

Visualizing Uncertainty in Reasoning -

A Bayesian Network-enabled Visual Analytics Approach for Geospatial Data

Ekaterina Chuprikova

Vollständiger Abdruck der von der Ingenieurfacultät Bau Geo Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:

Univ.-Prof. Dr. Urs Hugentobler

Prüfende der Dissertation:

1. Univ.-Prof. Dr.-Ing. Liqiu Meng
2. Univ.-Prof. Dr. Alan M. MacEachren
3. Univ.-Prof. Dr.-Ing. Jochen Schiewe

Die Dissertation wurde am 26.09.2018 bei der Technischen Universität München eingereicht und durch die Ingenieurfacultät Bau Geo Umwelt am 24.12.2018 angenommen.

Abstract

The growing importance of data-driven science and advances in computational capacity offer new opportunities for the analysis and visualization of geospatial and heterogeneous data. In recent years visual analytics has emerged as a relevant approach for gaining insights into various datasets. This includes embedding statistical methods in the interactive environment of visual analytics to help analysts understand and explore data. Apparently, much of the data explored using visual analytics is inherently uncertain due to limited knowledge, randomness and indeterminism, and vagueness. To address this challenge, we propose to integrate a probabilistic graphical model, namely the Bayesian Network, into a visual analytical system that allows us to model uncertainty and combine qualitative and quantitative data for reasoning.

This dissertation addresses the challenges of analytical reasoning under conditions of uncertainty when working with spatial data. It serves three research objectives: (1) to evaluate the feasibility of the Bayesian Network in representing conditional dependencies among heterogeneous spatial data; (2) to implement visual analytics scenarios that can demonstrate human-data discourses; (3) to build a prototype of a Bayesian Network-enabled visual analytical system dedicated to geospatial data classification tasks.

The literature review has shown that the Bayesian Network provides an effective framework for knowledge representation and reasoning under conditions of uncertainty. Despite the wide application of this technique in various fields of scientific research, its visual exploration is still limited to cause-effect relationships among variables, and only scant attention has been paid to the development of visualization support for the spatial component of the data. The development of visual interfaces can support users who work on spatial data integration, probabilistic methods, and visualization facility within a single interface and, thus, experience straightforward human-computer interaction processes. This research enhances the usability of probabilistic modelling within visual analytics and opens new perspectives for the application of Bayesian reasoning in GIScience and cartography. The applicability of the Bayesian Network-enabled visual analytics is illustrated based on two scenarios using land cover classifications and video surveillance data.

Although further usability studies are needed, the research work reported in this thesis marks a step forward towards innovative visualization approaches based on probabilistic methods, which treat the uncertainty in the reasoning process as an integral part of geospatial data analysis.

Zusammenfassung

Die zunehmende Bedeutung der datenbasierten Wissenschaft sowie der Einsatz datengetriebener Algorithmen verbunden mit dem technologischen Fortschritt in der Rechenkapazität bieten neue Möglichkeiten für die Analyse und die Visualisierung von heterogenen raumbezogenen Daten. Um Erkenntnisse aus verschiedenen Datensätzen gewinnen zu können und den Nutzern die explorative Datenanalyse weiter zu erleichtern, hat sich gerade die visuelle Datenanalyse in den letzten Jahren als zielführender Ansatz bewährt. Vor allem wurde auch die Einbindung statistischer Methoden in interaktive Umgebungen weiter entwickelt. Es ist bewiesen, dass viele der mit Hilfe der visuellen Analyse untersuchten Daten aufgrund von begrenztem Wissen, Zufälligkeit, Unbestimmtheit sowie Unschärfe von Natur aus unsicher sind. Zur Lösung dieses Problems wird vorgeschlagen, ein probabilistisches grafisches Modell - das Bayes'sche Netz - in die visuelle Datenanalyse zu integrieren. Die Integration dieses Modells ermöglicht Unsicherheiten zu modellieren und gleichzeitig qualitative und quantitative Daten so zu kombinieren, dass entsprechende Schlussfolgerungen aus den Daten gezogen werden können.

Diese Dissertation befasst sich mit den Herausforderungen des analytischen Schlussfolgerns unter Beachtung von Unsicherheiten bei der Arbeit mit Geodaten. Die drei Forschungsziele sind: (1) die Verwendbarkeit des Bayes'schen Netzes beim Darstellen bedingter Abhängigkeiten zwischen heterogenen raumbezogenen Daten; (2) die Implementierung visueller Datenanalyse-Szenarien, die dem Nutzer eine Interaktion mit dem Datensatz ermöglichen; (3) die Entwicklung eines Prototypen eines Bayes'schen netzwerkfähigen visuellen Analysesystems, speziell für die Klassifizierung von Geodaten.

Die Literaturrecherche hat gezeigt, dass das Bayes'sche Netz einen effektiven Rahmen für die Wissensrepräsentation und das Schlussfolgern unter Unsicherheits-Bedingungen bietet. Obwohl das Bayes'sche Netz in vielen verschiedenen Forschungsbereichen angewendet wird, ist die visuelle Exploration immer noch auf die Ursache-Wirkungs-Beziehung von Variablen beschränkt. Zudem wurde der Entwicklung einer Schnittstelle zur Visualisierung der räumlichen Komponente der Daten bislang nur wenig Aufmerksamkeit geschenkt. Die Entwicklung solcher visuellen Schnittstellen kann Nutzern dabei helfen direkte Mensch-Computer-Interaktionsprozesse auszuführen. Folglich können die Geodatenintegration, die probabilistischen Methoden und die Datenvisualisierung innerhalb einer einzigen Schnittstelle durchgeführt werden. Diese Forschungsarbeit zeigt, wie die Verwendbarkeit probabilistischer Modellierung in der visuellen Datenanalyse verbessert und neue Perspektiven für die Anwendung des Bayesschen Denkens in den Bereichen GIScience und speziell Kartographie eröffnen. Die Anwendbarkeit

der Bayes'schen netzwerkfähigen visuellen Datenanalyse wird in dieser Arbeit anhand von zwei Szenarien veranschaulicht: Klassifikation von Bodenbedeckung und Analyse von Daten aus Kameraüberwachungen.

Obwohl weitere Usability-Studien zur Nutzung des Bayes'schen Netzes im Bereich der visuellen Datenanalyse erforderlich wären, ist diese Forschungsarbeit ein Beitrag zu innovativen Visualisierungsansätzen auf der Grundlage probabilistischer Methoden, die die Unsicherheit im Argumentationsprozess als integralen Bestandteil der Geodatenanalyse behandeln.

List of abbreviations

DAG	Directed Acyclic Graph
CPT	Conditional Probability Table
JPD	Joint Probability Distribution
OSM	Open Street Map
CORINE	Co-Ordination of Information on the Environment
AI	Artificial Intelligence
GIS	Geographic Information Systems
VGI	Volunteered Geographic Information

Acknowledgements

This doctoral dissertation has been a life-changing experience, and it was only possible because of the support of many people.

Firstly, I would like to thank the TUM International Graduate School of Science and Engineering (IGSSE) and the people involved behind the scene. This excellent program made this dissertation an outstanding experience for me.

I would like to express my sincere gratitude to my advisor Prof. Dr.-Ing. Liqiu Meng for the continuous support of my doctoral work and related research, for her motivation, and immense knowledge. Her guidance and enthusiasm helped me in all the time of my research.

I would also like to thank Prof. Alan M. MacEachren, who became my co-advisor, for his valuable guidance. Thank you for your time and the opportunity to visit GeoVISTA center at Penn State University. He provided me with the suggestions I needed to choose the right direction and complete my dissertation.

Besides my advisors, I would like to thank my thesis committee: Prof. Dr. Urs Hugentobler and Prof. Dr.-Ing. Jochen Schiewe, for their insightful comments and encouragement, but also for the hard question which encouraged me to widen my research from various perspectives.

My sincere thanks also go to Prof. Jun Chen and Dr. Hao Wu, who provided me an opportunity to join their team as a visiting scholar. I appreciate this excellent chance to stay at the National Geomatics Center of China in Beijing that helped me to sharpen my research ideas.

I would like to acknowledge my colleagues from the Chair of Cartography, Technical University of Munich, who were always so helpful in numerous ways.

Last but not least, I would like to thank my family and friends for their continuous support. Love you.

Contents

Abstract	i
Zusammenfassung	iii
List of abbreviations	v
Acknowledgements	vii
1 Introduction	1
1.1 Motivation	1
1.2 Research aims and objectives	3
1.3 Research tasks	4
1.4 Thesis outline	4
2 Foundations and State-of-the-Art	7
2.1 Understanding uncertainty	7
2.1.1 Concepts and definitions	8
2.1.2 Sources of uncertainty	11
2.1.3 Uncertainty visualization	13
2.1.4 Uncertainty in reasoning process	14
2.2 Approaches for representing knowledge in uncertain domain . .	15
2.2.1 Human reasoning	15

2.2.2	Soft computing approaches	16
2.2.3	Evidence-based and probabilistic approaches	19
2.3	Design approaches for promoting analytical reasoning under uncertainty	31
2.4	Limitations of existing approaches and open challenges	38
3	Uncertainty in the reasoning process for geospatial data	41
3.1	Introduction	41
3.2	Reasoning under conditions of uncertainty	42
3.2.1	Human reasoning: historical perspective and current interpretation	42
3.2.2	Bayesian epistemology	46
3.3	Bayesian Network for reasoning under conditions of uncertainty	47
3.3.1	Events and their probabilities	48
3.3.2	Bayesian updating	50
3.3.3	Bayesian Network structure	51
3.3.4	A measure of uncertainty	54
3.4	A visual approach to human and Bayesian reasoning	55
3.4.1	Visualization and heuristics	55
3.4.2	Visualization and Bayesian reasoning	60
3.5	Summary	62
4	Visual analytics framework for supporting reasoning under uncertainty	65
4.1	Introduction	65
4.2	Requirements for visual analytics when using a probabilistic model on geospatial data	67
4.3	Visual analytics set-up	68
4.3.1	Visual analytics process	68

4.3.2	Visual analytics for reasoning under uncertainty	70
4.3.3	Bayesian Network-enabled visual analytics workflow	71
4.4	Graphical set-up	72
4.4.1	Bayesian Network mapping	73
4.4.2	Conditional probability tables	75
4.4.3	Outcome of the reasoning process	79
4.5	Summary	83
5	Prototypical implementations of Bayesian Network-enabled visual analytics	85
5.1	Scenario 1. Global land cover classification	86
5.1.1	Data	86
5.1.2	Experimental set-up	92
5.1.3	Visual encoding and interaction design	99
5.1.4	Results	103
5.2	Scenario 2. Video surveillance	106
5.2.1	Data	106
5.2.2	Experimental set-up	107
5.2.3	Visual encoding and interaction design	110
5.2.4	Results	111
5.3	Summary	113
6	Conclusion and outlook	115
6.1	Summary of thesis achievements	115
6.2	Research applicability to other domains	119
6.3	Outlook	121
	Bibliography	122

List of Tables

2.1	Sources of uncertainty in spatiotemporal data	12
2.2	Uncertainty in reasoning process. Based on Zuk & Carpendale (2007).	14
2.3	Summary of human reasoning and soft computing approaches: rule-based expert systems, fuzzy logic and fuzzy sets, rough sets, non-monotonic logics.	24
2.4	Summary of evidence theory (Dempster-Shafer theory) and probabilistic theory based techniques.	27
3.1	Joint probability for an event intersection.	49
3.2	Marginalization for a variable.	50
3.3	Visual variables.	55
5.1	Land cover classification scheme based on GlobeLand30.	88
5.2	Reclassification of CORINE land cover classes relevant for the study area based on the GlobeLand30 scheme.	92
5.3	Land cover variables and their instances.	94

List of Figures

2.1	Gradient of uncertainty components	9
2.2	Uncertainty visualization cube	13
2.3	Range of logical values	18
2.4	Applications of Bayesian Networks	21
2.5	Visual analytics process model	32
2.6	Visualization approaches for Bayesian Networks (GeNIe/SMILE and HUGIN Expert).	35
2.7	Visualization approaches for Bayesian Networks (BayesiaLab and Netica).	36
2.8	BayesiaLab extension for spatial data.	37
3.1	Variants of uncertainty	45
3.2	Combination of different events.	49
3.3	A representation of causal dependencies using Bayesian Networks.	52
3.4	An illustration of d-separation (direction-dependent separation) in the network.	53
3.5	The Müller-Lyer illusion	57
3.6	Error cone: hurricane Katrina forecast path	58
3.7	Availability heuristic applied to the visualization domain	59
3.8	Reasoning about diagnosis paradigm	60
3.9	Iconic symbols	61

3.10	Venn (Euler) diagram	61
3.11	Decision tree	61
3.12	Conditional probability table	61
3.13	Heatmap	62
3.14	Mosaic plot	62
3.15	"Beam cut" diagram	62
3.16	Probability curve	62
4.1	Uncertainty-aware visual analytics design	69
4.2	Bayesian Network visual representation.	73
4.3	Visualization approaches for Bayesian Networks.	74
4.4	Visualization example of conditional probability tables from GeNIe/SMILE software package.	76
4.5	Visualization of conditional probability tables using heat maps .	77
4.6	Visualization of conditional probability tables using heat maps. .	78
4.7	Visualization of joint probability distribution	78
4.8	Visualization of conditional probability tables using bar charts. .	79
4.9	Grid structure to visualize the outcome of the Bayesian Network analysis	80
4.10	Visualizing information entropy through point-based grid	81
4.11	Visualizing likelihood through point-based grid	82
5.1	Scenario 1: A workflow for assessing land cover data uncertainty using Bayesian Networks.	87
5.2	Scenario 1: An overview of pre-processing steps for the converting OSM data into land cover	90
5.3	Scenario 1: Bayesian Network structure for the land cover data analysis.	93
5.4	Scenario 1: Probability maps	96
5.5	Scenario 1: A map of information entropy based on Shannon Index	97

5.6	Scenario 1: An example of uncertainty in classification among land cover products.	98
5.7	Scenario 1: Three-tier architecture of the framework for land cover uncertainty analysis	99
5.8	Scenario 1: Uncertainty-Aware visual analytics interface.	100
5.9	Scenario 1: Conditional probability distribution panel	101
5.10	Scenario 1: Joint probability distribution panel	101
5.11	Scenario 1: Map view	102
5.12	Scenario 1: Case 1. Missing Lake	103
5.13	Scenario 1: Case 2. Forest and Shrubland	104
5.14	Scenario 1: Case 3. Grassland and cultivated areas	105
5.15	Scenario 2: Video surveillance data pre-processing	107
5.16	Scenario 2: Bayesian Network structure for reasoning about public surveillance data.	108
5.17	Scenario 2: Conditional probability table	109
5.18	Scenario 2: Uncertainty-aware visual analytics interface	110
5.19	Scenario 2: Map panel	111
6.1	Research applicability to other domains	120

Chapter 1

Introduction

Recognizing the increasing importance of data-driven science, it is crucial to develop methods to handle heterogeneous data and address the issues related to reasoning under uncertainty when using this data. This thesis is concerned with developing a visual analytics-enabled Bayesian Network approach to reasoning about spatial and heterogeneous data when a classification task is performed. A method of probabilistic graphical modeling was adopted and integrated within a visual analytics tool. Next is an introduction to the motivation, followed by the research aims and objectives, research tasks, and the thesis outline.

1.1 Motivation

The current state of technologies has triggered the acquisition of large amounts of data. Thus, the term *Big Data* has become common among various fields of scientific research, as data of diverse nature is generated by sensors, machines, networks, and social media. *Big Data* is characterized by its variety, velocity, validity, volatility, and veracity. The data veracity refers to the degree to which this data is certain and reliable. However, uncertainty is a natural outcome of scientific research and an inevitable characteristic of data handling processes; therefore, assessment and visualization of data uncertainty have been the focus of a substantial body of research. Both data and knowledge about the data are associated with uncertainty, which should be integrated and visualized in a manner that supports the reasoning process (Zuk & Carpendale 2007). While statistical methods used to be a main driving force for data analysis, a visual analytics approach has emerged as a solution that can visualize various data and operate it using innovative data- and knowledge-driven modelling

methods while taking advantage of qualitative and quantitative analytical techniques.

Daniel Kahneman and Amos Tversky explored the measurement of probability perception in the late 1970s (Tversky & Kahneman 1974) and proposed that humans generally exaggerate the probability of rare events such as earthquakes and floods, while at the same time they tend to underestimate the probability of rather likely events as car accidents. For example, the danger posed by nuclear power plants is often reported by media, although major reactor accidents caused by the failure of cooling systems are rare and the probability of a major accident occurring once in 20 - 25 years is roughly 1 in 100 000 (Lelieveld et al. 2012). Due to the fact that such a catastrophe would be a threat to any population and is often reported by the mass media, we might perceive the chances as being higher. At the same time, the odds of dying from stomach cancer are 3 in 1 000 (*Cancer Stat Facts: Stomach Cancer* 2018). Thus, sometimes the probability of different events might be perceived equally only because we have more information about some infrequent events. And when we evaluate the probability of an event, we base our knowledge on journalists' choices of topics and to our reliance on the availability heuristic (Kahneman 2011). In the same manner, when we analyze data, we often make judgments based on our availability heuristic, as it is problematic to interpret the numerical values. As a better alternative way to communicate information about probability is achieved by data visualization.

Although visual analysis of uncertain spatial data has been a significant topic that has drawn increasing attention in recent years from different research communities, the research outcomes have been limited by lack of a complementary focus on uncertainty in reasoning processes (MacEachren 2015). Uncertainty in the reasoning process is a persistent challenge in GIScience, as we are all unsure whether we perceive the visualized information correctly. But despite the uncertainty in the data and in our understanding of this data, a good visualization should support our cognitive processes in drawing the correct conclusions. But when it comes to the procedure of reasoning under uncertainty, how can we support users and provide an iterative process for decision-making? How can we place a human in the loop approach when it comes to reasoning under uncertainty?

In this work, we tackle the problem from a multidisciplinary perspective, extracting heterogeneous spatial data from two different domains (land cover classification and locations of surveillance cameras), and investigating a visual analytics approach as a combination of automated analysis and interactive visualization for analytical reasoning under uncertain conditions.

1.2 Research aims and objectives

In particular, this research aims to cover the following objective:

To support the reasoning process under uncertain conditions when performing a classification task on spatial data through developing a visual analytics application.

This main research objective can be further detailed by three interconnected sub-questions:

1. How can a classification task on heterogeneous spatial data best be performed within visual analytics?
2. Is it possible to combine analytical methods and Bayesian Network techniques to enhance the reasoning process using spatial data?
3. How can we integrate Bayesian Network and visualization techniques to address the complexity of decision-making processes under uncertainty using spatial data?

To answer these research questions we define the following sub-objectives:

Objective 1: To test the feasibility of using a probabilistic graphical model, namely the Bayesian Network, to represent conditional dependencies among heterogeneous data in order to perform a classification task.

Objective 2: To develop a visual analytics framework that can facilitate the understanding of data and uncertainty in the reasoning process using Bayesian Networks.

Objective 3: To build a prototype of a visual analytics interface that can integrate data, visualization, and computational capacity of Bayesian Networks to facilitate human-computer interactions for data analysis given subjective beliefs used in a selected application domain.

1.3 Research tasks

This dissertation addresses the following research tasks:

- To define an approach to explicitly model uncertainty in reasoning using a probabilistic graphical model, namely a Bayesian Network. This approach should be anchored in a Bayesian interpretation of relationships and dependencies among heterogeneous spatial data sets. Besides, it should combine notions of human judgment, confidence, belief, and evidence.
- To approach a classification task focused on heterogeneous spatial data under conditions of uncertainty, we combine a visual analytics approach and Bayesian Networks.
- To support users in introducing subjective beliefs for characterizing conditional probabilities within a single user interface.
- To provide visualization support when results of the inference can be observed in a spatial context.
- To exploit advances in visual analytics development by means of the integration of spatial data and computational capacity.

1.4 Thesis outline

Chapter 2, *Foundations and State-of-the-Art* provides an introduction to the subject and its value and relevance today, including some foundational understanding of the theoretical and practical basis of data visualization, visual analytics, and Bayesian reasoning.

Chapter 3, *Uncertainty in the reasoning process for geospatial data* introduces approaches for reasoning under conditions of uncertainty and, in particular, the methodology of probabilistic graphic models, namely Bayesian Networks, which is concerned with reasoning under uncertain conditions in the case of spatial data. This chapter provides the definition of reasoning under uncertainty, and shapes a design approach that is adopted by the visual analytics framework introduced in further sections.

Chapter 4, *Visual analytics framework for supporting reasoning under uncertainty* introduces the visual analytics methodology and the step-by-step approach of Bayesian Network integration. This chapter takes the reader

beyond the theoretical aspects of the methodology towards the design issues involved in establishing an effective visual analytics solution.

Chapter 5, *Prototypical implementations of Bayesian Network-enabled visual analytics* goes hand-in-hand with the previous chapters, as it explores the visual analytics prototype to communicate the data, probabilistic reasoning, and outcomes of the reasoning process based on the selected application domains.

Chapter 6, *Conclusion and outlook* come to a close by underlining the summary of the thesis achievements and sharing some of the opportunities for future work.

Chapter 2

Foundations and State-of-the-Art

“ *As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.* ”

Albert Einstein, *Geometry and Experience*, 1921

This chapter identifies a context of uncertainty in GIScience by presenting definitions and design approaches for visualization. Section 2.2 introduces conventional methods for reasoning under uncertainty, which include Naive Bayes, rule-based expert systems, Dempster-Shafer theory, Bayesian Networks, influence diagrams, fuzzy logic and fuzzy sets, rough sets, non-monotonic logics, and neural networks. Moreover, this chapter investigates several conceptual issues in the utilization of Bayesian logic for reasoning on heterogeneous data and the role of visual analytics in the development of uncertainty-aware applications.

2.1 Understanding uncertainty: concepts and definitions, sources, reasoning, and visualization

Uncertainty is a natural outcome of scientific research, and it is an inevitable characteristic of data handling processes. Although there are established approaches in GIScience, data acquisition, processing, and visualization are constantly affected by factors such as the dynamic nature of the measured

phenomena, the limited capacity of the measuring instruments and processing devices, and the human factor. Even if information from raw data is extracted using sophisticated algorithms, this information still has some degree of uncertainty. It occurs not only because of a lack of knowledge but also disagreement over the knowledge that currently exists. Nevertheless, well-informed decisions and decisions that particularly meet the problem of coping with uncertainty in environmental modeling depend essentially upon adequate data and knowledge of uncertainty. The literature on data uncertainty shows a variety of approaches to defining it. And since uncertainty touches most aspects of life (Tannert et al. 2007), it might be interpreted in different ways. Uncertainty can refer to noise in the information, statistical variability, non-deterministic relationships between action and consequences, or even the psychological reaction to difficult problems (Kirschenbaum et al. 2014).

2.1.1 Concepts and definitions

In GIScience, data uncertainty is often defined as a difference between the contents of a spatial database and the corresponding phenomena in the real world (Goodchild 2008). Thomson et al. (2005) have proposed a typology for categorizing uncertainty. Thomson et al. (2005) suggested categories such as accuracy/error, precision, completeness, consistency, lineage, currency, credibility, subjectivity, and interrelatedness. At the same time, Unwin (1995) has claimed that accuracy and error are two different things as they deal with distinctive aspects of data quality. Data uncertainty can therefore be considered as an umbrella term, where all the elements are closely related and in each phase of the spatiotemporal data collection, processing, modeling and visualization, different aspects of uncertainty may be introduced

Several authors have accessed uncertainty components in a systematic way (Thomson et al. 2005, Fisher et al. 2010) to describe different qualitative and quantitative aspects of data quality and its understanding. The uncertainty within categories proposed by Thomson et al. (2005) might be seen from two perspectives: uncertainty associated with the measurements of location, time, attributes, and uncertainty associated with the understanding that includes credibility of a data source, or the subjectivity of information. In principle, data uncertainty can be seen as a function that combines multiple factors whose relevance changes depending on a particular usage. Various data uncertainty components are rather loosely structured as some of them can represent both measurements and understanding. These components can be illustrated as a gradient from measurements to understanding (see Fig. 2.1).

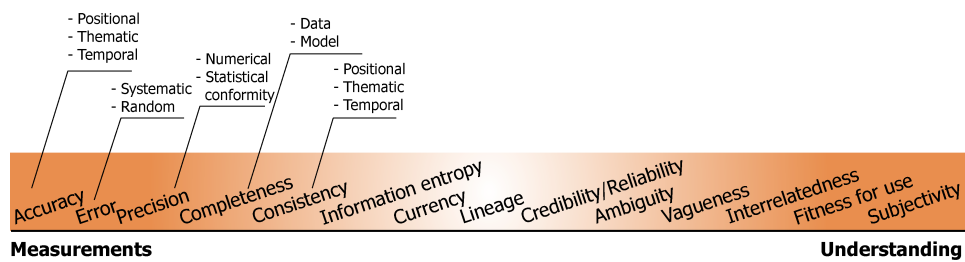


Figure 2.1: Gradient of uncertainty components.

Furthermore, a description is given that represents various components related to the data quality and uncertainty as they are defined in the literature.

Uncertainty "may be defined as a measure of the user's understanding of the difference between the contents of a dataset, and the real phenomena that the data are believed to represent" (Longley et al. 2011).

Accuracy is defined as closeness of agreement between a measured quantity value and a true quantity value (De Bièvre 2012). In spatial data three accuracy types may be distinguished:

- Positional:
 - Absolute shows "how closely all positions on a map or data layer match, corresponding to the positions of features represented on the ground in a desired map projection system" (Stanislawski et al. 1996);
 - Relative represents "how closely all the positions on a map or data layer represent their corresponding geometrical relationships on the ground" (Stanislawski et al. 1996);
- Thematic: how closely an object type is mapped compared to the object type on the ground;
- Temporal: how closely an object is mapped at a particular time compared to what is on the ground at the time of mapping.

Error is measured quantity value minus a reference quantity value (De Bièvre 2012). Error types:

- Systematic (bias);
- Random (noise).

Precision refers to the accuracy with which a measurement can be made, recorded, or calculated (Fisher et al. 2010). It is estimated in terms of standard deviation, variance, or coefficient of variation. Precision types:

- Numerical: number of significant digits of a measurement or observation;
- Statistical conformity of repeated measurements to the reported value.

Completeness is closeness of the data to be "100%" complete. The completeness may refer to the data and/or models.

- Data completeness is a measurable error of omission observed between the database and the specification (Veregin 1999).
- Model completeness is the agreement between the database specification and the abstract universe that is required for a particular database application (Veregin 1999).

Logical Consistency refers to the absence of apparent contradictions in a database that can be related to the topology, redundancy in thematic attributes, and inconsistency in time when different entities appear at the same location on two maps of the same date (Veregin 1999).

- Positional;
- Thematic;
- Temporal.

Lineage describes the history of a data set and the processing steps used to create the data.

Currency defines temporal gaps between actual time, data acquisition, and occurrence.

Credibility/Reliability is the extent to which it is based on the reliable sources or provided by trusted organizations.

Subjectivity is the amount of interpretation or judgment included (Thomson et al. 2005).

Interrelatedness deals with source independence from other information (Thomson et al. 2005).

Ambiguity deals with diverging perceptions in classification of an object (Fisher 1986).

Vagueness refers to the poor definition of class of object or individual object (Fisher 1986).

Fitness for use defines the suitability of data for a particular use and user (Dodge et al. 2008).

Information entropy expresses the amount of information that can serve as quantitative estimator of complexity (Jost 2006).

Generally, uncertainty can be classified into two main types aleatoric and epistemic (Kiureghian & Ditlevsen 2009). Aleatoric uncertainty represents statistical uncertainty and reflects inherent randomness in natural processes. Thus, aleatoric uncertainty represents unknowns during observations. Epistemic uncertainty is systematic and occurs due to the lack of knowledge and limited capability of models to describe a phenomenon. Moreover, epistemic uncertainty is characterized by alternative models. It emerges from parameter estimations and may be represented using probability values, which is why this type of uncertainty can be reduced if data of higher accuracy or additional information are available. Hence, uncertainty is characterized as epistemic if an analyst sees an opportunity to reduce it, and aleatoric if the analyst cannot reduce it. Conventionally, both types of uncertainty occur in real-world applications, and their influence is widely discussed in the scientific community (Kiureghian & Ditlevsen 2009). Uncertainty sources are especially difficult to detect as there is often a combination of different factors that influence data collection, processing, and visualization.

2.1.2 Sources of uncertainty

Pang et al. (1997) identifies three primary sources of uncertainty: (a) acquired from the measurements, numerical models or statistical variation, (b) introduced after data transformation, unit conversion, and data fusion, and (c) introduced through visualization processes. From a practical point of view, these sources of uncertainty can be seen as random and non-random. Table 2.1 provides an extended view of sources of uncertainty considering their random or non-random nature.

Table 2.1: Sources of uncertainty in spatiotemporal data

Uncertainty source	Random	Non-random
Observed in sampled data	Dynamic processes; incomplete measurements; statistical variations; errors in existing maps used for digital data creation;	Equipment errors; unsuitable data collection techniques.
Measures generated by models, simulations and data processing	Inaccuracies in digitizing (operator); errors in model coefficients; discretization of geographic entities; errors in attribute entry; misclassification	Data conversion algorithms; inaccuracies in digitizing (equipment); data storage (numerical precision, data format); uncertainty propagation in multiple overlay operations; interpolation; rescaling; re-sampling.
Introduced by visualization processes	Seeing patterns that do not exist, failure to notice patterns and relationships	Rendering on a device screen, approximations.

2.1.3 Uncertainty visualization

According to Brodlić et al. (2012), it is possible to distinguish two main issues related to uncertainty and visualization: (1) visualization of uncertainty, which considers how we depict uncertainty specified with the data, and (2) uncertainty of visualization, which examines how much inaccuracy occurs as we process data through the visualization pipeline.

Adequate visualization of uncertainty may help answer scientific questions and support decision-making processes; it is an ongoing research problem in the GIScience community, and it has been placed in the focus of research by many others. Current techniques for handling spatial-temporal uncertainty typically rely on treating data and its uncertainty as separate features when represented visually, through either intrinsic or extrinsic visualization, coincident/adjacent display, or static/dynamic views (see Fig. 2.2) (Kinkeldey, MacEachren & Schiewe 2014).

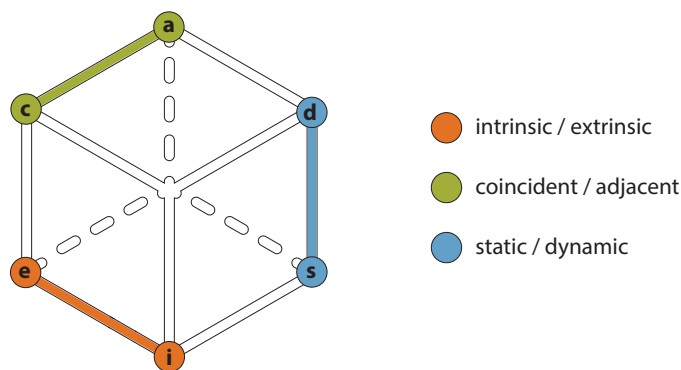


Figure 2.2: Uncertainty visualization cube according to Kinkeldey, MacEachren & Schiewe (2014).

Recent metastudies of GIScience literature have focused on visualizing the uncertainty of spatial data. MacEachren et al. (2005) and Smith Mason et al. (2016) have shown a variety of approaches, including glyphs (Wittenbrink et al. 1996), isolines and isosurface (Rhodes et al. 2003), grid structures (Kinkeldey, Mason, Klippel & Schiewe 2014), 3D (Wellmann & Regenauer-Lieb 2012), and choropleth maps (Lucchesi & Wikle 2017). The importance of visualization of uncertainty is often associated with the process of decision-making. Although current visualization techniques offer various approaches to handling uncertainty, it is still a challenging problem, as work to date deals primarily with visualizing ambiguous data rather than with reasoning under uncertainty (MacEachren 2015).

2.1.4 Uncertainty in reasoning process

In practice, all spatial data accrues some degree of uncertainty during acquisition, processing, and visualization, which is multiplied by the uncertainty in the reasoning process. Zuk & Carpendale (2007) have extended Thomson et al. (2005)'s topology to the reasoning to guide the development of visual representations for uncertainty. Table 2.2 demonstrates Zuk & Carpendale (2007)'s extension.

Table 2.2: Uncertainty in reasoning process. Based on Zuk & Carpendale (2007).

