COLLABORATIVE GEOVISUAL ANALYTICS

Gustavo Adolfo García Chapeton

COLLABORATIVE GEOVISUAL ANALYTICS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. ir. A. Veldkamp,
on account of the decision of the Doctorate Board,
to be publicly defended
on Thursday 7 July 2022 at 12.45 hours

by

Gustavo Adolfo García Chapeton
born on the 31st of March, 1985
in Quetzaltenango, Guatemala

This dissertation has been approved by:

Supervisors:
prof. dr. M.J. Kraak
dr. ir. R.A. de By

Co-supervisor:
dr. F.O. Ostermann

ITC dissertation number 416
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN:	978-90-365-5367-4
DOI:	10.3990/1.9789036553674

Cover designed by Job Duim
Printed by QTRL-P Hengelo

ITC
UNIVERSITY OF TWENTE.
FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

Graduation Committee:

Chair / secretary:	prof. dr. F.D. van der Meer
Supervisors:	prof. dr. M.J. Kraak University of Twente, ITC, Department of Geo-information Processing
	dr. ir. R.A. de By University of Twente, ITC, Department of Geo-information Processing
Co-supervisor:	dr. F.O. Ostermann University of Twente, ITC, Department of Geo-information Processing
Committee Members:	prof. dr. R. Zurita Milla University of Twente, ITC, Department of Geo-information Processing
	prof. A.D. Nelson University of Twente, ITC, Department of Natural Resources
	prof. dr. F.J. Villalobos University of Cordoba & IAS-CSIC
	prof. dr. N. Andrienko Fraunhofer IAIS Bonn / City University London

Acknowledgments

Siting here, writing these lines, I am thinking about the many individuals that in one or another way were part of my long journey as a PhD candidate, there are no words to thank them enough.

First, I would like to express my gratitude to Dr. Frank Ostermann, Dr. Rolf de By, and Prof. Menno-Jan Kraak. I greatly appreciate your support throughout my research. It was not always easy for me to advanced my research, but your trust and encouragement helped me to reach the end of this chapter in my academic life.

I would like to recognize and thank the important contributions of many persons in Málaga, Spain. I am especially grateful with Jesús Olivero, Emilio García, Antonio Sánchez, and Gonzalo Reina. It would not be possible to conduct my research without the data, knowledge and expertise that you provided. Thanks for all your kindness and support during my visits to Málaga.

I would also like to thank all the ITC staff that were always available to provide support in diverse matters. I am especially thankful with Lyande Eelderink and Javier Morales, for their friendship and support in personal and work related matters. An special thanks to Thomas Groen and Bert Toxopeus for their contributions to my thesis, and the great time we spent in Guatemala working together in a Tailor-made Training. I am also thankful with Jolanda Kuipers and Tonny Boeve, for their support with administrative procedures.

As the saying goes “Friends are the family you choose.” There are so many friends that were part of this journey, naming all of them is impossible. My life in Enschede (during the MSc and PhD studies) was truly enjoyable thanks to many people. Special thanks to Tatjana, Eduardo, Andrés, Parya, Manuel, Valentina, Magnus, Leila, Alby, Sheila, Abhishek, Yolla, Irene, Emma, Vero, Andre, Ana, Alfredo, Liliana, Diana, Hamed, Siddhi, Riddhi, Sana, Eva, Rosa, and Marleny. Thanks for filling my life with incredible memories. I would also like to thank my dear friends in Guatemala. Special thanks to Adita, Claudia, Coca, Maridalia, Dhaby, Margarita, Cristy, Héctor, Javier, Ana, Jesús, Alberto García (RIP), Karla, Gaby, María, and José. Thanks for keeping in touch when I was away, and for your support in the last years of this journey when I was already back in Guatemala.

Lastly, but most important to me, I would like to thank my family. To my always loving and supporting parents Ramiro and María, I owe you so much that two lives will not be enough to pay you back. Thanks for always encouraging me to pursue my dreams, every success in my life is in great part yours. To my brother and sister, Carlos and Cyndi, thanks for your unconditional support, I always have you in my mind and heart.

Gustavo Adolfo García Chapeton
Quetzaltenango, Guatemala

Contents

List of Figures

List of Tables

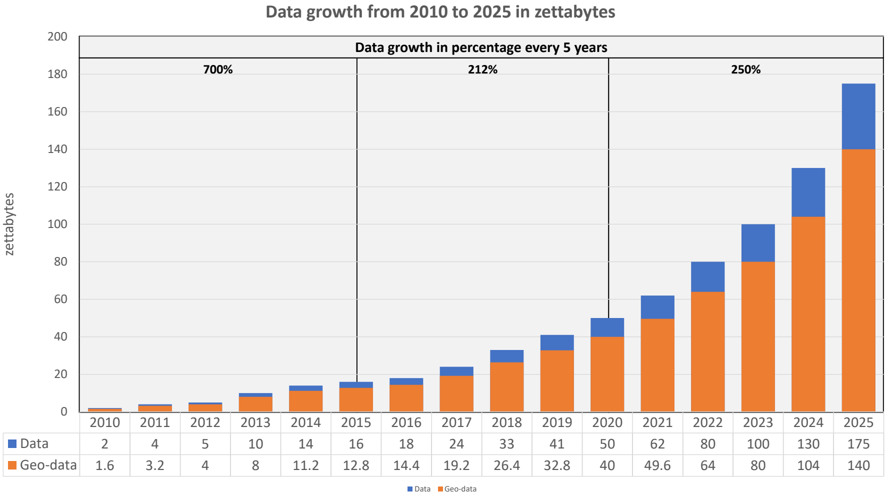
1 The need for collaborative geovisual analytics

In the last two decades, advancements in Information and Communication Technologies (commonly referred to as ICT) such as database technologies, fast Internet connections, miniaturization of sensors, development of mobile and wearable devices, and Internet of Things (IoT), in combination with several ground-breaking advances in geospatial technologies such as small GPS-enabled devices, high-resolution remote sensors, and linking of geoweb services, have led to an unprecedented abundance of geodata [24, 22, 149]. In a recent report, the International Data Corporation (IDC) estimated that the global data will grow from 33 ZB[1] in 2018 to 175 ZB by 2025 [127]. In this regard, although hard to measure, it is broadly claimed that about 80% of all data includes some type of geographic reference. For this reason, geoinformation science is a relevant research domain, with expanding potential applications as geodata for diverse disciplines becomes available. Figure 1.1 illustrates this data growth.

This boom of data availability is about the amount of data, the sources and types of data, and the speed of data production. This phenomenon of large, rapidly growing and heterogeneous data sets is frequently called Big Data. One of the many available definitions is "data sets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time" [22, p. 173]. It is commonly characterized by its volume (ever-increasing data sets), variety (from a diversity of sources and in varied formats), and velocity (produced at an increasingly fast pace) [174, 87], the so-called three V's of Big Data. As the Big Data concept evolved, new V's emerged, one is of particular relevance: value. It emphasizes that the greatest challenge of Big Data is to extract the information contained in those data sets, and at the same time that the value of such information is worth the hassle [22].

This abundance of big geodata presents a unique opportunity to increase our knowledge and understanding of natural and artificial processes. However, it also presents a challenge for analysts who need to make sense of increasingly large, heterogeneous, and multivariate geodata

[1] A zettabyte (ZB) is 1,000,000,000,000,000,000,000 bytes.



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
■ Data	2	4	5	10	14	16	18	24	33	41	50	62	80	100	130	175
■ Geo-data	1.6	3.2	4	8	11.2	12.8	14.4	19.2	26.4	32.8	40	49.6	64	80	104	140

Figure 1.1 Data growth from 2010 to 2025 in zettabytes. The values for data amount were estimated from a figure presented in [127], and the values for geodata were calculated as 80% of those values.

sets [127]. In this context, two problems emerge: first, the limited capacity of humans to work with large amounts of data, therefore requiring support from computers to transform the data into manageable representations [156]; second, the complexity of some data sets that renders the analysis by a single person infeasible, thus requiring a collaborative data analysis approach [66, 56]. Geovisual Analytics (GVA) aims to address both problems. To date, the research in this field has focused mainly on developing data transformation algorithms, visualization and interaction methods. However, limited attention has been given to the support for collaborative analysis [97, 66].

Regarding the need to support collaborative work in GVA, Thomas and Cook argue that while computer capacity keeps improving at an ever-increasing pace, the human analytic capacity is somehow fixed [156]. For this reason, they argue that “We must develop techniques that gracefully scale from a single user to a collaborative (multi-user) environment” [156, p. 27]. Additionally, they envision that such environments would support collaboration within and between organizations, “we envision, users may be collaborating from within the same team in an organization, at different levels of an organization, or even in different organizations” [156, p. 27]. The support for collaboration is crucial because problem-solving often requires the combination of input from persons with broad and varied expertise, and diverse perspectives [66].

1.1 Problem identification

As argued above, the abundance of geodata presents a unique opportunity to increase our knowledge and understanding of natural and artificial processes. However, it also presents a challenge for analysts who need to make sense of increasingly large, heterogeneous, and multivariate geodata sets. Analyzing such data sets is a complex task that benefits from combining human and machine analysis capabilities. Computers are capable of storing and processing large data sets and can help to identify patterns, trends, and outliers. However, unless guided by theory and domain knowledge, such 'blind' data mining is likely to produce spurious correlations and meaningless results [141, 79, 2, 156]. Therefore, human skills are needed to formulate a guiding hypothesis, parameterize algorithms, select evidence, validate results, draw conclusions, and ultimately make decisions. GVA enables and exploits this combined intelligence.

Many analytical problems are complex, ill-defined, and broad in scope [2, 66, 156]. In this regard, researchers from diverse domains agree that such analytical problems will benefit from approaches and tools that support reproducible, multidisciplinary collaborative work [46, 56, 58, 66, 116, 170]. In the domain of Visual Analytics (VA), from which GVA is a sub-field, the claim to support collaborative analysis dates back to the very definition of the research field in 2005:

> "Analytical reasoning must be a richly collaborative process and must adhere to principles and models for collaboration. Collaborative analysis provides both the human and computational scalability necessary to support reasoning, assessment, and action." [156, p. 33]

Further, the authors identify theoretical foundations to advance the support for collaborative analysis in VA:

> "Build upon theoretical foundations of reasoning, sense-making, cognition, and perception to create visually enabled tools to support collaborative analytic reasoning about complex and dynamic problems." [156, p. 63]

In this regard, Mathisen [97] analyzed the open visualization publications data set [67] showing that the topic of collaboration only received attention for a few years after the definition of the research field in 2005, and since then, it has declined. Although the data set only includes the publications of the two leading conferences in VA research, InfoVis and VAST, as the author claims, it illustrates the limited focus on collaboration. Currently, many GVA systems are single-user environments or offer limited support for collaborative work, and in consequence, collaboration remains a challenge for GVA research [20, 33, 54, 66, 79, 2, 156].

The lack of support for collaborative work in a GVA system limits the analysis to the input by a single analyst, which may lead to biased results due to the limited knowledge and expertise utilized during the analysis process. This research aims to enable collaborative analysis of geographic phenomena by analysts with diverse knowledge and expertise, which requires analyzing large or ever-increasing data sets. The selected application domain is pest management, which is important due to the significant adverse effects of pests on the environment, economy, and human health. This application domain was selected based on the following rationale: first, the spatiotemporal distribution of pests is heterogeneous because of variations in topographic, environmental, and weather conditions and human intervention of ecosystems. Therefore, pest management needs to be addressed as a geographic problem. Second, pest monitoring and control efforts require continuous data collection, which produces ever-increasing data sets, and offer the possibility of better understanding the pest population dynamics and designing pest management strategies. Third, making sense of those data sets requires the input of stakeholders from diverse backgrounds such as farmers, pest management experts, and scientists.

The specific application case is the monitoring and control of the Olive Fruit Fly (Bactrocera oleae, OFF in the sequel) in Southern Spain. The selection of this case is due to the worldwide importance of the Mediterranean region in olive production and Spain's leading role as an olive producer. Additionally, the pest management effort is well-organized, including the active collection of monitoring and control data. There is also plenty of publicly available data relevant to the analysis of species populations, such as topographic, environmental, and weather data. Finally, it involves a variety of stakeholders such as authorities, field technicians, landowners, and researchers, which are actively collaborating to improve the OFF management.

1.2 Key concepts

1.2.1 Collaborative data analysis

Collaborative data analysis means a joint effort of two or more analysts to achieve shared or intersecting goals regarding transforming data into actionable information. Collaboration is beneficial to solve complicated analysis tasks [69]; it allows to combine diverse expertise and perspectives [66], enabling a broader and deeper analysis of data [56], which produces better quality analytic results that support efficient decision-making and planning processes [66].

Collaborative data analysis can be described as a distributed cognition process. Two theoretical principles characterize Distributed Cognition (DC): first, the boundary of the analysis unit for cognition is not limited to an individual; instead "distributed cognition looks for cognitive processes, wherever they may occur, based on the functional relationships of elements that participate together in the process" [61, p. 175], therefore, a cognitive system may include several analysts working together to solve analysis problems. Second, the range of mechanisms that may participate in cognitive processes. Hollan, Hutchins and Kirsh [61] state that “traditional views look for cognitive events in the manipulation of symbols inside individual actors, distributed cognition looks for a broader class of cognitive events and does not expect all such events to be encompassed by the skin or skull of an individual” (p. 175-176). Therefore the interactions between analysts (i.e., social interaction) and of them with (physical and/or virtual) objects play a relevant role in the cognitive system. In brief, DC provides a framework in which cognition is conceived as a social process, involving human actors as thinking entities and artifacts as means for knowledge exchange and shared memory [152]. In this research work, those artifacts refer to the data and its multiple representations such as charts and maps arranged into interactive visual interfaces and the approaches that enable collaboration over them, such as discussion tools (e.g., chat and discussion forum).

To effectively support collaborative data analysis in analytic environments, it is necessary to understand how collaboration works and how to design, develop, use, and evaluate the analytic environments to support it. The multidisciplinary research field called Computer-Supported Cooperative Work (CSCW) addresses those concerns. There are many definitions for CSCW. Koch, Schwabe and Briggs [84] state that those definitions converge to a similar concept captured in the definition by Bowers and Benford "in its most general form, CSCW examines the possibilities and effects of technological support for humans involved in collaborative group communication and work processes" [15, p. 5]. Given that this research is concerned with collaborative work in GVA environments, the CSCW literature is of key importance.

1.2.2 Geovisual Analytics

VA emerged from the need to analyze increasingly big, conflicting, and dynamic data sets [156]. Wong and Thomas introduced the concept of VA as “the formation of visual abstract metaphors in combination with a human information discourse (interaction) that enables detection of the expected and discovery of the unexpected within massive, dynamically changing information spaces” [167, p. 20]. Later, Thomas and Cook further discussed and defined it as “the science of analytical reasoning assisted by interactive visual interfaces” [156, p. 4]. More recently, Keim et al. defined it as “visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets” [79, p. 7]. All these definitions convey a common goal of VA, to “maximize human capacity to perceive, understand, and reason about complex and dynamic data and situation” [156, p. 6].

VA research and application are multi-disciplinary and integrate knowledge from diverse domains, such as information and scientific visualization, interaction techniques, data management, statistics, data mining, and machine learning [59, 79]. VA strongly relies on visual interfaces to connect the human and computer parts of the analytic system. For this reason, VA is often confused with information visualization. While there is certainly an overlap between both, the difference is that information visualization focuses on developing techniques to visualize data effectively, and VA concerns itself with enhancing the analysis processes by combining the strengths of humans and computers through visual interfaces. Therefore, while information visualization plays a crucial role in VA, they are fundamentally different research and application domains.

GVA can be described as a sub-field of VA that deals with the specific issues related to the analysis of geographic phenomena [2, 59]. It enables analytical reasoning and decision-making of geographic phenomena by producing a synergy of the human analytical skills, with computer's storage and processing power, coupled through interactive geovisual interfaces [2, 158]. GVA integrates knowledge from diverse fields, such as VA, geographic information science, geovisualization, and perception and cognition [59, 158]. Given that phenomena in geographic space occur or evolve in time, GVA has emphasized the relationship between space and time [3].

GVA has a broad field of applications that ranges from counter-terrorism and disaster management to strategic business decision making [59]. Such applications usually involve multiple stakeholders with a diversity of interests, knowledge, and skills. For this reason, GVA pays special attention to the issues of collaboration, communication, and flexibility [59, 2]. Examples of such applications are the analysis of criminal activity [154, 136], human mobility [173], and road accident accumulation zones [125].

