, enabling the user to interact and explore different visualisations for different uncertainties stemming from the different components of the system, will enrich the user's understanding of the true nature of the data, and additionally, how different propagated uncertainties influence the final output. Further, the ability to use a variety of visualisations may also help with illusion type cognitive biases such as clustering and correlation. It is also important to give the user control to decide which of these uncertainties should influence the final output, or with how much importance it should influence the output. Providing the user with the possibility of giving weighted measures for each uncertainty component would be a realistic approach. Furthermore, Correa et al. (2009) presented several approaches including uncertainty projections and visualisations that enable the user to explore the uncertainties of individual data items and the impacts of different uncertainties. Chapter 4 specifically demonstrates how such uncertainties are reduced with the help of interactive exploration with the data.

2.3 Discussion & Future Work

The proposed framework in Figure 2.7 enables users to be informed with uncertainty information and can prevent users from falling into traps concerning mistaken uncertainties and unaware uncertainties. The different guidelines given in the framework for source uncertainties and propagated uncertainties can be tailored to concrete, individual cases where the scope of uncertainties, users, and their tasks are known. These guidelines will be useful to estimate the dynamics of uncertainties in developing visual analytics applications. Depending on its use, the quantification of source/ propagated uncertainties will help users to determine effective visualisation techniques by thinking of the trade-offs between gaining insights and showing uncertainties. Being informed of these uncertainties plays a role in trust building on the user end of the extracted knowledge. Thereby, uncertainty analysis processes are encouraged to be incorporated into visual analytics applications in order to increase awareness, reduce errors (e.g., cognitive biases), and therefore derive trustworthy conclusions fit for the task at hand. In addition to the uncertainty awareness, analytic provenance methods can be used to infer human measures that may give hints on trust building processes.

Combining measures/methods from both sides have the potential to identify relations between uncertainty propagation and human trust building. We have also identified some limitations and open questions that should encourage researchers to investigate this topic further.

First, uncertainties are difficult to be quantified and categorised into a single process. In visual analytics systems, uncertainties can be propagated and implied through the pipelines, as discussed above. Thus, combination of uncertainties from multiple sources could be larger than the sum. Our framework does not provide a quantified model of such intertwined process of uncertainty propagation just yet. As outlined in the guidelines, some efforts have been made to quantify and aggregate different subsets of uncertainty propagation within visual analytics process. Future researchers may need to integrate such efforts using our overarching framework and predict such uncertainty propagation in a specific context.

Second, another open question is whether the transparency of uncertainty propagation is always good and how much of it is beneficial to users. The framework builds upon an assumption that making the uncertainty propagation transparent will let users be aware of variation in their outcomes. However, providing too much information could always confuse, overwhelm, and mislead users, thereby making unwanted human errors. Furthermore, it is also a trade-off between efficiency and accuracy. For instance, applications for human safety, where uncertainty can result in catastrophic results, may need to consider as much transparency as possible. On the other hand, some business analytics may require fast and reasonable analysis results. Thus, it will be interesting to investigate what are proper amounts and methods to communicate uncertainty information to common users of visual analytics. Third, in line with previous points, it is also an open question whether the *awareness* of uncertainties leads to increasing or decreasing user trust in the outcomes. This question may be from the human's trust building process. To build trust in visual analytics outcomes, human users may need to build trust in the visual analytics system first. In this visual analytics knowledge generation process , the awareness of uncertainties may lead to increasing the awareness of the visual analytics process but not to increasing trust in the outcomes. Future research may study further these steps in human trust building.

2.4 Conclusions

Uncertainty is prevalent in our day to day life. Uncertainty has been largely researched in the space time paradigm over the last three decades. At the beginning of this

chapter we give a prelude to the various analysis approaches of uncertainty that use extrinsic, intrinsic, coincident, adjacent, static, and dynamic visualisations. Among the plethora of works on the topic of uncertainty communication, a major gap has been identified that inhibits the users from exploring context-specific uncertainties and to assess their data fitness-for-use. Visual analytics fill this gap by giving the users the ability to interactively analyse their data uncertainties.

In its second part, this chapter describes a conceptual framework that introduces uncertainty to the visual analytics knowledge generation process distinctly in two stages: uncertainties that are inherent in the data, and uncertainties that are propagated from data, to data models, to data visualisations, to data model-visualisation couplings. As an outcome this chapter identifies guidelines for uncertainty analysis within a visual analytics knowledge generation process. The following chapters instantiate some of these guidelines for selected types of data, thereby introducing novel visual analytics approaches for uncertainty analysis. The following chapters further demonstrate the usefulness of uncertainty analysis within several use-cases.

Chapter 3

Uncertainty Analysis of Image-based Volunteered Geographic Information

Contents

3.1 Background & Related Work 66

3.2 Reverse-viewshed Analysis for Assessing the Positional Accuracy of Image-based VGI 69

3.2.1 Flickr Metadata Retrieval with the FlickrMetaCrawlr . . . 71

3.2.2 Reverse-viewshed Analysis for POIs 72

3.3 Credibility as an Uncertainty Indicator for Flickr Images 75

3.4 Discussion & Future Work 82

3.5 Conclusions 86

With the increased availability of user generated data, assessing the uncertainty of such data becomes important. In this chapter, a novel technique to interactively assess the uncertainty of image-based VGI is introduced. Existing state of the art work for uncertainty analysis of image-based VGI mostly use textual tags together with visual cues from the image content to infer the uncertainty of these geographic data. These approaches either require the analyst to be familiar with the surrounding geography that they are analysing for the images, or are based on the assumption that the textual tags in all cases are highly accurate. Learning from the guidelines in Chapter 2 Section 2.2.1, due to limitations in spatial knowledge of contributors, such textual tags of images alone cannot be used to infer the uncertainty of images. These guidelines also show us that the quality of the textual labels help to determine the credibility of the contributors. Based on these guidelines, approaches are introduced to

(1) assess the positional accuracy of image data, and (2) utilise the positional accuracy of images to infer the credibility of contributors.

In our approach the positional accuracy of images is determined based on the *visibility* from the camera position (observer point) to the target position (target point). The visibility is assessed by computing a line-of-sight between the observer point and the target point based on in-between surface elevation data. If the location of the target lies within the visibility from the observer point, then the image is considered to be correctly geotagged. This technique, which we call the *reverse-viewshed analysis* can be used to assess the positional accuracy for every image. The positional accuracy is further used as a reference measure and is inspected against the textual tags of each image to infer the credibility of the images and image contributors.

This chapter unfolds as follows: In Section 3.1 related work on visual analytics approaches for uncertainty analysis is reviewed, along with issues that attribute to uncertainties in image-based VGI. Section 3.2 introduces the novel approach of a reverse-viewshed analysis for the assessment of positional accuracy in image-based VGI. Section 3.3 describes how the reverse-viewshed analysis is used to infer the credibility of image-based VGI.

This chapter is based on the publications Senaratne et al. (2013a) and Senaratne et al. (2013b)[1].

3.1 Background & Related Work

Image-based VGI, such as Flickr[2] are community contributed images with spatial references to them. A spatial reference is added in terms of a geotag, which could be either geographic coordinates (e.g., 47.660941, 9.181073 for the Emperia statue in Konstanz) and/or a textual label (e.g., "Emperia in Konstanz"). Mostly due to a lack of spatial knowledge or expertise the contributors of these images often incorrectly geotag these images, giving rise to uncertainties. As of May, 2015, Flickr has reported to host over 10 billion images, where around 3% of these Flickr images are geotagged, and Rinner et al. (2008) identified an exponential growth for such VGI.

In case of Flickr, as an example of a platform for visually generated VGI, contributors can upload photographs to share them with others. A Flickr user can maintain

[1]Appears in Sections 3.2 and 3.3. Both of these publications are a result of a collaboration with A. Broering from the University of Muenster and T. Schreck from the University of Konstanz. My contributions as the first author within these publications were the definition and implementation of the reverse-viewshed analysis as an uncertainty assessment technique for image-based VGI.

[2]`https://www.flickr.com/`

a profile to which uploaded photos are linked and to state metadata such as his/her real name, the date of registration, hometown, or contacts to other contributors/users. Also, metadata for the picture itself can be specified, such as title, caption, textual tags describing the photo (label), or the dates of capture and upload. Additionally, a spatial reference of the photo can be given in form of geographic coordinates. This geotag can be either produced by an external GPS device, automatically recorded with a camera built-in GPS, or it can be manually located using Flickr's map interface at varying levels of resolution (i.e., neighbourhood, city, country).

Additionally to the geotag that consists of geographic coordinates, Flickr contributors often specify the place of interest to which the picture relates, as textual tags. The map shown in Figure 3.1 displays all geotags of Flickr photos annotated with the textual tags "Angkor" and "Cambodia". Although most of the photos of this dataset are geotagged within the area of the ancient city in Cambodia, this visual analysis shows that there are also many pictures being located far away from it. For example[3], one photo displaying a site of Angkor on Flickr is geotagged at a location in California. Becker and Bizer (2009) further demonstrated through their work on the Flickr Wrappr, how images on Flickr are incorrectly geotagged.

Figure 3.1: Geotags of Flickr photos that were textually tagged as "Angkor" and "Cambodia". This figure appeared in Senaratne et al. (2013a).

To understand how uncertainties occur, in this section we discuss the tagging behaviour on such image-based VGI platforms and the existing semi-automatic approaches for uncertainty analysis in image-based VGI.

Flickr images have been explored in a multitude of geographical analyses. For instance, Jankowski et al. (2010) and Crandall et al. (2009) explored spatial and temporal patterns in user movement and, their interests in landmark and events captured through Flickr. These Flickr images are organised or searched with the help

[3] `http://www.flickr.com/photos/rbleib/5030263322/in/set-72157624911484519/`

of their accompanying tags that come in various forms. Ames and Naaman (2007) have comprehensively discussed the concept of tagging and have identified two main incentives that motivate contributors to tag: (1) *sociality*: describing who is intended to use the tag, (2) *function*: describing the intended usage of the tag, which could be either for organisational or retrieval purposes, and also to gain attention for the tagged content. Tagging an image is a means of adding metadata to the content in form of specific keywords to describe the content (Golder and Huberman, 2006), or in the form of geographic coordinates (Geotagging) to identify the location linked to the image content (Valli and Hannay, 2010). Friedland et al. (2011) and Moxley et al. (2008) developed semi-automatic tools that suggest tags for a given image, based on the geographic context and visual relevance. The algorithm of Moxley et al. (2008) gives weights to labels that refer to events, neighbourhoods, pertinent objects, and activities in a region, that help eventually to improve the tag suggestions.

Crandall et al. (2009) analyse the content of a photo based on text labels and image data, and the structure based on the geospatial data. They further assert that within a street level scale, text tags alone can be a useful source to estimate the location, but in combination with visual cues it can be an even stronger component in validating the location. Furthermore, Girardin et al. (2008) analysed tags of Flickr photos to explore how people perceive their environment, and the underlying semantics of how they describe the urban space. In a similar study, Sigurbjörnsson and Van Zwol (2008) found that most frequently tags represent a location followed by artifacts/objects.

When consuming such community contributed images, it is important to keep in mind that the content is not quantified by the objective notions of data quality, nor does it rely on traditional authorities who enforce data quality standards. Instead, uncertainty indicators such as the credibility of the data depends on the personal accuracy of the data contributors, and Bishr and Kuhn (2007) indicate that trusted contributors provide more useful data. This contributor-trust issue led Goodchild (2009) and Coleman et al. (2009) to categorise contributors of VGI into different groups based on their knowledge and experience with geographic information, and the motivations that drive the contributions.

Goodchild (2009) classified data producers as falling into either *Neo Geography* or *Academic Geography*. Neo Geography is where the role of the contributor intersects between the roles of subject, producer, presenter, and consumer. I.e., there is no clear role of the contributor belonging to any one of these distinguished roles. However in contributing to VGI, they are all experts in their own local communities. On the contrary, contributors falling into academic geography are involved in professional

geography, such as surveyor or cartographer. Coleman et al. (2009) classified data contributors as overlapping between *Neophytes*, *Interested Amateur*, *Expert Amateur*, *Expert Professional*, and *Expert Authority*. They analysed these groups based on what motivates contributors to produce data on VGI platforms. Coleman et al. (2009) further implied that contributors fall into the above categories depending on three different contexts: *Market driven*, *Social networks*, and *Civic/Governmental*. Contributors who fall into the category of Market driven contribute data on commercial databases or services such as TomTom[4] or Garmin[5]. Contributors falling into Civic/Governmental contribute data out of concern to their city/society, for example to PPGIS[6]. Contributors falling into Social Networks contribute to platforms such as OpenStreetMap, Flickr etc.

In a characterisation of contributor behaviour on Flickr, Van Zwol (2007) shows that the number of contacts per contributor and the number of pools an image belongs to can be used to predict the popularity of a photo. He further asserted that the social affiliation which is sustained by the network of contacts within Flickr, is important for the popularity of their photos.

Building up on these works, a novel approach for assessing the positional accuracy of geotagged Flickr images based on the *line-of-sight visibility* is introduced here. A reverse viewshed analysis is proposed as an objective baseline measure for positional accuracy which can further serve for additional investigations on what *characteristics* of a VGI volunteer influence the *credibility* of his/her contributions. Flickr is taken as the experimental data source, however, the approach is more generic and applicable to estimate the uncertainty of any image-based VGI source where geographic coordinates and textual tags, which denote an object or place of interest, occur.

3.2 Reverse-viewshed Analysis for Assessing the Positional Accuracy of Image-based VGI

A *viewshed* analysis can be conducted to determine the total area that is visible from a given point (O'Sullivan and Unwin, 2010). The parameters that are used to control the viewshed calculation between two given points are shown in Figure 3.2. Viewshed analysis is carried out in a variety of applications including but not limited to urban environment planning (Lake et al., 1998), locating telecommunication

[4] `www.tomtom.com`
[5] `www.garmin.com`
[6] `www.ppgis.net`

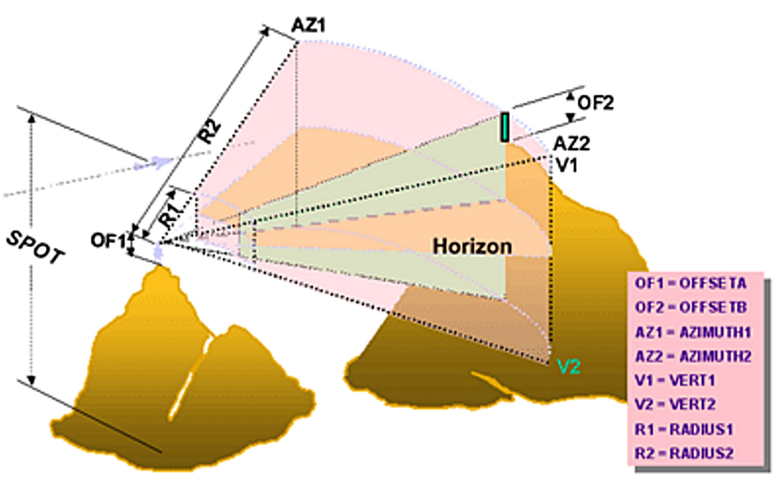


Figure 3.2: The parameters for a viewshed calculation. (Source: `http://www.esri.com/software/arcgis`).

towers (De Floriani et al., 1994), or tree cover conservation (Sherren et al., 2011). A viewshed of a particular point is calculated from surface elevation data around the region, which is employed in an algorithm that estimates the difference of elevation of the intermediate pixels between the viewpoint and the target pixels. In order to determine the visibility of the target pixel, the intermediate pixels are analysed for their line of sight (LoS). A line of sight determines if the target pixel is visible from the viewpoint, or obscured. If visible then the target pixel is included in the viewshed, if obscured then the target pixel is not included in the viewshed (Kim et al., 2004).

Amongst many who developed efficient viewshed algorithms (Fisher, 1991, 1993b; Wang et al., 1996), Fisher (1996); Kidner et al. (1999); Ralling et al. (1999) also discussed reverse viewshed analyses. A reverse viewshed analysis holds the same principles as the viewshed analysis. However, it is utilised to determine the visibility of a given target point from many observer points (Fisher, 1996). Fisher (1996) distinguished between the area which can be seen from the location (viewshed) and the area from which a location can be viewed (reverse viewshed), based on the height differences between the viewing point and the viewed object.

Taking this into consideration, the same technique to generate a viewshed is utilised here, but a different procedure is employed. I.e., instead of taking one viewshed from the target point, multiple viewsheds from the *observer points* are created to validate if the target falls within the visibility of the observer. This reverse viewshed analysis is used to determine the visibility of two prominent points of interest (POI) in Berlin (Germany), the *Brandenburg Gate* and the *Reichstag*, from the surrounding observer

points. The subsequent steps of the analysis work flow are shown in Figure 3.3. These steps are discussed in more detail in the following sections.

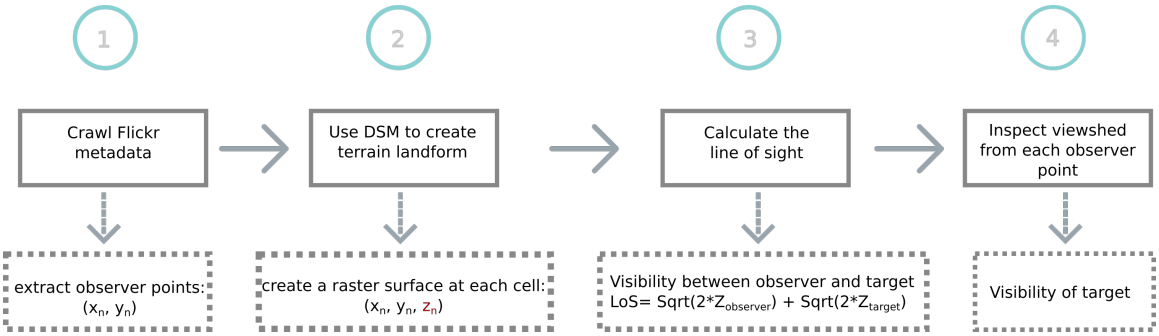


Figure 3.3: Work flow diagram for positional accuracy analysis within image-based Flickr.

3.2.1 Flickr Metadata Retrieval with the FlickrMetaCrawlr

The approach is developed and tested by experimental analysis. As a first step (Figure 3.3), Flickr is crawled and metadata of images for the two POIs, which are textually tagged as "Brandenburg Gate", "Berlin" and "Reichstag", "Berlin", is extracted. For each POI, 100 images from Flickr are considered.

To make metadata of Flickr images available for the developed process and the viewshed analysis, a tool has been implemented, the so-called *FlickrMetaCrawlr*. This tool is able to programmatically download metadata of Flickr photos and its contributors. The FlickrMetaCrawlr therefore relies on the open Flickr API[7] and fetches metadata of Flickr photographs for a specified set of tags. The Flickr API restricts applications to access a maximum of 5,000 photos in a single API query execution. However, a certain tag combination may result in a much larger number of photos - e.g., searching for "Times Square" and "New York" results in around 15,000 geotagged photos. Hence, a mechanism has to be included that divides the initial query into sub-queries which result in less than 5,000 photos. Therefore, to facilitate access to all photographs that confine to a tag query, the FlickrMetaCrawlr utilises a quadtree algorithm (Samet, 1984).

The quadtree is essentially applied to the geographic space and subdivides it recursively into four quadrants starting with the maximum extent (the bounding box of between 180°W, 90°S and 180°E, 90°N). A division into four quadrants is performed in case more than 5,000 photos are contained within a bounding box. Finally, for all defined quadrants (each containing less than 5,000 photos) separate API queries

[7] `https://www.flickr.com/services/api/`

can be executed. This way selected metadata such as geotags of images, tag count per image, image accuracy, user ID, user contact count, and number of photos per user were downloaded (from the public photo pool) for images textually labelled as "Brandenburg Gate" and "Berlin" as well as "Reichstag" and "Berlin".

The retrieved metadata for images for the POIs are further filtered based on the scale at which the images were geotagged. This scale is called accuracy in Flickr which is derived from the zoom level of the map. The accuracy varies between 1 and 16, while 1 being at the world level and 16 being at the street level and representing the highest accuracy in Flickr. We extracted the metadata for Flickr images which have been geotagged at street level. The retrieved geotags of the images are considered as *observer points* from where the photographs were taken.

3.2.2 Reverse-viewshed Analysis for POIs

In the second step of the work flow (Figure 3.3), the Spatial Analyst tool-box from Esri's ArcGIS suite (McCoy et al., 2001) is utilised together with a Digital Surface Model (DSM) to create the terrain landform in Berlin City, which is required for the reverse-viewshed calculation. An excerpt of the study area with the available DSM data is shown in Figure 3.4. The DSM represents the earth's surface, including the elevation of man-made buildings as well as the heights of the surrounding vegetation in our area of interest. These surface heights are derived from IRS-P5 Cartosat-1 in-flight stereo data with a 5m post spacing and a relative vertical accuracy of 2.5m with linear error of 90% (LE90). With these data, a raster surface is created, where each cell contains a geographical coordinate pair and an elevation value (x, y, z). For an observer n it is (x_n, y_n, z_n).

In the third step (Figure 3.3), the LoS between each observer point (geotag of each image) and the target point (Bradenburg Gate or the Reichstag) is calculated. LoS is essentially determined by the following formula:

$$LoS_(n, t) = \sqrt{2 * z_n} + \sqrt{2 * z_t}$$

where, n is a given observer, t is the target point, z_n is the surface elevation of the observer point and z_t is the surface elevation of the target point. Based on the LoS from a given observer point it is determined if the target point was in the vicinity to the observer or not. This calculation creates a viewshed raster layer. The resulting viewshed raster layer indicates in a binary form which cells are visible and which are not: visible cells with a value of 1 and non-visible cells with a value of 0. As an

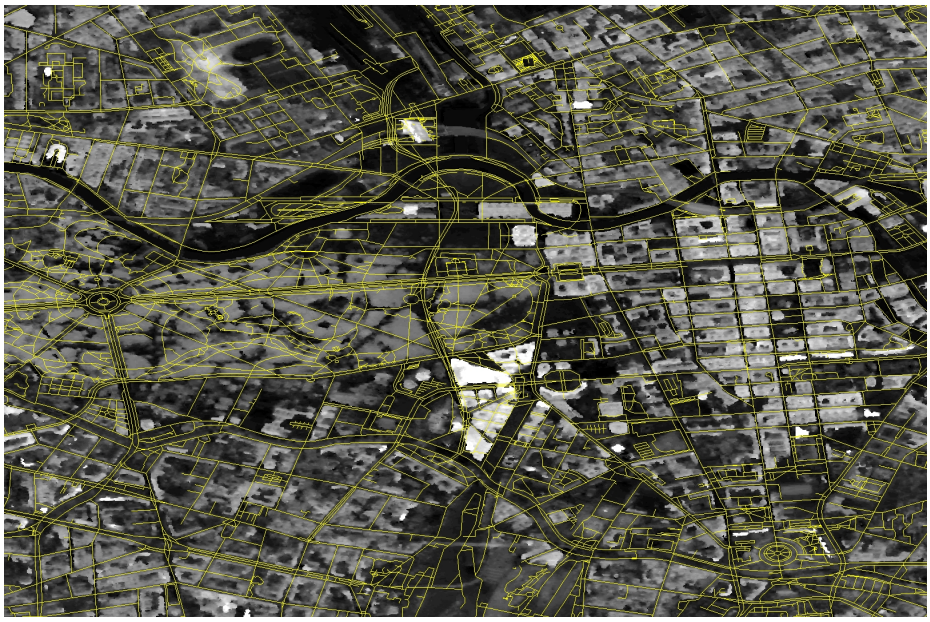


Figure 3.4: An excerpt of the study area in Berlin overlayed with the DSM.

example, in Figure 3.5 the non-visible cells altogether are indicated in pink colour. When the cells are within the LoS of the observer, then these cells are indicated in green colour. Thereby, in the fourth step of the work flow (Figure 3.3), for each of the images' observer points a viewshed is created and these are manually inspected to verify the positional accuracy.

If the calculated area of visibility includes the target position (Brandenburg Gate or Reichstag), the image is considered to be correctly geotagged (Figure 3.5 and Figure 3.6; green polygons). If the image content further represent the POI, the image is also considered as correctly labelled. If the calculated area of visibility does not include the target position, the image is considered to be incorrectly geotagged (Figure 3.5 and Figure 3.6; pink polygons), as according to the LoS measurements the observer could not have seen the POI. If the image content does not represent the POI, it is considered as incorrectly labelled. These considerations result in four different categories an image can belong to: (a) images that are incorrectly geotagged and incorrectly labelled, (b) images that are incorrectly geotagged, but correctly labelled, (c) images that are correctly geotagged, but incorrectly labelled, and (d) images that are correctly geotagged and correctly labelled. These four categories within the Brandenburg Gate and the Reichstag use cases are depicted in Figure 3.5 a-d and Figure 3.6 a-d, respectively. It should also be noted here however, that photographs that were taken from an elevated location such as a higher floor of a building are disregarded in our analysis, as only the surface elevation of the ground are considered

Figure 3.5: Reverse-viewshed from four exemplar observer positions (indicated with the arrow head, and the image taken from the position) to the Brandenburg Gate (indicated with the red rectangle). (a) image is incorrectly geotagged and incorrectly labelled, (b) image is incorrectly geotagged, but correctly labelled, c) image is correctly geotagged, but incorrectly labelled, and d) image is correctly geotagged and correctly labelled.

for each observer point in the viewshed computation.

Images that are geotagged out of the visibility range (POI falls in pink coloured areas) are considered to either misrepresent the location from where the image was taken, or the image content represents something else other than the POI but tagged as the latter. Images belonging to either of these two categories are considered to be representing incorrect location of the point of interest, and therefore lacks positional accuracy.

A reverse-viewshed successively determines from which observer points the point of interest is visible. This allows cross validating if an image was taken within the vicinity to the point of interest. Images belonging to observers whose line of sight does not include the position of the POI are regarded as incorrectly geotagged, and images belonging to observers whose line of sight includes the positions of the POI are regarded as correctly geotagged. In a successive step (Section 3.3), the various *user/photo metadata features* of the images are analysed to explore how these features can be used in association with the reverse-viewshed, with the aim of automatically classifying VGI contributors concerning their *credibility*.

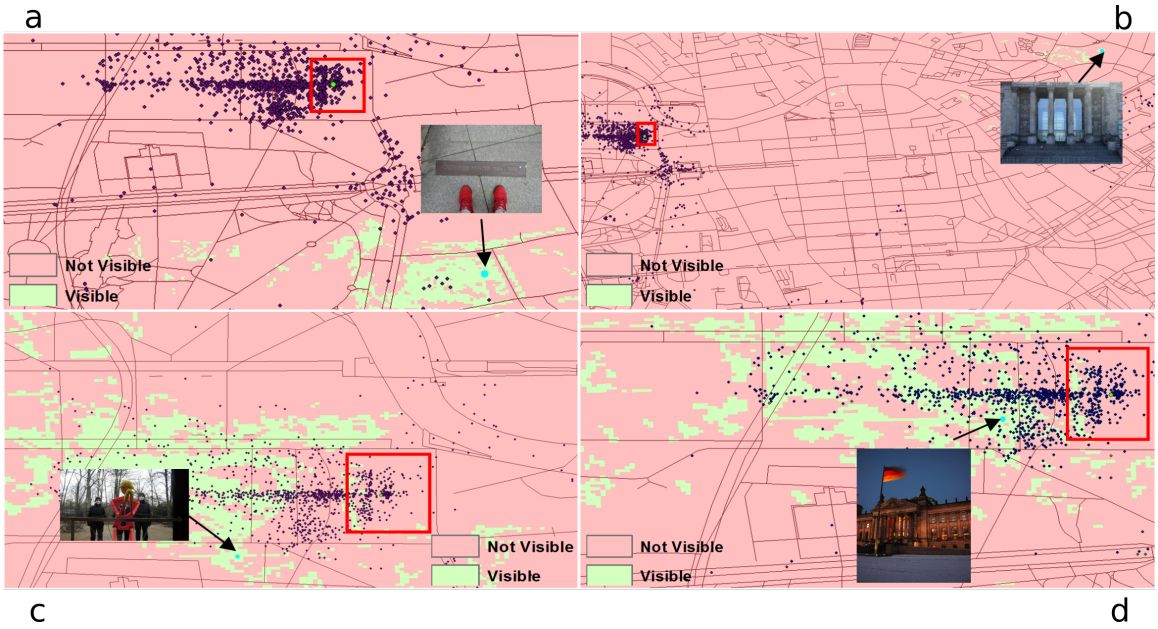


Figure 3.6: Reverse-viewshed from four exemplar observer positions (indicated with the arrow head, and the image taken from the position) to the Reichstag (indicated with the red rectangle). (a) image is incorrectly geotagged and incorrectly labelled, (b) image is incorrectly geotagged, but correctly labelled, (c) image is correctly geotagged, but incorrectly labelled, and (d) image is correctly geotagged and correctly labelled. These figures appeared in Senaratne et al. (2013a).

3.3 Credibility as an Uncertainty Indicator for Flickr Images

With massively increased production and availability of user generated geospatial data, considering the data *credibility* becomes a pressing issue. Flanagin and Metzger (2008) expressed the importance of assessing the subjective and objective nature of data credibility, which is a combination of trust and expertise. Frew (2007) described how metadata about VGI can provide a basis for the judgment of quality of these data sources.

In this section we explore how we can build up on the described approach for assessing the location correctness of image-based VGI in Section 3.2, towards inferring the credibility of VGI contributors in Flickr. As a result, credibility indicators for image-based VGI contributors out of these user and image features are derived.

To achieve this, we propose analysing the variability of selected user and photo metadata features of geotagged Flickr photos in reference to the location correctness of these images as derived in Section 3.2. We investigate which metadata of photographs (e.g., tag count of photographs, comments count of photographs, etc.) as well as

metadata about contributor (e.g., the number of photos, the number of contacts, or the used camera) can be utilised to eventually infer the *credibility* of contributors (credible contributors produce trustworthy content), using the location correctness of the images as the reference measure.

Related research such as Van Zwol (2007); Castillo et al. (2011); Gupta et al. (2012) utilised various VGI contributor metadata to derive conclusions and to characterise the contributor. Van Zwol (2007) takes the number of contacts of a contributor as the predictor for the expected popularity of a photo within the Flickr data source. Therefore, it can be assumed that the contributor contacts number characterises to a certain degree the popularity of the user.

Further, Castillo et al. (2011) and Gupta et al. (2012) showed for Twitter data how contributor-based features, such as the friend count and contribution frequency, associate with information credibility (As also shown in Chapter 4). This shows that contributor features can be used as a rich source of information to derive characteristics about the contributor and their produced content.

Based on these works, and in combination with the reverse-viewshed as a reference uncertainty measure, we can explore which metadata features show a causal relationship with correctly and incorrectly geotaged images. For each of the two selected POIs (the Brandenburg Gate and Reichstag), we analysed 100 geotagged Flickr images, each for its image content together with its photo and contributor metadata. The analysis is summarised in Table 3.1 and Table 3.2.

Table 3.1: The categories of images within the sample dataset falling into correct/incorrect geotagging and labelling.

Category	Correct Geotag	Correct Label
a	No	No
b	No	Yes
c	Yes	No
d	Yes	Yes

Table 3.2: The statistics of each metadata feature for image categories a, b, c, and d.

	Brandenburg Gate				Reichstag			
	a (30%)	**b (19%)**	**c (11%)**	**d (40%)**	**a (27%)**	**b (11%)**	**c (25%)**	**d (37%)**
Avg. user tag count	18	8	13	11	35	12	22	10
Avg. user photo count	19	4	18	5	8	8	10	3
Avg. user contact count	338	111	134	132	108	141	153	110
Avg. distance to the target (m)	626.5	402.9	299.1	161.6	1321	735.9	510.5	436.6

In Table 3.1 the photos are classified as **a** (wrong geotag and wrong label), **b** (wrong geotag but correct label), **c** (correct geotag but incorrect label), and **d** (correct geotag and correct label).

Table 3.2 presents the variation of each metadata feature within the four image categories a, b, c, and d for Brandenburg Gate and Reichstag. The descriptive statistics of these metadata features are presented in Figures 3.7, 3.8, 3.9, 3.10 for Brandenburg Gate, and in Figures 3.11, 3.12, 3.13, and 3.14 for Reichstag.

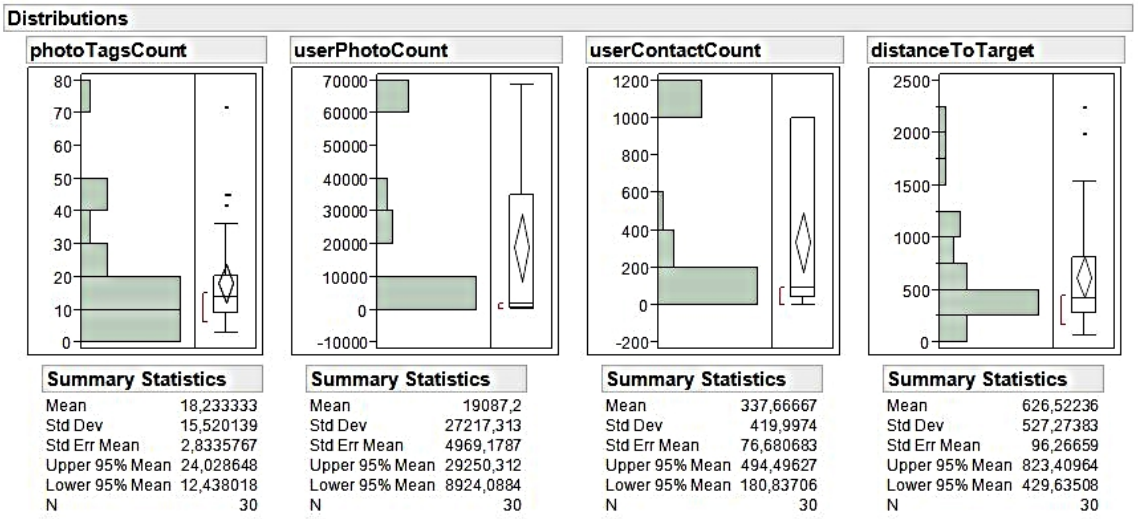


Figure 3.7: Distribution of data for category 'a' within the Brandenburg Gate use case.

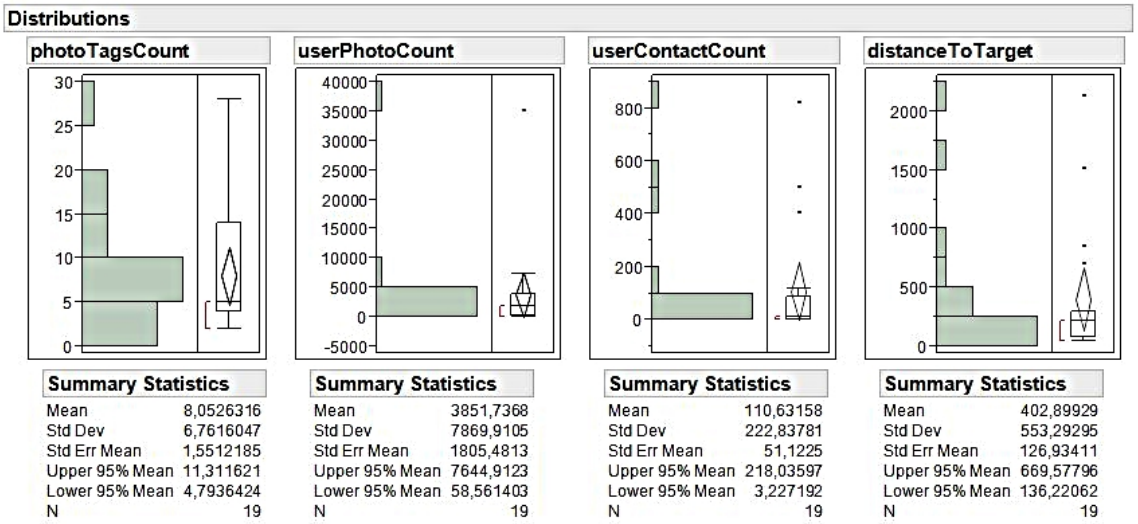


Figure 3.8: Distribution of data for category 'b' within the Brandenburg Gate use case.

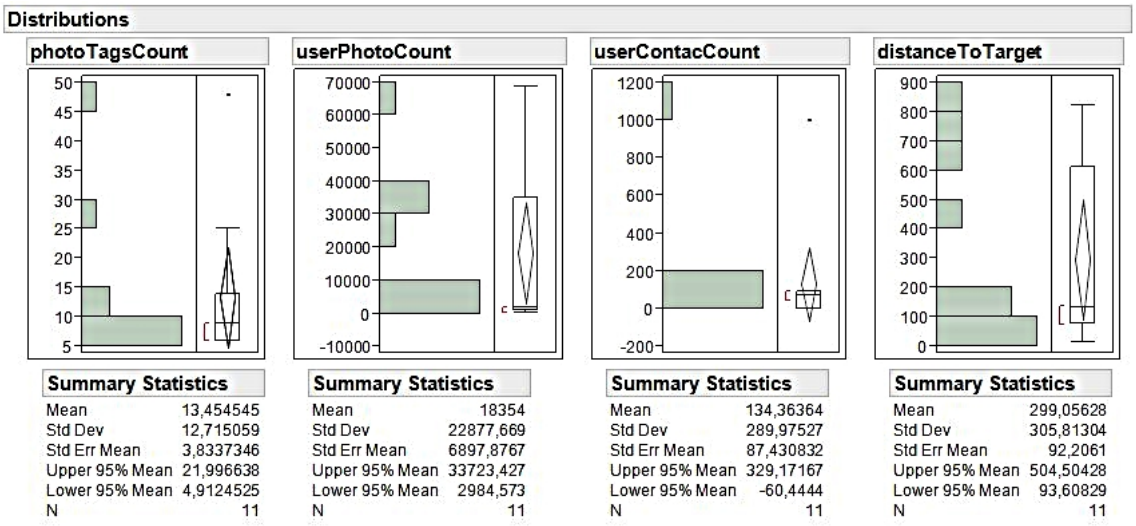


Figure 3.9: Distribution of data for category 'c' within the Brandenburg Gate use case.

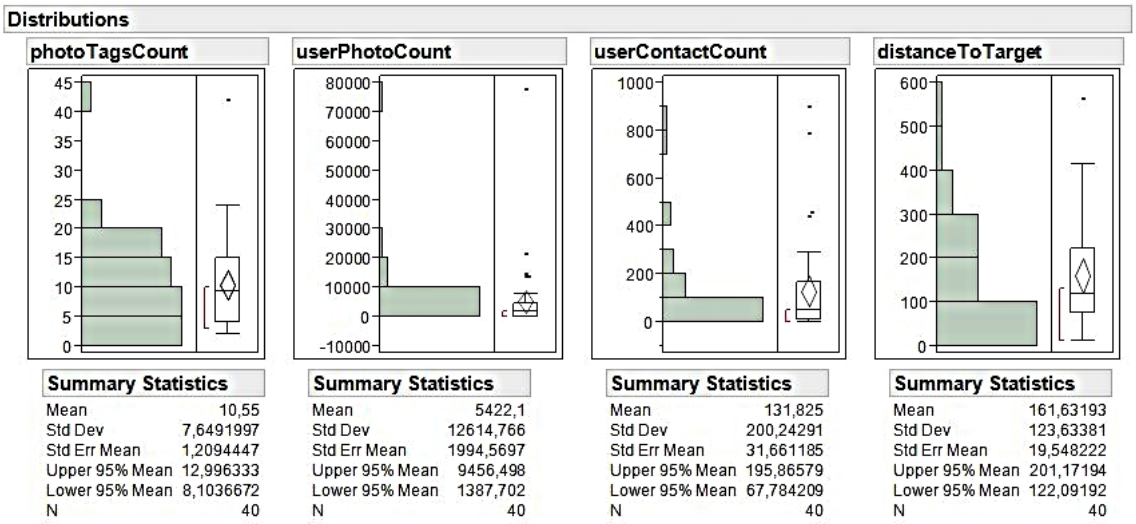


Figure 3.10: Distribution of data for category 'd' within the Brandenburg Gate use case.

We can observe interesting patterns within the gathered data. Contributors of photos within category (b) and (d) (Figures 3.8, 3.10, 3.12, 3.14) for both POIs have on average the lowest number of contacts (on average 121 contacts for "Brandenburg Gate" images and 125 contacts for "Reichstag" images), as compared to contributors of photos with incorrect labels in categories (a) and (c) (Figures 3.7, 3.9, 3.11, 3.13) who have on average 236 contacts within "Brandenburg Gate" images and 130 contacts within "Reichstag" images. This may explain the motivation and thus different priorities of contributors when contributing to VGI as also described by Coleman

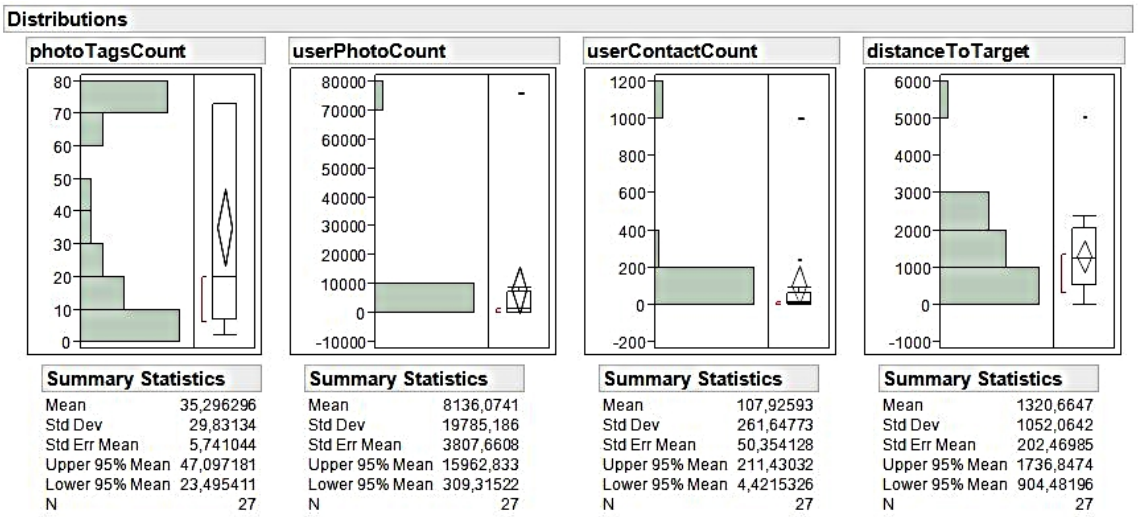


Figure 3.11: Distribution of data for category 'a' within the Reichstag use case.

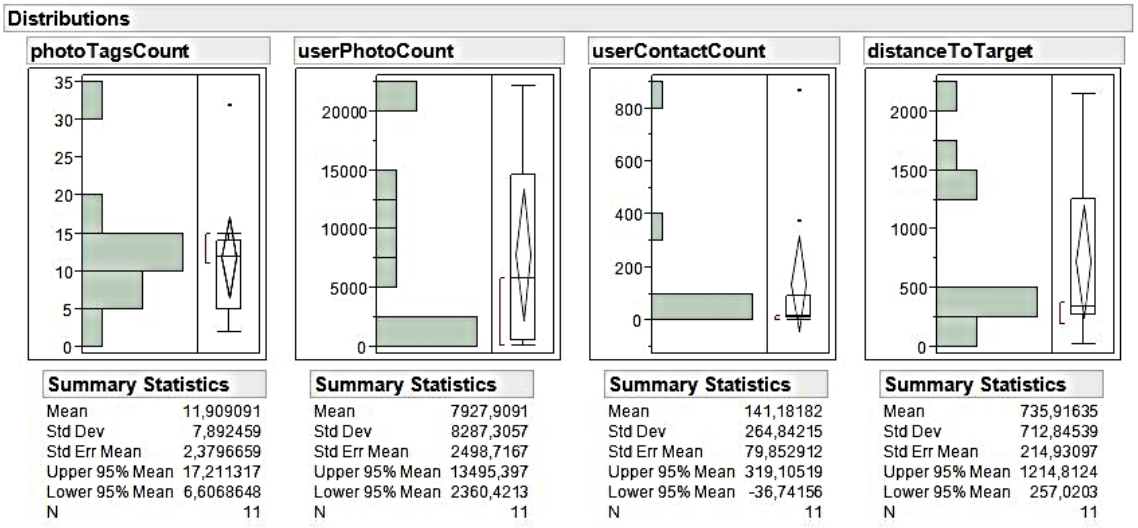


Figure 3.12: Distribution of data for category 'b' within the Reichstag use case.

et al. (2009). Contributors who have correctly labelled their images tend to have on average lower number of contacts in comparison to contributors falling in to the remaining categories. Hence, popularity in Flickr may not be a priority for this group of contributors, while priority in quality is.

Furthermore, the average number of photos produced by contributors within each category was analysed. This also revealed a pattern of correct and incorrect image labelling. Contributors of photos of category (a) and (c) (Figures 3.7, 3.9, 3.11, 3.13), with incorrect labels, have contributed significantly more photos over the years of

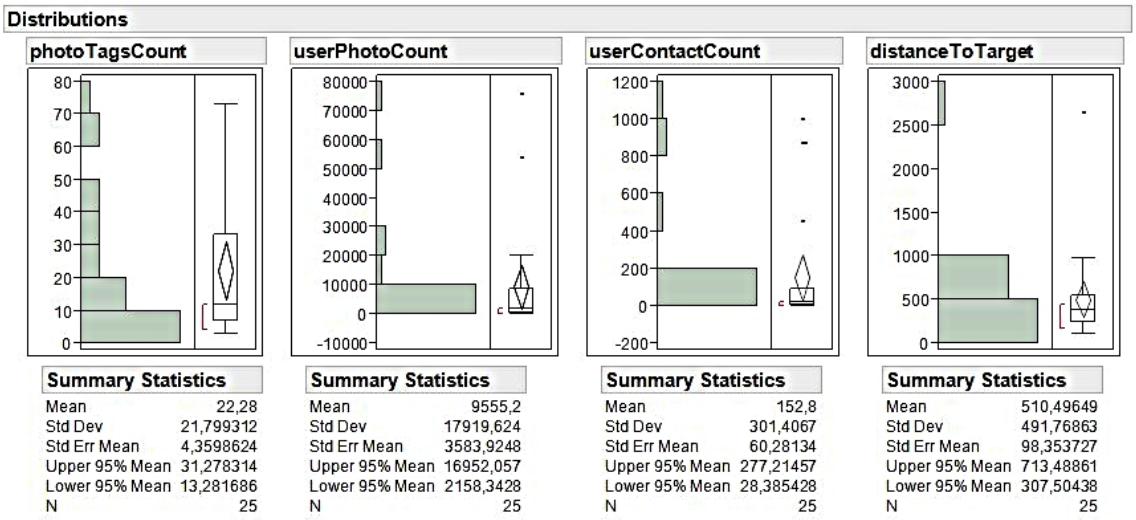


Figure 3.13: Distribution of data for category 'c' within the Reichstag use case.