Uncertainty Category	Reasoning Definition
Currency/Timing	Temporal gaps between assumptions and reasoning steps
Credibility	Heuristic accuracy and bias of analyst
Lineage	Conduit of assumptions, reasoning, revision, and presentation
Subjectivity	Amount of private knowledge or heuristics utilized
Accuracy/Error	Difference between heuristic and algorithm (e.g. Bayesian)
Precision	Variability of heuristics and strategies
Consistency	Extent to which heuristic assessments agree
Interrelatedness	Heuristic and analyst independence
Completeness	Extent to which knowledge is complete

In contrast to the standard approaches to uncertainty in GIScience, in Bayesian reasoning, uncertainty is often associated with such factors as ignorance (due to limits of our knowledge about a phenomena); randomness and indeterminism (as not all events are determined by causal relationships and there is always room for physical randomness); and vagueness (as the statements we make are often vague) (Korb & Nicholson 2010). Bayesian reasoning considers probability as a measure of our subjective degree of belief based on our present state of knowledge. Therefore, the reasoning process is subject to uncertainty when conditional relationships among variables are given to quantify the likelihood of events.

Recently, researchers have shown an increasing interest in developing techniques for decision-making, considering the states of uncertainty.

Reasoning approaches vary from human reasoning to sophisticated computational algorithms. For instance, Bayesian inference has gained considerable recognition as an effective method for supporting decision-making practices. Thus, various methods for representing and reasoning with uncertainty have been developed, including fuzzy set theory, formal logics (Bloch 2006), probabilistic clustering (Lin et al. 2017), and Bayesian Networks, also called belief networks and causal probabilistic networks (Stassopoulou et al. 1998). The summary of recently available methods is given in section 2.2.

2.2 Approaches for representing knowledge in uncertain domain

Data-driven science has boosted the development of various methods to support the reasoning process on data and spatial data in particular. Providing reliable decisions leading to intelligent actions by stakeholders who use the data, it is crucial to address the uncertainty in the reasoning process. The application of various methods to support decision-making practices has grown exponentially over the last decade, and the complexity and the amount of the data have also increased. This has created a high demand for automated reasoning approaches. The decision-making under uncertainty can be addressed from various perspectives (see, e.g., human reasoning, rule-based expert systems, fuzzy set theory, rough sets, non-monotonic logics, evidence theory (Dempster-Shafer theory), probabilistic models including naïve Bayes, Bayesian Networks, influence diagrams, Neural Networks, etc.). All these models involve inferences from the data with some degree of uncertainty. The uncertainty might be given through descriptive IF-THEN, non-monotonic statements, quantitative formalization as probability values, membership function, or degrees of belief.

As pointed out by Box (1979) "all models are wrong, but some are useful", and there is no method that can perform ideally on every given assignment. The summary is given in Tables 2.3 and 2.4 indicates that every method has advantages and drawbacks. Thus, the method selection essentially depends on the aim of the analysis and data available.

2.2.1 Human reasoning

Human reasoning is a complex process that involves existing knowledge and experience to make assumptions and draw conclusions. Although there are various definitions of the human reasoning process, from the general point of

view it can be divided into three kinds based on logical validity: deductive (from a general knowledge to a specific case), inductive (from a specific case to a general conclusion), and abductive (reasoning to the best explanation from the observations).

Deductive reasoning makes valid conclusions, which must be true given that their premises are true (Johnson-Laird 1999). In other words, deductive reasoning is a top-down approach leading to a conclusion. A deduction is a common approach in scientific research; a hypothesis is tested, and predictions about possible consequences are made. Deduction is a stepwise approach, where several premises can precede the inference. Example: "An airport is an artificial object." Another assumption could follow this premise, "all artificial objects are represented on a land cover map with a pixel value of 80". Those statements would lead to the conclusion that "an airport is represented on a land cover map with pixel value 80". In deductive reasoning, if the premise is true for a class of things, it is also true for all members of this class.

By contrast, inductive reasoning starts with an observation, but it guarantees neither that there is a conclusion nor that where there is a conclusion that it is logically reached. Inductive reasoning gathers evidence, searches for patterns, and forms a theory or hypothesis. This kind of argument is often used in science, too. For example, based on a collection of observations about a land cover over a long period of time, researchers may be able to ascertain whether the land cover is shrinking.

Abductive reasoning typically begins with an incomplete set of observations and proceeds to the best possible explanation for the set. This reasoning type generates the kind of daily decision-making that does its best with the information available, which often is incomplete and uncertain. Despite the fact that the approaches based on logical validity are commonly accepted, the human reasoning process is not simple and it does not follow a single path of logical thinking. It draws no clear distinction between deduction, induction, and abduction because it tends to exploit what we already know (Johnson-Laird 2010). Artificial Intelligence (AI) combines inductive, deductive, and abductive reasoning in a human-like manner and established structured solutions to analyze complex problems even in the presence of missing, incomplete and noisy information.

2.2.2 Soft computing approaches

Rule-based expert systems

Rule-based expert systems are a common technique in knowledge-based AI and are widely used in GIScience. For instance, Choi & Usery (2004) has

suggested a rule-based expert system for urban mapping, where an interactive question-and-answer sequence is integrated within Geographic Information Systems (GIS) for spatial data mapping. By using a natural language, an expert replies to a specific question in the manner Yes/No answer. However, despite the common use, rule-based expert systems have an obvious drawback due to limited ability to handle vague associations. For rule-based systems, only exact reasoning is possible. They assume that perfect knowledge exists and each question might be answered as true or false. In addition, this approach follows a sequence of rules, so the reasoning outcome depends on the rule order that may lead to divergent results if the order is changed.

Fuzzy logic and fuzzy sets

Given the limitation of rule-based approaches, fuzzy logic, introduced by Zadeh (1965), is able to describe vagueness through a combination of symbolic and numeric values. The fuzzy logic approach is based on the belief that all premises are associated with a degree of membership. For example, fuzzy logic represents the way people would think and which statements they might make, as for instance "It's drizzling", "It's pouring" "It's raining cats and dogs". Defining the continuum of the logical values between 0 (false statement) and 1 (completely true), it is possible to describe the multi-value statements as degrees of truth, or degrees of membership. In contrast to Boolean logic (rule-based approaches) (Fig. 2.3(a)), fuzzy logic (Fig. 2.3(b)) can describe a variation by describing assumptions as being partly true or partly false in the interval [0,1]. Foody (1996) addressed the benefits of fuzzy sets classification for a better representation of some vegetation from remotely sensed imageries as they support continuum of land cover classes' distribution. Therefore, the fuzzy set approaches can capture the natural uncertainty and imprecision of class definitions.

Rough set theory

Rough set theory was proposed by Pawlak (1982) in order to deal with imprecise, inconsistent, incomplete information and knowledge. Rough set theory represents a mathematical approach to describe our knowledge and perception of data (Walczak & Massart 1999). In contrast to fuzzy set theory, where the selection of the membership function is uncertain, rough set theory uses two precise boundary lines to describe uncertain concepts. This approach defines a set of lower and upper approximations, where the lower approximation includes all the objects that definitely belong to the set, whereas the upper set consists of the objects that may belong to the set. The

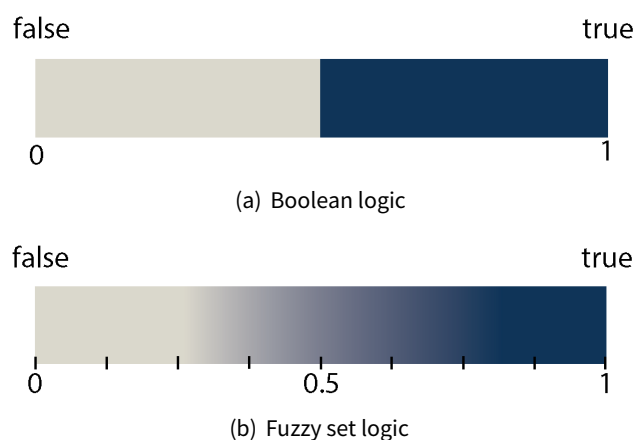


Figure 2.3: Range of logical values. According to boolean logic an element is either in or outside of the range. The fuzzy set proposes a membership function which assign the elements in the range of the interval $[0,1]$.

data is often represented as tables, also called information systems, or information tables, where condition and decision attributes are included. The application of rough set theory is crucial in AI and cognitive science applications which support decision-making and data mining. Sikder (2016) proposed using the rough set approach for classification and prediction of land cover classes, where he introduces approximate reasoning to support knowledge discovery in land cover classification. Moreover, the extensive use of rough set theory in GIScience includes such active domain applications as landslide hazard (Peng et al. 2014), map generalization, and optimization of spatial databases. Although rough set theory is an effective method for knowledge acquisition, it is restricted by its dependence on complete information tables, because in real-time applications, missing values and errors might occur (Nabwey 2011).

Non-monotonic logics

The term "non-monotonic logic" (NML) refers to a family of formal frameworks devised to capture and represent defeasible inference (Strasser & Antonelli 2018), in which reasoners draw conclusions approximately with the ability to revise the assumptions based on new evidence. The deductive reasoning is monotonic, which means that the arguments preserve the truth and are not allowed to be changed. However, in the real-life most of the problems are non-monotonic. In NML, the facts and rules are dynamic, which means that they can be changed at any time. Therefore, in contrast to monotonic logic, in NML, previous conclusions may be invalidated by new knowledge. This type of

reasoning is similar to human reasoning in its approach. Although the NML approach is widely used for reasoning in everyday environments, expert and scientific systems, there are number of drawbacks such as disbelief propagation, derivation of new beliefs from existing ones, contradictions in beliefs that cannot be overcome within some application.

2.2.3 Evidence-based and probabilistic approaches

Evidence theory

Evidence theory, also called the Dempster-Shafer theory of evidence, was introduced as an alternative approach to probabilistic techniques. At the qualitative level, Dempster-Shafer theory provides a graphical description of a knowledge base by modeling variables and their relations (Cobb & Shenoy 2003). Moreover, at the numerical level, a Dempster-Shafer theory of evidence assigns a belief function to subsets of the variables in the domain of each relation (Cobb & Shenoy 2003). Furthermore, to update knowledge within a belief function, further evidence may be provided. Thus, this mathematical model is based on reasoning with belief that defines all available evidence and plausibility that refers to all evidence consistent according to a hypothesis. Thus, the combination of the two aspects characterizes the probability interval of the hypothesis. The mathematical approach is described through a belief function, which can include combination of information from different resources.

A substantial number of applications of Dempster - Shafer Theory of evidence in GIScience indicate that this approach is able to handle uncertainty, imprecision, ignorance, and lack of data, and it can perform particularly well with raster-based input for various environmental applications (Malpica et al. 2007). Although Dempster - Shafer Theory and Bayesian Networks are similar in certain aspects, there are differences in graphic representations, numerical details, semantics and methods of performing inference (Cobb & Shenoy 2003). As two approaches have different semantics, they are used for different scenarios, where Dempster - Shafer evidence theory better facilitate representation of non-causal knowledge whereas causal relationships are better represented by conditional probabilities, thus Bayesian Networks. Moreover, despite the extensive usage, computational complexity is the main concern for developing Dempster - Shafer evidence theory within stand-alone applications for geospatial data.

Probabilistic theory based techniques

Probabilistic theory-based techniques are enabled to handle both epistemic and aleatoric uncertainties, where uncertainty is modeled through the degree of belief which substitutes knowledge about the whole system. The degree of belief is expressed as a value of conditional probability for all possible events.

Due to wide applicability to various domains – for example, pattern recognition data mining, data exploration and optimization, along with the ability to deal with sparse training data, provide interpretable results, and establish mathematically rigorous inference – probabilistic methods have a considerable advantage over deterministic methods. Some common techniques are summarized in Table 2.4.

Bayesian Network

Graphical models represent a large group of probabilistic methods. Among these methods Bayesian Networks are distinguishable for their ability to be versatile, extensible, and understandable for wide scientific community. Bayesian Networks are well established modeling tool for expert systems, where real world problems can be described through easily comprehensible graphical representation of conditional dependencies, while based on solid rules of probabilistic calculus.

Moreover, several studies have indicated that Bayesian Networks provide an effective approach to assess data uncertainty. Thus, the use and visualization of Bayesian Networks has received explicit attention (Chiang et al. 2005, Koiter 2006, Li et al. 2010, Cossalter et al. 2011, Taalab et al. 2015, Pietro et al. 2017, Drury et al. 2017, Champion & Elkan 2017, Abebe et al. 2018). The Bayesian Networks graphically represent uncertain quantities and decisions that reveal probabilistic dependencies among the variables and related information flows (Villa & Cozzani 2016). Bayesian Networks provide an effective mechanism to deal with uncertainties and information from different sources such as expert judgment and observable patterns. In addition, Bayesian Networks allow analysts to not only model the causal relationships, but also to visualize them, which gives a transparent inferential mechanism. Despite the given advantages, Bayesian Networks have some limitations too such as computational complexity and difficulty in specifying priors. When performing Bayesian inference, the computational complexity is growing exponentially. Thus, it requires both processing facility and effective visual support as specification of priors may involve specialized knowledge about the data and the context where this data is used.

The literature on Bayesian Networks shows a variety of practical solutions in different domains (see Fig. 2.4). Pourret et al. (2008) offered a comprehensive guide through diverse applications and demonstrated through case studies how can expert knowledge be combined with computational capacity of Bayesian Networks. The application examples include medical diagnosis, genetic models, crime risk factors, spatial dynamics, terrorism attack management, classification of Chilean wines, risk management, and human cognition. Moreover, Pourret et al. (2008) highlighted versatility and modeling power of Bayesian networks for users that come from various spheres of research as engineering, computer science, medicine, bioinformatics, real estate, finance, psychology and cognitive science. Due to computational power and ability to model complex systems under uncertain conditions, Bayesian networks became an increasingly popular method in GIScience too. To represent constant interest towards Bayesian Networks within the GIScience community we selected some publications over the last decade.

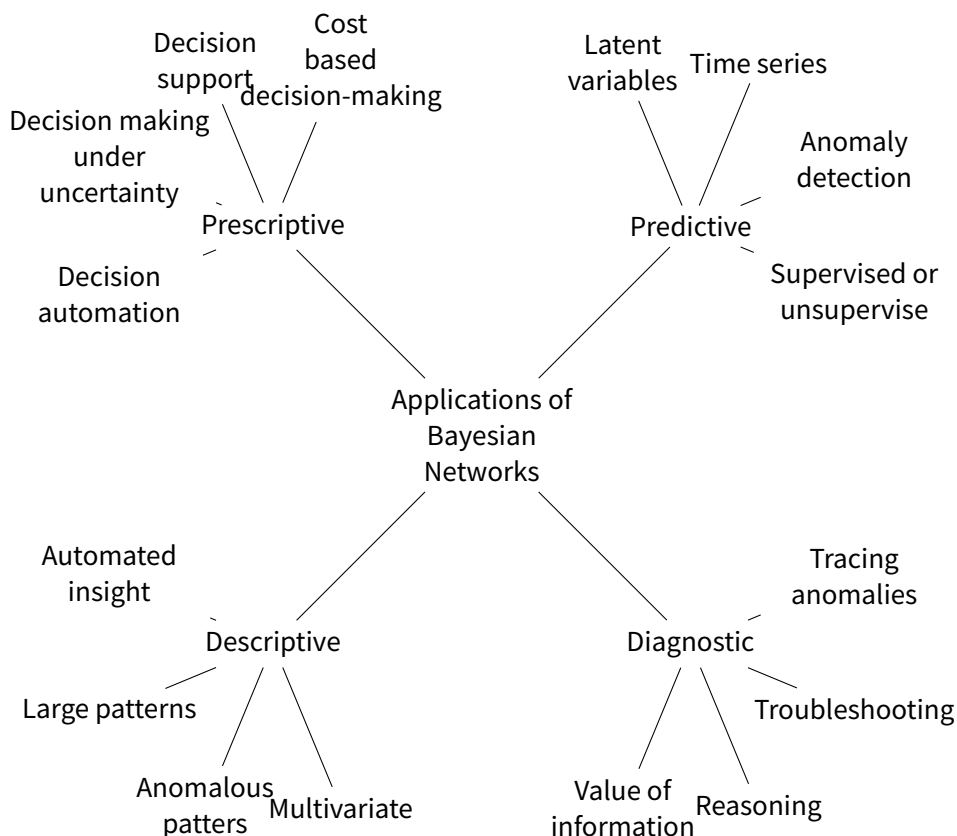


Figure 2.4: Applications of Bayesian Networks: Descriptive, diagnostic, predictive, and prescriptive analytics.

Source: Bayes Server: www.bayesserver.com

Stassopoulou et al. (1998) investigated how the Bayesian Networks can be used for assessing desertification risk of burned forest in Mediterranean region combining information from different sources. The study proved that the results obtained were worse when the uncertainty in the data and uncertainty in the inference were ignored.

Straub (2001) suggested to assess natural hazard risk such as rock-fall on roads using Bayesian Networks. Moreover, Straub (2001) underlined that Bayesian Networks are flexible and intuitive and they provide transparent inference mechanism and simultaneously incorporate the concept of risk through uncertainty definition.

Körner et al. (2009) proposed to analyze and visualize large trajectory data using Bayesian Networks within GIS MapInfo software package. The approach has applied to Scalable Sparse Bayesian Network Learning algorithm in order to generate a compact model of trajectory dependency structure which can be used for efficient visualization.

Landuyt et al. (2014) suggested to incorporate Bayesian Networks for ecosystem service modeling. To realize the modeling, Landuyt et al. (2014) have developed a Quantum GIS plug-in which enables pixel-based application of the Bayesian Networks to map ecosystem service delivery and associated uncertainties. The reported results are maps that can be used for decision-making and regional ecosystem service accounting.

Celio et al. (2014) integrated spatially explicit Bayesian Networks to model land use decisions in a pre - Alpine area in Switzerland along with biophysical data and local expert knowledge. Celio et al. (2014) have used questionnaires in order to collect information from local stakeholders and update the Bayesian Network model with important characteristics for land use decision-making in the case study region. The outcome of the research indicated path-dependencies of land use change that can be served as information for planners and policy makers.

Taalab et al. (2015) proposed Bayesian Networks for digital soil mapping that integrate measured data and expert opinion. The results reported indicate that such approach is effective for prediction of soil physical property and qualitative prediction of soil taxonomic class under consideration of uncertain expert knowledge. Additionally, Taalab et al. (2015) suggested the Bayesian Network method as a feasible alternative to black-box data mining techniques for soil properties modeling.

Chee et al. (2016) have demonstrated the use of spatial Bayesian Networks for two study cases - one developed for adaptive management of eucalyptus woodland restoration in south-eastern Australia, and another developed to manage the encroachment of invasive willows into marsh ecosystems in

east-central Florida. The proposed approach uses object-oriented concept which supports the complex Bayesian Network modeling.

To sum up, the efficiency of Bayesian Networks lays in providing uncertainty estimates about the given alternatives when human reasoning can be integrated. Therefore, the combination of computational consistency and human comprehension makes this method favorable for the further integration within a visual analytics application prototype that is proposed in Chapter 4. Further details on Bayesian Network approach are given in Chapter 3.

The executive summary for prominent reasoning methods is given in Tables 2.3 and 2.4 and indicates that every method has advantages and drawbacks.

Table 2.3: Summary of human reasoning and soft computing approaches: rule-based expert systems, fuzzy logic and fuzzy sets, rough sets, non-monotonic logics.

Method	Description	Dealing with uncertainty	Advantages	Disadvantages
Human reasoning	A complex system that is using knowledge, experience, a rule of thumb and heuristics for decision-making.	Using approximate reasoning and being able to deal with uncertain, incomplete, and fuzzy information.	Ability to learn and practice.	A slow and expensive process of learning. Can make mistakes in uncertain data.

Ruled-based expert systems	The knowledge is represented through IF-THEN statements. IF part is premise or condition and THEN is consequent. The connection among rules is given through logical operators AND and/or OR. The condition may have a numerical or symbolic value.	Classic logic permits only exact knowledge (TRUE OR FALSE). Unable to handle inconsistent knowledge.	Easy to explain. Separation of knowledge and processing system.	Inconsistent and unrealistic assumptions.
Fuzzy logic and fuzzy sets	Based on the set of mathematical approaches that use degrees of membership for the variable definition. Fuzzy logic reflects how people think.	Representation of vagueness of natural language.	Descriptive. Enabled to combine symbolic and numerical values.	Lack of clear semantics that leads to an arbitrariness in the application. No clear criterion that can be used for all the application domains (Díez & Druzdzel 2003).

Rough sets	A mathematical approach for soft computing that classifies the data through decision data tables (also called information system). The data can be acquired based on measurements or human experts.	Indiscernibility between objects is considered as uncertainty.	No prior information about the data is needed. A straightforward interpretation of the results. Data reduction.	Dependence on the complete decision table. Limited performance for incomplete and missing values.
Non-monotonic logics	A method for human-like reasoning that begins with making plausible, but fallible assumptions and revises them in the light of new evidence. Non-monotonic means that revision of the proposed assumption can be made.	Based on qualitative and logical propositions.	Easy to build. Adaptable.	Unable to weight conflicting evidence.

Table 2.4: Summary of evidence theory (Dempster-Shafer theory) and probabilistic theory based techniques.

Method	Description	Dealing with uncertainty	Advantages	Disadvantages
Dempster-Shafer theory	Generalization of the classic probabilistic framework also called a theory of belief function. It consists of specific calculation for modeling and reasoning about propositions of uncertainty.	Makes a distinction between uncertainty and imprecision by enabling to handle composite hypotheses.	Suitable for combining information from different sources.	A computational complexity grows exponentially with the number of hypotheses. Lack of robustness of the combination of evidence (Díez & Druzdzel 2003).

Naive Bayes	Based on Bayes' theorem. This method considers each variable's value as independent and assumes no correlation or relationship among the features.	Bayesian interpretation of uncertainty.	Suitable for high-dimensional datasets. Fast for training and prediction, easily interpretable. Few parameters.	Unrealistic assumptions. Conclusions are exclusive, findings are conditionally independent.
Bayesian Networks	A probabilistic graphical model built on Bayes' theorem and described by a mathematical model with qualitative and quantitative components (Darwiche 2008).	Allow modeling of both types of uncertainty: in the reasoning rules and in data sources.	Evidence combining approach. Easy to extend and refine. Easy to explain.	Computational complexity and difficulty to build the network. A large number of probability values to be filled.

Influence diagrams	Extension of Bayesian Networks. Include decision options and preferences. Include other options for the reasoning process as decision nodes and utility nodes. May be used for goal-oriented reasoning.	Uncertainty is represented as an interval of $[0, 1]$.	Enabled to accommodate complex systems. Straightforward. Evidence-based. Graphical output.	Computational complexity and difficulty to build the network. A large number of probability values to be filled.
--------------------	---	---	--	--

Neural networks	<p>Group of various approaches inspired by biological neural systems that can compute, learn, remember and optimize the data. The architecture is composed of many simple processing nodes whose function is determined by network structure and connection strengths (Sucharita 2016).</p>	<p>Some models are unable to understand the uncertainty. Two main groups of uncertainty are distinguished: epistemic (model uncertainty) and aleatoric (random, e.g. occlusions, lack of visual features).</p>	<p>High performance. Scalable models as the ability to understand data can increase according to the model size and the training data availability. An adaptable system can deal with incomplete data.</p>	<p>So-called "black-box" approaches. NN work well to describe data, but provide little to none understanding of generating mechanisms. Require a large amount of data to be trained for another application. Require large computational capacity.</p>
-----------------	---	--	--	--

2.3 Design approaches for promoting analytical reasoning under uncertainty

Although uncertainty is essential part in any analytical and reasoning process, there is a lack of research regarding the transparency of visual representation and the inference algorithms used. The transparency is crucial for establishing a reliable system and enabling an analyst to improve the decision-making. The transparency process may be facilitated when the reasoning about data supports elicitation of users' prior knowledge in visualization interaction. In this regard, the visual analytics may promote an understanding of the reasoning process and provide mission-appropriate interactions that allow analysts to have a true discourse with their information (Thomas & Cook 2005). Therefore, the visual analytics approach has become an important tool for gaining insights on various data sets. Typically, the visual analytics process (see Fig. 2.5) combines automated analysis and visual means along with human interaction in order to gain knowledge from data (Keim et al. 2010).

Keim et al. (2010) have defined a visual analytics process model that illustrates the relationship between the data, model used by the analytical system, visualization outcome and knowledge retrieved to support decision-making process. The visual analytics process starts with data pre-preprocessing and transformation. Further, either visual data exploration or automated data analysis can be performed. In case of data analysis, a potential analyst can refine the parameters involved in the modeling process, evaluate the findings, and provide new input. When the visual data exploration is performed first, the analyst can evaluate the visualized information, apply different styles, zoom, pan, and request details. Based on the findings from visual exploration the model parameters can be also refined. Therefore, in both cases, the analyst can gain knowledge from data, analytical reasoning, and visualization.

Uncertainty-aware visual analytics

In the literature, several approaches have been proposed to incorporate concepts of uncertainty within visual analytics. Previous research has demonstrated the potential of developing uncertainty-aware visual analytics. Kinkeldey (2014) reported on the development of ICchange, an interactive visual analytics application for exploratory analysis of land cover change. Correa et al. (2009) suggested an uncertainty-aware visual analytics framework targeted for housing data analysis in Boston. Their model is established based on statistical methods of uncertainty modeling, propagation, and aggregation. In complementary work, Bastin et al. (2013) investigated an approach based on the Monte Carlo simulation for uncertainty propagation. This includes tools

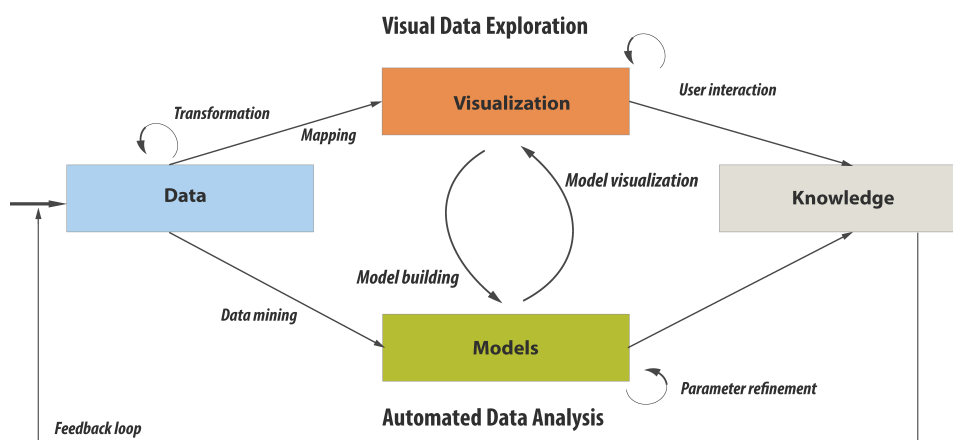


Figure 2.5: Visual analytics process model. Adopted from Keim et al. (2010)

that support elicitation, aggregation/disaggregation, visualization, and uncertainty/sensitivity analysis. One of the first examples of a framework that merges Bayesian statistics and visual analytics is proposed by House et al. (2015). Their solution focuses on the human-computer interaction that helps experts synthesize information in the data, interact with the data, and guide automated, analytical procedures. In one recent study, integration of uncertainty (confidence) visualization with computational methods in a visual analytics application demonstrated how confidence in an estimate on multiple interacting factors (based on weather predictions) can be simultaneously visualized (Kumpf et al. 2018). Squicciarini et al. (2014) have demonstrated an analytical framework called Abuse User Analytics (AuA) aimed to provide information about the behavior of on-line social network users. The AuA processes data users' discussions, and renders information about users' abusive activities. The analysis and visualization implemented within AuA utilize Bayesian Networks to model the users' choices and monitor changes in their behavior in text-based communities.

As it has been mentioned above (section 2.2.3), probabilistic methods, and Bayesian Networks in particular, have been proved as effective in assessing uncertainty in the human reasoning and in the data sources. Consequently, visual analytics approach might facilitate abstraction of numerical details from Bayesian statistics and represent the modeling through qualitative characteristics that promote human reasoning, and yet preserve the semantics underlying Bayesian Networks. The implementation of uncertainty-aware visual analytics using Bayesian Networks that considers spatial data contexts, seems to offer potentially useful tools for spatial data analysis.

This research contributes to efforts that move beyond traditional

intrinsic/extrinsic uncertainty visualization approaches by developing and demonstrating methods that incorporate the visualization within a model-based design using Bayesian Networks to describe data with inherent uncertainty. Thus, the integration of visual analytical tools into contexts such as classification tasks, where uncertainty is an important factor, could provide an additional dimension that may facilitate more informed data exploration and analysis that results in better decisions.

Software packages for Bayesian Network modeling

Development of Bayesian Network software has been successful and several packages (see Fig. 2.6 and 2.7) are available for commercial and educational use. These include GeNIe/SMILE (Koiter 2006), HUGIN Expert (Madsen et al. 2003), BayesiaLab (Conrady & Jouffe 2015), and Netica (Woodberry & Mascaro 2012).

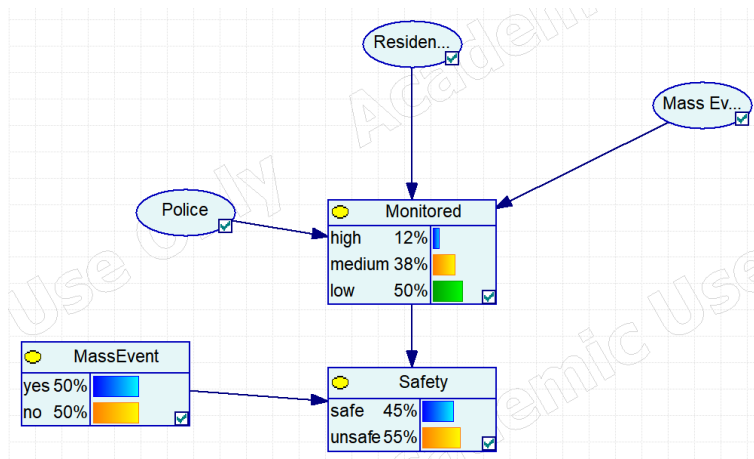
GeNIe is a graphical user interface to SMILE (Structural Modeling, Inference and Learning Engine) (Koiter 2006). The development environment GeNIe was created on the basis of Decision Systems Laboratory, University of Pittsburgh between 1995 and 2015, and since 2015 it has been supported by commercial company BayesFusion. Being free for use for academic purpose, this tool has been widely exploited for various scientific projects. Although GeNIe graphical modeling tool is powerful, it lacks the extension for spatial data.

HUGIN Expert is a general purpose tool for probabilistic graphical models such as Bayesian networks and influence diagrams (Madsen et al. 2003). Along with graphical user interface Hugin Expert offers API provided in the form of a library that can be linked into applications written using the C, C++, Java, or Python programming languages.

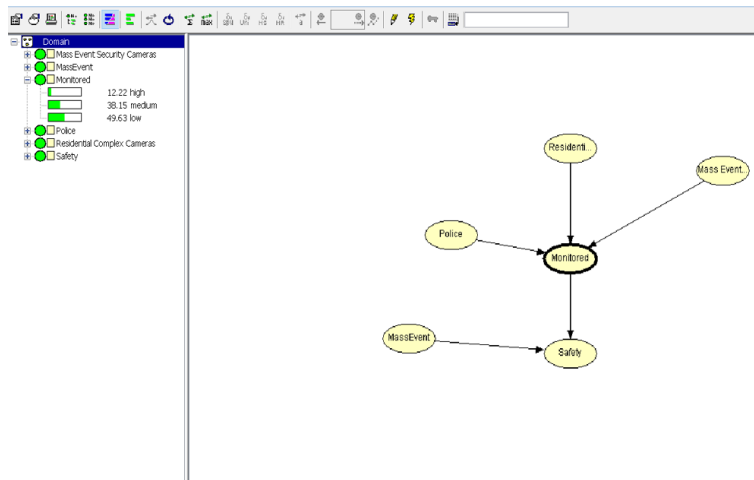
BayesiaLab (<http://www.bayesia.com/>) is a commercial software that facilitates Bayesian inference through visual environment. BayesiaLab provides an effective graphic representation by adopting various symbols to map the changes and characteristics of a network. Only recently BayesiaLab has released an extension that deals with spatial data. The newest software enables the visualization of the values of nodes on Google maps and provides facility to handle optimization problems such as travel, transportation, and logistics (see Fig. 2.8).

Netica is another powerful software incorporating Bayesian Networks for the inference (Woodberry & Mascaro 2012). Due to the interest within GIScience community, Netica extended the software for dealing with spatial data, and they provided facility for dealing with raster files.

Although various tools for Bayesian Network inference exist and provide rigorous modeling procedures, the complexity and lack of support for spatial data prompt the development of further solutions that can assist users with visual and analytical data analysis. These solutions should integrate the computational power of Bayesian Networks and data visualization in visual analytics form, which enables expert input and iterative approaches in the reasoning processes applied to data analysis.