1.2.3 Pest Management

A *pest* can be defined as an organism that conflicts with human welfare because it may affect crops, stored products, animals, or people [123]. It is important to emphasize that an organism is considered a pest only when its abundance reaches a level that seriously affects human welfare [57]. Pest outbreak events can threaten local flora, and fauna [117], and especially some agricultural pests can cause damages with significant economic impact for producers and the food supply chain, possibly threatening food security [36]. Additionally, pests threaten human health when they serve as infection vectors for diseases or create allergies [166]. Despite such adverse effects, these species are part of a natural ecosystem. For this reason, proper pest management strategies are needed to minimize their adverse effects without disrupting that natural ecosystem [43].

Pest management includes three stages: first, *the monitoring stage*, in which the species' presence or abundance is measured at several locations and with a set temporal frequency. This sampling aims to represent the population dynamics of the pest in the area of interest. Second, *the control stage*, where countermeasures are taken to keep the species population within acceptable geographic areas or abundance levels. Third, *the evaluation stage*, in which the effectiveness of the monitoring and control actions is assessed [57]. The Food and Agriculture Organization of the United Nations (FAO) defines Integrated Pest Management (IPM) as:

> "The careful consideration of all available pest control techniques and subsequent integration of appropriate measures that discourage the development of pest populations and keep pesticides and other interventions to levels that are economically justified and reduce or minimize risks to human health and the environment. IPM emphasizes the growth of a healthy crop with the least possible disruption to agro-ecosystems and encourages natural pest control mechanisms [38]"

However, agronomic production sees high levels of chemical treatments, and the lack of effective methods to determine when and where to apply a treatment can lead producers to incur unnecessary expenses. Further, overuse of pesticides can cause problems such as reduction of the effectiveness of the treatment due to increased resistance of the pest, reduction of biodiversity by affecting other species, and accumulation of chemical residues in crops, soil, and water bodies [57]. For this reason, it is of key importance to develop methods and tools that aid stakeholders in a better understanding of pest dynamics and support decision-making in pest management.

1.3 Research objectives

This research aims to address the long-standing challenge of supporting collaborative analysis in GVA systems. It acknowledges that due to the ever-increasing availability of geodata and the complexity of analytical problems, the need to enable collaborative work among analysts from diverse backgrounds (e.g., domain experts, data analysts, scientists, and laypersons) is becoming more pressing and prominent. To address this objective, the research was guided by the following specific objectives and research questions:

1. Review the state-of-the-art of collaborative geovisual analytics, and propose a research agenda
   a) What are the characteristics of geovisual analytics systems that support collaborative work?
   b) Which are the collaboration techniques available in geovisual analytics systems?
   c) What are the research challenges to effectively support collaborative work in geovisual analytics systems?
2. Design a software reference architecture for collaborative geovisual analytics systems
   a) Which architectural patterns provide a viable starting point to design an architecture for collaborative geovisual analytics systems?

  b) What are the implications of the research challenges to effectively support collaborative work in geovisual analytics systems in the design of the architecture?
  c) What should be the components of the architecture?

3. Design an approach for collaborative analysis in geovisual analytics systems
   a) How to make the approach flexible enough to support collaborative analysis in diverse application domains?
   b) How to accommodate the approach into the software reference architecture?
4. Implement the software reference architecture and the collaborative analysis approach in a prototype for the monitoring and control of the Olive Fruit Fly and evaluate its usability and utility
   a) Does the software reference architecture allow to accommodate the stakeholders' requirements in the prototype?
   b) Which software technologies are suitable to develop the prototype?
   c) Does the prototype enable stakeholders from diverse backgrounds to participate in the analysis of the pest dynamics?

1.4 Research methodology

Four main activities compose this research work: 1) literature review, design of 2) a software reference architecture and 3) a collaborative analysis approach, and 4) design, development, and evaluation of a proof-of-concept prototype for the architecture and the collaboration approach. Each of these activities addresses one of the specific research objectives stated in the previous section. Figure 1.2 shows the research workflow, which also shows how the activities relate to the thesis chapters. The activities are shown in a logical sequence but not a strict chronological sequence. The activities are not in strict chronological order because the execution of some activities provided feedback to improve the results of previous ones. For example, the experience designing and developing the prototype provided feedback to improve the collaborative analysis approach.

The first activity was a literature review to describe the state-of-the-art of collaborative analysis in GVA systems. The execution of this activity followed the guidelines for systematic literature review proposed by Kitchenham and Charters [83]. The reported results describe Collaborative GVA (CGVA) systems, collaboration techniques, and research challenges. This knowledge was a valuable input to design the software reference architecture, the collaborative analysis approach and the prototype.

The software reference architecture aims to provide a generic model for the design and development of GVA systems with features to support

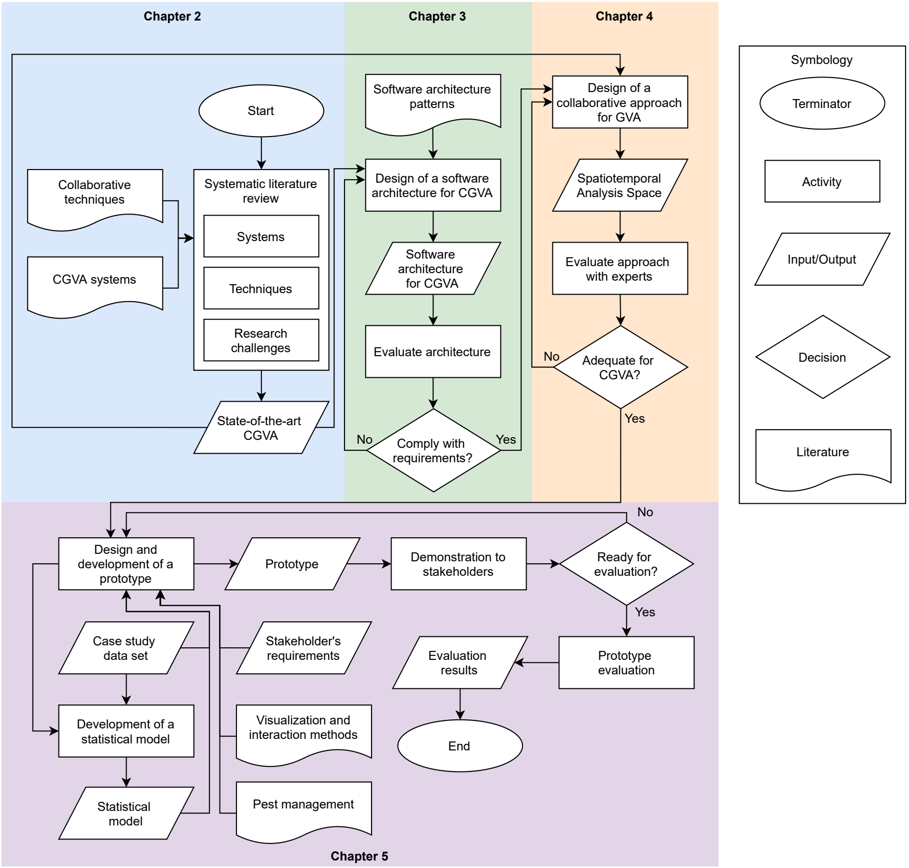

Chapter 2
Chapter 3
Chapter 4
Symbology
Terminator
Activity
Input/Output
Decision
Literature
Software architecture patterns
Design of a collaborative approach for GVA
Start
Collaborative techniques
Systematic literature review
Design of a software architecture for CGVA
Spatiotemporal Analysis Space
Systems
CGVA systems
Techniques
Software architecture for CGVA
Evaluate approach with experts
Research challenges
Evaluate architecture
No
Adequate for CGVA?
Yes
State-of-the-art CGVA
No
Comply with requirements?
Yes
No
Design and development of a prototype
Prototype
Demonstration to stakeholders
Ready for evaluation?
Yes
Case study data set
Stakeholder's requirements
Evaluation results
Prototype evaluation
Development of a statistical model
Visualization and interaction methods
End
Statistical model
Pest management
Chapter 5

Figure 1.2 Research workflow

collaborative analysis. Its design was guided by the analysis of the identified research challenges, and it is based on proven software architectural patterns. The design process was iterative, and on each iteration, the design at the moment was assessed to determine whether it complied with the design criteria or if adjustments were needed.

The collaborative analysis approach was designed based on the identified research challenges and the characteristics of the analysis process required to monitor and control pests. Those characteristics are: 1) support for long-term analysis of continuously growing geodata sets; 2) by stakeholders from varied backgrounds; and 3) contributing asynchronously. The result of this activity is the collaboration approach called Spatiotemporal Analysis Space, which can be described as an approach for long-term distributed asynchronous collaborative analysis in GVA environments. Additionally, the approach was mapped to the software reference architecture to provide an example of distribution for the approach functionality into software components.

Throughout the research process, the case study served as a reference to assess whether the results of the diverse activities are applicable to a real-world scenario. In this context, a web-based prototype was designed, developed and evaluated, as a proof-of-concept for the software reference architecture and the collaboration analysis approach. These activities were conducted with the support of seven stakeholders of the OFF management in Southern Spain.

1.5 Thesis outline

This thesis is composed of six chapters. The following paragraphs summarize the contents of each chapter:

Chapter 1 identifies the need to conduct research to advance the support for collaborative analysis in GVA systems. It describes the research objectives and questions that guided this work and the methodology to address those. Additionally, it identifies the analysis of pest populations dynamics as an application domain that can benefit from GVA and collaborative analysis, and briefly describes the case study for this research, which is the monitoring and control of the OFF in Southern Spain.

Chapter 2 addresses the first specific objective: "Review the state-of-the-art of collaborative geovisual analytics, and propose a research agenda." This objective was tackled with a systematic literature review focused on identifying and describing CGVA systems, collaboration techniques, and research challenges. The results of the review, especially the research challenges, had a direct impact on the design of the software architecture and the collaboration approach, presented in Chapters 3 and 4, respectively.

Chapter 3 addresses the second specific objective: "Design a software reference architecture for collaborative geovisual analytics systems." This chapter starts with a description of software architecture and the different software architectural patterns used to design the proposed architecture. Later, it describes the design criteria, which are based on the research challenges identified in Chapter 2. Finally, it describes the software reference architecture.

Chapter 4 addresses the third specific objective: "Design an approach for collaborative analysis in geovisual analytics systems." This chapter describes the design of the Spatiotemporal Analysis Space approach, which can be described as an approach for long-term distributed asynchronous collaborative analysis in GVA environments. Additionally, the chapter describes a mapping of the approach into the software reference architecture proposed in Chapter 3 and the implementation of the approach in the context of the application case.

Chapter 5 addresses the fourth specific objective: "Implement the software reference architecture and collaborative analysis approach in a prototype for the monitoring and control of the Olive Fruit Fly and evaluate its usability and utility." This chapter describes the design and development of a CGVA prototype, composed of a processing application (which implements a statistical model developed for the case study) and an interactive visual interface based on case-specific user requirements. The prototype is based on the software architecture and implements the collaborative analysis approach proposed in Chapters 3 and 4. Additionally, this chapter presents the results of a user evaluation conducted with the participation of the stakeholders of the case study.

Chapter 6 synthesizes and discusses the activities and results presented from Chapters 2 to 5, draws conclusions for the thesis, elaborates on the contributions of this research and proposes future research directions.

2 Collaborative geovisual analytics: state-of-the-art

The content of this chapter is published on: *García-Chapeton, G.A., Ostermann, F.O., de By, R.A., Kraak, M-J. Enabling Collaborative GeoVisual Analytics: Systems, techniques, and research challenges. Transactions in GIS. 2018; 22: 640– 663.* `https://doi.org/10.1111/tgis.12344`

Chapter 1 recognizes collaboration as one of the grand challenges for research in GVA. Considering the increasing availability of geodata and the increasing complexity of analytical problems, the need to advance the support for collaborative analysis is becoming more pressing and prominent. This chapter addresses the first specific objective stated in Chapter 1: "Review the state-of-the-art of collaborative geovisual analytics, and propose a research agenda." For this aim, a systematic review was conducted, identifying thirteen collaborative systems, six distinct collaboration techniques, and three research challenges.
This review follows the guidelines for systematic reviews proposed by Kitchenham and Charters [83]. These guidelines were originally designed for the field of software engineering but have been adopted successfully in other domains, such as information visualization [171], spatiotemporal analysis [147], and educational resources [5]. A systematic review has three phases: planning, conducting, and reporting. *The planning phase* (Section 2.1) defines the objective of the review, the process to identify the information sources, and the information to be obtained from them. *The conducting phase* (Section 2.2) includes the acquisition of information sources and the extraction, organization, and synthesis of the information. *The reporting phase* (Section 2.3) prepares a comprehensive document with the review results. Finally, Based on the results, three research challenges and strategies to address these are described.

2.1 Planning the systematic review

The following specific objectives guided this review:

1. To identify GVA systems that support collaborative analysis and describe their characteristics regarding collaboration scenarios and

technological platforms;

2. To identify and describe collaboration techniques implemented in GVA systems;
3. To identify research challenges to effectively support collaborative work in GVA and propose strategies to address these.

2.1.1 Information sources

The information for the literature review was obtained from several well-known electronic databases (as listed in Section 2.2) in the domain of geo-information science. The common search and selection criteria[1] were:

- Search keywords: (collaborative OR cooperative) AND ("geovisual analytics" OR "geospatial visual analytics" OR geoanalytics)
- Inclusion criteria
  - Publication date: between January 2004 and June 2017 inclusive
  - Publication type: journals, proceedings, transactions, and book chapters
  - Article type: full text and reviews
  - Language: English
- Exclusion criteria
  - Duplicated papers (identified using EndNote X8)
  - Non-relevant papers (determined by manual paper screening)

The search keywords were defined based on an investigation of the terms used by authors when referring to GVA systems; commonly used terms are: "GeoVisual Analytics" [2], "GeoSpatial Visual Analytics" [26], and "GeoAnalytics" [71]. Given that the interest is on systems supporting collaborative analysis, the term "Collaborative" was included. Additionally, since electronic collaborative systems are based on the Computer-Supported Cooperative Work (CSCW) principles, the keyword "Cooperative" was also included. The review covers from the introduction of VA as a research field in 2004 [167] until June 2017, when the review started.

The exclusion criteria aimed to retain only the papers that relate to the review objectives. Those papers describe CGVA systems, collaboration techniques, and research challenges to support collaboration in GVA systems effectively. For example, some papers use the term "collaborative" regarding a general collaborative effort or work, not a collaborative system, hence those papers were discarded.

The multidisciplinary nature of GVA research complicates a fully comprehensive literature review. However, the search approach aims to be sufficiently exhaustive to include the most relevant information. The

[1] See Annex A for details on the search and selection process on each database.

focus is on the collaborative capacity of the systems based on the supported collaboration scenarios, technological platforms, and collaboration techniques. Other perspectives might be adopted, and given that GVA is a fast-evolving field, a similar study will yield different search results in the future. However, the extracted research challenges will persist and thus ensure that these findings remain relevant for a significant period. To support comparison with future studies, a theoretical framework on information extraction and organization was adopted as described in Section 2.1.2.

2.1.2 Extraction and organization of information

The Knowledge Generation Model for VA (KGM-VA) proposed by Sacha et al. [137] is used as the theoretical framework to structure and organize this review. This model explicitly separates the human and computer components to highlight their role in VA, and it incorporates the notion of the analytical process, as shown in Figure 2.1. The model provides a clear theoretical separation and order for the stages in the analysis process. However, the stages may overlap in actual analysis processes, and analysts move back and forth in a dynamic knowledge generation process [137].

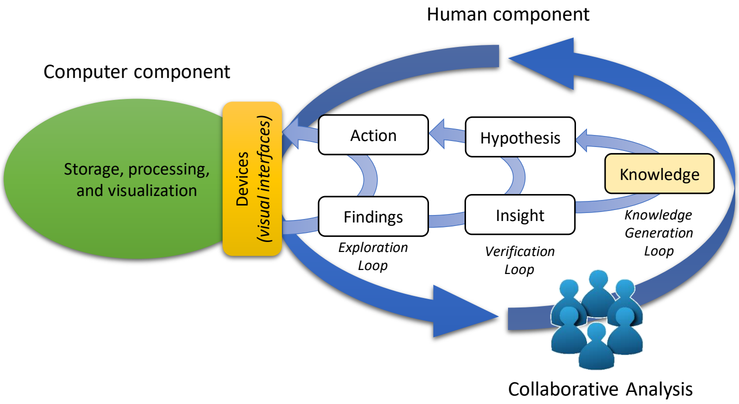


Figure 2.1 The Knowledge Generation Model for Visual Analytics explicitly separates the systems into human and computer components, and conceptualize the analysis process with three loops: exploration, verification, and knowledge generation. Illustration based on [137]

The KGM-VA models the analysis process with three stages termed loops: exploration, verification, and knowledge generation. These loops occur in the human component, while the computer component provides storage and processing power to support them. *The exploration loop* represents the interactions of analysts with the system, which produce findings; these are relevant observations about the phenomenon under study. The analysts' actions are guided by an analytical goal, or in its absence, to define one. *The verification loop* guides the exploration loop to confirm hypotheses or to generate new ones. In this loop, the analysts gain insights as the findings are interpreted in the context of the analysis domain and may contribute to verify or falsify a hypothesis. Finally, in *the knowledge generation loop*, the analysts combine their expertise with the identified evidence to accept or reject a hypothesis and generate new knowledge or to suggest further analysis if the evidence is not conclusive.