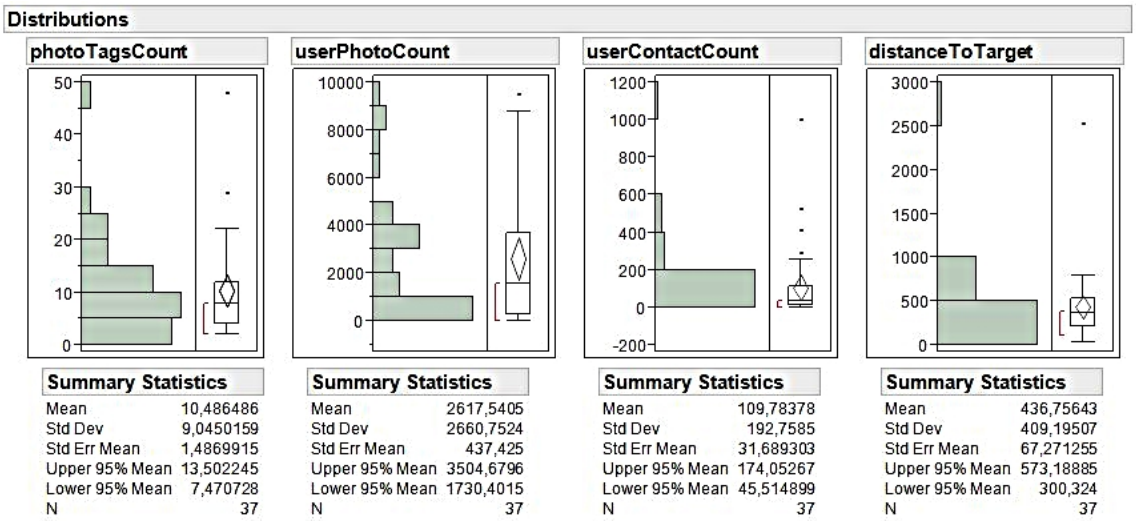


Figure 3.14: Distribution of data for category 'd' within the Reichstag use case. These figures appeared in Senaratne et al. (2013b).

their participation on Flickr. The average photo count of photo contributors for POI Brandenburg Gate in category (a) is 19,087 and for category (c) is 18,354, while for category d it is (5,422) and for category (b) it is 3,852. The average photo count of photo producers for POI Reichstag in category (a) is 8,136, category (c) is 9,555 while for category (b) and (d) it is 7,928 and 2,618 respectively.

Looking into the photo metadata, the average number of tags per photo further reveals a pattern in the above image categories. Photos for Brandenburg Gate within

categories (a) - 18 tags and (c) - 13 tags, have on average the highest number of tags. These photos are incorrectly labelled. Whereas photos in category (b) - 8 tags and (d) - 11 tags have the lowest number of tags on average and are also correctly labelled. Likewise, photos for Reichstag within categories (a) - 35 tags and (c) - 22 tags have on average the highest number of tags per photo, and photos in category (b) - 12 tags and (d) - 10 tags have the lowest number of tags on average and are also correctly labelled.

Further, we have computed the distance to the target by taking the orthodrome between the geotag and the actual geographical coordinates of a point of interest. This reveals that the average distance to the target decreases for images from (a) to (d) within the use cases for Brandenburg Gate as well as Reichstag. Images in category (a) have the highest averaged distance to the target and in category (d) have the lowest averaged distance to the target (Table 3.2). The closer to the point of interest a person is, the more focused the object would be in the image, thus, allowing the person to geotag/label more precisely. The further away from the point of interest, the person might become more imprecise when geotagging and labelling the image.

To determine the significance of the above observations we conducted a Chi-Square test for independence (Pearson, 1900) for both use cases. As seen in Figure 3.15 a pearson coefficient less than 0.05 can be observed in both use cases. Therefore, by rejecting the null hypothesis we conclude that there is a significant dependency relationship between the selected user metadata features and the observed patterns in the photo categories a - d. These meatadata features can be used generically to validate similar image-based VGI data sources.

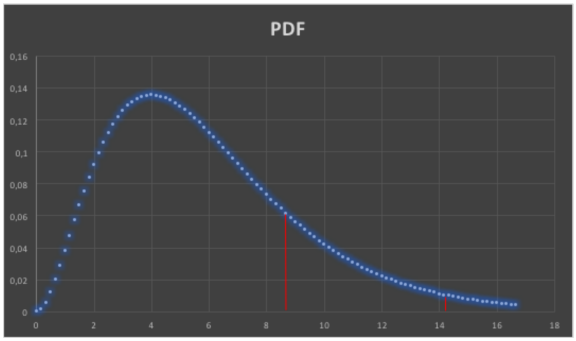
The above observations can be considered as triggers to look further into these findings. They will enable us to infer the user credibility within similar VGI sources, and in general to understand qualitative aspects in contributor provided data much better. In addition to the location correctness, other features such as the label precision, or image content can be used to evaluate the user credibility. Methods to utilise these features in combination to assess user credibility are discussed in the following sections.

3.4 Discussion & Future Work

A reverse-viewshed is carried out to assess the location correctness of geotagged Flickr images that confirm to a particular point of interest through the geotag and the image label. Images placed within a visibility region that do not include the

Chi-sq test for independence Use case: Brandenburg Gate	
X^2 Test statistic =	14,3388597
Critical X^2 value =	8.60
P-Value =	0.026071661
Decision is	Reject H_0

Chi-sq test for independence Use case: Reichstag	
X^2 Test statistic =	23.70322333
Critical X^2 value =	14.2
P-Value =	0.00059213
Decision is	Reject H_0



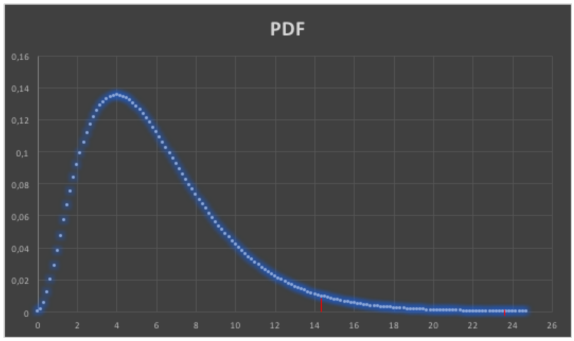


Figure 3.15: Result analysis for the Brandenburg Gate and Reichstag use cases using the Chi Square test for independence.

position of the point of interest are determined to be either incorrectly geotagged, or incorrectly labelled, or both. We have to consider possible reasons for these outliers. An obvious reason are mistakes made by the contributor when geotagging a photo. Such mistakes can for example result from either manually adding wrongly measured coordinates as a geotag to the photo, or coordinates measured by a malfunctioning GPS device. Another reason might be that, while the geotag is correct, a user lacks sufficient knowledge about what is shown on the photograph and provides incorrect place describing tags. Also, we have seen cases within the data sets, where it seems that contributors have made touristic round trips and collectively tagged their taken photos with all places visited during that trip. For example, a tourist visiting several places in Germany defines the same tags (including "Brandenburg Gate") for all taken photos during his/her trip and bulk uploads them as a photo set to Flickr.

The cases above can be clearly considered as wrongly tagged photos and lowering the credibility of the producers of such photos would be valid. Other outliers cannot be as easily considered as being wrongly tagged. In particular, when extracting data for a particular place of interest based on their textual tagging, we have to encounter outliers that are duplicates and referred to by the same name. One such example is the Eiffel Tower replica in Las Vegas (a replica of the original in Paris), which also attracts many visitors. Another example are photos that show miniatures of important sights. They are validly tagged by a user with the name of that sight while being located far away from the original place of interest. An example is the photo of

a miniature Eiffel Tower on someone's desk. A difficult case are photos of a certain place and a user draws comparisons to other sights by also adding the compared place of interest as a tag. An example could be a photo of the Shibuya crossing in Tokyo where the contributor wants to point out that it looks similar to the Times Square in New York and provides according tags.

Hence, a complement to our approach would be to utilise image recognition and classification techniques to automate the manual inspection process that is currently incorporated in our approach.

Ciregan et al. (2012) introduced multi-column deep neural networks for image recognition and classification after training them over several datasets including, images, hand writing, and traffic signs. They show that after iteratively training the random weights of the deep neural networks the outcome classification error is reduced. Furthermore, Vedaldi and Fulkerson (2010) introduced the *VLFeat* library that combines several computer vision algorithms to implement feature detectors, feature extractors, clustering, randomised kd-tree matching, and super-pixelisation. For example one such algorithm is the SIFT feature detector and descriptor. This library helps users to try out various such algorithms for image recognition and classification tasks. These techniques can programmatically identify the image content and compare it to the point of interest to find (dis)similarities, and then associate it with the reverse-viewshed to determine the uncertainty. This would already filter out images that are irrelevant to our query (e.g., those that are textually/geographically tagged as the Brandenburg Gate but represent a bus stop in the nearby region), and show us images that represent the target within the reverse viewshed.

Text analysis algorithms can also aid us in filtering out relevant and irrelevantly labelled photographs. For instance, in a geographic information retrieval context Wang and Stewart (2015) improves the extraction of semantics from web texts by combining Natural Language Processing (NLP) techniques with ontologies, using GATE[8] (General Architecture for Text Engineering) as their primary tool for extracting the spatial dimension and the semantics from text data. Also, Šarić et al. (2012) assess short text semantic similarity by using a support vector regression model that considers word-overlap similarity and syntax similarity as features. An incorporation of these semantic technologies for analysing image labels for their coherence with the image content may pave for future extensions of our approach.

Such works together with the reverse-viewshed analysis would strengthen the uncertainty analysis of images-based VGI.

[8] `https://gate.ac.uk/`

Thus far, we have considered only one aspect with which the reliability of a photograph can be assessed: the location correctness. In addition to this there are further aspects, as described above, that attribute to the reliability of an image, such as the label completeness, content relevance, user profile completeness etc. A weighted score for each of these aspects could give us a complete reliability score for each user, with which the user credibility can be evaluated.

Regarding data accuracy, when computing the reverse-viewshed analysis, one has to encounter issues of output quality variability that were emphasised by Fisher (1991). For example, in calculating the LoS for all observer points we considered the height of observers as 2m, but this varies from individual to individual. Such quality issues are due to data errors, data resolution, as well as errors in the viewshed analysis algorithm. Thus, in our work we limit our approach to calculating a reverse-viewshed upon which the location correctness of geotagged Flickr images are assessed. We suggest to use other additional user/photo metadata in combination with the location correctness to infer the credibility of users.

Future extensions of this work should focus towards a mechanism for *automatically* inferring the user credibility through analysing the dependency between user metadata and the location correctness determined with the reverse-viewshed. Thereby, the influence of viewshed sensibility should be studied and optimized, e.g., by investigating vectorised city models based on CityGML. Further, credibility-related measures can be extracted from analysing free-text comments that users provide for photos. An example is sentiment analysis, which computes polarity scores regarding the expressed opinions. Another direction will be to look into the temporal trends of photo capturing and uploading behaviours. Looking into these additional aspects and giving them a weighted score to find the complete reliability of geotagged images will allow one to evaluate the user's credibility within these visually generated VGI sources. Furthermore, to automate the process of user credibility assessment we can envisage trained statistical prediction algorithms for classifying the users according to the above mentioned weighted reliability parameters. These observations are a starting point to heuristically assess expected image credibility relating to location and description correctness. In the future, this approach can be refined to a full prediction model. Considering content-based analysis functions and multivariate regression analysis could provide advanced quality predictions. The proposed approach can be extended to larger data sets by considering additional data sets from the VGI domain- such as panaramio images. These results will eventually enable new applications and improve drawing usage from mass VGI data.

3.5 Conclusions

With the exponential growth of geo-referenced image-based data on various VGI platforms, the need for uncertainty analysis approaches of such data has become a pressing issue for maintaining the quality. This chapter contributes to the research and discussion on quality control of image-based VGI. We have investigated through experimental analysis how a reverse viewshed analysis can be utilised to assess the location correctness of image-based VGI. In doing so, we have first programmatically downloaded metadata of photographs for a certain point of interest by querying the open Flickr API for all geotagged photos, which are textually tagged (labelled) with the place description (e.g, with the tags "Brandenburg Gate" and "Berlin"). As a next step, we have computed the area of visibility from each observer point (geotag) based on surface elevation data, to the given points of interest, the Brandenburg Gate and the Reichstag in Berlin. With the help of this reverse-viewshed analysis we were able to determine if the position of the POI lies within the visibility from a given observer point. If it lies outside of the visibility region, the photograph captured by the observer is considered as incorrectly geotagged. We duly note that all images that do correspond to the point of interest through the geo/text tag do not necessarily visually represent the point of interest. This is also exhibited through analysing a sample dataset. We suggest in the future work to conduct image recognition techniques to filter out images that are irrelevant to the point of interest.

Within the sample dataset for Brandenburg Gate and Reichstag we have categorised the photographs into four groups based on the geotag and label correctness. On those categories we made observations in user and photo metadata to derive credibility indicators. In particular, we have found that users producing photos for category a and c (both wrongly labelled) have on average higher numbers of photos (for both use cases). Also, we found that photos in category a and c (both incorrectly labelled) have higher numbers of tags. Further, the producers of photos in category b and d (correctly labelled) together have on average lower number of contacts as compared to the other photo categories. As we insinuate that these are valuable indications for assessing the credibility of users based on the reliability of their contributions, these further imply on investigating the tagging behaviour of users beyond their motivational aspects.

Chapter 4

Uncertainty-aware Movement Analysis in Text-based Volunteered Geographic Information

Contents

4.1 Background and Related Work 89
4.2 Movement Detection in Implicitly Referenced Spatial Data 91
4.2.1 Keyword-based and #hashtag-based Data Gathering 92
4.2.2 Hostspot & Cluster Analysis with KDE & DBSCAN 93
4.2.3 Conversation Movement Trajectories 97
4.3 Sructural Characterisation of Movement Trajectories . . 99
4.3.1 Geospatial Structure-based Characterisation 99
4.3.2 Content Structure-based Characterisation 102
4.4 Feature-based Trajectory Ranking 109
4.4.1 Example Scenario 1 110
4.4.2 Example Scenario 2 111
4.5 Conversation Movement Analysis for Sports Journalism 116
4.6 Discussion & Future Work 118
4.7 Conclusions 123

Movement of phenomena can occur through many modalities such as space, time, content, or a combination thereof. Detecting such movement patterns, especially from implicitly referenced spatial data is a challenging yet important task. Exploring meaningful movement trajectories based on implicitly referenced spatial data such as Twitter data can be efficiently achieved by data analysis and visualisation. Current state of the art tools mostly incorporate time-series and clustering approaches, and

keyword-based queries to filter out the relevant events and movement trajectories of interest.

This chapter introduces a novel visual analytics approach for movement detection in text-based microblog data, that is termed here as *MovingOnTwitter*. The approach is among the first to introduce the usage of geospatial movement of Twitter data for analysis and event detection, as opposed to only using textual context data as seen in most existing methods. The developed approach is two-tiered: (1) Movement is detected, on the one hand through a keyword-based approach that relies on the *episodic sequence of spatio-temporal hotspots*, and on the other hand through a grouping strategy based on the *geospatial and content structure*. I.e., characteristics such as the uncertainty in the content of Twitter microblogs filter out the interesting and meaningful trajectories. (2) The observed movement trajectories are ranked through a user-defined interestingness measure.

Both of these tiers are presented in a visual interface and allow the user to explore Twitter text streams without having to have extensive prior knowledge. The user benefits from the pure exploratory capabilities of the tool that implements the developed approach. This chapter further demonstrates how the user interaction in trajectory characterisation and ranking help to reduce the uncertainty of the resulting trajectories. The usefulness of the approach is validated within appropriate use cases.

The remainder of this chapter unfolds as follows: in Section 4.1 the related work on movement detection through microblog data as well as the works of using characteristics of such data to analyse movement are reviewed. Section 4.2 presents a systematic framework that utilises a keyword-based approach and a #hashtag-based approach to detect movement patterns in implicitly referenced spatial data. To meaningfully analyse these movement patterns and to uniquely distinguish movement patterns based on their inherent characteristics, Section 4.3 characterises these movement patterns based on their geospatial and content structure. A feature-based ranking approach for the identified movement trajectories is introduced in Section 4.4. The developed approaches are validated through use case findings in Section 4.5.

The contents of this chapter are based on the publications Senaratne et al. (2014a)[1] and Senaratne et al. (2016, under review)[2].

[1]Appears in Sections 4.2 and 4.3

[2]Appears in Sections 4.2, 4.3, 4.4, and 4.5. Both of these works are a result of a collaboration with A. Broering from ESRI GmbH, D. Lehle from the University of Konstanz and T. Schreck from the University of Konstanz. My role as the first author was defining together with D. Lehle the systematic framework for detecting movement trajectories through episodic sequential hotspots, approaches for identifying the characteristics of detected trajectories through their various geospatial and content structures, and the formalisation of the feature-based trajectory ranking approach.

4.1 Background and Related Work

Since the Web 2.0 has emerged, humans can be considered as virtual sensors who are able to collect and contribute spatially referenced data in the form of images (as seen in Chapter 3), maps, text, audio, or video on the Web, making the consumers of data also the producers. O'Reilly (2005) described such user generated content as the *wisdom of the crowds* thereby emphasising the potential of the data. Event detection has long been practiced using news articles (Allan et al., 1998), or raw sensor data (Guralnik and Srivastava, 1999). The plethora of social media portals, such as Twitter, enables laypersons, domain experts as well as news broadcasters to turn to these social media portals to search and detect events in near real-time. Chunara et al. (2012) showcased how Twitter was used complimentary with HealthMap[3] and official data (from the Haitian Ministry of Public Health) to detect the outbreak of the cholera epidemic and estimate the disease dynamics which resulted as an aftermath of the Haitian earthquake in 2010.

Visual analytics has been helping analysts to visually explore and derive circumstantial evidences of events from text-based data sources. In an extensive survey, Wanner et al. (2014) review the state of the art visual analytics approaches for event detection in text data streams. As an outcome of their survey they formulate guidelines for building successful visual analytics approaches for various types of events in text-based data. Twitter microblogs, as many other VGI sources, have been utilised in various use cases such as disaster management (Cameron et al., 2012), situational awareness (MacEachren et al., 2011), or *movement detection* (Adrienko and Adrienko, 2011), thereby proving its immense potential. Movement detection in Twitter is not a trivial task, mostly due to the implicit spatial dimensions contained in the text (e.g., "I'm in New York enjoying Lady Gaga's concert" implies that Lady Gaga is performing at that specific time in New York) or attached to the text (e.g., geotaging content with the location of the event rather than the Tweeter's position) (this is described in detail in Chapter 2). These implicit spatial dimensions make it challenging to derive location based services without including additional content. Therefore deriving accurate trajectories from these implicit spatial dimensions is difficult.

Spatial movement and trajectory detection using VGI however is a pressing topic. Research in this area can lead to methods and technologies which are valuable for various applications ranging from marketing (e.g., how is the word about a new product spreading) to disaster management (e.g., what is the path of the hurricane). Thereby,

[3] `http://www.healthmap.org`

the key benefit of using VGI in such applications is its real-time character. Fruitful efforts can be seen through various works for movement analysis in text data. The work of Andrienko et al. (2013a) constructed trajectories of Twitter users from tweeting locations by computing the trajectory medoid (i.e., the cluster point of a dataset whose average dissimilarity to all objects in the cluster is minimal) for each spatially referenced tweet. In another work Adrienko and Adrienko (2011) introduced a method for spatial generalisation and aggregation of movement trajectories by extracting only the significant points in a trajecotry, that also retains the essential characteristics of the movement. Through parameterisation of the movement model they allow enough leeway to the user to control the extent of abstraction. They further introduce quality metrics for assessing the quality of the generalisation. von Landesberger et al. (2014) have been working on large complex time-dependent data, introducing time-dependent movement analysis features particularly for group movement, and methods to automatically analyse and filter interesting sub-parts of a dataset for in-depth inspections. Further, Sakaki et al. (2010) used a classifier that considered features such as keywords, number of words, and the context to approximate the trajectory of a moving Typhoon via Twitter. They utilised a particle filtering to assess the geographic locations of the typhoon path with a weighted average of latitudes and longitudes, and median as a baseline.

Another approach by Rinzivillo et al. (2008) used distance functions to determine the similarity between multiple trajectories, and further introduced a progressive clustering technique which was applied to analyse large sets of trajectories. Fuchs et al. (2013) demonstrated how Twitter can be used in combination with other social media data sources and mobile network metrics to determine events that occur through space and time. Andrienko et al. (2008a) classified position recordings to determine the location of moving phenomena. In Andrienko and Andrienko (2010) and Andrienko et al. (2009), methods were developed to use VGI for exploring the interests, behaviour, and mobility of people. They presented a conceptual framework that allowed the aggregation of movement data, with a focus on situation-oriented and trajectory-oriented movement data. Their research emphasised the importance of aggregating movement data for supporting visual exploration, and Andrienko et al. (2013b) discussed the need for appropriate visualisation methods to analyse such movement data.

Demonstrating the usefulness of geo-visual analytics, MacEachren et al. (2011) developed the *SensePlace2* tool to support situational awareness during crisis events using Twitter microblogs. Their map-based web application essentially incorporates

overview and detail on demand, and a visual interface that enables the user to understand place, time, and thematic components of emerging situations. This search-by-query application relies on keyword inputs by the user (requiring the user to possess prior knowledge of the situation they want to explore), to then visualise a list of Tweets pertaining to the keywords. Their text content analysis, although limited to the keyword frequency, provides an overview of how often the selected keyword occurs in the dataset in a given time frame. They further allow the user to specify the temporal range to analyse the dataset. But no further temporal analysis, such as temporal structural analysis is possible.

In another similar visual analytics tool called *ScatterBlogs2*, Bosch et al. (2013)focus mainly on the filtering accuracy where the users are enabled to interactively build task-tailored message filters. At the filter creation stage the user is able to visually create classifiers and filters to train and test them on archived event-based messages. Further, the user has to define keywords that describe the event-relevant messages in order to maximise the search query.

Both above tools are used in situations where the user possesses to some extent prior knowledge of what they want to explore. They have demonstrated the usefulness of their tools within appropriate use cases. In our approach in addition to the keyword-based relevance method, we rely on a grouping strategy to query and thereby filter the text-based Twitter datset based on the *geospatial and content structure* to derive trajectories. These structural analyses further help the user to detect changes in evolving conversations through the spatial, temporal, and contextual modalities.

The key advantage of the visual analytics approach presented here, which we aptly call *MovingOnTwitter*, is that it caters to wider analysis and explorative possibilities that do not necessarily require the user to possess prior knowledge of the events. The presented approach further aims at reducing the uncertainties that inherently come with such data, thereby allowing the user to take well informed analytic decisions. The work flow for the proposed MovingOnTwitter approach is shown in Figure 4.1. Each of the steps depicted in the work flow is described in detail in the following sections.

4.2 Movement Detection in Implicitly Referenced Spatial Data

A spatial trajectory is defined by Zheng and Zhou (2011) as a trace generated by a moving object in geographical space, usually represented by a series of chronologically ordered points $(p_1, p_2, ...p_n)$, where each point consists of a geospatial coordinate set

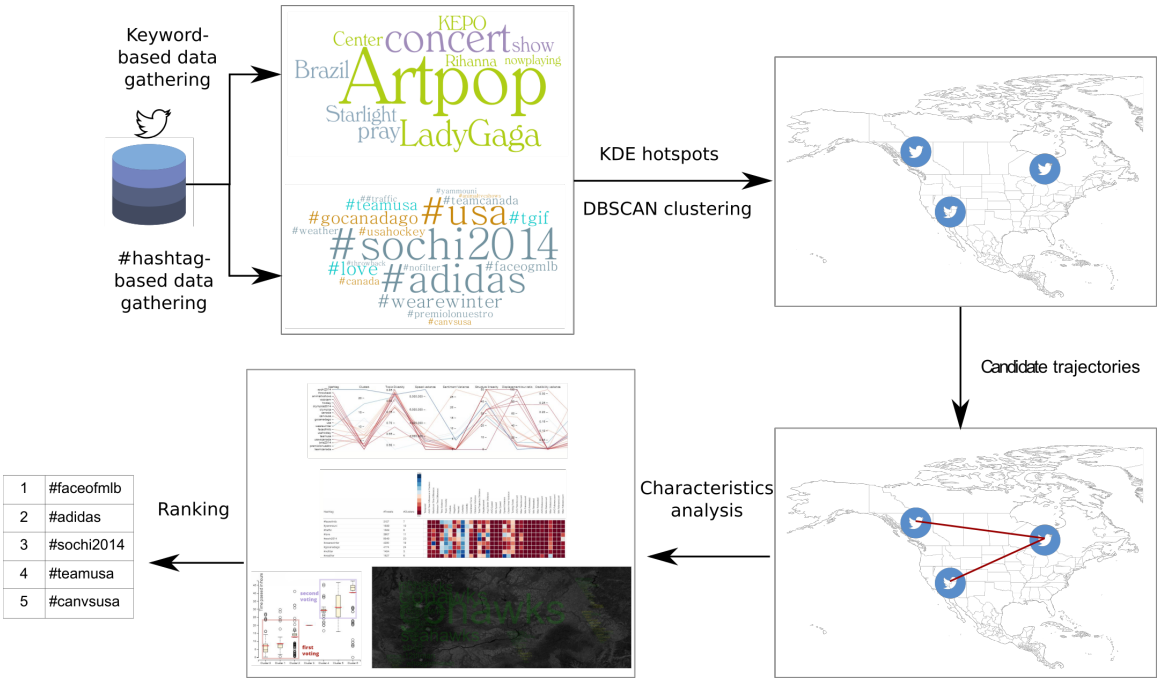


Figure 4.1: The work flow for MovingOnTwitter. This figure appears in Senaratne et al. (2016, under review).

and a time stamp such as $p = (x, y, t)$. In relation to the explicit and implicit nature of VGI (Chapter 1 Section 1.1), trajectories that are derived from VGI can also be categorised into what we call here as *directly observed trajectories* and *indirectly observed trajectories*. Directly observed trajectories have an explicit spatial reference, e.g., tracking of a sensor enabled twittering parcel package as described by Bröring (2013), while indirectly observed trajectories have an implicit spatial reference, e.g., extracting the movement of a flood by analysing the information given by contributors on Twitter and other VGI sources as described by Fuchs et al. (2013).

4.2.1 Keyword-based and #hashtag-based Data Gathering

As shown in Figure 4.1 in a first step, as in any data analysis approach, the dataset for movement analysis needs to be gathered. On microblogs, such as Twitter, one technique to detect indirectly observed trajectories from implicit spatial dimensions is by keyword filtering over a set of geotagged tweets. This simple technique can be utilised to derive context specific information based on the contributions on Twitter. In the next filtration step, a time frame is chosen for which the geotagged tweets shall be extracted. After this stage a Term Frequency-Inverse Document Frequency (tf-idf) analysis (Phelan et al., 2009) is run on the extracted Tweets with a stop word list

specifically for tweets to generate a pre-selection of the frequently used keywords in the dataset. These keywords are then ranked based on their frequency count. After sampling the data as necessary, a keyword is chosen to classify the data based on their spatial extent (as described in Section 4.2.2).

A keyword-based gathering of data however requires prior knowledge of events that the user wants to analyse. This approach will not suffice for tasks where the user does not possess prior knowledge. Instead they rely on *pure exploratory* features of analysis tools to explore patterns and trends. The #hashtag-based filtering technique relies on the trending topics on Twitter for a given period of time, to narrow down the search-space (as shown in stage one of Figure 4.1). Contributors are able to add the symbol '#' as a prefix to the topic they want to discuss on Twitter. If other Twitter contributors want to discuss the same topic they can join in and use the same #hashtag to continue the discussion of the said topic, e.g., #faceofmlb to discuss major-league baseball. When many contributors use this #hashtag in a given period of time, and therefore the #hashtag gains momentum it is considered to be *trending* on Twitter. Therefore the #hashtag-based filtering technique for example extracts the twenty most trending #hashtags and the following conversations from Twitter for exploratory analysis. This allows the user to explore and analyse Twitter data without necessarily possessing prior knowledge.

4.2.2 Hostspot & Cluster Analysis with KDE & DBSCAN

To derive more reliable spatial trajectories this simple keywords- and #hashtag-based filtering techniques are not sufficient, a spatial and a temporal dimension are required and more sophisticated methods need to be applied to effectively detect spatial trajectories. Therefore in a second step as shown in Figure 4.1, the gathered data are classified based on their density distribution using the KDE hotspot analysis approach, and the DBSCAN clustering approach.

Contrary to point data mapping, which focuses on mapping the location of individual events, *hotspot mapping* focuses on highlighting areas that have higher than average incidence of events. These hotspot areas can exist in different scales of interest. In order to estimate these hotspots of events corresponding to a chosen keyword, a smooth, continuous, and differentiable Gaussian Kernel Density Estimation (KDE) is utilised at each time step, which in principle creates a surface based on the distribution and density of Twitter message geotags. Then, a heat map visualisation is used to display the results of the KDE. The work of Chainey et al. (2008) shows that KDE has

a higher Prediction Accuracy Index (PAI) when it comes to performance in comparison to other methods.

To demonstrate the KDE approach with an example dataset, we implemented it as an HTML5 application based on Bootstrap[4] and additional JavaScript libraries, such as D3.js[5], jQuery[6], and the Google Maps API[7]. For storing and processing of large amounts of Tweets we used a MySql[8] database where we store the data with all its metadata.

As a proof of concept for the developed approach, we applied this implemented tool to an example dataset- concert route of pop music artist *Lady Gaga*. This artist has over 41.2 Million followers[9] on Twitter, and is known for controversial performances that constantly make headlines on Twitter as well as in other social media platforms. She planned a North American tour between 11.1.-16.3 of 2013. We chose a dataset that represented the tour on Twitter in order to determine her tour trajectory based on what Twitter contributors had to say.

In preparation of the dataset, we initially filtered the Twitter stream[10] for geotagged Tweets, based on the keywords 'lady gaga' and 'ladygaga', as well as for the selected time frame. This resulted in 26,000 tweets that contained any term referring to Lady Gaga. We consider this as sufficient for our initial analysis. We then generated a pre-selection of top-ten most frequently used keywords as a result of a *tf-idf* analysis on the extracted Tweets with a stop word list specifically for tweets. The ranked keywords resulting from this term frequency calculation were: 'Artpop', 'LadyGaga', 'concert', 'Starlight', 'nowplaying', 'Brazil', 'KEPO', 'Center', 'Rihanna', 'show'. For our analysis we chose the keyword 'concert', and therefore the dataset is filtered on a second round based on this selected keyword.

To sample our dataset we associate each Tweet with the geographic coordinates of the closest larger city which has a population greater than 100,000. This served as a pre-clustering and to remove noise which is necessary for the next step, the KDE.

The gathered dataset has a temporal resolution of one day, as there was a minimum time gap of one day between each performance. Further it has a spatial resolution at city level, as the artist performed in different cities in North America on each day of

[4] `http://getbootstrap.com/`
[5] `http://d3js.org/`
[6] `http://jquery.com/`
[7] `https://developers.google.com/maps/documentation/javascript/`
[8] `http://mysql.de/`
[9] `https://twitter.com/ladygaga`
[10] Note that we could only capture 10% of all geotagged Tweets, due to Twitter's policies regarding download limits.

her tour. We ran the KDE based on a Gaussian distribution for the geotagged Tweets point data at each time step (day). The resulting hotspot clusters were visualised as a heat map layer for every day on top of a geographic map. As you can see in the resulting map visualisation in Figure 4.2, higher activity of tweeters are clustered around particular locations. These dense locations are called hotspots.

Figure 4.2: Hotspots detected with Kernel Density Estimation for the Lady Gaga dataset are visualised with a heat map. This figure appeared in Senaratne et al. (2014a).

The resulting KDE for every day already indicates an episodic sequence of hotspot clusters, these hotspot clusters are an approximation for the cities where the concert took place during the tour as they show the highest densities. The implementation of the MovingOnTwitter approach explicitly allows for an iterative trajectory detection. I.e., if the user is not content with the results, the input parameters can be changed. This process further helps to reduce the uncertainty of the outcome trajectories (as described in the framework in Chapter 2 Section 2.2.2). Also, a user can refine the filtration step and adjust the specified time frame, or keywords as input to the KDE. This way, after multiple iterations, the detection of trajectories can be optimised.

As an alternative to KDE, data can be clustered using the density-based clustering algorithm *DBSCAN* (Density-based Spatial Clustering of Applications with Noise), where the user can specify the noise in terms of the radius (in km) and the minimum points in a cluster. We used this approach on the same Lady Gaga dataset and achieved clusters similar to the KDE hotspot results (Figure 4.3). The sequential direction from one cluster to the other is indicated by the arrow heads in Figure 4.3.

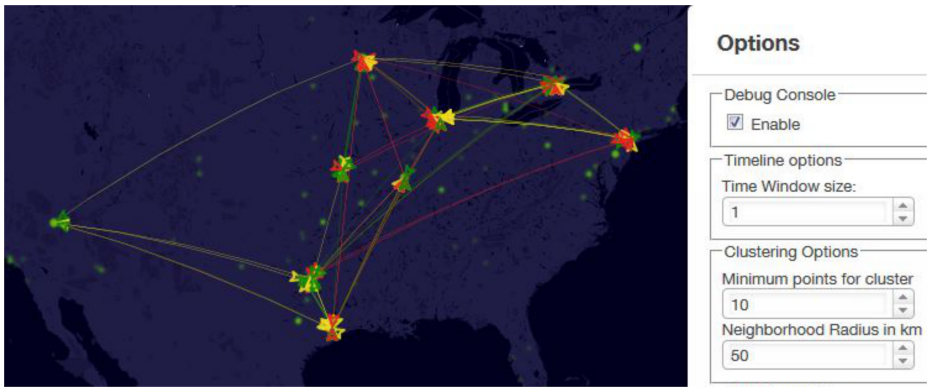


Figure 4.3: Clustered routes before averaging the time. The arrows indicate the sequential direction from one cluster to the other. This figure appeared in Senaratne et al. (2014a).

In a third step as shown in Figure 4.1, we derive the trajectories of the Lady Gaga concert tour data. To achieve this, for each of the cluster centroids we computed the average time, based on the assumption that people tweet about the concert around that particular geographic region on the day of the concert. Connecting the averaged time at each progressive cluster centroid gives us what represents a trajectory. This is shown in Figure 4.4.

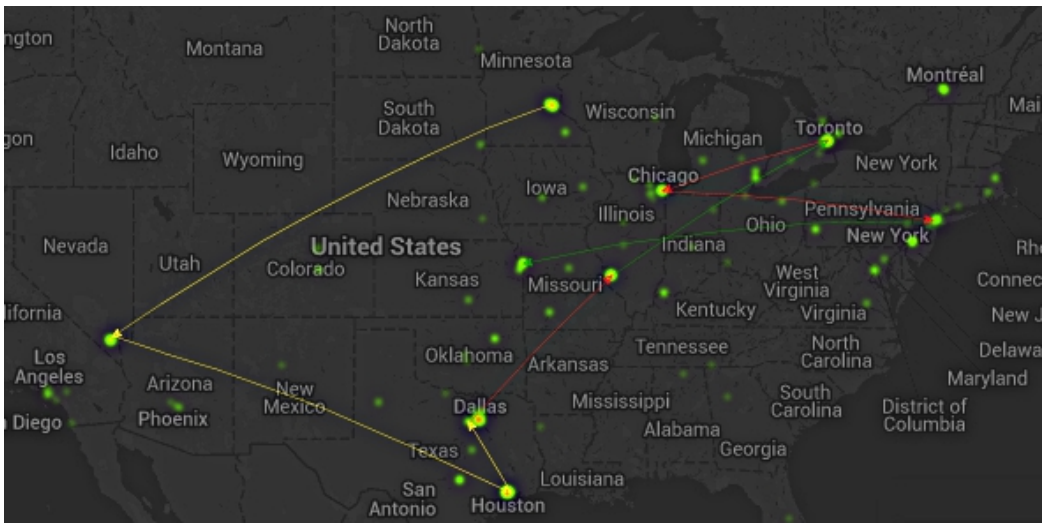


Figure 4.4: The KDE of the Tweets relating to the Lady Gaga concert tour, and the clustered routes after averaging the time. Cities (e.g., Las Vegas, Dallas, Houston, or Toronto) where the concert took place are already visible as hotspots. Also cities where the concert was later canceled, such as in New York and Florida are evident through the hotspots. The arrows in the clustered routes indicate the sequential direction from one hotspot to the other. The colours of the trajectories depict the average sentiments from the respective hotspots. This figure appeared in Senaratne et al. (2014a).

Figure 4.5 shows the actual route of the tour (in Black colour) overlayed on top

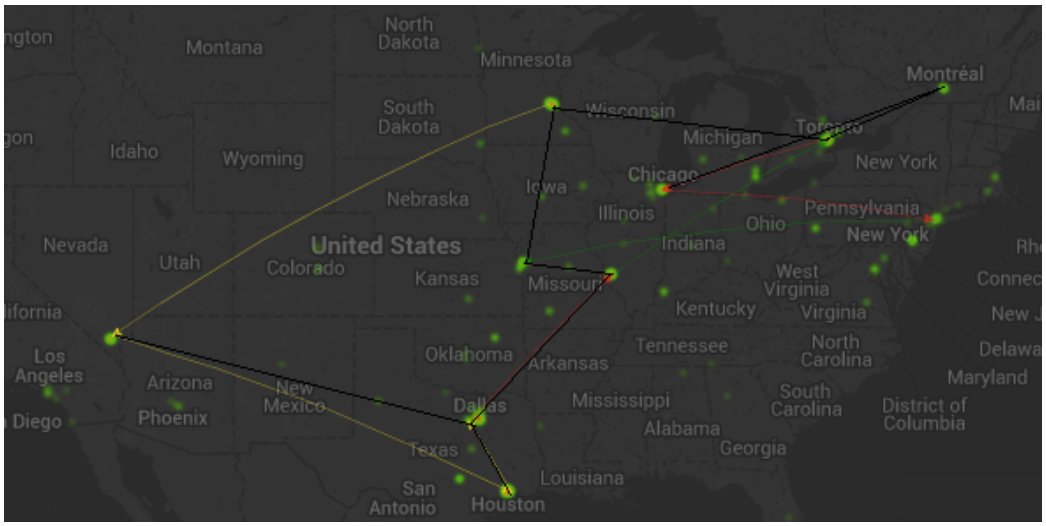


Figure 4.5: Actual route (in Black colour) over the approximated trajectory (in Green, Yellow, and Red colour that depicts the averaged positive, neutral and negative sentiments from the respective hotspots). The route indicates the following concerts: Las Vegas (NV) on 25.01., Dallas (TX) on 29.01., Houston (TX) on 31.01., St. Louis (MO) on 02.02., Kansas city (MO) on 04.02., St. Paul (MN) on 06.02., Toronto (ON) on 08.02., and Montreal (QC) on 11.02., before the concert got cancelled for the remaining leg of the tour starting from Chicago (IL) which was supposed to take place on the following 13.02. This figure appeared in Senaratne et al. (2014a).

of the approximated trajectory resulting from our MovingOnTwitter approach. The actual route as well as the determined trajectory start in Las Vegas and correspond in several other cities. A time slider helps to navigate through the produced heat maps at every day of the tour. The user can change the time window of the time slider if she wishes to change the temporal resolution. Time steps are weighted differently on the heat map to enhance the observations of the currently selected visualised time step (i.e., tweets from previous days have lower weighting).

4.2.3 Conversation Movement Trajectories

Analysing the episodic changes in *conversations* that move through various modalities such as space, time, and context can be useful to detect interesting and meaningful incidents. A conversation within the context of this work refers to an exchange of thoughts, news, or ideas about a particular topic between two or more people. As we learned in Section 4.2.1, on Twitter #hashtags allow its users to collate all conversations pertaining to a particular topic, allowing the user to collect conversations of news, thoughts, and ideas about a trending topic.

In this section we are further interested in the temporal and spatial progression

of Tweet messages belonging to a user-chosen #hashtag. We call this progression a *conversation trajectory*, and propose methods for the identification of interesting trajectories and their analysis.

To group similar hashtags based on their relative distribution in space and time, we performed a density-based DBSCAN clustering (Ester et al., 1996) on each hashtag. The key advantages of using DBSCAN are that (1) the user does not need to specify the number of clusters and (2) it can find non-linearly separable clusters. Due to the episodic nature of conversations we added a maximum temporal distance to the DBSCAN algorithm that helped us to cluster hashtags that are closer in time as well as in space. Further, we set a minimum number of Tweets as a parameter setting for DBSCAN. The average time in each of these hashtag clusters are connected sequentially to derive the episodic conversation trajectory.

As described at the beginning of Section 4.2 , a trajectory T can be denoted as follows:

$$T = p_1(x_1, y_1, t_1), p_2(x_2, y_2, t_3), ..., p_n(x_n, y_n, t_n) \tag{4.1}$$

In our approach, the trail of points p of a trajectory represents the centroids of the episodic clusters that are derived by DBSCAN. A cluster C can be denoted as follows:

$$C = TW_1, TW_2, ...TW_n, \forall tc \in TW, \exists w = \# \tag{4.2}$$

where TW is a Tweet contained in that given cluster C, and all of the Tweets have the same #hashtag mentioned in one of the words w within the Tweet content tc.

Thus, a trajectory for a given #hashtag can also be represented as:

$$T_{\#} = C_1, C_2, ..., C_3 \tag{4.3}$$

where each C represents a cluster centroid.

In the following sections, we demonstrate how we can structurally characterise these derived trajectories (Section 4.3), rank them based on a user-defined interestingness measure (Section 4.4), and demonstrate an implementation of the approach with an example dataset in Section 4.5.

4.3 Sructural Characterisation of Movement Trajectories

The fourth step of our MovingOnTwitter approach as shown in Figure 4.1 is the characteristics analysis of the trajectories. For a *meaningful* analysis of trajectories, we need to be able to distinguish them in terms of their inherent characteristics. We have identified several such characteristics based on the geospatial structure and the content structure of movement trajectories. To exemplify these characterisations within this section, we have taken conversation movement trajectories for chosen #hashtags. Characterisation of conversation trajectories is helpful to the analyst to filter out interesting and meaningful patterns, as well as to distinguish them from noise in the data, thereby reduce the uncertainty of the results. The usefulness of trajectory characterisation is further demonstrated within a use case scenario in section 4.5.

An excerpt of the implementation of the MovingOnTwitter approach for hashtag selection and characteristics analysis for chosen hashtag-based trajectories is shown in Figure 4.16

In the following sections we break down the structural characterisations into (1) geospatial structure-based characterisation in Section 4.3.1, and (2) content structure-based characterisation in Section 4.3.2.

4.3.1 Geospatial Structure-based Characterisation

Characteristics based on the geographic structure of a trajectory path are important to indicate how the episodic clusters are generated, how they differ between different trajectories, and to detect interesting changes in the trajectory path. Such characteristics include the overall distance of the trajectory that represents the overall coverage of the tweets per topic, change of direction of the trajectory, and the speed of propagation of topics in clusters that represent the episodic hotspots.

Distance Variance

Distance variance calculates the distance between two consecutive clusters of a trajectory. This is useful to determine the overall impact of a given topic on Twitter. To calculate the distance variance we calculate the great-circle distance between two centroids using the Haversine formula (Robusto, 1957). The Harvesine formula is as

follows:

$$
\begin{aligned}
a &= \sin^2(\Delta\varphi/2) + \cos\varphi_2 * \sin^2(\Delta\lambda/2) \\
c &= 2 * \arctan 2(\sqrt{a}, \sqrt{(1-a)}) \\
d &= R * c
\end{aligned}
\tag{4.4}
$$

Where φ and λ represent the geographical coordinates latitude and longitude respectively, and R is the radius of the earth.

These distance variance values are visualised in MovingOnTwitter through a heat map visualisation as shown in the Figure 4.17. In Figures 4.6 and 4.7 we demonstrate the distance variance for the hashtags "#melfest" and "#chinesenewyear" through the trajectory visualisation. #melfest refers to the Swedish song contest "Melodifestivalen"[11] which is the pre-selection for the Eurovision song contest. #melfest has a much lower distance variance than #chinesenewyear, as it has a lower global spread (audience is mainly coming from Sweden and Europe). Whereas the Chinese new year is celebrated around the world, and it has a larger spread with clusters of tweets coming from many corners of the globe.

Figure 4.6: Trajectory visualisation for #melfest. The circles represent the clusters, and their colours represent the dominating topics in each cluster. This figure appears in Senaratne et al. (2016, under review).

Trajectory Linearity

Trajectory linearity indicates the directional characteristics of trajectories. We calculate the ratio of turning points of each trajectory segment to calculate the linearity. A

[11] `http://www.eurovision.tv/tag/expand/Melodifestivalen`

Figure 4.7: Trajectory visualisation for #chinesenewyear. The circles represent the clusters, and their colours represent the dominating topics in each cluster. This figure appears in Senaratne et al. (2016, under review).

segment of a trajectory is assumed to have a turning point when the bearing angle between the subsequent segments is higher than a predefined threshold of 90 degrees. Therefore, the ratio indicates how many times more than 25% of its initial angle a trajectory is heading in a different direction. In the examples shown in Figures 4.6 and 4.7 the hashtag #chinesenewyear has significantly less turning points, resulting in a more linear trajectory, as compared to #melfest song contest trajectory which has a more dynamic nature. The reason for this is that the song contest is a live event very popular in Europe that takes place in the course of 4 hours, in contrast to the Chinese New Year that is celebrated all over the world at different time zones at the dawn of the new year. This characteristic feature, which is also known as *turning point* is calculated for each hashtag trajectory and is visualised through a heat map visualisation in MovingOnTwitter as shown in Figure 4.17.

Speed Variance

The speed variance determines how fast a particular topic on Twitter propagates between locations. While some topics have a high peak time soon followed by a drop, others propagate over a steady speed at a longer time interval. The analyst can use this characteristic to detect the virality of a topic on Twitter and further analyse the content therein. The variance is calculated first by averaging the Tweet creation date of all tweets in a each cluster, and then by dividing the distance of each subsequent trajectory by the difference in time for each cluster. This characteristic is visualised in MovingOnTwitter through a heat map visualisation as shown in Figure 4.17.