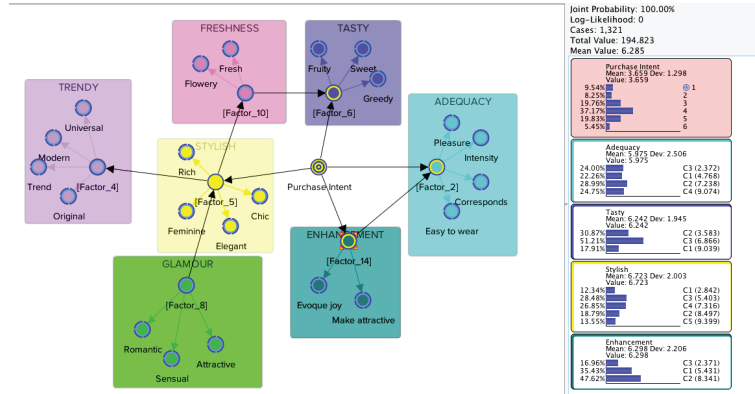


(a) GeNIe/SMILE. The nodes can be visualized via bar chart or icon.

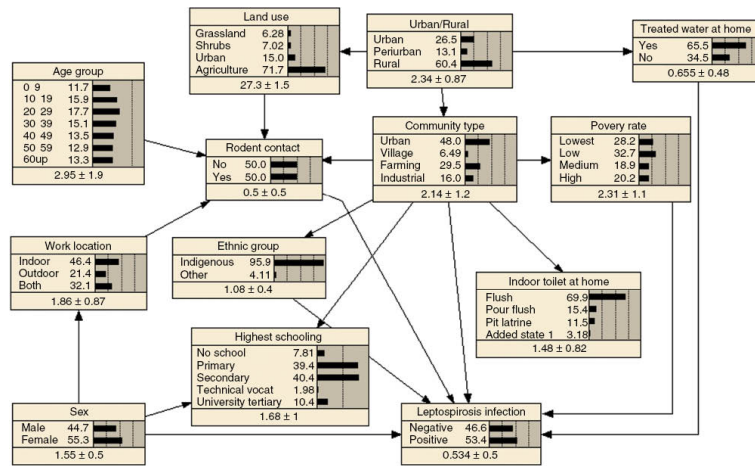


(b) HUGIN Expert

Figure 2.6: Visualization approaches for Bayesian Networks (GeNIe/SMILE and HUGIN Expert).



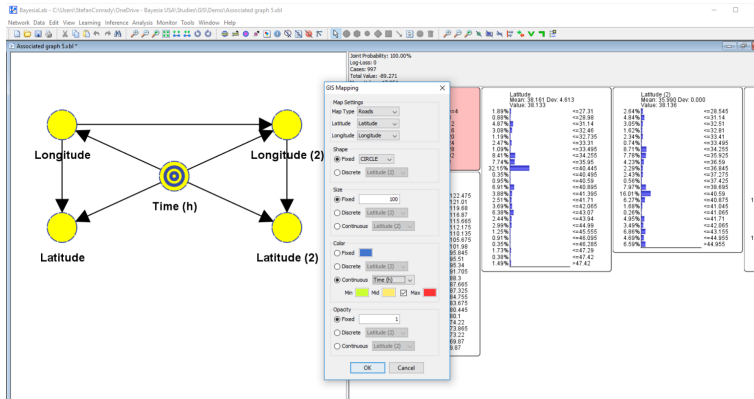
(a) BayesianLab (Conrady & Jouffe 2015)



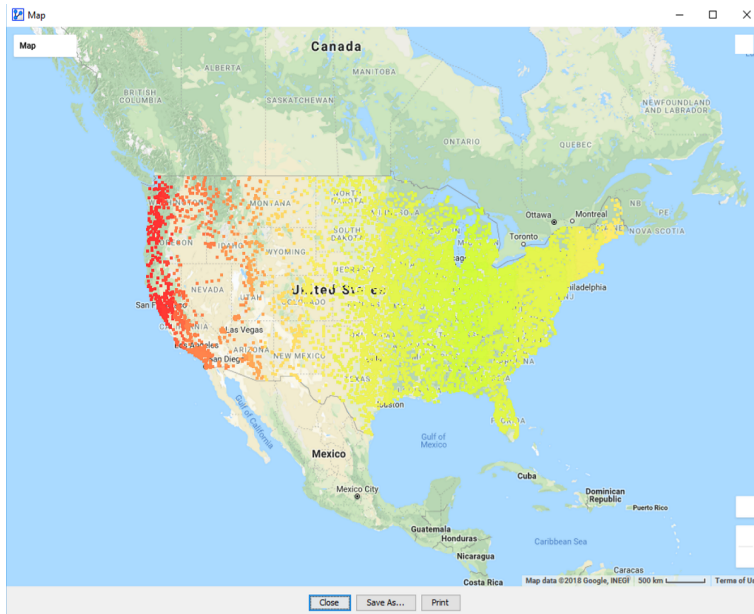
(b) Netica (Woodberry & Mascaro 2012)

Figure 2.7: Visualization approaches for Bayesian Networks (BayesiaLab and Netica).

2.3. Design approaches for promoting analytical reasoning under uncertainty 37



(a) Graph view



(b) Map view

Figure 2.8: BayesianLab extension for spatial data.

2.4 Limitations of existing approaches and open challenges

Over the past decade, a significant body of research in GIScience and cartography was related to uncertainty assessment and its visualization. The researchers typically deal with such complex and challenging topics as climate change, location-based services, environmental impacts, and application of deep learning algorithms for pattern recognition. Explanations and predictions of investigated phenomena, or an area of interest, are fraught with uncertainties, and it is a major challenge for scientists to manage these. Similarly, uncertainty in data and reasoning is also a challenge for cartographers, who visually communicate data to the broader public.

As the saying goes, a picture is worth a thousand words, and so data visualization is a powerful way to convey information in a concise manner and it can explain results of complex data analysis. However, as the amount of data and the complexity of the required analysis grow, visual analytics has become a common tool for gaining insights from the data. And given that uncertainty is an imminent part of data and human knowledge, the necessity of its integration has been highlighted by several authors (Correa et al. 2009, Bastin et al. 2013, Kinkeldey 2014, MacEachren 2015).

Despite the significant advances in the approaches to assessing and visualizing data uncertainty, the challenges for the development of uncertainty-aware visual analytics are seen in the following:

- While data visualization techniques may represent data and uncertainty, a visual analytics approach can enhance user experience with an interactive interface and computational capacity. Therefore, it is crucial to develop effective visual analytics that can accommodate data, visualization, computational power, and human knowledge to support the reasoning process transparently for the user.
- When different analysts explore the same visual representation, they may still have different involvement and, thus, they would draw different conclusions as the previous experience and level of expertise vary. In cases when decision-making is a central purpose of the visual representation, the importance of analyst input should not be underestimated.
- Commonly, deterministic approaches in data analysis do not take variability into account, whereas probabilistic techniques enable uncertainty to be quantified. The potential of the integration of

probabilistic methods within visual analytics in the spatial context can open new perspectives for the reasoning about spatial data.

Chapter 3

Uncertainty in the reasoning process for geospatial data

“ *All our knowledge begins with the senses, proceeds then to the understanding, and ends with reason. There is nothing higher than reason.* ”

Immanuel Kant, *The Critique of Pure Reason*, 1781

This chapter introduces the theoretical foundations of reasoning under conditions of uncertainty uncovering the human and Bayesian theories behind this process. Furthermore, this chapter is mainly focused on the following objective:

To give an overview of human reasoning processes and to test capability of using a probabilistic graphical model, namely Bayesian Network, to represent conditional dependencies among heterogeneous data in order to perform a classification task.

3.1 Introduction

Reasoning is a crucial human brain operation that provides us with methods, algorithms, beliefs, and logic for solving problems and making decisions. In the process of human reasoning, visualization plays an important role, as people often use such various representations as maps, diagrams, charts, and figures

to support and/or generate their hypothesis. By examining the size, shape, orientation, and color people can make judgments in their minds, even though it may be simplified considering assumptions made. Thus, human reasoning is often associated with some degree of fuzziness and uncertainty due to our cognitive biases (see section 3.4.1). The judgments made may be expressed by means of a natural language or visual representation.

Over the last decade, the amount of data generated on the daily basis has increased dramatically due to the acquisition and processing advances of current information communication technology. The data-driven approaches in science can help to bridge the gap between the environment and society; therefore, researchers from different scientific fields have started paying considerable attention to the challenges of working with the data and making sense out of it. By applying sophisticated computational algorithms scientists can extract important patterns from the data and make predictions for future development.

With the scientific advances, numerical approaches to quantify the uncertainty have become prominent, however, the qualitative information and expert knowledge are still of paramount importance when dealing with heterogeneous data. Furthermore, we describe how human reasoning is perceived from historical perspectives and now, and how Bayesian Networks can be utilized to indicate the probabilistic uncertainty as well as entropy when dealing with heterogeneous and spatial data.

3.2 Reasoning under conditions of uncertainty

3.2.1 Human reasoning: historical perspective and current interpretation

The role of human reasoning and certainty has been always a topic of significant importance for philosophers, cognitive scientists, psychologists, and scientists from other fields. And despite the progress in the methods and theories, how human reasoning works is still a puzzle. To give a short overview, some crucial ideas in epistemology that changed the course of philosophy of science are briefly reviewed below.

The French philosopher and scientist René Descartes (1596 - 1650) was a founder of the method in philosophy that defines knowledge in terms of doubt. In Descartes' works, doubt is understood as the opposite of certainty, where the knowledge of the nature of reality was derived from the ideas of the intellect and not from the senses (Newman 2016).

Gottfried Wilhelm von Leibniz (1646 - 1716) also followed the rational approach in his work "Meditations on Knowledge, Truth, and Ideas" (1684) about the possibility and nature of human knowledge. Leibniz addressed a distinction between truths of reasoning (reason or explanation can be discovered by analysis) and truths of fact (reason cannot be discovered through a finite process of analysis) (Look 2017). This distinction defines the nature of human knowledge or cognition. According to Leibniz, human knowledge is intuitive only for primary notations and propositions (Look 2017).

In contrast to the rationalist view, John Locke argued in his work "An Essay Concerning Human Understanding" (1689) that we can know for certain, these would include our own existence or the nature of mathematics and morality (Uzgalis 2018). Locke defined probability as follows: "Probability is nothing but the appearance of such an Agreement or Disagreement, by the intervention of Proofs, whose connection is not constant and immutable, or at least is not perceived to be so, but is or appears, for the most part to be so, and is enough to induce the Mind to judge the Proposition to be true, or false, rather than the contrary." Like Descartes, Locke based his account on intuition, where the connection among ideas is presumed in belief.

David Hume (1711 - 1776) in his most influential work "A Treatise of Human Nature" (1748) proposed an empirical investigation into human nature and argued that a passion, rather than reason controls human choice. He claims that inductive reasoning and our beliefs regarding cause and effect cannot be justified by reason, but by mental habit and custom. According to Hume, everything is subject to some degree of uncertainty, and science has to deal with the fact that absolutely certain is not possible.

The German philosopher Immanuel Kant (1724 - 1804) in his contribution to metaphysics, "Critique of Pure Reason" (1781) and its summary "Prolegomena to Any Future Metaphysics" (1783), distinguished a priori from a posteriori cognition and between analytic and synthetic judgments. According to Kant and his "Critique", a priori is what one knows independently from experience, and a posteriori is a knowledge one obtains from experience. An analytic judgment is explanatory and does not contain new information, whereas a synthetic judgment is one whose assertion includes information not contained in the subject. Innovative ideas from "Critique of Pure Reason" redefined the approach to how knowledge is structured and argued that scientific reasoning is always uncertain, although technical reasoning can be pragmatic.

Rudolf Carnap (1891 - 1970) in his book "Philosophy and Logical Syntax" (1935) redefined a method of logical analysis "to analyze all knowledge". Carnap (1945) addressed the two different probability interpretations: frequency and degree of confirmation. Working on the latter concept, he defined the probability of a statement as the degree of confirmation the empirical evidence

gives to the statement.

Clarence Irving Lewis (1883 - 1964) argued that necessary truths are knowable a priori, which means that we form expectations and make predictions about a future outcome, conditional on actions we might make. Our beliefs constitute empirical knowledge and past experience gives us a good reason (largely inductive) for making these predictions (Hunter 2016). Therefore, Lewis believed that knowledge is possible only where there is a possibility of error. In his work "The Analysis of Knowledge and Valuation" (1946), he distinguished three classes of empirical statements: expressive statements (something certain and present in experience), terminating judgments (statements verified empirically), and non-terminating or objective judgments (certain judgments if terminating judgments accept them).

Hilary Putnam (1926 - 2016) was a Harvard-based philosopher, who contributed to various fields of philosophy of science, and in particular metaphysics, epistemology, logic and mathematics. Although Putnam has revisited his thoughts on earlier made arguments from realism to pragmatism, he defined knowledge as being framework-relative and the idea of truth as related to a set of investigative interests and priorities.

By examining the history of skeptical approaches, we see that it is not easy to define human knowledge. Moreover, emerging methodological advances in psychology and cognitive science offer new perspectives on human cognition, reasoning, and decision-making.

An innovative approach to explaining the reasoning process under uncertainty was proposed by Tversky & Kahneman (1974) and Kahneman & Tversky (1982). In their research (Kahneman & Tversky 1982) they addressed variants of uncertainty through psychological analysis that distinguishes two levels of responses. The first reflects perceptual expectancies and surprise, and, the second represents phenomenological examination. The latter deals with experiences of doubt and uncertainty in judgments of subjective probability. Kahneman and Tversky suggested that the uncertainty in judgments of subjective probability can be differentiated into internal and external attributions (see Fig. 3.1). For example, color, size and texture are normally experienced as properties that belong to external objects, whereas feelings are attributed to internal. Kahneman & Tversky (1982) emphasized that external uncertainty can be approached two ways: (a) a distributional model, where the case in question is an instance of a class of similar cases and its relative frequencies are known or can be assessed; (b) a singular mode, where probabilities can be estimated by the propensities of a particular case in hand. Internal uncertainty is also seen two ways: (a) a reasoned mode, where analysis and weighting of evidence are performed; (b) an introspective, where the judgment is based on an introspective association.

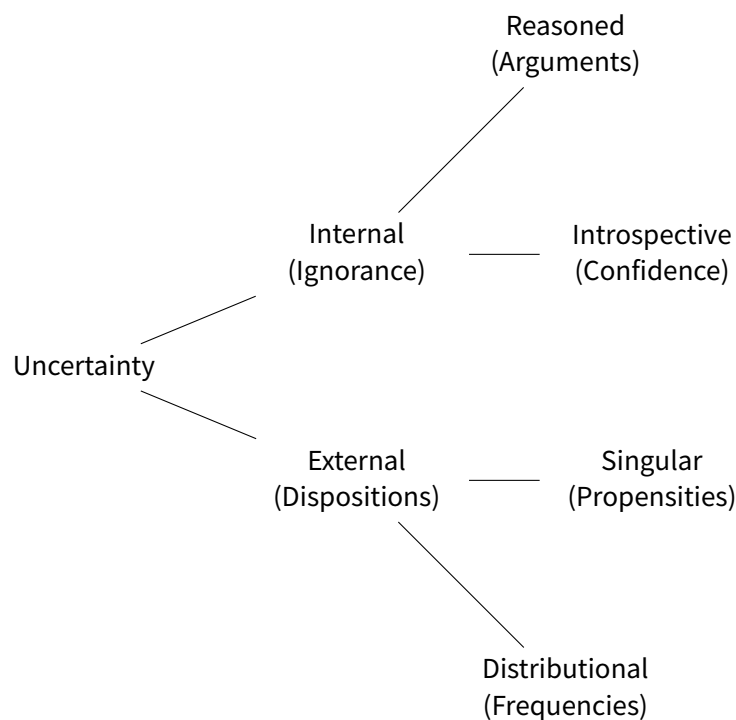


Figure 3.1: Variants of uncertainty suggested by Kahneman & Tversky (1982)

In another study, Tversky & Kahneman (1974) conducted several experiments on how people assess the probability of an uncertain event or the value of an uncertain quantity. In the course of their research, Tversky and Kahneman aimed to discover the heuristic principles and reasoning process based on previous experience and current belief. The results of these experiments revealed three heuristics that are employed to assess probabilities and to predict values. These heuristics include representativeness, availability, and adjustment from an anchor.

The representativeness heuristic determines how people evaluate the probability that an event A belongs to a class or process B. In other words, this heuristic involves judging the probability of an event based upon how similar it is to our actual thinking of such an event. Just like other heuristics, it is a mental shortcut that helps humans make decisions. For instance, people often judge the probability that they can win the lottery next time based on whether or not they won the last game. In reality, the lotto results are not dependent upon each other and winning or losing is random.

The Availability heuristic involves judging about frequencies or probabilities of a particular event based upon how fast and well we can recall similar events. For example, people might believe that catastrophes are more common than other

transportation accidents because they can immediately recall some examples of crashes.

The adjustment from an anchor explains how people access numerical predictions and why they tend to overestimate the probability of conjunctive events and to underestimate the probability of disjunctive events. Tversky & Kahneman (1974) suggested that the stated probability of the elementary event provides a natural starting point for the estimation probability of any event, which means that the initial anchor gives a great deal of influencing on the future assessments.

The studies of cognitive heuristics reveal how individuals reason and make decisions, and they can provide a solid ground for harmonizing the perceptual features of visualizations. Although Tversky & Kahneman (1974) examined their ideas on various common situations from everyday life, heuristics can be also applied to visual and cartographic design, as often people have to take a decision based on presented visual interface, where some information is missing or uncertain. When it comes to sophisticated data or complex analytical task personal knowledge and experience are often key features of heuristic judgments. Technological advances offer computational power to emulate the reasoning process, decision-making, visual perception, and linguistic comprehension. MacEachren (2015) addressed the importance of integration of the principles about judgments under uncertainty within visual analytics domain.

3.2.2 Bayesian epistemology

Bayesian epistemology emerged as a philosophical movement in the 20th century, though its main formalization can be traced back to Reverend Thomas Bayes (1701 - 1761) (Talbot 2016). Bayesian epistemology is a formalization of inductive logic in combination with a new inductive test for epistemic rationality, which involves updating on received new evidence and examining the reasoning outcome. Reverend Thomas Bayes found a simple mathematical formula (see Eq. 3.1) used for estimating conditional probabilities that now plays a central role in Bayesian approaches for data analysis.

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}, \quad (3.1)$$

where likelihood $P(Y|X)$ represents how probable is the evidence given that the hypothesis is true; prior $P(X)$ shows how probable is the hypothesis before the evidence observation; marginal $P(Y)$ represents how probable is the new

evidence under all possible hypothesis; and posterior $P(X|Y)$ represents how probable is the hypothesis given the observed evidence.

Bayesian approach to epistemology is based on principles, methods, and problem-solving mechanisms that are consolidated by three main aspects (Colombo et al. 2018): (1) uncertainty should be captured by a real-valued function that measures degrees of belief; (2) degrees of belief, at any given time, should satisfy the axioms of probability theory; (3) degrees of belief, represented by determinate probabilities, should be updated in the light of new information, typically by the canonical rule of conditionalisation.

Progress in the development of computational statistics in the last decade has offered new tools for Bayesian inference and prediction, that are capable to embrace the complexity of environmental modeling under uncertain conditions. Imagine a situation when we have two or more alternative responses to an environmental change. A farmer with a large land has noticed that his land is affected by soil degradation. The major problems for soil degradation might be caused by water and wind erosion, soil fertility decline, water-logging, salinization, lowering of groundwater recharge. Besides, there are underlying causes of degradation as for instance, land shortage, economic pressure, poverty, and population increase. If the reason behind the soil degradation lies in improper crop rotation which contributes to soil fertility decline, then the land management strategy considering this factor might be changed accordingly. But what if both improper crop rotation and deforestation of nearby territories that leads to water erosion, contribute to the soil degradation? A decision when several aspects might have the influence is uncertain, and one should take into account possible outcomes weighted by their impacts. Therefore, there is a great need to address the importance of effective decision-making considering the uncertainty quantification. From a practical perspective, Bayesian approaches provide an efficient inferential framework that allows experts to updating knowledge considering the expert input.

3.3 Bayesian Network for reasoning under conditions of uncertainty

This research is focused on a particular type of probabilistic models named Bayesian Networks. When the term Bayesian Network was introduced by Pearl (1985), the main motivation was to develop a model for human inferential reasoning. Being important in different research fields, Bayesian Networks are both mathematically rigorous and intuitively understandable (Kenett 2016). Bayesian Networks are commonly used for applications where heterogeneous

data is involved. Just to name few examples, Bayesian Networks are used in the analysis of airport check-in processes (Pietro et al. 2017), agriculture (Drury et al. 2017), urban area's vulnerability to flooding (Abebe et al. 2018), and prediction of natural disasters (Li et al. 2010). Taalab et al. (2015) studied Bayesian Network application for prediction of soil properties and showed that it is an effective technique for quantitative and qualitative prediction and a reasonable alternative to black-box data mining techniques.

Before going into further details some key definitions are given in sections 3.3.1, 3.3.2, 3.3.3 and 3.3.4.

3.3.1 Events and their probabilities

In probability theory, an event is defined as an outcome of an experiment. When you toss a coin, the outcome is an event. The same can be valid for a spatial event in a broad sense. For instance, the land cover product Globeland30 shows forest as a land cover class assigned to a given location, and this class at this location can be considered as an event.

The probability of an event may be seen from three main fundamental perspectives that are commonly accepted in modern statistics:

- frequentist interpretation views the probability from the point of view of the frequency of occurrence of the event in a sample;
- propensity interpretation determines the probability by physical properties;
- subjectivist interpretation determines the probability as a subjective personal measure of the belief in that event.

The implication of the subjectivist view on probability is adopted in this work as it allows analysts to deal with decision analysis. In other words, the subjectivist interpretation of probability is defined from the perspective of a personal belief, which can vary among different experts or systems involved in the analysis.

An event can be a combination of different events. If we are interested in probability of union of two events (Fig. 3.2(a)), we consider combined probability of these events $P(Z) = P(X \cup Y)$.

When the main interest is the probability of events occurring together, we are interested in the intersection of these events $P(Z) = P(X \cap Y)$ (Fig. 3.2(a)). And eventually, if events are mutually exclusive, this combination is disjointed (Fig. 3.2(c)).

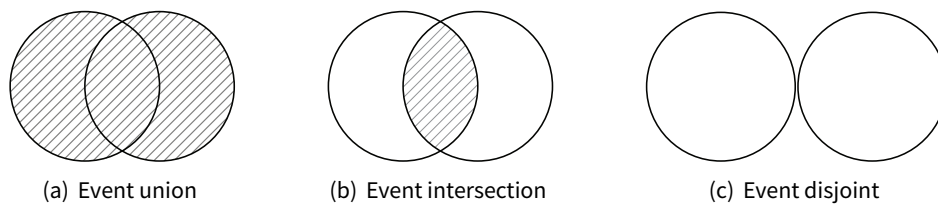


Figure 3.2: Combination of different events.

Joint probability

A joint probability refers to a statistical estimate that measures the likelihood of two events occurring together (Fig. 3.2(b)). For instance, joint probability is the probability of event X occurring together with event Y denoted as $P(X, Y)$, $P(X \cap Y)$ or $P(X, Y)$, which reads as the joint probability of X and Y .

It can be quantified as a number in an interval between 0 and 1 inclusive, where 0 is assigned to an impossible chance of occurrence and 1 indicates the certain outcome of an event. An example joint probability distribution for variables Snowing and Windy is shown in Table 3.1. The probability of wind and no snow is 0.10 (or 10%). As the joint probability indicates only events that happen at the same time, it is only applied to situations when more than one event can occur. The joint probability values can be summed to one for discrete variables. If two variables are independent (i.e. unrelated) then $P(X, Y) = P(X)P(Y)$.

Table 3.1: Joint probability for an event intersection.

Snowing	Windy = False	Windy = True
False	0.70	0.10
True	0.1	0.1

Conditional probability

A conditional probability refers to a statistical estimate that measures the likelihood of an event based on the occurrence of another event denoted as $P(X|Y)$. Conditional probability is estimated based on multiplication of probabilities of the preceding event by the updated probability of the succeeding or conditional event. For example, the probability of Windy being True, given that Snowing is True can be equal to 50%. This would be denoted as $P(Windy = True|Snowing = True) = 50\%$.

Marginal probability

A marginal probability refers to a statistical estimate that measures the probability of an event occurring unconditionally from other events. For instance, in order to know the probability of snowing (from the example above), we need to access the probability of $P(\text{Snowing})$ and sum up all the probability values for Snowing = False, and Snowing = True (see Table 3.2). For discrete variables, marginalization is performed by means of a probability summary, and for continuous variables by means of integration.

Table 3.2: Marginalization for a variable.

Snowing	Windy = False	Windy = True	Sum
False	0.70	0.10	0.80
True	0.1	0.1	0.2

	Snowing = False	Snowing = True
P (Snowing)	0.8	0.2

3.3.2 Bayesian updating

Bayesian Networks rely on the conditional probabilities of events that can be updated if new information is available, as for example to represent the conditional probability such statement can be made: given an event X , the probability of the event Y is z . This statement can be represented in mathematical notation as $P(X|Y) = z$. This statement is true if events are independent of each other.

The basic rule of conditional probability is $P(X|Y)P(Y) = P(X, Y)$, where $P(X, Y)$ is the probability of the joint event when both events have occurred. For example, one's belief in whether or not a city district has high criminality is independent of his belief in whether or not it has a high percentage of families that earn less than a minimum wage. However, one's belief in whether or not a scarcity of job opportunities is not conditionally independent of either of these two factors since either could be stimulant for criminal behavior. Additionally, if one happens to know that a district has high criminality level, his belief in whether or not a high percentage of families that earn less than a minimum wage it is no longer independent of his belief in whether or not there is a

scarcity of job opportunities. If the job market is poor that gives one a reason to believe that the criminality is economically induced.

Updating mechanism in Bayesian Networks is based on Bayes' theorem (see Eq. 3.2). Using this theorem, the probabilities of event $P(X)$ are updated based on the new evidence related to event Y , in other words, it describes the probability of an event based on the prior knowledge about events related to the event that has been analyzed. The reasoning is based on the subjective degree of belief (posterior probability). Using probabilistic theory allows to reason about events and make the best decision from the probabilistic point of view.

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} \quad (3.2)$$

Based on equation 3.2 the posterior probability of $P(X|Y)$ can be estimated using prior beliefs in the occurrence of event $P(X)$ and event $P(Y)$ and the probability of event Y given that X has occurred $P(Y|X)$. This mechanism is the basis for Bayesian inference. A substantial benefit of Bayesian Network is also seen in the ability of this approach to combine frequentist and subjectivist views of probability, thus the numerical probabilities can be extracted from the databases and combined with the expert judgments.

3.3.3 Bayesian Network structure

The Bayesian Network is a probabilistic graphical model built on Bayes' theorem (see equation 3.2) and described by a mathematical model with qualitative and quantitative components such as Directed Acyclic Graph (DAG) and Conditional Probability Tables (CPTs), respectively (Darwiche 2008).

Directed acyclic graph

The effectiveness of the Bayesian Networks lies in their graphical model structure, where the qualitative component is realized within a DAG. The nodes in the DAG represent random variables and edges connect the variables. The terms node and variable are used interchangeably. The nodes in the network are commonly drawn as circles (see Fig. 3.3) labeled by the variable's names (Kenett 2016). The edges define probabilistic relations among the nodes.

Consider an example of a Bayesian Network given in Fig. 3.3, a network structure is represented with nodes of two types, so called parents and children. For instance, the node X_1 is a parent node with two children X_2 and X_3 .

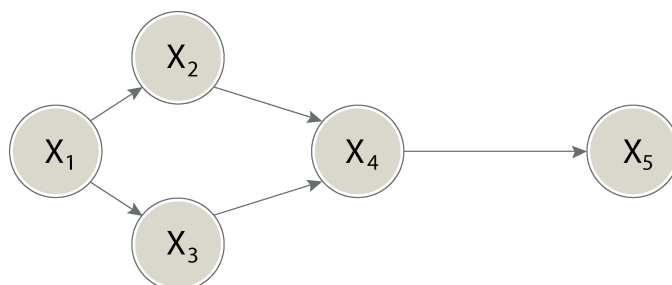


Figure 3.3: A representation of causal dependencies among five variables using Bayesian Network. Nodes represent variables and edges represent dependencies among them.

Joint probability distribution

The DAG defines a factorization of the Joint Probability Distribution (JPD) of $Z = X_1, X_2, \dots, X_n$, often called the global probability distribution, into a set of local probability distributions, one for each variable (Scutari 2009). The factorization is performed applying Markov property, which characterizes the process when the conditional distribution at each node depends only on its parents. The equation 3.3 is valid for discrete variables, and 3.4 for continuous (Scutari 2009), and define the probability of a state of a given child (X_i) node quantified under conditions of the states of the parent nodes Π_{X_i} .

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \Pi_{X_i}) \quad (3.3)$$

$$f(X_1, \dots, X_n) = \prod_{i=1}^n f(X_i | \Pi_{X_i}) \quad (3.4)$$

D-separation

D-separation (from direction-dependent separation) is a property of DAG that indicates whether a node X is conditionally independent of another node Y , given a third node Z . Considering the example given in Fig. 3.4, the nodes X and Y are not connected, as there is a node Z in between. Besides, between X and Y there is another node A .

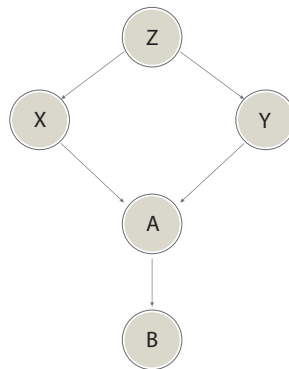


Figure 3.4: An illustration of d-separation (direction-dependent separation) in the network.

Conditional probability tables

The quantitative component of the Bayesian Network is determined by CPTs. Each CPT represents a table with a list of possible states for a node. CPTs provide a mechanism to replace the uncertainty state with a numerical value in the interval $[0, 1]$ representing degrees of truth, belief or plausibility (Ortega 2010). The size of CPTs depends on the number of parents and grows exponentially. A large number of nodes require additional data and/or expert knowledge. CPTs express the probability of the state of each variable given its parents and essentially represent the strength of the belief in the causality (Luxhøj 2014) defined by the experts or learned from the data structure.

Learning algorithms

There are various structures utilized for the Bayesian Network modeling. Naive Bayesian Networks are typically used for classification under assumptions that all the variables are independent of each other and information about the predicted variable is given. Naive Bayesian Network might be updated with the training data to improve the prediction outcome, and such a network would be called naive optimized Bayesian Network. More complex learning algorithms could be divided into constraint-based and score-based algorithms (Scutari 2009). The Bayesian Network structure can be also defined based on an expert knowledge.

3.3.4 A measure of uncertainty

In order to quantitatively describe data uncertainty, various measures have been proposed. These include scalar values like probability, error percentage, distance (e.g. from the true value), standard deviation (Griethe & Schumann 2006), and the Shannon diversity index which is a quantitative estimator of complexity (Jost 2006). According to Jost (2006) the Shannon Index (also called Shannon-Wiener Index) is the most common diversity measure and it plays a central role in information theory as a measure of information, choice, and uncertainty (Spellerberg 2008). This index represents entropy, giving uncertainty as to the outcome of a sampling process. In other words, the higher the entropy (the higher the uncertainty) of a given location, the higher the diversity across all the classes at this location.

$$-\sum_{i=1}^n P_i \cdot \ln(P_i), \quad (3.5)$$

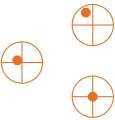











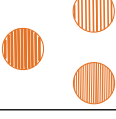


where p_i is the probability associated with the state i of a target node. One of the advantages of using the Shannon Index of diversity is that it represents a large amount of information in one expression (Spellerberg 2008). Therefore, when adopting this measure of uncertainty it is possible to communicate information by giving only one estimate (see equation 3.5).

3.4 A visual approach to human and Bayesian reasoning

3.4.1 Visualization and heuristics

With an increasing amount of data, the information visualization has become a conventional way of communication for wide public and research community. The information visualization, as well as geospatial information visualization, scale down the data into meaningful and abstract forms in order to convey the relevant, significant, and informative phenomenon. Bertin (1983) proposed a set of visual variables that enables a visual communication researchers to represent the information about quantitative and qualitative data characteristics into a visual format that allows the users to understand it intuitively. Bertin identified seven "retinal variables" (Table 3.3): (1) location, (2) size, (3) color hue, (4) color value, (5) grain, (6) orientation, and (7) shape. Later on, this semiotics was extended with such variables as (8) color saturation (Morrison 1974), (9) arrangement (Morrison 1974), (10) clarity (fuzziness) (MacEachren 1992), (11) resolution (MacEachren 1992), and (12) transparency (MacEachren 1992).