For this review, the KGM-VA provides an effective framework to analyze and describe the human and computer components and their interaction (the system level), and the role of a technique in the analysis process for enabling collaboration among participants (the technique level).

At the system level, the aim is to describe the organization of the human component (i.e., the analysts) to address the collaborative effort, the provision of computer storage and processing power to support the collaborative effort, and how they are linked.

A commonly used approach to characterize the organization of participants in a collaborative effort is to consider the time and space in which participation takes place, usually distinguishing four collaborative scenarios: synchronous co-located (same time and space), synchronous distributed (same time and different space), asynchronous co-located (different time and same space), and asynchronous distributed (different time and space) [75]. These scenarios are not mutually exclusive, and the support for multiple scenarios (or hybrid scenarios) in a system is a desirable characteristic [66] because it allows a more flexible collaboration workflow. Additionally, a collaborative effort can also be characterized based on its duration; in this case, two scenarios can be defined, time-critical (or short-term) and long-term [2, 66, 79, 156]. Having a well-defined set of scenarios allowed to identify the existence of patterns regarding the organization of the participants.

Information and communication technologies (commonly referred to as ICT) enable the collaborative effort. Here, the focus was on two defining characteristics: first, the provision of storage and processing power, which was addressed by reviewing the deployment options for a system; and second, the supported devices because they link the system's human and machine components.

At the technique level, the interest is to describe the defining characteristics of the techniques and describe similarity and co-occurrence among these. This information allows us to understand why some techniques are more popular than others and identify patterns regarding their roles in the analysis process and combination with other techniques.

The collaboration techniques constitute the mechanism for the analyst to externalize and communicate findings and insights to other analysts, which occurs within and across the loops of the analysis process, and enables knowledge generation. This process of collaborative knowledge generation is grounded in the theory of DC, which considers cognition as a social process, involving human actors as thinking entities and artifacts as means for knowledge exchange and shared memory [61].

2.2 Conducting the systematic review

2.2.1 Acquisition of information sources

The search for information sources resulted in 124 papers. From the identified papers, 99 were unique, and 28 were selected for the review process. Table 2.1 shows the search and selection results.

Table 2.1 Results of the search and selection process for information sources. The column “Total” accounts for the total of papers identified in a database, the column “Unique” accounts for non-duplicated papers, and the column “Selected” accounts for the papers that contribute to address the review objectives

Source	**URL**	**Total**	**Unique**	**Selected**
ACM digital library	dl.acm.org	19	11	4
GeoBase	www.engineeringvillage.com	3	3	2
IEEEXplore	ieeexplore.ieee.org	10	10	2
Science direct	www.sciencedirect.com	12	12	3
Scopus	www.scopus.com	32	22	7
Springer Link	link.springer.com	39	34	6
Web of Science	apps.webofknowledge.com	9	6	4
Total	—–-	124	99	28

The search was extended to the web using the Google Search Engine[2] to include relevant CGVA systems not featured in academic literature. After some iterative refinement, the search query was: ‘collaboration AND “visual analytics software” -paper -book -conference.’ The search was limited to the last two years (from July 1st, 2015 to June 30th, 2017) to target active projects only. The term “Geo” was not included in the query because despite being used with geodata, some systems might not mention it explicitly enough to appear in the search results. The search produced 56 results[3]. This extended search found three additional systems (i.e., SAP BusinessObjects, Oracle BI Visual Analytics, and SAS Visual Analytics). A potential reason for the absence of those systems in academic literature is that they all are commercial.

2.2.2 Extracted information

The systematic review identified thirteen CGVA systems and six distinct collaboration techniques; summaries are presented in Tables 2.2 and 2.3, respectively. In the following sections, the CGVA systems and collaboration techniques are described based on the criteria outlined in Section 2.1.2.

[2] www.google.com

[3] The total number of matches for the query was 740, but Google Search Engine detected that 56 were the most relevant results, and the others were very similar to those. This result is not reproducible because Google Search results are based on many unknown variables, including individual user search history.

Table 2.2 Summary of collaborative geovisual analytics systems. The “Year” column shows the year of the first reference to a collaborative feature in the system, if known; the “Description” and “Applications” columns are based on a study of the literature referencing the system, and freely available information from the software publisher; the “Collaborative scenario” column follows the widely adopted categorization proposed by Johansen [75], who distinguishes between synchronous and asynchronous collaboration, and co-located and distributed work places

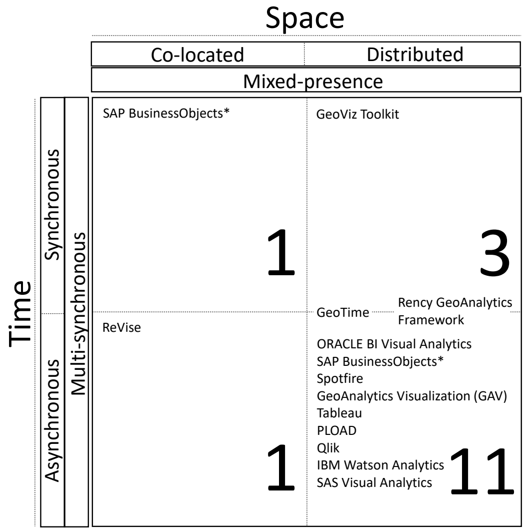
System name	**Year**	**Description**	**Collaborative scenario**	**Applications**	**References**
ORACLE BI Visual Analytics	Unknown	Commercial business intelligence VA system, available as cloud-based service; supports multiple devices	Asynchronous distributed	Business Intelligence	[115]
SAP Business Objects	Unknown	Commercial business intelligence VA system; supports multiple devices, including a special setup for large screens	Synchronous co-located, and Asynchronous distributed	Business Intelligence	[138]
ReVise	2006	Desktop-based prototype based on the Improvise Toolkit; implements the method “Re-Visualization", which allows users to generate and review analysis session logs	Asynchronous co-located	Analysis of U.S. Census data	[134, 158]
GeoTime	2007	Commercial law enforcement GVA system; support multiple devices (desktop, mobile and web-based)	Synchronous distributed, and Asynchronous distributed	Criminal intelligence analysis	[30, 121, 159]
Spotfire	2007	Commercial general-purpose VA system; available as desktop-based, web-based and cloud-based system; supports multiple devices	Asynchronous distributed	General purpose; reported applications in domains such as energy, financial services, manufacturing and telecommunications	[31, 157, 162]

GeoAnalytics Visualization (GAV) Framework	2008	Framework and class library for rapid development of web-based GVA applications	Asynchronous distributed	Exploration and analysis of statistical indicators, analysis of volumetric data, and analysis of flood data, among others	[59, 70, 71, 72, 88]
GeoViz Toolkit	2009	Desktop-based general purpose GVA system	Synchronous distributed	Analysis of health-related data, and of terrorist attack data	[48, 60]
RENCI GeoAnalytics Framework	2011	Cyber-infrastructure for development of web-based GVA systems; supports multiple devices	Synchronous distributed, and Asynchronous distributed	Hurricane damage assessment (CyberEye) and, emergency management and response (Big Board)	[50, 51, 81]
Tableau	2012	Commercial general-purpose VA system; server version offers collaborative features and integration for desktop-based, web-based and mobile versions; successor of Polaris system	Asynchronous distributed	General purpose. Reporting applications in domains such as banking, communications, education, government, insurance, and sports	[32, 60, 153]
PLOAD	2013	Environmental decision support system (EDSS); available as web-based and cloud-based system; focus on the management of watersheds	Asynchronous distributed	Watershed management	[150]
QLik	2013	Commercial general-purpose VA system; available as desktop-based, web-based and cloud-based; supports multiple devices	Asynchronous distributed	General purpose. Reported applications in domains such as health care, financial services, energy and utilities, and life sciences	[122]

IBM Watson Analytics	2014	Commercial general-purpose VA system; cloud-based	Asynchronous distributed	General purpose. Reported applications in domains such as banking, insurance, retail, telecommunication and education	[65, 102]
SAS Visual Analytics	2015	Commercial general purpose VA system; offers on-premises and cloud-based deployments; supports multiple devices	Asynchronous distributed	General purpose. Reported applications in domains such as banking, communications, defense and security, health care, and high-tech manufacturing	[139, 140]

2.3 General findings

In a collaborative system, participants may interact either at the same (synchronous) or different (asynchronous) moment in time, and at the same (co-located) or different (distributed) location, which results in four different collaboration scenarios [75]. Figure 2.2 shows those scenarios and the identified systems that support them. The most commonly supported scenario is asynchronous distributed (85% of the systems). This scenario is popular because it promotes participation by eliminating the constraints for analysts to synchronize in time and space, significantly increasing the potential scalability of the collaborative effort [55]. Additionally, the study by Benbunan-Fich, Hiltz, and Turof [12] found that asynchronous collaboration resulted in higher-quality outcomes because participants have time to generate and reflect on new ideas and can contribute regardless of their location. Synchronous and co-located (8% of the systems) requires specific hardware for parallel input from multiple sources and parallel output to a potentially diverse audience to enable it, requiring a more demanding design of the interface. Asynchronous and co-located (8% of the systems) means effectively sharing the same input and output devices, which has become rare with falling hardware costs. Lastly, distributed but synchronous (23% of the systems) is more common, but it requires special coordination across different locations and potentially time zones.



Collaborative scenario**	**No**	**%**
Synchronous co-located	1	8%
Synchronous distributed	3	23%
Asynchronous co-located	1	8%
Asynchronous distributed	11	85%
Hybrid collaborative scenario	**No**	**%**
Mixed-presence	0	0%
Multi-synchronous	2	15%
Synchronous co-located and asynchronous distributed	1	8%

No = Number of systems that support the scenario
% = Percentage of systems that support the scenario (out of 13 systems)

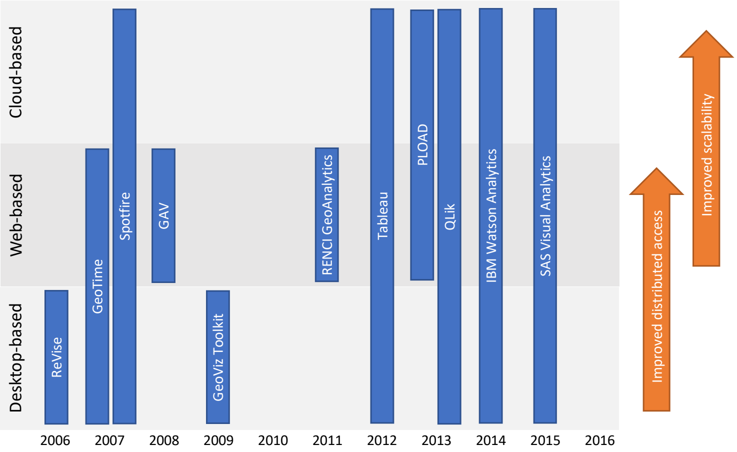
* SAP BusinessObjects appears twice, it was done to avoid confusion by placing it in a diagonal between synchronous co-located and asynchronous distributed.
** Collaborative scenarios are not mutually exclusive, for this reason the column **No** do not sum up to 13 and **%** to 100%.

Figure 2.2 Collaborative analysis in GVA can occur in four scenarios defined by space and time. The figure shows those scenarios and the systems that support them. Additionally, it shows hybrid collaboration scenarios (e.g., mixed-presence and multi-synchronous)

These collaboration scenarios are not mutually exclusive, and a combination is sometimes called a hybrid collaboration scenario. As shown in Figure 2.2, a mixed-presence scenario has co-located and distributed users [95], and a multi-synchronous scenario features synchronous and asynchronous interactions [120]. Isenberg et al. [66] claim the need to expand the research in hybrid collaborative scenarios, which is consistent with the finding that only 23% of the identified systems support some hybrid collaboration scenarios.

The literature argues that the support for time-critical and long-term analysis is of key importance for CGVA [2, 66, 79, 156]. The duration of the analysis effort characterizes these analysis scenarios. In a time-critical scenario, the analysis must be completed as rapidly as possible to minimize undesirable consequences. Examples are analyses in response to natural disasters or terrorist attacks. Among the identified systems, only the RENCI GeoAnalytics Framework [50] supports collaborative time-critical analysis of emergency situations. In a long-term scenario, the analysis extends over a longer time span and usually aims to generate understanding and/or enable strategic decisions. Examples are analysis of climate change and species conservation. None of the identified systems supports long-term analysis scenarios.

GVA environments are increasingly using cloud-based platforms, as shown in Figure 2.3, which is a general trend in analytical systems [161]. Cloud-based platforms offer two advantages: first, flexible and scalable storage capacity and processing power to work with large and complex data sets, enabling users to work from thin clients; Second, distributed access to the system enabled by the Internet, which improves the potential for multi-disciplinary and cross-domain collaboration among geographically separated participants.



Note: The location in this timeline is based in the first reference to a collaborative feature in the systems.

Figure 2.3 The use of cloud technology to deploy CGVA systems is increasing, which improves the scalability and distributed access to the system

Finally, most of the identified systems support multiple device types, such as PCs, smartphones, tablets, touch tables, or large screens (explicit claims were identified for 62% of the systems). The support for multiple device types facilitates reaching a broader audience, which is further promoted by most systems supporting asynchronous distributed collaboration and the increasing use of cloud-based deployment. This combination of technologies removes time and space constraints to participate in analysis efforts and eliminates the need for specialized hardware to access the system. The concept of contributing irrespective of location, device, or time is called Ubiquitous Analytics [34]. This convergence of technologies that dramatically improves the potential for effective collaboration in the analysis of geographic information was already predicted almost two decades ago by MacEachren [90].

2.4 Techniques

The review identified six collaboration techniques: annotation, discussion board, instant messaging, interaction history, snapshot, and storytelling. Table 2.3 offers a summary of the advantages and limitations of each technique. Among these techniques, snapshot, storytelling, and annotation are the most popular ones, implemented by 85%, 62% and 54% of the systems, respectively (See Table 2.4). Further, they are the techniques that co-occurred more often, as shown in Table 2.5a. The combination of these three techniques offers a flexible working environment that allows analysts a seamless combination of independent and collaborative analysis, and it produces self-explanatory results that can be immediately communicated. Finally, Figure 2.5b shows a cross-tabulation of collaboration techniques and scenarios through the systems supporting them. Noteworthy is the absence of annotation technique in the system supporting synchronous co-located scenario, because it can certainly help co-located participants to support claims during a discussion session. Additionally, snapshot is the only technique that occurs in all four scenarios, which provides evidence of its flexibility.

Table 2.3 Advantages and limitations of the identified collaboration techniques

Technique	Advantages	Limitations
Annotation	• Enables analysts to point at, describe and delineate features of interest in the data products • Can carry semantics that link the annotation with the underlying data	• Lack of guidelines to regulate its use may lead to an overload of irrelevant contributions
Discussion board	• Enables topic-centered discussion among geographically distributed analysts • Topics are organized in threads	• Synthesizing discussion results is not trivial

Instant messaging	• Enables discussion among geographically distributed analysts	• Private discussions may lead to lack of awareness of others' work and to fragmentation of the known information • Discussion board is more flexible and better organized
Interaction history	• Documents automatically the analysis as a continuous process • The interaction logs can be stored and accessed based on different models • Allows to review and extend the analysis process	• An interaction history may require revisions before it can be disseminated • Snapshot offers an alternative to document the analysis process as discrete states and is deemed sufficient in most use cases
Snapshot	• Allows to store discrete states of the analysis process on-demand • Stored states can be reconstructed for further analysis • Can be applied to independent visual products or the whole analytical environment	• Unlike interaction history, snapshot cannot reconstruct the interactions that led to the stored states
Storytelling	• Organized in chapters • Supports a flexible analysis process by allowing to update the story • Specific focus in communication of analytical results • Effective, engaging, and easy to understand for specialists and laypersons	• Doesn't incorporate identification of individual's contributions • Doesn't offer provenance of the story

2.4.1 Annotation

Annotation means any piece of information in the form of text or graphic attached to an information product such as a data table, illustration, or map. An annotation may be a mere overlay on a visual product, but it may also be a data-aware artifact carrying semantics that links the annotation with the underlying data [54, 128]. Annotations can be made on the aggregate level of the information product or on individual features comprising them [128]. In (geo)visualization, annotation facilitates access to and recall of contributions (i.e., external memory), document ideas in private and public analysis spaces and elicit information from all participants in a collaborative effort [54, 62].