4.3.2 Content Structure-based Characterisation

Analysing the content of Tweets is paramount for context-aware information foraging. To alleviate misconstrues, noisiness, and fuzzy language we can use sentiment and topic analyses techniques. While term-usage analysis is used to find general patterns and topic terms, keyword based analysis of tweets help to find what people are talking about, where, when, and how often in the clusters. Using sentiments and topic analyses within large amounts of trajectories help us to to determine which current episodic conversations are worth exploring. In the following sections we develop methods to explore characteristics such as *topic diversity*, *sentiment linearity*, and *credibility/certainty variance* that describe the changes of content.

Topic Diversity

The first step of analysing the diversity of topics in the tweets in the episodic clusters is to determine the thematic categories that the tweets fall into. Many techniques have been used in the state of the art for topic classification, such as named entity extraction (NER), or latent derichlet allocation (LDA), which require to know the number of topics in advance. Due to the non-regulated nature of Twitter messages (use of abbreviations or slang), we need more multi-modal language features and specific classifier training to achieve effective topic classifications. The work of Fiaidhi et al. (2013) uses the Java library LingPipe[12] which relies on computational linguistics for topic classification. Their work further shows a significant improvement in the accuracy of topic classification. Based on their work, we use a hierarchical feature subset selection algorithm to classify the tweets. The training of this language model is done by categorising character sequences. For each classified topic, conditional and joint probabilities are calculated and a score is given. We take the topic with the highest score to classify the tweet. A character based $n-gram$ is used to classify the tweets, where n is set to the average length of a word in a tweet. Based on the work of Bochkarev et al. (2015), we use an $n-gram$ the size of 5. During the labelling process we filtered out URLs, unicode characters, usernames, punctuation etc., and stop words. Accordingly, the tweets are classified into 12 topics: *computers & technology, education*, *family*, *food*, *health*, *marketing*, *music*, *news & media*, *pets*, *politics*, *recreation & sports*, *other*. We use their training data of pre-labeled tweets[13]

[12] `http://alias-i.com/lingpipe`

[13] `http://flash.lakeheadu.ca/~maislam/Data/`

to train our classifier. The topics are mapped to a colour scheme using ColorBrewer[14] as seen in Figure 4.8 and visualised accordingly in the clusters.

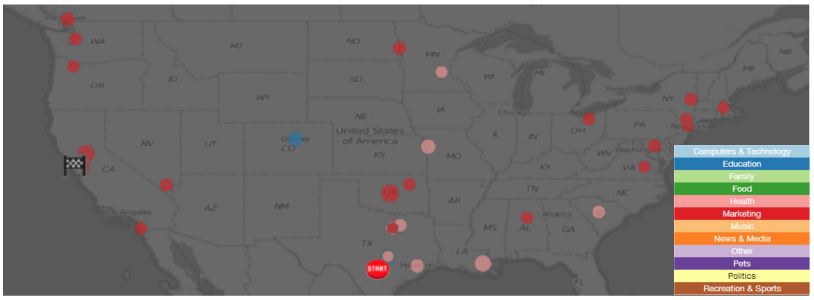


Figure 4.8: The episodic clusters of #skilledtrade. The circle radius indicates the number of tweets in the cluster, and the colour hue indicates the most frequent topic observed in the cluster. These colour hues are used only to create the primary visual differences between the classes of topics, and they do not indicate any similarity between the topics. Colours are allocated to the topics using Colorbrewer. This figure appears in Senaratne et al. (2016, under review).

To determine the diversity of topics in the dataset, we calculated the Simpson-Index (Simpson, 1949) which assesses the probability of two tweets from random clusters having the same topic. It is expressed as:

$$\lambda = 1 - \sum_{i=1}^{s} p_i^2 \tag{4.5}$$

p_i represents the relative amount of the topic i to the sum of all individual topics. This indicates the topic diversity along a given trajectory. To get an overview of these probability values of topic diversity for each hashtag trajectory, they are visualised in MovingOnTwitter using a heat map visualisation as shown in Figure 4.17. Further, parallel coordinates are used to observe these values for trajectories that are filtered out for a lower cluster distance as shown in Figure 4.18. Figure 4.8 shows a low topic diversity for the #skilledtrade trajectory based on the topic classification. The circles represent the clusters, and the circle radius represents the size of the cluster (tweet density). The colour of the circles represent the topic category accordingly. Therefore, the topics covered in the clusters are marketing (red), health (pink), and education (blue). Evidently, #skilledtrade is used for job offers in skilled trades such as welders, electricians, machinists etc.

[14] `http://colorbrewer2.org`

Topic Drift

A drift is a gradual change of phenomena that can be observed across many modalities. The topic drift analysis helps the user to identify the most relevant keywords of the tweets without having to read all the tweets. This enables a quick visual analysis of huge amounts of tweets. This part of our approach is related to the Nokia Internet Pulse (Kaye et al., 2012) and Wordle (Viegas et al., 2009) techniques to represent the most relevant keywords in a time frame. However, contrary to Kaye et al. (2012) where a vertical axis was used, we use a word cloud to represent the keywords in a time frame chosen by the user. This leaves the angles of the keywords in the word cloud and their colour to represent additional features.

The word cloud created as part of MovingOnTwitter shows the most occurring keywords during a chosen time frame. We used the D3.js library[15] and combined it with the algorithms in Viegas et al. (2009) to visualise the tag cloud. The positions of the words in the word cloud is based on Wordle (Viegas et al., 2009) , which positions the words at a random starting point, and if overlaps occur then the word is moved a step along an increasing spiral. This is repeated until no overlaps occur. The word cloud is drawn at the end when all the words have been positioned (Figure 4.9). The shown keywords of the word cloud are generated from all tweets that appear in that time frame, by filtering out a list of stop words. We map the average sentiment on a colour scale ranging from red, over yellow to green to indicate the varying emotions of Tweeters regarding the specific keywords. Further, we use the font size to map the number of occurrences of a keyword by calculating with a logarithmic function using the D3.js Scales functionality[16]. Consequently, the most occurring words appear larger than the least occurring words.

In addition, we map the optional parameters such as credibility indicators (described in Section 4.3.2) to the angle of the keywords in the word cloud. The rotation angles are between 0°, implying higher credibility, and 90°, implying lower credibility. This design choice was taken to communicate higher credible keywords with ease (when it is more horizontal) and lower credible words with lesser ease (when the keywords are dangling at an angle). The angle of these keywords with credibility is calculated based on a logarithmic scale.

[15] `http://www.jasondavies.com/wordcloud/`
[16] `https://www.dashingd3js.com/d3js-scales`

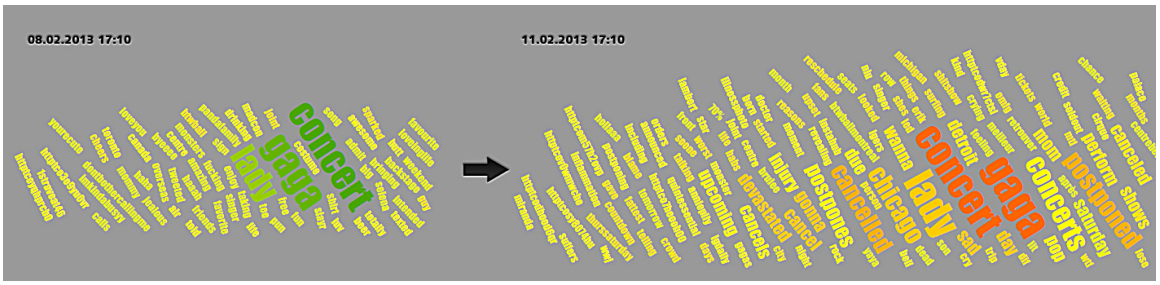


Figure 4.9: The sentiment and topic change from 08.02.2013 to 11.02.2013 in the Lady Gaga concert tour dataset. The positive to mostly negative change of sentiment together with the topic change indicate that the concert was canceled just before it was supposed to air in Chicago, USA. The actual tour information confirms this. The weighted credibility features *contains URL*, and *is Retweet* are used in combination to compute the average credibility of the tweets that pertain to the topics, and this is indicated through a 0° - 90° angle of each topic. More horizontal words indicate more credible topics (therefore easier to read) and more angular words towards 90° indicate non- credible topics (therefore more difficult to read). This figure appeared in Senaratne et al. (2014a).

Sentiment Drift

The sentiment drift analysis of MovingOnTwitter helps the user to determine the subjective opinion and the emotions of the contributors and how it changes across geographic space over time. The sentiment drift is visually analysed in our approach by linking a word cloud with the map visualisations. Therefore, we compute the average sentiment for each hotspot cluster. To achieve this we first classify each tweet with the help of Sander's annotated Twitter data set which has been evaluated by Saif et al. (2013) for significant results. By using this dataset, we trained a classifier using the LingPipe Java toolkit[17] which uses computational linguistics for processing the text. The polarity of these tweets were annotated as either positive, negative, or neutral sentiments. To obtain the collective sentiment of each hotspot cluster we averaged the sentiments of each tweet belonging to the hotspot clusters. We visualise these sentiments using different colours (red for negative sentiments, yellow for neutral sentiments, green for positive sentiments) as seen in Figure 4.9 for an exemplary sentiment drift in the Lady Gaga concert tour dataset.

The trajectory segments are coloured based on the avearge sentiment of the next occurring hotspot cluster (for a trajectory segment going from time t_1 at hotspot cluster C_1 to t_2 at hotspot cluster C_2, the colour of the segment would represent the sentiment at t_2 and hotspot cluster C_2). This is depicted in Figure 4.3 and Figure 4.4.

[17] `http://alias-i.com/lingpipe/`

Sentiment Linearity

Sentiment analysis allows the analyst to observe the majority attitude and opinion of people regarding a particular topic, brand, product etc., thereby enriching the content analysis process. In the previous Sections 4.3.2 and 4.3.2 we demonstrated how content analysis together with sentiment analysis helped to discover the cancellation of a concert tour in a particular city. In this section, we look into the sentiment linearity which indicates the *contradictory changes* of sentiments in the course of a trajectory. This is especially useful to detect controversial events, where people discussing these events have opposing opinions, surprise, or disbelief (Popescu and Pennacchiotti, 2010). To calculate the sentiment linearity (S_l), we first calculate the number of positive, negative, and neutral tweets in each episodic cluster. Next, we calculate a sentiment score for the subsequent cluster by using the following measure of contradiction by Tsytsarau et al. (2010):

$$S_l = \frac{\theta * \sigma^2}{\theta + (\mu)^2}$$

σ^2 is the variance, and μ is the mean of the sentiments in a given cluster, and θ allows us to add a small value that limits the level of contradictions when the aggregated sentiments is close to zero. Therefore, we set the value for θ at 0.05 (similar to Tsytsarau et al. (2010)). This sentiment variance for each hashtag trajectory is indicated through a parallel coordinates visualisation as shown in Figure 4.18. To indicate how often significant changes occur between subsequent episodic clusters, we calculate the *sentiment turns*, in addition to the work of Tsytsarau et al. (2010). A sentiment turn occurs whenever the change of the sentiment score S_l of the cluster $C_n + 1$ differs from the cluster C_n by more than a user-defined threshold δ (by default this is set to $\delta = 0.5$). In Figure 4.10 we use a horizon chart to show the sentiment change for #6nations, which was trending for the annual Northern hemisphere rugby union championship[18] during the 19.02.2014 - 20.02.2014 time frame. For a hashtag to be trending, is to be among the most popular topics discussed on Twitter at a given time. One indication for this is the number of Tweets that are mentioning a particular hashtag. In Figure 4.10 colour blue on the far left shows a slightly increasing positive sentiment with the beginning of the game (e.g., tweet: "First weekend I have not worked in 2014, just in time for the start of the #6nations"), and the gradual red colour shows negative sentiments from England fans towards the first point for France (e.g., tweet: "31 seconds and France score #WTF #6nations #englandrugby

[18] `https://en.wikipedia.org/wiki/2015_Six_Nations_Championship`

#fail"). To scalably visualise these horizon charts we use the Cubism.js library[19]. The small multiples aligned by time enables the analyst with rapid comparisons to increase discovery. The values for the sentiment turns are further indicated in MovingOnTwitter with a heat map visualisation as shown in Figure 4.17.

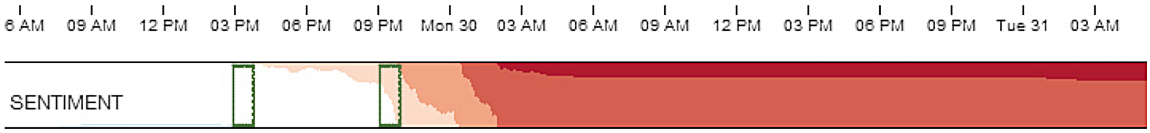


Figure 4.10: Sentiment horizon chart for #6nations rugby tournament. A blue (far left) to red (far right) diverging colour scheme indicates the progression of positive to negative sentiments. Sentiment change along the 48 hour time frame can be clearly detected at two specific instances as highlighted in the green boxes. First instance is right after the game has started, second instance is when France scored its first point. This figure appears in Senaratne et al. (2016, under review).

Certainty Variance

Words of Estimative Probability (WEP), as coined by Kent (1964), indicate the certainty in people-to-people dialogues, and was used for military intelligence analysis reports to deduct the probability of events occurring. We use this probability estimation of words to *signal* the certainty of conversations on MovingOnTwitter. While not an always robust measure, we use it as an estimation to nudge the analyst in the right direction. As such, we adapt their WEP to indicate an overall certainty score for the trajectories, and to indicate whenever the uncertainty changes. By following the work of Campbell et al. (2011), we classify signal words into five categories as seen in Figure 4.11. Subsequently, we assign several signal words under each category with a certainty score, as seen in Table 4.1. Tweets that contain any given signal word will be assigned its corresponding certainty score. If a given tweet does not contain any of the signal words, we handle it as 'certain' and assign a 1.0 certainty score. The resulting average of the certainty scores characterises the clusters, and the variance values indicate the changes of certainty. Figure 4.12 shows how the certainty is visualised for selected hashtags, and how the hashtags are sorted according to the certainty value.

Credibility

[19] `http://square.github.io/cubism/`

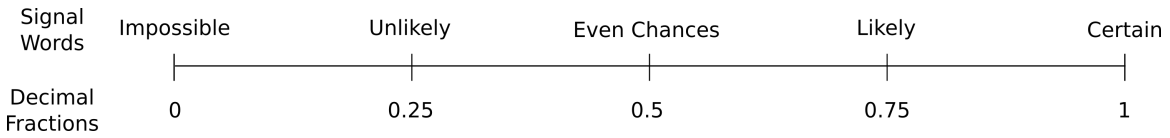


Figure 4.11: The five classes of uncertainty signal words adapted from Kent (1964).

Table 4.1: Uncertainty signal words and their corresponding scores.

Signal Words	**Score**
impossible, unthinkable, unreasonable, cannot, infeasible, unreliable	0.00
unlikely, odd, uneven, diverse, unsure, implausible, improbable	0.25
even chance, believe, estimate, guess, maybe, suppose, think, perhaps, eventual, assume, presume	0.50
likely, possibly, high chance, expected, expect, anticipated, potential, potentially, supposably, belike, presumably, reasonable, probable, plausible	0.75
certain, certainly, sure, safe to say, of course, confident, definitly, certainly, most likely, most probably, assured, reliable	1.00

In assessing the credibility of Twitter data, the source of information plays a crucial role, as it is what credibility is primarily based upon. However, this is not straight forward. Due to the non-authoritative nature of Twitter data, the source maybe unavailable, concealed, or missing (this is avoided by gatekeepers in authoritative data). Hovland et al. (1953) defined credibility as the *believability of a source or message, which comprises primarily two dimensions, the trustworthiness, and expertise.* Expertise contains objective characteristics such as accuracy, authority, competence, or source credentials (Flanagin and Metzger, 2008). Therefore, in assessing the credibility of data one needs to consider factors that attribute to the trustworthiness and expertise. Metadata about the origin of Twitter data can provide a foundation for the source credentials of Twitter data (Frew, 2007). In our work we utilise message- and user-based credibility features derived by Castillo et al. (2011) in a supervised classification. Based on the weighting given for these credibility features in Castillo et al. (2011), and the credibility impact for features found in the study of Morris et al. (2012), we derive the credibility impact on a scale between 0 - 7 for 9 message-based features, and 5 user-based features (Table 4.2). Credibility of information is also context specific, therefore based on the context at hand, the analyst can change these weighting criteria in our approach to fit their purpose. Using these credibility features we deduce the believability of conversation hotspot clusters in Section 4.5.

Figure 4.9 further shows how such weighted credibility features are applied to the Lady Gaga concert dataset. Figure 4.17 further shows how these credibility values for each hashtag trajectory are visualised within the heat map visualisation.

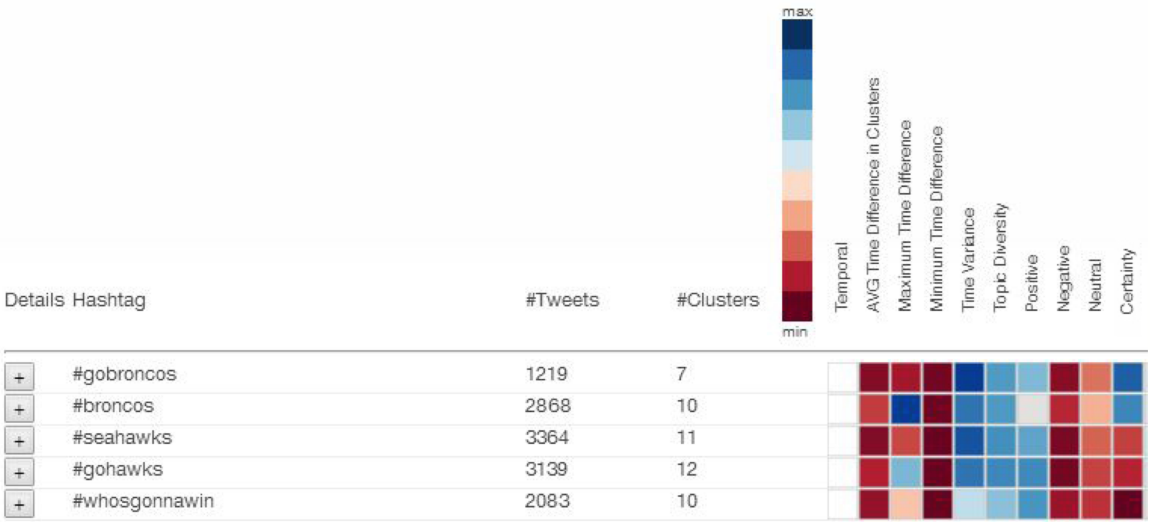


Figure 4.12: The certainty value used for sorting the Superbowl related hashtags. #whosgonnawin has the lowest certainty value due to many low certainty signal words. This figure appears in Senaratne et al. (2016, under review).

Table 4.2: Credibility features used in MovingOnTwitter2.

Type	**Features**	**Credibility Impact**
Message	contains question mark (?)	3.5
	contains exclamation mark (!)	3.5
	contains emoticon smile (:-), ;-), ...)	2.71
	contains emoticon frown (:-(, ;-(, ...)	2.71
	contains URL	4.9
	contains user (@cnnbrk)	3.5
	contains hashtag (#melfast)	3.5
	contains stock symbol ($APPL)	3.5
	is retweet (contains "RT")	5.12
User	registration age (days passed since registration)	5.46
	status count (no. of tweets user has posted)	5.18
	count followers (no. of people following this author)	5.13
	count friends (no. of people author is following)	5.13
	has description (a non-empty "bio" 1)	5.0

4.4 Feature-based Trajectory Ranking

As part of MovingOnTwitter, to further allow the analyst to have control over deriving trajectories that are *meaningful* to the task at hand, and reduce the search space size, we developed a ranking technique based on an *interestingness measure* according to the task at hand. Ranking of trajectories means to sort the resulting trajectories

based on an appropriate interestingness measure relevant to the use case at hand. Our ranking algorithm takes into consideration an interestingness measure based on the characteristic features derived in Sections 4.3.1 and 4.3.2 to sort the resulting hashtag-based trajectories. The chosen characteristic features are as follows: The topic diversity as it aids to observe trajectories, the sentiment variance that gives us insights to the discrepancies in discussions, high structural linearity to indicate movement, and speed variance to determine the virality of the topic. These features are considered as individual dimensions in the following interestingness measure calculations.

In an initial step we define the following method to calculate the distance for trajectories.

$$D_n = (1 + \frac{((D_n - min(D_n)) * (100 - 1))}{(max(D_n)) - min(D_n))})$$

Each dimension D_n is re-scaled to a value between 0 and 100 to make sure that there are no dominant dimensions. Then we define an interestingness value I_n for each of the dimensions between 0-100. For each trajectory the distance is then calculated by the Euclidean distance of each dimension to its interestingness value. Then the average difference is calculated to get the overall proximity according to the user defined interestingness. The resulting formula, where D is defined as the number of dimensions is shown below.

$$d = \sqrt{\frac{\sum_{n=1}^{D}(D_n - I_n)^2}{D}}$$

Although we calculate the distance considering the topic diversity, sentiment variance, structure linearity, and speed variance, it can consider other descriptive trajectory characteristics as well.

In the following, we demonstrate how our filtering and ranking approach works under different exemplary clustering parameters and different interestingness measures.

4.4.1 Example Scenario 1

For this example, we use a geotagged Twitter dataset collected during a 48 hour time frame from 01-02.02.2014 that was a weekend. To retrieve trajectories with clusters that contain a larger group of people discussing topics in a non-dense area, we set the following clustering parameters: geographic radius as the distance threshold (eps) = 50 km, minimum number of Tweets (MinTws) = 50, maximum temporal distance (tf)

Figure 4.13: Visualisation of the top 6 trajectories ranked left to right by high topic diversity and high structure linearity, low sentiment variance, and low speed variance. These trajectories further indicate a cyclic structure.

= 60 min. Furthermore, in this example we want to retrieve conversation trajectories that tend to have highly diverging topics (0 = lowest diversity, 100 = highest diversity), high linearity of the trajectory structure (0 = highest linearity, 100 = lowest linearity), low sentiment variation (0 = lowest variation, 100 = highest variation), topics that are less viral (0 = lowest virality, 100 = highest virality). Therefore, the interestingness measures are specified with the following values: topic diversity = 100, structure linearity = 0, sentiment variance = 0, speed variance = 0. The resulting top six conversation trajectories are shown in Figure 4.13. The clusters that are visualised in different colours indicate the diversity in the dominant topics that are discussed in each cluster. The first five trajectories refer to discussions related to sports, and the respective hashtags stand for the acronyms of the different sports events that are being discussed: #6nations / #sixnations = rugby championship tournament, #nufc = newwcastle united football club, #safc = scarborough athletic football club, #mufc = manchester united football club. The sixth ranked trajectory with #oomf, being an exception to the rest of the sports related conversations, stands for 'one of my friends / followers', and is used for expressing secret flirtations and fights. Interestingly, most of these conversation trajectories further indicate a cyclic structure that starts and ends at the same location.

4.4.2 Example Scenario 2

For this example, we use a geotagged Twitter dataset collected during a 48 hour time frame from 19-20.02.2014 that was during the week. To retrieve conversation trajectories from more clusters, where each cluster contains smaller crowds of people, discussing similar topics in a very dense area, we set the following clustering parameters:

Table 4.3: The top five hashtag conversations

Rank	Hashtag	No. of Clusters	No. of Tweets
1	#traffic	67	1589
2	#london	14	1047
3	#faceofmlb	139	3157
4	#fml	4	999
5	#gold	32	1049

Figure 4.14: Visualisation of the #fml trajectory.

eps = 5 km, MinTws = 5, tf = 15 min. Further, in this example we are interested to see trajectories that tend to have less diverging topics, high linearity in the trajectory structure, highly contradictory, and less viral. Therefore, to filter and rank the resulting trajectories, we specify the following interestingness measures: topic diversity = 20, structure linearity = 20, sentiment variance = 80, speed variance = 10. The resulting top five hashtag conversations are listed in Table 4.3. Upon visualising these trajectories, #fml trajectory particularly showed episodic clusters that are spatially closer to each other, and that contained mostly negative sentiments. This is shown in Figure 4.14. #fml is a popular term referred to by Tweeters when they have a particularly bad day or are in a bad mood.

To have a closer look at this trajectory, we further refine our DBSCAN parameters to eps = 5 km, MinTws = 1, tf = 120 min. From the resulting trajectories, we found a sub-trajectory belonging to a cluster from the #fml trajectory. This is visualised in Figure 4.15. This sub-trajectory indicates the negative experiences of one particular tweeter that keeps evolving over the course of time (refer to the tweets that are indicated in Figure 4.15). This trajectory at its very fine granularity further indicates a very linear structure that starts and ends at two locations.

The above two example scenarios show exemplary conversation trajectories that are extracted under two different DBSCAN clustering parameter settings as well as different interestingness measures. The conversations from both scenarios differ on the one hand in terms of their spatial distribution and their geospatial structure,

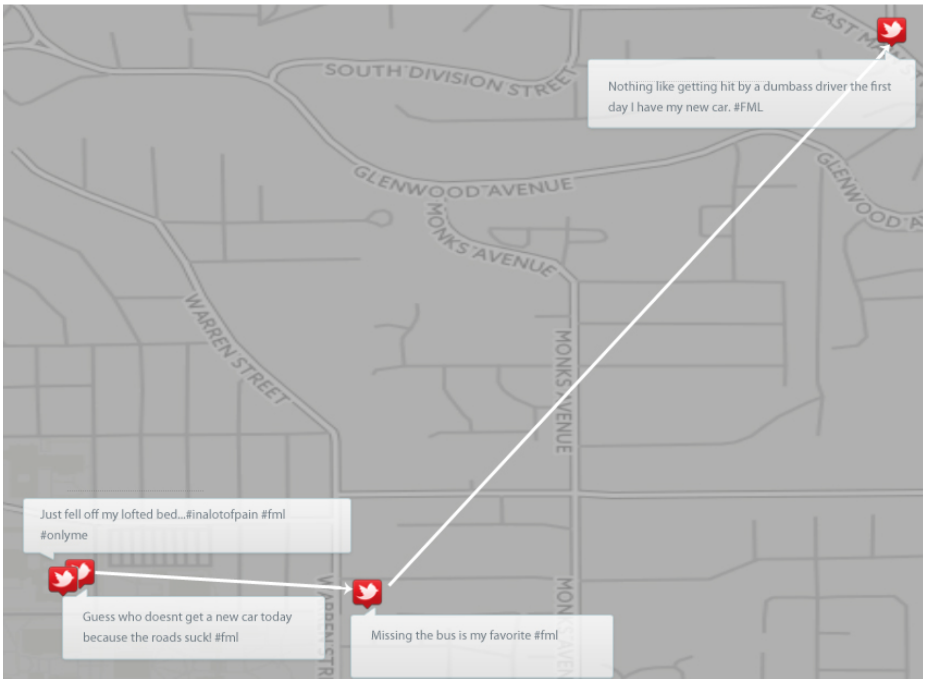


Figure 4.15: Visualisation of the #fml sub-trajectory with refined DBSCAN parameters to look at an even more dense area. This trajectory indicates one particular tweeter's negative experience that keeps evolving over the course of time. The tweets contained in each cluster are indicated next to each cluster.

and on the other hand in terms of the thematic nature of the conversations. The trajectories in scenario 1 that consider a larger radius for clusters, are spatially distributed over a larger geographic area, and further exhibit a cyclic structure for the trajectories. Trajectories in scenario 2 however consider a smaller cluster radius and therefore are spatially distributed over a smaller geographic area. These trajectories further indicate a more linear structure. In terms of the thematic nature, scenario 1 mainly contains conversations that discuss several sports clubs (hence high topic diversity), and scenario 2 mainly contain personal opinions about the various *current* situations, such as traffic, Brit Awards 2014 held in London, a competition for major league baseball (#faceofmlb), or simply expressing negative experiences with #fml. A third example scenario that incorporates another set of clustering parameters and interestingness measures to filter and rank trajectories is presented in Section 4.5. This example further showcase other visualisation views that are utilised for the interactive interestingness-based conversation trajectory extraction. Differences can further be identified in this example scenario, where particularly a back and forth structure for the trajectories can be observed.

The Figure 4.16 shows an overview of the conversation trajectory ranking and

characterisation functions of the MovingOnTwitter approach.

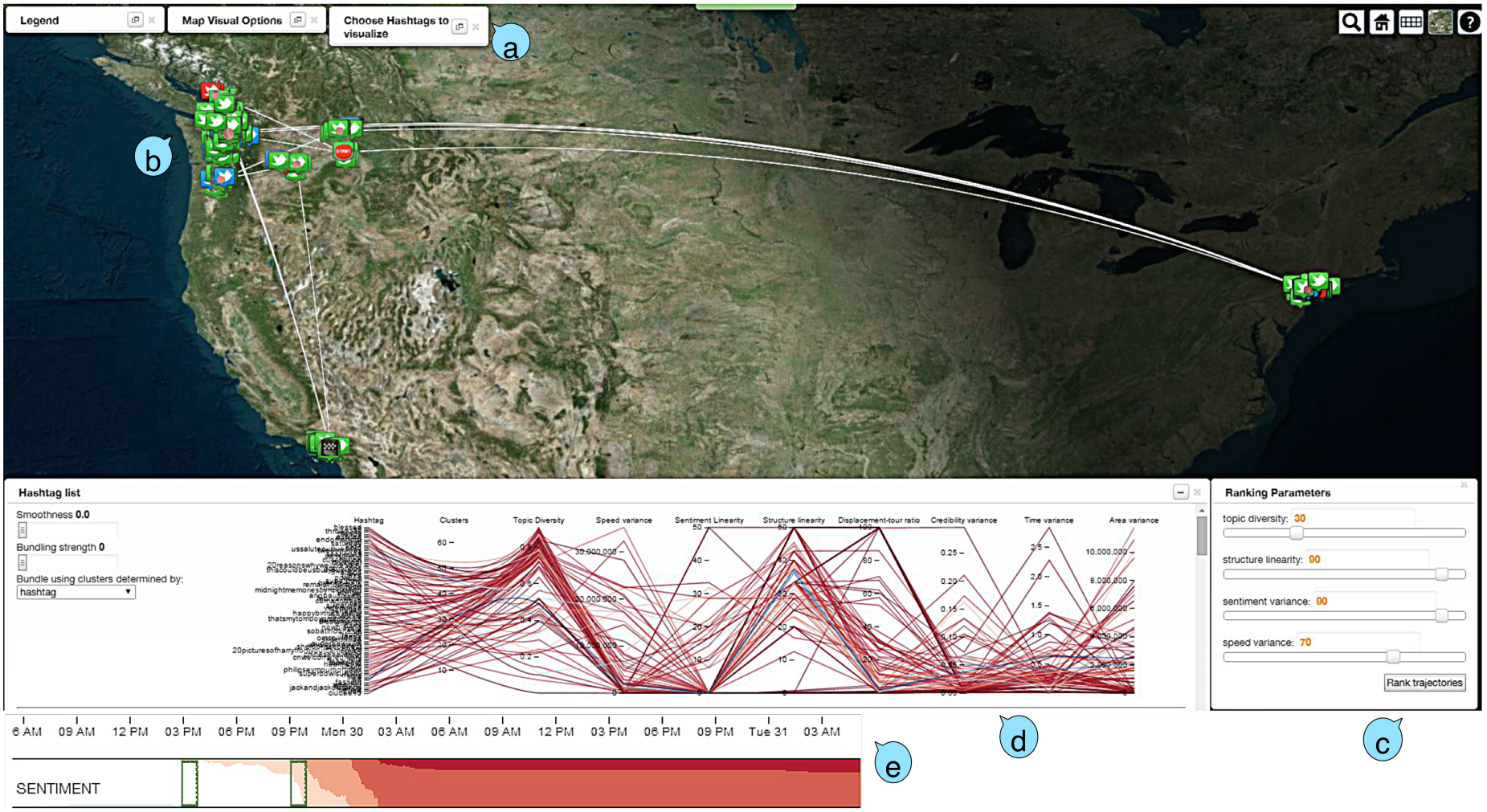


Figure 4.16: An excerpt of MovingOnTwitter overview. (a) drop down menu of the list of hastags in the data, (b) Tweets pertaining to the selected hashtag are visualised on the map, (c) parameterisation, (d) parallel coordinates plot for analysing trajectory characteristic features, (e) trajectory sentiment analysis. This figure appears in Senaratne et al. (2016, under review).

4.5 Conversation Movement Analysis for Sports Journalism

In this section, the usefulness of characterisation and ranking of conversation trajectories using our MovingOnTwitter approach is further demonstrated within a sports journalism use case. We showcase how our approach aids to exploratively analyse Twitter social media data to report on current topics, trends, sportsman popularity, and *interesting* changes in trends following the work flow as seen in Figure 4.1. In the following we first describe the dataset and our Twitter mining process, followed by an illustrative use case example using our developed MovingOnTwitter system.

Several features are used to detect and analyse the drifts in hashtag-based trajectories that are derived from episodic sequential clusters. With hashtags, Twitter analysts are able to categorise tweets based on the topic of interest. By analysing our example dataset, we have seen that the resulting trajectories can have a different level of interest to an analyst. These different structures can be observed through the characteristics of the trajectory. Our developed system enables the user to flexibly detect, analyse, and visually explore these trajectories.

Our dataset consists of all geo-tagged tweets from 19. February 2014, 00:00:00 to 20. February 2014, 23:59:59. The geotag of these Tweets comes from the location where the Tweet is originating from (either recorded by the in-built GPS - Global Positioning System, or the WiFi location). Overall, there are 8,607,490 tweets. In a sports journalism use case, the main task of the analyst is to explore trending discussions for sports.

In order to define the appropriate DBSCAN clustering parameters, we assume that this type of discussion mostly originates from more populated cities. Therefore, we set the neighborhood minimum to 50 tweets (MinTws = 50), and the radius to 15km (eps = 15km). Because many sport events have a length of over one hour, we begin with a maximum time frame of 45 minutes (tf = 45min). As a result we found 163 hashtag-based trajectories within the dataset. However, we do not know what kind of hashtags are used, but we know that sports events are often discussed with difference of opinions. Therefore in a next step, we begin by setting the ranking parameters, the topic diversity and sentiment variance to 100. We are not sure about the expected speed and structure of the dispersion, so we set these parameters to 50. The system provides us a sorted list of the hashtag-based trajectories according to these parameters. The top 20 trajectories are shown in Figure 4.17.

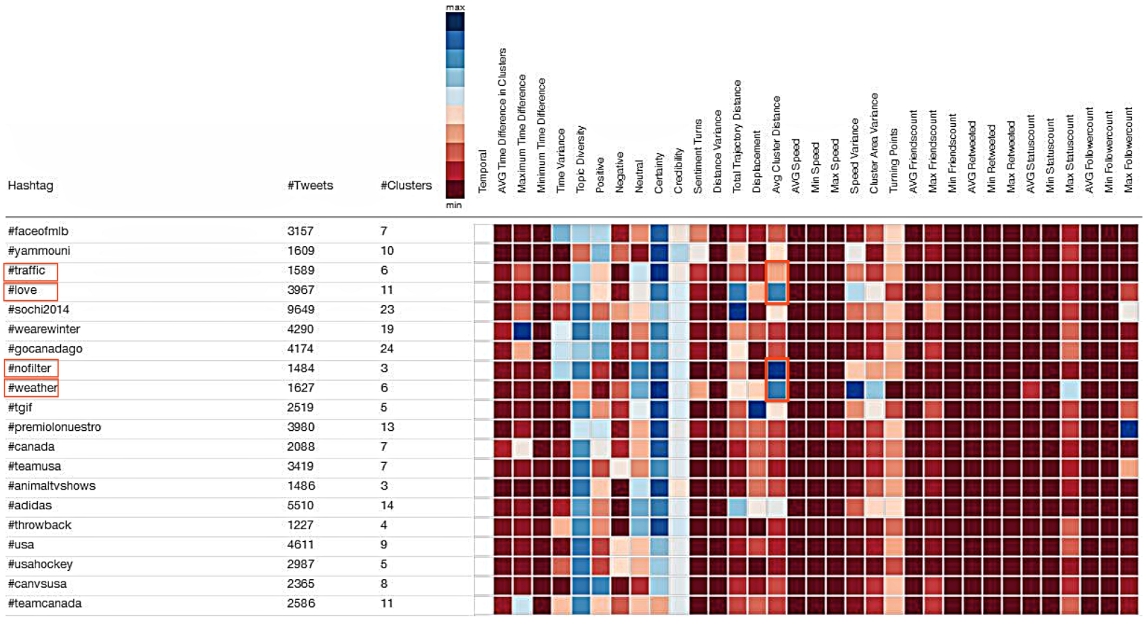


Figure 4.17: The list of hashtag-based trajectories sorted according to the parameters. The statistical values of the trajectory features are indicated in a heat map visualisation. The unusual high avg. cluster distance for non-sports related hashtags is highlighted in red. This figure appears in Senaratne et al. (2016, under review).

Some of the hashtags appear to be not sports related (#traffic, #love, #nofilter, #weather...), and they seem to have higher average cluster distance compared to the rest of the hashtags as seen in Figure 4.17 (highlighted cells). By further filtering the results with a lower average cluster distance threshold these non-relevant hashtags disappear. These results which now comprise of the top ranked 19 trajectories are shown in Figure 4.18 with a parallel coordinates plot to visualise the variance values of the trajectory characteristics derived in Section 4.3 (topic diversity, speed, sentiment, structure linearity, credibility etc.).

As one can observe in Figure 4.18, #faceofmlb has a very high sentiment variance. A background check[20] on #faceofmlb reveals that the marketing department of the Major League Baseball (MLB) initiated a Twitter contest to allow Twitter users to vote on which MLB player should be the face of the 2014 MLB season. The cluster boxplots for #faceofmlb in Figure 4.19 show two similar structures, each within a span of 24 hours, referring to the voting. Evidently, the voting was carried out in two sessions. First, Twitter users could nominate one player from each of the 30 teams to compete in the contest. Next, the winning player with most nominations from the first round goes on to the second round to compete with another winner.

[20] `http://www.copypress.com/blog/face-of-mlb-highlights-dangers-of-social-contests/`

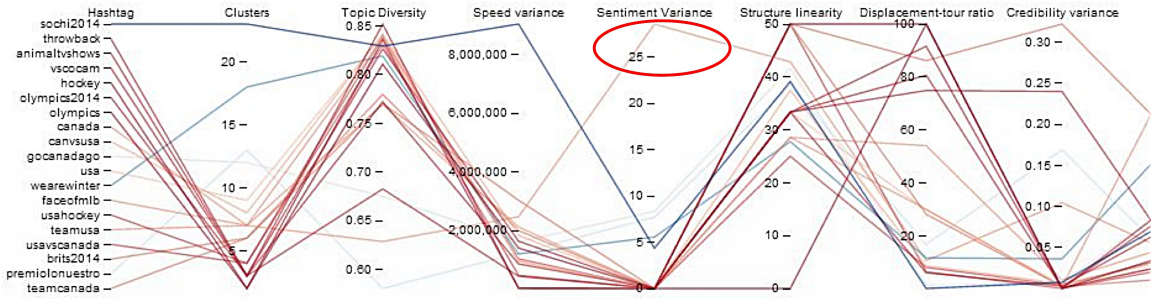


Figure 4.18: Parallel coordinates visualisation helps to extract the top ranked hashtag-based trajectories based on a lower cluster distance. The vertical axes of the plot further represents the derived trajectory characteristics. The high sentiment variance for #faceofmlb is highlighted in red. This figure appears in Senaratne et al. (2016, under review).

When the #faceofmlb trajectory is mapped on a geographic map as seen in Figure 4.20, we can observe that at the beginning of the contest two specific players Joey Votto and Felix Hernandez representing Cincinnati Reds and Seattle Mariners teams respectively were being voted for. Upon a sentiment and text content analysis as seen in Figure 4.21 we can clearly derive that Joey Votto lost against Felix Hernandez.

However, upon analysing the average credibility of the clusters appearing in Seattle (representing the winning team player) in Figure 4.22 we can identify a lower credibility for the second cluster. The text representing this cluster indicates that some voters cheated by voting many times using many fake accounts.

4.6 Discussion & Future Work

This chapter presents an approach to visually explore and analyse movement trajectories based on implicitly referenced spatial information in text-based VGI. We have also encountered a number of limitations that pave path for future improvements and further exploitation of such implicitly referenced spatial data, such as Twitter microblogs. First and foremost, the biggest limitation is the size of the data sample that we have worked with. The Twitter streaming API allows us less than 3% of the total volume of Tweets, and further the geotagged Tweets statistically represent as little as 3% of the allowed volume. Therefore, we have worked on a minority of the entire dataset that is not sufficient to derive any generic conclusions. One way to maximise the location analysis of these Tweets is to extract spatial references from the text. With proper uncertainty analysis approaches in place (as described in the

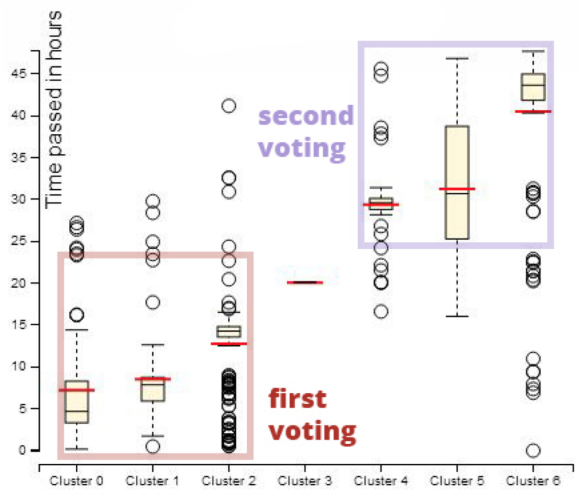


Figure 4.19: Cluster boxplots for #faceofmlb. The two similar structures are evidently resembling the two tiered voting sessions during the face of MLB contest. This figure appears in Senaratne et al. (2016, under review).

Figure 4.20: #faceofmlb conversation movements at the beginning of the contest. Tweeters are talking about two specific players representing the Cincinnati Reds and Seattle Mariners teams. This figure appears in Senaratne et al. (2016, under review).

guidelines in Chapter 2), we can derive accurate location information from text analysis. A combination of traditional event detection approaches with our hashtag-based approach may improve the accuracy of movement trajectory analysis, in particular the detection of conversation movement trajectories. Using the hashtags-based approach further limits the data search space. While this could be a good aspect for narrowing down the data, this also removes a lot of Tweet candidates from the dataset that may not use hashtags. Furthermore, the uncertainty measures and the parameterisation of the model (e.g., ranking by interestingness features) for our use cases were done based on a heuristic nature. Such heuristics can adapt to the anecdotal use cases at hand. A compilation of such anecdotal use cases may serve as a body of knowledge to learn from in future work.

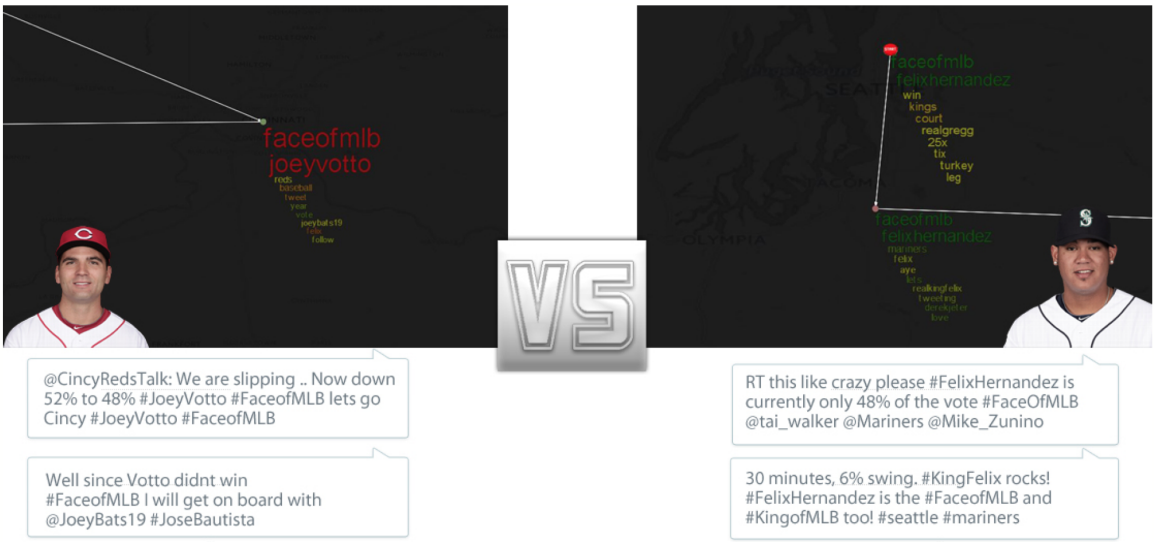


Figure 4.21: Content and sentiment analysis for the tweets discussing the two players, Joey Votto and Felix Hernandez. This figure appears in Senaratne et al. (2016, under review).

Social media analysis tools, such as the MovingOnTwitter tool are sometimes questioned regarding their usefulness based on their impact of discovery. A frequent question we receive especially from the geo-sciences field is what can these visual analytics tools discover that we otherwise cannot discover from the conventional information foraging sources such as newspapers or television. We can answer this question by highlighting the main advantages of MovingOnTwitter. In comparison to the conventional information foraging sources, our tool allows the users to explore and discover trajectories and incidents by also tapping into the additional dimensions attached to it- such as the sentiments regarding the incident, or the uncertainty regarding the incident.

Furthermore, the user can *infer* information based on these explorations. For example in Section 4.3.2 we discovered a cancellation of the Lady Gaga concert purely based on the topic and sentiment analysis. Or in Section 4.5 with the help of credibility analysis we discovered that the winner of the Face of MLB 2014 tournament may have been due to falsified Tweets. Such discoveries are rare and time consuming when the user has to rely on conventional mediums alone.