Table 3.3: Visual variables.

	Visual variable	Point	Line	Area
1	Location			
2	Size			
3	Hue			
4	Value			
5	Grain			

6	Orientation			
7	Shape			
8	Saturation			
9	Arrangement			
10	Fuzziness			
11	Resolution			
12	Transparency			

Practically, the visual variables are utilized to convey information in an efficient manner allowing the user to derive meanings from the visual representations. Analyzing a list of coordinates, reading a table with hundreds of values, interpretation of a bar chart or localized points on the map require concentration and some mental work. Though, some of the interpretation tasks require more effort than others. And being familiar with visualization means, one can interpret length of a bar charts intuitively in comparison to analyzing a long list of values, where a cognitive task is much harder.

Thus, the visual representations can lessen the cognitive effort that readers need to expand when interpreting complex data by converting that data into visual genres (Jones 2015). Despite the beneficial aspect of the visual

representation, the human visual system may also experience perceptual conflicts due to physical illusions (disturbance of light, or our eyes) or/and cognitive misinterpretations (see an example on Fig 3.5).

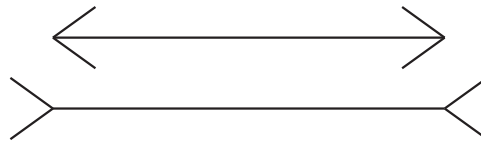


Figure 3.5: The Müller-Lyer illusion. Despite the equal length these lines are perceived to be different.

Information visualization and principles of judgments under uncertainty present a challenge: **How can we apply the studies about heuristics to the visualization of data (in particular, geospatial data)?**

Jones (2015) addressed this issue through exploration of heuristics of representativeness, availability, and affect in information graphics. The representativeness heuristic is applied when one explores a visual representation and interpret this visual design based on the similarity to other designs of the same genre (Jones 2015).

For instance, visualization of hurricane data (see example in Fig. 3.6) has been discussed among visual communication researchers, as many people misinterpret the probabilistic concepts that are being communicated by the error cone (Broad et al. 2007, Cox et al. 2013). The researchers reported that people tend to perceive the cone coverage as an exaggerated chance of being in the hurricane's path, whereas the outer side of the cone provided people with a false sense of security (Cox et al. 2013). This interpretation seems to follow the representativeness heuristic, where a typical reader might assume that the map is like others of the same type shows definitive classes.

Klayman & Brown (1993) and Shafir et al. (1993) argue that humans are not intuitive Bayesian statisticians, thus the affect heuristic (adjustment from an anchor) explains why people tend to overestimate the probability of conjunctive events and to underestimate the probability of disjunctive events. A visualization example on reasoning about conjunctive or disjunctive events can be given using adjacent displays.

The availability heuristic considers the available options in our memory and imagination in order to make judgments about particular events, or in this discussion visual representation. Thus, this heuristic type can lead us to react, at least at the beginning, to the information that a visual representation conveys and examine it based on the available knowledge about frequencies and probabilities of a particular event based upon how fast and well they can recall similar events. For instance, biases due to the retrievability of instances

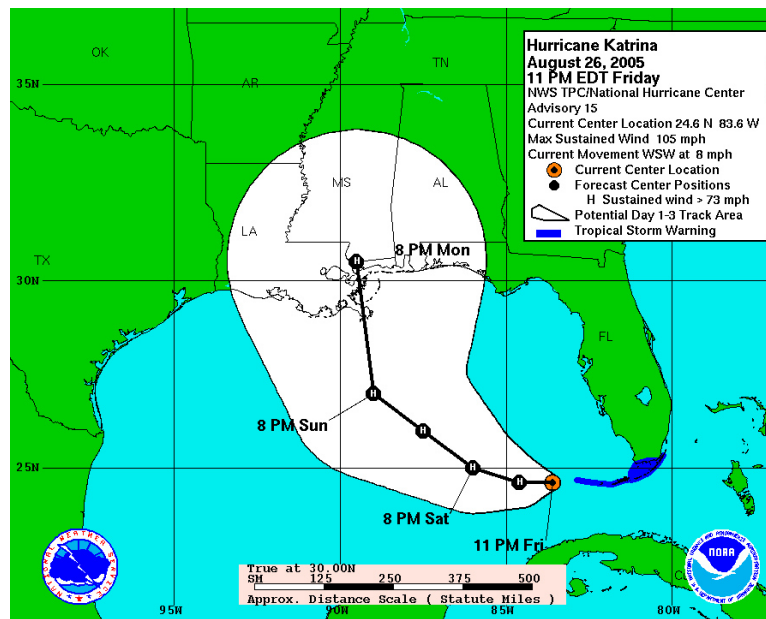


Figure 3.6: Error cone represents predicted hurricane track. Hurricane Katrina forecast path on August 26, 2005 at 11 p.m. EDT. Source: National Hurricane Center.

represent how the size of a class is judged by the availability of information that can be easily retrieved (Tversky & Kahneman 1974). Such classes would appear more numerous than other. When a certain issue is visualized, it can be positively framed, therefore, people would see more evidence on this. Although further studies are required to describe this heuristic type applied to visualization perception, an example of an election map might help to uncover some insights. According to availability heuristic, we overestimate the frequency and magnitude of events that happened to occur more vivid. For instance, when one has to decide to which candidate to give his/her vote, the opinion will be formed based on the available information about the candidate, exposure to the mass media information and personal experience. When a prediction of winning is visualized with chances given without additional information of the possible distribution, it can appear more numerous than other (see Fig. 3.7). The election forecast, in this example, is visualized according to the results of a simulation based on the poll data with some degree of uncertainty (according to fivethirtyeight.com) due to ambiguous poll data, undecided voters, and pooling errors. Despite the data uncertainty, Fig. 3.7 gives only a little information about an alternative outcome, thus provides the users with a false impression about the election results.

Prevalent number of visualizations and maps have been created without considering the uncertainty component, however, it is an inevitable issue if we

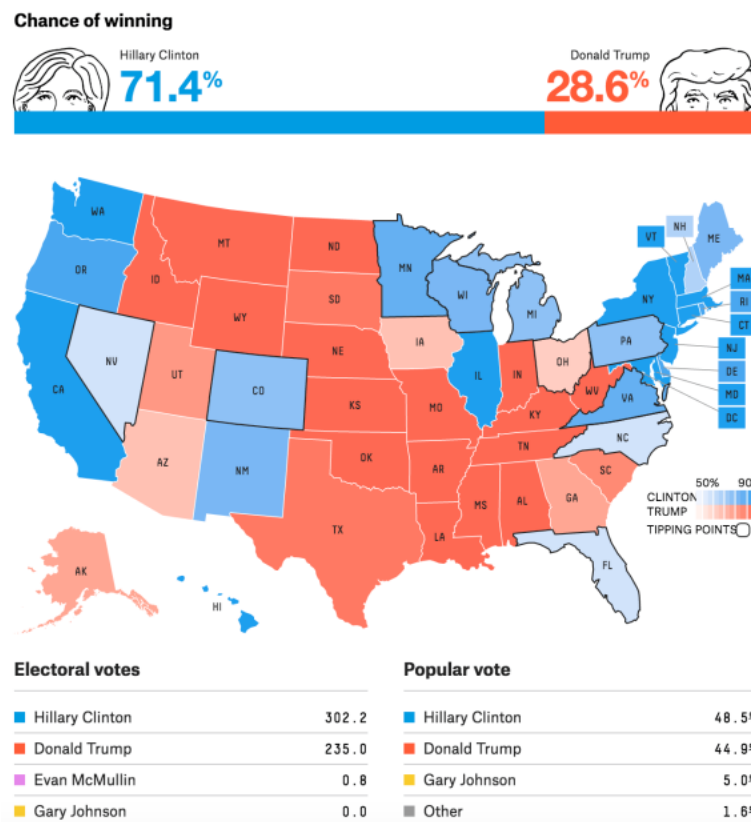


Figure 3.7: Availability heuristic applied to the visualization domain. Source: FiveThirtyEight.com.

want to improve the way how people understand the statistics. In fact, a wide audience that consumes data visualization is exposed to probabilities and statistical inference on a daily basis through mass media, newspapers, and mobile applications. The misunderstanding of the processes behind summary statistics and statistical inference can lead to misunderstanding of the data visualized if the uncertainty is not communicated. One of the potential means for addressing this problem is to integrate human and Bayesian reasoning within a visual environment.

3.4.2 Visualization and Bayesian reasoning

Previous studies indicate that despite the commonly applied task of updating knowledge in the light of new evidence, people may perform inadequately. For instance, we make a subjective probability judgment about the time we may spend on the road going from A to B. And imagine you got informed about a major car accident happened, how would these information change your probability judgment? This would be a classical task of Bayesian reasoning. A commonly used example is reasoning about diagnosis paradigm, which proves that most people fail to give a correct answer in a Bayesian reasoning task (as visualized in Fig. 3.8). Casscells et al. (1978) conducted a survey with a question: "If a test to detect a disease whose prevalence is 1/1000 has a false-positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?". The most frequent answer was 95%, whereas the correct answer is about 2%. If we consider that 1000 people are tested, 1 person has the disease, and the test shows a positive result in 50 cases (5%). Therefore, the probability that a person with a positive result has the disease is 1/50 (2%). As the Bayesian reasoning approaches have opened up new perspectives for the evidence-based decision-making, information visualization community proposed to introduce a visual support for probabilistic data in order to effectively communicate Bayesian algorithms.

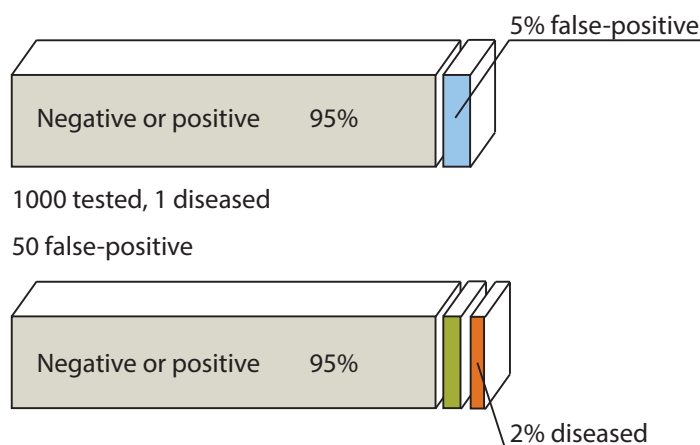


Figure 3.8: Reasoning about diagnosis paradigm

Integrated pictorial representations such as iconic symbols (Fig. 3.9) (Cole 1989, Brase 2009, Ottley et al. 2016)), Venn circles/Euler diagram (Fig. 3.10) (Brase 2009, Rodgers 2014), and Venn circles with dots are ported to be an effective means of representation for Bayesian reasoning problems (Brase 2009). Although, according to studies (Brase 2009), people perform better at Bayesian

reasoning tasks when they deal with pictorial representations as icon arrays in comparison to continuous fields like Venn (Euler) diagram. Another way to represent possible scenarios from a given state is decision tree visualization (Gigerenzer & Hoffrage 1995, Martignon & Wassner 2002, Friederichs et al. 2014). Each edge of the decision tree (Fig. 3.11) corresponds to a decision made, where the probability value can be indicated. Each node in the decision tree is responsible for representing a specific prediction.

Conditional probability judgments are often presented by means of a table (Fig. 3.12), also called contingency or conditional probability table (Cole 1989). In such table, probabilistic values are usually given in numeric or frequency format. In a similar manner, a heat map visualization approach can represent probabilities using color value and hue (Fig. 3.13). A mosaic plot (Fig. 3.14) is another commonly used graphical method that enables visualization of frequencies of a contingency table where each box has a proportional size according to the probability value (Chiang et al. 2005). To describe a simple Bayesian task, a "beam cut" diagram (Fig. 3.15) can be used where each "slice" represents the information in proportion (Gigerenzer & Hoffrage 1995). The resulting proportion can also be visualized using probability curves (Cole 1989) that indicate two populations and a vertical line that shows the threshold for the test score (Fig. 3.16).

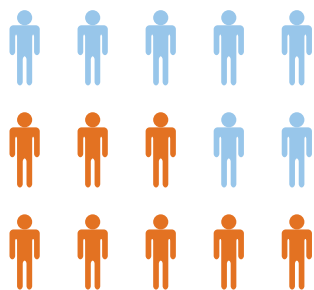


Figure 3.9: Iconic symbols

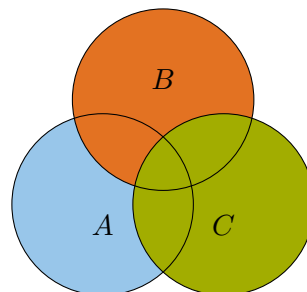


Figure 3.10: Venn (Euler) diagram

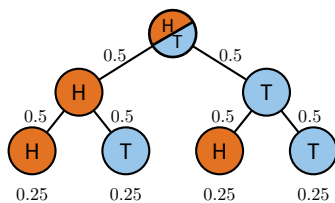


Figure 3.11: Decision tree

	A	B
C	0.70	0.30
D	0.1	0.9

Figure 3.12: Conditional probability table

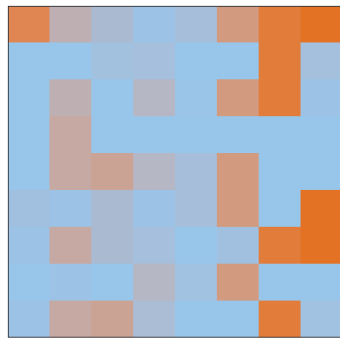


Figure 3.13: Heatmap

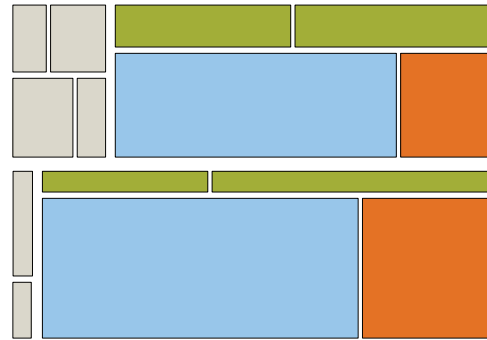


Figure 3.14: Mosaic plot

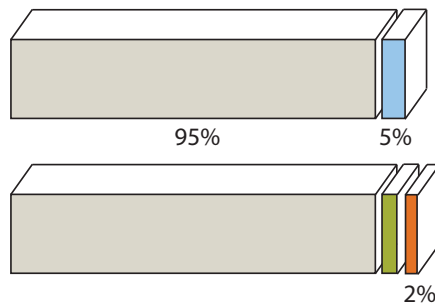


Figure 3.15: "Beam cut" diagram

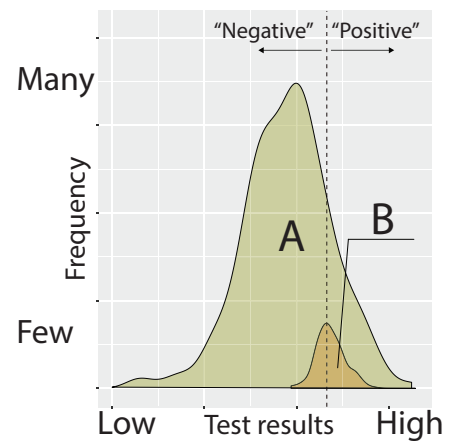


Figure 3.16: Probability curve

3.5 Summary

This chapter broke down the topic of uncertainty in the reasoning process into foundations of human and Bayesian reasoning, and summarized the visual approaches to support the reasoning process.

This chapter introduced approaches for computational reasoning under uncertainty and provided theoretical foundations for a method of Bayesian Networks that is adopted for the visual analytics development (see Chapter 4).

Despite the fact that the Bayesian Networks are a commonly applied modeling technique in different applications, the visual exploration is limited to cause-effect relationships among the variables, and only scant attention has been paid to the development of visualization support for the spatial component of the data. Therefore, the development of visual interfaces that allow users to iteratively deal with Bayesian Network analysis in a spatial

context might enhance the usability of this method and open new perspectives for application of Bayesian reasoning in GIScience and cartography. The next chapter will reveal how the computational capacity of Bayesian Networks can be integrated with visual interfaces in order to support data analysis in spatial context.

Chapter 4

Visual analytics framework for supporting reasoning under uncertainty

“ *Bad reasoning as well as good reasoning is possible; and this fact is the foundation of the practical side of logic.* ”

Charles Sanders Peirce, 1877

This chapter takes a step towards uncertainty-aware visual analytics that integrates a Bayesian Network for the reasoning about geospatial data. And, thus, it is intended to address the following research objective:

To develop a visual analytics framework that can facilitate the understanding of data and uncertainty in the reasoning process using Bayesian Network.

4.1 Introduction

Plato's cave allegory (Plato 380 B.C.) gives a vivid illustration of the uncertainty in the human perception, according to which, people live in the real world like in a cave where they can only believe to their senses to understand the true reality. They can judge the true world by the vague shadows on the walls of the cave. In other words, we can only rely on portions of information and many decisions we make are established on beliefs concerning the likelihood of

uncertain events (Tversky & Kahneman 1974). Apart from the daily uncertainty that human beings deal with, it is also a natural outcome of scientific research and it is an inevitable characteristic of data handling processes. Although data uncertainty and its visualization have been the focus of a substantial body of research, the research outcomes have been limited due to a lack of a complementary focus on uncertainty in reasoning processes within visualization and visual analytics domain (MacEachren 2015).

Analytical reasoning is a crucial component of visual analytics along with computation and interactive visual representation. The goal of analytical reasoning is to gain insight from data, and visual analytics facilitates this process and enables human reasoning about complex problems through visual interfaces and computational methods that can process large, messy, and heterogeneous data (Keim et al. 2010). Thomas & Cook (2005) have underlined that visual analytics facilitate computational support for human reasoning realized through software development.

Significant research has been conducted to integrate statistical methods in the interactive environment of visual analytics where data visualization provides support to analysts in understanding and exploring the data. However, much of the data explored with visual analytics is inherently uncertain due to limits of our knowledge about a phenomenon, randomness and indeterminism, and vagueness. And since uncertainty and its complement certainty are fundamental parts of any analytical or reasoning process, it plays a crucial role in the visualization process as it adds a cognitive constraint in the visual perception (Zuk & Carpendale 2007). Accordingly, the impact of uncertainty in reasoning process has a high potential to be exploited within the visual support. Analog to human operators who are able to analyze the data based on incomplete and noisy pieces of information, Bayesian statistics enables the systematic probabilistic inference where estimation can be performed on uncertain or incomplete data.

Moreover, uncertainty in the analytical reasoning might drive an interest to explore the visualized data through an iterative mechanism according to which the visualization outcome is empowered by the users and is represented in a geographic context. Hence, the combination of visual analytics, Bayesian models, and geospatial context can be seen as a constructive environment that provides potential analysts with the interface that promotes choice and quantifies uncertainty in the reasoning process. The Bayesian Networks, as a type of probabilistic graphical models, provide a powerful technique for the reasoning under uncertainty where both computer and expert input can be taken into account. The mathematical basis behind Bayesian models (see the Eq. 3.2) facilitates the assessment of an independent event occurrence given the observed information. While common software packages are focused on the computational capacity and internal improvements of the processing

algorithms, the geovisualization support can complement the inferential process. Furthermore, we set requirements for the further development and describe visual analytics and graphical set-ups that contribute to the framework of Bayesian Network-enabled visual analytics for reasoning under uncertainty on geospatial data.

4.2 Requirements for visual analytics when using a probabilistic model on geospatial data

As it has been underlined in the Chapter 3, Bayesian Networks are graphical models that provide techniques for reasoning in a consistent and mathematically rigorous manner under conditions of uncertainty. The integration of Bayesian Networks in visual analytics with a geographic context where the user input is considered for the inference aims at:

- Integrating expert knowledge and reasoning in the states of uncertainty in an interactive and iterative manner;
- Generating new knowledge from the inference outcome visualized on a digital map;
- Putting emphasis on the visual significance of reasoning under uncertainty.

From the above discussion, we set the following requirements for the analytical reasoning process on geospatial information when using a probabilistic model:

- The Bayesian Network needs to be implemented in a way that it performs inference on the data jointly with the expert input;
- The visual analytics needs to promote the awareness that the data can be interpreted with certainty levels, and that the alternatives can be specified;
- The visual analytics interface needs to represent an efficient tool where the data can be interactively integrated within its geographic extent.

To meet these requirements we propose to provide following visualization tools:

- Data model view: visual representation of the cause-effect relationships between observed and unobserved variables;

- Computational view: visual representation of the model parameters which can be edited by the analyst. This view facilitates the iterative data exploration with formulated subjective beliefs about the data;
- Map view: visual representation of the reasoning outcome in the geographic context with optional overlays with other related data sets that can support the knowledge discovery.

4.3 Visual analytics set-up

4.3.1 Visual analytics process

As suggested by Keim et al. (2010), the visual analytics process combines automatic and visual analysis methods along with human interaction. The process aims to gain new knowledge through data exploration and support for analytical reasoning. Building upon an overview of the visual analytics process presented by Keim et al. (2010) (see Fig. 4.1), we propose a visual analytics approach for spatial, heterogeneous data analysis that uses integrated Bayesian Networks to support reasoning about spatial patterns in the context of uncertain conditions. More specifically, we address the issue of decision-making under uncertainty when performing a classification task. When decision makers deal with classification problems, they often quantify the likelihood of a given event on the basis of their personal knowledge. The Bayesian Network-enabled visual analytics proposed here can facilitate interaction between the analyst and a probabilistic model through visual interface supporting expert input. Moreover the approach presented here facilitates the iterative workflow between the user and visual content through an interaction with data, visualization and models in order to extract knowledge.

This research builds upon the related work and integrates Bayesian Network within a new environment for analysis of spatial heterogeneous data where a user is placed in the center of the analysis. The visual analytics approach proposed here combines the analytics, spatial data, uncertainty in the reasoning, and Bayesian inference in one visual interface.

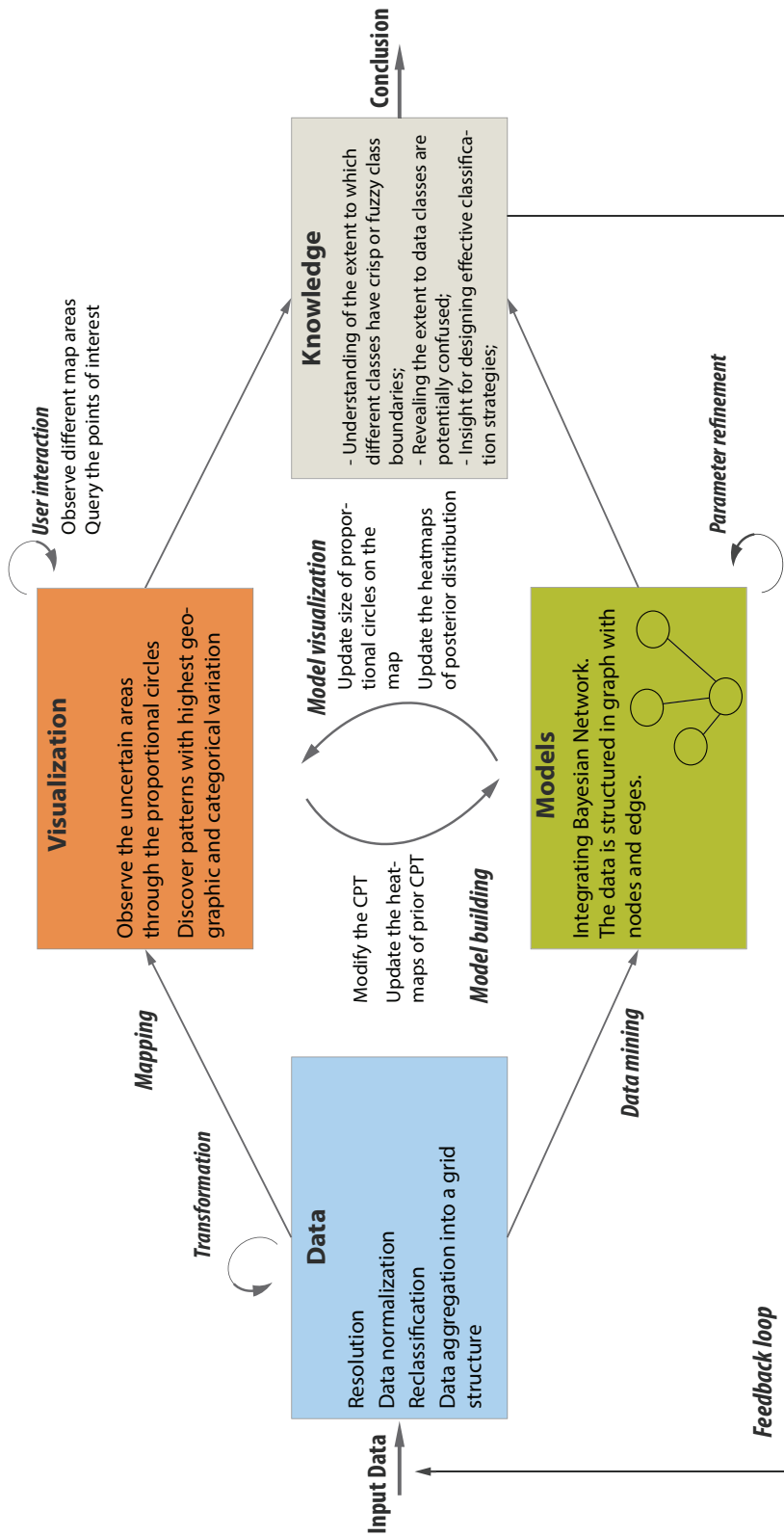


Figure 4.1: Uncertainty-aware visual analytics design. Based on Keim et al. (2010)

4.3.2 Visual analytics for reasoning under uncertainty

Even though uncertainty visualization techniques have been broadly discussed in the data visualization community, it is still a disputed topic. Does uncertainty visualization help infer from data? Or does it impede reasoning more than help it?

Data uncertainty is often represented using such measures as probability, entropy, confidence intervals, a margin of error, and variance, which can be depicted using various visualization techniques (see, e.g., pie chart, bar chart, line graph, violin plot, scatter plot, and cone of uncertainty). In practice, many data visualization scenarios can be handled without representing data uncertainty. According to Morss et al. (2008), when it comes to temperature forecast, most people expect the temperature to fall within a range around a predicted value; thus, they can infer uncertainty into the deterministic value. In contrast to the temperature forecast that people may experience daily, the visualization of uncertainty for hurricane forecasts is crucial as it can help make emergency response decisions (Cox et al. 2013). Although various methods to explicitly depict data uncertainty have been established in different domains, studies in cognitive science suggest that non-experts have an intuitive understanding of uncertainty and the choice of uncertainty visualization technique is task dependent (Hegarty et al. 2016). However, the persistent challenge is related to uncertainty in reasoning on data.

As MacEachren (2015) has underlined, explicit attention should be focused on the role of visual interfaces in reasoning under uncertainty, where an analyst needs support in such scenarios when not only data but the problem itself, options, potential outcomes, and implications of findings are uncertain. Thus it is crucial to move the visualization component beyond explicit uncertainty signification towards a visual and analytical process that may enable an analyst to reason about hard problems in the context of uncertainty.

Visual interfaces play a significant role in information communication as they accommodate recent advances in the graphics display along with cognitive studies. In contrast to data visualization, visual analytics interfaces are enabled to not only visually represent the data and its uncertainty, but it also provides a data-driven and user-driven environment. This environment involves the user in the process of visualization and analysis, which makes it easier to understand the data complexity, reason about it and eventually reach an evidence-based conclusion.

Bayesian inference aims to quantify the evidential strength of a hypothesis and has been proven to be a powerful tool for modeling human cognition (Holyoak & Morrison 2012). Bayesian inference workflow includes several stages such as data learning, inference, and validation, where visual interfaces are practical in

every step. By integrating Bayesian inference in the visual analytics process, the uncertainty in the reasoning is operated through probability elicitation from the expert input or data attributes.

4.3.3 Bayesian Network-enabled visual analytics workflow

Step 1: Data

The data processing circle includes such stages as: data collection, preparation and input, storage, and eventually visualization. The data collection process needs to satisfy the quality requirements and provide the basis for the spatial analysis. Although it is common to use raw information for the visual analytics, some steps of preprocessing are necessary, for instance, reclassification, normalization and discretization. Despite the fact that Bayesian Networks work with both data types, discrete and continuous, for purpose of complexity reduction, the data used within the proposed visual analytics was aggregated within a regular grid and, thus, discretized. The input data was integrated as delimited text file; therefore, even large information volume can be analyzed. The data storage is supported by Database Management System, which provides reliable working environment. The data visualization is supported by Java Script visualization libraries.

Step 2: Build Bayesian inference

In the approach proposed here the DAG structure of Bayesian Network is defined for each particular application (as it is discussed in the Chapter 5). Thus, in order to find out the distribution $P(X|Y)$ with the prior probabilities $P(X)$, which are given internally, one should define the posterior distribution $P(Y|X)$ within the visual panels in the application. The inference is built upon Bayes' theorem (see Eq. 3.2). In order to define meaningful results one should provide the probability values that reflect one's knowledge about the data.

Step 3: Construct a model visualization

Here we propose a solution, which is a combination of three views: network structure view, CPT view and map view. The Bayesian Network structure, DAG, is visualized within the application to facilitate the understanding of the conditional dependencies among variables. The CPT view enables an analyst to introduce subjective beliefs as conditional probabilities with the help of graphical user interface. A significant challenge in using Bayesian inference within visual analytics is the priors' elicitation, or in other words, the way how subjective knowledge is determined through prior probabilities. The subjective beliefs are provided for the selected variables that can influence the outcome of the reasoning process. They represent the user knowledge considering the

degree of uncertainty that is inevitably involved. The users can explore the map and artifacts of the data. Based on the visual representation the user knowledge might be updated with the new information, or questioned about the validity of the introduced conditional probabilities. The inference output is visualized using regular grid overlaid with the base map and thematic data.

Step 4: User feedback and parameter refinement

Bayesian Networks can be specified using expert knowledge or learned from the data, or with combinations of both (Kenett 2016). The parameters of the local distributions are elicited from experts via an interactive facility provided by the visual analytics interface. Naturally, data uncertainty derived from the data fusion method depends on the quality of the input datasets and expert judgments. By refining the model parameters the reasoning outcome might change.

Step 5: Update the model visualization based on the feedback

Based on the refined parameters, the model can be updated with the expert feedback that can be included in the posterior analysis. Bayesian updating facilitates the updating of probability distribution of unobserved variables.

4.4 Graphical set-up

As several studies have indicated, Bayesian Networks provide an effective approach to access data uncertainty, visualization of Bayesian Networks received considerable attention. Zapata-Rivera et al. (1999) reported on the utility of temporal order, color, size, proximity (closeness), and animation techniques for mapping cause-effect relationships in a Bayesian Network model. Chiang et al. (2005) proposed to integrate heat maps for visualizing the conditional probability tables. The results obtained by Cossalter et al. (2011) have demonstrated the utility of a "thought bubble line" to connect nodes in a graph representation and their internal information at the side bar view. The visualization approach suggested by Cossalter et al. (2011) aims to improve the ability of experts to analyze large Bayesian Network models. Champion & Elkan (2017) introduced two visualization techniques: inference diffs, for comparisons of effects of evidence using concentric pie and ring charts, and relevance filtering to guide the user to variables of interest in the model.

As it has been defined in section 3.3 the inference in Bayesian Networks allows calculation of posterior probability distribution of unobserved variables in the network. Thus, this posterior probability distribution is used to draw conclusions according to the model structure. But the inference process is complex and it is hard to understand for less experienced users. Apart from the

visualization of the inference parameters, the outcome of the reasoning process can be also visualized, and in case of spatial data the outcome, it can be visualized in a geospatial context.