Annotation has three main functions [54, 62]: 1) to highlight a feature of interest in a visual product, for example, a potentially suitable location for facilities; 2) to provide information on the feature of interest, for example, describing building status during post-disaster damage assessment; 3) and to act as boundary object, for example, delineating areas affected by a natural disaster. Figure 2.4 shows examples of annotation in the context of pest management.

Table 2.4 List of the identified collaboration techniques and the GVA systems implementing them.

	Annotation	Discussion board	Instant messaging	Interaction history	Snapshot	Storytelling	Techniques / System
ORACLE BI Visual Analytics					✓	✓	2
SAP BusinessObjects		✓			✓	✓	3
ReVise	✓			✓	✓		3
GeoTime	✓				✓	✓	3
Spotfire	✓	✓			✓		3
GeoAnalytics Visualization (GAV) Framework					✓	✓	2
GeoViz Toolkit			✓		✓		2
RENCI GeoAnalytics Framework	✓		✓				2
Tableau	✓				✓	✓	3
PLOAD	✓						1
QLik					✓	✓	2
IBM Watson Analytics		✓			✓	✓	3
SAS Visual Analytics	✓				✓	✓	3
Systems / Method	7	3	2	1	11	8	
% systems with method	54%	23%	15%	8%	85%	62%	

The annotation technique is implemented in 54% of the identified systems, and it applies to every loop of the analysis process. During the exploration loop, it enables analysts to highlight, describe and communicate findings. Later, in the verification loop, these findings are interpreted in the problem's domain and constitute insights that may lead to hypothesis generation or identify evidence for existing hypotheses. The annotations create awareness of the findings and insights, and these constitute documentation for accepting or rejecting a hypothesis in the knowledge generation loop. Two aspects for efficient use of annotation are: first, guidelines to moderate the usage of annotations, with the lack of them possibly creating an overload of irrelevant contributions [62]; second, functionality to track existing annotations, create links between annotations to understand their relationships, and synthesize them [19, 168].

2.4.2 Discussion board

Discussion board (also known as discussion forum) enables users to exchange text messages on a chosen topic. Since users are allowed to reply directly to any message, the communication is not necessarily linear but follows a hierarchical structure, in which each branch is called

Table 2.5 a) Co-occurrence of collaboration techniques in the identified systems; values in bold show the two highest co-occurrences of techniques. b) Number of systems that implement a technique and support a collaborative scenario.

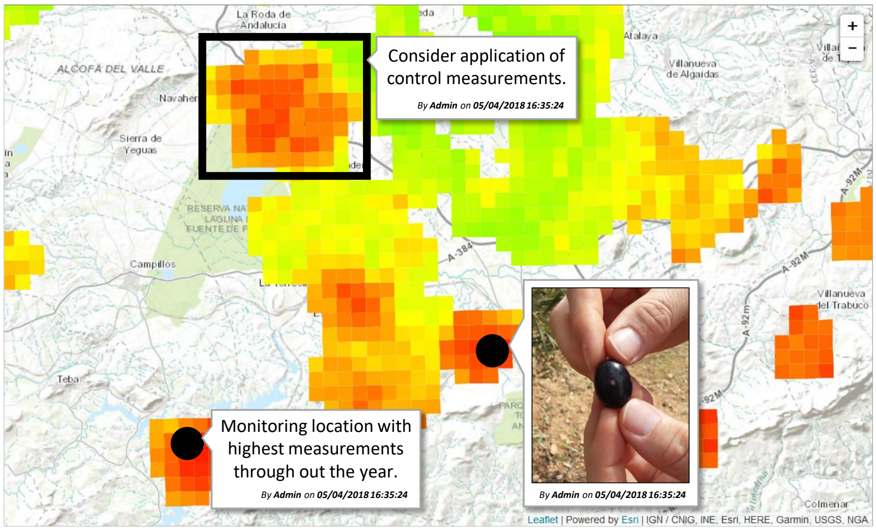
	a						b			
	Annotation	Discussion board	Instant messaging	Interaction history	Snapshot	Storytelling	Synchronous co-located	Synchronous distributed	Asynchronous co-located	Asynchronous distributed
Annotation	-	-	-	-	-	-	0	2	1	6
Discussion board	1	-	-	-	-	-	1	0	0	3
Instant messaging	1	0	-	-	-	-	0	2	0	1
Interaction history	1	0	0	-	-	-	0	0	1	0
Snapshot	**5**	3	1	1	-	-	1	2	1	9
Storytelling	3	2	0	0	**8**	-	1	1	0	8

a thread [164, 171]. Figure 2.5 shows an example of a discussion board in a GVA environment.

The discussion board technique is applicable to all the loops of the analysis process. It provides a mechanism to discuss ideas, generate hypotheses, share findings, reach agreements and plan further actions. Analysts can create threads to discuss findings (exploration loop). These threads document the arguments to understand the findings in the context of the analysis' domain, which constitutes insights (verification loop). These insights can help generate hypotheses or identify evidence, and analysts can create threads to organize and document them. Once enough evidence is available, the analysts can use the content of the threads as input to draw conclusions (knowledge generation loop).

Compared with instant messaging, a discussion board enables analysts to engage in different discussions around the same data view without mixing the topics because each topic has its thread. Such improved organization for discussions through threads makes a message board more suitable to support larger groups. Additionally, the discussions are public, which ensures transparency among all the analysts. Due to the public nature of discussions, it is common to find some moderation mechanism in discussion boards; For example, a participant with privileges to remove inappropriate content or automatic deletion of contents based on a list of forbidden words. While both techniques are based on the idea of message exchange, these differences make message boards slightly more popular, with 23% of the identified systems implementing it, against 15% implementing instant messaging.

On the downside, synthesizing multiple threads can be a cumbersome



Consider application of control measurements.

By Admin on 05/04/2018 16:35:24

Monitoring location with highest measurements through out the year.

By Admin on 05/04/2018 16:35:24

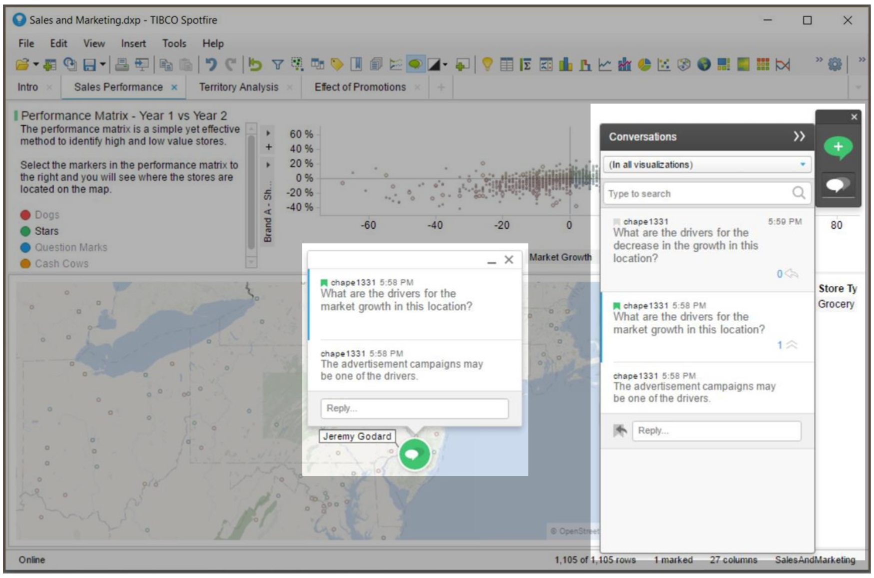
By Admin on 05/04/2018 16:35:24

Figure 2.4 Examples of annotation technique to point at a feature of interest (i.e., location with highest measurements), to describe a feature of interest (i.e., damage in olive fruit), and as boundary object (i.e., area that require control measurements)

task due to unstructured contributions, ambiguity, and unclear references. This issue can be addressed by adopting a formal argumentation model, examples include [89], [132], [133] and [165]. The work by Rinner et al. [133] is particularly relevant for us, as it describes 'Argumentation maps' or 'Argumaps,' which combines the strengths of argumentation modeling and detailed geographic location to support any argumentation process that has a spatial component.

2.4.3 Instant messaging

Instant messaging enables users on a network to exchange text messages [21]. It was originally designed as a one-on-one synchronous communication method, but nowadays, it also serves for discussions among more than two participants and for asynchronous communication. It was popularized originally in the late 1990s by systems such as America Online's Instant Messenger (AIM), Microsoft Messenger, Yahoo! Messenger, and more recently by Facebook Messenger and WhatsApp [27, 118]. In addition to text-based communication, current implementations allow the inclusion of hypertext, multimedia elements, and file exchange. In the context of collaborative analysis, instant messaging enables analysts to work together solving analytic problems regardless of their locations [48]. Like discussion boards, instant messaging provides a mechanism to discuss ideas, generate hypotheses, share findings, reach agreements, and plan further actions during all analysis phases. However, the private and directed nature of instant messaging communications limits its usefulness in a collaborative setting. The reasons are: first, lack of awareness of



Sales and Marketing.dxp - TIBCO Spotfire

Performance Matrix - Year 1 vs Year 2

The performance matrix is a simple yet effective method to identify high and low value stores.

Select the markers in the performance matrix to the right and you will see where the stores are located on the map.

chape1331 5:58 PM
What are the drivers for the market growth in this location?

chape1331 5:58 PM
The advertisement campaigns may be one of the drivers.

Conversations

(In all visualizations)

chape1331 5:59 PM
What are the drivers for the decrease in the growth in this location?

chape1331 5:58 PM
What are the drivers for the market growth in this location?

chape1331 5:58 PM
The advertisement campaigns may be one of the drivers.

1,105 of 1,105 rows 1 marked 27 columns SalesAndMarketing

Figure 2.5 Spotfire allows to create and access discussion boards linked to a visual product or from a ‘Conversations’ panel. Screenshot created using the Sales and Marketing example included in Spotfire.

others’ work that may lead to duplicated efforts; and second, fragmentation of the known information that may difficult to share a common ground for the analysis. Both awareness and common ground are crucial elements for effective collaborative analysis [52]. From the point of view of DC, a key element is the external representation of information/knowledge and its propagation [152]. While the exchanged messages constitute externalization and propagation of information/knowledge, it is limited to the participants of the instant messaging session. Therefore, it does not propagate (at least directly) to the other analysts and limits the cognitive activity of the system as a whole.

Figure 2.6 shows the interface of the GeoViz Toolkit and its instant messaging component called GeoJabber[4]. An interesting feature of GeoJabber is that analysts can share the current status of the visual interface using the technique of snapshot (which is described later) to support claims during discussion sessions [48].

Despite instant messaging being a widely-used communication technique, this literature review shows that its use in CGVA environments is very limited, and only 15% of the identified systems implement it. Three reasons for this are that: awareness and common ground are of key importance for collaborative work, and instant messaging may disrupt them; message boards offer equivalent functionality but are more flexible, and; it is possible that external tools that enable (video)chat communication are used in parallel with the analytics system.

[4] Screen shot created using the Health example included in the software.

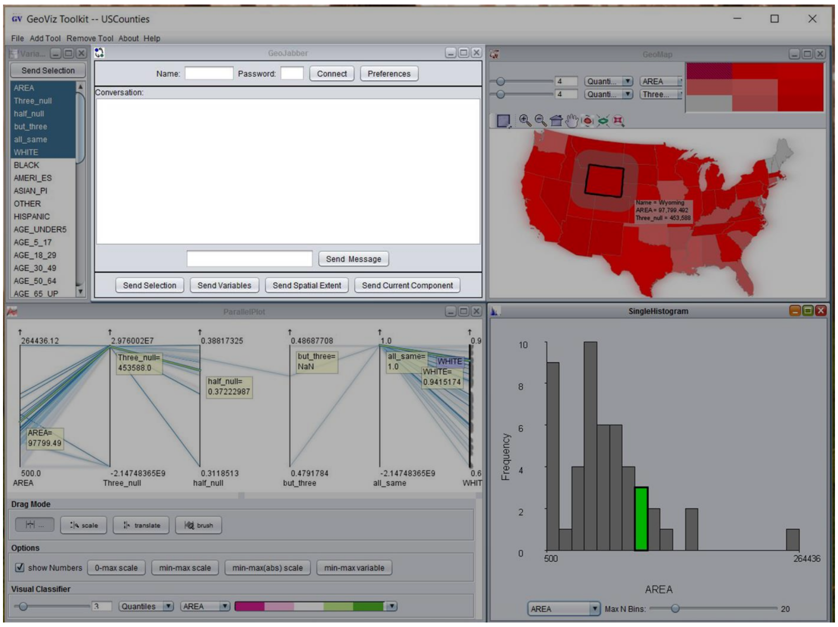


Figure 2.6 GeoViz Toolkit offers instant messaging functionality through the component called GeoJabber.

2.4.4 Interaction history

Interaction history provides analysts with the capacity to save, review, and reuse analytical work. To do so, this technique creates logs of the user's actions and/or state changes in a GVA environment during analysis sessions [53]. These logs can be organized based on different models such as stack, linear, and branching [53, 88]. The stack model is the simplest and enables analysts to undo and redo actions/states. The linear model stores the actions/states in the order of occurrence and enables analysts to transverse the analysis as a linear continuum. The branching model stores the actions/states in a tree-like structure, which allows to document multiple analysis paths.

The review only identified one implementation of interaction history, the Re-visualization technique in ReVise [134]. It allows to save and revisit session logs, and offers the options called 'jump in' and 'breadcrumbs.' Jump in allows to resume an analysis session at any given moment and extend it. Breadcrumbs allows to create indicators of key moments in the analysis session, and attach annotations (based on text or audio) to those indicators to describe why the analyst considers them of importance.

Interaction history supports all the loops of the analysis process by allowing to document the analysts' interaction with the system. Among the identified techniques, this is the only one that automatically and unobtrusively documents the analysis as a continuous process. Additionally, it allows to review and extend the analysis. The general role of this technique is to enable analysts to document and review the analysis process that led to findings, insights, hypotheses, evidence and conclusions. Additionally, it can help in better understanding individual and collaborative strategies during data analysis [53].

This technique is the least popular among the identified techniques, with only one implementation (8% of the systems). A likely reason is that the conceptually similar but technically less complex snapshot technique is implemented by most systems and deemed sufficient for most use cases. They both allow to document the evolution of the analysis process, but they differ in the level of detail. While interaction history documents the analysis as a continuous process and allows to review all interactions and state changes, snapshot only captures discrete states. Another difference is that interaction history automatically documents the analysis process, while snapshot documents the states on-demand. Another factor limiting its utility is the need to edit the exploration logs and analysis process (e.g., select only relevant parts and add narrative) before it can be disseminated [88].

2.4.5 Snapshot

The *snapshot* technique captures the state of a visual product at a given moment, which can be used to document and share findings and insights (e.g., patterns and outliers) for further analysis or communication [88]. A simple and common approach is to capture it as a static image (e.g., ReVise [134]). However, more advanced approaches store different parameters to reconstruct the captured state (e.g., GeoJabber [48]), allowing further exploration and analysis from the stored state. Furthermore, this concept can be extended to capture the whole analytical environment's state, composed of several visual products, as implemented in the GAV framework [59]. Figure 2.7 shows a schematic view of the mechanism to capture and restore snapshots in the GAV framework.

The snapshot technique is commonly used for asynchronous collaboration, allowing analysts to reconstruct saved states on-demand. However, GeoJabber implements a synchronous version to support claims during analysis sessions supported by instant messaging [48]. To do so, GeoJabber includes a mechanism to capture, encode and transfer the snapshot as a particular type of message, which modifies the receiver's viewer to reconstruct the sender's view.

Like interaction history, the snapshot technique can be used in all the loops of the analysis process and enables the analysts to document the evolution of the analysis. Unlike interaction history, snapshot only captures specific states of the analysis process and usually is manually triggered.