Furthermore, our approach MovingOnTwitter enables the user to sort and rank outcome trajectories based on a user-defined interestingness measure that relies on carefully thought-out geographic and content characteristics of trajectories. This considerably saves discovery time in comparison to parameter browsing and searching. An additional advantage here is that the user can narrow down the search space

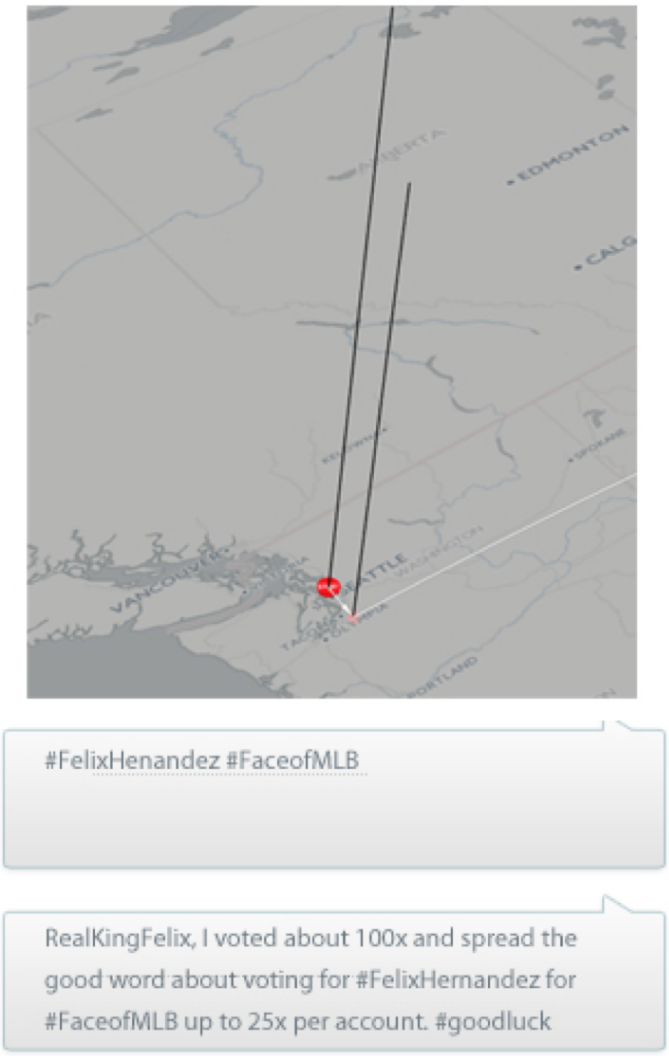


Figure 4.22: Credibility of clusters mapped to the length of the bars (shorter bar indicates lower credibility). Cluster with lower average credibility also indicates that tweeters have made many fake profiles. This figure appears in Senaratne et al. (2016, under review).

specific to the analysis interests, thereby reduce the uncertainties of the outcome. This is demonstrated in Section 4.4.

Both ranking and characteristics analyses in this chapter show that simply clustering the data alone is not enough to derive useful and meaningful insights. Although we evaluate our approach through the anecdotal findings (Munzner, 2009) from ranking and characteristics analyses that attest the ground-truth, we need to further evaluate our approach in order to generically apply the developed methods in other use cases as well as other fields. In future work a comprehensive evaluation of our approach can be carried out for example by following the evaluation framework presented by Stavrakantonakis et al. (2012). Their framework is specifically developed for evalu-

ating social media monitoring tools, addressing three issues: (1) *the main concepts related to social media monitoring such as, analysis, insights, engagements etc.*, (2) *the technology used*, and (3) *user interface*. Although they evaluate social media monitoring tools particularly for enterprises, their framework could be followed to evaluate the MovingOnTwitter tool in the future addressing the same issues.

Together with the validation guidelines by Munzner (2009) within such a framework we can evaluate MovingOnTwitter.

A conceptual example of how to carry out such an evaluation on the basis of the three issues identified in the framework by Stavrakantonakis et al. (2012) is as follows.

(1) *Insights gained*: ScatterBlogs2 (Bosch et al., 2013), SensePlace2 (MacEachren et al., 2011), and MovingOnTwitter all three tools are developed for Twitter data analysis and therefore handle the space, time, and content dimensions. For a given dataset and a main task, by utilising the various workflows, methods, and visualisations one can qualitatively compare how the task is solved and the insights gained. This can be done for example with target users. This is similar to what Munzner (2009) call *abstraction validation*.

(2) *Technology used*: This can be achieved by analysing the computational complexity by measuring the system time and memory.

(3) *User interface*: One way to evaluate the user interface would be to follow the guidelines in Chapter 2 that have been identified to assess the uncertainties of the visual mapping (Section 2.2.2), visualisation (Section 2.2.2), and the coupling between the model and the visualisation (Section 2.2.2). Also as identified in Munzner (2009), the user interface can be evaluated using qualitative or quantitative result image analysis or by measuring the human time and error for operation within a lab study.

In future work either one or all of the above issues can be evaluated.

Furthermore, the parameterisation in the ranking process are done relying on our background knowledge. However for a generic application where the user may not have necessary background knowledge or misspecifies the parameters, our approach could further extend in the future to include methods to assess the suitability of parameters. For example, as described in the guidelines in Chapter 2 Section 2.2.2 we can quantify the uncertainties of the paramerisation in place of distance functions to measure the distance of parametrisation from the true value. Additionally, in a review of feature

selection methods, Kumar and Minz (2014) point out to probability of error measures (Devijver and Kittler, 1982), dependency measures (Hall, 2000), inter-class distance measures and consistency measures (Liu and Motoda, 1998) to evaluate the derived features. These can be incorporated into a future version of the MovingOnTwitter tool.

4.7 Conclusions

Text-based VGI have proven to be useful e.g. for event detection, individual movement tracking, or disease tracking. However, due to the implicit nature of the spatial references in such data an accurate observation of above said phenomena becomes a challenging task. In this chapter we introduce the episodic sequential hotspot analysis approach for detecting movement patterns of phenomena from large implicitly referenced text-based VGI. Two approaches are utilised to gather and group data based on two instances: (1) keyword-based relevance approach for instances where the user possess prior knowledge of the phenomena she/he wants to track, (2) hashtag-based approach for instances where the user does not possess prior knowledge.

The hashtag-based grouping strategy further considers the geospatial and content structures as the characteristics to filter out the meaningful and interesting trajectories. As geospatial characteristics we have derived *distance variance*, *trajectory linearity*, and the *speed variance*, and as content characteristics we have derived *topic diversity*, *sentiment linearity*, *cretainty variance*, and *credibility variance* to identify interesting as well as accurate movement trajectories. Furthermore, to allow more flexibility for the user to control the filtration of results, the proposed approach consists of an interestingness-based ranking mechanism for the movement trajectories. The features that define the interestingness of the trajectories are: the topic diversity as it aids to observe trajectories, the sentiment variance that gives insights to the discrepancies in discussions, structural linearity to indicate movement, and speed variance to determine the virality of the topic. This ranking and sorting approach further help the users to reduce the uncertainties of the outcome by parameterising the model to their specific needs. The usefulness of text-based VGI analysis under these characteristics and interestingness-based ranking is demonstrated within a sports journalism use case using Twitter data. Finally, we discussed a number of limitations of the approach and resulting options for interesting future work.

Chapter 5

Uncertainty-aware Space-time Exploration from Location-based Mobile Communication Data

Contents

5.1 Background & Related Work 127

5.2 Location-based Mobile Communication Data 128

5.3 Spatio-temporal Analytics through Movement Patterns 129

5.3.1 Movement Trajectory Extraction from Location-based Mobile Internet Usage Data 129

5.3.2 Spatial and Temporal Movement Similarity of Users 132

5.3.3 Place Classification based on Home and Work Area Detection 135

5.3.4 Regional Partitioning with Origin-Destination Analysis . . 137

5.3.5 Spatial Area Analysis based on Temporal Usage Patterns . 140

5.4 Uncertainty in Movement 141

5.4.1 Space-time Prisms to Analyse Uncertain Movement Path Segments 143

5.4.2 Positional Uncertainty Reduction in Mobile Communication-based Movement Data 146

5.5 Discussion & Future Work 148

5.6 Conclusions 150

The previous chapter looked at how movement can be detected through implicitly referenced spatial data, and how characteristics such as uncertainty can be explored using visual interactive analysis approaches. In this chapter we explore visual analytics approaches for finding spatio-temporal patterns through user movement from *explicitly referenced location-based spatial data*. Furthermore, this chapter explores approaches

for the analysis of uncertainty in such explicitly referenced location data based on the guidelines identified in Chapter 2. Explicitly referenced location data comprises geographic coordinates. These coordinates can come for example from GPS measurements. Those GPS data are carrying uncertainties due to GPS precision errors, temporary unavailability of GPS data (e.g., when a user is indoors), or interpolation of crisp locations.

In case of collected *mobile communication* data from users, location can be approximated by the location of the closest receiver antenna. This results in even higher positional uncertainty than GPS data. Hence, in this chapter we take such mobile communication data as an example. We develop a visual analytics approach to (1) extract movement trajectories from the mobile communication data, (2) explore spatio-temporal patterns though the derived movements, and (3) assess the uncertainty of the movement trajectories derived from location-based mobile communication data. Thereby, an interpolation over the sequential change of locations over a period of time indicates a movement trajectory of a given user (as seen in Chapter 4).

This chapter unfolds as follows: Section 5.1 reviews the related work on visual analysis of mobile communication data. In Section 5.2 the location-based mobile communication dataset is briefly reviewed. In Section 5.3 the integration of data visualisations with suitable data analysis algorithms for the exploration of spatio-temporal patterns are described. Under this section, approaches are presented for: the extraction of movement trajectories (Section 5.3.1), the analysis of spatial and temporal similarity (Section 5.3.2), place classification (Section 5.3.3), regional partitioning through origin and destination analysis (Section 5.3.4), and the spatial area analysis based on temporal patterns of mobile communication (Section 5.3.5). Section 5.4, describes on the one hand an approach that adapts the space-time prism for uncertain markers to analyse the uncertainty of location-based movement patterns, and on the other hand an approach for the *reduction* of uncertainties in movement patterns that are extracted from location-based mobile communication data. This chapter ends with a discussion in Section 5.5 and a conclusion of the presented approach in Section 5.6.

The contents of this chapter are based on the publication Senaratne et al. (2017b)[1].

[1]This work is a result of a collaboration with M. Mueller from the University of Konstanz, M. Behrisch from the University of Konstanz, F. Lalanne from the Inria Institute Chile, J. Bustos from the University of Chile, J. Schneidewind from O_2 Telefonica Germany, D. Keim from the University of Konstanz, and T. Schreck from the Technical University of Graz. My contribution as the first author within this publication was defining together with M. Mueller the approaches in all four aspects of this chapter: (1) extract movement trajectories from mobile communication data, (2) explore spatio-temporal patterns though the derived movements, (3) assess the uncertainty of the

5.1 Background & Related Work

With the increase of urbanisation in cities and more than 50% of the world's population living in urban areas[2], it is imperative to foster careful planning in cities. Ubiquitous data such as *mobile communication data* can be utilised as alternatives to officially surveyed data, and therefore used as digital footprints to understand the urban dynamics (e.g., movement of people), city infrastructure, spatial connectedness, or segregation within a city. The derived insights can then for example be utilised to optimise city planning (Calabrese et al., 2011a).

Deriving actionable, useful knowledge from such large amounts of complex location-based data requires appropriate data analysis methods and models. Visual-interactive data analysis methods can help to support such data analysis by effective integration of automatic data analysis methods, interactive steering of the analysis process, and result visualisation (Keim et al., 2008). Those methods can be particularly helpful when analysing mobile communication data. The large scale that network systems can reach in terms of users, nodes and traffic, as well as the large number of variables that can influence services, create a challenge. Deriving insights of data collected from citizens' mobile communication through visual analytics can facilitate various city management tasks. For example, city planners can analyse which strategic decisions are to be taken to improve a city's infrastructure, the transportation system, or general urban planning.

Different works have utilised such mobile network data to explore spatial and temporal patterns in user communities. Sagl et al. (2012) developed a visual analytics approach for analysing collective human mobility based on spatio-temporal patterns. For the visual analysis of their data (e.g., using time series visualisations) they used Tableau[3] together with SPSS Statistics[4] software, and for the spatial visual analysis of the data (e.g., using heat map visualisations) they used ESRI's ArcGIS[5] software. In Shen and Ma (2008) MobVis is presented, a tool that incorporates spatial and social data heterogeneously into one network to analyse individual and group behavioral patterns. They use interactive time charts and ontology graphs to support the temporal and semantic filtering of their data. They further introduce a *behavior ring* visualisation method to represent and compare individual and group behaviors of

movement trajectories derived from location-based mobile communication data, and (4) derive an approach to reduce uncertainty in the movement trajectories.

[2] `http://www.demographia.com/db-worldua.pdf`

[3] `http://www.tableau.com/`

[4] `http://www-01.ibm.com/software/analytics/spss/`

[5] `http://www.esri.com/software/arcgis`

people. Furthermore, Gao et al. (2013) utilise one week's mobile phone data to analyse the structures of spatial interaction communities of a city in China. Their approach mainly utilises agglomerative clustering techniques together with heat map and graph visualisations to discover the structures of spatial interaction in communities. Further, the work of Ferreira et al. (2013) used taxi data to explore urban activities in New York, USA. They developed a visual analysis system that allows the users to visually query various taxi trips around the city. Tasks such as origin-destination analysis of human mobility were performed by enabling adequate spatio-temporal querying, using heat map and time-series visualisations.

In this chapter, a novel visual analytics approach for pattern exploration and search in GSM (Global System for Mobile Communications) data of metropolitan area networks is introduced. The approach defines interactive visualisation methods allowing the exploration of mobile communication based on antenna location information and user movement. We define geospatial and matrix based representations of data which can be interactively navigated. The approach integrates data visualisation with suitable data analysis algorithms, allowing to spatially and temporally compare mobile usage, identify regularities, as well as anomalies in daily movement patterns across regions and user groups. Specifically, we have developed methods to analyse movement trajectories and spatio-temporal similarities of user movements, place classifications based on home and work area, regional partitions based on origin and destination of movements, temporal mobile usage patterns, and further methods to assess the uncertainties of these location-based mobile communication data. We demonstrate the effectiveness of our visual analytics approach by applying it to a large, community-provided data set of mobile communication data from Santiago (Chile).

5.2 Location-based Mobile Communication Data

The dataset comes from the Adkintun Mobile project (Bustos-Jiménez et al., 2013), where the goal of the project was to measure the quality-of-service of different Chilean mobile Internet operators. The data was collected on a voluntary basis by 358 users, who were required to install an application on their Android phones. In return, the users gained insight into their quality-of-service and were able to review their personal mobile Internet usage. The data was recorded in a period of seven months from October 2012 to May 2013, where the mobile application logged device and network usage data every ten seconds. Each record contains information about the network traffic (received/sent bytes), the antenna operator, the Location Area Code (LAC) of the

currently logged-on antenna, the Cell Identifier (CID), the signal strength (measured in dBm) and a measure which indicates whether the phone was in active use at the time of the measurement. A second table contains the geographical coordinates of over $60,000$ antennas across Chile along with their respective LAC and CID. However, within the scope of this study we use data only for the city limits of Santiago. The data is stored in a PostgreSQL 9.3[6] database with the spatial extension PostGIS[7].

We developed a visual analytics approach with six encompassing views to analyse this dataset within five urban analyses tasks. In the following our visual analysis system is described based on these analysis perspectives.

5.3 Spatio-temporal Analytics through Movement Patterns

Exploration of mobile phone datasets can provide us insights into for example the urban dynamics of a city that help various stakeholders to make better decisions and services to their targeted audiences. We carry out our analysis tasks in two tiers. Firstly, we focus on the *spatial changes* of mobile Internet usage patterns. For this we extract user movement trajectories through the digital footprints left behind by mobile users, and use these changing movements of users to uncover spatial patterns in the metropolitan area of Santiago de Chile. Secondly, we focus on the *temporal changes* of mobile Internet usage and use these to uncover spatial patterns and events.

5.3.1 Movement Trajectory Extraction from Location-based Mobile Internet Usage Data

A trajectory is a trace generated by user movement in geographical space, usually represented by a series of chronologically ordered points $(p_1, p_2, \ldots, p_n)$, where each point consists of a geospatial coordinate set and a time stamp, such as $p = (x, y, t)$. A movement analysis can help to classify the users based on their movement, and thereby use these classifications, e.g., for targeted marketing. Further, it helps mobile operators to improve their services on the most traversed routes, thereby increasing customer satisfaction. For the following analyses we use map and matrix views.

The monitor data allows us to approximate trajectories of users with antenna positions as consecutive locations. This means that a user can be anywhere within the operational range of an antenna. This low spatial granularity of locations aids

[6] `https://www.postgresql.org/`
[7] `http://www.postgis.net/`

us in analysis through general regions as opposed to exact locations. This aspect is important also for the anonymisation of the data.

The trajectory extraction algorithm from these data consists of two steps. First, we define and detect a *movement session*. A session is defined as a time interval of continuous Internet usage of one user, that contains a set of n tuples $(t|s)$ where t is a point in time and s a point in space. A session is considered to be over when the time difference of two consecutive monitor records exceeds a user defined threshold. Therefore, a session S can be defined as follows:

$$S = (t_1 s_1), (t_2 s_2), \ldots, (t_n s_n) \tag{5.1}$$

In the second step we extract the trajectory out of these sessions where each session gets divided into one or more trajectories. To decide where one trajectory ends and another starts, the spatial element of the $(t|s)$ tuples is examined. Whenever a sequence of tuples without spatial change exceeds a certain duration threshold, the current trajectory is considered to be over and a new one is created. This is expressed in the following trajectory extraction algorithm:

ALGORITHM 1: Algorithm for extracting trajectories from mobile Internet sessions

Data: S, threshold k
Result: A set of trajectories
Start a new trajectory from $tuple_i (i = 0)$;
foreach *following tuple* **do**
    Find closest cluster C_i;
    if $t_i - t_i + 1 < k$ **then**
        add $tuple_i + 1$ to the current trajectory;
    else
        end current trajectory and start a new trajectory from $tuple_i + 1$;
    end
end

Consecutive tuples of a session which have the same spatial component are grouped together. These trajectories can be extracted at different granularity. Highest granularity is when antenna locations are used as spatial elements of the tuples (Figure 5.1), and lowest when the cluster centers are used as spatial elements (bigger clusters therefore result in coarser trajectories).

The highest granularity is when we use the antenna locations for the spatial component in the tuples. For more abstract locations, the antennas are further spatially clustered using Hartigan's Leader algorithm (Hartigan, 1975; Isaacman et al., 2011). The pseudo code for the Hartigan's Leader algorithm which requires a maximum

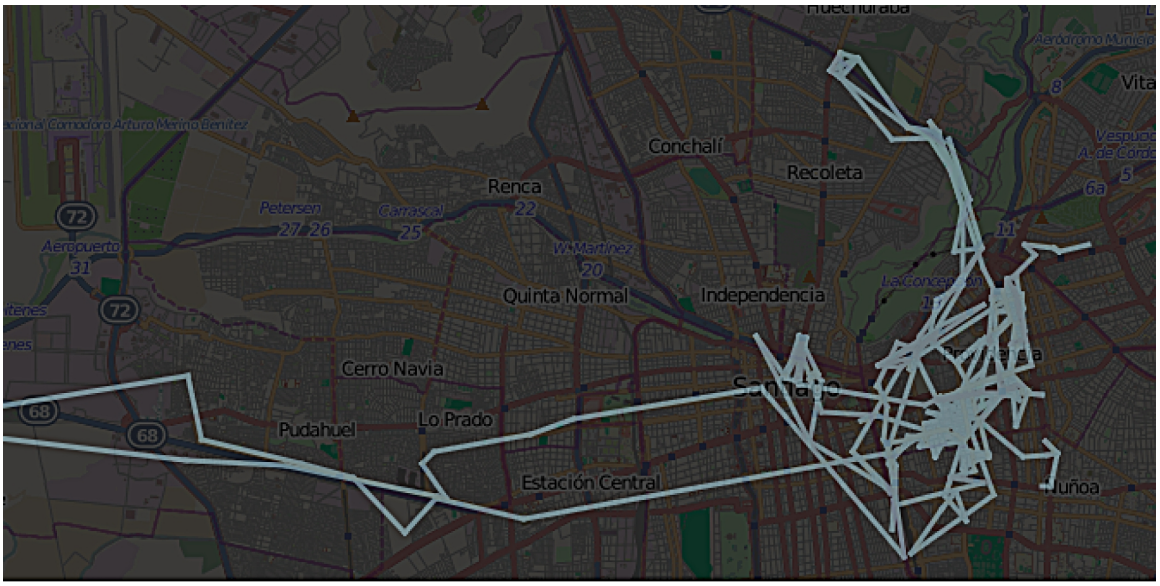


Figure 5.1: Trajectory of a selected user with antenna locations taken as the consecutive user locations. This figure appears in Senaratne et al. (2017b).

user-defined distance threshold (for the cluster radius) is as follows:

ALGORITHM 2: Pseudo code for Hartigan's Leader algorithm – adapted from Hartigan (1975)

Data: Antennas with their geographic coordinates, maximum distance threshold d
Result: A set of clusters
Assign the first antenna instance to a new cluster C_i;
foreach *antenna a* **do**
    Find closest cluster C_i;
    if $dist(a, ci) < d$ **then**
        add a to cluster C_i;
    else
        create a new cluster and assign a;
    end
end

Several radii were used to pre-calculate and generate the clusters to avoid the long computations times. The chosen radii from highest to lowest granularity were at 500m, 1000m, 2000m, 4000m. Trajectory of a selected user at these different granularity levels is shown in Figures 5.2, 5.3, and 5.4.

We have used these movement patterns in the following analysis tasks: spatial and temporal similarity analysis of users in Section 5.3.2, residential and work area detection from user movements in Section 5.3.3, spatial area analysis in Section 5.3.5, and origin-destination analysis of users in Section 5.3.4. All these behavioral analyses were carried out with anonymized user data.

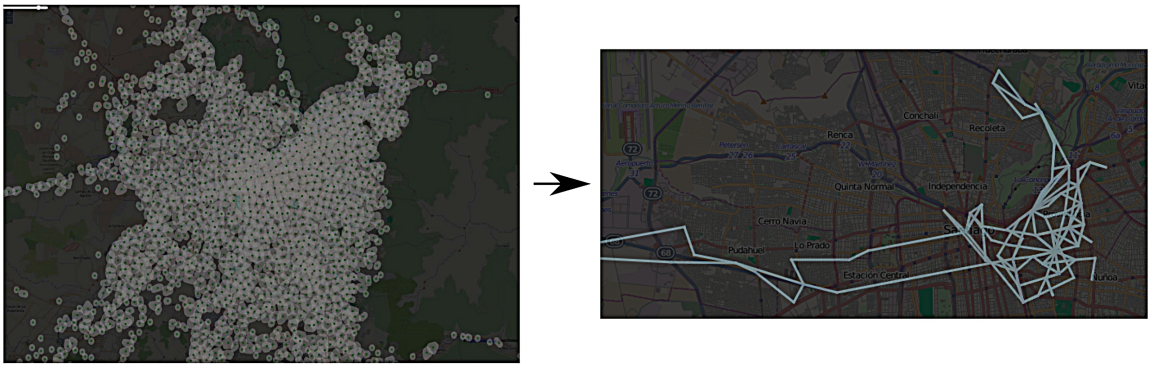


Figure 5.2: Trajectory of a selected user with 500m radius cluster centroids taken as the consecutive user locations.

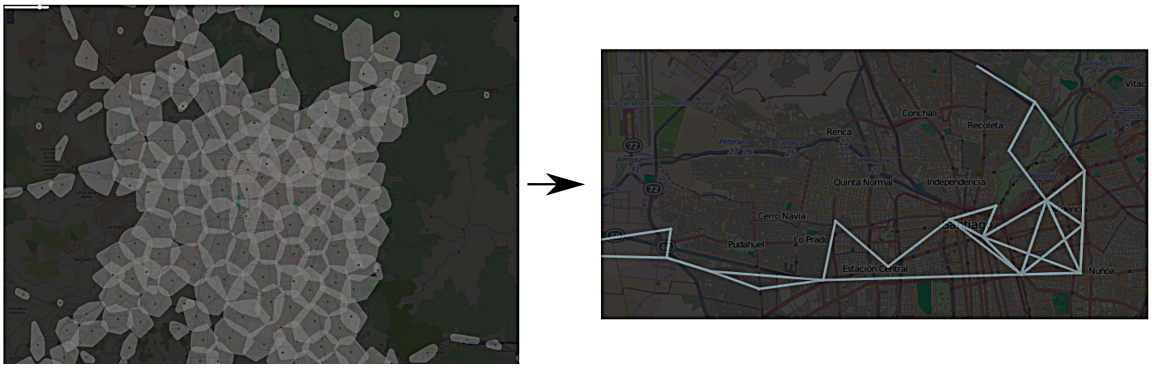


Figure 5.3: Trajectory of a selected user with 2000m radius cluster centroids taken as the consecutive user locations.

5.3.2 Spatial and Temporal Movement Similarity of Users

Similarity between two objects is said to be estimated with a *similarity measure* which may contain a cost of changing one feature into another, or the *distance between the two objects* (Faloutsos et al., 1997).

Several works have introduced and discussed approaches for the analysis of similarity in movement trajectories. The work of Dodge et al. (2012) presents a two-tiered approach for assessing the similarity between movement trajectories. As a first step they decompose trajectories into segments based on parameters such as speed, acceleration, or directions of movement. In a second step they modify the Levenshtein edit distance technique (Levenshtein, 1966) to assess the similarity of decomposed trajectory segments.

Ranacher and Tzavella (2014) on the other hand present a taxonomy of approaches for similarity analysis based on the different parameters of a moving object, such as the speed, spatial path, and time duration. Similarity analysis within two or more movement trajectories help in various demographic studies for focus group marketing,

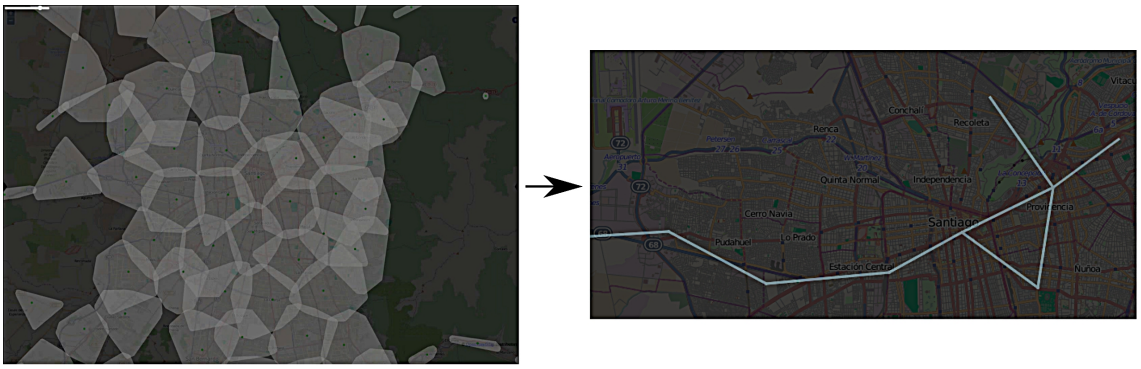


Figure 5.4: Trajectory of a selected user with 4000m radius cluster centroids taken as the consecutive user locations.

urban planning, or city traffic flow management – for example to detect which areas in a city are most connected. Andrienko and Andrienko (2007) suggested several ways in which two or more movement trajectories may be similar. These are: (1) *Similarity of overall characteristics* which considers the geometry, distance, traveled time, and movement vectors as features , (2) *Co-location in space* which considers the partial and full overlaps in geographic positions, (3) *Synchronisation in time* which considers the similar movements occurring at the same time or slightly delayed time, (4) *Co-incidence in space and time* which considers full overlaps in geographic positions occurring at the same time, or at slightly delayed times. The authors further gave suggestions of computational and visualisation techniques for detecting similarities between movement trajectories.

To assess and compare the spatial and spatio-temporal similarity between users' movement trajectories, we introduce a similarity-based matrix visualisation. This extends the classification of techniques suggested by Andrienko and Andrienko (2007) for similarity analysis in movement trajectories. We consider *co-location in space* and *co-incidence in space and time* for similarity assessment in our derived movement trajectories of mobile users. To calculate the spatial similarity of users, we compare the movement trajectory of each user for possible co-locations. To determine the locations that the edges of the trajectory passes though (as demonstrated in Figure 5.5), we calculated for each edge the intersection with other locations of the same level of granularity. The results are stored in a separate table. Using this table, we built a spatial profile for each user by extracting their distinct locations.

Accordingly, the spatial profile P for a user x at granularity g can be denoted as $P(x, g) = a, b, c, d, e, f, g$. Then, the similarity of two users x and y at granularity g can be calculated as:

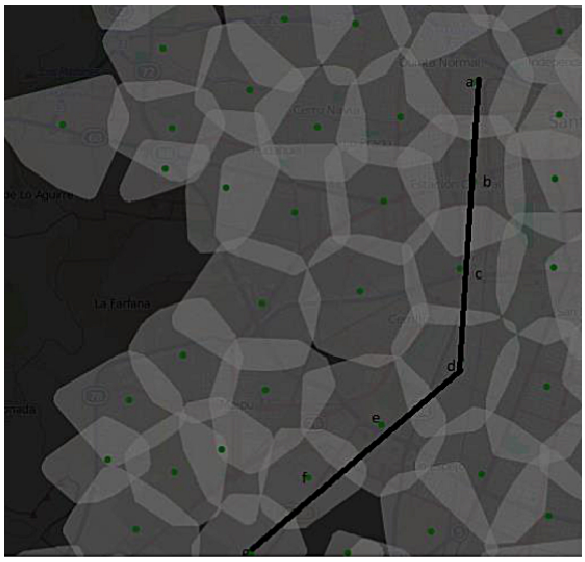


Figure 5.5: Example of two edges (a-d and d-g) of a trajectory segment passing through several locations (b,c,e,f). This figure appears in Senaratne et al. (2017b).

$$sim(x, y, g) = \frac{P(x, g) \cap P(y, g)}{min(|P(x, g)|, |P(y, g)|)}$$

To visualise the similarity values of each user pair, we map them to the size of a circular gradient (as shown in Figure 5.6 a, b). The size of the gradients is scaled between 0 (smallest gradient) and 1 (largest gradient) to visualise the normalised low-to-high similarity values respectively. These circular gradients are then drawn in the corresponding matrix cell. The adjacent map view visualises the movement profiles of the corresponding user pair, with green coloured spatial movement coverage indicating the spatial similarity (i.e., the co-locations) between the selected user pair (Figure 5.6 c).

To analyse if the movement of two users coincide in space *and time*, we check if any two spatially similar users (denoted by $L(x, y, g)$) have as well an overlap in time (using a time interval of 5 minutes) at the given location. If there is an overlap, we add the location to the set $O(x, y, g)$. Therefore, we calculate temporal similarity as follows:

$$sim(x, y, g) = \frac{|O(x, y, g)|}{|L(x, y, g)|}$$

We add this temporal component to the similarity matrix. The temporal similarity values are mapped to the opacity of the corresponding matrix cell. The results are shown in Figure 5.7. The visualisation of temporal similarity with the spatial similarity helps city planners for user demographic analysis.

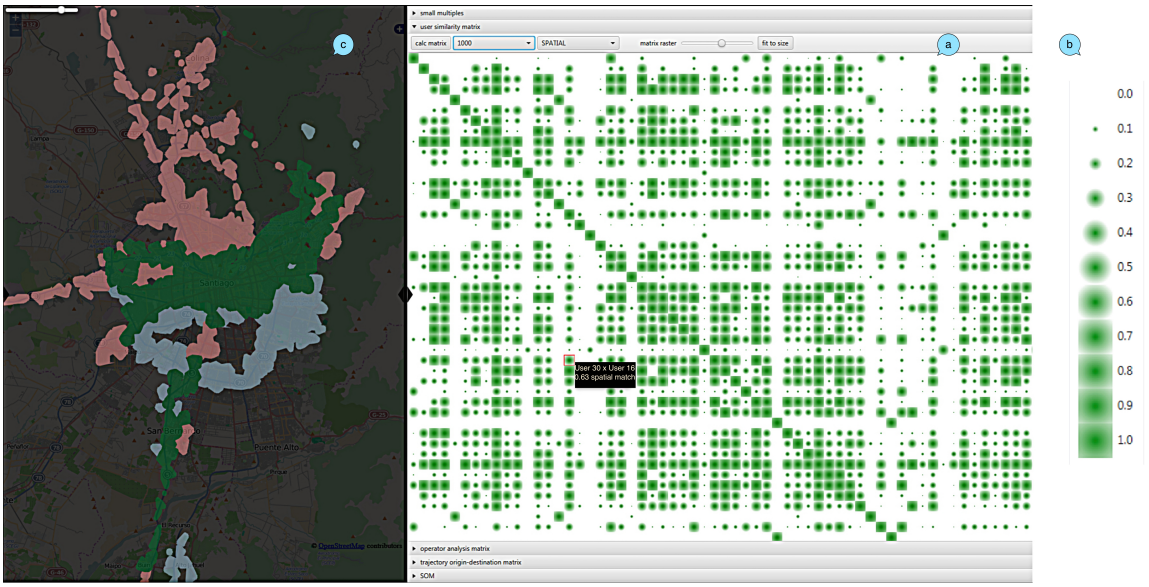


Figure 5.6: Spatial similarity matrix of users. (a) The matrix overview facilitates the analysis of unique movements which are represented through blank lines. The larger the circular gradient, the higher is the similarity. (b) Legend shows the scaling of the circular gradient. Specific user similarities can be analysed by zooming in. (c) Each matrix cell can be clicked to visualise the movement profile of the corresponding user pair on the map view. The movement profiles of the selected User 30 is shown in red and User 16 is shown in blue. The intersection of both user movements with a spatial similarity value of 0.63 is shown in green. This figure appears in Senaratne et al. (2017b).

5.3.3 Place Classification based on Home and Work Area Detection

Place classifications help to identify geographic space in terms of it's unique characteristics. Thus far semantic technologies have been very popular to identify *place* using topological and semantic descriptions (Pronobis et al., 2009; Rottmann et al., 2005). Furthermore, the movement of people can also give us hints to the characteristics of the the underlying geographic space. For example, three main forces have been identified by Hultman and Wärneryd (2001) for what drives humans to move within space and time. These are (1) *attendance*: decision to attend an event at a certain time and place, e.g., attending a funeral or a concert, (2) *existence*: for day-to-day survival, e.g., travelling to work, and (3) *imitation*: following what someone else does, e.g. touristic travels for leisure. If we can identify one, or all of these cues of human movement we can generate and identify characteristics of space that can lead to a place classification.

In our work we use human movement behaviour derived from the mobile Internet usage dataset to identify characteristics of geographic space. For this we use Hultman

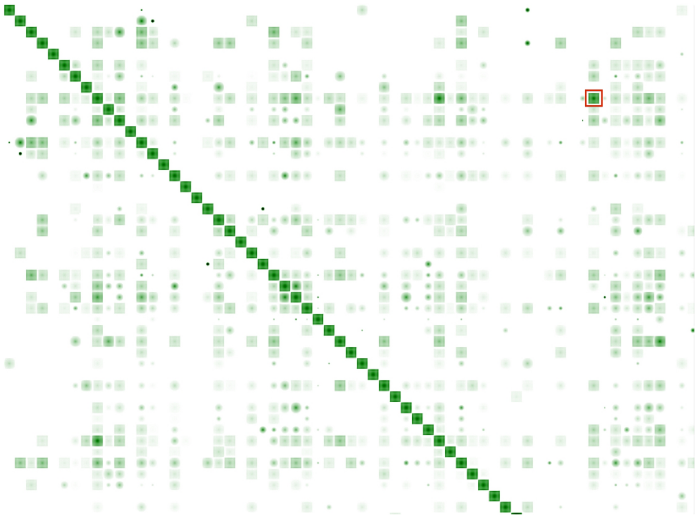


Figure 5.7: An excerpt of the spatial similarity (circular gradient) and temporal similarity (opacity) matrix of users. The highlighted cell depicts a user pair with high spatial as well as temporal similarity. The diagonal cells only reflect the self-similarity, and therefore ignored in the analysis. This figure appears in Senaratne et al. (2017b).

and Wärneryd (2001)'s *existence* as a movement cue. Since humans are creatures of habit, we can exploit their daily movement patterns to observe when they are travelling to work, and when they are travelling home. With these identified patterns, we want to classify geographic space according to residential and industrial areas. One possible application of this place classification is for real estate developers to identify their target customer groups within residential (home) and work (industry) areas. In the following we demonstrate how we achieve this for the city of Santiago, using visual analytics methods.

Forming user groups is helpful to use as an abstraction to conduct analysis on, and to help speeding up analysis in contrast to the long computation times required for ungrouped data. For this analysis we used an interactive tabular view.

Following the works of Isaacman et al. (2011), we first extract the home and work locations of users based on their spatio-temporal mobile data usage behaviors. We performed the detection using a 500m clustering granularity. For each user, the following steps were performed.

Work area detection

1. Determine the set of distinct antenna clusters visited by the user during weekdays between the time 13:00 and 17:00 (which is a core working time for most people according to Isaacman et al. (2011)).
2. Rank clusters by the number of days when the cluster was visited. If the ranking is

not clear, take the total number of records into account.
3. Pick the cluster with the highest rank as work cluster.

Home area detection
1. Determine the set of distinct antenna clusters visited by the user during weekdays between the time 00:00 and 06:00 and between 20:00 and 00:00, or during weekends.
2. Rank clusters by the number of days when the cluster was visited. If the ranking is not clear, take the total number of records into account.
3. Pick the cluster with the highest rank as home cluster

In a subsequent step, we used a reverse geotagger which takes a geo-location as input and outputs information such as administrative hierarchies of a location, and the name of the closest city. The analyst can either use the home location or work location as input. These hierarchies of locations are then mapped to the tree nodes as seen in Figure 5.8, administrative hierarchies are shown through ID values. To classify these areas according to the home/work locations we colour the tree nodes according to the work place density for each area. For every area, we count the amounts of user homes (h) and the amount of user work locations (w) (as derived earlier in this section). We then divide (w) by ($h + w$) to get a ratio value which indicates whether the area is more a residential area (values < 0.5) or more an industrial area (values > 0.5). We map the value for each node on a diverging colour scale from green (0) over orange (0.5) to red (1). The result are shown in Figure 5.8. Accordingly, the area Barrio Brasil (value of 0.92, shown in red in Figure 5.8) which is quite central in Santiago is a work/industrial area, popular for its student and night life, and the area Santa Carolina (value of 0.29, shown in green in Figure 5.8) which is 20 km outside of the city center is a more residential suburb of Santiago. Both verify our findings. The analyst can further explore which users belong to each of these home/work area nodes.

5.3.4 Regional Partitioning with Origin-Destination Analysis

In Section 5.3.3, we have seen how geographic areas can be identified in terms of their unique characteristics by analysing the movement behaviour of people. On the other hand works such as Lambiotte et al. (2008) and Ratti et al. (2010) have shown that the geography around us also has an effect on the human interactions.

In our work we want to explore the *effects of geography on the movement of humans*. To achieve this we successively partition the geographic areas based on the *origin*

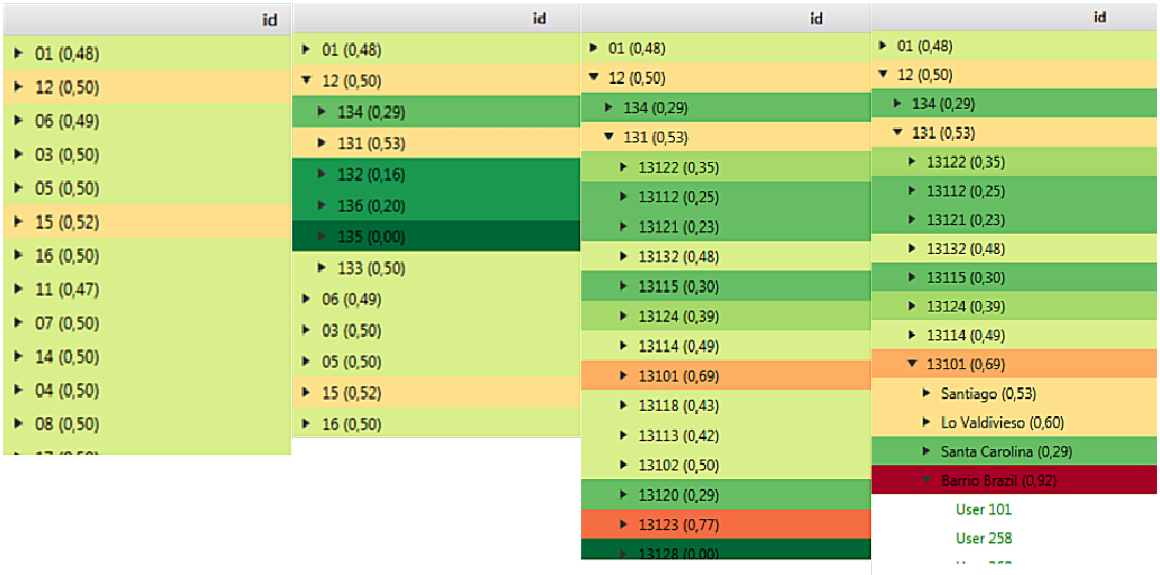


Figure 5.8: User classification by home and work areas. The tree node areas are coloured according to the home (Green)and work (Red) area ratio. This figure appears in Senaratne et al. (2017b).

and destination of movement trajectories. For example, the recent work of Tolouei et al. (2015) used mobile data collected for weekdays over a period of one month with static map visualisations to observe the origin-destination of movements within Leicestershire, UK.

The users' origin-destination movement analysis can help to uncover the attractiveness of an area, number of different places that people come from, the number of traversed trips for any time of day, or to capture the patterns of urban mobility in different areas of a city. Such analyses are useful for example in use cases within the city planning, transportation management, or emergency response domains (Calabrese et al., 2011b).

Within our approach we have developed an interactive *origin-destination matrix* to *effectively* explore and analyse where user movement trajectories (described in Section 5.3.1) start and end. To achieve this task we utilise the table view - to choose individual users (if needed), the matrix view - to view the origin and destination of users, additionally a raster map view - to partition an area into multiple spatial segments and indicate the corresponding origin / destination locations.

Using the map, an arbitrary area of interest can be selected for investigation, which is then rasterised by partitioning the map into multiple spatial segments. These segments are coordinated with the rows and columns of the matrix. Each matrix cell is associated with an origin (rows) and a destination area (column). The

origin-destination analysis for aggregated trajectories of all users is shown in Figure 5.9 component (a). The cells are coloured according to the amount of trajectories traversed within a particular origin and destination. As we can observe from Figure 5.9 component (a), most traversed routes appear to be between the center and the North-West of Santiago, and fewer routes within the South-West regions in Santiago. While this analysis looks at aggregated trajectories, the same analysis can be performed to observe the origin-destination per user.

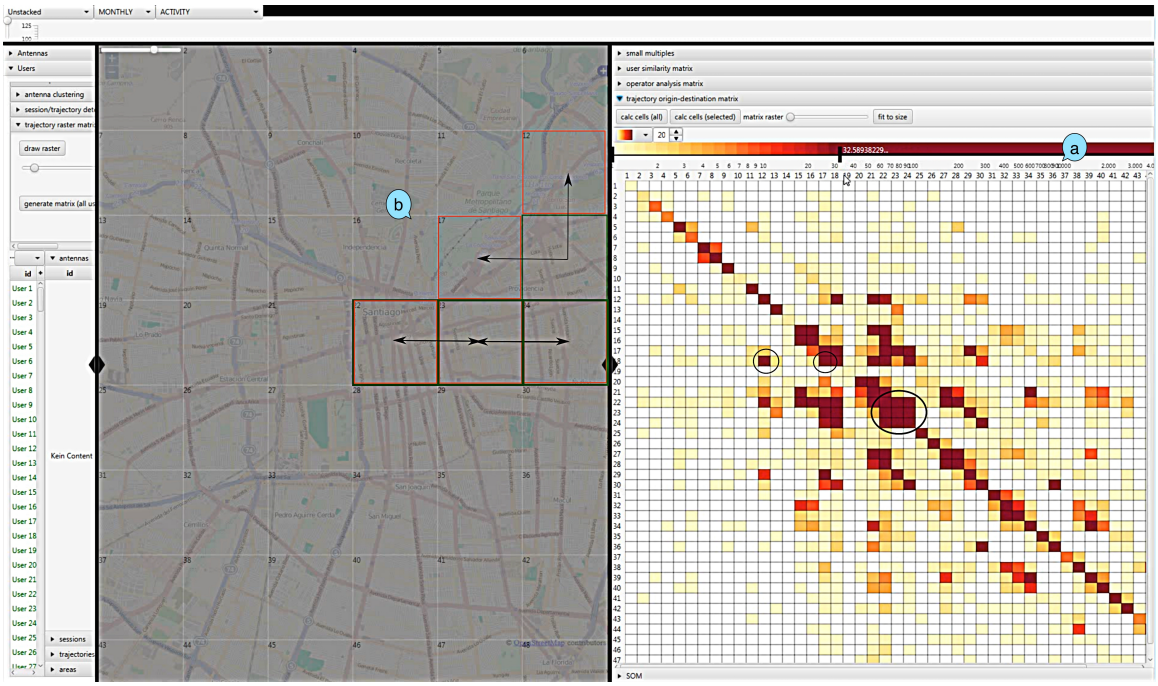


Figure 5.9: Origin-destination analysis of aggregated user trajectories. (a) Each row and column represents an origin and destination respectively for the aggregated trajectories. The colour scale represents low (lighter) to high (darker) no.of aggregated trajectories, (b) The selected cells in the matrix are coordinated with the rasterized map, as shown in the green (origin) and red (destination) spatial segments. This figure appears in Senaratne et al. (2017b).

Characteristics of the surrounding geography that may impact such movement patterns are the disparities that one can observe within various modalities such as economic, health, sanitation etc. One aspect that explains the observed movement patterns within Santiago in Figure 5.9 is the economic division in the city (Agostini and Brown (2007)), which according to the communal human development index for Santiago shows a higher concentration of high standard public and private facilities in the central and Eastern parts of Santiago, than in the Southern and Western parts of Santiago. Agostini and Brown (2007) shows in their study that the Southern counties in Chile have the highest inequality in terms of income.

Detecting such disparities most often involve surveys and census studies in the state of the art research. The presented visual analytics approach helps the user to interactively explore such disparities as geographic effects on human movement behaviour.

5.3.5 Spatial Area Analysis based on Temporal Usage Patterns

Spatial area analysis can help urban planners, e.g., to observe various attributes that influence the socio-economic factors in different areas of a city. For this purpose we analyse the aggregated Internet usage data for a selected antenna over a given period of time. For this analysis we use the raster map view, a small multiples view, and a cluster layout obtained by applying the Self Organizing Map (SOM) algorithm (Kohonen, 1998; Andrienko et al., 2010). The raster map approach, similar to spatial clustering, facilitates the user to analyse areas rather than single antennas. Instead of a cluster hull to differentiate the areas, we use here a raster with a user-defined cell size. The cells are then coloured according to the aggregated number of records transmitted through the antenna, which indicates the overall Internet usage from the antenna for that time period.

Figure 5.10 shows an example of the method in the center of Santiago. In the main map view (component (c) in Figure 5.10), the aggregated data for the whole time period is considered for the calculation of the values for each cell. The daily usage patterns are visualised in the small multiples view (component (d) in Figure 5.10), allowing the user to compare patterns of different time periods (in addition to daily patterns, the user can also switch to weekly or monthly patterns). Evidently, the daily usage patterns indicate that the most number of usage records are distributed within the city center to the East of Santiago. These results once again resonate with the findings in Section 5.3.4, and therefore as a possible explanation can attribute to the economic disparity in the city.