4.4.1 Bayesian Network mapping

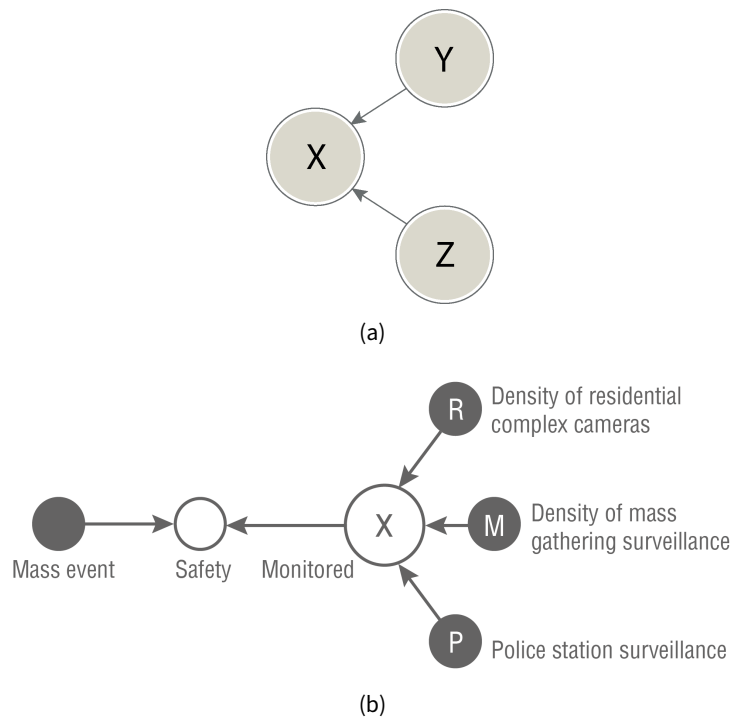
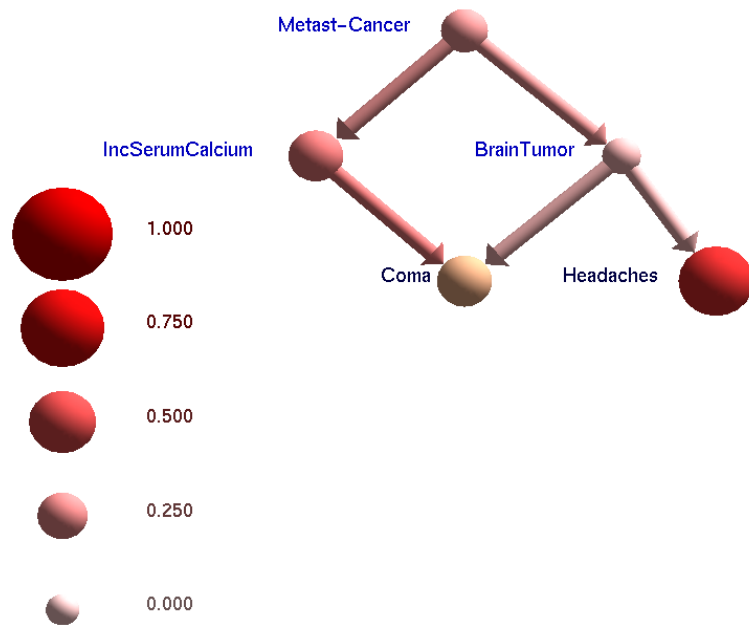


Figure 4.2: Bayesian Network visual representation.

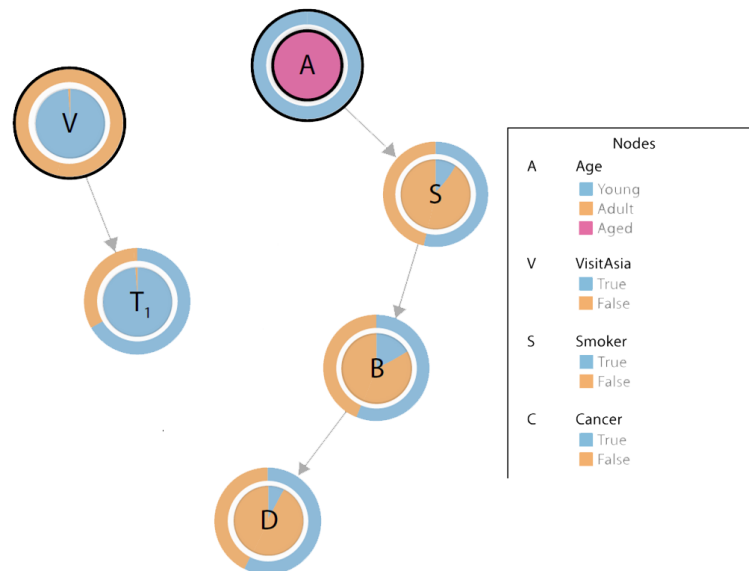
Traditionally, Bayesian Networks are depicted using a graph representation where variables (nodes) are drawn as circles or ovals and the edges, which connect the nodes, represent probabilistic dependencies (see Fig. 4.2). The visualization of Bayesian Networks may externalize the probabilistic dependencies, thus facilitate the understanding of cause-effect relationships among the variables. By highlighting a node, the system can make a user aware of a hierarchical structure that is important for the inference procedure. The good practice for the network visualization is to show all the nodes without overlays, minimize the edge crossings, and keep it compact, but readable. Moreover, the color and size can also be used to illustrate the hidden properties of a node, for instance, hierarchical position (parent/child).

Various visualization approaches have been proposed to represent the structure of Bayesian Networks (Cossalter et al. 2011, Koiter 2006, Champion & Elkan 2017,

Chiang et al. 2005). These approaches vary from simple representation as nodes with connecting edges (see Fig. 4.2) to more complex ones where each node has color and size to represent the information it carries as it can be seen on Fig. 4.3.



(a) Bayesian Network visualization in VisNet based on Zapata-Rivera et al. (1999)



(b) Bayesian Network visualization based on Champion & Elkan (2017)

Figure 4.3: Visualization approaches for Bayesian Networks.

In this research we adopt the simple representation in order to reduce the visual complexity at this stage of the development. Moreover, we set the main visual focus on the reasoning process which allows user input through CPT as it is introduced in the section 4.4.2.

4.4.2 Conditional probability tables

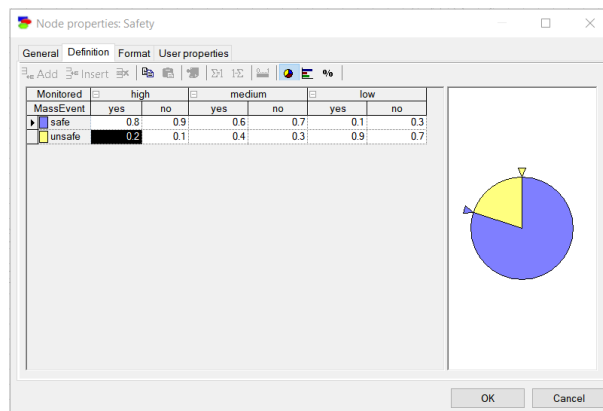
The quantitative component of the Bayesian Network is determined by Conditional Probability Tables (CPTs) and, in the system proposed, it can be elicited from the analyst via visual interface. The Bayesian probability theory entails a consistent mechanism to replace the uncertainty state with a value that represents degrees of truth, belief or plausibility (Ortega 2010). The interaction process within the application takes place through the CPT panel, where the conditional probabilities can be adjusted in order to see how the distributions change.

In general view, the belief that can be assigned within the CPTs and described as a degree to which one believes in a statement X after observing a statement Y . This basically defines the degree to which one believes in the statement X before seeing Y with prior probability $P(X)$, and how it effects the data if Y is observed with the probability $P(X|Y)$. This inference is based on Bayes' theory (Eq. 3.2). The benefit of providing the conditional probabilities about the data within visual analytics application is seen in bridging the gap between the users' knowledge about the data and the visualization outcome. The user input (subjective belief) can be specified using numerical value in the interval $[0, 1]$. The aggregated sum of conditional probabilities equals to 1.

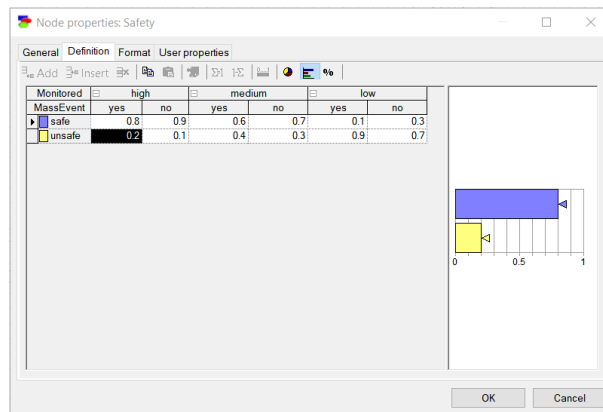
Commonly, in the software packages that deal with Bayesian Networks, CPTs are visualized as tables where probability values can be assigned through tables, bar charts, pie chart, or heat maps. When it comes to visualization approaches for CPTs, two main challenges might occur. First, the CPTs can have large size and they grow exponentially with the increasing number of considered variables. And second, CPTs should be simultaneously presented with a sufficient scope for users to filter and focus on one event combination. The filtering is seen as a solution to reduce the visual clutter.

As it has been implemented within software package GeNIe/SMILE, the CPT can be visualized as a common table, but a selected column could be enriched with a pie or bar chart visualization, where the probability distribution can be interactively manipulated (see Fig. 4.4).

Chiang et al. (2005) proposed to visualize CPTs with heat maps (see Fig. 4.5(a)), where the probability value of each cell is represented by a color tone. This type of visualization provides the users a quick comparison among all entries within



(a) CPT with pie chart visualization for a selected column



(b) CPT with bar chart visualization for a selected column

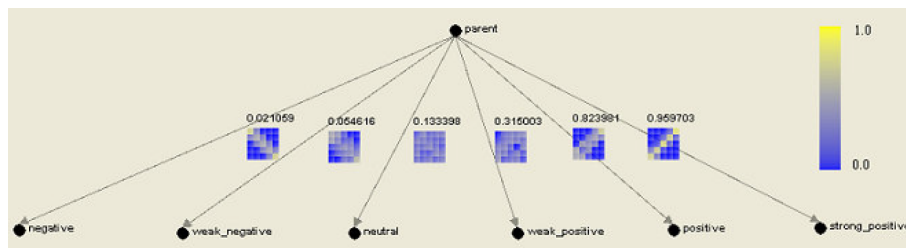
Figure 4.4: Visualization example of conditional probability tables from GeNIe/SMILE software package.

CPTs. Furthermore, it can separate the CPTs and provide a compact representation for each event combination. However, with the increasing number of CPTs visual cluttering might occur (see Fig. 4.5(b)).

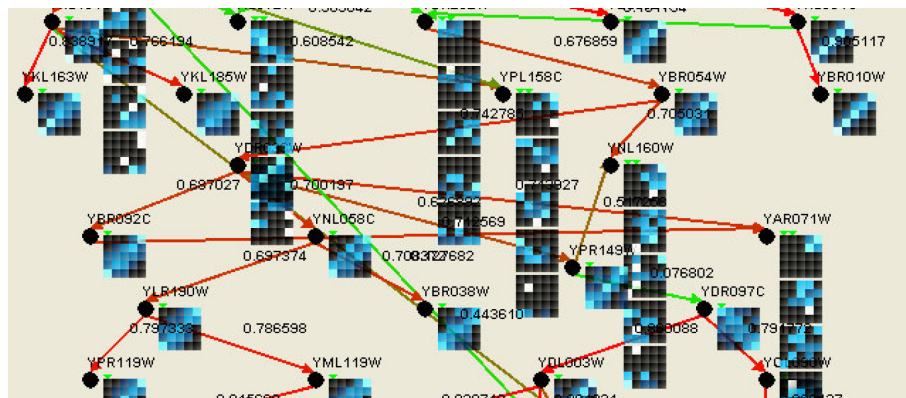
Heat maps

Following the advances for the CPT visualization, we propose to combine heat maps with a probability value given in each cell and filtering functionality which helps to overcome the problem of visual cluttering.

As mentioned in the section 3.3.1, the Bayesian Network modeling enables users to access the JPD. Thus, we propose to use heat maps for the visual representation of JPD tables too. The aim of JPD visualization is to access the



(a) Visualization of CPTs using heat maps according to Chiang et al. (2005)



(b) Visual cluttering of CPT heat maps. Based on Chiang et al. (2005)

Figure 4.5: Visualization of CPTs using heat maps according to Chiang et al. (2005).

probability distribution of every possible event as defined by the combination of the values of all variables (nodes). Thus, based on the visual representation, the analyst may find out the information about the co-occurrence of classes across the analyzed data. As it can be seen from Fig. 4.7, JPD is visualized as a heat map showing the difference across values of variable combinations. In the example given (Fig. 4.7), two heat maps represent data in two-dimensional views, where each cell visualizes the product of the probabilities of the land cover classes for GlobeLand30 and Co-Ordination of Information on the Environment (CORINE), and for GlobeLand30 and Open Street Map (OSM). These distributions provide a visual summary and may be used as a probabilistic statement of interest. For example, the land cover derived from the OSM includes some values of No Data. Thus, an analyst might be interested in finding a corresponding class in GlobeLand30 for the missing values. It can be concluded from Fig. 5.10 that most missing values in the OSM are classified in GlobeLand30 as cultivated and bareland. Considering this information as additional evidence, the strength of belief in these two classes for the OSM data might be changed.

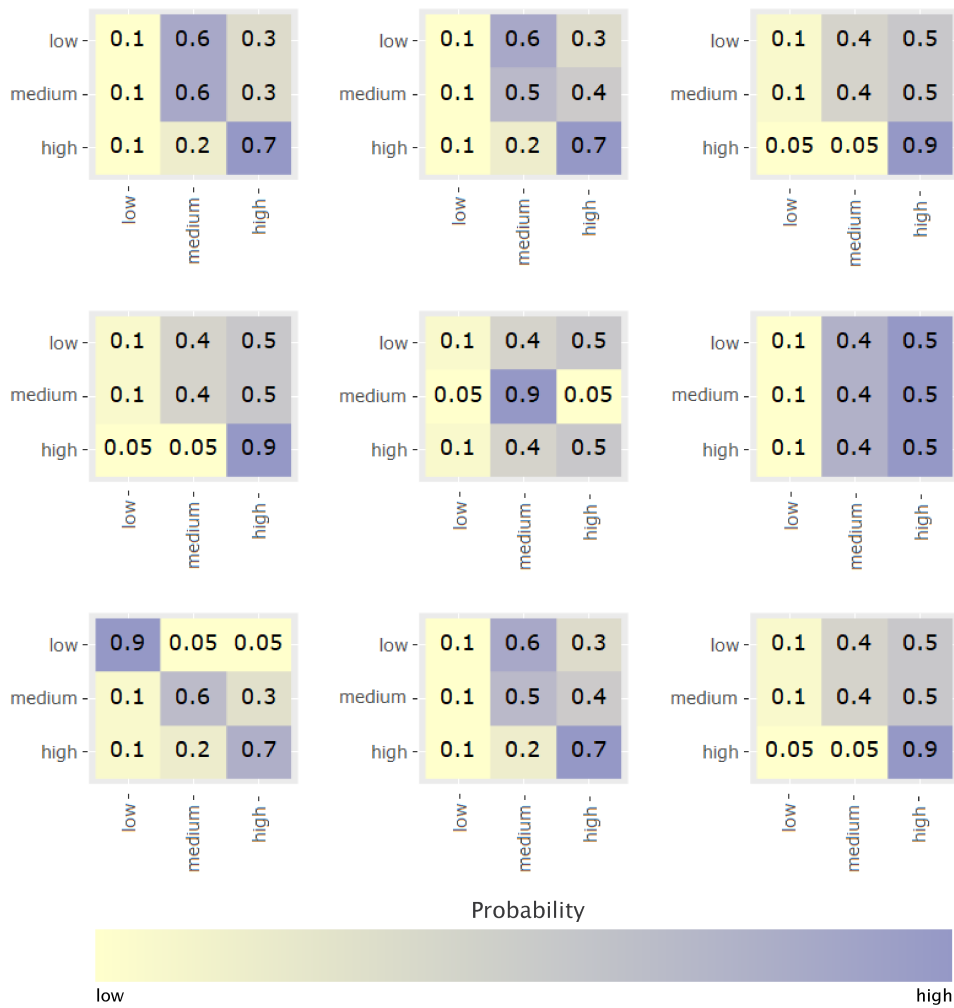


Figure 4.6: Visualization of conditional probability tables using heat maps.

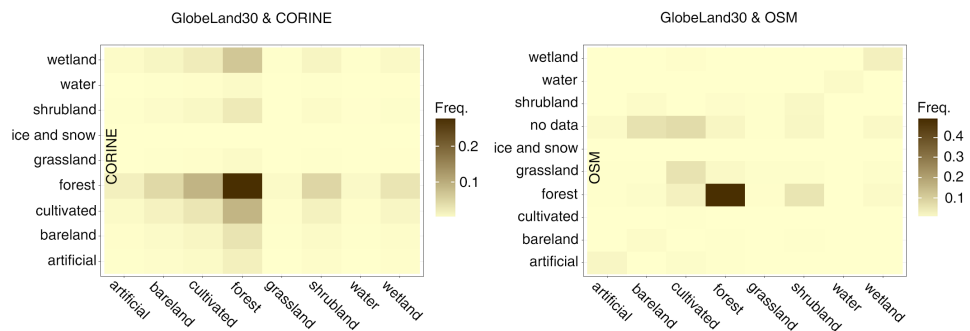


Figure 4.7: Visualization of joint probability distribution: Each cell shows the product of the probabilities of the land cover classes. The higher color intensity highlights the higher probability of class combinations.

Distribution

Conditional probabilities may be shown through the spread of possible values in the interval $[0,1]$ using an interactive bar chart, or a slider. This approach is supported by a filtering functionality; therefore, it is suitable when the number of value combinations is large. By implementing an interaction facility for the definition of probability values, we provide an interface for the user to define subjective beliefs. The probability values are given in the interval $[0,1]$ with a total sum of 1 for each variable combination. The pros of this approach are seen in the simple representation and high interactivity.

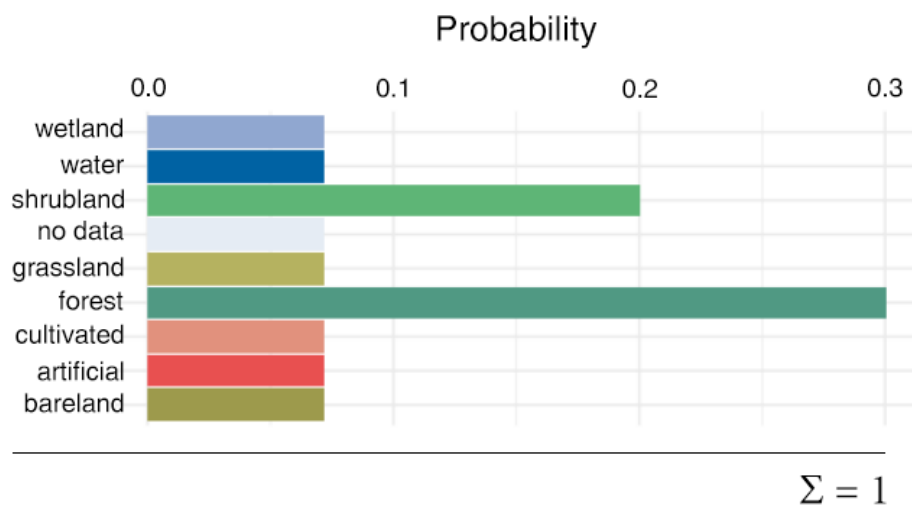


Figure 4.8: Visualization of conditional probability tables using bar charts.

4.4.3 Outcome of the reasoning process

In general, the spatial data used for the Bayesian Network modeling can be either continuous or discrete. The discrete data can be seen as an approximation to describe the real world data. Therefore, for the reasoning purpose, the data is discretized and aggregated within a grid structure. The grid structures have different forms and can be summarized by the following types: polygon, diamond, hexagon, line, regular point, and random point grids (see Fig. 4.9).

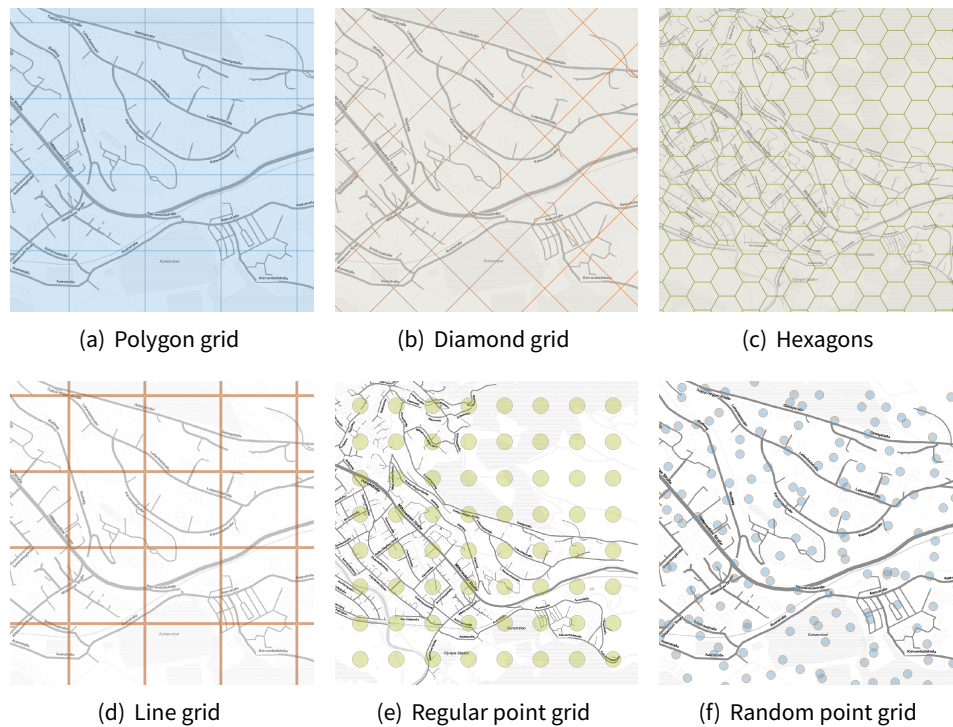


Figure 4.9: Grid structure to visualize the outcome of the Bayesian Network analysis

It is crucial to minimize the chance that uncertainty is perceived in a negative way. Accordingly, the uncertainty representation should be considered as a recommendation for better understanding of the classification process. Thus, finding an appropriate visualization technique for the uncertainty, as it is defined in this research, is a challenging task. As it has been mentioned, diverse approaches have been suggested to visually encode the spatiotemporal uncertainty information and Kinkeldey, MacEachren & Schiewe (2014) summarized them into three main categories: intrinsic/extrinsic visualization, coincident/adjacent display, and static/dynamic views. As there is no one-suits-all method, we take into account the data type we work with and the analytical approach we apply as main fitness criteria. Thus, extrinsic visualization techniques are chosen because the uncertainty visualization can be integrated as an overlay with different datasets.

The extrinsic visualization techniques incorporate additional graphical objects to signify uncertainty (Kinkeldey 2014) and mainly make use of grid structures overlaid with the data (Kinkeldey, Mason, Klippel & Schiewe 2014), glyphs (Wittenbrink et al. 1996), or isolines. In this research we also chose to apply extrinsic technique utilizing modified version of a regular point grid, where the

circle radii represent an assigned value. Thus, the grid of proportional circles characterize the geographic and thematic variability of a given location. Moreover, the proportional circles are localized and can handle both numerical and categorical data. Although we consider two different scenarios of using Bayesian Network within visual analytics, a similar visualization approach is used. More details about the practical solutions can be found in Chapter 5.

The reasoning outcome of the Bayesian Network analysis can be represented in different forms, as for example it can be information entropy measured as Shannon index or a likelihood of grid cell for belonging to a particular class. When considering the representation of information entropy on a digital map, the size of the circles can directly identify the amount of entropy measured as the Shannon Index at a given location and may be updated when new evidence is set within the conditional probability panel. The larger the entropy, the larger the circle symbol will be. This visualization approach is widespread, and easily comprehensible. Each circle can be explored individually and analyzed based on the classes that occurred at a location (Fig. 4.10). The circles can be also analyzed in groups and provide the experts with information about a bigger pattern within the given datasets.

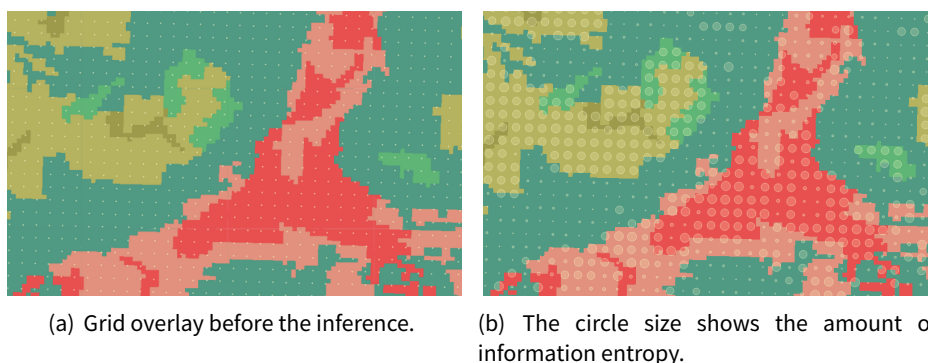
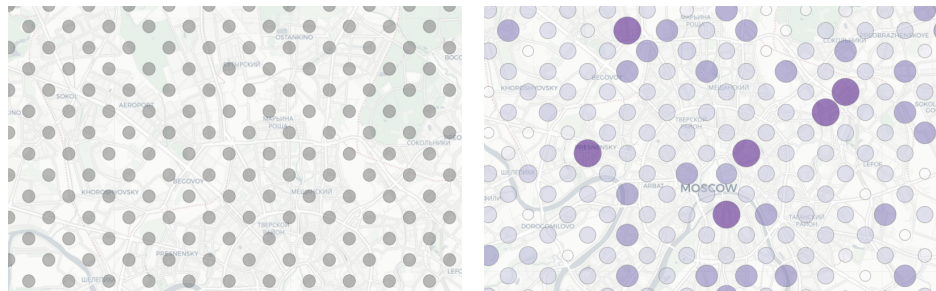


Figure 4.10: Visualizing information entropy through point-based grid. The proportional circles identify the amount of entropy measured as the Shannon Index. The larger the entropy, the larger the circle symbol will be.

Moreover, the proportional circles can characterize the likelihood of different city zones being assigned to a particular characteristic. They represent the data set discretely, showing patterns of interest, rather than representing an interpolated value over the city district. The size and color of circles directly signify the probability value at a given location and may be updated when new evidence is set.



(a) Grid overlay before the inference.

(b) The circle size shows the likelihood of different city zones being assigned to a studied characteristic

Figure 4.11: Visualizing likelihood of different city zones being assigned to a particular characteristic through point-based grid. The size and color of the proportional circles characterize the likelihood of different city zones being assigned to a particular characteristics given the probability value in the interval $[0, 1]$.

4.5 Summary

The development of Bayesian Network-enabled visual analytics with a geospatial component has a potential to bridge the gap between current spatial data analysis and visualization domain where uncertainty in reasoning process may be of interest. The Bayesian Network approach, adopted and implemented within a visual analytics application, facilitates Bayesian analysis of multiple datasets and integrates expert knowledge to reveal the hidden patterns considering the uncertain user knowledge.

In addition, this research contributes to the growing set of extrinsic visualization approaches by developing and demonstrating a method of proportional circles that incorporate the results of conditional reasoning within a model-based design using Bayesian Network to infer from the data with inherent uncertainty.

Chapter 5

Prototypical implementations of Bayesian Network-enabled visual analytics

“ *That all our knowledge begins with experience there can be no doubt.* ”

Immanuel Kant, *The Critique of Pure Reason*, 1781

To examine the potential of visual analytics techniques with embedded Bayesian Network described in the Chapters 3 and 4, we developed two scenarios that represent a classification task relying on (a) land cover and (b) locations of surveillance cameras. The implementation is based on Bayesian Network inference algorithms and realized with such open-source software and libraries as R statistics, R shiny, Leaflet.js, Geoserver, PostGIS, and PostgreSQL.

This chapter is focused on the following objective:

To build a prototype of a visual analytics interface that can integrate data, visualization, and computational capacity of Bayesian Networks to facilitate human-computer interactions for data analysis given subjective beliefs used in a selected application domain.

Each scenario is divided into four parts: data, experimental set-up, visual encoding and interaction design, and results. The visual encoding and interaction design describe how users see the visual representation and the

techniques that allow them to interact with what they see. As the visualizations and interactions are connected they are described in the same section.

5.1 Scenario 1. Global land cover classification

In this scenario, we demonstrate the development of a prototype for the exploration of spatially heterogeneous data and uncertainty in classification of land cover. We develop a visual analytics interface that leverages integrated Bayesian Network model capabilities to support land cover analysis. Our solution accommodates multiple land cover datasets and serves as a modeling tool where patterns with the highest geographic and categorical variation might be discovered through expert input. The prototype empowers human operators to assess multiple land cover classifications, uncover uncertain information and assign conditional probabilities based on expert beliefs, thus perform inference using Bayesian Network.

Bayesian Network is a powerful tool for reasoning based on heterogeneous data, as it allows modeling of uncertainty in the reasoning rules (through conditional probability tables) and the uncertainty in data sources (through the priors) (Stassopoulou et al. 1998). Several researchers (Aitkenhead & Aalders 2009, Krüger & Lakes 2014, Celio et al. 2014) have proved that Bayesian Networks provide a versatile method for assessing uncertainty in land use - land cover modeling, and it can combine diverse knowledge sources. Accordingly, using a Bayesian Network enables analysts to define relations between land cover variables in terms of the conditional distributions for each land cover class and allows reasoning under the uncertainties that are associated with these conditional distributions.

To assess the land cover classification uncertainty in different datasets of the same region, we introduce a scenario based on several land cover products. In this scenario the following questions are asked:

- **What is the most likely land cover class at each given location if three datasets are examined?**
- **What are the uncertainties associated with the modelled outcome of the experiment?**

5.1.1 Data

To demonstrate an application of the Bayesian Network method for analysis of land cover classifications, three major datasets were chosen, namely

GlobeLand30, CORINE (GLC2006) and Volunteered Geographic data based on OSM. These datasets are independently acquired and show classification of land cover, with overlap for European Union (EU countries). This data is commonly used for various environmental studies.

The prototype is introduced here and its potential is demonstrated through a case study analysis of land cover data for Upper Bavaria, Germany. The study area of Upper Bavaria reveals a sufficient diversity in the distribution of land cover classes, thus provides a solid background for exploring the probability of each land cover class. By analyzing different products, it is possible to reveal their diversity based on conditional dependencies among the classes and provide an estimate of geographic and categorical variation. Fig. 5.1 represents the intended workflow with the view to analyze how different land cover maps correspond to each other in order to find the areas that have the highest degree of uncertainty in the classification. Further, the datasets used for this analysis are introduced in the details.

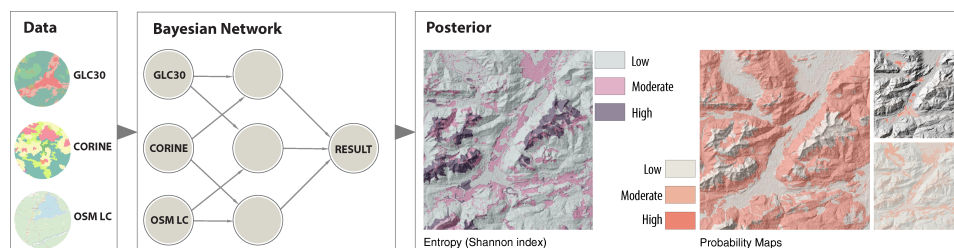



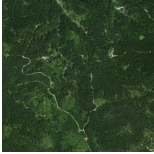





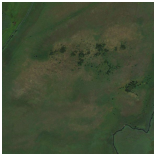

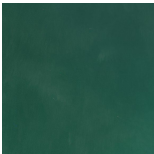

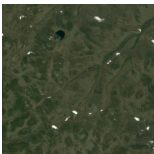






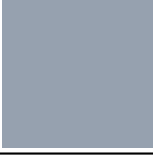
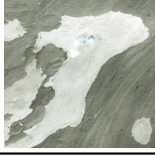
Figure 5.1: A workflow for assessing land cover data uncertainty using Bayesian Networks. Data: GlobeLand30 (GLC30), CORINE (GLC2006), and land cover derived from Open Street Map (OSM).

GlobeLand30

In 2010 China launched a project with the aim to identify global land cover classes in resolution of 30 meters. The GlobeLand30 data set is freely available and comprise 10 major classes of land cover, including cultivated areas, forests, grassland, shrub land, wetland, water bodies, tundra, artificial surfaces, bareland and permanent snow and ice (see Table 5.1). The classification is available for two base-line years, 2000 and 2010. The GlobeLand30 was produced based on more than 20,000 Landsat and Chinese HJ-1 satellite images (see www.globallandcover.com). Previous studies indicate the overall classification accuracy can range from 46% (Sun et al. 2016) and up to 80% (Brovelli et al. 2015, Chen et al. 2015, Arsanjani et al. 2016). Therefore, the data is of heterogeneous quality. The color coding is altered from the original colors offered by GlobeLand30 to satisfy the design needs of the visual analytics prototype introduced in Chapter 4.