Snapshot is the most popular among the identified techniques (85% of the systems) and is commonly used in combination with storytelling. In this review, all eight systems that implement storytelling also implement snapshots, representing the highest co-occurrence of techniques. Additionally, the second-highest co-occurrence is between snapshot and annotation (five systems). This combination allows describing the ana-

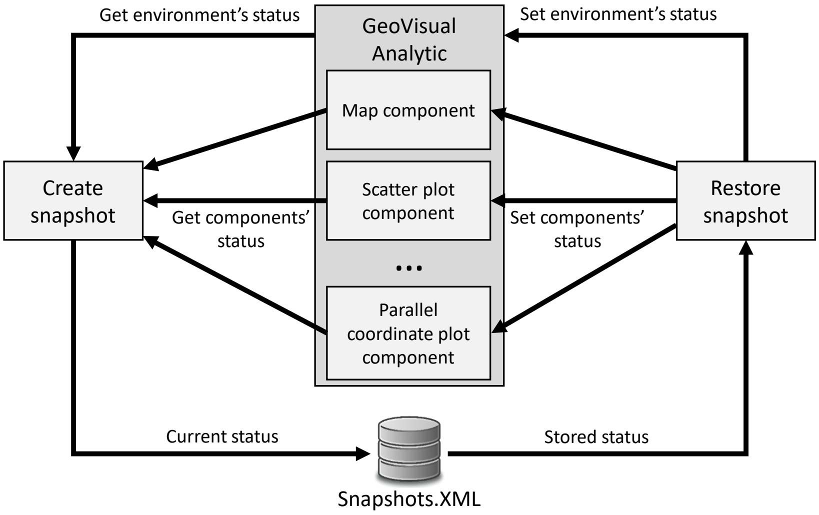


Figure 2.7 The GAV framework is capable of recording the state of all components in the analytical environment as XML; later this file can be used to reconstruct the state of the environment for further analysis, illustration based on [60]

.

lysis process through a 'story' [72] and support claims by using snapshots to document findings and annotation to highlight and describe specific aspects of them [162].

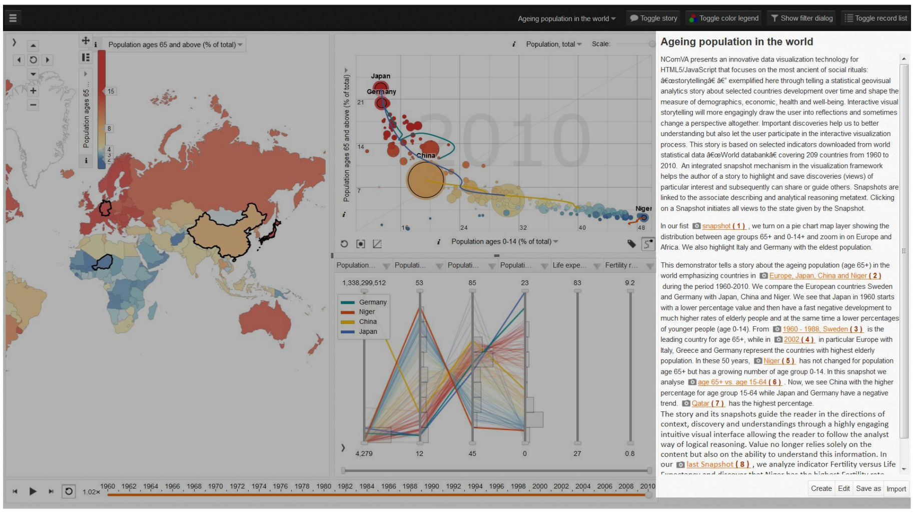
2.4.6 Storytelling

Storytelling is a comprehensive approach combining methods to tell a story about data exploration and the analysis process that led to certain findings or conclusions through interactive visualization [72]. A story may be organized in chapters and include descriptions, multimedia elements, annotations, and snapshots, all facilitating a reader's understanding of the original analytical process [88].

Storytelling can be seen as a communication technique. However, by allowing the reader to interact with the snapshots in the story, he or she can explore further, create new snapshots and modify the story with new findings and insights. In this case, storytelling is not only a communication technique, but also a method for collaborative knowledge building [88]. Figure 2.8 shows the storytelling technique in an application[5] developed with the GAV Framework.

Storytelling (62% of the systems) is applicable to every loop of the analysis process, and it is often used in combination with snapshot and annotation techniques. This combination provides a working space where

[5] Accessible online on `http://mitweb.itn.liu.se/GAV/dashboard/#story`.



Ageing population in the world

NComVA presents an innovative data visualization technology for HTML5/JavaScript that focuses on the most ancient of social rituals: â€œstorytellingâ€ â€" exemplified here through telling a statistical geovisual analytics story about selected countries development over time and shape the measure of demographics, economic, health and well-being. Interactive visual storytelling will more engagingly draw the user into reflections and sometimes change a perspective altogether. Important discoveries help us to better understanding but also let the user participate in the interactive visualization process. This story is based on selected indicators downloaded from world statistical data â€œWorld databankâ€ covering 209 countries from 1960 to 2010. An integrated snapshot mechanism in the visualization framework helps the author of a story to highlight and save discoveries (views) of particular interest and subsequently can share or guide others. Snapshots are linked to the associate describing and analytical reasoning metatext. Clicking on a Snapshot initiates all views to the state given by the Snapshot.

In our fist snapshot (1) , we turn on a pie chart map layer showing the distribution between age groups 65+ and 0-14+ and zoom in on Europe and Africa. We also highlight Italy and Germany with the eldest population.

This demonstrator tells a story about the ageing population (age 65+) in the world emphasizing countries in Europe, Japan, China and Niger (2) during the period 1960-2010. We compare the European countries Sweden and Germany with Japan, China and Niger. We see that Japan in 1960 starts with a lower percentage value and then have a fast negative development to much higher rates of elderly people and at the same time a lower percentages of younger people (age 0-14). From 1960 - 1988, Sweden (3) is the leading country for age 65+, while in 2002 (4) in particular Europe with Italy, Greece and Germany represent the countries with highest elderly population. In these 50 years, Niger (5) has not changed for population age 65+ but has a growing number of age group 0-14. In this snapshot we analyse age 65+ vs. age 15-64 (6) . Now, we see China with the higher percentage for age group 15-64 while Japan and Germany have a negative trend. Qatar (7) has the highest percentage.
The story and its snapshots guide the reader in the directions of context, discovery and understandings through a highly engaging intuitive visual interface allowing the reader to follow the analyst way of logical reasoning. Value no longer relies solely on the content but also on the ability to understand this information. In our last Snapshot (8) , we analyze indicator Fertility versus Life Expectancy and discover that Niger has the highest Fertility rate

Figure 2.8 Applications developed with the GAV Framework can include Storytelling.

the analysts describe the analysis process through a story [72], and document relevant observations using snapshot and annotation [162]. Given that the story can be updated with new findings, insights, hypotheses and evidence as needed, storytelling supports a more flexible analysis process. This flexibility is not possible with instant messaging or message boards because they are based on the idea of appending contributions and not modifying them.

The convenience of this combination of techniques can be explained from the point of view of DC. In this context, the cognitive artifacts are highly important because they represent an individuals' internal representation of information/knowledge externally, thereby enabling cognition across individuals and time [61, 152]. The combination of storytelling, snapshot and annotation constitutes an effective medium for externalization and propagation. This combination offers the flexibility of snapshots to externalize the context with the attention to specific details and relationships of annotations while exposing the entire reasoning process through storytelling. Further, unlike instant messaging, it is publicly available, which ensures propagation of the externalized representations. Lastly, it is succinct because the story evolves to keep only the relevant information, unlike a discussion board that appends all the contributions or interaction history that records all interactions and state changes.

Among the identified techniques, storytelling is the only one with specific focus on effective communication of analytical results. It offers a flexible working space for independent and collaborative analysis, where results (i.e., stories) can be communicated immediately to a broader audience. Authors agree that results presented with storytelling are more effective, engaging and easy to understand for specialists and laypersons [44, 148, 151].

2.5 Synthesis of results

The results at the system level show that the most common collaborative scenario is asynchronous distributed. The reason is that this collaboration scenario does not restrict analysts' participation on time or space, which increases the potential for scalability of the collaborative effort, and improves the quality of analytical results. The review also identified limited support for hybrid collaboration scenarios and for time-critical and long-term analysis scenarios. Neumayr [109] provides an explanation for this finding. The author argues that the lack of support for hybrid collaboration scenarios is due to the absence of theoretical frameworks to inform their design. In their absence, users adopt multiple tools to fulfill their needs in an ad hoc manner. Regarding the technological platform, the use of cloud technology is increasing, which improves the scalability and accessibility of the system. In terms of supported devices, web-based interfaces enable most systems to support multiple devices, such as PCs, smartphones, tablets, touch tables, and large screens.

The most commonly supported collaboration techniques are snapshot, storytelling, and annotation, which also co-occur often. This combination of techniques offers a flexible setup to build knowledge through an iterative process. Storytelling offers a working space where analysts describe the analysis process through a story and document relevant observations using the snapshot and annotation techniques. Snapshot allows to capture the context of an observation, and annotation to highlight and describe specific aspects of it. During the analysis process, the story evolves as new findings and insights are produced, and once completed, the story serves to communicate results to a broader audience.

The identified collaboration techniques do not have well-defined roles within the analysis process's loops (i.e., exploration, verification, and knowledge generation). The review results show that all the techniques are helpful in all the loops. The reason is that the loops overlap, and there is continuous feedback between them. Although the identified techniques support the three loops of the analysis process, the review did not identify any other mechanisms than storytelling to aid in the synthesis of analytic results, which is important to support the knowledge generation loop. Additionally, the review did not identify any mechanism to summarize the level of agreement about the evidence and conclusions, which would provide certainty when results are communicated.

2.6 Challenges of contemporary collaborative geovisual analytics

The literature review identified three research challenges to support collaborative analysis in GVA systems. The challenges are the lack of support for hybrid collaboration scenarios, cross-device collaboration, and time-critical and long-term analysis. These are described in the following sections.

2.6.1 Challenge 1: hybrid collaboration scenarios

Collaborative systems are commonly characterized by the time and space in which collaboration takes place. This characterization of the systems defines four collaboration scenarios (See Figure 2.2). Any combination of these scenarios is a hybrid collaboration scenario, e.g., mixed-presence supporting co-located and distributed participants [95, 96], or multi-synchronous supporting synchronous and asynchronous contributions [120]. In 2011, Isenberg et al. [66] claimed the need to further research hybrid scenarios in information visualization and raise the expectation to see more systems supporting them in the following years. This review shows that it remains a challenge in GVA, with only 23% of the identified systems supporting some hybrid collaboration scenario.

To effectively support collaborative analysis, it is necessary to consider that a typical analysis effort comprises of many different tasks, each of which may benefit or even require more than one collaboration scenario. Therefore, instead of forcing analysts to work in a specific collaboration scenario, GVA systems should enable them to move seamlessly between scenarios. For example, in emergency management, co-located and distributed analysts need to collaborate in real-time during an emergency situation. However, during the relief stage, asynchronous collaboration may be more suitable. In this example, the analysts benefit from a hybrid scenario related to the location (i.e., mixed-presence) and another one related to the time of collaboration (i.e., multi-synchronous).

To address this challenge, we need to research: the suitability of collaboration scenarios for specific types of tasks; evaluate the advantages and disadvantages of hybrid scenarios which may depend on the application domain; and design mechanisms that allow analysts to seamlessly move from one collaboration scenario to another while keeping all the analysis contributions/results available and ensuring awareness of others' work. Additionally, special attention is necessary for scenarios in which analysts may work offline, which may require local temporal storage and a versioning system so that the client devices can synchronize when a connection is available.

2.6.2 Challenge 2: cross-device collaboration

The support for multiple types of devices can provide the analysts with a more flexible analysis workflow, engage a more diverse audience, and facilitate collaboration [33]. 62% of the identified systems explicitly claim to support multiple types of devices (see Table 2.2). However, there is little evidence of those systems taking advantage of the unique characteristics of each type of device. For example, most of the identified systems support smartphones, but only as viewers; in contrast, Big Board [51] allows the use of integrated sensors on smartphones to capture information (e.g., photos, videos and sounds) and create geo-located annotations to share it.

Badam, Fisher, and Elmqvist [8] recognize the need for cross-device collaboration in visualization systems, identify the potential in current technologies to realize it and propose the concept of ubiquitous analytics and visualization spaces. These are collaborative visualization environments that support multiple types of devices connected through a network. The support for cross-device collaboration has the potential to improve multi-disciplinary and cross-domain analysis by enabling actors from diverse backgrounds (e.g., scientists, domain experts, and laypersons or citizen scientists) to participate without requiring specialized devices or specific hardware [8, 34]. To effectively support cross-device collaboration, the user interface needs to take advantage of the unique characteristics of diverse types of devices. For example, in the analysis of species distribution, an in-office analyst may benefit from using desktop workstations to identify features of interest, such as patterns and outliers, and to develop a hypothesis. In contrast, an in-field analyst may benefit from using smartphones to check for the status of the species' population in his or her surroundings (using location-based services) and to capture information that may act as evidence. Especially in the growing field of Citizen Science, it is a crucial element of any project to let participants with diverse skills, capabilities, interests, and hardware collaborate.

To address this challenge, we need to research: the capacity and limitations of each type of device in the context of CGVA; analyze the activities that each type of device may support; design mechanisms that allow participants to seamlessly move from one device to another while all contributions remain available; and develop an infrastructure that ensures responsiveness regardless of the device in use. Additionally, research is needed to evaluate the devices' suitability concerning the diverse collaboration scenarios.

2.6.3 Challenge 3: time-critical and long-term analysis

The support for time-critical and long-term analysis scenarios has been claimed to be of key importance for (G)VA [2, 79, 156]. The duration of the analysis effort defines these scenarios. In a time-critical scenario, the analysis effort lasts for only a short time, and timely results are required to minimize undesirable consequences. Examples are analysis in response to natural disasters, terrorist attacks and cyber-attacks. In a long-term scenario, the analysis effort extends for a much longer time span and aims to generate understanding and/or enable strategic decisions regarding such phenomena. Examples are analysis about climate change, urban dynamics and species conservation.

Only one of the identified systems supports time-critical analysis, and none support long-term analysis. While it can be argued that any system can work in both scenarios, the lack of specialized functionality hampers the flow of the analysis process. For example, in an emergency situation (a time-critical scenario), analysts need specialized tools to deal with rapidly changing analysis conditions (i.e., real-time data updates, dynamic planning and coordination, and awareness of others' work) and tools to negotiate and reach consensus in an agile manner [2]. These are not expected characteristics from a general-purpose GVA system and potentially not the ones required for long-term analysis.

There are diverse applications domains that can benefit from time-critical and long-term analysis scenarios. However, to support these analysis scenarios, specialized functionality is needed. To address this challenge, we need to research how to design and evaluate systems for time-critical analysis that prevent conflicting interaction due to the concurrent access to resources during the analysis sessions, ensure awareness of all the participants regarding the progress of the analysis, and facilitate timely communication of results. Additionally, we need to research how to design and evaluate systems for long-term analysis that adequately summarize the analysis progress, allow participants to work with multiple projects and multiple working hypotheses in parallel, and facilitate the communication of partial and final results.

2.7 Chapter summary

A literature review was conducted to describe the state-of-the-art regarding the support for collaborative analysis in GVA. The review followed the guidelines for systematic literature review proposed in [83], which resulted in the identification of thirteen systems, seven collaboration techniques and three research challenges.

The review shows that the most common collaborative scenario is asynchronous distributed. It is common because removing the constraints of place and time to participate promotes participation in the analysis effort. Additionally, the review shows that cloud-based deployments are increasing, which improves the scalability and accessibility of the system.

Six collaboration techniques were identified: annotation, discussion board, instant messaging, interaction history, snapshot, and storytelling. The most commonly implemented techniques are snapshot, storytelling, and annotation, which also co-occur often. This combination of techniques offers a working space where analysts describe the analysis process through a story and document relevant observations using snapshot and annotation. The story evolves as the analysts add new findings and insights, and once concluded, the story serves to communicate analytical results to a broader audience.

Finally, the literature review identified three prominent and pressing research challenges. These are the lack of support for hybrid collaborative scenarios, cross-device collaboration, and time-critical and long-term analysis scenarios.

3 Designing a software architecture for collaborative geovisual analytics

In Chapter 2, three prominent and pressing research challenges for CGVA were described; these are the lack of support for: hybrid collaboration scenarios, cross-device collaboration, and time-critical and long-term analysis. This chapter addresses those challenges by proposing a reference model to design and develop systems that offer such features. Therefore, this chapter addresses the second specific research objective stated in Chapter 1: "Design a software reference architecture for collaborative geovisual analytics systems."

The chapter is structured as follows: Section 3.1 provides a brief introduction to software architecture and to the architectural patterns used in designing the proposed architecture; Section 3.2 describes the design criteria, which are based on the analysis of the CGVA systems and the research challenges identified in Chapter 2; and Section 3.3 describes the proposed software reference architecture.

3.1 Software architecture

A system's software architecture is an abstract high-level description of the structures needed to reason about the system. An architecture comprises of elements such as classes, processes, devices, and protocols; their relationships such as 'shares data with,' 'provides services to,' and 'executes on'; and the properties of both elements and relationships, that together form a software system [11, 104]. An architecture defines the fundamental structures (i.e., architectural design) that address functional, non-functional, and performance requirements, but not the software system's implementation details (i.e., detailed design). However, even for experienced software engineers, it is often hard to separate them clearly [6]. Fundamental structures have an impact on a significant part of the system, or even the entire system, and are therefore hard and expensive to change once implemented.