For longer time intervals, the amount of small multiple instances could lead to overcrowded displays. Therefore, we cluster the small multiple instances such that similar patterns are grouped and the user can focus on the analysis of the group instead of each pattern. Again we use the SOM algorithm for the cluster layout by using the daily usage patterns as the input variable. This is shown as component (f) in Figure 5.11. This additionally arranges the results in a grid, so that similar clusters are closer to each other. The time of occurrence of a selected pattern from the SOM

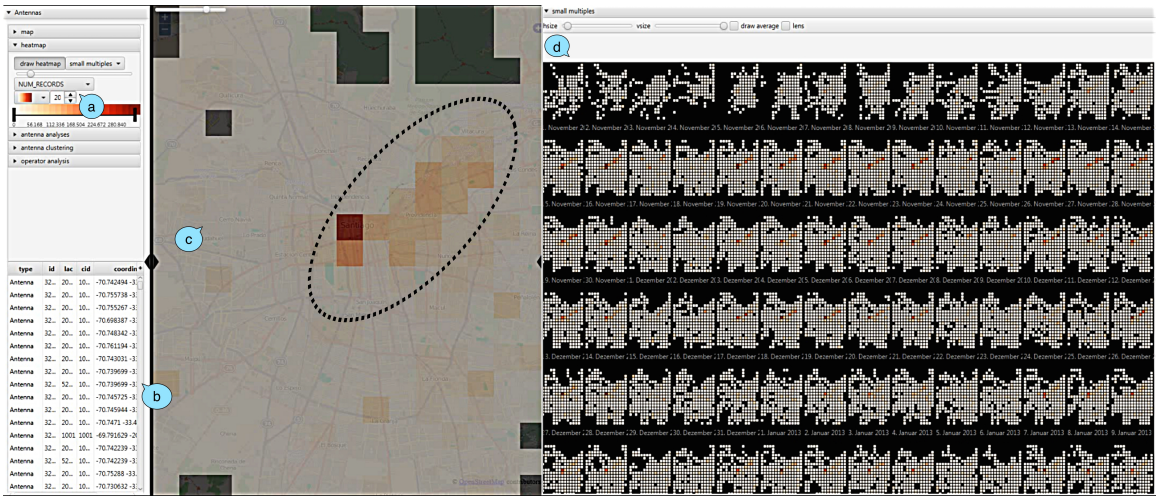


Figure 5.10: Component (a) selects the number of records as the attribute for analysis, (b) chooses an antenna from the table for analysis, and (c) shows a raster map of Santiago created with a user-defined cell size, where each cell indicates the aggregated number of records for the entire duration of time with a heat map visualisation. The colour scale can be interactively adjusted to the users' preference. (d) shows the small multiples view for the daily usage patterns of mobile Internet for the chosen antenna. This figure appears in Senaratne et al. (2017b).

view is indicated in Figure 5.11 as component (e). This further depicts the assumed correlation.

5.4 Uncertainty in Movement

When it comes to the antenna location-based movement data that is used for movement analysis in this chapter, the unavailability of antenna signals, for example due to indoor movements (such as moving through a tunnel, or a building), or the large operational range of an antenna where the user could be anywhere within this area, cause positional uncertainty of movement approximations.

Furthermore, as seen in the algorithm 1 the movement *approximations* that are derived in this chapter consider an interpolations of *crisp locations* in the form of chronologically ordered points, $p = (x, y, t)$. In our analysis we drew a straight line from one crisp point to another, assuming the user took a direct route from one position to another. However, in reality this may not be true, and further gives rise to positional uncertainty.

To model these positional uncertainties within the movement data in our visual analytics approach we utilise the *space-time cube*. The space-time cube which originates from the work of Hägerstraand (1970), is an interactive visualisation technique that

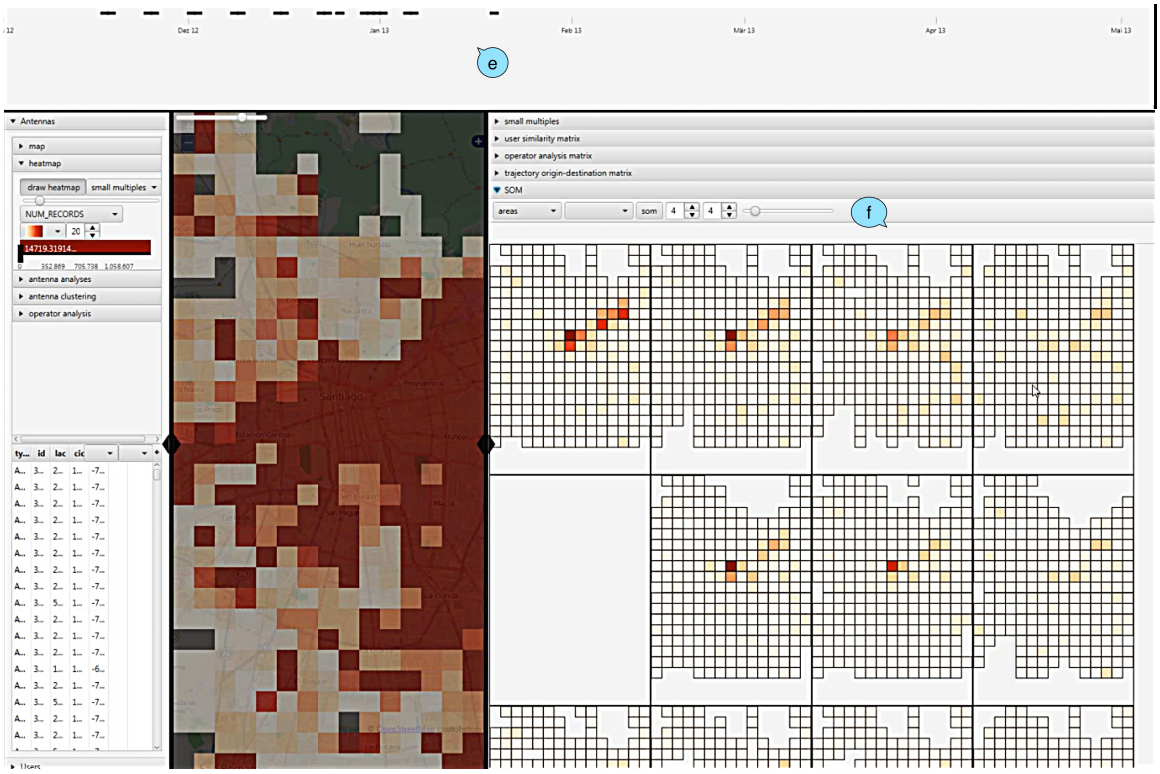


Figure 5.11: Component (e) shows the time of occurrence of chosen patterns of SOM, and (f) shows the SOM view with a 4 x 4 grid. The daily patterns are used as the input to the SOM algorithm. The clusters depict 16 similar daily patterns.

incorporates a three dimensional cube for representing data on a horizontal plane and a vertical plane. We use this technique for visualising the movement trajectories where the horizontal plane is used for depicting the spatial extent (x, y) of the data and the vertical plane is used for depicting the temporal extent (t) of the data. These time-ordered sequences of samples called *markers* make up the *space-time path* for each of the users of the mobile Internet usage dataset. An example of the space-time path of a selected user within a 24 hour time frame is shown in Figure 5.12. As one can observe from this figure, the selected user is somewhat stationary during the morning and evening hours, but in between spends a longer duration of time also in the northern commune of Santiago.

Figure 5.13 shows the inter-sample time interval for the same space-time path. However, the data was collected at a 10-second sampling rate. Therefore ideally the inter-sample time duration should be 10 seconds at each path segment. This variation of the time interval duration can be due to a lack of antenna signal reception, or technical issues with the mobile phone devices. It therefore indicates already the uncertainties in the data.

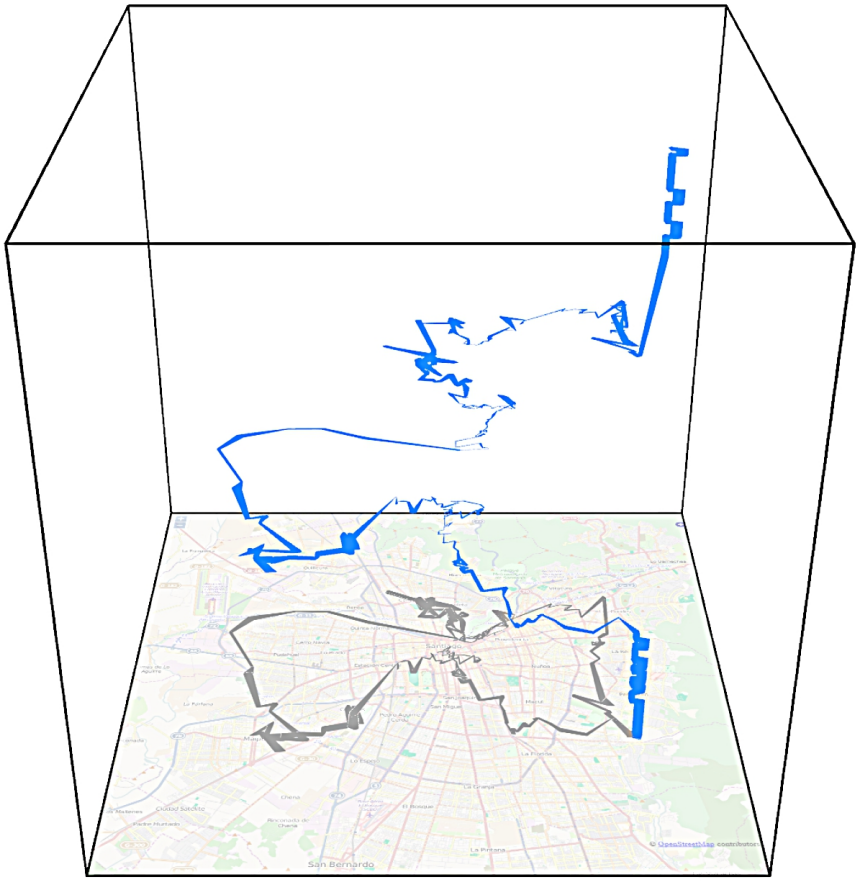


Figure 5.12: The space-time path of a selected user within a 24 hour time frame visualised using the space-time cube. This figure appears in Senaratne et al. (2017b).

5.4.1 Space-time Prisms to Analyse Uncertain Movement Path Segments

The uncertainty of the movement data as described above can be represented as *volume* in the space-time cube. This volume represents all the space-time points within an area that any user *may* have traversed based on any given marker and a maximum travel velocity. Therefore the volume between two given markers is called the *space-time prism*. It's projection onto the horizontal geographic plane is called the *Potential Path Area* (PPA). The work of Miller (2005) demonstrated the use of these space-time prisms for crisp markers. The works of Kuijpers et al. (2010) and Kuijpers and Othman (2009) modeled the uncertainty of locations in between crisp markers on a road network.

Due to the inherent positional uncertainties in our derived movement data, we have *regions* instead of crisp markers. Those region-based markers denote several possible markers of user movement; and we call those here *uncertain markers*. Therefore in

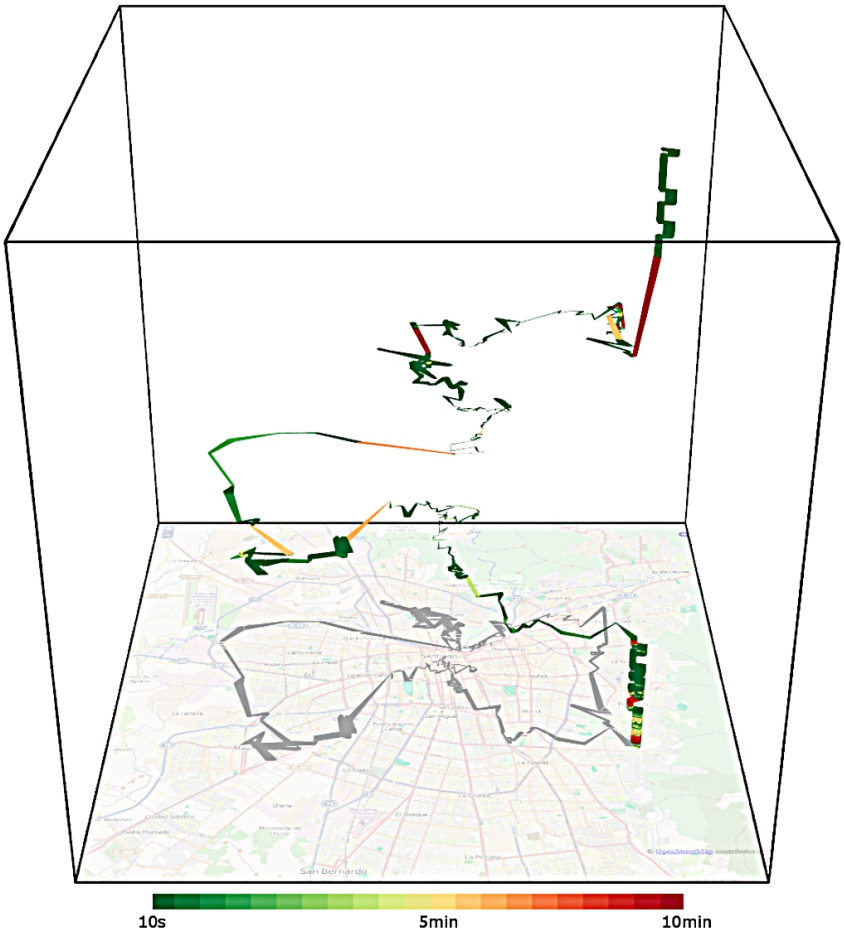


Figure 5.13: The space-time path of a selected user with the path segments coloured according to the inter-sample time interval duration. This figure appears in Senaratne et al. (2017b).

our work we define and adapt the space-time prisms for uncertain markers to visually analyse the positional uncertainties in the approximated movement data.

Based on the work of Miller (2005), a space-time prism Z between two crisp markers $m_i = (p_i, t_i)$ and $m_j = (p_j, t_j)$, and a maximum velocity of $v_max > \frac{\text{u}p_i - p_j\text{u}}{t_j - t_i}$ can be formulated as a parametric function of time. This is shown below.

$$Z_ij(t) = \begin{cases} SE_i(t) & \text{if } t \in (t_i, t'] \\ SE_i(t) \cap SE_j(t) & \text{if } t \in [t', t''] \\ SE_j(t) & \text{if } t \in [t'', t_j) \end{cases}$$

$SE_i(t)$ is the spatial extent that can be reached from point p_i in the time interval $t - t_i$ with a maximum velocity v_max. This spatial extent can be defined as a constrained set of points as follows:

$$SE_i(t) = p\text{ıı}p - p_i\text{ı} \leq (t - t_i)v_max$$

Similarly, $SE_j(t)$ is the spatial extent that can be reached from point p_j in the time interval $t_j - t$ with a maximum velocity v_max. This spatial extent can be defined as a constrained set of points as follows:

$$SE_ij(t) = p\text{ıı}p_j - p\text{ı} \leq (t_j - t)v_max$$

$SE_i(t)$ and $SE_j(t)$ have two possible topological relations where t' and t'' are time points between t_i and t_j. These possibilities are, when at t' $SE_j(t)$ *contains* $SE_i(t)$ and at t'' $SE_i(t)$ contains $SE_j(t)$.

Further, t' and t'' are calculated as follows:

$$t' = \frac{t_i + t_j - t_i^*j}{2} t'' = \frac{t_i + t_j + t_i^*j}{2} t_i^*j = \text{ıı}p_i - p_j\text{ı} * v^-1_max$$

Given the velocity v_max, t_i^*j is the minimum travel time from p_i to p_j. This parametric function is used for the space-time prism between uncertain markers as well, however the key difference is in the way we define the spatial extents $SE_i(t)$ and $SE_j(t)$, and the maximum travel time t_i^*j. Therefore we determine the spatial extents for the uncertain markers by using the hull at the extreme points of the regions. This is exemplified in Figure 5.14. For convex regions we use the convex hull.

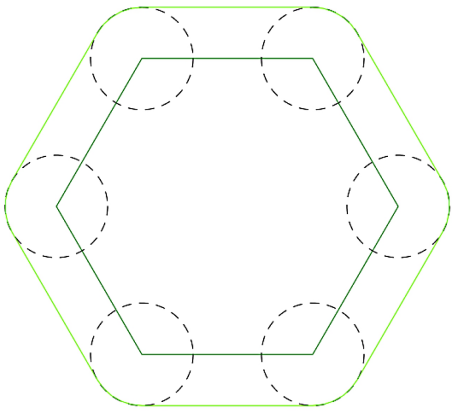


Figure 5.14: Region expansion for defining the spatial extents within uncertain markers. A hexagon is taken as an example region which is indicated in dark green. At each edge of the hexagonal region the spatial extents are indicated with the dashed circles. The hull is indicated with the light green polygon which therefore is the spatial extent for uncertain markers at time t.

Therefore, the spatial extents of two uncertain markers $m_i = (region_i, t_i)$ and $m_j = (region_j, t_j)$ can be represented as follows.

$$SE_i(t) = pıp \in CER(region_i, t - t_i) SE_j(t) = pıp \in CER(region_j, t_j - t)$$

$CER(region, t)$ is the expanded region at time t and a given v_max. For defining the maximum travel time for uncertain markers, we divide the distance between the points by the assumed speed. Therefore this is expressed as follows:

$$t_i^* j = \frac{maxDist(region_i, region_j)}{v^-1_max}$$

The Figure 5.15 shows examples of potential path areas and their corresponding space-time prisms for uncertain markers.

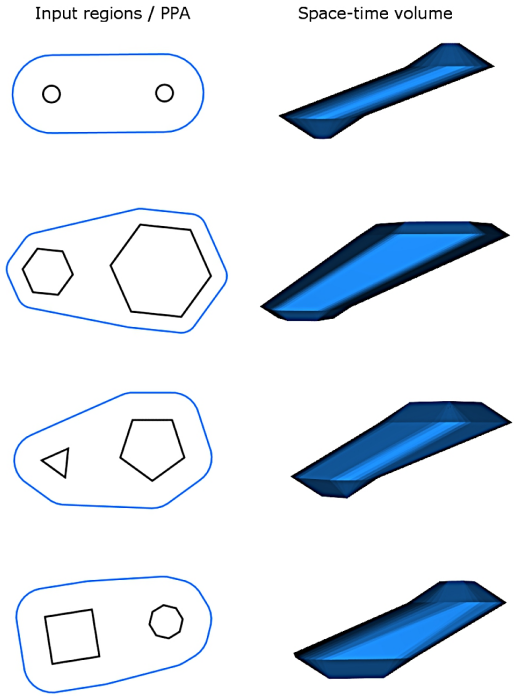


Figure 5.15: Potential path area for uncertain markers and their corresponding space-time prism representations.

For the space-time path example shown in Figure 5.13, we have now estimated the potential path area and the corresponding space-time prism for two uncertain path segments. This is show in Figure 5.16.

5.4.2 Positional Uncertainty Reduction in Mobile Communication-based Movement Data

As important as it is to acknowledge the underlying uncertainties in data, it is also imperative to *reduce uncertainties* with the right tools and methods in place. Luck

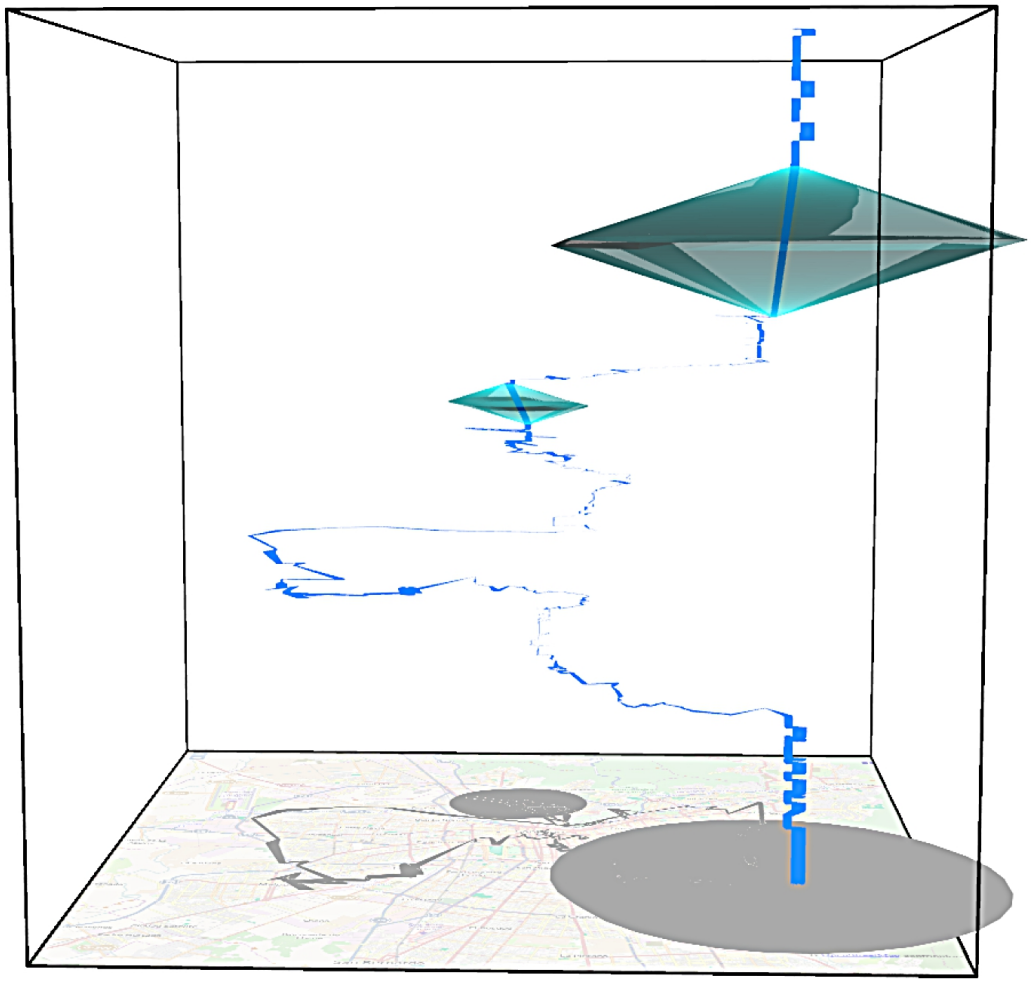


Figure 5.16: Potential path area for two uncertain path segments and their corresponding space-time prism representations. The maximum velocity is assumed to be 50 km/h in this example. This figure appears in Senaratne et al. (2017b).

et al. (1996) found in their study that reducing uncertainty by reducing the number of noise sources increased the accuracy of target locations. In the mobile Internet usage dataset that we have used for movement analysis and exploration here, the location of users is determined at the highest resolution by the location of the antennas. However, it is very common for the mobile receptors to receive signal (or *jump*) from one antenna to another when the mobile is used in an area where the operational range of two antennas overlap. This causes positional uncertainty of users. By detecting the areas where such antenna jumps occur already paves path for uncertainty reduction in the data, and thereby reduce the uncertain space size for user locations. To achieve this we propose a two-tiered approach: (1) identify the area segments of a space-time path where the antenna jump occurs, (2) interpolate the jumped area segment candidate

with the preceding region of the space-time path. Replace these area segments with the newly interpolated area segments.

The algorithm is as follows:

ALGORITHM 3: Antenna jump detection and interpolation

Data: Time-ordered sequence *Input* of sample data Threshold for maximum stay duration *maxDuration*

Result: Time-ordered sequence *Output* of sample data with interpolated antenna jumps

Extract time-ordered (*region, samples, duration*) sequence *Stays* from *Input*;

for $i \leftarrow 1$ **to** $|Stays| - 1$ **do**
    if $Stays[i-1].region == Stays[i+1].region \wedge Stays[i].duration \leq maxDuration$ **then**
        Output.add(interpolate(Stays[i-1].region,Stays[i].region,0.5),
        (Stays[i].firstSampleTime + Stays[i].lastSampleTime/2);
    end
    else Output.addAll(Stays[i].samples);
    ;
end

A sliding window (size of 3) is used to slide over the path sequences, and at each step the area of the first and last position of the window are checked if they are equal. If they happen to be equal, then the duration of stay in the middle of the window is checked if it is less than or equal to the user specified maximum duration. If these two conditions hold, then the middle of the window is considered as an antenna jump location. In the second step a linear interpolation with a parameter value of 0.5 is carried out from the preceding region to the next, replacing the previously identified antenna jump location. The 0.5 parameter value corresponds to a region in-between the input region and the immediate successive region. Figure 5.17 depicts this approach for four different *maxDuration* thresholds. Too low and high thresholds will give higher false-negative classifications and false-positive classifications respectively.

5.5 Discussion & Future Work

In this chapter several algorithms are presented together with various visual analysis approaches for (1) movement trajectory extraction out of mobile telecommunication data, (2) identification of patterns in urban environments based on the movement trajectories of users, (3) investigating the positional uncertainties in movement trajectories, and (4) reducing uncertainties in movement trajectories extracted from such mobile telecommunication data.

Our findings obtained from our prototype on the data set from Santiago is indicative for a number of analysis cases of interest to mobile data analysis. While our findings

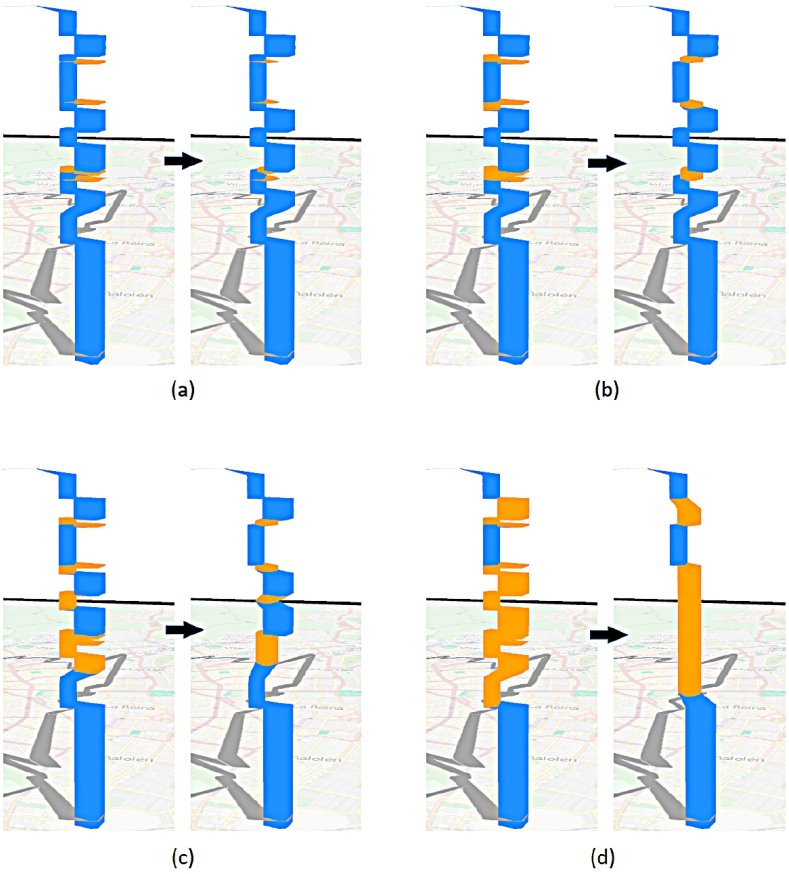


Figure 5.17: Positional uncertainty reduction of space-time path segments of users through (left) identifying the antenna jump locations and (right)the interpolation of antenna jump locations. The maximum time duration is (a) 3 minutes, (b) 6 minutes, (c) 15 minutes, (d) 20 minutes. This figure appears in Senaratne et al. (2017b).

can be regarded plausible, we state that they are based on a sample of several hundred users from a much larger user basis, and may incur a bias. Finer-grained analysis of user trajectories requires more user data and ideally equally distributed over the whole region of the metropolitan area.

Our analysis considers aggregates of mobile data consumed. More detailed analysis would be possible by enhancing the data collection mechanisms, e.g., distinguishing between data used for different applications such as email, web, voice, or video. The correlation of service usage with temporal and spatial aspects might lead to further insights of usage with great potential for applications such as transportation optimisation or marketing.

An important aspect in any visual interactive analysis scenario is to guide the analyst to quickly find interesting and relevant views. Due to the large data volumes in mobile data analysis, a fully manual search through all possible data sections and views

is practically not affordable. Hence, methods for automatic identification of interesting views are needed. An idea for future work is to interactively learn from the user which situations are interesting, and use such a learning scheme to suggest relevant situations in unseen data. To this end, approaches for visual relevance feedback (Behrisch et al., 2014) could be helpful.

5.6 Conclusions

This chapter showcases how geospatial movement patterns, as extracted before (see Chapter 4), help greatly to identify human bahavioural patterns. Those derived human bahavioural patterns can give us clues about the *arrangements* of the surrounding urban environments. To explore and identify these patterns this chapter utilises a GSM telecommunication dataset where the consecutive locations for movements are extracted from the antenna locations.

The encompassing visual analytics approach first introduces a matrix visualisation to interactively explore the spatial and temporal similarity of movement between users. The movement patterns are further employed to classify the surrounding geographic space into industrial and residential areas based on the movement of users between home and work. Next, utilising the same matrix view we have computed the origin and the destination of users' movement, thereby identifying the local spatial segregation patterns in the surrounding urban environment. Using a raster map view, we observed temporal usage patterns of mobile Internet, which further clarified the previously observed spatial segregation within the surrounding urban environment. In the last section of this chapter we define and adapt the space-time prism for *uncertain markers* as an extension to the space-time-cube method to assess and visualise the uncertainty of the approximated user movement patterns. This is followed by an approach to reduce the positional uncertainties in the user movement data, derived from GSM antennas.

Several interesting questions for future work occur. For example, the developed approaches could be extended towards embedding an interactive learning within the developed tools. Great potential lays in designing innovative applications based on the presented approaches, e.g., in the direction of user profiling for marketing, improving network management, optimising public transportation, or supporting urban planning in general.

Chapter 6

Uncertainty Analysis of Bi-dimensional Numerical Data

Contents

6.1 Background & Related Work 152
6.2 Numerical Data from a Smart Grid Network 155
6.3 Applying Monte Carlo Simulation and Sampling Method for Uncertainty Assessment 157
6.4 Glyph Design for Bi-dimensional Uncertainty Analysis in Smart Grid Environments 159
6.5 Performance and Preference of Glyph Designs 162
6.5.1 Design of the Usability Study 164
6.5.2 Study Results 169
6.6 Discussion & Future Work 173
6.7 Conclusions 175

Numerical data is collected in the form of numbers as opposed to the image-, text-, and location-based data, which have been analysed in the previous chapters. Numerical data is utilised in abundance for various analytical tasks due to the elemental ways of collecting such data, and therefore many methods have also been practiced to assess the uncertainty of numerical data. However, the challenge remains that many uncertainty assessment methods such as simulation based methods that we encountered in Chapter 2 are chosen out of context of the data. In case of visual analytics approaches, this can be avoided by enabling the user to intuitively choose appropriate uncertainty assessment methods. The next challenge is, when we have two dimensions of uncertainty to communicate to the user. As opposed to one dimensional uncertainties that we saw in the previous chapters, we need appropriate visualisation designs to communicate both types of uncertainties efficiently to the user, at the same

time. Vigorous evaluations of combined visual metaphors for the context at hand can alleviate design failures.

To address both of these challenges, this chapter presents a visual analytics approach that enables the comparison of two uncertainty assessment methods and the selection of glyph designs to represent bi-dimensional uncertainties. Based on the guidelines of Chapter 2, the sampling and the Monte Carlo simulation techniques are chosen here to assess the uncertainties within the numerical dataset, and their performances are compared to select the most appropriate method. Visualisations are then designed to communicate these quantified uncertainties in a visual analytics environment. Both are implemented as a tool for a smart grid environment that aids users to interactively choose *uncertainty assessment methods* as well as *pre-selected visual metaphors* to visually analyse bi-dimensional uncertainties. By conducting first a pilot study and then a larger user study, these visualisation designs are evaluated.

This chapter unfolds as follows: In Section 6.1 the related works on theories, and limitations of visualisation designs for uncertainty analysis are discussed. Section 6.2 gives a brief overview of the numerical data that are used in this chapter. Section 6.3 presents how two uncertainty assessment methods are compared for choosing a best-fit method for the data and task at hand. Section 6.4 presents the glyph design that incorporates various visual metaphors in combination to represent bi-dimensional uncertainties. These alternative visualisations are evaluated within a user study. The step-by-step approach of the user study as well as its results are presented in Section 6.5.

The contents of this chapter are partially based on the publication Senaratne et al. (2014b)[1].

6.1 Background & Related Work

With Bertin (1983)'s introduction to the *visual variables*, uncertainty visualisation gained significance to present various uncertain elements in data through the variables *size*, *value*, *colour*, *shape*, *orientation*, *location*, and *texture*. Based on these grounding principles many state of the art visualisation methods to represent spatio-temporal uncertainty have been developed and tested on their usability in different settings.

[1]This work is a result of a collaboration with S. Mittelstaedt, C. Jacob, and T. Schreck, all affiliated to the University of Konstanz during this work. My contributions as the first author were guiding the comparison between the uncertainty assessment methods, collaboratively designing the glyph visualisations, defining the step-by-step approach for the user study, and overseeing the implementations.

Uncertainty visualisation presents quantified uncertainty elements of data in a visual context. Extending the work by Bertin (1983) on visual variables, Morrison (1974) took a slightly modified approach and introduced *colour saturation* which was later applied to depict spatial-temporal uncertainty of data. In addition to Bertin's variables, MacEachren (1992) studied *focus* as an additional variable: higher uncertainty is depicted out of focus and lower uncertainties in focus. Furthermore, MacEachren (1992) manipulated the focus variable into four metaphors; *contour crispness* - varying the crispness or fuzziness of the edges of the symbols used, *fill clarity* – manipulating fill clarity to depict high and low uncertainty, *fog* – using transparency to depict uncertainty, *resolution* – adjusting the resolution of geographic detail to convey varying uncertainty. These visual variables are presented in Figure 6.1.

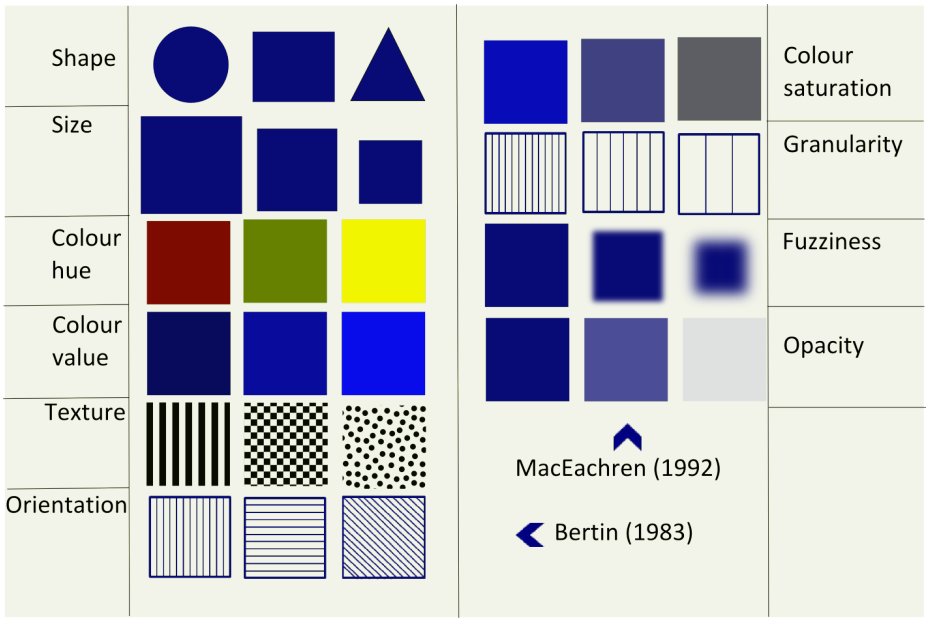


Figure 6.1: The visual variables introduced by Bertin (1983) and MacEachren (1992).

Many theories have evolved regarding the applications and the cognitive limitations of visualisations created through these visual variables.

Tufte and Graves-Morris (1983) in their contributions to graphical excellence have provided six principles that can be followed in order to promote graphical integrity. These are: (1) graphics representing numbers need to be directly proportional to the quantities represented, (2) clear and detailed text should be used wherever needed to avoid ambiguity, (3) show data variation and not design variation, (4) the depiction of money development in time series must be adjusted to inflation, (5) number of dimensions used to read data should not exceed the number of data dimensions that

are being presented, (6) data should not be shown out of context. They have further suggested that following these principles will help to overcome misinterpreting the representations of reality.

Ware (2000) grounded his research on visual perception and comprehension in terms of physiological, perceptual, and cognitive psychological theories. Following the *Gestalt Laws* (Koffka, 1935), Ware (2000) explained that properties such as proximity, similarity, continuity, symmetry, closure, connectedness, and relative size substantially influence the perception of patterns, and that they can be used as basic design principles to create visualisations. Extending Tufte's (Tufte and Graves-Morris, 1983) graphical excellence principle of integrating text descriptions with a graphic, Ware (2000) further emphasised that text might prevail over images when it comes to presenting abstract ideas, logic, and conditional information.

Adding to that, Agrawala et al. (2011) presented a set of design principles for visual communication, in which they foremost described the importance of human perception, cognition and communicative intent of visualisations in order to create meaningful designs. These principles are to be used to either emphasise important information or de-emphasise irrelevant information. They pointed out to an example of an early implementation of these principles; the London underground map from 1933 by Harry Beck. In this the curved subway lines are straightened and the stops are evenly spaced in order to simply convey the sequence of stops to the viewer, allowing the human cognitive system to perceive the underlying information. They developed a three stage approach for creating meaningful visualisation design systems. The first stage is to identify the domain specific design principles. Their aim is to create visualisations that cater to different information domains (e.g., cartographic visualisation and technical illustrations) based on perception and cognition of these visualisations. The second stage is to implement these design principles by encoding them into algorithms and interfaces, thereby creating the visualisation. The third stage involves the evaluation of these design principles by measuring their usefulness. This is achieved by getting qualitative user feedback, quantitative usage statistics, and conducting formal usability studies. These procedures help to assess how well the design principles serve the purposes of information processing, communication, and decision making.

Emphasising the importance of interactive and exploratory capabilities of such visualisation systems, Andrienko et al. (2008b) discussed three approaches to analytical and exploratory visualisation. These are: (1) direct depiction: user identifies the patterns of measured phenomena through interacting with the visualisation that

directly depict the measurements recorded in their data, (2) summarisation: user identifies the patterns of measured phenomena through visualisations of data summaries that are computed from the original data measurements, (3) pattern extraction: the computer identifies data patterns or summaries and represent these visually to be interpreted by the user.

Statistical propagation and visualisation of uncertainty in spatio-temporal data have long been studied in various works such as Ehlschlaeger et al. (1997); van de Kassteele and Velders (2006); Verstegen et al. (2012); Pang (2001); Hengl (2003); Kardos et al. (2006). These visualisations have been evaluated in previous works such as Evans (1997); MacEachren et al. (2012); Kardos et al. (2004); Cliburn et al. (2002). Lee and Chen (2009) examined several widely used uncertainty propagation techniques in order to understand the characteristics and limitation of these methods, and further compare heir performances.

Also, previous works by the author of this thesis have contributed to the evaluation of visualisation techniques for *uncertainty* in general spatio-temporal data; see Senaratne and Gerharz (2011); Senaratne et al. (2012); Reusser et al. (2012). These works served as a conceptual basis towards developing this thesis.

6.2 Numerical Data from a Smart Grid Network

The numerical data used as an example case in this chapter consist of two years of measurement data from selected transformer stations within a smart grid network. Each of these transformer stations contain several modules that collect the measurements of power in watt, voltage, and electric current in ampere in a 15 minute interval. Within the work presented in this chapter, only the power data is taken as an example for numerical data. The distribution of the transformer stations in Germany is shown in Figure 6.2.

Electric power grids are the backbone of our society, since failures in the electricity supply have a strong impact on the fundamental societal structures such as life/health, environment, or economy. The rise of renewable energy of small energy producers, e.g. photo-voltaic plants, increase the system's complexity. To integrate these producers and to transfer their energy to other regions communication infrastructures and energy infrastructures are tightly coupled, which increases the effectiveness, however, also increases the vulnerability, since failures in one infrastructure can cascade into the other. Visualisation systems are needed that abstract the complex information of both infrastructures in case of a crisis to enable crisis response by decision makers.

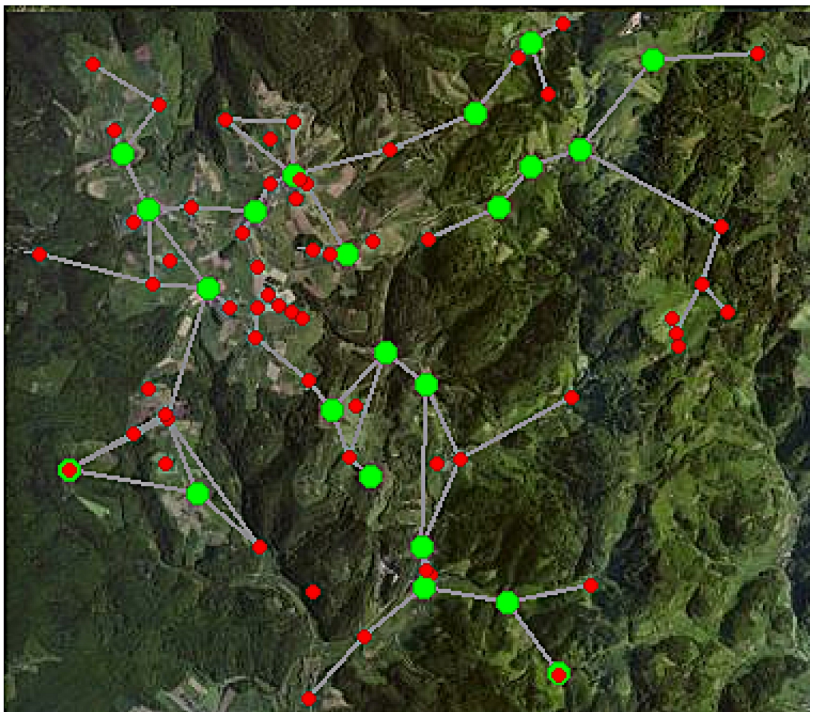


Figure 6.2: The smart grid network of transformer stations in Germany. Red nodes indicate the transformer stations where measurement data is unavailable, and the green nodes indicate the transformer stations where the measurement data is available.

In Mittelstädt et al. (2013) the authors abstract the incoming data from each infrastructure element and apply a set of rules to map this information to a colour encoded scale that highlights which elements are in normal, danger, or alarm mode. An example of this encoding is shown in Figure 6.3. This mapping is consistent over all infrastructures and thus, allow interdisciplinary teams to "perceive" and "understand" a crisis situation. Further, the system predicts based on the past data, the currents status, and the users' actions, a possible future subsequent development of the situation and of all infrastructure elements. This allows the evaluation of alternative actions and therefore supports the crisis managers in the decision making process. The decider will draw a decision based on the visualisation of the current and future state (alarm level) of infrastructures and the detail information of elements of interest based on the propagated subsequent development of alternative actions. However, such analysis systems are error-prone. Errors propagated by the measurement modules or the discrepancy between simulation models and reality reduce the confidence of decision makers for such systems in general. Her/his trust into measurements and predictions is of major importance and thus, such systems must highlight how uncertain or certain some predictions or measurements are.

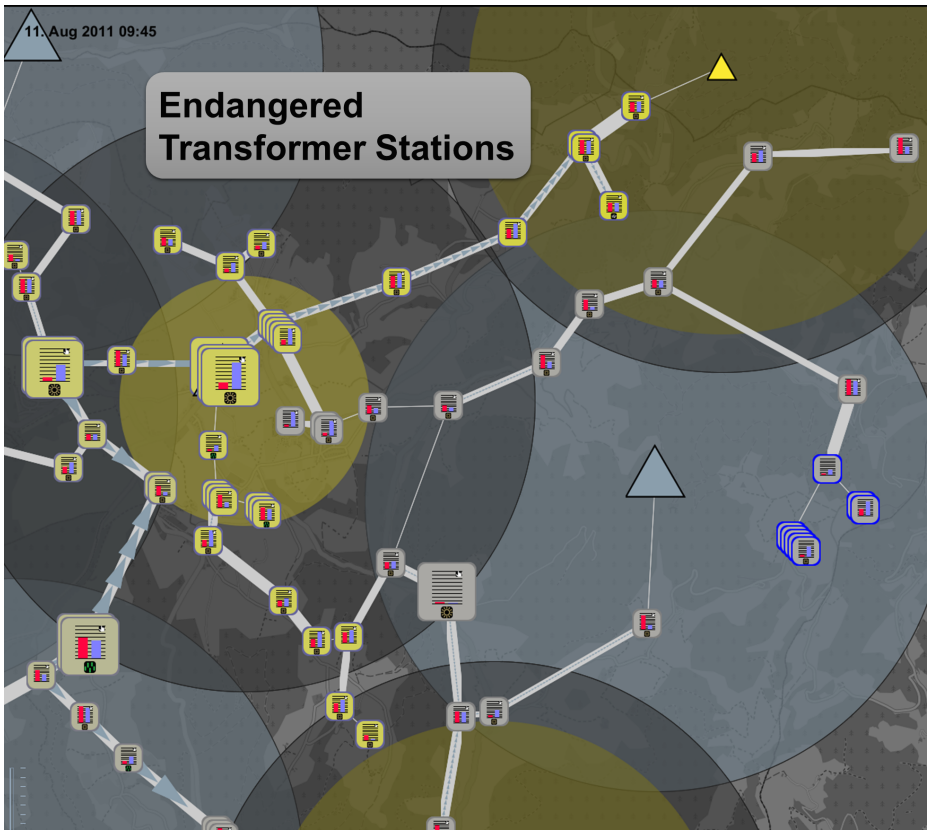


Figure 6.3: Transformer stations (rectangles) are connected via power lines and are also connected to the communication infrastructure (triangles), which transfers the information to the central control room. The transmission range of the mobile stations is visualised as concentric circles. While gray indicates normal operation mode, the yellow elements on the screen reveal a severe situation. High deviations in voltage cascaded from the energy grid into the mobile grid due to failures of the power supply. This figure is taken from Mittelstädt et al. (2013).