Table 5.1: Land cover classification scheme based on GlobeLand30.

Color code	Land cover class	Description
		Cultivated lands. Lands used for agricultural purposes, gardens, dry and irrigated farmlands.
		Forest. Lands covered with woods more than 30%, including deciduous and coniferous forest. Sparse woodland 10 - 30%.
		Grassland. Lands covered with natural grass with cover over 10%
		Shrubland. Lands covered with shrubs over 30%, including deciduous and evergreen shrubs. Desert step with cover over 10%
		Wetland. Wetland plants, peat bogs, salt and mangrove marsh, floodplains, lake marsh.
		Water. Rivers, lakes, natural and fish reservoir.
		Tundra. Vegetated lands in polar regions. This class is not present in the study area.

		Artificial surface. Lands modified by human activities: industrial and mining areas, settlements, urban green zones, artificial water bodies.
<hr/>		
		Bareland. Lands with less than 10% vegetation, sandy fields, bare rocks.
<hr/>		
		Permanent ice and snow.

Land cover derived from Open Street Map

Since the term Volunteered Geographic Information (VGI) was introduced by Goodchild (2007), knowledgeable amateurs have contributed large amounts of spatially referenced data to different Web portals. Data acquired from VGI has attracted much attention within remote sensing community too, as it can significantly contribute to the validation of land cover classifications. VGI-based approaches have been proposed, and on-line communities such as GEO-Wiki (geo-wiki.org) are contributing to land cover data collection (Fritz et al. 2009). Thus far, however, the coverage of the contributed data doesn't allow for exhaustive validation of large region datasets. In contrast to that, the OSM contributors are very actively collecting a broad range of thematic data with close to complete spatial coverage in certain areas (Ribeiro & Fonte 2015). The OSM (openstreetmap.org) is, without any doubt, a massive and well recognized project. The database consists of vector data, which is attributed with a variety of geospatial labels and might serve as data source for different cartographic products. Since every contributor can freely edit the database, the OSM data is heterogeneous in terms of quantity and quality. Assessing the accuracy of the OSM is, hence, an essential task to facility the scientific usage of this data source. Several studies have reported encouraging results in terms of the overall accuracy and completeness (Helbich et al. 2012, Neis et al. 2012). Thus, utilizing the OSM as a source for land cover data is a promising approach. Since the OSM data is not specifically tailored to the needs of land cover map validation, various methods for transforming the original data into a more suitable representation have been developed (Fonte et al. 2015). While only a portion of the OSM attributes is valuable for a derived land cover map, the

coverage is still high enough to be usable, especially in urban areas (Ribeiro & Fonte 2015).

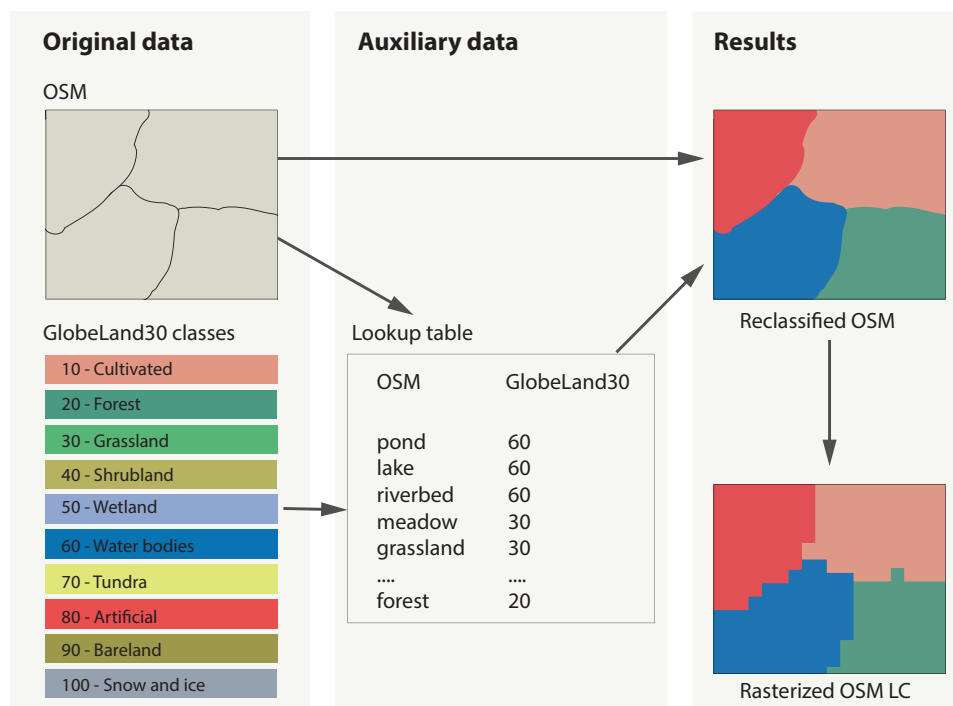


Figure 5.2: An overview of pre-processing steps for converting original OSM data into land cover. Source: Chuprikova et al. (2017).

In this research, we established a workflow (Fig. 5.1.1) to convert raw OSM data into land cover coverage. In order to preserve the entire content of the database, a complete XML-encoded extract of the OSM database is used, representing the study area, instead of pre-processed ESRI Shape Files, as suggested by Fonte et al. (2016). For an efficient processing of the large data amount a PostGIS database is established. The OSM vector database is attributed with a great variety of labels and serves as source information for various cartographic products (Chuprikova et al. 2017). For the derivation of the land cover map, a subset of the OSM tags, namely "amenity", "building", "historic", "land use", "leisure", "natural", "shop", "tourism", and "waterway" are considered. A mapping from the OSM attributes to the classes used in the GlobeLand30 classification scheme is conducted for polygon features, since point and line features don't provide immediate information about the coverage of an area. Exploiting additional information implicitly contained in point and line features, might be possible in general, using assumptions about specific feature classes, such as empirically determined road widths. In order to keep the used data as noise-free as possible, this is omitted in this experiment.

Due to the fact that GlobeLand30 classification consists of 10 classes, some degree of ambiguity is possible as well as areas with missing data. In spite of the heterogeneity of both OSM data and its user community, OSM provides a dynamic source to evaluate the land cover classification and assist experts in finding land cover patterns that are better mapped by citizens than by using remote sensing techniques. In the final step, the vector data is rasterized to a 30m grid with an appropriate minimum mapping unit, merging small features with their neighboring features. The total coverage of the obtained reference map is about 71% of the total study area.

CORINE

The land cover mapping for the countries of European Union is realized within the programme CORINE. Based on CORINE Land Cover (CLC) 2000 for Germany, the data was updated and land cover product CLC2006 was produced as a vector database. According to Keil et al. (2011) the main data sources of the land cover and land use mapping were satellite images of Landsat 7 (for 2000) and IRS-P6 LISS III as well as SPOT-4 and SPOT-5. This product is following common European wide CLC nomenclature and consists of 44 classes, where 37 classes are relevant for Germany. Therefore, CLC2006 is characterized as result of the GIS derivation considering the land cover changes. The data is released in the projections of Gauss-Kruger Zone 3, Gauss-Kruger Zone 4 or UTM Zone 32.

Several authors (Gallego 2001, Brovelli et al. 2015, Arsanjani et al. 2016) have proposed to assess the land cover quality using CORINE datasets for different study areas. Arsanjani et al. (2016) studied the validation of GlobeLand30 against CORINE for the area of Germany and showed that overall accuracy is 92%. Another solution was described in Brovelli et al. (2015) who indicated that the overall accuracy values of third level CORINE Land Cover are generally higher than 80%. In both cases the validation was made using confusion matrix that represents comparisons of a land cover map against the referenced dataset. In this study CLC2006, further called CORINE, is used as a variable for constructing Bayesian Network. Therefore, the vector map CLC2006 is gridded for use in the modeling process. The resolution of the gridded map is 30m. Moreover, the complexity of land cover classification is scaled down in order to provide consistent classification for all the data sources. That is to say, 44 CORINE land cover classes are assigned to 10 classes in line with the GlobeLand30 classification (see Table 5.2).

Table 5.2: Reclassification of CORINE land cover classes relevant for the study area based on the GlobeLand30 scheme.

Land cover classes	CORINE, Pixel values	GlobeLand30, Pixel values
Cultivated	32 - 41	10
Forest	42 - 45	20
Grassland	46 - 47	30
Shrubland	48 - 49	40
Wetland	55 - 59	50
Water bodies	60 - 64	60
Tundra	-	70
Artificial surfaces	21 - 31	80
Bareland	50 - 53	90
Permanent ice and snow	54	100

5.1.2 Experimental set-up

Preprocessing

The details of the preprocessing procedure is listed below:

1. All datasets are cropped to the municipal boundary of Upper Bavaria;
2. The classes of OSM and CORINE are normalized to match the 10 classes specified by GlobeLand30;
3. The datasets are then rasterized with 30m pixel resolution and aligned to GlobeLand30 to ensure pixels covered the same land area.

Processing

The workflow for implementation of Bayesian Network for land cover classification is illustrated in Fig. 5.1. As described in the section 3.3.3 the underlying mathematical model of a Bayesian network is based on components such as DAG and CPTs. The nodes in DAG represent random

variables, and arrows among them describe dependencies among these variables. The terms node and variable are used interchangeably. In this study we treat the land cover classifications GlobeLand30 ($A = a_i$), CORINE ($B = b_i$), and land cover derived from OSM ($C = c_i$), as nodes for constructing the Bayesian Network. Thus, the edges between the land cover nodes (parent nodes) and the resulting node (child node R , where r_1, \dots, r_n are all possible values) indicate the causality between values of the land cover variables (A, B , and C) and the values of the resulting node R . Based on the DAG structure, the quantitative part is described through conditional distribution within CPTs (see Fig. 5.3). CPTs express the probability of the state of each variable given its parents and essentially represent the strength of the belief in the causality (Luxhøj 2014) defined by the experts or learned from the data structure. In this study, we consider the application of Bayesian Networks to discrete data; therefore the CPT lists the local probabilities that the child node R will acquire from each combination of values of the parent nodes A, B , and C .

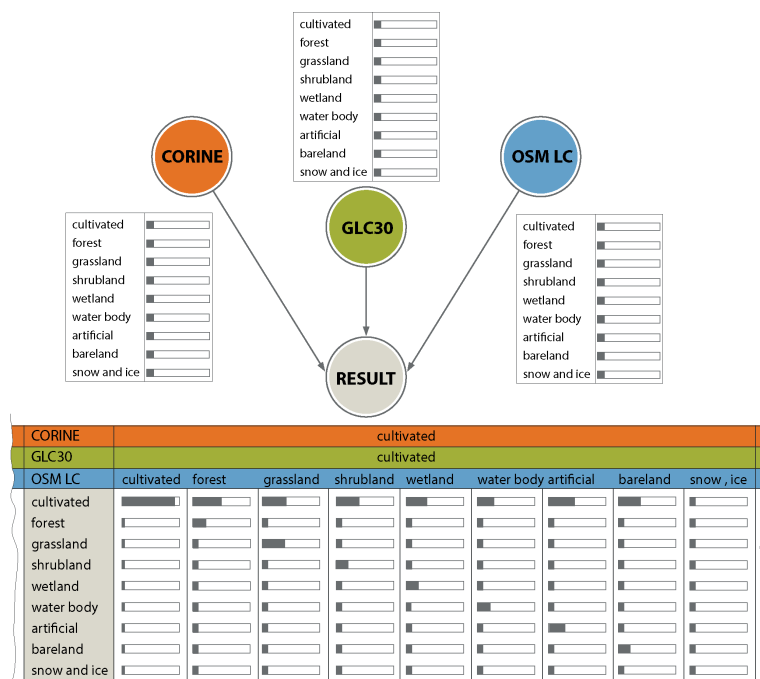


Figure 5.3: Bayesian Network structure for the land cover data analysis. Example for the land cover class cultivated.

Thus, the approach is developed based on the following setting. If we consider that one of the nodes doesn't have a parent, for instance GlobeLand30 $A = a_i$ (named GLC30 in Fig. 5.3), the prior probability for this variable is assigned based on the possible values: $P(A = a_i)$, where a_1, \dots, a_n are all possible values (instantiations) of variable A_i . The possible values for GlobeLand30 ($A =$

a_i) are the land cover classes assigned to cultivated areas, forest, grassland, shrubland, wetland, water bodies, artificial surfaces, bareland, and snow and ice coverage. A similar process is applied for other nodes of the network that do not have parents (nodes B and C). The resulting node R at the next level has parents (A , B , and C); therefore, the conditional probability is assigned applying the Markov property (Stassopoulou et al. 1998). Consequently, the node *Result* ($R = r_i$) contains the conditional probabilities of each node value for a conditioning case. A conditioning case is a possible combination of values (land cover classes) for the parent node, as can be seen in equation 5.1. The process is organized in one-way direction, so that the child doesn't transfer any feedback to the parent. On this note, the Bayesian approach is able to model the proportions of true values in selected pixels at each location across the whole study area.

$$P(R = r_i | A = a_i, B = b_i, C = c_i) \quad (5.1)$$

The CPTs are defined based on expert knowledge and fundamental laws of geography (see Table 5.3). Land cover classes which are likely to represent similar Earth's characteristics are assigned to a high probability, while classes unlikely to be overlapped - based on region, geography and expert assumptions, are assigned to a low value. The self occurrence probability of each class $P(a_i, b_i, c_i)$ is assigned the highest value, for example when $a_i, b_i, c_i = forest$.

Table 5.3: Land cover variables and their instances.

Land cover classification	
	GLC30, $A = a_i$
	CORINE, $B = b_i$
	OSM, $C = c_i$
Result, $R = r_i$	$P(R = r_i A = a_i, B = b_i, C = c_i)$

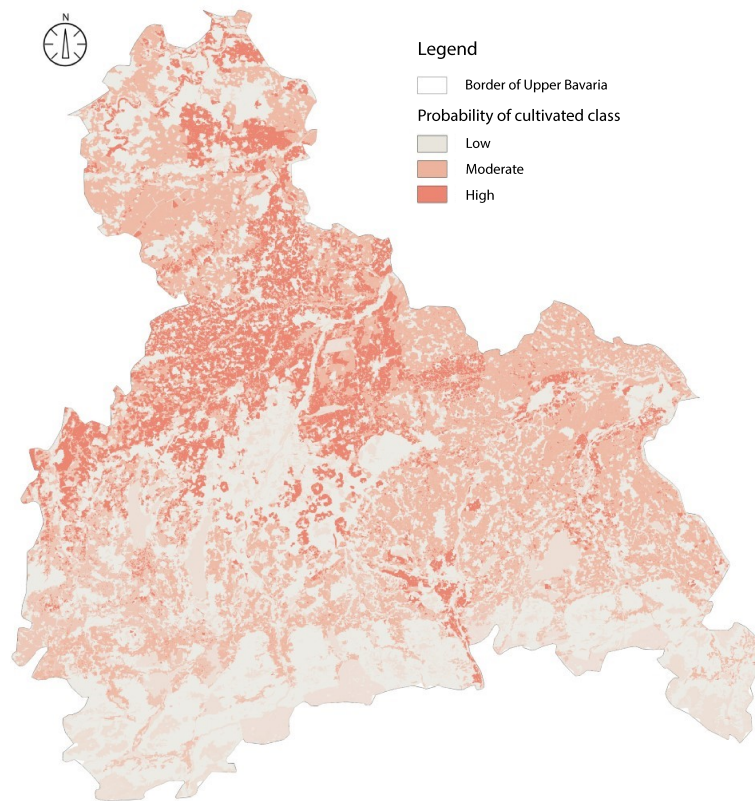
Based on the described analysis, the values of probabilities for each land cover class at each location of the area of interest might be extracted. Using this information, it is possible to predict which land cover class has a higher potential to occur. Although a fused land cover map might be of interest in some applications, here the main focus on the information about uncertainty. According to Jost (2006) the Shannon Index (also called Shannon-Wiener Index) is the most common diversity measure and it plays a central role in information theory as a measure of information, choice, and uncertainty

(Spellerberg 2008). This index represents entropy, giving the uncertainty as the outcome of a sampling process.

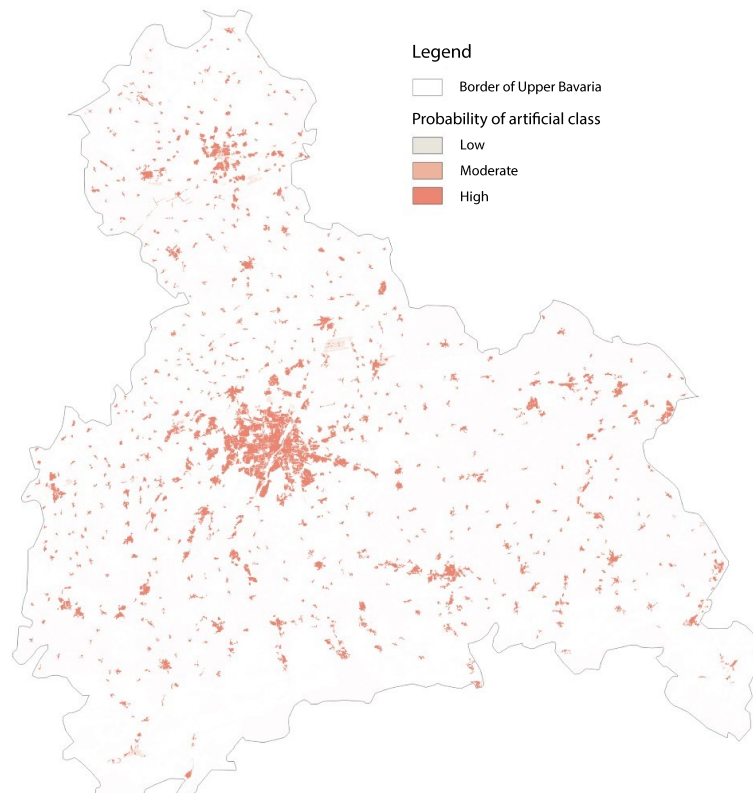
Another important advantage of the Bayesian Network is that it provides an efficient representation of JPD. We utilize this measure in the application as it can express all the probabilities of all combinations of different values. The JPD may answer any probabilistic questions of interest and provide information about the co-occurrence of land cover classes.

Outcome

The simulation is realized using R statistics and package for the spatial implementation of Bayesian Networks and mapping "bnsatial". The outputs of this simulation procedure are posterior probability maps (see Fig. 5.4), and the map of uncertainty measured as Shannon index (entropy) (see Fig. 5.5). The maps of posterior probability depict updated prior probability of land cover class occurrence considering the information from different input datasets. The map of uncertainty is elaborated to depict the diversity in the data and it illustrates Shannon entropy applied for the land cover classification. Therefore, the latter map shows the information entropy at each location. The aim of the output maps is to highlight the areas with the highest degree of the uncertainty and provide visual guidance for the further land cover validation as well as for a deeper understanding of the dynamics related to the high uncertainty. As it can be seen from Fig. 5.6, the uncertainty of a polygon is high, and when we compare to the original data sets, it is evident that the highlighted area was defined in inconsistent manner. Hence, the attention for the validation should be placed at such areas. Moreover, the output data may be utilized to guide the implementation of decision support tools.



(a) Probability map of cultivated area



(b) Probability map of artificial area

Figure 5.4: Probability maps represent expected state of a target node (i.e. the state with the highest relative probability).

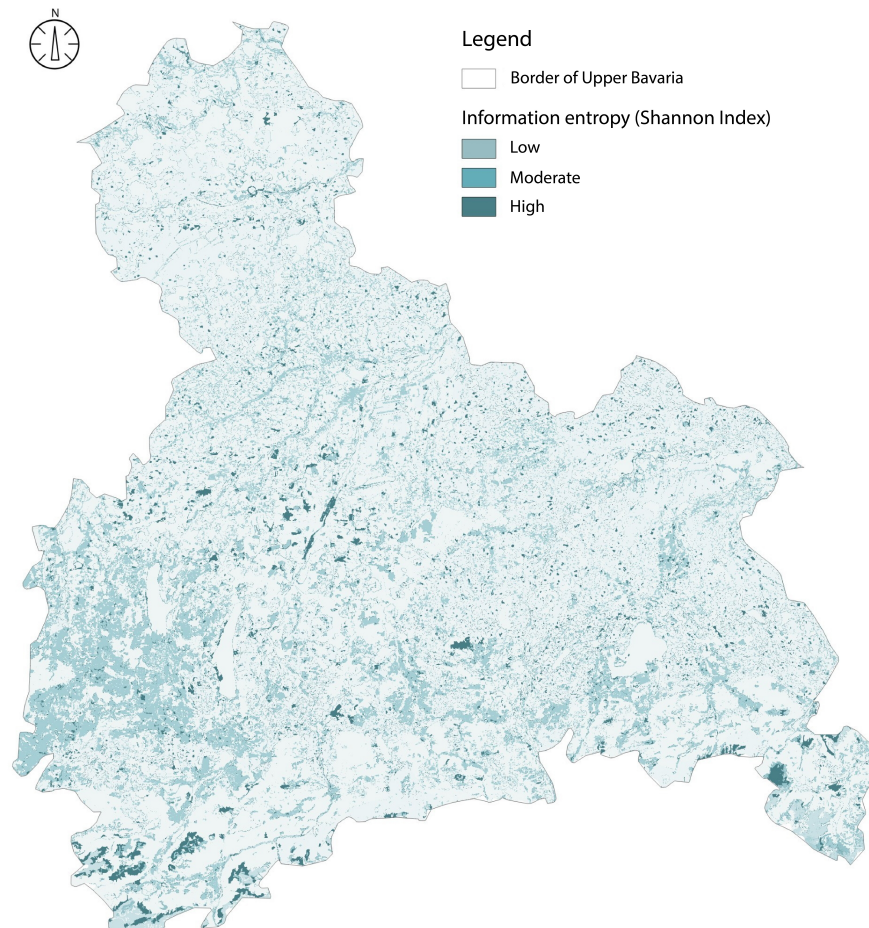


Figure 5.5: Information entropy map based on Shannon Index. The map represents uncertainty quantified as diversity in the information outcome. A higher degree of uncertainty means greater diversity in the land cover classes among the three datasets.

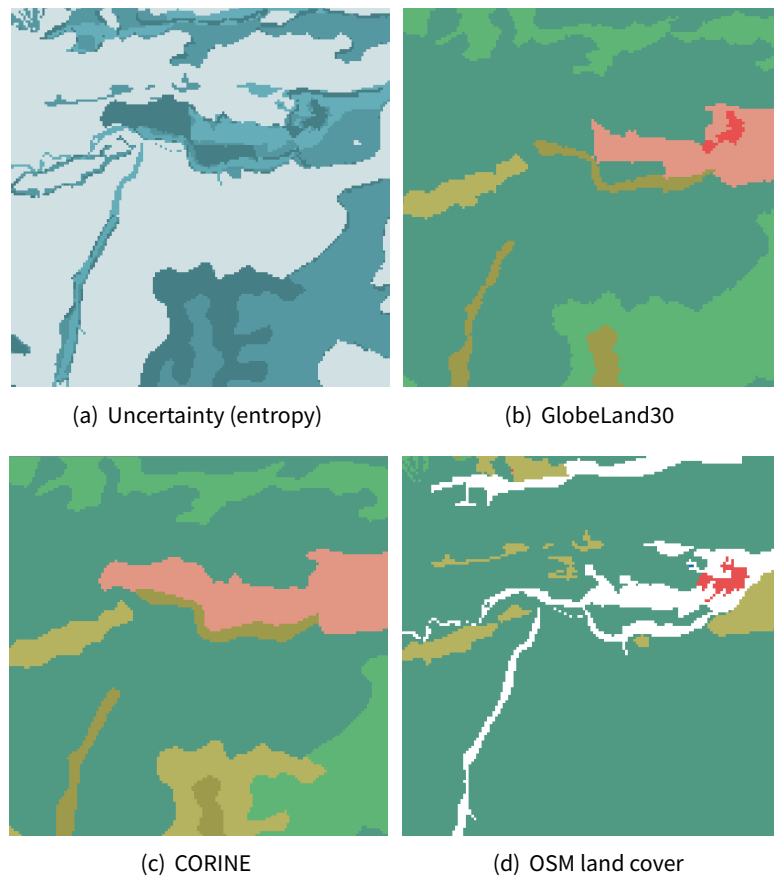


Figure 5.6: Scenario 1: An example of uncertainty in classification among land cover products.

5.1.3 Visual encoding and interaction design

Although Bayesian inference is an effective method for decision-making practices, it commonly proceeds without reference to modern developments in statistical graphics made possible by greater computational capacity (Kerman et al. 2008). To demonstrate the connection between modeling and statistical graphics, we combine the computational and visualization components for uncertainty analysis of classified land cover data. The mechanism needed to model, create, and interpret a visualization of uncertain patterns in land cover classification is realized through the visual analytics application introduced here.

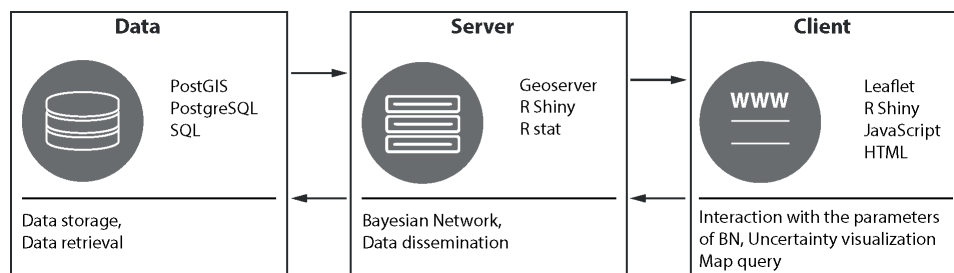


Figure 5.7: Scenario 1: Three-tier architecture of the framework for land cover uncertainty analysis

An implementation of the uncertainty-aware visual analytics application requires interoperability of methods and data in an open system setting. The application for land cover uncertainty analysis is established based on the three-tier architecture illustrated in Fig. 5.7. This includes: (a) a Database Management System (DBMS), namely PostgreSQL, (b) a Servlet Engine (Geoserver and R server) and (c) a Client side tool (implemented with R Shiny). The data is stored in the DBMS and rendered on the Web map, where the analyst can explore the map layers and interact with the output of the reasoning process. The prototype design is depicted in Fig. 5.8. The graphical User Interface (UI) adopts a visual analytics approach to provide an interactive visual interface, linked to data and computational methods, with support for reasoning about land cover classification. The interface is divided into three main sections illustrated in Fig. 5.8: Conditional Probability Panel, Map panel, and Joint Distribution plots.



Figure 5.8: Uncertainty-Aware visual analytics interface. The graphical UI includes Conditional Probability Panel, Joint Probability Distribution plots, and Map panel.

Conditional Probability Tables

The Conditional Probability Distribution panel is located in the upper left part of the application and it facilitates the interaction between the user and the Bayesian Network. As discussed earlier, the probability of any land cover class occurring is expressed as a prior or unconditional probability (see section 3.3.3). Due to the fact that the Bayesian Network structure is predefined and the state of each node is described within the server side, the user interaction takes place through definition of the conditional probabilities among the classes within the resulting node R . The Bayesian Network structure can be seen in Fig. 5.3. This process provides new evidence for the modeling. For instance, consider a situation when two nodes (A and B) define that the land cover class at a given location is Cultivated and another node (C) states that the land cover class should be assigned to Grassland. In this case, the conditional probability of the class Cultivated in the resulting node (R) should be updated to a lower value if the subjective belief assumes that value of the node C is reliable. By doing it, a user can compare the land cover classifications in a non-deterministic manner considering the random nature of the phenomena.

The conditional probabilities are defined from the subjective point of view when experts express their strengths of belief about the quality of land cover classifications as probabilities. The subjective belief, for example, may assume that urban structures are well mapped in the OSM data, so all the combinations where the OSM data show urban structures should be considered as more

probable. To assign meaningful beliefs, it is crucial to be familiar with the data used for the analysis.

Conditional Probability Distribution

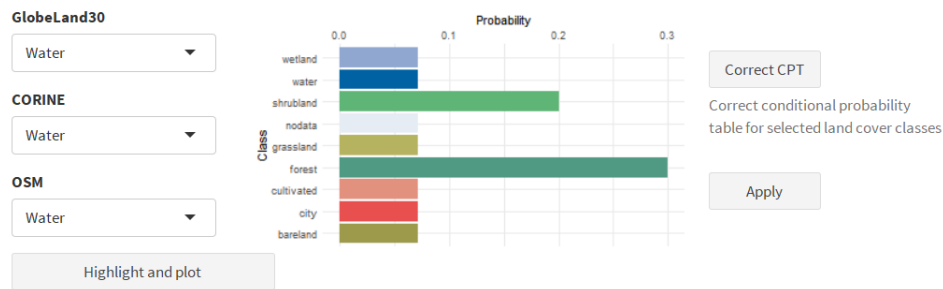


Figure 5.9: Conditional probability distribution panel facilitates the interaction between the user and the Bayesian Network through a definition of strengths of belief about the quality of land cover classifications as probabilities.

Joint Distribution plots

The Bayesian Network modeling enables experts to access the JPD. The JPD is the probability distribution of every possible event, for example a land cover class, as defined by the combination of the values of all the variables (nodes). Thus, based on the visual representation of the JPD, an analyst may find out the information about the co-occurrence of classes across the land cover products.

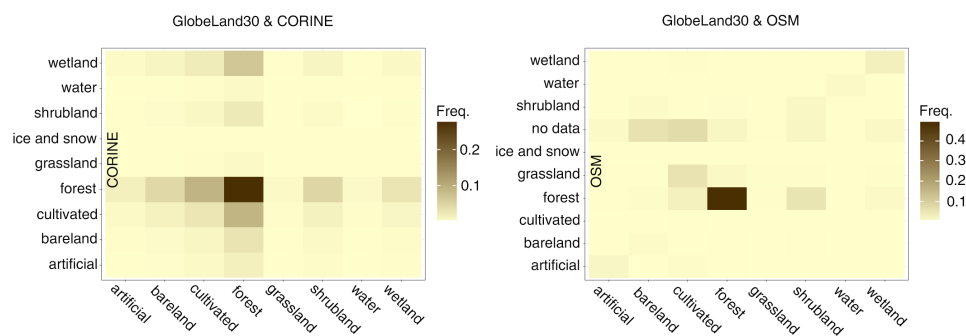


Figure 5.10: Joint probability distribution panel: The visual representation of the JPD allows us to reason about the relationship between multiple land cover classes. The higher color intensity highlights the higher probability of class combinations.

The table structure of JPD is visualized as a heat map showing the difference across land cover classes' combinations. Thus, as can be seen in Fig. 5.10, two

heat maps represent data in two-dimensional views, where each cell visualizes the product of the probabilities of the land cover classes for GlobeLand30 and CORINE, and for GlobeLand30 and OSM. These distributions provide a visual summary and may answer a probabilistic statement of interest. For example, the land cover derived from the OSM includes some values of No Data. Thus, an analyst may want to find a corresponding class in GlobeLand30 for the missing values. As can be seen in Fig. 5.10, most of missing values in the OSM are classified in GlobeLand30 as cultivated and bareland. Considering this information as additional evidence, the strength of belief in these two classes for the OSM data might be changed.

Map panel

Fig. 5.11 represents the outcome of the modeling in its spatial context, where results of the inference are registered as a new temporal attribute, which is further used for the visualization of estimated uncertainty on the digital map. In this scenario, we chose to visualize information entropy among three datasets, measured as the Shannon Index, by applying proportional circles that characterize the geographic and thematic variability of a given location. The larger the entropy, the larger the circle symbol will be. This visualization approach is well-known, and easily comprehensible. Each circle can be individually explored and analyzed based on the classes that occurred at the corresponding location. Circles can be also analyzed in groups and provide experts with information about apparent patterns within the land cover classification. By comparing the land cover data of different sources using Bayesian Network modeling with an expert input, a user can discover the principal differences across the land cover classifications and obtain a visual representation of the uncertainty among different datasets of the same theme.



Figure 5.11: Map view illustrates the categorical maps of land cover and with an overlay of proportional circles that identify the amount of entropy measured as the Shannon Index. The larger the entropy, the larger the circle symbol will be.

5.1.4 Results

The prototype demonstrates the feasibility of accommodating both human and Bayesian reasoning for the analysis of remote sensing data, especially for the reasoning based on diverse sources with varying degrees of reliability. The output of the analysis introduced is visualized on the digital map as an overlay and illustrates uncertainty measured as Shannon entropy. Based on the given visualization, it is possible to localize the places with high uncertainty in land cover classes and provide a visual guide for further land cover validation as well as for a deeper understanding of the land cover classification from Earth Observation products. Further, we discuss three cases that can help to interpret the results.