The process of designing a software architecture is complex because there are many concerns to address, such as the user's interface, data processing, security, data storage, and the communication and coordination between the components that perform each function. Furthermore, functional and non-functional requirements (which might be contradictory) set constraints on the design. In this regard, a key concept for software architecture is *architectural patterns*, which are reusable solutions with well-understood properties for commonly occurring problems in software architecture design [11, 6]. Some examples of architectural patterns are *client-server*, *layered*, and *microservices* [130], which are described in the following sections. Software architects commonly design an architecture by selecting, combining, and fine-tuning several patterns [11]. For example, a web-based system might use a three-tier client-server pattern (i.e., two client-server relationships as shown in Figure 3.2), and within this pattern use layers to organize its modules [78, 11].

The design of a system's architecture aims to provide the system with specific *quality attributes*. "A quality attribute is a measurable or testable property of a system that is used to indicate how well the system satisfies the needs of its stakeholders" [11] such as scalability, extensibility, and security, which have a direct impact on how well the intended functionality can be provided. Nevertheless, developers often start coding without defining a formal architecture, which may lead to what is frequently called *the big ball of mud* architecture anti-pattern. This anti-pattern means unorganized modules that lack roles, responsibilities, and relationships to one another [130]. Such software systems are composed of tightly coupled modules, making them difficult to maintain, update, or scale. It is important to highlight that a detailed system architecture design is a requirement but not a guarantee for quality attributes. A well-designed architecture that is (partially) not realized in the implementation may lead to a software system that does not achieve some or any of the desired quality attributes, which is known as *architecture erosion* [155].

It is important to understand two key terms to discuss and document software architectures: *structure* and *view*. A structure is a set of elements as they exist in software or hardware. Structures can be categorized as: *module structures*, which deal with the (static) organization of code or data units, such as classes and layers; *component-and-connector structures*, which deal with the (dynamic) organization of the system at run-time, such as services and processes; and *allocation structures*, which deal with the relationships between software elements and the environments in which these are created and executed, such as development teams and computer networks [11]. In contrast, a view is a representation of a coherent set of elements (according to a chosen notation) which allows us to focus our attention on a small number of the system elements [25] such as the classes (in design time) or objects (in execution time) that form a user interface.

Software architectures can have different goals and scopes. The one presented in this chapter is a *software reference architecture*, which can be described as “a reference architecture (RA) is used for the design of concrete architectures in multiple contexts serving as an inspiration or standardization tool” [4, p. 417]. In comparison with a *concrete software architecture* (i.e., for the development of a single software system), “their design and application take place in a broader and, hence, less-defined context with a larger and less-defined stakeholder base” [4, p. 417]. The proposed architecture aims to provide a reference model to facilitate the design and development of CGVA systems by multiple organizations. In the classification by Angelov, Grefen, and Greefhorst [4], this is a reference architecture of type 3, which is described as a “classical, facilitation architecture designed for multiple organizations by an independent organization” (p. 423).

In the following sections, three relevant architectural patterns are discussed: client-server, layered, and microservice. These patterns were used to design a software reference architecture for CGVA, described in Section 3.3. The discussion here shows that a combination of these patterns allows to design software that is easy to develop, test, and maintain because components are organized into tiers and layers with clearly defined and limited scope (i.e., *separation of concerns*, which is defined later, provided by client-server and layered patterns) [142, 130]. Additionally, we discuss how the microservices pattern can fit in the architecture. This pattern offers flexibility (i.e., support for multiple programming languages and ease to accommodate new functionality), maintainability (i.e., small independent units), and scalability (i.e., independently deployed units) [130].

3.1.1 Client-Server architecture

Client-Server is a distributed architecture pattern, which consists of two software levels, *client* and *server*. These commonly reside on different hardware and communicate over a network, but they can also reside on the same hardware [142]. The two components have clearly differentiated roles in the architecture. The server offers services (sometimes referred to as resources), for example, access to data and processing capabilities, which can be used by the client on-demand [11]. The server is always listening for requests from the client, with the latter being in charge of starting the communication [6]. The centralization of the resources in the server improves security and facilitates managing authentication and authorization [142]. Another characteristic is that the server component can process concurrent requests from several clients, as shown in Figure 3.1.

This architectural pattern offers good scalability, meaning that the system can easily accommodate new clients and servers [11]. To this end, the server-side can include a component that distributes the workload (i.e., a *load balancer*) among several servers. In this scenario, servers can be added or removed to cope with fluctuations in the number of clients and service requests. The load balancer component and the changing number of servers are not visible to the clients, who do not need to take

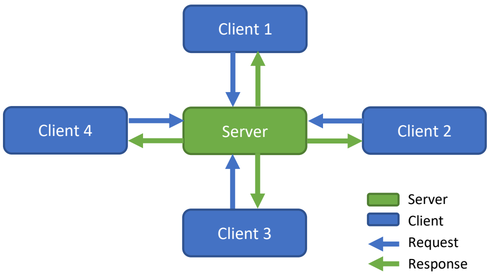


Figure 3.1 In a Client-Server architecture, a server can attend concurrent clients.

any action when the server-side needs to scale up or down. This scaling process may even be controlled by an algorithm that automatically adjusts the number of servers according to the changing workload.
The Client-Server pattern can be applied multiple times to build an arbitrarily complex software architecture. By creating multiple client-server relationships, a server can act itself as a client to other servers, resulting in what is called an *n-tier architecture* [6]. Currently, a typical application for this type of architecture is web-based systems. Figure 3.2 shows an architecture for a web-based system, in which the Client-Server pattern is applied two times. In this architecture, the tier implementing the system's logic (i.e., a web server) participates in both functions: server and client.

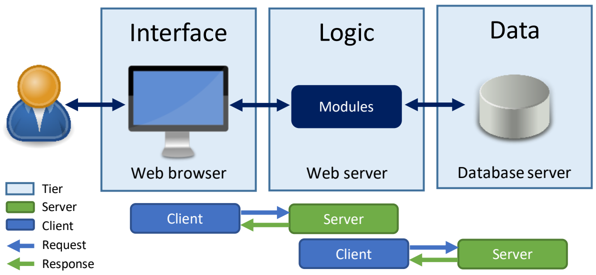


Figure 3.2 A typical architecture for web-based systems includes three tiers: interface, logic and database. On this architecture the web server acts as the server component for the interface, and as client for the database server.

So far, we have discussed the characteristics and advantages of this architectural pattern; however, every pattern has disadvantages. There are two potential issues related to the server in this pattern: first, it can be a performance bottleneck; if the server does not have enough capacity to cope with the service demand, the whole system would suffer from bad performance. Second, the server is a single point of failure, which means that if the server fails, the whole system fails, because the clients' requests cannot be attended [11]. These disadvantages can be addressed with diverse strategies, but there is always a trade-off to be considered.

For example, to address the bottleneck problem, a load balancer can be added (as described earlier), which in turn adds complexity to the system.

3.1.2 Layered architecture

In the layered architecture pattern, components are grouped into layers with well-defined roles and responsibilities within the application, and commonly each layer is only allowed to use the functionality (through an interface) of the layer immediately below [130, 11]. Layered is the most commonly used pattern [130, 11] because grouping components based on their functionality (i.e., separation of concerns, described below) naturally leads to this pattern. While the pattern itself does not specify the number of layers to be included, architectures with four layers are most common: presentation, application logic (sometimes also called business logic), data access, and database.

A powerful feature of this architecture pattern is the *separation of concerns*, meaning that each layer has a well-defined scope, which limits the functionality provided by the layer. For example, the components in the presentation layer are responsible only for the user interface functionality, while matters such as how the data is processed (application logic), accessed (data access), and stored (database) are out of its scope and are actually unknown and irrelevant to it. The separation of concerns makes it relatively easy to define clear roles and responsibilities in the architecture and further to develop, test, govern, and maintain the software [130].

A layered architecture is commonly represented as a stack of boxes, where layers are allowed to use the functionality of the layer immediately below. This allowed-to-use relationship is not represented by arrows as in other patterns but by simple adjacency, where the top layer is allowed to use the one immediately below [11]. Because of this communication pattern, a call to a function in the top layer that requires data may have to transverse all the layers to get it from the database, unless the data is already available in some layer, perhaps in cache memory. Figure 3.3A shows a layered architecture with the previously mentioned four layers and a request that goes through all the layers. In some cases, it is convenient to allow what is called *layer bridging*, which means that layers are allowed to use other layers than the one immediately below; hence requests can avoid passing unnecessarily by some layers [130, 11]. For example, in Figure 3.3B, the Logic layer can directly access two layers, Services, and Data access. Additionally, using the layers to the right is allowed. Therefore, all the layers can use the Security layer. It is advisable always to specify the rules for the allowed-to-use relationships because this avoids the reader guessing the semantics of the diagram.

An important consideration when using layered architecture is to be aware of the *architecture sinkhole anti-pattern*. This anti-pattern is the situation in which requests go through the layers with little to no processing happening [130]. Every layered architecture will have some

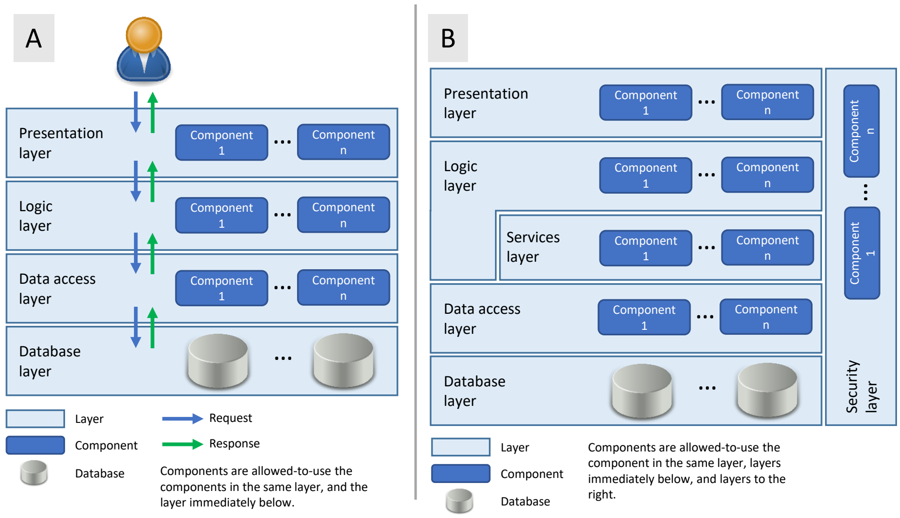


Figure 3.3 A) In a strict layered architecture, layers are only allowed to use the layer immediately below, therefore, a request for data on the top layer has to pass through all the layers to get it from the database; B) The allow-to-use relationship can be modified to allow a layer to utilize multiple layers which is known as layer bridging. Illustration based on [130, 11]

requests showing such behavior, but if they represent a high percentage of the requests, actions to solve it must be taken. It is normal to apply the Pareto principle (also known as the 80/20 principle) for this purpose, meaning that it is considered acceptable to have 20% of requests showing this behavior. In an extreme situation, it may be the case that the architecture pattern is not adequate for the project at hand.

The main disadvantage of this pattern is that the number of layers directly impacts the performance of the system [130, 11]. The reason is that each layer adds some overhead to the management of a request; even if there is no processing happening in a layer, there is still latency for context switching between layers. Additionally, the decision about in which layer to locate functionality during the architectural design is crucial because once the system is built, it is hard and costly to change it. Finally, a wrong decision regarding the scope of the layers might compromise the development of future functionality.

3.1.3 Microservices architecture

Microservices is a distributed architecture pattern in which the functionality of the software system is divided into small, independently-deployable, loosely-coupled, collaborating units called *microservices*. While this architecture simplifies the design, development, and maintenance of the functional units, it requires a more complex hardware and software infrastructure when compared to a monolithic architecture (i.e., all functionality implemented as one deployable unit). Two recent descriptions of this architectural pattern are:

> “The microservice architectural style is an approach to developing a single application as a suite of small services, each running in its own process and communicating with lightweight mechanisms, often an HTTP resource API. These services are built around business capabilities and independently deployable by fully automated deployment machinery. There is a bare minimum of centralized management of these services, which may be written in different programming languages and use different data storage technologies.” [131]

> “At its core, microservices is a method of developing software applications as a suite of independently-deployable, modular services. Each service is configured to run as a unique process and communicates through a well-defined, lightweight mechanism to serve a business goal.” [146, p. 10]

This architecture pattern favors the notion of “you build, you run it.” This notion means that software development teams are fully responsible for developing and supporting a set of services throughout their life cycle. This working pattern is possible because the services are independent of each other, as illustrated in Figure 3.4. The illustration shows that each service has its data, logic, and entry point (i.e., API); therefore, they are fully decoupled. In some scenarios, information and functionality need to be shared between components. The former can be addressed by using a shared database, but this might lead to services coupling, which is undesirable in a microservices architecture. The latter can be addressed by copying a small portion of business logic between components, which violates the DRY (Don’t Repeat Yourself) principle, but which is acceptable for the sake of component independence [130].

As mentioned before, the system functionality is provided by collaborating services, and this is achieved by either *orchestration* or *choreography*. In the former, a component coordinates the cooperation between services, i.e., centralized coordination. In the latter, the collaboration occurs through message exchange among the microservices without centralized coordination [29]. Given that orchestration leads to service coupling due to the centralization of the coordination, choreography is preferred [29].

The main disadvantage of this pattern is that it adds upfront cost and complexity to the system, particularly in the form of hardware and software infrastructure to execute the system, which is often not justifiable on small projects; therefore, its use might be restricted to large systems [130, 146]. Additionally, it suffers from known downsides of distributed architectures, such as issues related to security, integration, and network reliability [29].

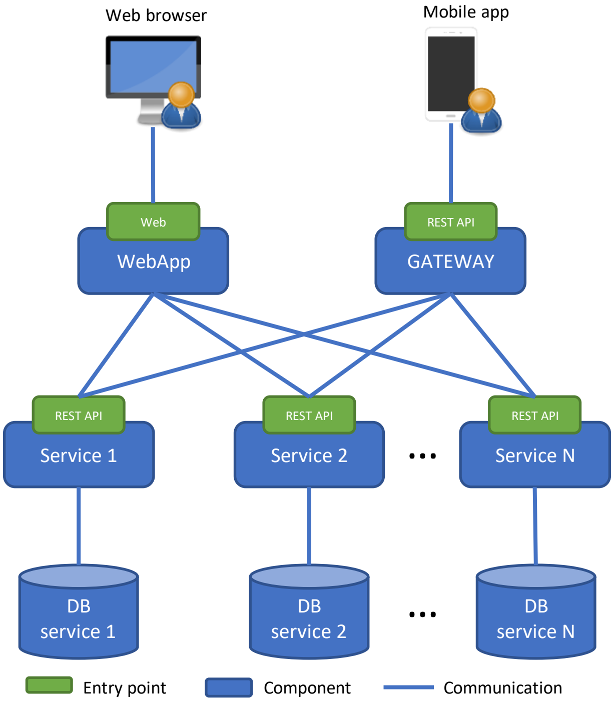


Figure 3.4 In a microservice architecture, the system's functionality is divided in small, independent units called microservices. Illustration based on [131]

3.2 Design criteria

Based on the literature review presented in Chapter 2, a software reference architecture for CGVA systems was designed, which is described in Section 3.3. This section discusses how the identified systems and challenges shape the architectural design, which provides the rationale for choosing client-server, layered, and microservices patterns as the building blocks for the software reference architecture.

3.2.1 General characteristics of CGVA systems

An effective collaborative analytics system must support a combination of individual and collaborative analysis [69, 92]. Despite that real-world analysis often includes a combination of individual and collaborative activities, most VA systems are either single-user or purely collaborative systems [69]. For this reason, the reference architecture includes individual and collaborative workspaces and defines a continuous analysis workflow across them (See Figure 3.5A). In an individual workspace, the user can analyze data and document his or her findings in a private space. If deemed relevant, the analyst can make those findings available to others by exporting them to a collaborative workspace. Additionally, if the analyst is interested in working individually on a contribution available in a collaborative workspace, this can be imported. On the other hand, in a collaborative workspace, analysts work together analyzing data and shared individual contributions.