6.3 Applying Monte Carlo Simulation and Sampling Method for Uncertainty Assessment

For assessing the uncertainty in the numerical dataset, two methods are compared with each other in order to choose the best suitable method for the given dataset. The two methods are: (1) Sampling, and (2) Monte Carlo Simulation (MCS).

The sampling method (Helton et al., 2006) typically takes the distribution of a selected subset of the data and estimates the characteristics of the whole dataset. Within our smart grid dataset we incorporated this method and further constructed a 95% confidence interval with which we demonstrate the uncertainty in the modular power data.

The Monte Carlo simulation technique (Mooney, 1997) models the statistical errors

in the data by the use of ordinary statistics and random variables, assuming that the errors have a Gaussian probability distribution function. Continuous repetition of the simulation removes the variations in the probability distribution function. The uncertainty in the data is therefore propagated by the mean error and the standard deviation for each data point. Once again with a 95% confidence interval we demonstrate the uncertainty in the modular power data.

We carried out the above two methods for two random weeks of the data that are available for three transformer stations within the smart grid network. For the comparison of the two methods we counted how many data points are within and outside of the 95% confidence interval. The method that shows more data points within the 95% confidence interval is considered more appropriate for our dataset. Table 6.1 shows the comparison results for the three transformer stations, and Figure 6.4 shows the uncertainty for a selected transformer station in Helgenreute (Germany) assessed through sampling and MCS methods. This comparison shows us that the Monte Carlo simulation method works better for the dataset.

The aggregated uncertainty for the alarm levels is estimated by the classifier that maps the incoming field information, the detected anomalies, and expected behavior of an element to discrete alarm levels. The distance to the decision boundary indicates how sure the classifier is in the assignment, e.g., the element was assigned to "danger" class but it was also close to the boundary "normal" and thus, the classifier is more uncertain.

Table 6.1: Comparison results of Sampling and Monte Carlo Simulation (MCS) methods for the three transformer stations.

Week	**Sampling**	**MCS**	**Transformer station**
1st week	in:514 out158	in: 566 out: 106	Helgenreute
	in:413 out: 259	in:505 out: 167	Rathaus
	in: 492 out: 180	in: 643 out: 29	Eckle
2nd week	in: 1230 out: 786	in: 1513 out: 503	All the transformer tations

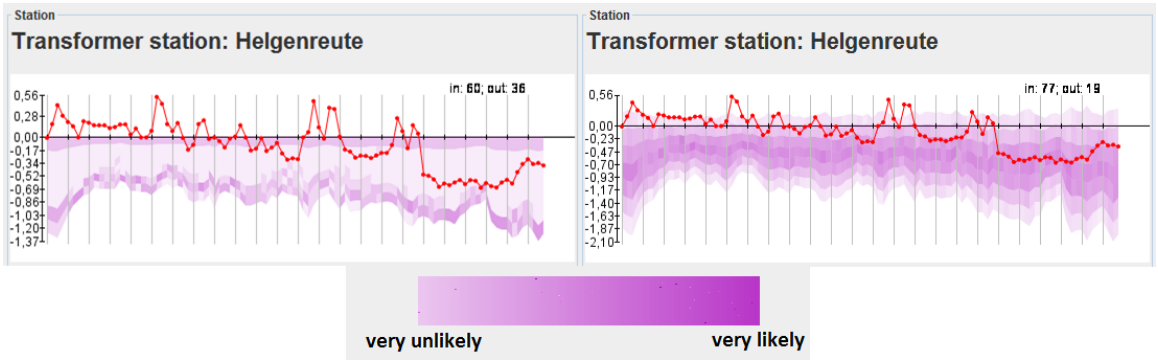


Figure 6.4: Sampling method (left) and MCS method (right) to assess power uncertainty at the Helgenreute transformer station. Low to high Purple colour saturation indicates the high and low uncertainties. This figure appeared in Senaratne et al. (2014b).

6.4 Glyph Design for Bi-dimensional Uncertainty Analysis in Smart Grid Environments

Glyphs have become most popular among extrinsic visualisations due to their multivariate nature, and are utilised to represent variables through a multitude of parameters such as location, shape, size, colour, orientation, aspect ratio, or curvature (Borgo et al., 2013). Works by Pang (2001) and Cliburn et al. (2002) have demonstrated the use of glyphs for uncertainty visualisation in geo-spatial data under various settings. Related to the study by MacEachren et al. (2012) and based on the findings in Senaratne and Gerharz (2011), visual metaphors such as filling, colour transparency, colour saturation, noise lines, fuzzy borders, border colour hue, and icons can be chosen to design the glyphs to depict the two kinds of uncertainty: (1) modular power uncertainty and (2) aggregated power uncertainty. These candidate glyphs were designed to stay consistent with the existing visualisation infrastructure as seen in Figure 6.8. The above visual metaphors were randomly assigned (to avoid pre-selection bias) to depict the two kinds of uncertainty within a glyph. The glyph candidates for modular power uncertainty visualisation are shown in Figure 6.5 and the aggregated power uncertainty visualisation candidates are shown in Figure 6.6.

Considering the smart grid monitoring requirements and with a goal of reducing the solution space, we pick the most effective visualisations through assessing the usability of the visualisation candidates. As a first step we conduct a pilot study to assess the feasibility of the visualisation candidates through a usability study for the smart grid environment. For this pilot study, 4 participants were chosen who had a background in visual analytics and possessed knowledge regarding the smart grid environment, which is in focus within this chapter. Throughout the study, we used

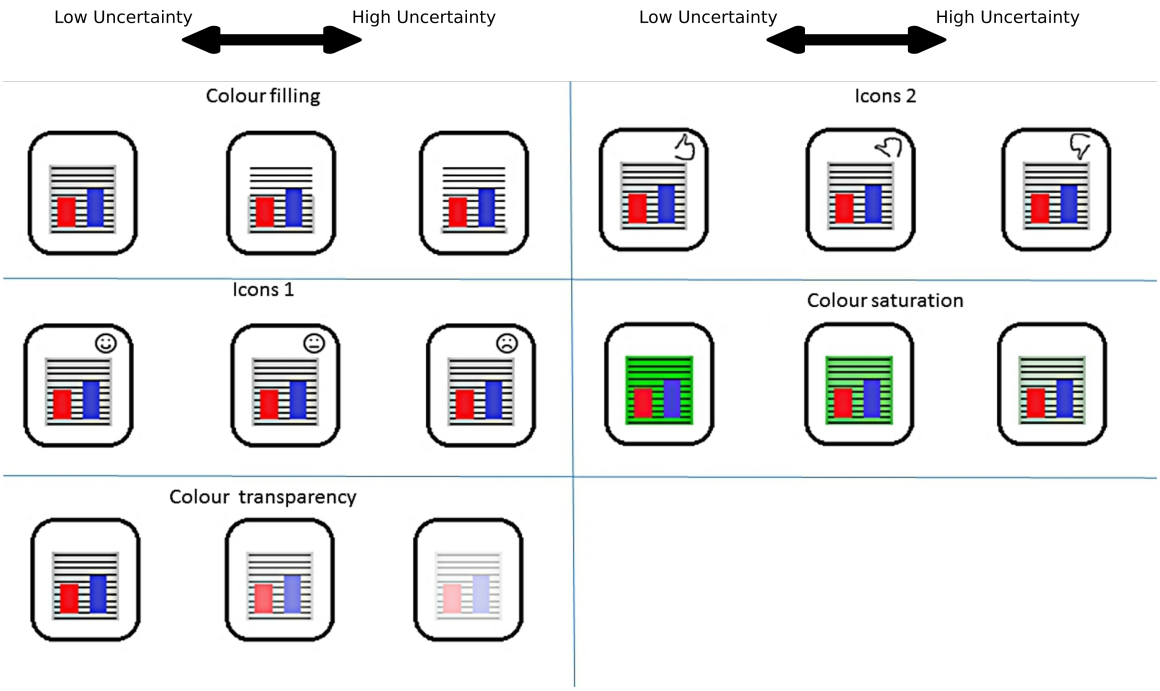


Figure 6.5: Candidates for modular power uncertainty visualisation. The inner rectangle of the glyph indicates the modular uncertainty. Uncertainty increases from left to right. This figure appeared in Senaratne et al. (2014b).

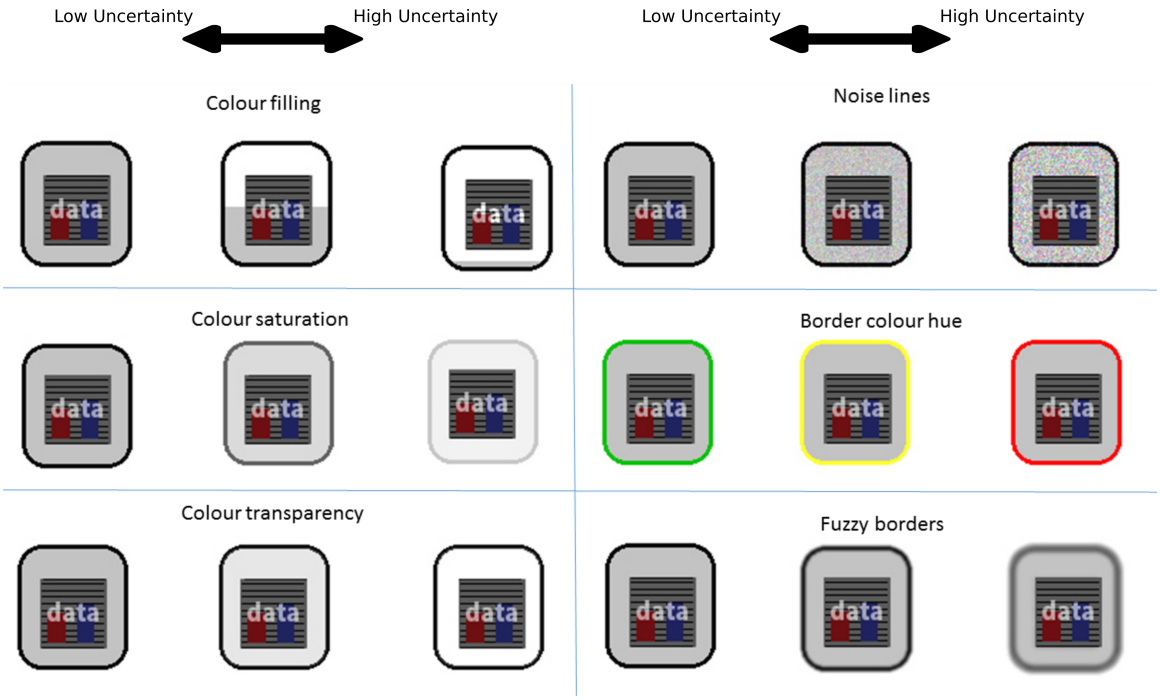


Figure 6.6: Candidates for aggregated power uncertainty visualisation. The outer boarder of the glyph indicates the aggregated uncertainty. Uncertainty increases from left to right. This figure appeared in Senaratne et al. (2014b).

the think-aloud method (Nielsen et al., 2002) where the participants were asked to verbalise their thought process when solving the given tasks . With the permission of the participants these thoughts and comments were recorded.

Figure 6.7 shows the three-tiered procedure of the pilot study. Accordingly, in a first step we give the participants a brief introduction to the context of the dataset and the tasks. In a second step in the pilot study we showed a randomised order (1, 2, 3) of low, medium and high uncertainty depictions of chosen visualisations in the foreground of a geographical map (as seen in Figure 6.8). Then we asked the participants to give the correct order of the uncertainty visualisations from low, medium to high (the correct order for the given example would be 3, 1, 2). At the next step we conducted qualitative interviews to rank the visualisations based on the preference, and understanding on a lickert scale from 1 - very bad to 5 - very good. This allows us to assess the intuitiveness of the different visualisations.

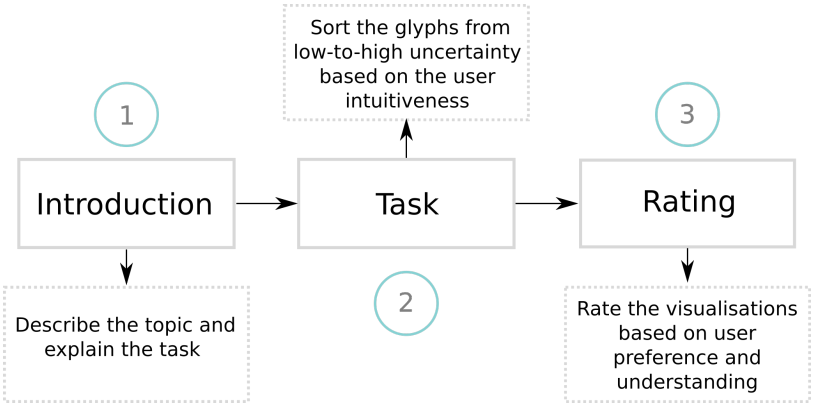


Figure 6.7: The three-tiered work flow of the pilot study.

The results from the pilot study are shown in Figure 6.9. Based on these results for depicting the modular power uncertainty, the visual metaphors *icon*, *saturation of green*, and *transparency* are selectedd for the main study. Although the *thumbs up icon* is slightly less understanding compared to the *smiley icon*, in terms of preference and performance they both have similar results. However, since the *thumbs up icon* can be scaled smaller in a visualisation while still being identifiable, it was chosen for the main study . Because the participants had difficulties in identifying the thumbs up icon at a glance, the bordering area around this icon (within the inner rectangle of the glyph) was filled with a neutral gray colour to improve the contrast effect for better identification of the icon. The *green saturation* visual metaphor was well received by the participants, where it supposedly signals trustworthiness. This metaphor further yielded good results in terms of preference, understanding, and performance.

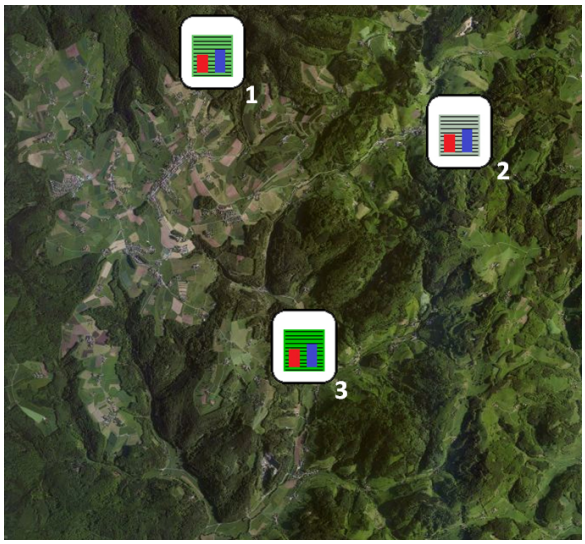


Figure 6.8: Example of the randomised uncertainty depictions of the modular power uncertainty.

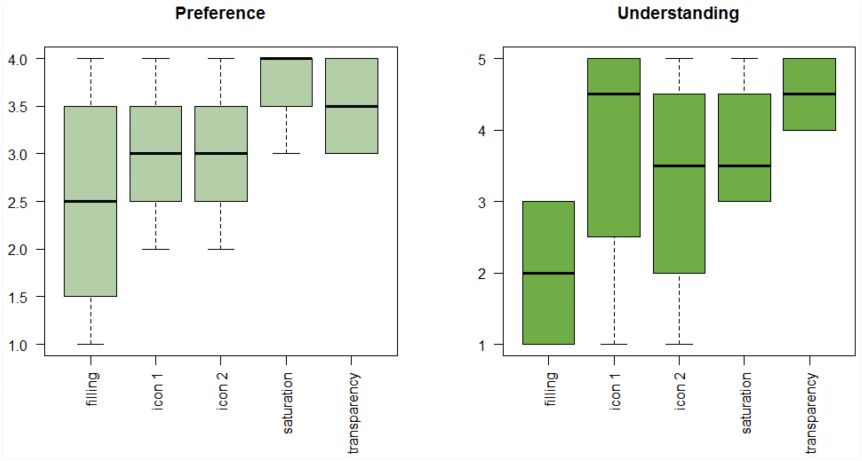
The *transparency* visual metaphor, as also evident from related studies, yielded good results.

For depicting the aggregated power uncertainty, the *transparency*, *noise*, and *fuzziness* visual metaphors were chosen. The *noise* metaphor, although it was difficult to distinguish between different levels of uncertainty according to the participants, yielded good results in terms of preference and understanding. As a solution to this, instead of *noise lines*, the metaphor was changed into *noise holes* similar to the work of Payne (2009) on colour mixing visualisations. In addition, the *transparency* and *fuzziness* visual metaphors yielded good results, and therefore were also included in the main study.

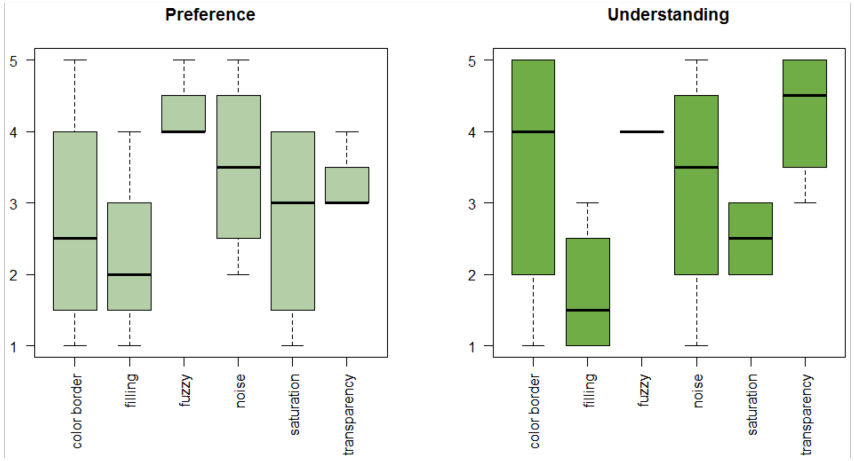
6.5 Performance and Preference of Glyph Designs

In this section we evaluate the usability of the bi-dimensional uncertainty visualisation candidates, as selected in the pilot study (Section 6.4), based on the performance and preference within a usability study.

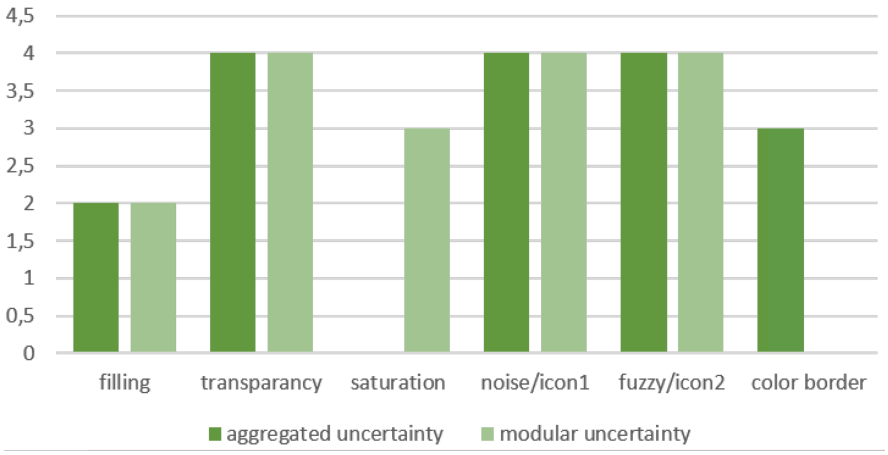
According to Nielsen (1994), when assessing how users interact with an interface, there are particular components of usability, which can be tested. These components are: *easy to learn*, *efficient to use*, *easy to remember*, *minimal errors*, and *subjectively pleasing*. To test the usability of the modular and aggregate uncertainty visualisation candidates the *easy to learn* and *subjectively pleasing* components were assessed though an online study. Based on the work of Senaratne et al. (2012), in the context of



a. Preference and understanding for modular uncertainty visualisations



b. Preference and understanding for aggregated uncertainty visualisations



c. Performance- the count of correct answers

Figure 6.9: Results of the pilot study.

uncertainty visualisation methods, *easy to learn* is the ability of different visualisations to communicate uncertainty to the users. Therefore, this component was assessed based on the users' performance, i.e., the correct answers to the specified tasks. The *subjectively pleasing* component was assessed based on the users' preference ratings on the visualisations. The goal of this study was to investigate the best combinations of visual metaphors for bi-dimensional uncertainty visualisation in the context of the example of smart grid numerical data. The following sections outline the design of the online study and its outcomes.

6.5.1 Design of the Usability Study

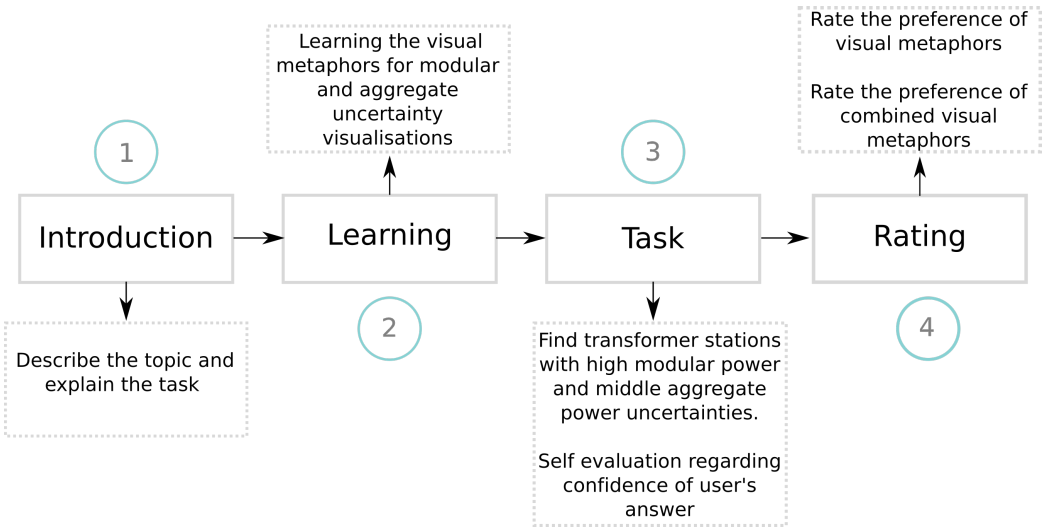


Figure 6.10: Four-tiered work flow of the main usability study.

Figure 6.10 shows the four-tiered work flow of the usability study that was conducted on the uncertainty visualisation candidates. The nine combinations of the modular and aggregate uncertainty visualisations that were used in the study are shown in Figure 6.11 and are as follows in the order of "aggregated power uncertainty / modular power uncertainty": (1) Fuzziness / Icon, (2) Fuzziness / Saturation, (3) Fuzziness / Transparency, (4) Noise / Icon, (5) Noise / Saturation, (6) Noise / Transparency, (7) Transparency / Icon, (8) Transparency / Saturation, (9) Transparency / Transparency. Based on a *within subjects design*, these nine combinations could be correctly indicated four times by each participant. To make this study sustainable, the alarm level of the data visualisation (shown through the bar charts in the inner rectangles of the glyph) was kept at a constant value.

The 20 participants who took part in the study consisted of 8 female and 12 male, all within the ages between 18 - 27 years, and all had a background in either computer science or mathematics. In order to acquire these participant, an email with a link to the online study was circulated through mailing lists. As an incentive to take part in the study, we offered the chance of winning a 25 Euro gift voucher through a raffle draw. The participants also had the chance to stay anonymous by not signing up for the raffle draw.

The hypotheses that the study is based upon are given below. The results of the pilot study (Section 6.4) on each of the individual visual metaphors helped to derive these hypotheses.

Hypothesis 1 (H1): The combinations (for aggregate uncertainty / modular uncertainty) of Transparency / Transparency, Transparency / Icons, and Fuzziness / Icon are more intuitive than Noise / Saturation combination. Further, Transparency / Transparency, Transparency / Icons are more intuitive than Fuzziness / Saturation.

Hypothesis 2 (H2): Noise / Saturation combination has a higher bias than Transparent / Transparent, Transparent / Icons, Fuzziness / Icons, and Fuzziness / Transparency.

In the following sections each of the four steps of the work flow (Figure 6.10) is explained.

Step 1: Introduction

At the beginning of the study, the participants were informed about the main purpose of the study and then given a brief description of the meaning of uncertainty within the context of the smart grid environment as well as the two types of uncertainty that are visualised through the glyph design. Figure 6.12 shows how the description looks like in the survey. To better grasp the smart grid environment, the network of transformer stations was visualised on a map along with the power data indicated on the glyphs at each transformer station.

Step 2: Learning

In the learning step, the goal was to get the participants to learn the visual metaphors, and where the modular and aggregated uncertainties are situated in the glyph design. This was an important step before continuing on to the tasks. The participants were

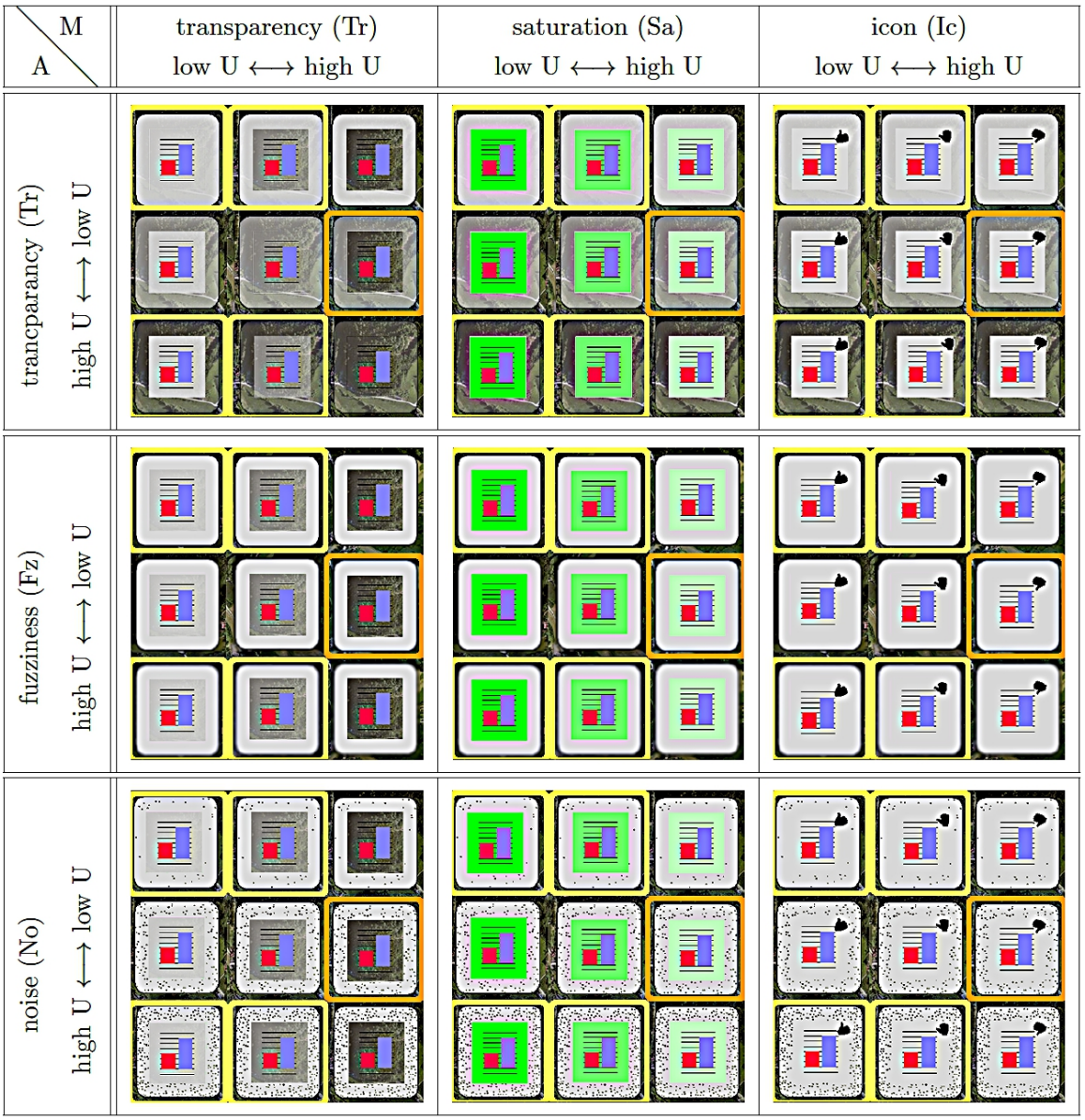


Figure 6.11: The combinations of the visual metaphors for modular uncertainty (shown in columns- M in the matrix), and aggregated uncertainty (shown in rows- A in the matrix). The low, middle, high uncertainty values for modular uncertainty are shown from left to right of the columns, and the low, middle, high uncertainty values for aggregated uncertainty are shown from top to bottom of each row.

shown examples of aggregated uncertainty (Figure 6.13 a) and modular uncertainty (Figure 6.13 b) along with instructions to learn how the individual visual metaphors depict the scales of the two types of uncertainties. This was repeated for each visual metaphor.

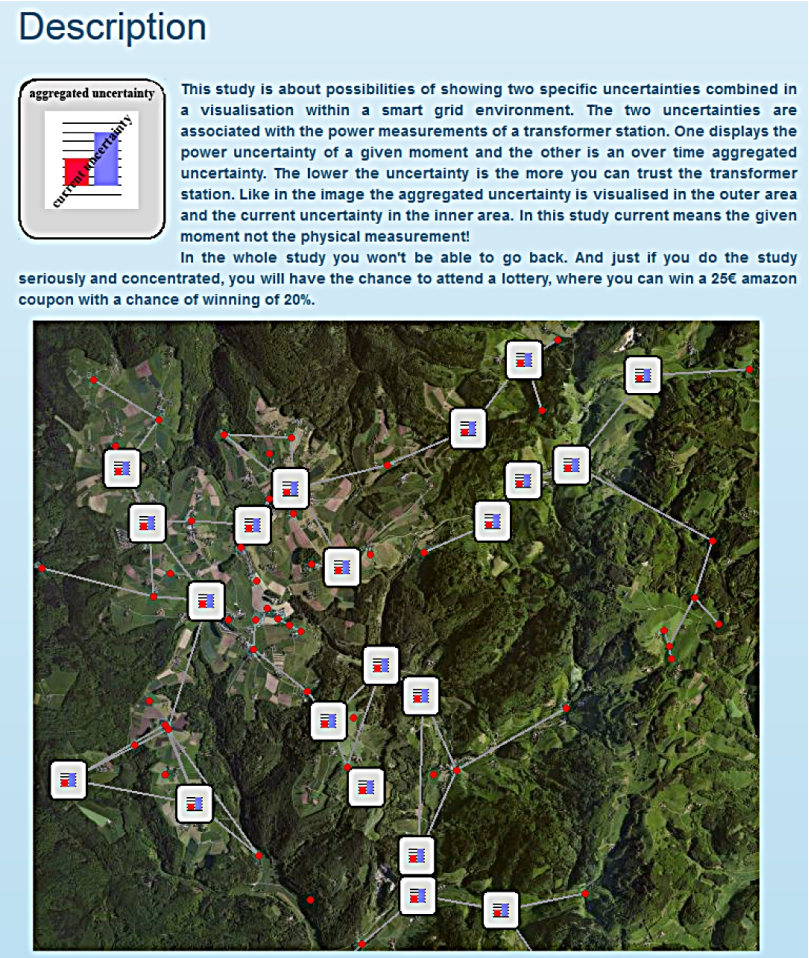


Figure 6.12: An excerpt of the online study that introduces the context and the two types of uncertainty and their respective visualisations to the participants.

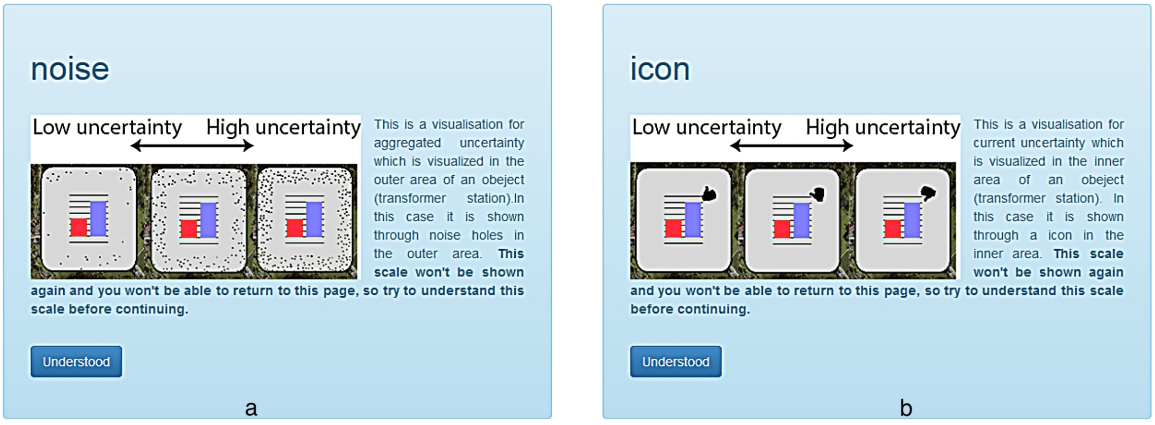


Figure 6.13: An excerpt of the online study that instructs the participants to learn the individual visual metaphors. (a) noise metaphor depicting the three scales of aggregated uncertainty. (b) icon metaphor depicting the three scales of modular uncertainty.

Step 3: Task

The participants were given the task of finding the transformer station with given modular and aggregated power uncertainty values among many transformer stations with varying modular and aggregated power uncertainty depictions. An example task is shown in Figure 6.14. After selecting the combination of visual metaphors, that the participants think is the correct one, the participants are further asked to indicate how confident they are of their answer.

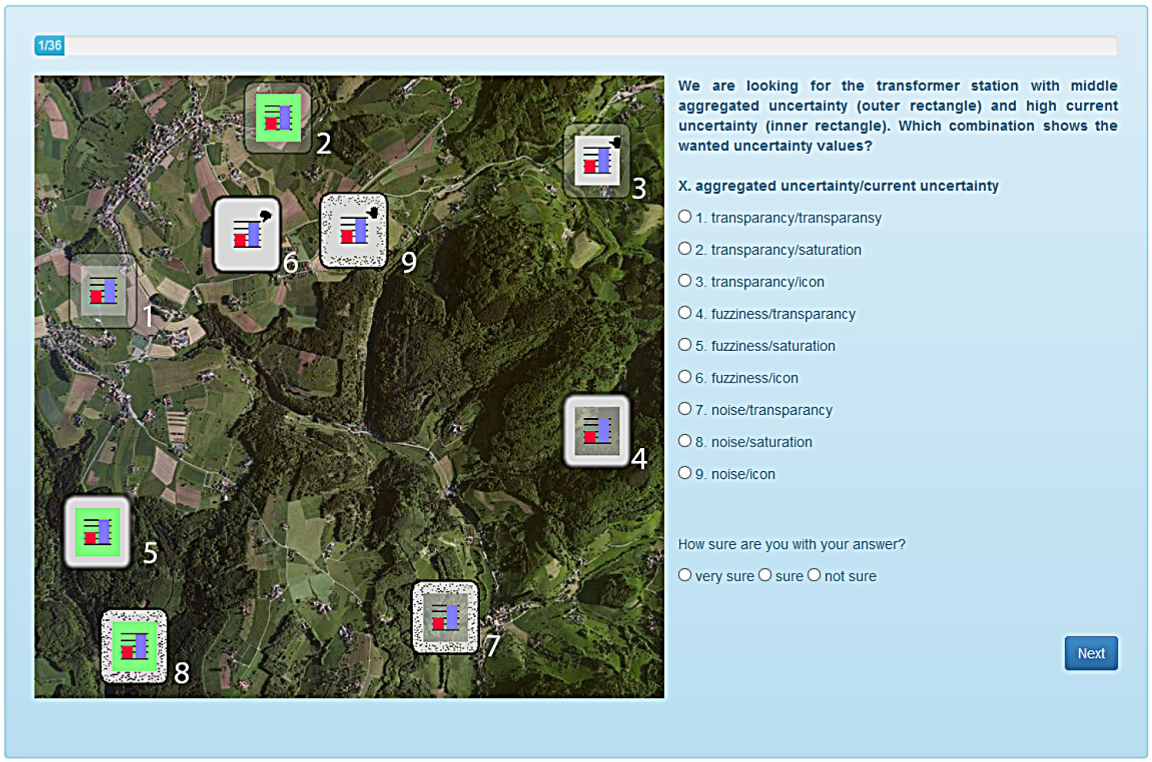


Figure 6.14: An excerpt of the online study that asks the participant to find the transformer station with middle aggregated power uncertainty and high modular power uncertainty. 'Current' here means the present modular uncertainty.

Step 4 : Rating

At the final stage of the study, the participants were asked to rate the modular uncertainty visualisations, the aggregated uncertainty visualisations, and then the combinations of the different visual metaphors based on their preference. This preference is based on the visually pleasing aspects of the chosen visual metaphors and their combinations.

Table 6.2: Significance analysis for the *performance* of participants within the visualisation combinations using the Wilcoxon matched pairs signed rank test.

Hypothesis for the combination of visual metaphors	**p-value**	**Null hypothesis**
Transparency / Transparency is more intuitive than Noise / Saturation	0.03588	rejected
Transparency / Transparency is more intuitive than Fuzziness / Saturation	0.1737	not rejected
Transparency / Icons is more intuitive than Noise / Saturation	0.1285	not rejected
Transparency / Icons is more intuitive than Fuzziness / Saturation	0.4756	not rejected
Fuzziness / Transparency is more intuitive than Noise / Saturation	0.04168	rejected
Fuzziness / Icons is more intuitive than Noise / Saturation	0.02035	rejected
Noise / Transparency is more intuitive than Noise / Saturation	0.01664	rejected
Noise / Icons is more intuitive than Noise / Saturation	0.04654	rejected

6.5.2 Study Results

For assessing the correctness of participant answers on the specified tasks, the *Wilcoxon Matched Pairs Signed Rank* test (Wilcoxon, 1945) was used. Following a one-tailed test, two dependent samples were compared. The correctness is assessed based on the number of times a participant selected the correct combination for the given modular and aggregated uncertainties. The null hypothesis is that there is no difference among the compared data values for the two considered samples. The alternative hypothesis is that there is a difference in a specified direction (Wilcoxon, 1945). The following Table 6.2 summarises the test results for the hypothesis H1.

As evident from the results, hypothesis H1 holds true for combinations with Transparency, Fuzziness, Noise holes, and Icons. They significantly outperform the combinations with Saturation in terms of user performance (correctness). These results are in line with existing findings from previous empirical evaluations of uncertainty visualisation methods such as MacEachren et al. (2012). It can be concluded that the above combinations are the best combinations for correctly analysing bi-dimensional uncertainties within the smart grid numerical data example use case. For testing the bias of the combinations, the *correctness* was analysed against the participant confidence in answering the task oriented questions. A combination is considered to

have high bias if the said combination was selected incorrectly, but was indicated as *very sure* by the participant in terms of confidence.

Therefore the bias for a given combination was calculated as follows:

$$correctess = \begin{cases} -1, & whenCorrectlyIdentified \\ 1, & whenNotCorrectlyIdentified \end{cases} \tag{6.1}$$

$$participantconfidence = \begin{cases} 1, & whenConfidence = notSure \\ 2, & whenConfidence = sure \\ 3, & whenConfidence = verySure \end{cases} \tag{6.2}$$

and

$$bias = correctness * confidence \tag{6.3}$$

An excerpt of the bias results for the visualisation combinations, where the modular and aggregated uncertainties were both kept at low are shown in Table 6.3. The Wilcoxon matched pairs signed rank test was once again used to test hypothesis H2. To do this, first, the bias of the different environments was summed up (Table 6.3):

$$bias = \sum_{i=1}^{n} correctness_i * confidence_i \tag{6.4}$$

Same as for correctness, the null hypothesis is that there is no difference between the matched data values within the two compared samples. The alternative hypothesis is that *there is a difference in a specific direction.* In the assessment of bias, the first sample is greater than the second one. The Table 6.4 summarises the test results for the hypothesis H2.

The results from Table 6.4 reveal that the combination Noise / Saturation has high bias, which led the participants to select any other combination as the correct answer to the task oriented questions, and also indicate "very sure" in terms of confidence. The combinations with Transparency, Icons, and Fuzziness on the other hand were mostly correctly chosen, with "very sure" in terms of confidence.

For assessing the preference of the individual as well as combined visual metaphors for the aggregated / modular uncertainties, the participants were asked to rate them. The ratings from the individual visual metaphors for modular power uncertainty and aggregated power uncertainty are shown in the Figure 6.15.

Based on these individual ratings a hypothesis H3 is derived for the combined visualisations for aggregated / modular uncertainty:

Table 6.3: Excerpt of two participants' bias for the different combinations of visualisations.

Participant	Combination	correctness	confidence	bias
1	Transparency / Transparency	1	1	1
1	Transparency / Sauration	-1	2	-2
1	Transparency / Icons	-1	2	-2
1	Fuzziness / Transparency	1	1	1
1	Fuzziness / Saturtion	-1	2	-2
1	Fuzziness / Icons	-1	1	-1
1	Noise / Transparency	-1	1	-1
1	Noise / Saturation	-1	2	-2
1	Noise / Icons	-1	1	-1
2	Transparency / Transparency	-1	3	-3
2	Transparency / Saturation	-1	3	-3
2	Transparency / Icons	-1	3	-3
2	Fuzziness / Transparency	-1	3	-3
2	Fuzziness / Saturation	-1	3	-3
2	Fuzziness / Icons	-1	3	-3
2	Noise / Transparency	-1	2	-2
2	Noise / Saturation	-1	2	-2
2	Noise / Icons	-1	2	-2

Table 6.4: Significance analysis for the *bias* of visualisation combinations using the Wilcoxon matched pairs signed rank test.

Hypothesis for the combination of visual metaphors	p- value	Null hypothesis
Noise / Saturation has higher bias than Transparency / Icons	0.01377	rejected
Noise / Saturation has higher bias than Transparency / Transparency	0.001522	rejected
Noise / Saturation has higher bias than Fuzziness / Icons	0.01797	rejected
Noise / Saturation has higher bias than Fuzziness / Transparency	0.003598	rejected

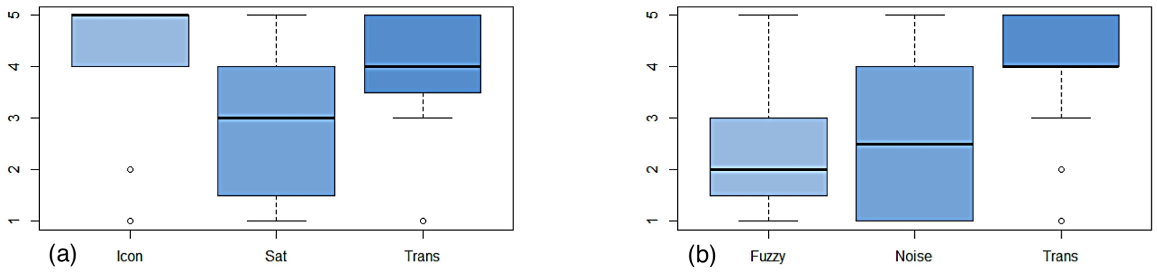


Figure 6.15: The participants' preference rating on the individual uncertainty visualisations. (a) preference rating for modular uncertainty visualisations - Icons and Transparency had the highest ratings. (b) preference rating for aggregated uncertainty visualisations - Transparency had the highest ratings.

Table 6.5: Significance analysis for the *preference* of visualisation combinations using the Wilcoxon matched pairs signed rank test.

Hypothesis for the combination of visual metaphors	**p-value**	**Null hypothesis**
Transparency / Transparency is preferred over Noise / Saturation	0.003049	rejected
Transparency / Transparency is preferred over Fuzziness / Saturation	0.002811	rejected
Transparency / Icons is preferred over Noise / Saturation	0.005238	rejected
Transparency / Icons is preferred over Fuzziness / Saturation	0.004212	rejected

Hypothesis 3 (H3): The combinations of Transparency / Transparency and Transparency / Icons have higher preference over the combinations with Saturation with Fuzziness and Noise, in terms of the visual appeal of the visualisations.

Once again the Wilcoxon matched pairs signed rank test was used to test the hypothesis H3, where it is assumed that the combinations of Transparency / Transparency and Transparency / Icons have significantly higher preference over the combinations of Saturation with Fuzziness and Noise. The null hypothesis is that there is no difference in the values of both samples. Table 6.5 sums up the test results.

Evidently, the most preferred visualisation combinations for depicting the aggregated / modular uncertainties are Transparency / Transparency as shown in Figure 6.16 and Transparency / Icons as shown in Figure 6.17.

Finally, the chosen visual metaphors are integrated into an implementation of the overall infrastructure as seen in Figure 6.18.

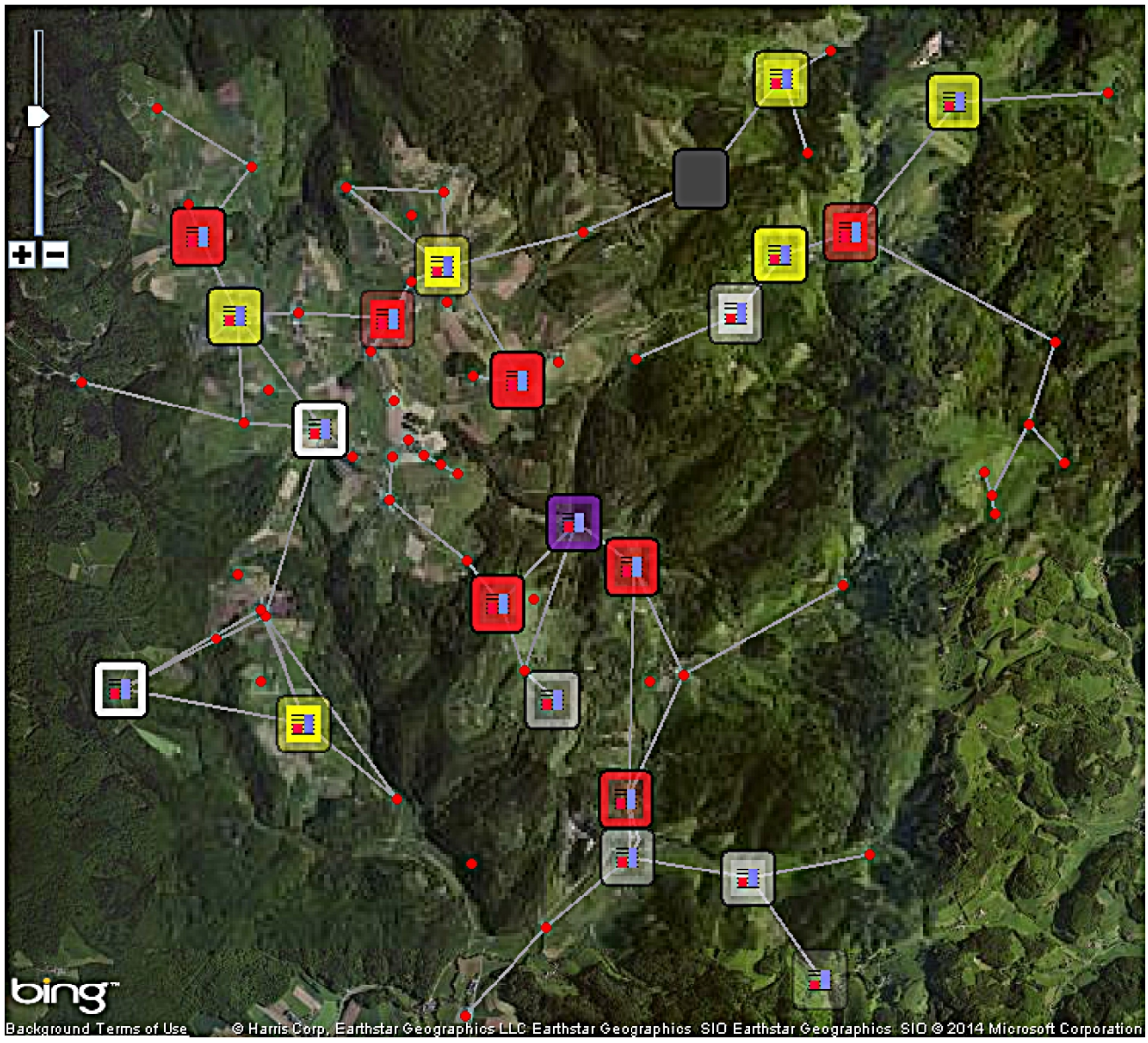


Figure 6.16: The aggregated power uncertainty and the modular uncertainty at the different transformer stations depicted through Transparency / Transparency visualisation combination.