Case 1

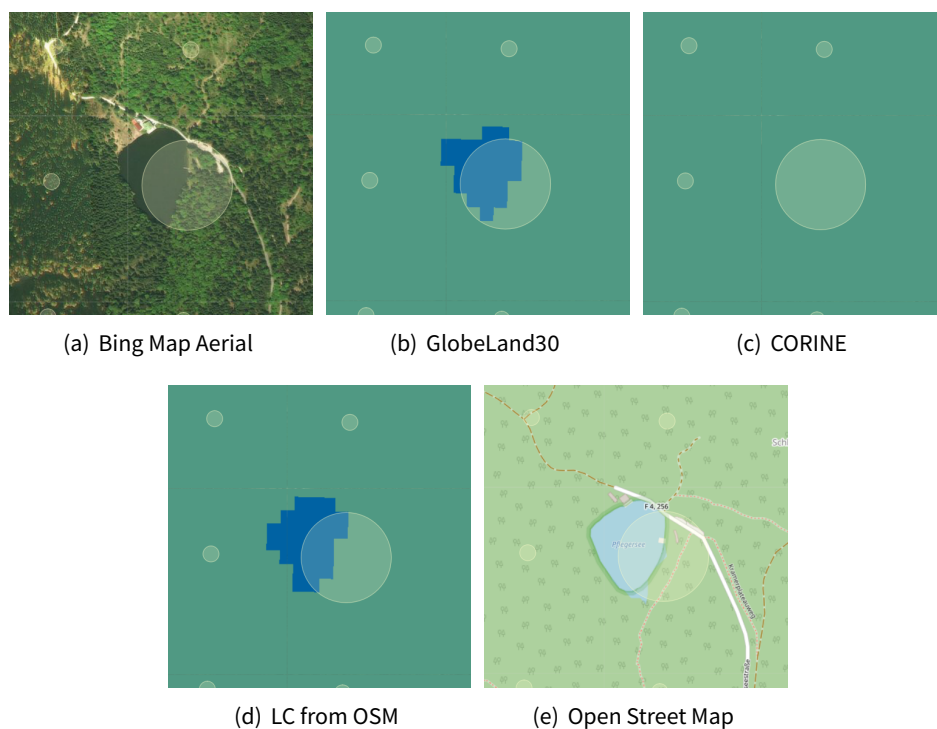


Figure 5.12: Case 1. Missing lake: The Bayesian Network analysis facilitated the discovery of water body area where the data sources disagreed in land cover definition.

To illustrate the first case (Fig. 5.12), one part of a highly uncertain pattern visualized on the map is selected. The specified location was interpreted as

uncertain due to the high heterogeneity in the underlying data. Two of the datasets (GlobalLand30 and land cover from OSM) appear to show that there is a small lake and another dataset suggests a forest class. Even though CORINE is a local dataset, and the minimum mapping unit is smaller than the lake size, this information was not captured in CORINE.

Case 2

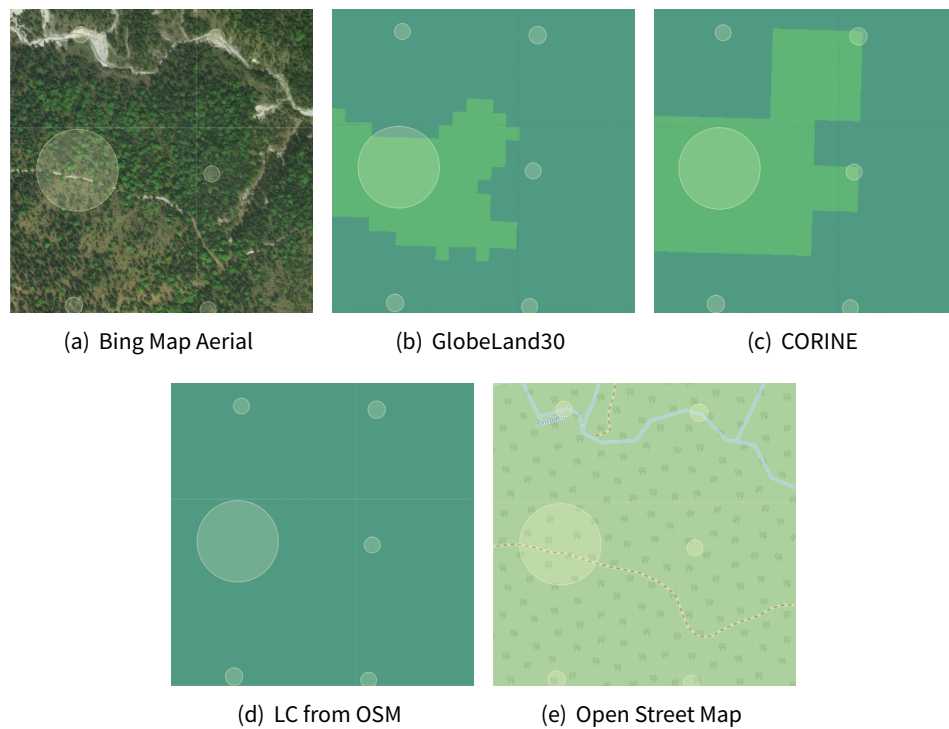


Figure 5.13: Case 2. Forest and shrubland: The analysis indicated that the OSM data miss to define some land cover classes if the definition is vague for a common user.

The second case (Fig. 5.13) shows an area covered by forest and/or shrubland. Such areas are hardly detectable from the Earth Observation data, as they can fall into both categories during the classification process depending on the threshold set. However, the evidence obtained from Globeland30 and CORINE states that this area has mixed characteristics, and one part of it is classified as shrubland and another part as forest. In contrast, OSM data shows that this area is assigned to forest. Considering the nature of OSM data, it can be assumed that it is hard for a common user to distinguish the difference between similar natural phenomena. Moreover, the information from JPD of

GlobeLand30 and the OSM illustrates that class shrubland mostly co-occurs with forest, bareland and shrubland. Therefore, there is some degree of ambiguity in definition of the shrubland in OSM data.

Case 3

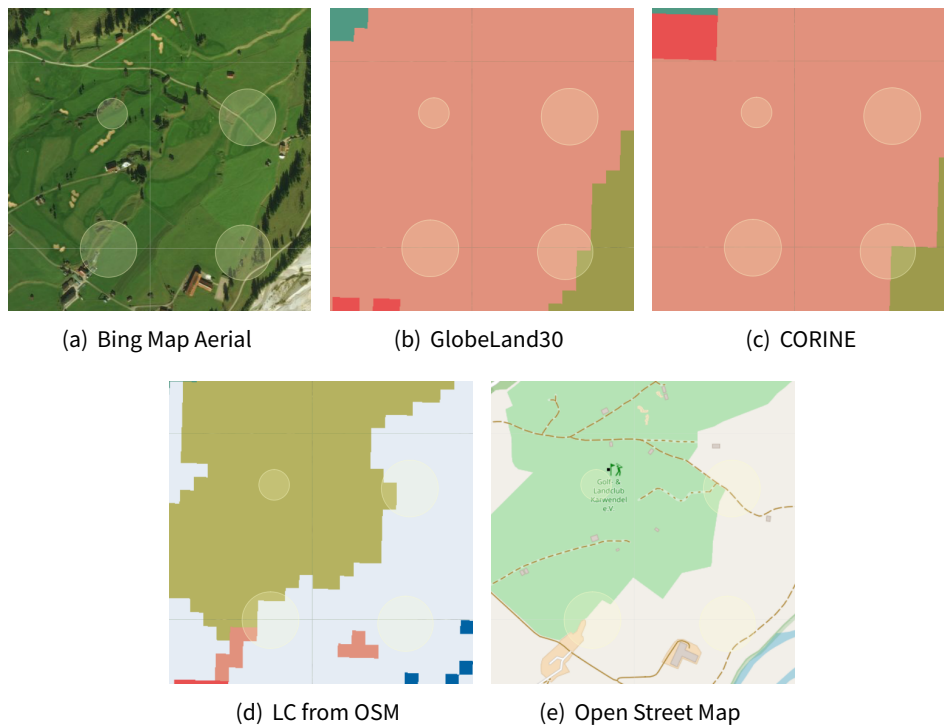


Figure 5.14: Case 3. Grassland and cultivated areas: the OSM data is often more detailed and precise in recognizing particular land structures. For example a golf course should be labeled as grassland rather than cultivated area.

The third case (Fig. 5.14) shows that OSM data is more detailed and precise in recognizing particular land structures. Some areas of vegetation are hard to classify from Earth Observation data due to their similarity. For instance, cultivated areas represent lands for agriculture, gardens, and irrigated and dry farmlands, whereas the lands covered by natural grass belong to the class of grassland. The land cover classification process might introduce some degree of ambiguity. Thus, as can be seen in Fig. 5.14, a misinterpretation of grassland occurs, a golf course in this example was classified as cultivated area, though it should be rather labeled as grassland or artificial land. Such cases, may be better resolved when volunteered geographic data is used to improve training sites for the classification procedure.

5.2 Scenario 2. Video surveillance

In the second scenario, we present a visual analytics application for spatial and heterogeneous data analysis that uses integrated Bayesian Network to support reasoning about classification task of surveillance camera distribution. The application is based on the publication for EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3) (Chuprikova et al. 2018). More specifically, we address the issue of decision-making in terms of classification under uncertainty. When decision makers deal with classification problems, they often quantify the likelihood of a given class on the basis of their personal knowledge. The application provides a visual analytics interface that supports expert input to the Bayesian Network's probabilistic model. The classification task is conducted using locations of surveillance cameras in public places in the city of Moscow. The provision of multiple kinds of surveillance cameras may increase the chance that if anything happens in a location it will be possible to: (a) notice that something happened, and (b) gain some information about what happened and perhaps who was responsible for the happening. The application would also allow an analyst to explore the likelihood of an undesirable/unsafe happening without it being noticed (or without it being noticed soon enough to counteract). Interactive visualization enables an analyst to participate in the data exploration and reason about the information provided by the data considering the state of uncertainty that is inherently incorporated during the data exploration (Chuprikova et al. 2018).

5.2.1 Data

As firstly described in the novel 1984 by George Orwell, camera surveillance turned out to be a reality for most of the big cities, where privacy was sacrificed for technological advances that high-tech can give to the public. Despite the debates of whether closed-circuit television (CCTV), also known as camera surveillance should be installed in public spaces since decades it has become a critical tool for a variety of tasks such as law enforcement, personal safety, traffic control, resource planning, and security of assets (Upmanyu et al. 2009). Being a large city Moscow has got more than 2,500 surveillance cameras installed in mass gathering places, such as subway exits, squares, shopping malls, to name a few. These cameras operate seven days a week and 24 hours a day to provide material to the authority and citizens to ensure the safety of residents when necessary. The cameras gather information in the form of video and screenshots. The citizen can request the data by contacting a relevant authority within 5 days after an accident. In order to localize the camera position, the data can be visualized on a map, thus in an event of an incident,

the camera location and coverage can be found. Apart from mass gathering places, more than 20 000 of video surveillance cameras are installed in common areas of residential complexes to monitor the efficiency of public utilities in terms of area cleaning, garbage disposal, gardening and land improvement, control compliance with the rules of parking vehicles, and increase the level of safety of residents in the city of Moscow. The current research leverages data acquired from the open portal of the municipality of Moscow. The primary input information includes data from video surveillance cameras in the city installed in: (a) mass gathering places, (b) common areas of residential complexes, and (c) public points of police assistance. The point data from the cameras was aggregated to a hexagonal grid (see Fig. 5.15), with a size selected to be as geographically precise as possible while containing one or more cameras from each category in most cells. The density of the cameras in each grid cell is described as high, medium and low. The qualitative description of the cameras' density is given in order to process this information in the Bayesian Network. In addition, this description facilitates human reasoning with uncertain knowledge (Osseiran 2001). The point grid, which is used for the final visualization, represents centers of the hexagons and provides an overview over the whole city area.

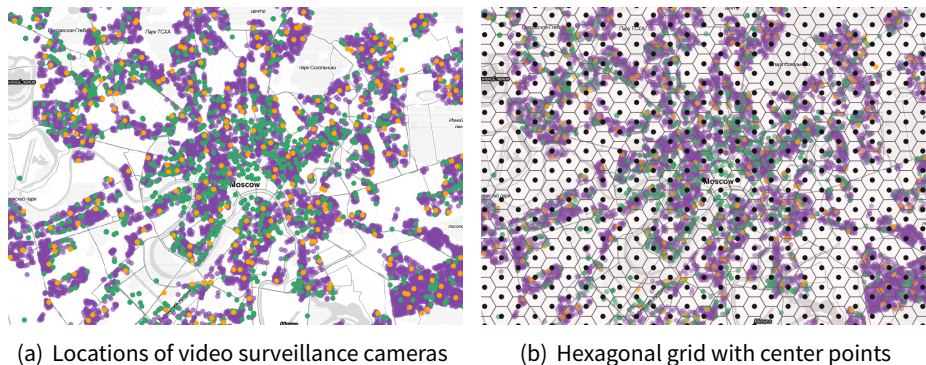


Figure 5.15: Video surveillance data preprocessing. The point data from the video surveillance cameras (left) and is aggregated to hexagonal grids (right).

5.2.2 Experimental set-up

In this scenario, we apply Bayesian Networks to represent probabilistic knowledge about spatially referenced locations of video surveillance cameras.

The qualitative component is a graphical model structure of the dependencies among the variables. The DAG consists of nodes (random variables) and edges that connect these nodes (see Fig. 5.16). Nodes are the labeled circles. Edges define probabilistic relations among nodes. Here, our aim to explore the

geographic coverage provided by three kinds of public security cameras, particularly for use in monitoring activity associated with potential mass gathering events. The three data sets are: residential complex cameras (R), mass gathering cameras (M), and police station surveillance cameras (P). Each is depicted in the Bayesian Network as a node. Such discrete nodes are also called parent nodes, as they do not have predecessors. Furthermore, the parent nodes R, M, P, and the "Mass event" node are described by prior probability distributions. For the nodes, R, M, and P the prior probability distributions are given based on the density of the locations within a given grid cell and characterized using qualitative values: high, medium, and low.

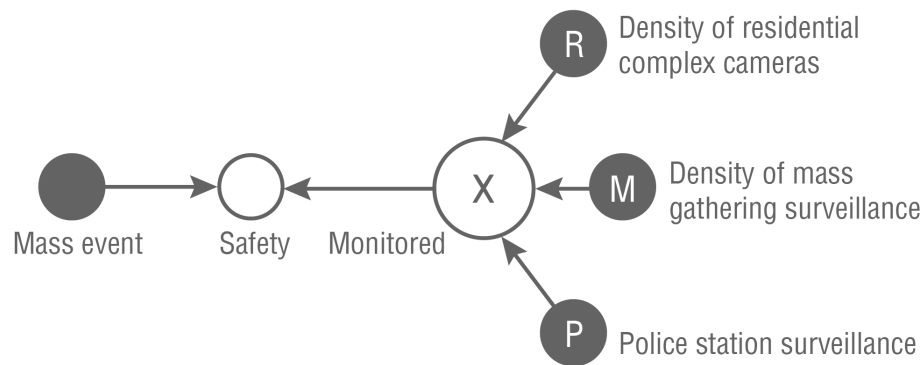
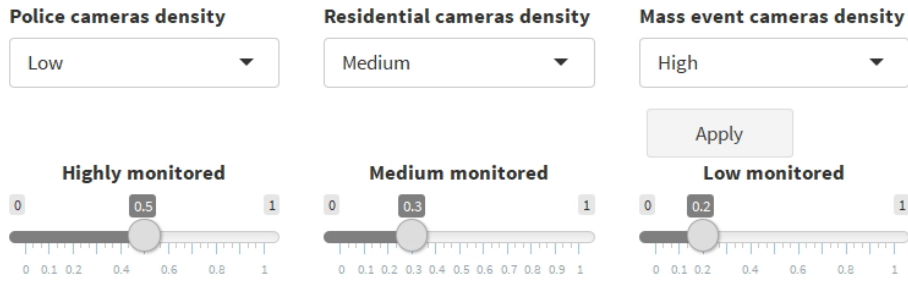


Figure 5.16: Scenario 2: Bayesian Network structure for reasoning about public surveillance data.

The quantitative component of the Bayesian Network is determined by CPTs and can be elicited from an analyst via the visual analytics interface. The interaction process is realized through the CPT panel, where the conditional probabilities can be adjusted in order to see how the distributions change. The node X (Monitored) is a child node of R, M, and P; therefore, it is defined by a probability distribution over its outcomes (highly monitored, medium monitored, low monitored), conditional on the outcomes of its predecessors (nodes R, M, and P) (see Fig. 5.17). The size of the CPTs in the Bayesian Network is exponential in the number of parent nodes. In the case given, three nodes can generate a table with 27 columns. Given that the node X includes a prior probability distribution with three outcomes (highly monitored, medium monitored, low monitored), the CPT for this node has three rows, therefore, there are 81 possible intersections.

Monitored

Selected conditional dependencies:



High Police camera density:

High Residential Camera D

low-	0.1	0.6	0.3
medium-	0.1	0.6	0.3
high-	0.1	0.2	0.7

Medium Residential Camera D

low-	0.1	0.6	0.3
medium-	0.1	0.5	0.4
high-	0.1	0.2	0.7

Low Residential Camera D

low-	0.1	0.4	0.5
medium-	0.1	0.4	0.5
high-	0.05	0.05	0.9

Medium Police camera density:

High Residential Camera D

low-	0.1	0.4	0.5
medium-	0.1	0.4	0.5
high-	0.05	0.05	0.9

Medium Residential Camera D

low-	0.1	0.4	0.5
medium-	0.05	0.9	0.05
high-	0.1	0.4	0.5

Low Residential Camera D

low-	0.1	0.4	0.5
medium-	0.1	0.4	0.5
high-	0.1	0.4	0.5

Low Police camera density:

High Residential Camera D

low-	0.9	0.05	0.05
medium-	0.1	0.6	0.3
high-	0.1	0.2	0.7

Medium Residential Camera D

low-	0.1	0.6	0.3
medium-	0.1	0.5	0.4
high-	0.1	0.2	0.7

Low Residential Camera D

low-	0.1	0.4	0.5
medium-	0.1	0.4	0.5
high-	0.05	0.05	0.9

Figure 5.17: Conditional probability table for the node X "Monitored". The CPT can be interactively elicited from the analyst.

5.2.3 Visual encoding and interaction design

The prototypical design of the visual analytics application is illustrated in Fig. 5.18. The numerical values of subjective belief in the interval $[0, 1]$ are introduced as interactive heat maps within panels "Monitored" (Fig. 5.17) and "Safety". In the panel "Monitored", a potential analyst can insert values within the CPT based on how the given data are related. For example, one assumes that a high density of police and residential complex cameras and a medium density of mass gathering cameras result in the city zone being classified as "highly monitored" with a given probability. Such an assumption may be inferred from the fact that mass gathering cameras have a wider angle of observation, thus fewer cameras are needed to cover an area seamlessly. The same procedure is valid for the node "Safety" as it is defined by outcomes of its predecessors: "Monitored" and "Mass Event". The numerical parameters of the CPTs are visualized and users can interactively change them according to their own subjective beliefs. The inference outcome is registered in the data attribute and it includes a categorical value "noticed/unnoticed" and a numerical value, which is the value of the conditional distribution.

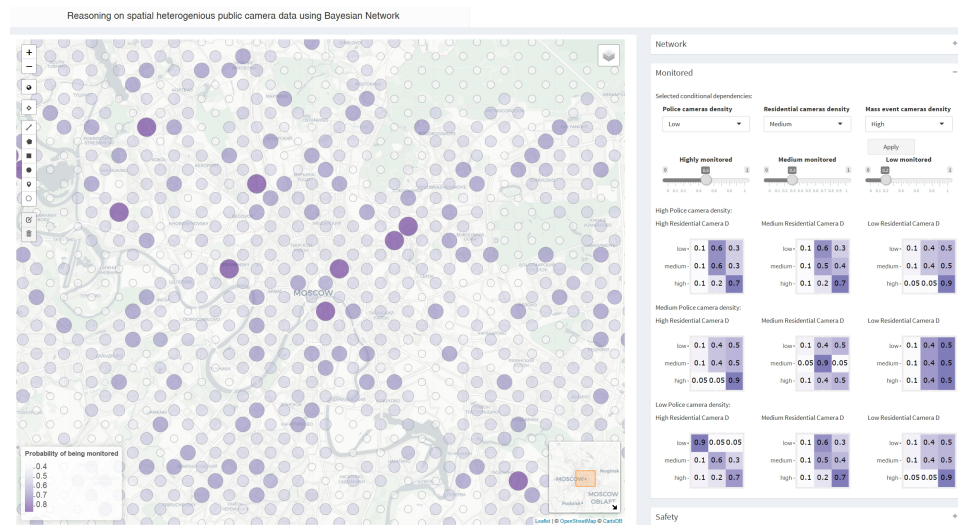


Figure 5.18: Uncertainty-aware visual analytics interface. The interface is developed using data of public surveillance cameras installed in mass gathering places, common areas of residential complexes, and points of police assistance.

The map panel, shown in Fig. 5.19, represents the outcome of the modeling in its spatial context. In this scenario, we chose proportional circles to indicate the likelihood of different city zones being monitored. The proportional circles represent the data set discretely, showing patterns where the cameras' coverage is dense, rather than representing an interpolated value over the city district. The size and color of circles directly signify the probability value at a

given location and may be updated when new evidence is set (for instance, evidence of a mass gathering event or updated conditional distribution with CPT for "Monitored"). The map panel also includes zoom, selection, and layer controls, map legend, and an overview map for a rapid navigation. Additionally, the user can browse the map and obtain details about each point.

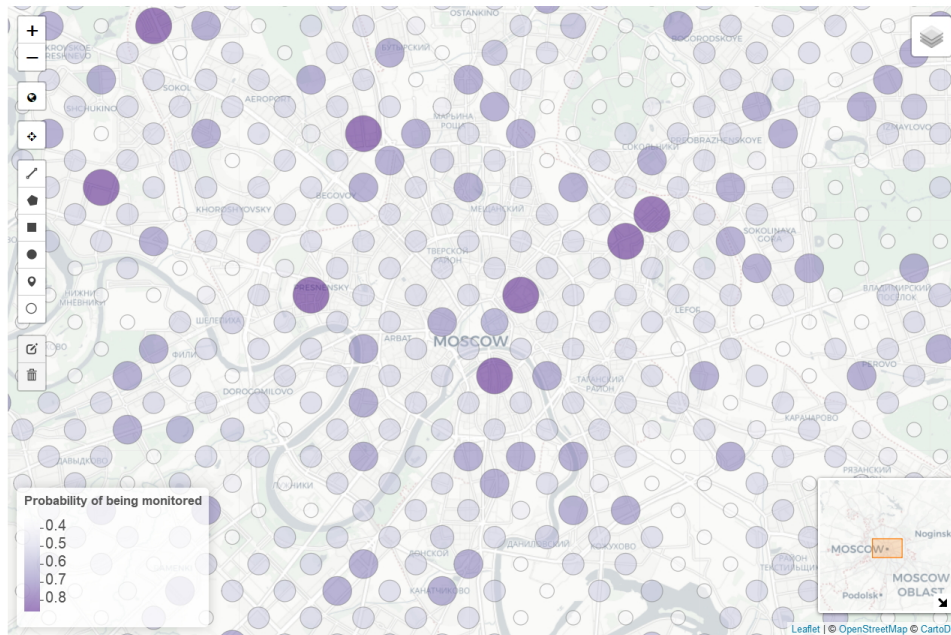


Figure 5.19: Map panel. The size and color of the proportional circles characterize the likelihood of different city zones being monitored given the probability value in the interval $[0, 1]$.

5.2.4 Results

This visual analytics application has been prototypically implemented using interoperable methods from an open system setting. The prototype is supported by a three-tier architecture including Database Management System (DBMS), namely PostgreSQL, Servlet Engine (GeoServer and R) and Client side components (R Shiny). It adopts R packages bnlearn, that provides Bayesian Network structure, parameter learning and inference (Scutari 2009), and gRain that implements propagation algorithm in Bayesian Network (Højsgaard 2013).

This scenario demonstrates the usefulness of an uncertainty-aware visual analytics application for reasoning about surveillance camera data from three sources in Moscow: mass gathering places, residential complexes, and points of police assistance. The connections among spatial data, Bayesian Network

modeling, visualization, and the users allows interactive examination of city areas monitored by public camera facilities under the subjective judgment of the spatial data availability, quality, and relevance. The application has the flexibility of providing probabilistic predictions based on user input and the Bayesian Network model structure; therefore it combines both human and Bayesian reasoning. The user interaction involves user-defined quantitative input and an interactive map. The analysis is directly driven by inputs of the analysts. The output of the analysis is visualized using proportional circles to signify the probability of public safety under consideration of a mass event occurrence. The results of the analysis of heterogeneous public security data aim to complement the existing city models with insights for designing effective city planning strategies and supporting decision-making in environmental modeling.

5.3 Summary

In spite of the awareness of spatial data uncertainty and its significance, little attention has been given to reasoning/decision-making under conditions of uncertainty (MacEachren 2015). Taking classification tasks based on uncertain and heterogeneous spatial data as case studies, we combine a visual analytics approach with Bayesian statistics to address four objectives:

- Evaluate the feasibility of using a probabilistic graphical model, namely a Bayesian Network, to represent conditional dependencies utilizing heterogeneous spatial data;
- Provide visualization support where results of the inference can be observed in a spatial context;
- Support users to express their subjective beliefs as conditional probabilities;
- Advance the power of visual analytics by integrating spatial data and computational capacity.

The capabilities of our approach are demonstrated by means of two classification scenarios based on different datasets. The two scenarios are intended to support the land cover classification on the one hand and the city zone classification the other hand. The visual analytics applications combine probabilistic inferential method, namely Bayesian Network, applied to spatial data and an interactive visual interface that supports human reasoning. The uncertainty-aware prototypes are able to integrate data, computational capacity, and visualization facility to support human-computer interaction processes for each application, and have thus revealed the potential of advanced visual analytics.

However, the findings have a number of possible limitations. These include the low quality of existing reference data, low quality of volunteered contributions to the OSM, and the vagueness of prior expert knowledge about the data, as Bayesian Network requires robust domain knowledge to assign meaningful probabilities. In spite of these shortcomings, Bayesian Network has been proved versatile and useful for handling incomplete and diversified data. Its integration with visual analytics approach enhances the computational capability to discover patterns of interest under uncertain reasoning process.

In future research, we plan to conduct a usability study with a group of analysts to evaluate the feasibility of the approach in different scenarios. Another issue to be tackled is scalability of the interface design when more data is fused into the Bayesian Network as the conditional probabilities will grow exponentially.

Chapter 6

Conclusion and outlook

6.1 Summary of thesis achievements

The main focus of this research is on the visual analytics approach that integrates Bayesian data analysis and human reasoning about spatial data. More specifically, we proposed a design framework to support the reasoning process under uncertain conditions, such as performing a classification task on spatial data and developed a visual analytics prototype, which is capable of obtaining probabilistic predictions based on user input and Bayesian updating; therefore it combines both human and Bayesian reasoning.

Foundational research works provided relevant definitions of uncertainty, introduced the available methods for assessing uncertainty in the reasoning process, and approaches for visual representations within software packages and analytical applications. Based on the literature reviewed, we have concluded that Bayesian Networks can provide an effective framework for knowledge representation and reasoning. Unlike "black-box" techniques (i.e, neural networks), Bayesian Networks are capable of considering a large number of interrelated variables and their instances that are cognitively meaningful and interpretable. In comparison to rule-based systems, Bayesian Networks are logical in their probability calculus; they manage uncertainty and reflect new evidence. Although Bayesian Networks are a widely applied modeling technique in different research fields, there a lack of coherent approaches that combine probabilistic reasoning and visualization means, in particular for the spatial data component.

The applicability of the Bayesian Network-enabled visual analytics is demonstrated on two scenarios using land cover classifications and video surveillance data. The prototyped systems integrate data, computational

capacity, and the human-computer interaction process. The analysis is driven directly by inputs of analysts through the user interface. The outcomes of the analysis are visualized on interactive maps with proportional circles to signify the probability of an unobserved event.

Targeting a classification task focused on heterogeneous spatial data under conditions of uncertainty, we combine a visual analytics approach and Bayesian statistics to serve the following objectives:

Objective 1: To test the feasibility of using a probabilistic graphical model, namely the Bayesian Network, to represent conditional dependencies among heterogeneous data in order to perform a classification task.

Chapter 3 presents the main methodological driving force behind the approach of modeling uncertainty in a reasoning process when dealing with spatial data. It creates the following findings:

- Bayesian Networks are an effective communication approach as it can aggregate knowledge from diverse sources in a graphical structure. The graphical structure is able to map cause-effect relationships among variables, which reduces the need for human involvement in probability computing.
- Bayesian Networks are able to handle missing or incomplete data efficiently. Due to the fact that Bayesian Networks are based on the probability method, which is basically a mathematical perspective on uncertainty, even if an event is unpredictable, the laws of probability can help assess the possible outcome.
- By using probabilistic methods, it is not possible to determine the future outcome, but we can reduce uncertainty and gain a better understanding about an event under examination. Uncertainty can be assessed as entropy or probability associated with a targeted node of the network.
- The mathematically rigorous method of Bayesian Networks can be potentially empowered by visualization for reasoning under conditions of uncertainty.
- Apart from the advantages, Bayesian Networks undergo some limitations in terms of the computational complexity and the large number of parameters required to be assigned by an analyst.

Objective 2: To develop a visual analytics framework that can facilitate the understanding of data and uncertainty in the reasoning process using Bayesian Networks.

Chapter 4 investigated the potential of visual analytics approach to integrate computational facility of Bayesian Networks with visualization

methods, thus facilitate the understanding of the processes and results of spatial data analysis with inherent uncertainty. To this end, the visual analytics system enables domain experts to compute the joint probability distribution between all combinations of variables given in the network. Exploration and assessment of the complex estimation is facilitated by a visual interface comprising multiple views: map, CPTs, and network panels.

The visual analytics framework made the following contributions:

- Benefiting analysts with the power of visual analytics of diverse spatial data and their relationships;
- Supporting users in overcoming the limitations of Bayesian Networks and providing a visual interface that allows the input of subjective beliefs;
- Visualizing the output of the analysis on a digital map using proportional circles to signify the probability (or measure of entropy) of a predicted variable/event.
- Enabling analysts to share insights with others.

Objective 3: To build a prototype of a visual analytics interface that can integrate data, visualization, and computational capacity of Bayesian Networks to facilitate human-computer interactions for data analysis under subjective beliefs in a selected application domain.

Chapter 5 illustrates the capabilities of the proposed uncertainty-aware visual analytics prototype. The approach extends an established technique for reasoning about spatial data, where uncertainty is integrated through visualization support. We illustrate the capabilities of our proposed uncertainty-aware application visual analytics by presenting two case studies.

- The prototype proposed is realized using spatial data, Bayesian Network modeling, visualization, and the users as components in order to examine the spatially heterogeneous data under the subjective judgment of spatial data availability, quality, and relevance.
- The first prototype dedicated to exploring the differences of land cover classifications considering user input on how similar or dissimilar the land cover classes are.
- The second prototype serves the classification task using surveillance camera data from three sources. The results of this analysis may provide new insights for effective city planning.

- The prototypical case studies demonstrate how an analyst can be involved in statistical inference by means of visual interface. The results of the analysis aim to complement existing spatial models with insights, thus support decision-making in environmental modeling.
- The prototypes developed were tested in a small group of users and require further investigation of their performance.

6.2 Research applicability to other domains

This research has demonstrated that the Bayesian Network method has the potential to provide an effective and versatile technique for spatial data analysis. Moreover, due to the broad applicability of Bayesian Network-enabled visual analytics, we expect that the results described will be of interest to other system designers who are considering similar problems. See some proposed scenarios in Fig. 6.1.

Geomarketing solutions

An increasing amount of social and economic data available for geospatial analysis creates a demand for new visual and analytical approaches to analyze it. Thus, Bayesian Network-enabled visual analytics may offer a novel technique that supports prior knowledge awareness and user involvement using interactive visualization. Bayesian Network-enabled visual analytics may utilize data to uncover the structure of the relationships and dependencies among demographic, social, and economic factors. For instance, a classification task can be conducted using socio-economic data to analyze how urban demography affects the distribution of cars in the city (Chuprikova & Meng 2019). The application would also allow an analyst to explore the data under an alternative perspective, where interactive visualization enables an analyst to participate in the data exploration and reason about the information provided by the data considering the state of uncertainty that is inevitably incorporated during the data exploration.