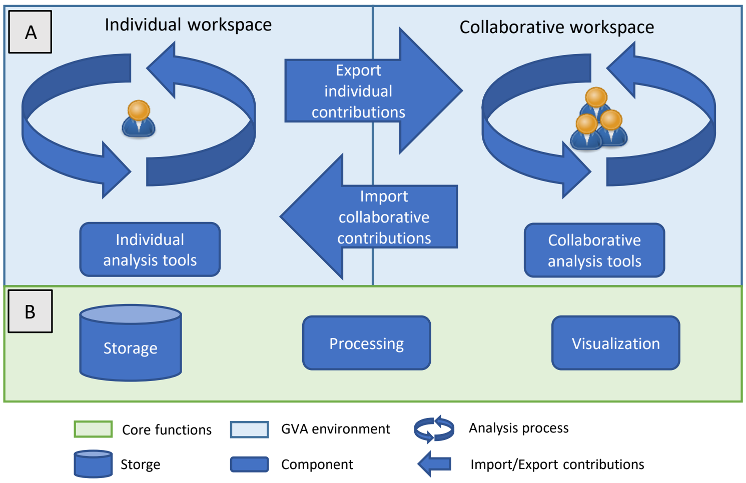

A
Individual workspace
Collaborative workspace
Export individual contributions
Import collaborative contributions
Individual analysis tools
Collaborative analysis tools
B
Storage
Processing
Visualization
Core functions
GVA environment
Analysis process
Storge
Component
Import/Export contributions

Figure 3.5 A) Combining individual and collaborative analysis. B) The workspaces to support individual and collaborative analysis are GVA environments, therefore, they require functionality to store, process, visualize and analyze the data. Illustration based on [69]

Despite the specific functionality offered by the individual and collaborative workspaces to fulfill their roles in a system, they are both GVA environments. Therefore, their core includes functionality to store, process, visualize and analyze the data (See Figure 3.5B). This functionality may vary greatly depending on the application domain. For example, while the analysis of pest populations' dynamics requires time series of fixed monitoring locations, the analysis of animal migrations requires time series of moving objects. This difference in the type of data affects the requirements for functionality to store, process, visualize and analyze the data. For this reason, the reference architecture does not define the specific functionality for a system but proposes an organization for the components that provide such functionality.

As in any other system type, security is an essential matter for CGVA systems. Security commonly includes functionality such as authentication, authorization, encryption, and activity logging. The system components require access to security functionality to ensure that all activities are performed by authorized users, and to maintain a historical record of those activities. To address this need, the software reference architecture includes components to provide security functionality.

3.2.2 Hybrid collaboration scenarios

In a collaborative system, the participants' interactions can be characterized by the time at and place in which they occur. These two dimensions allow to define four collaboration scenarios: synchronous and co-located (i.e., same time and place), asynchronous and co-located (i.e., different time but same place), synchronous and distributed (i.e., same time but different place), and asynchronous and distributed (i.e., different time and place) [75]. In practice, these scenarios are not mutually exclusive, and analysts often cross their boundaries during collaborative work [82]. A combination of these scenarios is sometimes called a hybrid collaborative scenario [109, 66], e.g., multi-synchronous and mixed-presence.

Literature offers various definitions for multi-synchronous collaboration. Two definitions are: "Multi-synchronous collaboration is a process in which some users work in real-time (e.g., desktop-based users) while other users work in isolation and commit updates when necessary (e.g. mobile users)" [143, p. 1], and "Multi-synchronous authoring tools allow simultaneous work in isolation and later integration of the contributions" [124, p. 6]. From the previous definitions, multi-synchronous collaboration enables delayed integration of contributions as shown in Figure 3.6, which is particularly relevant to support analysts to work offline. In this scenario, analysts might work simultaneously as in synchronous collaboration, but they do not necessarily submit their contributions immediately to a shared database. Later, the contributions can be reconciled and submitted to a shared data storage, so that other users see changes in a delayed manner as in asynchronous collaboration. For example, in the analysis of pest population dynamics, while some analysts might work in-office with a permanent connection to a shared database, others might work in the field with no (reliable) connection to the shared database. Therefore, to enable them to collaborate, the system should support multi-synchronous collaboration. There are two main concerns for a system to support this collaboration scenario: synchronization of contributions and analyst-stored copy of shared data required for offline work.

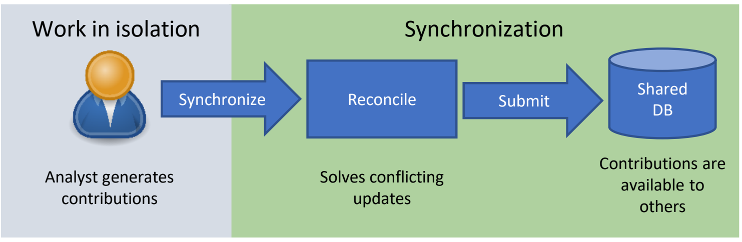

Work in isolation
Synchronization
Synchronize
Reconcile
Submit
Shared DB
Analyst generates contributions
Solves conflicting updates
Contributions are available to others

Figure 3.6 In multi-synchronous collaboration, analysts might be working simultaneously, but contributions are not available to others until they are submitted to a shared database.

Assuming that the system supports synchronous, asynchronous, and multi-synchronous collaboration, there is a need for a synchronization mechanism which behavior depends on the collaboration mode. Such a mechanism is described in [103]. For the synchronous and asynchronous mode, the analyst's contributions are immediately integrated into the shared database, although not necessarily viewed at the same moment because some analysts might be offline. Therefore, we need to consider the following situations: An online user should receive synchronous updates when contributions (e.g., messages, annotations, or snapshots) occur. For this mechanism to be effective, the updates should not disrupt the analyst's work, but at the same time, be evident to ensure awareness of their occurrence. A user that comes online should receive notifications about the contributions that occurred when he or she was offline. By including a timestamp with the contributions, the system can determine which occurred when a user was offline. Figure 3.7 shows the behavior of the synchronization mechanism in those two situations. Additionally, given that contributions are persistent and timestamped, a history log is always available.

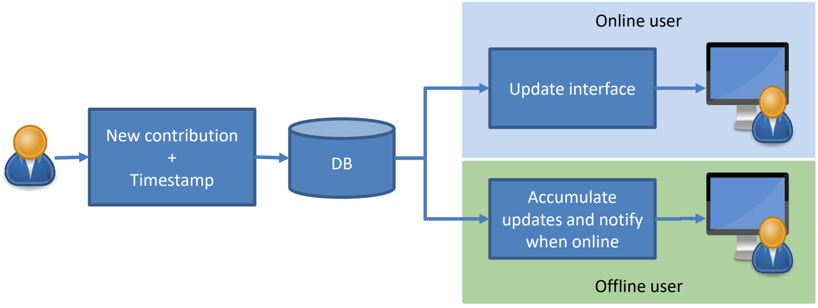


Figure 3.7 An online user receives synchronous updates of new contributions, while an offline user will be notified of new contributions on the next login.

Some tools work specifically in synchronous manner (e.g., video conference). The user needs to know who is online to allow starting a collaborative activity with such tools. To this end, an up-to-date list of online users is required. One mechanism to build such a list is the *heartbeat*, which is a periodic signal that informs the system about an analyst's online presence. Figure 3.8 shows a schematic view of the heartbeat mechanism to build the users' presence list. This functionality can be provided by the synchronization component.

The synchronization mechanism in multi-synchronous collaboration is complex because contributions are integrated in a delayed manner and in batches, which might create updating conflicts [103]. For this reason, an analyst might need to reconcile his or her contributions with the current state of the shared database before they can be integrated (see Figure 3.6). This collaboration scenario is appropriate when an analyst does not have a permanent connection to a shared database, e.g., during fieldwork in areas without a reliable Internet connection.

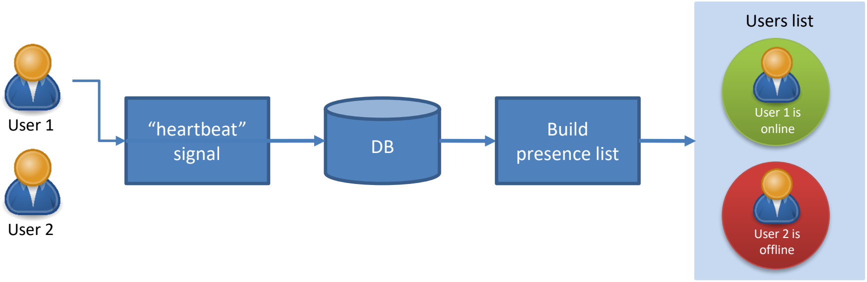


Figure 3.8 Through a heartbeat signal, the system knows whether a user is online or offline.

The second consideration is a component to manage local storage. This component has two roles: first, to store data to support offline work; and second, to serve as cache memory to improve the system's performance by reducing the need to contact a server. For the former, it is unlikely that this component might pull all the data from the server because it might be too big, which makes it impractical, but also because the amount of local storage is limited. Instead, the component should allow the analyst to specify which data to store locally before working in offline mode. For the role as cache memory, the component should specify a device-specific storage limit and a memory management algorithm, for example, *first-in-first-out* (commonly referred to as FIFO). This means that data is cached when retrieved from the server, and when the memory is full, the oldest content is removed to allocate memory to cache new data.
A mixed-presence scenario can be described as distributed collaboration among groups of co-located analysts [95, 96, 82], as shown in Figure 3.9. In the illustration, a site with a single analyst working individually in the local context and collaboratively in the system's context is included to account for such a scenario. For example, during an emergency situation, staff from institutions such as the police, firefighters, and army might work collaboratively in a co-located manner within the institutions' facilities and remotely (i.e., distributed) with the other institutions. There are two main concerns for a system to support this scenario: synchronization across sites/devices, and multiple interfaces. The former was already discussed in the previous paragraphs. The latter relates to the support for cross-device collaboration, which is discussed in the following section.

3.2.3 Cross-device collaboration

For cross-device collaboration to be effective, the system should provide an interface that takes advantage of the unique characteristics of each device type and a mechanism to ensure that contributions are synchronized across them. These requirements pose three main concerns: multiple interfaces, separation of the interface, processing and storage, and

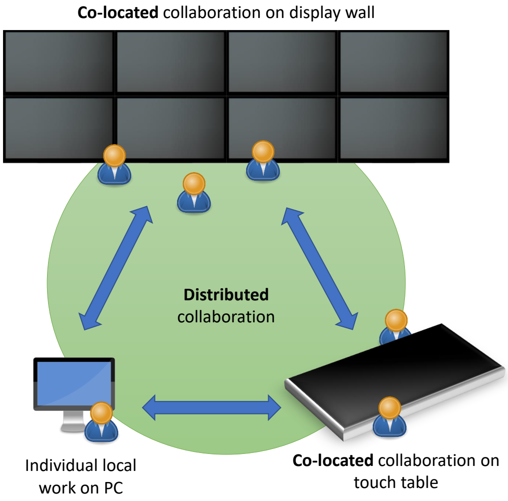

Co-located collaboration on display wall
Distributed collaboration
Individual local work on PC
Co-located collaboration on touch table

Figure 3.9 A system supporting mixed-presence enables distributed collaboration among groups of co-located analysts, and potentially individual ones. Additionally, this scenario is also an example of cross-device collaboration.

synchronization across devices. These concerns are discussed in the following paragraphs, except for synchronization across devices which was discussed in the previous section.

Differences in screen size, interaction capability, and integrated sensors make the development of a single interface that works appropriately for several device types a highly complex task. Let us, for example, think about a touch table and a smartphone. While they both provide a touch interface, they significantly differ in other characteristics. The screen size of a touch table provides a lot of room to display information and makes it a good choice for users to work concurrently, while the smartphone offers limited display space and is better suited for use by a single user at a time. Another important difference is mobility, where smartphones are designed to be carried around easily, touch tables are designed to be used at fixed locations. One must consider that devices may play a specific role as part of the system, hence depending on the device in use, the interface may offer specialized functionality. For example, in the analysis of species distribution, an in-office analyst may benefit from using desktop workstations to identify features of interest such as patterns and outliers and develop a hypothesis. In contrast, an in-field analyst may benefit from using a smartphone to check for the status of the species' population in his or her surroundings using location-based services and capture information that may act as evidence. Based on these reasons, the architecture should offer the possibility of having multiple specialized interfaces for a diversity of devices instead of a highly-complex one-fits-all solution. To this end, the architecture can divide the system's functionality into layers. The interface is a layer whose responsibility is to provide the link between a user and the rest of the system. Additionally, each implementation may offer specialized tools depending on the target device type. Figure 3.10 shows that the system may offer multiple implementations of the interface layer, each one specialized for a device type.

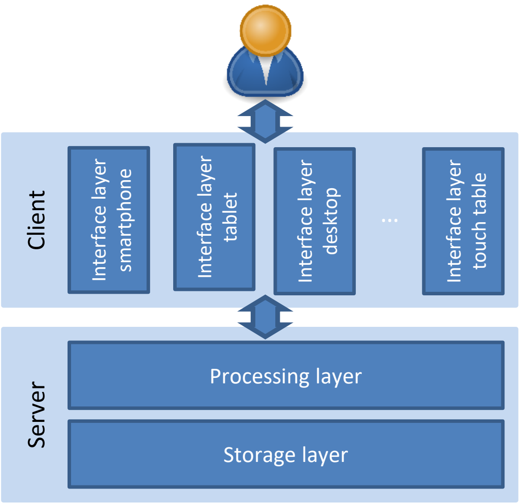


Figure 3.10 The architecture defines the user interface as a layer, therefore, multiple specialized interfaces can be developed without affecting the rest of the software system.

Another important difference between devices is concerning processing and storage capacity. Therefore, there is a need to separate the interface from the processing and storage. For example, a tablet can be convenient for mobility, but it is not a good option to run a heavy computational statistical model because of processing and storage limitations. The implication of this for the architecture is that the processing and storage load should be minimized in the user's device and take place somewhere else. Thus, the architecture can divide the system into client- and server-sides, as shown in Figure 3.10. The former provides the interface to work with the system, whose characteristics depend on screen size, interaction capability, and integrated sensors of the device in use. In other words, it implements the interface layer discussed earlier. The latter provides storage and processing capabilities and can be accessed on-demand from the client-side. This separation offers two advantages, analysts can work from thin clients, and the server-side can be scaled up or down to cope with processing and storage needs without requiring changes on the client-side.

Based on the previous considerations, the architecture can use two design patterns to divide the system's functionality: layered and client-server. By including an interface layer, the system can implement multiple interfaces specialized for diverse devices, and any required changes are isolated to that specific layer. Further, by separating the system into client and server sides, the heavy storage and processing tasks can be moved to the sever-side, which reduces the need for specialized hardware on the client-side.

3.2.4 Time-critical and long-term analysis

An analysis effort can be classified based on its duration. In this regard, the support for time-critical and long-term analysis scenarios has been argued to be important for (G)VA [2, 79, 156]. In a time-critical scenario, the analysis effort lasts for only a short time, and delivery of timely results are required to minimize undesirable consequences. For example, during a natural disaster, it is crucial to minimize the damage to humans, animals, and the environment, which requires tools for agile planning, awareness of the analysis's progress, and decision-making. In a long-term scenario, the analysis effort extends for a much longer time span and aims to generate understanding and/or enable strategic decisions regarding the phenomena under study. For example, in pest management, it is essential to reduce the impact of control measurements on the agroecosystem, which requires tools to organize and analyze during many years findings regarding the pest dynamics and the effects of human intervention.

To enable a GVA system to support time-critical and long-term analysis efforts, its architecture should be flexible to enable diverse collaboration setups and accommodate specialized functionality for each scenario. This is addressed by the design decisions for hybrid collaboration scenarios and cross-device collaboration. However, there is a scenario in which the criteria outlined so far fall short. During a time-critical analysis effort, there might be a need to deploy new processing functionality or update an existing one, which should occur with minimum disruption of the ongoing analysis effort. To address this, the system's logic might be designed based on the microservices pattern, which has two important advantages: first, faster deployment times due to the small size of the units to be deployed, which minimizes the downtime; and second, the work of analysts who are not using the specific functionality to be updated, will not suffer any disruption. The proposed architecture does not use the microservices pattern for three reasons: first, this scenario is specific to time-critical analysis; second, the complexity introduced by the microservices might not be justifiable for many systems [130, 146]; and third, microservices is still a recent pattern with potentially unknown characteristics [29]. However, a brief discussion on how microservices could fit in the architecture is provided.

3.3 A software reference architecture for collaborative geovisual analytics

This section describes a software reference architecture for CGVA based on the literature review presented in Chapter 2. The architecture is based on the architectural patterns described in Section 3.1 and the design criteria outlined in Section 3.2.

The architecture is described using the "4 + 1" view model [85]. This model proposes to document an architecture with the following (4 + 1) views: *development view*, which represents the programmers' perspective; *logical view*, which represents the functionality provided to the end-users; *process view*, which represents the dynamic aspects of the system such as communication among processes; *physical view*, which represents the deployment of the system; and *scenarios* (the plus one view) to illustrate and validate the architectural design [85]. This model enables a deep understanding of the architecture by providing diverse views with a well-defined target audience to address the concerns of the several stakeholders of a software system.