6.6 Discussion & Future Work

In this chapter several visual glyph designs for communicating bi-dimensional uncertainties in numerical data have been tested through a rigorous evaluation process. The outcome for the most effective glyph designs are very much resonant with previous studies in similar settings. Throughout the study several limitations have also been met. In the realistic scenario within the smart grid environment, four alarm levels are constantly indicated in the data – *normal* (requiring least attention shown in grey colour), *warning* (speculative shown in yellow colour), *alarm 1* (take necessary action to alleviate shown in red colour), and *alarm 2* (beyond the control of the operators shown in purple colour). We considered only the *normal* alarm level in

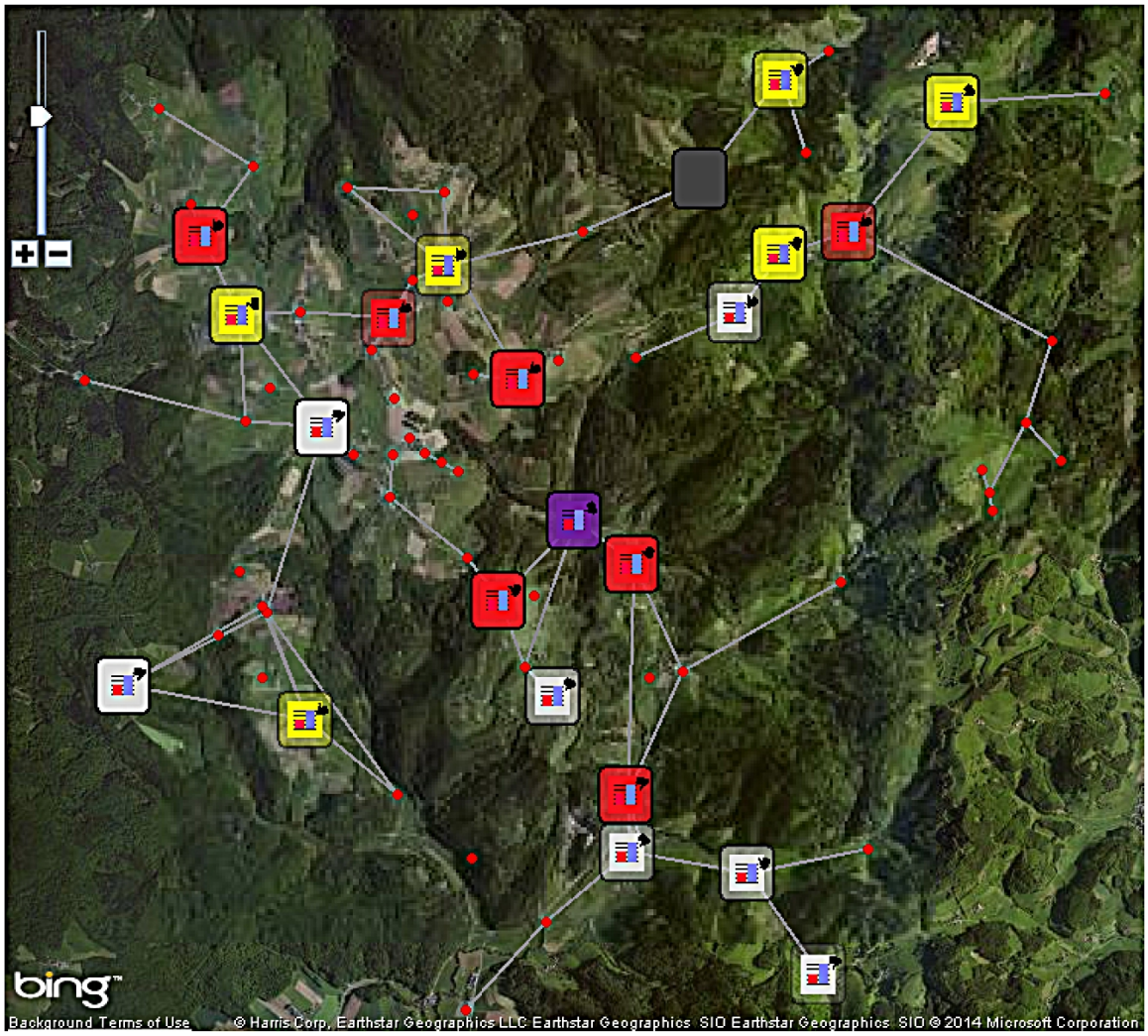


Figure 6.17: The aggregated power uncertainty and the modular uncertainty at the different transformer stations depicted through Transparency / Icons visualisation combination.

designing the bi-dimensional uncertainty glyphs, with a grey coloured background. This is because we considered uncertainty awareness in this context is most important when the operator thinks that everything is functioning properly, but in reality it may not. Therefore within the scope of the study it was enough to consider just one alarm level. In the future the designed glyphs can be manipulated to also incorporate other dimensions in data such as the other alarm levels in the smart grid environment scenario.

Furthermore, as in many usability studies, the participants were asked directly to select a given combination of uncertainties (e.g., middle aggregated power uncertainty and high modular power uncertainty). In the future such usability studies on uncer-

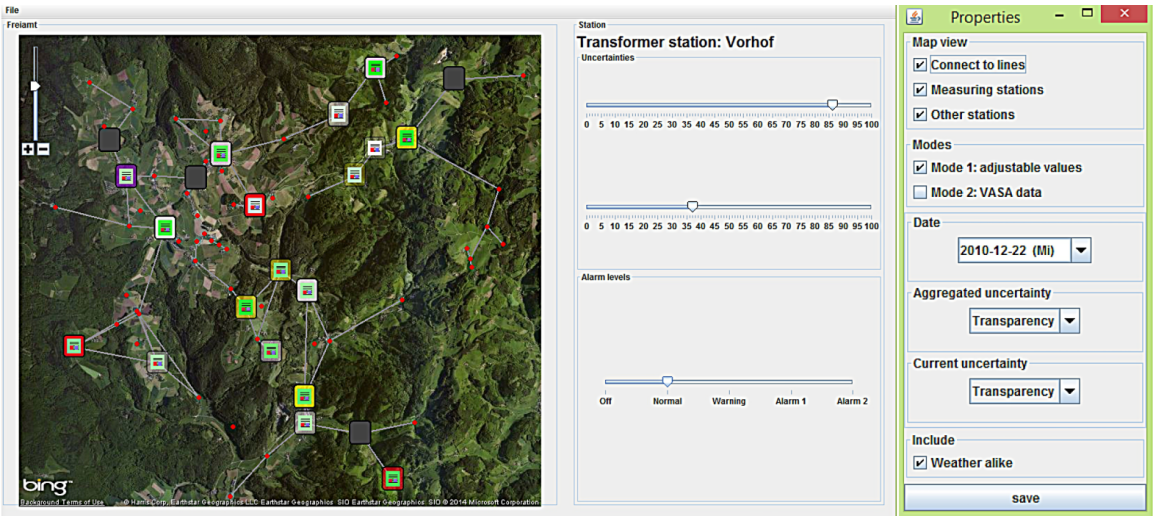


Figure 6.18: An implementation of the approach. On the left panel the user can choose the values for the two different types of uncertainties from a scale between 0 and 100, as well as the alarm levels they want to see for a chosen transformer station. The right side panel allows the user to choose the properties in terms of the map view, user- adjustable data or upload the dataset, as well as the visual variables for the aggregated power uncertainty and modular power uncertainty.

tainty visualisation designs can be designed to be more context-aware, by changing the direct uncertainty-based questioning to passive uncertainty-based questioning. For example, the participants can be given an example scenario, ask them to assume the role of a decision maker (or choose actual decision makers) and ask them context based questions where they have to consider uncertainty in data in order to make proper decisions. This will help researchers to design uncertainty visualisations most suitable for the tasks at hand. Due to the variation in data, context, users, or uncertainties, it is unfeasible to design generic visualisations. Such anecdotal evidences of visual designs for uncertainty will serve in the future as references for designing context-aware uncertainty visualisations.

6.7 Conclusions

Visual analytics systems are in place within smart grid environments to alleviate crisis situations by allowing decision makers to perceive and understand the severity of a crisis situation. However, errors in measurements that are propagated due to various reasons (such as data transformations, or errors in measurement devices) can make the decision makers less confident in deriving information. Therefore, analysis and visualisation of uncertainty within such data has become important. In this chapter we have utilised

two uncertainty assessment methods: sampling and Monte Carlo simulation, to assess uncertainties inherent in power data within a smart grid environment, and compared their performance to best fit our use case. We found that the Monte Carlo simulation method is most suitable for measuring uncertainty in our application domain. Further, through a usability study we identified most effective visual metaphors to communicate to crisis managers bi-dimensional uncertainties comprising modular power uncertainty and aggregated power uncertainty. A prototype that incorporates these visual designs has been implemented to be utilised within an uncertainty-aware smart grid monitoring application.

Chapter 7

Conclusion and Future Work

Contents

7.1 Summary of Conclusions 177
7.2 Interdisciplinary Visual Analytics Research 180
7.3 Future Work 181

This dissertation defines a conceptual framework for handling uncertainty in visual analytics. It demonstrates example visual analytics approaches for the analysis of uncertainties in different selected data types. At the end of each of the above chapters, the presented approaches are comprehensively discussed, their limitations are pointed out, possible developments in future work are outlined, and conclusions are drawn. In this chapter, a bird's eye view perspective is taken in Section 7.1 to discuss the impact of the presented research in the larger visual analytics field. Section 7.2 describes how the conducted interdisciplinary research helps to further the visual analytics agenda. Section 7.3 ends this chapter by outlining some future directions for uncertainty-aware visual analytics research.

7.1 Summary of Conclusions

An overview of the contributions of this dissertation is shown in Figure 7.1, which extends Figure 2.7 presented in Section 2.2. The blue coloured rectangles represent the developed visual analytics approaches for *assessing* the inherited uncertainty in data. The ovals represent the visual analytics approaches for *reducing* uncertainty in data (blue) and in models (green). Approaches for both authoritative and non-authoritative data were developed, as shown in white dotted rectangles in the figure.

Spatio-temporal data are inherently *uncertain*. Such uncertainties arise mainly during the data collection phase. Although sometimes quality control standards or

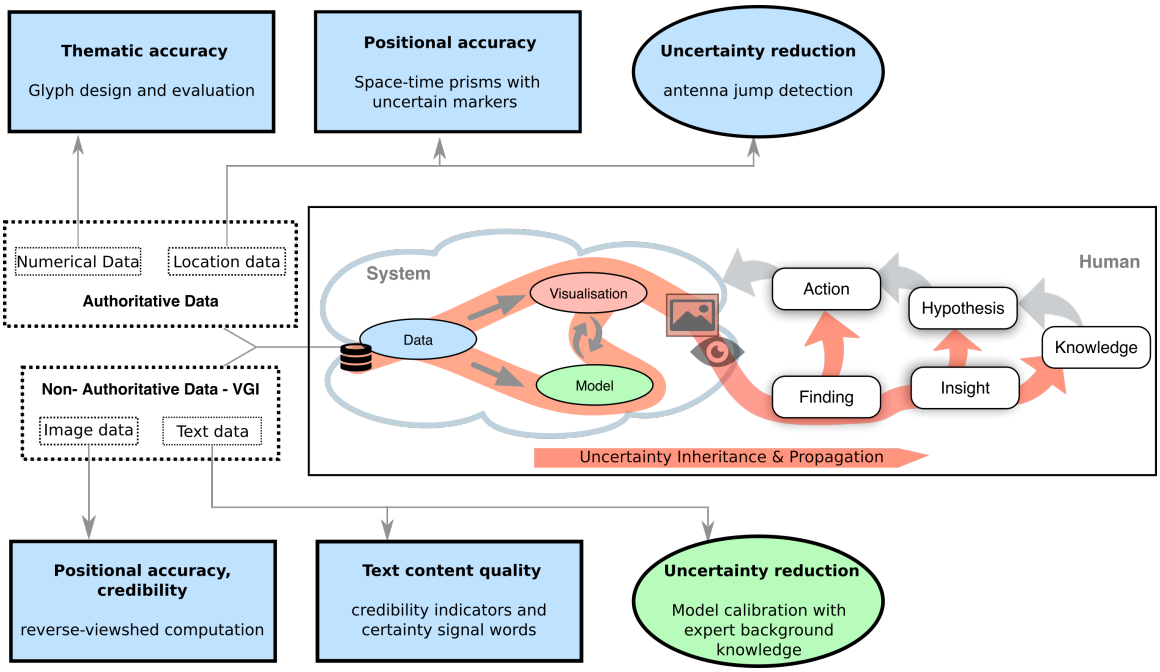


Figure 7.1: Summary of contributions.

gatekeepers are in place to minimise these uncertainties, an accidental omission, an unwanted commission of data, or a low-resolution mapping are unavoidable. We have found a number of reasons for this, a main one being that geographic entities are continuous in nature, and also due to the quality and limitations of spatial knowledge of humans. Since uncertainty has a strong impact on the decisions we make, incorporating uncertainty dimensions into data analysis can help users to reduce errors, derive more trustworthy results, or acknowledge the insufficiency of data to avoid making bad decisions.

Throughout the years we have seen how uncertainty visualisation has evolved from static, to dynamic, to interactive in nature (Section 2.1) in order to visually communicate uncertainties. Visual analytics approaches (Section 2.1.2) focused on enabling the users with more interaction and fact-finding to explore uncertainties in data. However, uncertainty is not only inherent in data. A drawback in many existing visual analytics frameworks is that it disregards the uncertainties that keep propagating in a system. In efforts to account for propagated uncertainties, some previous works focused on incorporating uncertainty propagation and communicating it in the information visualisation pipeline. We saw other works that go a step further and introduce workflows to identify inherent and propagated uncertainties in data and visualisations within visual analytics pipelines.

Extending these works, this dissertation has taken a more microscopic view on the role of uncertainty within the visual analytics knowledge generation process. As a result it systematically identified on the one hand that uncertainty inheritance in spatio-temporal data depends on how the data are collected – types of uncertainties that are inherited in authoritative data, and types of uncertainties that are inherited in non-authoritative volunteered data. On the other hand it identified uncertainty propagation further in the model building phase (e.g., bad parameters), model usage (e.g., model misspecification), visual mapping (e.g., inappropriate visual metaphors), in the visualisation (e.g., low-resolution visualisations), and model-visualisation coupling stages (e.g., poorly designed mappings between the model and the visualisation) in a data analysis system. To assess these inherited and propagated uncertainties this dissertation has extensively reviewed state of the art methods, and thereby produced thirteen *guidelines* for the assessment of inherited and propagated uncertainties. As a result we have established a comprehensive framework for the role of uncertainty in the visual analytics knowledge generation process (Section 2.2). This is shown in the rectangle in the center of the Figure 7.1.

As can be seen from the guidelines, most uncertainties are inherited in the data and therefore many works propose methods to assess these uncertainties. Inherited uncertainty in volunteered spatial data can be assessed following three approaches according to Goodchild and Li (2012). These are classified as geographic, crowd-sourced, and social approaches. Through a review of existing methods this dissertation further identified an additional approach to assess the inherited uncertainty in spatial data: *data mining*, thereby extending the classification of Goodchild and Li (2012).

One main drawback in the reviewed methods is that most of them are fully automatic, which hinders the users from weighing their knowledge in on the uncertainty analysis process. Questions such "to what extent do I want uncertainty to play a role?", "does uncertainty give more clarity or obscure the overall analysis process?", "how can I reduce uncertainties?" cannot be resolved using only fully automatic approaches. Visual analytics bridge this gap by enabling the user to interactively steer the whole analysis process.

Hence, this dissertation introduces four novel visual analytics approaches to handle inherited uncertainty in four distinct data types: image data (Chapter 3), text data (Chapter 4), location data (Chapter 5), and numerical data (Chapter 6) following the framework introduced in Chapter 2. This is shown in the blue rectangles in Figure 7.1. These presented approaches on the one hand utilise the spatial dimensions along with the unique characteristics of data, such as image content to assess the positional

uncertainty in the data. On the other hand they utilise the unique characteristics of data, such as the textual content to assess the thematic uncertainties in the data.

The heterogeneity of the methods used to visually and interactively analyse uncertainty of the different data types confirms that uncertainty analysis should be treated as context-specific. As further evident from the summary of conclusions in Figure 7.1, in analysing uncertainty, a distinction needs to be made between the authoritative data - where data features alone have been used to assess the uncertainty, and non-authoritative data - where data features together with contributor features are used to assess the uncertainty.

Throughout the uncertainty analysis process, the user can go back in the exploration loop and determine the causes of uncertainty in the data or models, and take necessary action to *reduce uncertainties* in order to achieve more trustworthy knowledge. This dissertation has introduced two approaches for reducing uncertainties in the analysis process (Sections 4.3, 4.4, and 5.4.2). An example for reducing uncertainties caused by data is shown in the blue oval, and an example for reducing uncertainties caused by the data model is shown in the green oval in Figure 7.1.

In addition to the uncertainty-related contributions, this dissertation has further introduced two approaches for deriving and analysing movement trajectories from implicitly referenced text data, and explicitly referenced mobile communication-based location data, under the influence of uncertainty. The algorithm that is utilised in the former case, extracts spatial movement patterns based on the episodic sequence of hotspots that are detected in Twitter text data. These trajectories are further contextually analysed to observe the sequence of sentiments and content that move along the trajectory. In the latter case for mobile communication-based location data, the movement detection algorithm creates segments of a trajectory based on the sessions of mobile communication for any given user with a user-specified threshold. Based on the identified movement patterns, the spatial arrangements in the surrounding urban environment can be visually analysed.

7.2 Interdisciplinary Visual Analytics Research

As visual analytics comprises methods that can help in various application domains, it is crucial to advance visual analytics research by involving knowledge from other domains. This dissertation combines the domain knowledge from the field of Geoinformatics and the specifics of spatio-temporal data with the methodologies of visual analytics.

The interdisciplinary way of conducting this research was fruitful in both directions. Solving uncertainty-related issues in spatio-temporal data using methods and approaches from visual analytics resulted on the one hand in visual analytics solutions for problems from the geographical information science domain. On the other hand, it resulted in an enhancement of the visual analytics knowledge generation process by introducing to it the role of uncertainty predominantly based on spatio-temporal data. This observation is supported by the fact that this research could contribute to computer science venues (e.g., IEEE TVCG, ACM SIGSPATIAL) as well as to venues from geographical information science (e.g., IJGIS, TGIS).

As user generated data is particularly prone to uncertainty, VGI was a focal area of this research. VGI is produced by common citizens and its creation can be fostered, e.g., through dedicated Citizen Science projects. This domain was analysed as part of this research. Close interaction with stakeholders from this domain was initiated through a temporary research visit at the Extreme Citizen Science lab at the University College London (UCL) as well as outreach to the community (e.g., through an engagement in the Vespucci Summer School on *VGI and Citizen Science: engaging, creating and understanding*), and involvement in various conferences (e.g., GIScience, GeoViz). Through these interdisciplinary activities, the issues of uncertainty in VGI were investigated and solutions that utilise visual analytics were developed. However, the challenges of handling uncertainty in VGI are still broad and this area holds many research questions to be addressed in future work.

Besides looking into the field of VGI as a specific kind of spatio-temporal data assessed by users, this research investigated authoritatively collected telecommunications data. During this part of the work, stakeholders from this domain (specifically from the company "Telefonica Germany") were incorporated as experts and sources for requirements and feedback. Through a collaboration with researchers at the University of Chile in Santiago real telecommunication data could be gathered, and visual analytics solutions could be developed to derive urban dynamics patterns from telecommunication data as well as visual analytics solutions for dealing with uncertainty in telecommunication data.

7.3 Future Work

In future work, further visual analytics approaches need to be developed to analyse propagated uncertainties in models, visualisations, model-visualisation couplings based on our framework. The availability of anecdotal evidences of similar visual analytics

approaches will create a body of knowledge that can act as references for handling uncertainty within the visual analytics knowledge generation process. This dissertation has shown that it is equally important to also reduce uncertainties in the knowledge generated. Therefore more mechanisms need to be developed along different data types and use cases to reduce uncertainties. For example, interpolation has shown to increase uncertainties in data. However, interpolation can also help to reduce uncertainty in data as seen in Section 5.4.2. Therefore, it is necessary to emphasise here that uncertainty in data should be looked at on a use case basis.

For some use cases, such as in disaster mitigation, it is important to look at many datasets before deriving important decisions. While future visual analytics work should look into the integration of heterogeneous data sources, semi-automatic uncertainty-aware merging processes need to be in place to improve the current state of the art research. Further, many more uncertainties may evolve out of such heterogeneous data fusion processes, and these need to be accounted for in analyses. As a final thought, with uncertainty taking center stage in future data analyses, human interaction methods need to be redefined to suite the uncertainty-aware data analysis approaches.

List of Abbreviations

API	Application Programming Interface
CID	Cell Identifier
DBSCAN	Density-based Spatial Clustering of Applications with Noise
DE-9IM	Dimensional Extended- nine Intersection Model
DSM	Digital Surface Model
DBSCAN	Density-based Spatial Clustering of Applications with Noise
GPS	Global Positioning System
GSM	Global System for Mobile Communications
KDE	Kernel Density Estimation
ISO	International Standardisation Organisation
ISO/TC	International Standardisation Organisation / Technical Committee
LAC	Location Area Code
LCA	Latent Class Analysis
LE90	Linear Error of 90%
LoS	Line of Sight
MPP	Most Probable Point
NLP	Natural Language Processing
OSM	OpenStreetMap
PM10	Particulate Matter 10 µm or less in diameter
PDF	Probability Distribution Function
POI	Point of Interest
PPGIS	Public Participation Geographic Information Systems
PPA	Potential Path Area
PTV	Possibilistic Truth Value
SOM	Self Organizing Map
SIFT	Scale-Invariant Feature Transform
tf-idf	term frequency-inverse document frequency
VGI	Volunteered Geographic Information
URL	Uniform Resource Locator

List of Figures

1.1 Classification of VGI based on the nature of volunteering and the spatial dimension (figure is adapted from M. Craglia (2012)). 2
1.2 An example of uncertainty depiction in the German daily news channel Tagesschau. (Source: `wetter.tagesschau.de`). 3
1.3 An example for errors caused through inappropriate visualisations used for uncertainty depiction. (Source: `http://www.nhc.noaa.gov/aboutcone.shtml`). 4
1.4 Work flow of the approaches followed in the subsequent chapters. . . 5
1.5 Structure of this dissertation. 6

2.1 Whitening. Colour hue represents top soil thickness data and the saturated intensity represents the uncertainty of top soil thickness. Figure is taken from Hengl and Toomanian (2006). 22
2.2 Uncertainty visualisation with glyphs. Wind velocity data is depicted through the length of the arrow plot and its angular uncertainty through width of the arrow head. Figure is taken from Pang (2001). 23
2.3 Uncertainty visualised through contouring method. PM10 concentration data is shown in the background with higher saturation of red corresponding to higher concentration and lower saturation corresponding to lower concentration. The uncertainty of PM10 concentration is shown in the foreground with thickening contour lines. Thicker contours correspond to higher uncertainty and thinner contours to lower uncertainty in PM10 concentration. Figure is taken from Senaratne et al. (2012). 25
2.4 Uncertainty of ground-level Ozone concentration data for each hour is visualised through the error bars. The height of each error bar represents the amount of uncertainty associated for the given hour. Figure is taken from Senaratne et al. (2012). 26

2.5 Adjacent Maps. Colour saturation is used to depict high/low PM10 values and high/low uncertainties of PM10. Figure is taken from Senaratne et al. (2012). 27
2.6 Uncertainty in PM10 concentration data for Europe is visualised through (a) a probability map with a rainbow colour scale, (b) PDF graph, and (c) the corresponding probability values. Figure is taken from Senaratne et al. (2012). 29
2.7 A framework for uncertainty inheritance and propagation within the visual analytics knowledge generation process. This figure appeared in Sacha et al. (2016). 31
2.8 Source uncertainty is inherent to the data. It further varies depending on the authoritative and non-authoritative nature, as well as the implicit and explicit geography that is captured in the data. 32
2.9 Example of an incorrectly geotagged photo on Flickr (Brandenburg Gate in Berlin is tagged in Jakarta). This figure appeared in Senaratne et al. (2017a). 33
2.10 Distribution of the surveyed papers. This figure appeared in Senaratne et al. (2017a). 35
2.11 Uncertainty propagation through data processing. 54
2.12 Uncertainty propagation through model building and model usage. . . 55
2.13 Uncertainty propagation through visual mapping and the visualisation. 57
2.14 Uncertainty propagation through the model-visualisation coupling. . . 59
2.15 Aggregation of inherited and propagated uncertainties. 60

3.1 Geotags of Flickr photos that were textually tagged as “Angkor” and “Cambodia”. This figure appeared in Senaratne et al. (2013a). 67
3.2 The parameters for a viewshed calculation. (Source: `http://www.esri.com/software/arcgis`). 70
3.3 Work flow diagram for positional accuracy analysis within image-based Flickr. 71
3.4 An excerpt of the study area in Berlin overlayed with the DSM. . . . 73

3.5 Reverse-viewshed from four exemplar observer positions (indicated with the arrow head, and the image taken from the position) to the Brandenburg Gate (indicated with the red rectangle). (a) image is incorrectly geotagged and incorrectly labelled, (b) image is incorrectly geotagged, but correctly labelled, c) image is correctly geotagged, but incorrectly labelled, and d) image is correctly geotagged and correctly labelled. 74
3.6 Reverse-viewshed from four exemplar observer positions (indicated with the arrow head, and the image taken from the position) to the Reichstag (indicated with the red rectangle). (a) image is incorrectly geotagged and incorrectly labelled, (b) image is incorrectly geotagged, but correctly labelled, (c) image is correctly geotagged, but incorrectly labelled, and (d) image is correctly geotagged and correctly labelled. These figures appeared in Senaratne et al. (2013a). 75
3.7 Distribution of data for category 'a' within the Brandenburg Gate use case. 78
3.8 Distribution of data for category 'b' within the Brandenburg Gate use case. 78
3.9 Distribution of data for category 'c' within the Brandenburg Gate use case. 79
3.10 Distribution of data for category 'd' within the Brandenburg Gate use case. 79
3.11 Distribution of data for category 'a' within the Reichstag use case. . . 80
3.12 Distribution of data for category 'b' within the Reichstag use case. . . 80
3.13 Distribution of data for category 'c' within the Reichstag use case. . . 81
3.14 Distribution of data for category 'd' within the Reichstag use case. These figures appeared in Senaratne et al. (2013b). 81
3.15 Result analysis for the Brandenburg Gate and Reichstag use cases using the Chi Square test for independence. 83

4.1 The work flow for MovingOnTwitter. This figure appears in Senaratne et al. (2016, under review). 92
4.2 Hotspots detected with Kernel Density Estimation for the Lady Gaga dataset are visualised with a heat map. This figure appeared in Senaratne et al. (2014a). 95

4.3 Clustered routes before averaging the time. The arrows indicate the sequential direction from one cluster to the other. This figure appeared in Senaratne et al. (2014a). 96
4.4 The KDE of the Tweets relating to the Lady Gaga concert tour, and the clustered routes after averaging the time. Cities (e.g., Las Vegas, Dallas, Houston, or Toronto) where the concert took place are already visible as hotspots. Also cities where the concert was later canceled, such as in New York and Florida are evident through the hotspots. The arrows in the clustered routes indicate the sequential direction from one hotspot to the other. The colours of the trajectories depict the average sentiments from the respective hotspots. This figure appeared in Senaratne et al. (2014a). 96
4.5 Actual route (in Black colour) over the approximated trajectory (in Green, Yellow, and Red colour that depicts the averaged positive, neutral and negative sentiments from the respective hotspots). The route indicates the following concerts: Las Vegas (NV) on 25.01., Dallas (TX) on 29.01., Houston (TX) on 31.01., St. Louis (MO) on 02.02., Kansas city (MO) on 04.02., St. Paul (MN) on 06.02., Toronto (ON) on 08.02., and Montreal (QC) on 11.02., before the concert got cancelled for the remaining leg of the tour starting from Chicago (IL) which was supposed to take place on the following 13.02. This figure appeared in Senaratne et al. (2014a). 97
4.6 Trajectory visualisation for #melfest. The circles represent the clusters, and their colours represent the dominating topics in each cluster. This figure appears in Senaratne et al. (2016, under review). 100
4.7 Trajectory visualisation for #chinesenewyear. The circles represent the clusters, and their colours represent the dominating topics in each cluster. This figure appears in Senaratne et al. (2016, under review). . 101
4.8 The episodic clusters of #skilledtrade. The circle radius indicates the number of tweets in the cluster, and the colour hue indicates the most frequent topic observed in the cluster. These colour hues are used only to create the primary visual differences between the classes of topics, and they do not indicate any similarity between the topics. Colours are allocated to the topics using Colorbrewer. This figure appears in Senaratne et al. (2016, under review). 103

4.9 The sentiment and topic change from 08.02.2013 to 11.02.2013 in the Lady Gaga concert tour dataset. The positive to mostly negative change of sentiment together with the topic change indicate that the concert was canceled just before it was supposed to air in Chicago, USA. The actual tour information confirms this. The weighted credibility features *contains URL*, and *is Retweet* are used in combination to compute the average credibility of the tweets that pertain to the topics, and this is indicated through a 0° - 90° angle of each topic. More horizontal words indicate more credible topics (therefore easier to read) and more angular words towards 90° indicate non- credible topics (therefore more difficult to read). This figure appeared in Senaratne et al. (2014a). . . 105
4.10 Sentiment horizon chart for #6nations rugby tournament. A blue (far left) to red (far right) diverging colour scheme indicates the progression of positive to negative sentiments. Sentiment change along the 48 hour time frame can be clearly detected at two specific instances as highlighted in the green boxes. First instance is right after the game has started, second instance is when France scored its first point. This figure appears in Senaratne et al. (2016, under review). 107
4.11 The five classes of uncertainty signal words adapted from Kent (1964). 108
4.12 The certainty value used for sorting the Superbowl related hashtags. #whosgonnawin has the lowest certainty value due to many low certainty signal words. This figure appears in Senaratne et al. (2016, under review). 109
4.13 Visualisation of the top 6 trajectories ranked left to right by high topic diversity and high structure linearity, low sentiment variance, and low speed variance. These trajectories further indicate a cyclic structure. 111
4.14 Visualisation of the #fml trajectory. 112
4.15 Visualisation of the #fml sub-trajectory with refined DBSCAN parameters to look at an even more dense area. This trajectory indicates one particular tweeter's negative experience that keeps evolving over the course of time. The tweets contained in each cluster are indicated next to each cluster. 113

4.16 An excerpt of MovingOnTwitter overview. (a) drop down menu of the list of hastags in the data, (b) Tweets pertaining to the selected hashtag are visualised on the map, (c) parameterisation, (d) parallel coordinates plot for analysing trajectory characteristic features, (e) trajectory sentiment analysis. This figure appears in Senaratne et al. (2016, under review). 115
4.17 The list of hashtag-based trajectories sorted according to the parameters. The statistical values of the trajectory features are indicated in a heat map visualisation. The unusual high avg. cluster distance for non-sports related hashtags is highlighted in red. This figure appears in Senaratne et al. (2016, under review). 117
4.18 Parallel coordinates visualisation helps to extract the top ranked hashtag-based trajectories based on a lower cluster distance. The vertical axes of the plot further represents the derived trajectory characteristics. The high sentiment variance for #faceofmlb is highlighted in red. This figure appears in Senaratne et al. (2016, under review). . 118
4.19 Cluster boxplots for #faceofmlb. The two similar structures are evidently resembling the two tiered voting sessions during the face of MLB contest. This figure appears in Senaratne et al. (2016, under review). 119
4.20 #faceofmlb conversation movements at the beginning of the contest. Tweeters are talking about two specific players representing the Cincinnati Reds and Seattle Mariners teams. This figure appears in Senaratne et al. (2016, under review). 119
4.21 Content and sentiment analysis for the tweets discussing the two players, Joey Votto and Felix Hernandez. This figure appears in Senaratne et al. (2016, under review). 120
4.22 Credibility of clusters mapped to the length of the bars (shorter bar indicates lower credibility). Cluster with lower average credibility also indicates that tweeters have made many fake profiles. This figure appears in Senaratne et al. (2016, under review). 121

5.1 Trajectory of a selected user with antenna locations taken as the consecutive user locations. This figure appears in Senaratne et al. (2017b). 131
5.2 Trajectory of a selected user with 500m radius cluster centroids taken as the consecutive user locations. 132

5.3 Trajectory of a selected user with 2000m radius cluster centroids taken as the consecutive user locations. 132
5.4 Trajectory of a selected user with 4000m radius cluster centroids taken as the consecutive user locations. 133
5.5 Example of two edges (a-d and d-g) of a trajectory segment passing through several locations (b,c,e,f). This figure appears in Senaratne et al. (2017b). 134
5.6 Spatial similarity matrix of users. (a) The matrix overview facilitates the analysis of unique movements which are represented through blank lines. The larger the circular gradient, the higher is the similarity. (b) Legend shows the scaling of the circular gradient. Specific user similarities can be analysed by zooming in. (c) Each matrix cell can be clicked to visualise the movement profile of the corresponding user pair on the map view. The movement profiles of the selected User 30 is shown in red and User 16 is shown in blue. The intersection of both user movements with a spatial similarity value of 0.63 is shown in green. This figure appears in Senaratne et al. (2017b). 135
5.7 An excerpt of the spatial similarity (circular gradient) and temporal similarity (opacity) matrix of users. The highlighted cell depicts a user pair with high spatial as well as temporal similarity. The diagonal cells only reflect the self-similarity, and therefore ignored in the analysis. This figure appears in Senaratne et al. (2017b). 136
5.8 User classification by home and work areas. The tree node areas are coloured according to the home (Green)and work (Red) area ratio. This figure appears in Senaratne et al. (2017b). 138
5.9 Origin-destination analysis of aggregated user trajectories. (a) Each row and column represents an origin and destination respectively for the aggregated trajectories. The colour scale represents low (lighter) to high (darker) no.of aggregated trajectories, (b) The selected cells in the matrix are coordinated with the rasterized map, as shown in the green (origin) and red (destination) spatial segments. This figure appears in Senaratne et al. (2017b). 139

5.10 Component (a) selects the number of records as the attribute for analysis, (b) chooses an antenna from the table for analysis, and (c) shows a raster map of Santiago created with a user-defined cell size, where each cell indicates the aggregated number of records for the entire duration of time with a heat map visualisation. The colour scale can be interactively adjusted to the users' preference. (d) shows the small multiples view for the daily usage patterns of mobile Internet for the chosen antenna. This figure appears in Senaratne et al. (2017b). . . . 141
5.11 Component (e) shows the time of occurrence of chosen patterns of SOM, and (f) shows the SOM view with a 4 x 4 grid. The daily patterns are used as the input to the SOM algorithm. The clusters depict 16 similar daily patterns. 142
5.12 The space-time path of a selected user within a 24 hour time frame visualised using the space-time cube. This figure appears in Senaratne et al. (2017b). 143
5.13 The space-time path of a selected user with the path segments coloured according to the inter-sample time interval duration. This figure appears in Senaratne et al. (2017b). 144
5.14 Region expansion for defining the spatial extents within uncertain markers. A hexagon is taken as an example region which is indicated in dark green. At each edge of the hexagonal region the spatial extents are indicated with the dashed circles. The hull is indicated with the light green polygon which therefore is the spatial extent for uncertain markers at time t. 145
5.15 Potential path area for uncertain markers and their corresponding space-time prism representations. 146
5.16 Potential path area for two uncertain path segments and their corresponding space-time prism representations. The maximum velocity is assumed to be 50 km/h in this example. This figure appears in Senaratne et al. (2017b). 147
5.17 Positional uncertainty reduction of space-time path segments of users through (left) identifying the antenna jump locations and (right)the interpolation of antenna jump locations. The maximum time duration is (a) 3 minutes, (b) 6 minutes, (c) 15 minutes, (d) 20 minutes. This figure appears in Senaratne et al. (2017b). 149

6.1 The visual variables introduced by Bertin (1983) and MacEachren (1992). 153
6.2 The smart grid network of transformer stations in Germany. Red nodes indicate the transformer stations where measurement data is unavailable, and the green nodes indicate the transformer stations where the measurement data is available. 156
6.3 Transformer stations (rectangles) are connected via power lines and are also connected to the communication infrastructure (triangles), which transfers the information to the central control room. The transmission range of the mobile stations is visualised as concentric circles. While gray indicates normal operation mode, the yellow elements on the screen reveal a severe situation. High deviations in voltage cascaded from the energy grid into the mobile grid due to failures of the power supply. This figure is taken from Mittelstädt et al. (2013). 157
6.4 Sampling method (left) and MCS method (right) to assess power uncertainty at the Helgenreute transformer station. Low to high Purple colour saturation indicates the high and low uncertainties. This figure appeared in Senaratne et al. (2014b). 159
6.5 Candidates for modular power uncertainty visualisation. The inner rectangle of the glyph indicates the modular uncertainty. Uncertainty increases from left to right. This figure appeared in Senaratne et al. (2014b). 160
6.6 Candidates for aggregated power uncertainty visualisation. The outer boarder of the glyph indicates the aggregated uncertainty. Uncertainty increases from left to right. This figure appeared in Senaratne et al. (2014b). 160
6.7 The three-tiered work flow of the pilot study. 161
6.8 Example of the randomised uncertainty depictions of the modular power uncertainty. 162
6.9 Results of the pilot study. 163
6.10 Four-tiered work flow of the main usability study. 164
6.11 The combinations of the visual metaphors for modular uncertainty (shown in columns- M in the matrix), and aggregated uncertainty (shown in rows- A in the matrix). The low, middle, high uncertainty values for modular uncertainty are shown from left to right of the columns, and the low, middle, high uncertainty values for aggregated uncertainty are shown from top to bottom of each row. 166

6.12 An excerpt of the online study that introduces the context and the two types of uncertainty and their respective visualisations to the participants. 167
6.13 An excerpt of the online study that instructs the participants to learn the individual visual metaphors. (a) noise metaphor depicting the three scales of aggregated uncertainty. (b) icon metaphor depicting the three scales of modular uncertainty. 167
6.14 An excerpt of the online study that asks the participant to find the transformer station with middle aggregated power uncertainty and high modular power uncertainty. 'Current' here means the present modular uncertainty. 168
6.15 The participants' preference rating on the individual uncertainty visualisations. (a) preference rating for modular uncertainty visualisations - Icons and Transparency had the highest ratings. (b) preference rating for aggregated uncertainty visualisations - Transparency had the highest ratings. 172
6.16 The aggregated power uncertainty and the modular uncertainty at the different transformer stations depicted through Transparency / Transparency visualisation combination. 173
6.17 The aggregated power uncertainty and the modular uncertainty at the different transformer stations depicted through Transparency / Icons visualisation combination. 174
6.18 An implementation of the approach. On the left panel the user can choose the values for the two different types of uncertainties from a scale between 0 and 100, as well as the alarm levels they want to see for a chosen transformer station. The right side panel allows the user to choose the properties in terms of the map view, user- adjustable data or upload the dataset, as well as the visual variables for the aggregated power uncertainty and modular power uncertainty. 175

7.1 Summary of contributions. 178

List of Tables

2.1 Classification of the reviewed papers according to the uncertainty measures and indicators. $\star$ = map-based, $\bullet$ = image-based, $\diamond$ = text-based, and $\bowtie$= all types of VGI. 37

2.2 Uncertainty measures/indicators classified according to type of method to assess them. Methods are grouped in geographic, social, crowdsourced (abbrv. 'C.'), and the newly found data mining approaches. Type of VGI indicated as: $\star$ = map-based, $\bullet$ = image-based, and $\diamond$ = text-based. 52

3.1 The categories of images within the sample dataset falling into correct/incorrect geotagging and labelling. 77

3.2 The statistics of each metadata feature for image categories a, b, c, and d. 77

4.1 Uncertainty signal words and their corresponding scores. 108

4.2 Credibility features used in MovingOnTwitter2. 109

4.3 The top five hashtag conversations 112

6.1 Comparison results of Sampling and Monte Carlo Simulation (MCS) methods for the three transformer stations. 158

6.2 Significance analysis for the *performance* of participants within the visualisation combinations using the Wilcoxon matched pairs signed rank test. 169

6.3 Excerpt of two participants' bias for the different combinations of visualisations. 171

6.4 Significance analysis for the *bias* of visualisation combinations using the Wilcoxon matched pairs signed rank test. 171

6.5 Significance analysis for the *preference* of visualisation combinations using the Wilcoxon matched pairs signed rank test. 172

Bibliography

N. Adrienko and G. Adrienko, "Spatial generalization and aggregation of massive movement data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 2, pp. 205–219, 2011.

J. C. Aerts, K. C. Clarke, and A. D. Keuper, "Testing popular visualization techniques for representing model uncertainty," *Cartography and Geographic Information Science*, vol. 30, no. 3, pp. 249–261, 2003.

E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the 2008 International Conference on Web Search and Data Mining, 11-12 February.* ACM, New York, NY, USA, 2008, pp. 183–194.

C. A. Agostini and P. Brown, "Geographic income inequality in chile," *Revista de Analisis Economico*, vol. 22, no. 1, 2007.

M. Agrawala, W. Li, and F. Berthouzoz, "Design principles for visual communication," *Communications of the ACM*, vol. 54, no. 4, pp. 60–69, 2011.

M. Al-Bakri and D. Fairbairn, "Assessing the accuracy of crowdsourced data and its integration with official spatial datasets," in *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, P. F. N.J. Tate, Ed., 2010, pp. 317–320.

A. L. Ali, F. Schmid, R. Al-salman, and T. Kauppinen, "Ambiguity and plausibility: Managing classification quality in Volunteered Geographic Information," in *Proceedings of the 22nd International Conference on Geographic Information Systems 4-7th November 2014.* ACM, New York, NY, USA, 2014.

J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and*

development in information retrieval, ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 37–45. [Online]. Available: http://doi.acm.org/10.1145/290941.290954

R. Allendes Osorio and K. W. Brodlie, "Contouring with uncertainty," *Theory and Practice of Computer Graphics 2008. Proceedings.*, pp. 59–66, 2008.

M. Ames and M. Naaman, "Why we tag: motivations for annotation in mobile and online media," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 971–980.

G. Andrienko and N. Andrienko, "A general framework for using aggregation in visual exploration of movement data," *The Cartographic Journal*, vol. 47, no. 1, pp. 22–40, 2010.

G. Andrienko, N. Andrienko, P. Bak, S. Kisilevich, and D. Keim, "Analysis of community-contributed space-and time-referenced data (example of flickr and panoramio photos)," in *IEEE Symposium on Visual Analytics Science and Technology (VAST) 12-13th October 2009*, IEEE. IEEE, New Jersey, USA, 2009, pp. 213–214.

G. Andrienko, N. Andrienko, S. Bremm, T. Schreck, T. Von Landesberger, P. Bak, and D. Keim, "Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns," in *Computer Graphics Forum*, vol. 29, no. 3. Wiley Online Library, 2010, pp. 913–922.

G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel, *Visual analytics of movement*. Springer, 2013.

G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom, "Thematic patterns in georeferenced tweets through space-time visual analytics," *Computing in Science and Engineering*, vol. 15, no. 3, pp. 72–82, 2013.

G. L. Andrienko, N. V. Andrienko, J. Dykes, S. I. Fabrikant, and M. Wachowicz, "Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research." *Information Visualization*, vol. 7, no. 3-4, pp. 173–180, 2008.

N. Andrienko and G. Andrienko, "Designing visual analytics methods for massive collections of movement data," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 42, no. 2, pp. 117–138, 2007.

N. Andrienko, G. Andrienko, N. Pelekis, and S. Spaccapietra, "Basic concepts of movement data," in *Mobility, Data Mining and Privacy*. Springer, 2008, pp. 15–38.

V. Antoniou and A. Skopeliti, "Measures and indicators of vgi quality: An overview," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, pp. 345–351, 2015.

V. Antoniou, J. Morley, and M. Haklay, "Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon," *Geomatica*, vol. 64, no. 1, pp. 99–110, 2010.

J. J. Arsanjani, P. Mooney, A. Zipf, and A. Schauss, "Quality assessment of the contributed land use information from openstreetmap versus authoritative datasets," in *OpenStreetMap in GIScience*. Springer, 2015, pp. 37–58.

A. Ather, "A quality analysis of openstreetmap data," Master's thesis, University College London, UK, 2009.

C. Barron, P. Neis, and A. Zipf, "A comprehensive framework for intrinsic OpenStreetMap quality analysis," *Transactions in GIS*, vol. 18, no. 6, 2014.

M. K. Beard, B. P. Buttenfield, and S. B. Clapham, *NCGIA Research Initiative 7: Visualization of Spatial Data Quality: Scientific Report for the Specialist Meeting 8-12 June 1991, Castine, Maine*. National Center for Geographic Information and Analysis, 1991.