Social communications

A potential field of research that may benefit from visual analytics application is social communication in the spatial context where reasoning under conditions of uncertainty is considered. Social communication, social connectivity, community involvement or social distance go beyond of what one can evaluate if only one source of information is considered. But these fields of research integrate such aspects as physical and human infrastructure, economic and education background, social networks, languages, and so on. In this regard, the involvement of uncertainty in the reasoning process qualitative and quantitative data plays an important role.

Innovation ecosystems development

A mapping of innovation ecosystems is another rising field of research where uncertainty in the reasoning process cannot be ignored and a visual analytics approach can help derive new insights from the data. To investigate the potential of entrepreneurship and economic development, approaches that integrate historical and current data, predictive analytics, expert knowledge, and visualization can support researchers and decision-makers in visualizing entrepreneurial ecosystems, anticipating the development, and interpreting its driving forces and impacts.

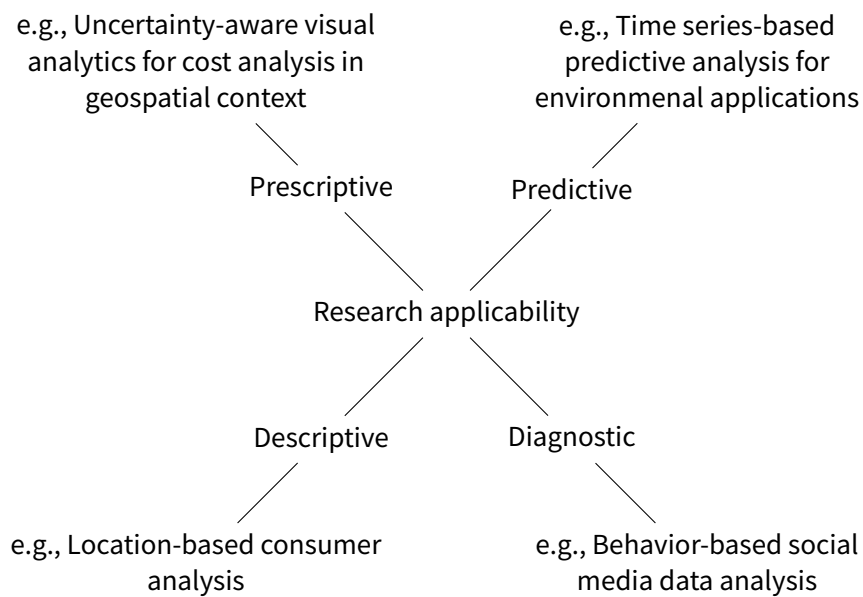


Figure 6.1: Research applicability to other domain
After: Bayes Server: www.bayesserver.com

6.3 Outlook

Extending the heuristics implications in cartography

Although the previous chapters have emphasized that some major advances have been made in addressing issues of reasoning under conditions of uncertainty in GIScience, there is still a long way to go to develop and test such systems that can accommodate analytical reasoning considering prior expert knowledge. The growing interest in uncertainty in reasoning and awareness of their importance and impact in cartography and geographical information systems will eventually drive the scientific community to integrate human-computer interaction, taking into account data, visualization, expert knowledge and statistics to build new approaches for deriving knowledge from large amounts of data. In this regard, the cognitive shortcuts, or heuristics, which simplify the decisions, especially when they are taken in instances of uncertainty, are importance aspects in the reasoning process to consider. Generally, there are different heuristics, some of which are of general purpose such as representativeness, availability, and adjustment from an anchor (for more details see Section 3.4.1). In order to explicitly address the potential negative implications of these three reasoning heuristics, there is a need to conduct further research that will apply cognitive psychology studies to user-interface design.

Extending the methodology for Big Data

The further development of Bayesian Network-enabled systems requires us to find a more informative way of representing the amount of information condensed in a Bayesian Network. The solution is seen in interactive approaches that visually synchronize the graphical network structure and conditional probability tables. To implement a fully adjustable modeling process, a user interface component can integrate visual tools for making changes in Bayesian Network nodes and edges.

Extending the design methodology for application that can include both uncertainty in data and in reasoning process

Another crucial challenge is to integrate data uncertainty along with uncertainty in the reasoning process into one coherent system. The reliability of the resulting system may significantly increase as it would combine different data aspects and expert input.

Usability testing

Moreover, an evaluation of Bayesian Network-enabled visual analytics is necessary to define how easy a design approach is for a group of representative users. To evaluate the user interface design and functionality strategies such as heuristics review and usability testing can be adopted and implemented one after another:

- Heuristics review is a prominent approach for the initial evaluation, which include looking at the user interface and expressing the evaluator's opinion on whether it meets the usability heuristics. The following workflow can be adopted when the heuristics review is conducted:
 1. Define a target group intended to use the application and tasks they may perform (e.g., ecologists for the Scenario 1 (Section 5.1); urban planners for the Scenario 2 (Section 5.2)).
 2. Determine the goals and set of heuristics. The heuristics should evaluate the goals set. Recent studies indicate various heuristics that can be applied during the evaluation. At least nine basic usability heuristics can be identified (Nielsen & Molich 1990): simple and natural dialog; speak the user's language; minimize user memory load; be consistent; provide feedback; provide clearly marked exits; provide shortcuts; good error messages; prevent error.
 3. Select evaluators among usability experts that have experience in data visualization. It is suggested to evaluate the user interface with around five experts, and additional resources spend on alternative usability testing methods.
- Cognitive walkthrough is used to evaluate the visualization product with new or occasional users. This approach is using a more specific procedure, which simulated the user's problem at each step of the evaluation when the user's response is registered. This method gives particular attention to the cognitive aspects of users' experience (Ning et al. 2019).
 1. Define the user and their goals when working with the user interface. For instance, ecologists within Scenario 1 (Section 5.1) may be interested in the uncertainty of the land cover products. The urban planners within the Scenario 2 (Section 5.2) may want to find out the city districts with the highest density of the monitoring facilities.
 2. Specify the representative tasks the target users may conduct.
 3. Describe the correct actions that the users should make to solve the intended task.
 4. Record the walkthroughs during the interface exploration.

Moreover, the next step in the development of Bayesian Network-enabled visual analytics is to test the application performance using performance metrics. As the computational software approach is developed to calculate and display the inference output, its prototype must be evaluated to ensure that it performs in a time-effective manner.

Bibliography

- Abebe, Y., Kabir, G. & Tesfamariam, S. (2018), 'Assessing Urban Areas Vulnerability to Pluvial Flooding Using GIS Applications and Bayesian Belief Network Model', *Journal of Cleaner Production* **174**, 1629–1641.
- Aitkenhead, M. & Aalders, I. (2009), 'Predicting Land Cover Using GIS, Bayesian and Evolutionary Algorithm Methods', *Journal of Environmental Management* **90**(1), 236 – 250.
- Arsanjani, J. J., See, L. & Tayyebi, A. (2016), 'Assessing the Suitability of GlobeLand30 for Mapping Land Cover in Germany', *International Journal of Digital Earth* **9**(9), 873–891.
- Bastin, L., Cornford, D., Jones, R., Heuvelink, G. B. M., Pebesma, E., Stasch, C., Nativi, S., Mazzetti, P. & Williams, M. (2013), 'Managing Uncertainty in Integrated Environmental Modelling: The UncertWeb Framework', *Environmental Modelling and Software* **39**, 116–134.
- Bertin, J. (1983), *Semiology of Graphics*, University of Wisconsin Press.
- Bloch, I. (2006), 'Spatial Reasoning under Imprecision Using Fuzzy Set Theory, Formal Logics and Mathematical Morphology', *International Journal of Approximate Reasoning* **41**(2), 77 – 95. Advances in Fuzzy Sets and Rough Sets.
- Box, G. (1979), Robustness in the Strategy of Scientific Model Building, in R. L. Launer & G. N. Wilkinson, eds, 'Robustness in Statistics', Academic Press, pp. 201 – 236.
- Brase, G. L. (2009), 'Pictorial Representations in Statistical Reasoning', *Applied Cognitive Psychology* **23**(3), 369–381.
- Broad, K., Leiserowitz, A., Weinkle, J. & Steketee, M. (2007), 'Misinterpretations of the "Cone of Uncertainty" in Florida during the 2004 Hurricane Season', *Earth Interactions* **88**(5), 651–667.
- Brodlie, K., Allendes Osorio, R. & Lopes, A. (2012), 'A Review of Uncertainty in Data Visualization', *Expanding the Frontiers of Visual Analytics and Visualization* pp. 81–109.

- Brovelli, M., Molinari, M., Hussein, E., Chen, J. & Li, R. (2015), 'The First Comprehensive Accuracy Assessment of GlobeLand30 at a National Level: Methodology and Results', *Remote Sensing* **7**(4), 4191–4212.
- Cancer Stat Facts: Stomach Cancer* (2018), <https://seer.cancer.gov/statfacts/html/stomach.html>. Accessed: 2018-08-01.
- Carnap, R. (1945), 'On Inductive Logic', *Philosophy of Science* **12**(2), 72–97.
- Casscells, W., Schoenberger, A. & B. Graboys, T. (1978), 'Interpretation by Physicians of Clinical Laboratory Results', *The New England Journal of Medicine* **299**(18), 999–1001.
- Celio, E., Koellner, T. & Grêt-Regamey, A. (2014), 'Modeling Land Use Decisions with Bayesian Networks: Spatially Explicit Analysis of Driving Forces on Land Use Change', *Environmental Modeling and Software* **52**, 222 – 233.
- Champion, C. & Elkan, C. (2017), 'Visualizing the Consequences of Evidence in Bayesian Networks', *Computing Research Repository* **ABS/1707.00791**.
- Chee, Y. E., Wilkinson, L., Nicholson, A. E., Quintana-Ascencio, P. F., Fauth, J. E., Hall, D., Ponzio, K. J. & Rumpff, L. (2016), 'Modeling Spatial and Temporal Changes with GIS and Spatial and Dynamic Bayesian Networks', *Environmental Modelling and Software* **82**, 108 – 120.
- Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X. & Mills, J. (2015), 'Global Land Cover Mapping at 30m Resolution: A POK-based Operational Approach', *ISPRS Journal of Photogrammetry and Remote Sensing* **103**, 7–27.
- Chiang, C.-h., Shaughnessy, P., Livingston, G. & Grinstein, G. (2005), 'Visualizing Graphical Probabilistic Models', *Technical Report, UML CS*.
- Choi, J. & Usery, E. L. (2004), 'System Integration of GIS and a Rule-Based Expert System for Urban Mapping', *Photogrammetric Engineering & Remote Sensing* **70**(2), 217 – 224.
- Chuprikova, E., Liebel, L. & Meng, L. (2017), 'Towards Seamless Validation of Land Cover Data', *Proceedings. ICC 2017 - International Cartography Conference* pp. 1–10.
- Chuprikova, E., MacEachren, A. M., Cron, J. & Meng, L. (2018), Visual Analytics-enabled Bayesian Network Approach to Reasoning about Public Camera Data, *in* K. Lawonn, N. Smit, L. Linsen & R. Kosara, eds, 'EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3)', The Eurographics Association.

- Chuprikova, E. & Meng, L. (2019), 'Reasoning about socio-economic data: a visual analytics approach to bayesian network', *International Journal of Cartography* **0**(0), 1–17.
URL: <https://doi.org/10.1080/23729333.2019.1613073>
- Cobb, B. R. & Shenoy, P. P. (2003), 'A comparison of Bayesian and belief function reasoning', *Information Systems Frontiers* **5**(4), 345–358.
- Cole, W. G. (1989), 'Understanding Bayesian Reasoning via Graphical Displays', *SIGCHI Bull.* **20**(SI), 381–386.
- Colombo, M., Elkin, L. & Hartmann, S. (2018), 'Being Realist about Bayes, and the Predictive Processing Theory of Mind', *The British Journal for the Philosophy of Science* p. axy059.
- Conrady, S. & Jouffe, L. (2015), *Bayesian Networks and BayesiaLab: A Practical Introduction for Researchers*, Bayesia USA.
- Correa, C. D., Chan, Y. H. & Kwan-Liu, M. (2009), 'A Framework for Uncertainty-Aware Visual Analytics', *VAST 09 - IEEE Symposium on Visual Analytics Science and Technology, Proceedings* pp. 51–58.
- Cossalter, M., Mengshoel, O. J. & Selker, T. (2011), 'Visualizing and Understanding Large-Scale Bayesian Networks', *AAAI-11 Workshop on Scalable Integration of Analytics and Visualization* **23**(August), 12–21.
- Cox, J., House, D. & Lindell, M. (2013), 'Visualizing Uncertainty in Predicted Hurricane Tracks', *International Journal for Uncertainty Quantification* **3**(2), 143–156.
- Darwiche, A. (2008), 'Chapter 11. Bayesian Networks - Handbook of Knowledge Representation', *Foundations of Artificial Intelligence* **3**(07), 467–509.
- De Bièvre, P. (2012), 'The 2012 International Vocabulary of Metrology: VIM', *Accreditation and Quality Assurance* **17**(2), 231–232.
- Díez, F. J. & Druzdzel, M. J. (2003), *Reasoning Under Uncertainty*, Nature Publishing Group, London, pp. 880–886.
- Dodge, M., McDerby, M. & Turner, M. (2008), *Geographic Visualization: Concepts, Tools and Applications: Concepts, Tools and Applications*, John Wiley and Sons Ltd, United Kingdom.
- Drury, B., Valverde-Rebaza, J., Moura, M.-F. & de Andrade Lopes, A. (2017), 'A Survey of the Applications of Bayesian Networks in Agriculture', *Engineering Applications of Artificial Intelligence* **65**(Supplement C), 29 – 42.
- Fisher, P., Comber, A. & Wadsworth, R. (2010), 'Approaches to Uncertainty in Spatial Data', *Fundamentals of Spatial Data Quality* pp. 43–59.

- Fisher, P. F. (1986), 'Models of Uncertainty in Spatial Data', *Geographical Information Systems: Principles, Techniques, Management and Applications* pp. 191–205.
- Fonte, C. C., Bastin, L., See, L., Foody, G. & Lupia, F. (2015), 'Usability of VGI for Validation of Land Cover Maps', *International Journal of Geographical Information Science* **29**(7), 1269–1291.
- Fonte, C., Minghini, M., Antoniou, V., See, L., Patriarca, J., Brovelli, M. & Milcinski, G. (2016), Automated Methodology for Converting OSM Data into a Land Use/Cover Map, in '6th International Conference on Cartography & GIS'.
- Foody, G. (1996), 'Fuzzy Modelling of Vegetation from Remotely Sensed Imagery', *Ecological Modelling* **85**(1), 3 – 12. Fuzzy Logic in Ecological Modelling.
- Friederichs, H., Ligges, S. & Weissenstein, A. (2014), 'Using Tree Diagrams without Numerical Values in Addition to Relative Numbers Improves Students' Numeracy Skills: A Randomized Study in Medical Education', *Medical Decision Making* **34**(2), 253–257.
- Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F. & Obersteiner, M. (2009), 'Geo-Wiki.org: The Use of Crowdsourcing to Improve Global Land Cover', *Remote Sensing* **1**(3), 345–354.
- Gallego, J. (2001), 'Comparing CORINE Land Cover with a More Detailed Database in Arezzo (Italy)', *JRC* pp. 1–8.
- Gigerenzer, G. & Hoffrage, U. (1995), 'How to Improve Bayesian Reasoning without Instruction: Frequency Formats', *Psychological Review* **102**(4), 684–704.
- Goodchild, M. F. (2007), 'Citizens as Sensors: the World of Volunteered Geography', *GeoJournal* **69**(4), 211–221.
- Goodchild, M. F. (2008), *Imprecision and Spatial Uncertainty*, Springer US, Boston, MA, pp. 480–483.
- Griethe, H. & Schumann, H. (2006), 'The Visualization of Uncertain Data: Methods and Problems', *Proceedings of SimVis VI*(August), 143–156.
- Hegarty, M., Friedman, A., Boone, A. P. & Barrett, T. J. (2016), 'Where Are You? The Effect of Uncertainty and its Visual Representation on Location Judgments in GPS-like Displays', *Journal of Experimental Psychology: Applied* **22**(4), 381.
- Helbich, M., Amelunxen, C., Neis, P. & Zipf, E. (2012), Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata, in 'In Jekel T, Car A, Strobl J and Griesebner G: Geospatial Crossroads, GI Forum 2012. Proceedings of the Geoinformatics Forum Salzburg'.

- Højsgaard, S. (2013), 'Bayesian Networks in R with the gRain package', *Aalborg University* **46**(2012), 1–17.
- Holyoak, K. J. & Morrison, R. G. (2012), *The Oxford Handbook of Thinking and Reasoning*, Oxford University Press.
URL: <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199734689.001.0001/oxfordhb-9780199734689>
- House, L., Leman, S. & Han, C. (2015), 'Bayesian Visual Analytics: BaVA', *Statistical Analysis and Data Mining* **8**(1), 1–13.
- Hunter, B. (2016), Clarence Irving Lewis, in E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2016 edn, Metaphysics Research Lab, Stanford University.
- Johnson-Laird, P. N. (1999), 'Deductive Reasoning', *Annual Review of Psychology* **50**(1), 109–135. PMID: 15012459.
- Johnson-Laird, P. N. (2010), 'Mental Models and Human Reasoning', *Proceedings of the National Academy of Sciences* **107**(43), 18243–18250.
- Jones, J. (2015), 'Information Graphics and Intuition: Heuristics as a Techne for Visualization', *Journal of Business and Technical Communication* **29**(3), 284–313.
- Jost, L. (2006), 'Entropy and Diversity', *Oikos* **113**(2), 363–375.
- Kahneman, D. (2011), *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York.
- Kahneman, D. & Tversky, A. (1982), 'Variants of Uncertainty', *Cognition* **11**(2), 143–157.
- Keil, M., Bock, M., Esch, T., Metz, A., Nieland, S. & Pfitzner, A. (2011), 'CORINE Land Cover - Aktualisierung 2006 für Deutschland'.
- Keim, D., Kohlhammer, J., Ellis, G. & Mansmann, F. (2010), 'Mastering the Information Age: Solving Problems with Visual Analytics, Eurographics Association'.
- Kenett, R. S. (2016), 'On Generating High InfoQ with Bayesian Networks', *Quality Technology and Quantitative Management* **13**(3), 309–332.
- Kerman, J., Gelman, A., Zheng, T. & Ding, Y. (2008), *Visualization in Bayesian Data Analysis*, Springer, Berlin, Heidelberg.
- Kinkeldey, C. (2014), 'Development of a Prototype for Uncertainty-Aware Geovisual Analytics of Land Cover Change', *International Journal of Geographical Information Science* **28**(10), 2076–2089.

- Kinkeldey, C., MacEachren, A. M. & Schiewe, J. (2014), 'How to Assess Visual Communication of Uncertainty? A Systematic Review of Geospatial Uncertainty Visualisation User Studies', *The Cartographic Journal* **51**(4), 372–386.
- Kinkeldey, C., Mason, J., Klippel, A. & Schiewe, J. (2014), 'Evaluation of Noise Annotation Lines: Using Noise to Represent Thematic Uncertainty in Maps', *Cartography and Geographic Information Science* **41**(5), 430–439.
- Kirschenbaum, S. S., Trafton, J. G., Schunn, C. D. & Trickett, S. B. (2014), 'Visualizing Uncertainty', *Human Factors* **56**(3), 509–520.
- Kiureghian, A. D. & Ditlevsen, O. (2009), 'Aleatory or Epistemic? Does it Matter?', *Structural Safety* **31**(2), 105 – 112. Risk Acceptance and Risk Communication.
- Klayman, J. & Brown, K. (1993), 'Debias the Environment instead of the Judge: an Alternative Approach to Reducing Error in Diagnostic (and Other) Judgment', *Cognition* **49**(1), 97 – 122.
- Koiter, J. (2006), 'Visualizing Inference in Bayesian Networks', p. 125.
- Korb, K. B. & Nicholson, A. E. (2010), 'Chapter 1: Bayesian Reasoning', *Bayesian Artificial Intelligence* pp. 3–28.
- Körner, C., Liebig, T. & May, M. (2009), Fast Visual Trajectory Analysis Using Spatial Bayesian Networks, in '2009 IEEE International Conference on Data Mining Workshops(ICDMW)', Vol. 00, pp. 668–673.
- Krüger, C. & Lakes, T. (2014), 'Bayesian Belief Networks as a Versatile Method for Assessing Uncertainty in Land-Change Modeling', *International Journal of Geographical Information Science* **29**(1), 111–131.
- Kumpf, A., Tost, B., Baumgart, M., Riemer, M., Westermann, R. & Rautenhaus, M. (2018), 'Visualizing Confidence in Cluster-Based Ensemble Weather Forecast Analyses', *IEEE Transactions on Visualization and Computer Graphics* **24**(1), 109–119.
- Landuyt, D., Broekx, S., Van der Biest, K. & Goethals, P. (2014), 'Probabilistic Mapping With Bayesian Belief Networks: An Application On Ecosystem Service Delivery In Flanders, Belgium', *7th International congress on Environmental Modelling and Software (iEMSs 2014)* .
- Lelieveld, J., Kunkel, D. & Lawrence, M. G. (2012), 'Global Risk of Radioactive Fallout after Major Nuclear Reactor Accidents', *Atmospheric Chemistry and Physics* **12**(9), 4245–4258.
- Li, L., Wang, J., Leung, H. & Jiang, C. (2010), 'Assessment of Catastrophic Risk Using Bayesian Network Constructed from Domain Knowledge and Spatial Data', *Risk Analysis* **30**(7), 1157–1175.

- Lin, X., Li, H., Zhang, Y., Gao, L., Zhao, L. & Deng, M. (2017), 'A Probabilistic Embedding Clustering Method for Urban Structure Detection', *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLII-2/W7**, 1263–1268.
- Longley, P. A., Goodchild, M. F., Maguire, D. J. & Rhind, D. W. (2011), *Geographical Information Systems and Science*, Wiley Publishing.
- Look, B. C. (2017), Gottfried wilhelm leibniz, in E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', summer 2017 edn, Metaphysics Research Lab, Stanford University.
- Lucchesi, L. R. & Wikle, C. K. (2017), 'Visualizing Uncertainty in Areal Data with Bivariate Choropleth Maps, Map Pixelation and Glyph Rotation', *Stat* **6**(1), 292–302.
- Luxhøj, J. T. (2014), 'A Conceptual Object-Oriented Bayesian Network (OOBN) for Modeling Aircraft Carrier-based UAS Safety Risk', *Journal of Risk Research* **18**(October), 1–29.
- MacEachren, A. M. (1992), 'Visualizing Uncertain Information', *Cartographic Perspective* **13**(13), 10–19.
- MacEachren, A. M. (2015), Visual Analytics and Uncertainty: Its Not About the Data, in E. Bertini & J. C. Roberts, eds, 'EuroVis Workshop on Visual Analytics (EuroVA)', The Eurographics Association.
- MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M. & Hetzler, E. (2005), 'Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know', *Cartography and Geographic Information Science* **32**(3), 139–160.
- Madsen, A. L., Lang, M., Kjærulff, U. B. & Jensen, F. (2003), 'The Hugin Tool for Learning Bayesian Networks', In: Nielsen T.D., Zhang N.L. (eds) *Symbolic and Quantitative Approaches to Reasoning with Uncertainty. ECSQARU 2003. Lecture Notes in Computer Science* **2711**, 594–605.
- Malpica, J., Alonso, M. & Sanz, M. (2007), 'Dempster - Shafer Theory in Geographic Information Systems: A survey', *Expert Systems with Applications* **32**(1), 47 – 55.
- Martignon, L. & Wassner, C. (2002), Teaching Decision Making and Statistical Thinking with Natural Frequencies, in B. Phillips, ed., 'Developing a Statistically Literate Society'.
- Morrison, J. L. (1974), 'A Theoretical Framework for Cartographic Generalization with the Emphasis on the Process of Symbolization', *International Yearbook of Cartography* **14**, 115–127.

- Morss, R. E., Demuth, J. L. & Lazo, J. K. (2008), 'Communicating Uncertainty in Weather Forecasts: A Survey of the U.S. Public', *Weather and Forecasting* **23**(5), 974–991.
URL: <https://doi.org/10.1175/2008WAF2007088.1>
- Nabwey, H. A. (2011), 'A Probabilistic Rough Set Approach to Rule Discovery', *International Journal of Advanced Science and Technology* **30**.
- Neis, P., Zielstra, D. & Zipf, A. (2012), 'The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007-2011', *Future Internet* **4**, 1–21.
- Newman, L. (2016), Descartes' Epistemology, in E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2016 edn, Metaphysics Research Lab, Stanford University.
- Nielsen, J. & Molich, R. (1990), Heuristic Evaluation of User Interfaces, in 'Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, pp. 249–256.
- Ning, W., Goodman-Deane, J. & Clarkson, P. J. (2019), Addressing Cognitive Challenges in Design—A Review on Existing Approaches, in 'Proceedings of the Design Society: International Conference on Engineering Design', Vol. 1, Cambridge University Press, pp. 2775–2784.
- Ortega, P. A. (2010), 'Logic, Reasoning under Uncertainty and Causality', *Neural Information Processing Systems math.ST*, 13.
- Osseiran, A. (2001), 'Qualitative Bayesian Networks', *Information Sciences* **131**(1), 87 – 106.
- Ottley, A., Peck, E. M., Harrison, L. T., Afergan, D., Ziemkiewicz, C., Taylor, H. A., Han, P. K. & Chang, R. (2016), 'Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability', *IEEE Transactions on Visualization and Computer Graphics* **22**(1), 529–538.
- Pang, A. T., Wittenbrink, C. M. & Lodha, S. K. (1997), 'Approaches to Uncertainty Visualization', *The Visual Computer* **13**(8), 370–390.
- Pawlak, Z. (1982), 'Rough Sets', *International Journal of Computer & Information Sciences* **11**(5), 341–356.
- Pearl, J. (1985), Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning, in 'Proceedings of Cognitive Science Society (CSS-7)'.
- Peng, L., Niu, R., Huang, B., Wu, X., Zhao, Y. & Ye, R. (2014), 'Landslide Susceptibility Mapping Based on Rough Set Theory and Support Vector Machines: A case of the Three Gorges Area, China', *Geomorphology* **204**, 287–301.

- Pietro, L. D., Mugion, R. G., Musella, F., Renzi, M. F. & Vicard, P. (2017), 'Monitoring an Airport Check-in Process by Using Bayesian Networks', *Transportation Research Part A: Policy and Practice* **106**(Supplement C), 235 – 247.
- Plato (380 B.C.), 'Republic'.
- Pourret, O., Naïm, P. & Marcot, B. (2008), *Bayesian Networks: A Practical Guide to Applications*, Statistics in Practice, Wiley.
- Rhodes, P. J., Laramée, R. S., Bergeron, R. D. & Sparr, T. M. (2003), 'Uncertainty Visualization Methods in Isosurface Rendering', *Eurographics 2003* pp. 83–88.
- Ribeiro, A. & Fonte, C. (2015), 'A Methodology for Assessing Openstreetmap Degree of Coverage for Purposes of Land Cover Mapping', *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* **II-3/W5**, 297–303.
- Rodgers, P. (2014), 'A Survey of Euler Diagrams', *J. Vis. Lang. Comput.* **25**, 134–155.
- Scutari, M. (2009), 'Learning Bayesian Networks with the bnlearn R Package', **VV**(li).
- Shafir, E., Simonson, I. & Tversky, A. (1993), 'Reason-Based Choice', *Cognition* **49**(1), 11 – 36.
- Sikder, I. U. (2016), 'A Variable Precision Rough Set Approach to Knowledge Discovery in Land Cover Classification', *International Journal of Digital Earth* **9**(12), 1206–1223.
- Smith Mason, J., Retchless, D. & Klippel, A. (2016), 'Domains of Uncertainty Visualization Research: a Visual Summary Approach', *Cartography and Geographic Information Science* **0406**(May), 1–14.
- Spellerberg, I. (2008), Shannon - Wiener Index, in S. E. Jørgensen & B. D. Fath, eds, 'Encyclopedia of Ecology', Academic Press, Oxford, pp. 3249 – 3252.
- Squicciarini, A. C., Dupont, J. & Chen, R. (2014), Online Abusive Users Analytics Through Visualization, in 'Proceedings of the 23rd International Conference on World Wide Web', WWW '14 Companion, ACM, New York, NY, USA, pp. 155–158.
- Stanislawski, L. V., Dewlitt, B. A. & Shrestha, R. L. (1996), 'Estimating Positional Accuracy of Data Layers within a GIS through Error Propagation', *Photogrammetric Engineering & Remote Sensing* **62**(4), 429–433.
- Stassopoulou, a., Petrou, M. & Kittler, J. (1998), 'Application of a Bayesian Network in a GIS-based Decision Making System', *International Journal of Geographical Information Science* **12**(1), 23–46.

- Strasser, C. & Antonelli, G. A. (2018), Non-Monotonic Logic, in E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', summer 2018 edn, Metaphysics Research Lab, Stanford University.
- Straub, D. (2001), 'Natural Hazards Risk Assessment Using Bayesian Networks'.
- Sucharita, G. (2016), *Artificial Neural Networks in Geospatial Analysis*, American Cancer Society, pp. 1–7.
- Sun, B., Chen, X., Zhou, Q., Tong, K., Kong, H. & Asia, C. (2016), 'Uncertainty Assessment of Globeland30 Land Cover Data Set Over', *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLI-B8**, 2–6.
- Taalab, K., Corstanje, R., Zawadzka, J., Mayr, T., Whelan, M., Hannam, J. & Creamer, R. (2015), 'On the Application of Bayesian Networks in Digital Soil Mapping', *Geoderma* **259-260**(October), 134–148.
- Talbott, W. (2016), Bayesian Epistemology, in E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2016 edn, Metaphysics Research Lab, Stanford University.
- Tannert, C., Elvers, H.-D. & Jandrig, B. (2007), 'The Ethics of Uncertainty', *EMBO reports* **8**(10), 892–896.
- Thomas, J. J. & Cook, K. A. (2005), *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, National Visualization and Analytics Ctr.
- Thomson, J., Hetzler, E., Maceachren, A. & Gahegan, M. (2005), 'A Typology for Visualizing Uncertainty', *Visualization and Data Analysis* **5669**(January), 146–157.
- Tversky, A. & Kahneman, D. (1974), 'Judgment under Uncertainty: Heuristics and Biases', *Science* **185**(4157), 1124–1131.
- Unwin, D. J. (1995), 'Geographical Information Systems and the Problem of "Error and Uncertainty"', *Progress in Human Geography* **19**(4), 549–558.
- Upmanyu, M., Namboodiri, A. M., Srinathan, K. & Jawahar, C. V. (2009), Efficient privacy preserving video surveillance, in '2009 IEEE 12th International Conference on Computer Vision', pp. 1639–1646.
- Uzgalis, W. (2018), John Locke, in E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', summer 2018 edn, Metaphysics Research Lab, Stanford University.
- Veregin, H. (1999), 'Data quality parameters', *Geographical Information Systems* **1**, 177–189.

- Villa, V. & Cozzani, V. (2016), 'Application of Bayesian Networks to Quantitative Assessment of Safety Barriers Performance in the Prevention of Major Accidents', *Chemical Engineering Transactions* **53**, 151–156.
- Walczak, B. & Massart, D. (1999), 'Rough Sets Theory', *Chemometrics and Intelligent Laboratory Systems* **47**(1), 1 – 16.
- Wellmann, J. F. & Regenauer-Lieb, K. (2012), 'Uncertainties Have a Meaning: Information Entropy as a Quality Measure for 3-D Geological Models', *Tectonophysics* **526-529**(Supplement C), 207 – 216. Modelling in Geosciences.
- Wittenbrink, C. M., Pang, A. T. & Lodha, S. K. (1996), 'Glyphs for Visualizing Uncertainty in Vector Fields', *IEEE Transactions on Visualization and Computer Graphics* **2**(3), 266–279.
- Woodberry, O. & Mascaro, S. (2012), *Programming Bayesian Network Solutions with Netica*, Bayesian Intelligence.
- Zadeh, L. (1965), 'Fuzzy Sets', *Information and Control* **8**(3), 338 – 353.
- Zapata-Rivera, J. D., Neufeld, E. & Greer, J. E. (1999), Visualization of Bayesian Belief Networks, in 'IEEE Visualization 1999 Late Breaking Hot Topics Proceedings', Press, pp. 85–88.
- Zuk, T. & Carpendale, S. (2007), 'Visualization of Uncertainty and Reasoning', *International Symposium on Smart Graphics* pp. 164–177.