C. Becker and C. Bizer. (2009) Flickr wrappr: Precise photo association. [Online]. Available: http://www4.wiwiss.fu-berlin.de/flickrwrappr

H. Becker, M. Naaman, and L. Gravano, "Selecting quality twitter content for events." *ICWSM*, vol. 11, 2011.

M. Behrisch, F. Korkmaz, L. Shao, and T. Schreck, "Feedback-driven interactive exploration of large multidimensional data supported by visual classifier," in *Proc. IEEE Conference on Visual Analytics Science and Technology*, 2014, pp. 43–52.

J. Bertin, *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin press, 1983.

M. Bishr and K. Janowicz, "Can we trust information? - the case of Volunteered Geographic Information," in *Proceedings of Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium, 20th September,*

2010, A. Devaraju, A. Llaves, P. Maué, and C. Kessler, Eds., vol. 640. CEUR-WS.org, 2010.

M. Bishr and W. Kuhn, "Geospatial information bottom-up: A matter of trust and semantics," in *The European information society*. Springer, 2007, pp. 365–387.

V. V. Bochkarev, A. V. Shevlyakova, and V. D. Solovyev, "Average word length dynamics as indicator of cultural changes in society," *Social Evolution & History*, vol. 14, no. 2, pp. 153–175, 2015. [Online]. Available: https://arxiv.org/abs/1208.6109

G. Bordogna, P. Carrara, L. Criscuolo, M. Pepe, and A. Rampini, "A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science," *Information Sciences*, vol. 258, pp. 312–327, 2014.

R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen, "Glyph-based visualization: Foundations, design guidelines, techniques and applications," *Eurographics State of the Art Reports*, pp. 39–63, 2013.

H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, and T. Ertl, "Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2022–2031, 2013.

C. Brando and B. Bucher, "Quality in user generated spatial content: A matter of specifications," in *Proceedings of the 13th AGILE international conference on geographic information science, 11-14th May, 2010*, H. P. M. Painho, M.Y. Santos, Ed. Springer Verlag, 2010, pp. 11–14.

K. Brodlie, R. A. Osorio, and A. Lopes, "A review of uncertainty in data visualization," in *Expanding the Frontiers of Visual Analytics and Visualization*. Springer, 2012, pp. 81–109.

A. Bröring, "Live and web-based parcel monitoring with low-cost sensors," in *Online Proceedings of the 16th AGILE International Conference on Geographic Information Science. Leuven, Belgium. 15.-17. May 2013*, D. Vandenbroucke, B. Bucher, and J. Crompvoets, Eds., 2013.

C. G. Bucher, "Adaptive sampling - an iterative fast monte carlo procedure," *Structural Safety*, vol. 5, no. 2, pp. 119–126, 1988.

J. Bustos-Jiménez, G. Del Canto, S. Pereira, F. Lalanne, J. Piquer, G. Hourton, A. Cádiz, and V. Ramiro, "How adkintunmobile measured the world," in *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, 2013, pp. 1457–1462.

B. Buttenfield and M. K. Beard, "Graphical and geographical components of data quality," *Visualization in geographic information systems*, pp. 150–157, 1994.

B. Buttenfield and R. Weibel, "Visualizing the quality of cartographic data," in *Third International Geographic Information Systems Symposium (GIS/LIS 88), San Antonio, Texas*, 1988.

H. Cai and Y. Lin, "Tuning trust using cognitive cues for better human-machine collaboration," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 54, no. 28. SAGE Publications, 2010, pp. 2437–2441.

F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in rome," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, 2011.

F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 0036–44, 2011.

M. A. Cameron, R. Power, B. Robinson, and J. Yin, "Emergency situation awareness from twitter for crisis management," in *Proceedings of the 21st international conference companion on World Wide Web*. ACM, 2012, pp. 695–698.

C. Campbell, L. F. Pitt, M. Parent, and P. R. Berthon, "Understanding consumer conversations around ads in a web 2.0 world," *Journal of Advertising*, vol. 40, no. 1, pp. 87–102, 2011.

R. Canavosio-Zuzelski, P. Agouris, and P. Doucette, "A photogrammetric approach for assessing positional accuracy of openstreetmap roads," *ISPRS International Journal of Geo-Information*, vol. 2, no. 2, pp. 276–301, 2013.

K. R. Canini, B. Suh, and P. L. Pirolli, "Finding credible information sources in social networks based on content and social structure," in *Proceedings of Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 9-11 October, 2011*. IEEE, NJ, USA, 2011, pp. 1–8.

F. Castanedo, "A review of data fusion techniques," *The Scientific World Journal*, vol. 2013, 2013.

C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 675–684.

A. Cedilnik and P. Rheingans, "Procedural annotation of uncertain information," in *Visualization 2000. Proceedings*. IEEE, 2000, pp. 77–84.

S. Chainey, L. Tompson, and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," *Security Journal*, vol. 21, no. 1-2, pp. 4–28, 2008.

C. Chatfield, "Model uncertainty," *Encyclopedia of Environmetrics*, 2006.

R. Chunara, J. R. Andrews, and J. S. Brownstein, "Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak," *The American Journal of Tropical Medicine and Hygiene*, vol. 86, no. 1, pp. 39–45, 2012. [Online]. Available: http://www.biomedsearch.com/nih/Social-news-media-enable-estimation/22232449.html

B. Ciepłuch, R. Jacob, P. Mooney, and A. Winstanley, "Comparison of the accuracy of openstreetmap for ireland with google maps and bing maps," in *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*, P. F. N.J. Tate, Ed. University of Leicester, UK, 2010, pp. 337– 340.

B. Ciepłuch, P. Mooney, R. Jacob, J. Zheng, and A. Winstanely, "Assessing the quality of open spatial data for mobile location-based services research and applications," *Archives of photogrammetry, cartography and remote sensing, ISSN 2083-2214*, vol. 22, pp. 105–116, 2011.

D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649.

W. S. Cleveland and R. McGill, "An experiment in graphical perception," *International Journal of Man-Machine Studies*, vol. 25, no. 5, pp. 491–500, 1986.

D. C. Cliburn, J. J. Feddema, J. R. Miller, and T. A. Slocum, "Design and evaluation of a decision support system in a water balance application," *Computers & Graphics*, vol. 26, no. 6, pp. 931–949, 2002.

M. Codescu, G. Horsinka, O. Kutz, T. Mossakowski, and R. Rau, "Osmonto-an ontology of openstreetmap tags," *State of the map Europe (SOTM-EU) 2011*, 2011.

D. J. Coleman, Y. Georgiadou, J. Labonte *et al.*, "Volunteered Geographic Information: the nature and motivation of produsers," *International Journal of Spatial Data Infrastructures Research*, vol. 4, no. 1, pp. 332–358, 2009.

P. Corcoran, P. Mooney, and A. Winstanley, "Topological consistent generalization of openstreetmap," in *Proceedings of the GIS Research UK 18th Annual Conference GISRUK 2010*. Maynooth University, 2010, pp. 353–357.

C. Correa, Y.-H. Chan, and K.-L. Ma, "A framework for uncertainty-aware visual analytics," in *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*. IEEE, 2009, pp. 51–58.

M. Craglia, F. Ostermann, and L. Spinsanti, "Digital earth from vision to practice: making sense of citizen-generated content," *International Journal of Digital Earth*, vol. 5, no. 5, pp. 398–416, 2012.

D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 761–770.

A. C. Cullen and H. C. Frey, *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*. Springer Science & Business Media, 1999.

A. C. Da Silva and S.-T. Wu, "Consistent handling of linear features in polyline simplification," in *Advances in Geoinformatics*, M. M. C.A. Davis Jr, Ed. Springer, Switzerland, 2007, pp. 1–17.

L. De Floriani, P. Marzano, and E. Puppo, "Line-of-sight communication on terrain models," *International Journal of Geographical Information Systems*, vol. 8, no. 4, pp. 329–342, 1994.

B. De Longueville, N. Ostländer, and C. Keskitalo, "Addressing vagueness in volunteered geographic information (vgi)–a case study," *International Journal of Spatial Data Infrastructures Research*, vol. 5, pp. 1725–0463, 2010.

G. De Tré, A. Bronselaer, T. Matthé, N. Van de Weghe, and P. De Maeyer, "Consistently handling geographical user data," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems, Applications, 28th June - 2nd July, 2010*, F. H. E. Huellermeier, R. Kruse, Ed. Springer, 2010, pp. 85–94.

P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*. Prentice Hall, 1982.

S. Dodge, P. Laube, and R. Weibel, "Movement similarity assessment using symbolic representation of trajectories," *International Journal of Geographical Information Science*, vol. 26, no. 9, pp. 1563–1588, 2012.

I. Drecki, *Spatial data quality*. Taylor & Francis New York, NY, 2002, ch. Visualisation of uncertainty in geographical data, pp. 140–159.

G. Dutton, "Handling positional uncertainty in spatial databases," in *Proceedings of the 5th International Symposium on Spatial Data Handling*, 1992, pp. 460–469.

C. R. Ehlschlaeger, A. M. Shortridge, and M. F. Goodchild, "Visualizing spatial data uncertainty using animation," *Computers & Geosciences*, vol. 23, no. 4, pp. 387–395, 1997.

A. Endert, M. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews, "The human is the loop: new directions for visual analytics," *Journal of Intelligent Information Systems*, pp. 1–25, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10844-014-0304-9

M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

B. J. Evans, "Dynamic display of spatial data-reliability: Does it benefit the map user?" *Computers & Geosciences*, vol. 23, no. 4, pp. 409–422, 1997.

C. Faloutsos, H. Jagadish, A. O. Mendelzon, and T. Milo, "A signature technique for similarity-based queries," in *Compression and Complexity of Sequences 1997. Proceedings.* IEEE, 1997, pp. 2–20.

H. Fan, A. Zipf, Q. Fu, and P. Neis, "Quality assessment for building footprints data on openstreetmap," *International Journal of Geographical Information Science*, vol. 28, no. 4, pp. 700–719, 2014.

E. Fauerbach, R. Edsall, D. Barnes, and A. MacEachren, "Visualization of uncertainty in meteorological forecast models," in *Proceedings of the International Symposium on Spatial Data Handling, Delft, The Netherlands, August*, 1996, pp. 12–16.

C. Fernandez, E. Ley, and M. F. Steel, "Model uncertainty in cross-country growth regressions," *Journal of applied Econometrics*, vol. 16, no. 5, pp. 563–576, 2001.

N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2149–2158, 2013.

J. Fiaidhi, S. Mohammed, A. Islam, S. Fong, T.-h. Kim, S. Sharma, and B. Bhushan, "Developing a hierarchical multi-label classifier for twitter trending topics," *International Journal of u-and e-Service, Science and Technology*, vol. 6, no. 3, pp. 1–12, 2013.

B. Fiessler, R. Rackwitz, and H.-J. Neumann, "Quadratic limit states in structural reliability," *Journal of the Engineering Mechanics Division*, vol. 105, no. 4, pp. 661–676, 1979.

D. Fisher, S. M. Drucker, and A. C. König, "Exploratory visualization involving incremental, approximate database queries and uncertainty," *IEEE Computer Graphics and Applications*, vol. 32, no. 4, pp. 55–62, 2012. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/MCG.2012.48

P. Fisher, "Animation and sound for the visualization of uncertain spatial information," *Visualisation in Geographical Information Systems, John Wiley & Sons, Chichester*, pp. 181–185, 1994.

P. F. Fisher, "First experiments in viewshed uncertainty: the accuracy of the viewshed area," *Photogrammetric engineering and remote sensing*, vol. 57, no. 10, pp. 1321–1327, 1991.

——, "Algorithm and implementation uncertainty in viewshed analysis," *International Journal of Geographical Information Science*, vol. 7, no. 4, pp. 331–347, 1993.

——, "Visualizing uncertainty in soil maps by animation," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 30, no. 2, pp. 20–27, 1993.

——, "Extending the applicability of viewsheds in landscape planning," *Photogrammetric Engineering and Remote Sensing*, vol. 62, no. 11, pp. 1297–1302, 1996.

A. Flanagin and M. Metzger, "The credibility of Volunteered Geographic Information," *GeoJournal*, vol. 72, no. 3, pp. 137–148, 2008.

G. Foody, L. See, S. Fritz, M. Van der Velde, C. Perger, C. Schill, D. Boyd, and A. Comber, "Accurate attribute mapping from volunteered geographic information: issues of volunteer quantity and quality," *The Cartographic Journal*, vol. 52, no. 4, 2014.

M. Forghani and M. R. Delavar, "A quality study of the openstreetmap dataset for tehran," *ISPRS International Journal of Geo-Information*, vol. 3, no. 2, pp. 750–763, 2014.

J. Frew, "Provenance and volunteered geographic information," University of California, Santa Barbara, Tech. Rep., 2007. [Online]. Available: http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Frew_paper.pdf

G. Friedland, J. Choi, H. Lei, and A. Janin, "Multimodal location estimation on flickr videos," in *Proceedings of the 3rd ACM SIGMM international workshop on Social media*. ACM, 2011, pp. 23–28.

G. Fuchs, N. Andrienko, G. Andrienko, S. Bothe, and H. Stange, "Tracing the german centennial flood in the stream of tweets: first lessons learned," in *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, 5-8, November, 2013*, A. V. D. Pfoser, Ed., ACM. ACM, New York, NY, USA, 2013, pp. 31–38.

S. Gao, Y. Liu, Y. Wang, and X. Ma, "Discovering spatial interaction communities from mobile phone data," *Transactions in GIS*, vol. 17, no. 3, pp. 463–481, 2013.

L. E. Gerharz and E. J. Pebesma, "Usability of interactive and non-interactive visualisation of uncertain geospatial information," in *Proceedings of Geoinformatik*, 2009, pp. 223–230.

L. E. Gerharz, C. Autermann, H. Hopmann, C. Stasch, and E. Pebesma, "Uncertainty visualisation in the model web," in *EGU General Assembly Conference Abstracts*, vol. 14, 2012, p. 4421.

F. Girardin, F. Calabrese, F. Dal Fiore, C. Ratti, and J. Blat, "Digital footprinting: Uncovering tourists with user-generated content," *IEEE Pervasive computing*, vol. 7, no. 4, pp. 36–43, 2008.

J.-F. Girres and G. Touya, "Quality assessment of the french OpenStreetMap dataset," *Transactions in GIS*, vol. 14, no. 4, pp. 435–459, 2010.

S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of information science*, vol. 32, no. 2, pp. 198–208, 2006.

M. Goodchild, "Neogeography and the nature of geographic expertise," *Journal of location based services*, vol. 3, no. 2, pp. 82–96, 2009.

M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.

M. F. Goodchild and L. Li, "Assuring the quality of Volunteered Geographic Information," *Spatial statistics*, vol. 1, pp. 110–120, 2012.

H. Griethe and H. Schumann, "The visualization of uncertain data: Methods and problems." in *SimVis*, 2006, pp. 143–156.

J. B. Guinée, "Handbook on life cycle assessment operational guide to the iso standards," *The international journal of life cycle assessment*, vol. 7, no. 5, pp. 311–313, 2002.

M. Gupta, P. Zhao, and J. Han, "Evaluating event credibility on twitter," in *SIAM International Conference on Data Mining (SDM 2012)*, 2012, pp. 153–164.

V. Guralnik and J. Srivastava, "Event detection from time series data," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '99. New York, NY, USA: ACM, 1999, pp. 33–42. [Online]. Available: http://doi.acm.org/10.1145/312129.312190

R. B. Haber and D. A. McNabb, "Visualization idioms: A conceptual model for scientific visualization systems," *Visualization in scientific computing*, vol. 74, p. 93, 1990.

T. Hägerstraand, "What about people in regional science?" *Papers in regional science*, vol. 24, no. 1, pp. 7–24, 1970.

M. Haklay, "How good is Volunteered Geographic Information? a comparative study of OpenStreetMap and Ordnance Survey datasets," *Environment and planning. B, Planning & design*, vol. 37, no. 4, p. 682, 2010.

M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," *Computer Science Working Papers, University of Waikato*, 2000. [Online]. Available: http://researchcommons.waikato.ac.nz/handle/10289/1024

O. Hartig, "Provenance information in the web of data." *LDOW*, vol. 538, 2009.

J. Hartigan, *Clustering Aloritms*. John Wiley, New York, 1975.

D. Hasan Dalip, M. André Gonçalves, M. Cristo, and P. Calado, "Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia," in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, 14-19 June, 2009*. ACM, New York, NY, USA, 2009, pp. 295–304.

P. Hashemi and R. A. Abbaspour, "Assessment of logical consistency in openstreetmap based on the spatial similarity concept," in *OpenStreetMap in GIScience*. Springer, 2015, pp. 19–36.

R. Hecht, C. Kunze, and S. Hahmann, "Measuring completeness of building footprints in openstreetmap over space and time," *ISPRS International Journal of Geo-Information*, vol. 2, no. 4, pp. 1066–1091, 2013.

M. Helbich, C. Amelunxen, P. Neis, and A. Zipf, "Comparative spatial analysis of positional accuracy of openstreetmap and proprietary geodata," *Proceedings of GI_Forum, 3-6 July, 2012*, pp. 24–33, 2012.

J. C. Helton, J. D. Johnson, C. J. Sallaberry, and C. B. Storlie, "Survey of sampling-based methods for uncertainty and sensitivity analysis," *Reliability Engineering & System Safety*, vol. 91, no. 10, pp. 1175–1209, 2006.

T. Hengl, "Visualisation of uncertainty using the hsi colour model: computations with colours," in *Proceedings of the 7th International Conference on GeoComputation*, 2003, pp. 8–17.

T. Hengl and N. Toomanian, "Maps are not what they seem: representing uncertainty in soil-property maps," in *Proc. Accuracy*, 2006, pp. 805–813.

T. Hengl, D. J. J. Walvoort, and A. Brown, "Pixel and colour mixture: Gis techniques for visualisation of fuzziness and uncertainty of natural resource inventories," in *Proc. Accuracy*, 2002, pp. 300–308.

L. Hollenstein and R. Purves, "Exploring place through user-generated content: Using flickr tags to describe city cores," *Journal of Spatial Information Science*, no. 1, pp. 21–48, 2010.

C. Hovland, I. Janis, and H. Kelley, *Communication and persuasion; psychological studies of opinion change*. Yale University Press, 1953.

D. Howard and A. M. MacEachren, "Interface design for geographic visualization: Tools for representing reliability," *Cartography and Geographic Information Systems*, vol. 23, no. 2, pp. 59–77, 1996.

D. Hoyle, *ISO 9000: quality systems handbook*. Butterworth and Heinemann, Oxford, UK, 2001.

K. L. Huang, S. S. Kanhere, and W. Hu, "Are you contributing trustworthy data?: the case for a reputation system in participatory sensing," in *Proceedings of the 13th ACM international conference on Modeling, analysis, and simulation of wireless and mobile systems, 17-21 October, 2010*. ACM, New York, NY, USA, 2010, pp. 14–22.

B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," *SSRN*, vol. 14, no. 1, 2008. [Online]. Available: http://dx.doi.org/10.2139/ssrn.1313405

J. Hultman and O. Wärneryd, "What about nature?" *Human Dimensions of Global Environmental Change*, pp. 14–20, 2001.

S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Pervasive computing*. Springer, 2011, pp. 133–151.

D. Jäckle, H. Senaratne, J. Buchmüller, and D. A. Keim, "Integrated Spatial Uncertainty Visualization using Off-screen Aggregation," in *EuroVA International Workshop on Visual Analytics*, 2015.

S. P. Jackson, W. Mullen, P. Agouris, A. Crooks, A. Croitoru, and A. Stefanidis, "Assessing completeness and spatial error of features in volunteered geographic information," *ISPRS International Journal of Geo-Information*, vol. 2, no. 2, pp. 507–530, 2013.

N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless, "Geolocating static cameras," in *IEEE 11th International Conference on Computer Vision (ICCV), 14-20 October, 2007*. IEEE, NJ, USA, 2007, pp. 1–6.

P. Jankowski, N. Andrienko, G. Andrienko, and S. Kisilevich, "Discovering landmark preferences and movement patterns from photo postings," *Transactions in GIS*, vol. 14, no. 6, pp. 833–852, 2010.

JCGM, "Evaluation of measurement data - guide to the expression of uncertainty in measurement, jcgm 100," Joint Committee for Guides in Metrology, 2008, joint document by JCGM member organisations: BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML. [Online]. Available: http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

B. Kang, J. O'Donovan, and T. Höllerer, "Modeling topic specific credibility on twitter," in *Proceedings of the ACM international conference on Intelligent User Interfaces, 14-17 February, 2012*. ACM, NJ, USA, 2012, pp. 179–188.

J. Kardos, A. Moore, and G. L. Benwell, "The trustree for the visualisation of attribute and spatial uncertainty: usability assessments," in *Proceedings of the 16th Annual Colloquium of the Spatial Information Research Centre (SIRC 2004: A Spatio-temporal Workshop)*, 2004, pp. 39–52.

J. Kardos, A. Moore, and G. Benwell, "Expressing attribute uncertainty in spatial data using blinking regions," in *Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Lisbon, Portugal*, 2006, pp. 814–824.

J. Kaye, A. Lillie, D. Jagdish, J. Walkup, R. Parada, and K. Mori, "Nokia internet pulse: a long term deployment and iteration of a twitter visualization," in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2012, pp. 829–844.

D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," in *Information visualization*. Springer, 2008, pp. 154–175.

M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

S. Kent, "Words of estimative probability," *Studies in Intelligence*, vol. 8, no. 4, pp. 49–65, 1964.

C. Keßler and R. T. A. de Groot, "Trust as a proxy measure for the quality of Volunteered Geographic Information in the case of OpenStreetMap," in *Geographic Information Science at the Heart of Europe*, J. C. D. Vandenbroucke, B. Bucher, Ed. Springer, Switzerland, 2013, pp. 21–37.

C. Keßler, P. Maué, J. T. Heuer, and T. Bartoschek, "Bottom-up gazetteers: Learning from the implicit semantics of geotags," in *GeoSpatial semantics*, S. L. K. Janowicz, M. Raubal, Ed. Springer, Berlin Heidelberg, 2009, pp. 83–102.

C. Keßler, J. Trame, and T. Kauppinen, "Tracking editing processes in volunteered geographic information: The case of openstreetmap," in *Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory, 12-16 September, 2011*, M. Duckham, A. Galton, and M. Worboys, Eds., 2011.

D. Kidner, A. Sparkes, and M. Dorey, "Gis and wind farm planning," in *Geographical information and planning*. Springer, 1999, pp. 203–223.

Y.-H. Kim, S. Rana, and S. Wise, "Exploring multiple viewshed analysis using terrain features and optimisation techniques," *Computers & Geosciences*, vol. 30, no. 9, pp. 1019–1032, 2004.

C. Kinkeldey, A. M. MacEachren, and J. Schiewe, "How to assess visual communication of uncertainty? a systematic review of geospatial uncertainty visualisation user studies," *The Cartographic Journal*, vol. 51, no. 4, pp. 372–386, 2014.

G. J. Klir and M. J. Wierman, *Uncertainty-based information: elements of generalized information theory*. Springer Science & Business Media, 1999, vol. 15.

K. Koffka, *Principles of Gestalt Psychology*. Routledge, 1935.

T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.

T. Koukoletsos, M. Haklay, and C. Ellul, "Assessing data completeness of vgi through an automated matching procedure for linear data," *Transactions in GIS*, vol. 16, no. 4, pp. 477–498, 2012.

O. Kounadi, "Assessing the quality of openstreetmap data," *Thesis (MSc), University College of London Department of Civil, Environmental And Geomatic Engineering*, 2009.

B. Kuijpers and W. Othman, "Modeling uncertainty of moving objects on road networks via space–time prisms," *International Journal of Geographical Information Science*, vol. 23, no. 9, pp. 1095–1117, 2009.

B. Kuijpers, H. J. Miller, T. Neutens, and W. Othman, "Anchor uncertainty and space-time prisms on road networks," *International Journal of Geographical Information Science*, vol. 24, no. 8, pp. 1223–1248, 2010.

V. Kumar and S. Minz, "Feature selection," *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.

I. R. Lake, A. A. Lovett, I. J. Bateman, and I. H. Langford, "Modelling environmental influences on property prices in an urban environment," *Computers, Environment and Urban Systems*, vol. 22, no. 2, pp. 121–136, 1998.

R. Lambiotte, V. D. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren, "Geographical dispersal of mobile communication networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 21, pp. 5317–5325, 2008.

S. H. Lee and B. M. Kwak, "Response surface augmented moment method for efficient reliability analysis," *Structural safety*, vol. 28, no. 3, pp. 261–272, 2006.

S. Lee and W. Chen, "A comparative study of uncertainty propagation methods for black-box-type problems," *Structural and Multidisciplinary Optimization*, vol. 37, no. 3, pp. 239–253, 2009.

V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," in *Soviet physics doklady*, vol. 10, 1966, p. 707.

T. Lindeberg, "Scale invariant feature transform," *Scholarpedia*, vol. 7, no. 5, 2012. [Online]. Available: http://dx.doi.org/10.4249/scholarpedia.10491

H. Liu and H. Motoda, "Feature selection for knowledge discovery and data miningkluwer academic publishers," *Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers*, 1998.

P. Longley, *Geographic information systems and science*. John Wiley & Sons, 2005.

S. J. Luck, S. A. Hillyard, M. Mouloua, and H. L. Hawkins, "Mechanisms of visual–spatial attention: Resource allocation or uncertainty reduction?" *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, no. 3, p. 725, 1996.

L. S. M. Craglia, F. Ostermann, "Digital earth from vision to practice: making sense of citizen-generated content," *International Journal of Digital Earth*, vol. 5, no. 5, pp. 398–416, 2012.

A. M. MacEachren, "Visualizing uncertain information," *Cartographic Perspectives*, vol. 13, no. 13, pp. 10–19, 1992.

A. M. MacEachren and J. H. Ganter, "A pattern identification approach to cartographic visualization," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 27, no. 2, pp. 64–81, 1990.

A. M. MacEachren, C. A. Brewer, and L. W. Pickle, "Visualizing georeferenced data: representing reliability of health statistics," *Environment and Planning A*, vol. 30, no. 9, pp. 1547–1561, 1998.

A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler, "Visualizing geospatial information uncertainty: What we know and what we need to know," *Cartography and Geographic Information Science*, vol. 32, no. 3, pp. 139–160, 2005.

A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "Senseplace2: Geotwitter analytics support for situational awareness," in *IEEE Conference on Visual Analytics Science and Technology (VAST), 23-28 October, 2011*. IEEE, NJ, USA, 2011, pp. 181–190.

A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan, "Visual semiotics & uncertainty visualization: An empirical study," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2496–2505, 2012.

H. O. Madsen, S. Krenk, and N. C. Lind, *Methods of structural safety*. Courier Corporation, 2006.

P. Maué, "Reputation as tool to ensure validity of vgi," in *Proceedings of the VGI specialist meeting, 13-14 December*, 2007.

J. McCoy, K. Johnston, and E. systems research institute, *Using ArcGIS spatial analyst: GIS by ESRI*. Environmental Systems Research Institute, 2001.

R. Melchers, "Importance sampling in structural systems," *Structural safety*, vol. 6, no. 1, pp. 3–10, 1989.

M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we rt?" in *Proceedings of the first workshop on social media analytics, 25-28 July*. ACM, New York, NY, USA, 2010, pp. 71–79.

N. Milholland and E. Pultar, "The san francisco public art map application: using vgi and social media to complement institutional data sources," in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on MapInteraction, 5-8 November*. ACM, New York, NY, USA, 2013, pp. 48–53.

H. J. Miller, "A measurement theory for time geography," *Geographical analysis*, vol. 37, no. 1, pp. 17–45, 2005.

S. Mittelstädt, D. Spretke, D. Thom, D. Jäckle, A. Karsten, and D. A. Keim, "Situational Awareness for Critical Infrastructures and Decision Support," in *In Proceedings of the NATO STO IST-116 Symposium on Visual Analytics*, 2013.

D. R. Montello, M. F. Goodchild, J. Gottsegen, and P. Fohl, "Where's downtown?: Behavioral methods for determining referents of vague spatial queries," *Spatial Cognition & Computation*, vol. 3, no. 2-3, pp. 185–204, 2003.

C. Z. Mooney, *Monte carlo simulation*. Sage Publications, 1997, vol. 116.

P. Mooney and P. Corcoran, "The annotation process in OpenStreetMap," *Transactions in GIS*, vol. 16, no. 4, pp. 561–579, 2012.

M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing?: understanding microblog credibility perceptions," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, 11-15 February.* ACM, New York, NY, USA, 2012, pp. 441–450.

J. L. Morrison, "A theoretical framework for cartographic generalization with the emphasis on the process of symbolization," *International Yearbook of Cartography*, vol. 14, no. 1974, pp. 115–27, 1974.

E. Moxley, J. Kleban, and B. Manjunath, "Spirittagger: a geo-aware tag suggestion tool mined from flickr," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval.* ACM, 2008, pp. 24–30.

L. N. Mummidi and J. Krumm, "Discovering points of interest from users' map annotations," *GeoJournal*, vol. 72, no. 3-4, pp. 215–227, 2008.

T. Munzner, "A nested model for visualization design and validation," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 6, pp. 921–928, 2009.

P. Neis, D. Zielstra, and A. Zipf, "The street network evolution of crowdsourced maps: Openstreetmap in germany 2007–2011," *Future Internet*, vol. 4, no. 1, pp. 1–21, 2011.

J. Nielsen, *Usability engineering.* Elsevier, 1994.

J. Nielsen, T. Clemmensen, and C. Yssing, "Getting access to what goes on in people's heads?: reflections on the think-aloud technique," in *Proceedings of the second Nordic conference on Human-computer interaction.* ACM, 2002, pp. 101–110.

R. O'Connor. (2009, January) Facebook and twitter are reshaping journalism as we know it. [Online]. Available: http://www.alternet.org/story/121211/facebook_and_twitter_are_reshaping_journalism_as_we_know_it

J. O'Donovan, B. Kang, G. Meyer, T. Hollerer, and S. Adalii, "Credibility in context: An analysis of feature distributions in twitter," in *Proceedings of the International Conference on Privacy, Security, Risk and Trust (PASSAT) and International Confernece on Social Computing (SocialCom), 3-5 September.* IEEE, NJ, USA, 2012, pp. 293–301.

C. Olston and J. D. Mackinlay, "Visualizing data with bounded uncertainty," in *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on.* IEEE, 2002, pp. 37–40.

T. O'Reilly, *What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software.* O'Reilly Media, September 2005. [Online]. Available: http://oreilly.com/web2/archive/what-is-web-20.html

F. O. Ostermann and L. Spinsanti, "A conceptual workflow for automatically assessing the quality of volunteered geographic information for crisis management," in *Proceedings of the 14th AGILE Conference on Geographic Information Science*, F. T. S. Geertman, W. Reinhardt, Ed., 2011.

D. O'Sullivan and D. J. Unwin, *Geographic information analysis, Second Edition.* Hoboken, NJ, USA: John Wiley & Sons, 2010, ch. The pitfalls and potential of spatial data, pp. 33–53.

A. Pang, "Visualizing uncertainty in geo-spatial data," in *Proceedings of the Workshop on the Intersections between Geospatial Information and Information Technology*, 2001, pp. 1–14.

A. T. Pang, C. M. Wittenbrink, and S. K. Lodha, "Approaches to uncertainty visualization," *The Visual Computer*, vol. 13, no. 8, pp. 370–390, 1997.

J. P. Payne, "Gis tools for cartographic representation of spatial data uncertainty," Master's thesis, University of Redlands, 2009. [Online]. Available: http://inspire.redlands.edu/gis_gradproj/121

K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

E. J. Pebesma, K. de Jong, and D. Briggs, "Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: an air quality example," *International Journal of Geographical Information Science*, vol. 21, no. 5, pp. 515–527, 2007.

O. Phelan, K. McCarthy, and B. Smyth, "Using twitter to recommend real-time topical news," in *Proceedings of the Third ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2009, pp. 385–388. [Online]. Available: http://doi.acm.org/10.1145/1639714.1639794

A. Popescu, G. Grefenstette, and P. A. Moëllic, "Gazetiki: automatic creation of a geographical gazetteer," in *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, 16-20 June*. ACM, New York, NY, USA, 2008, pp. 85–93.

A.-M. Popescu and M. Pennacchiotti, "Detecting controversial events from twitter," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1873–1876.

K. Poser and D. Dransch, "Volunteered geographic information for disaster management with application to rapid flood damage estimation," *Geomatica*, vol. 64, no. 1, pp. 89–98, 2010.

J.-S. Prassni, T. Ropinski, and K. Hinrichs, "Uncertainty-aware guided volume segmentation," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 16, no. 6, pp. 1358–1365, 2010.

A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *The International Journal of Robotics Research*, 2009.

B. Pross, L. Gerharz, C. Stasch, and E. Pebesma, "Tools for uncertainty propagation in the model web using monte carlo simulation," in *Proceedings of the iEMSs sixth Biennial meeting: Managing resources of a limited planet. International congress on environmental modelling and software (iEMS 2012), international environmental modelling and software society (iEMSs)*, 2012.

P. Ralling, D. Kidner, and A. Ware, "Distributed viewshed analysis for planning application," *Innovations in GIS*, vol. 6, pp. 185–199, 1999.

P. Ranacher and K. Tzavella, "How to compare movement? a review of physical movement similarity measures in geographic information science and beyond," *Cartography and geographic information science*, vol. 41, no. 3, pp. 286–307, 2014.

C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, "Redrawing the map of great britain from a network of human interactions," *PloS one*, vol. 5, no. 12, p. e14248, 2010.

P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman, "Reputation systems," *Communications of the ACM*, vol. 43, no. 12, pp. 45–48, 2000.

D. Reusser, H. Senaratne, L. Gerharz, T. Sterzel, T. Nocke, and M. Wrobel, "User preferences for the presentation of uncertainties on web platforms for climate change information," in *EGU General Assembly Conference Abstracts*, vol. 14, 2012, p. 4148.

C. Rinner, C. Keßler, and S. Andrulis, "The use of web 2.0 concepts to support deliberation in spatial decision-making," *Computers, Environment and Urban Systems*, vol. 32, no. 5, pp. 386–395, 2008.

S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko, "Visually driven analysis of movement data by progressive clustering," *Information Visualization*, vol. 7, no. 3-4, pp. 225–239, 2008.

S. Robinson, M. Jones, J. Williamson, R. Murray-Smith, P. Eslambolchilar, and M. Lindborg, "Navigation your way: from spontaneous independent exploration to dynamic social journeys," *Personal and Ubiquitous Computing*, vol. 16, no. 8, pp. 973–985, 2012.

C. Robusto, "The cosine-haversine formula," *The American Mathematical Monthly*, vol. 64, no. 1, pp. 38–40, 1957.

A. Rottmann, Ó. M. Mozos, C. Stachniss, and W. Burgard, "Semantic place classification of indoor environments with mobile robots using boosting," in *AAAI*, vol. 5, 2005, pp. 1306–1311.

D. Sacha, H. Senaratne, B. C. Kwon, and D. A. Keim, "Uncertainty propagation and trust building in visual analytics," in *IEEE VIS 2014 - Provenance for Sensemaking Workshop (poster paper)*, 2014.

D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, D. Keim *et al.*, "Knowledge generation model for visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1604–1613, 2014.

D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, "The role of uncertainty, awareness, and trust in visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 240–249, 2016.

G. Sagl, M. Loidl, and E. Beinat, "A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic," *ISPRS International Journal of Geo-Information*, vol. 1, no. 3, pp. 256–271, 2012.

H. Saif, M. Fern, Y. He, and H. Alani, "Evaluation datasets for twitter sentiment analysis a survey and a new dataset, the sts-gold," in *1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013*, 2013. [Online]. Available: http://oro.open.ac.uk/id/eprint/40660

T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World Wide Web*. ACM, 2010, pp. 851–860.

H. Samet, "The quadtree and related hierarchical data structures," *ACM Computing Surveys (CSUR)*, vol. 16, no. 2, pp. 187–260, 1984.

F. Šarić, G. Glavaš, M. Karan, J. Šnajder, and B. D. Bašić, "Takelab: Systems for measuring semantic text similarity," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2012, pp. 441–448.

G. S. Schmidt, S.-L. Chen, A. N. Bryden, M. A. Livingston, L. J. Rosenblum, and B. R. Osborn, "Multidimensional visual representations for underwater environmental uncertainty," *Computer Graphics and Applications, IEEE*, vol. 24, no. 5, pp. 56–65, 2004.

S. Schmitz, P. Neis, and A. Zipf, "New applications based on collaborative geodata - the case of routing," in *Proceedings of XXVIII INCA International Congress on Collaborative Mapping and Space Technology, 4-6 November*, 2008.

H. Senaratne and L. Gerharz, "An assessment and categorisation of quantitative uncertainty visualisation methods for geospatial data," in *14th AGILE international conference on geographic information science-advancing geoinformation science for a changing world. AGILE*, 2011.

H. Senaratne, L. Gerharz, E. Pebesma, and A. Schwering, "Usability of spatio-temporal uncertainty visualisation methods," in *Bridging the Geographic Information Sciences*. Springer, 2012, pp. 3–23.

H. Senaratne, A. Bröring, and T. Schreck, "Assessing the credibility of vgi contributors based on metadata and reverse viewshed analysis - an experiment with geotagged flickr images," in *16th AGILE international conference on geographic information science-advancing geoinformation science for a changing world. AGILE*, 2013.

——, "Using reverse viewshed analysis to assess the location correctness of visually generated vgi," *Transactions in GIS*, vol. 17, no. 3, pp. 369–386, 2013.

H. Senaratne, A. Bröring, T. Schreck, and D. Lehle, "Moving on twitter: Using episodic hotspot and drift analysis to detect and characterise spatial trajectories," in *7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN), 4-7 November*, ACM. ACM, New York, NY, USA, 2014, pp. 23–30.

H. Senaratne, S. Mittelstädt, C. Jacob, and T. Schreck, "Uncertainty visualization for crisis management in smart grid environments," in *GIScience 2014: Eighth International Conference on Geographic Information Science*, 2014.

H. Senaratne, D. Lehle, and T. Schreck, "Movingontwitter2: Detection, characterisation, and interest- based ranking of conversation trajectories," *Transactions on Spatial Algorithms and Systems*, 2016, under review.

H. Senaratne, A. Mobasheri, A. L. Ali, C. Capineri, and M. M. Haklay, "A review of volunteered geographic information quality assessment methods," *International Journal of Geographical Information Science*, vol. 31, no. 1, pp. 139–167, 2017. [Online]. Available: http://dx.doi.org/10.1080/13658816.2016.1189556

H. Senaratne, M. Mueller, M. Behrisch, F. Lalanne, J. Bustos, J. Schneidewind, D. Keim, and T. Schreck, "Urban mobility analysis with mobile network data: A visual analytics approach," *Transactions on Intelligent Transportation Systems*, 2017.

Z. Shen and K.-L. Ma, "Mobivis: A visualization system for exploring mobile data," in *Visualization Symposium, 2008. PacificVIS'08. IEEE Pacific.* IEEE, 2008, pp. 175–182.

K. Sherren, J. Fischer, J. Pink, J. Stott, J. Stein, and H.-J. Yoon, "Australian graziers value sparse trees in their pastures: A viewshed analysis of photo-elicitation," *Society and Natural Resources*, vol. 24, no. 4, pp. 412–422, 2011.

B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on.* IEEE, 1996, pp. 336–343.

L.-A. Siebritz, "Assessing the accuracy of openstreetmap data in south africa for the purpose of integrating it with authoritative data," Master's thesis, University of Cape Town, 2014.

B. Sigurbjörnsson and R. Van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proceedings of the 17th international conference on World Wide Web.* ACM, 2008, pp. 327–336.

E. H. Simpson, "Measurement of diversity." *Nature*, 1949.

T. A. Slocum, D. C. Cliburn, J. J. Feddema, and J. R. Miller, "Evaluating the usability of a tool for visualizing the uncertainty of the future global water balance," *Cartography and Geographic Information Science*, vol. 30, no. 4, pp. 299–317, 2003.

I. Stavrakantonakis, A.-E. Gagiu, H. Kasper, I. Toma, and A. Thalhammer, "An approach for evaluation of social media monitoring tools," *Common Value Management*, vol. 52, no. 1, pp. 52–64, 2012.

A. Tatu, G. Albuquerque, M. Eisemann, P. Bak, H. Theisel, M. A. Magnor, and D. A. Keim, "Automated analytical methods to support visual exploration of high-dimensional data," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 5, pp. 584–597, 2011. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TVCG.2010.242

M. Tenney, "Quality evaluations on canadian openstreetmap data," in *Proceedings of the 4th Conference on Spatial Knowledge and Information (SKI 2014).* McGill University, Montreal, Quebec, Canada, 2014.

J. Thomson, E. Hetzler, A. MacEachren, M. Gahegan, and M. Pavel, "A typology for visualizing uncertainty," in *Electronic Imaging 2005.* International Society for Optics and Photonics, 2005, pp. 146–157.

W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic Geography*, vol. 46, pp. 234–240, 1970.

R. Tolouei, P. Álvarez, and N. Duduta, "Developing and verifying origin-destination matrices using mobile phone data: the llitm case," in *European Transport Conference 2015*, 2015.

M. Tsytsarau, T. Palpanas, and K. Denecke, "Scalable discovery of contradictions on the web," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 1195–1196.

E. R. Tufte and P. Graves-Morris, *The visual display of quantitative information*. Graphics press Cheshire, CT, 1983, vol. 2, no. 9.

J. W. Tukey, *Exploratory data analysis*. Addison-Wesley, 1977.

C. Valli and P. Hannay, "Geotagging where cyberspace comes to your place." in *Proceedings of the International Conference on Security and Management (SAM '10), 12-15 July*. CSREA press, Athens, GA, USA, 2010, pp. 627–632.

J. van de Kassteele and G. J. Velders, "Uncertainty assessment of local no2 concentrations derived from error-in-variable external drift kriging and its relationship to the 2010 air quality standard," *Atmospheric Environment*, vol. 40, no. 14, pp. 2583–2595, 2006.

F. Van der Wel, R. M. Hootsmans, and F. Ormeling, "Visualization of data quality," *Visualization in modern cartography*, pp. 313–331, 1994.

F. J. Van der Wel, L. C. Van der Gaag, and B. G. Gorte, "Visual exploration of uncertainty in remote-sensing classification," *Computers & Geosciences*, vol. 24, no. 4, pp. 335–343, 1998.

M. Van Exel, E. Dias, and S. Fruijtier, "The impact of crowdsourcing on spatial data quality indicators," *Proceedings of GiScience 2011, 14-17 September*, 2010.

P. A. J. Van Oort and A. K. Bregt, "Do users ignore spatial data quality? a decision-theoretic perspective," *Risk Analysis*, vol. 25, no. 6, pp. 1599–1610, 2005. [Online]. Available: http://dx.doi.org/10.1111/j.1539-6924.2005.00678.x

R. Van Zwol, "Flickr: Who is looking?" in *Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence*. IEEE Computer Society, 2007, pp. 184–190.

A. Vandecasteele and R. Devillers, "Improving volunteered geographic data quality using semantic similarity measurements," *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, no. 1, pp. 143–148, 2013.

——, "Improving volunteered geographic information quality using a tag recommender system: The case of openstreetmap," in *OpenStreetMap in GIScience*, J. Arsanjani, A. Zipf, P. Mooney, and M. Helbich, Eds. Springer, Switzerland, 2015, pp. 59–80.

A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1469–1472.

J. A. Verstegen, D. Karssenberg, F. van der Hilst, and A. Faaij, "Spatio-temporal uncertainty in spatial decision support systems: A case study of changing land availability for bioenergy crops in mozambique," *Computers, environment and urban systems*, vol. 36, no. 1, pp. 30–42, 2012.

F. B. Viegas, M. Wattenberg, and J. Feinberg, "Participatory visualization with wordle," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1137–1144, 2009.

T. von Landesberger, S. Bremm, T. Schreck, and D. Fellner, "Feature-based automatic identification of interesting data segments in group movement data," *Sage Information Visualization*, vol. 13, no. 3, pp. 190–212, 2014, peer-reviewed article.

L. Vullings, C. Blok, C. Wessels, and J. Bulens, "Dealing with the uncertainty of having incomplete sources of geo-information in spatial planning," *Applied Spatial Analysis and Policy*, pp. 1–21, 2013.

D. Wang, Z. Huang, Q. Liu, X. Zhang, D. Xu, Z. Wang, N. Li, J. Zhang, and D. Zhang, "Using semantic techology for consistency checking of road signs," in *Web Information Systems Engineering–WISE 2013 Workshops, 13-15 October*, Z. Huang, C. Liu, J. He, and G. Huang, Eds. Springer, Berlin Heidelberg, 2014, pp. 11–22.

J. Wang, G. J. Robinson, and K. White, "A fast solution to local viewshed computation using grid-based digital elevation models," *Photogrammetric Engineering and Remote Sensing*, vol. 62, no. 10, pp. 1157–1164, 1996.

W. Wang and K. Stewart, "Creating spatiotemporal semantic maps from web text documents," in *Space-Time Integration in Geography and GIScience*. Springer, 2015, pp. 157–174.

F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, and D. A. Keim, "State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams," in *EuroVis - STARs*, R. Borgo, R. Maciejewski, and I. Viola, Eds. Swansea, UK: Eurographics Association, 2014, pp. 125–139.

C. Ware, *Information visualization*. Morgan Kaufmann San Francisco, 2000, vol. 2.

F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

J. Will, "Development of an automated matching algorithm to assess the quality of the openstreetmap road network: a case study in göteborg, sweden," *Student thesis series INES*, 2014.

D. Xiu and G. E. Karniadakis, "Modeling uncertainty in flow simulations via generalized polynomial chaos," *Journal of computational physics*, vol. 187, no. 1, pp. 137–167, 2003.

P. A. Zandbergen, D. A. Ignizio, and K. E. Lenzer, "Positional accuracy of tiger 2000 and 2009 road networks," *Transactions in GIS*, vol. 15, no. 4, pp. 495–519, 2011.

W. Zhang and J. Kosecka, "Image based localization in urban environments," in *3D Data Processing, Visualization, and Transmission, Third International Symposium on*. IEEE, 2006, pp. 33–40.

Y. Zheng and X. Zhou, *Computing with spatial trajectories*. Springer, 2011.

D. Zielstra and H. H. Hochmair, "Positional accuracy analysis of flickr and panoramio images for selected world regions," *Journal of Spatial Science*, vol. 58, no. 2, pp. 251–273, 2013.

T. Zuk and M. S. T. Carpendale, "Visualization of uncertainty and reasoning," in *Smart Graphics, 7th International Symposium, SG 2007, Kyoto, Japan, June 25-27, 2007, Proceedings*, 2007, pp. 164–177. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-73214-3_15