3.3.1 Development view

The architecture divides the system into client and server sides. Further, it divides the system into five layers: analytical environments, client-side logic, server-side logic, storage, and security. Each layer has a well-defined responsibility within the system and is allowed to use the functionality provided by the layer immediately below and the layer to the right (i.e., security layer). Figure 3.11 shows the components on each layer and the distribution of the layers into the client and server sides.

The analytical environments layer is responsible for the user interface. Its design, implementation, and deployment are device-dependent, and a system may have several implementations depending on the supported devices. To develop an interface for a specific type of device, it is necessary to consider its characteristics, including screen size, interaction capabilities, integrated sensors, storage and processing capacity, and system requirements. The user interface can offer two types of workspace: individual and collaborative, implementing either or both depends on the device's role in the system as shown in Figure 3.12. Both types of workspace are GVA environments, therefore composed of visualization, interaction, and coordination components, as described in Section 3.3.2.

The client-side logic layer provides functionality (to the analytical environments) that does not affect how data is visualized, the interaction possibilities, nor the coordination between visualization components. It enables the client-side to request data from the server and manages the local storage to work as cache memory and support offline work. It also enables multi-synchronous collaboration by providing logic to reconcile and submit contributions and provides client-side logic for the toolbox.

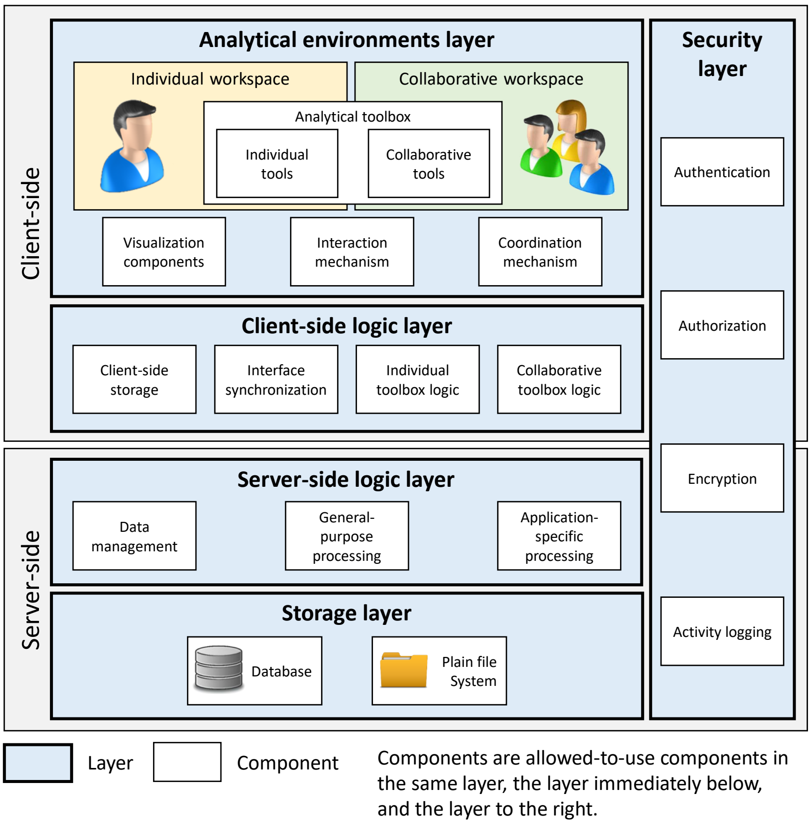


Figure 3.11 Partition of the system into client and server sides, and further into layers.

The server-side logic layer is responsible for enabling the client-side to access the data and provide processing capabilities. The data management component provides basic access routines to the stored data and includes functions to create, read, update and delete data (commonly referred to as CRUD functions). This component should enable uniform access to the data regardless of the underlying technologies on the storage layer. The processing functionality is divided into two categories: general-purpose, which includes processing capabilities of general application such as data aggregation and interpolation; and application-specific, which provides specialized functionality such as an urban-growth or a disease-spreading model, and which depends on the application domain of the system.

The storage layer provides persistence for the system. It may use a plain file system, database technologies such as relational, object-oriented, or document-oriented, and/or distributed file systems such as Google File System (GFS) or Hadoop Distributed File System (HDFS). The specific technology and database design depend on the system's requirements.

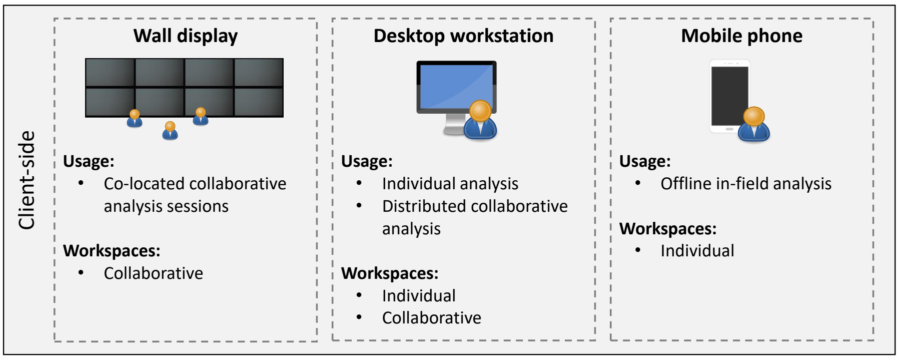


Figure 3.12 An implementation of the user interface might implement either or both workspaces depending on the role of the device within the system.

However, it will include at least security data (e.g., users, roles, and permissions), analytical artifacts (e.g., annotations, snapshots, and messages), and data representing the phenomenon to be analyzed.

Finally, the security layer provides functionality to safeguard the system. This layer provides the mechanisms to identify the user (i.e., authentication), enable him or her to perform actions for which permissions are granted (i.e., authorization), protect communications (if needed) by encrypting them, and keep records of users actions (activity logging).

3.3.2 Logical view

The system's functionality is exposed to the analyst through two types of analytical workspaces: individual and collaborative (See Figure 3.11). These workspaces are GVA environments, and having both aims to provide a flexible workflow that enables the combination of individual and collaborative activities during the analysis effort, which is common in practice [109, 69]. The environments are composed of visualization, interaction and coordination components, and analytical tools, as shown in Figure 3.13.

A visualization component maps one or more dimensions of geodata (i.e., space, time, and attributes) to a visual representation such as a scatter plot, choropleth map, or space-time cube. An interaction component enables a visualization component to react upon the user input. While the interactions depend on the visualization component, the input method depends on the device in use. For example, a map can offer object selection (i.e., an interaction) triggered by a mouse click on a PC or by the tap gesture in a touchscreen. Finally, a coordination component enables the communication between visualization components to coordinate their behavior. Following with the previous example, when an object is selected in the map, it notifies a coordination component, which informs other components about the selection; in response, those other components might highlight the object or show detailed inform-

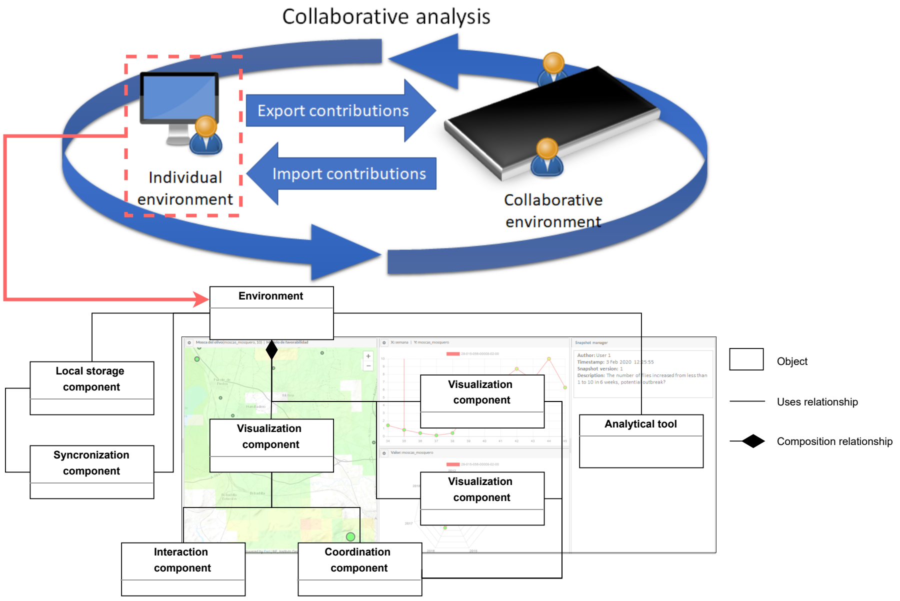


Figure 3.13 Simplified UML object diagram for an individual workspace in a Collaborative Geovisual Analytics system.

ation. These components might behave differently depending on the type of workspace. For example, while a visualization component in the individual workspace only needs to react to interactions from a single user, the same component in the collaborative workspace might need to react to interactions from multiple users working in a collocated or distributed setup. Figure 3.14 shows an example of these components working together. An elaborated discussion and examples regarding a component-based framework (i.e., GAV framework) for the design and development of GVA environments is presented in [59, 88, 60, 72, 71, 70].

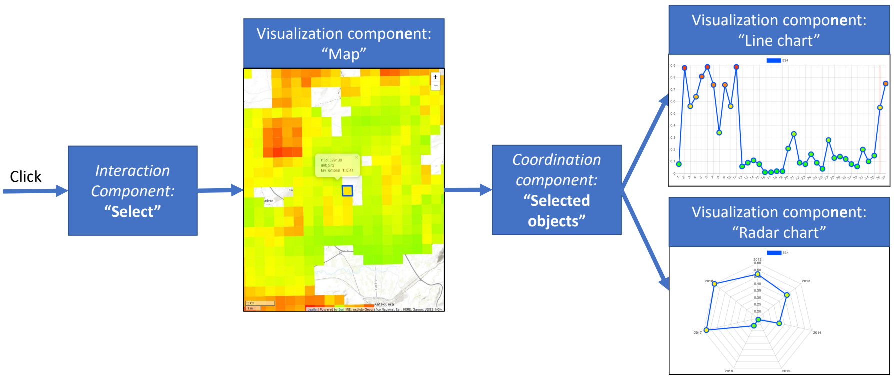

Click
Interaction Component: "Select"
Visualization component: "Map"
Coordination component: "Selected objects"
Visualization component: "Line chart"
Visualization component: "Radar chart"

Figure 3.14 Visualization, interaction and coordination components working together.

The workspaces also provide access to analytical tools. On the one hand, the individual environment enables an analyst to document, review, and extend his or her analytical work. On the other hand, the collaborative environment enables communication, coordination, and decision-making among analysts. Some of these tools produce artifacts that can support both individual and collaborative work; therefore, these artifacts can be imported/exported between the analytical environments. This enables an analyst to share individual findings for collaborative analysis and isolate collaborative findings to continue working on them individually. Additionally, the workspaces provide access to general-purpose and application-specific processing functionality.

3.3.3 Process view

The system enables collaboration among several analysts, who may interact in synchronous, asynchronous, or multi-synchronous fashion. Regardless of the synchronization type, an analyst's contributions are available to others through a shared data storage as shown in Figure 3.15A. Depending on the system's requirements, the synchronization mechanism can implement different behaviors. In synchronous and asynchronous collaboration, the contributions are immediately stored in the shared data storage, but they differ in the time at which other analysts see the contributions. In synchronous collaboration, the contributions are immediately communicated, perhaps by the server-side using push notifications as shown in Figure 3.15B, which would require that the client previously had opened a communication channel to receive notifications from the server. In asynchronous collaboration, the contributions are seen in a delayed manner, either because the analyst was offline when the contributions occurred or because the system uses an on-demand synchronization mechanism, which enables the analyst to decide when to pull updates as shown in Figure 3.15C.

As mentioned before, multi-synchronous collaboration differs from synchronous and asynchronous because the contributions are not submitted immediately to the shared data storage. The delayed submission of contributions may suffer from update (version) conflicts because data and analysis artifacts in the shared data storage may have changed from the time a contribution was made. Hence, the system needs a reconciliation process to solve any conflicting updates before they are submitted. Figure 3.16 shows a reconciliation process in which the analyst interactively corrects conflicting updates.

Regarding the system performance, the analyst might work with the system from diverse devices, and regardless of the device in use, the system should offer a level of performance matching the user requirements. Three design decisions accomplish this: first, the heavy processing takes place on the server-side, which enables the user to work from thin clients; second, the usage of load balancers and redundant servers, which eliminates the bottlenecks and single points of failure; and third, the client-side implements a cache memory, which reduces the need to con-

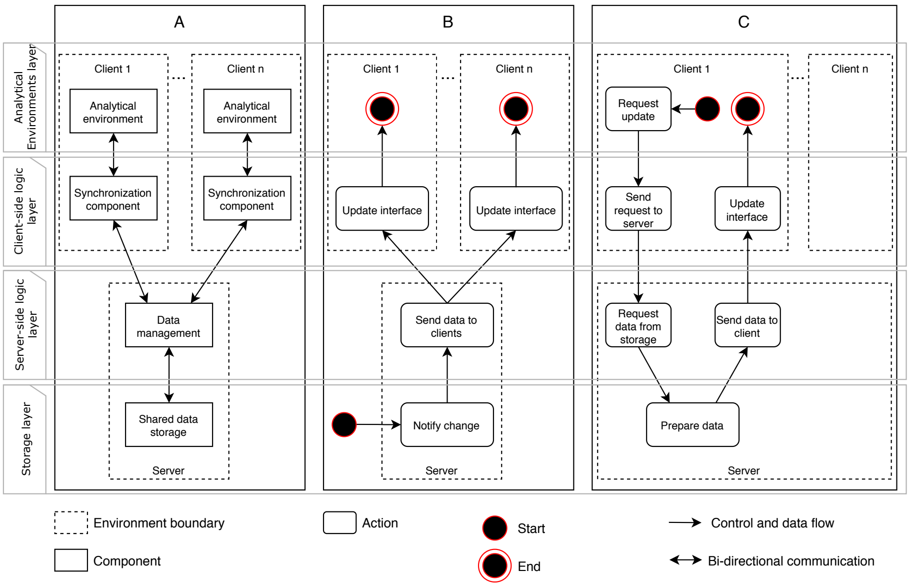


Figure 3.15 A) Analysts contributions are available to others through a shared data storage. B) In synchronous collaboration, the server can push updates to clients as soon as contributions occur. C) In asynchronous collaboration, the clients could pull the updates on-demand.

tact the server to retrieve data. Figure 3.17 shows a potential workflow for the cache memory. When a workspace needs data, it requests data to the interface synchronization component, which provides it to the workspace by retrieving it from the cache memory or the server-side. The retrieved data is stored in the cache memory, which may require to release space by deleting old data.

3.3.4 Physical view

Depending on the requirements for availability and performance, the system might be deployed using different infrastructures. Figure 3.18 shows two different deployment scenarios.

The scenario depicted in Figure 3.18A is typical for non-critical small systems. There are two main problems associated with this infrastructure: First, the servers in both the processing and storage tiers are bottlenecks because their latency affects the system's general performance. Second, those servers are also single points of failure, which means that if either fails, the whole system stops working. The scenario depicted in Figure 3.18B solves both problems. The system can offer better performance by adding multiple processing and storage servers and load balancers to access them. However, with only one load balancer on each tier, the problem of a single point of failure is not solved but moved to a different location (i.e., the load balancer). This can be addressed

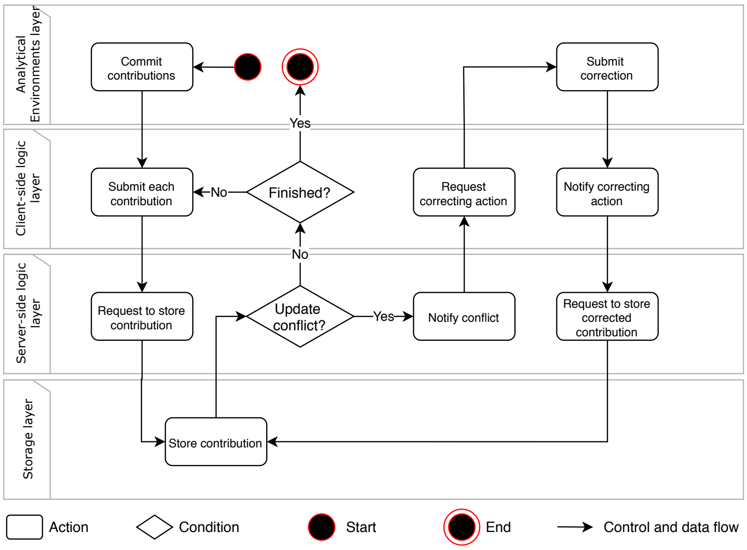


Figure 3.16 Workflow to submit contributions in multi-synchronous collaboration.

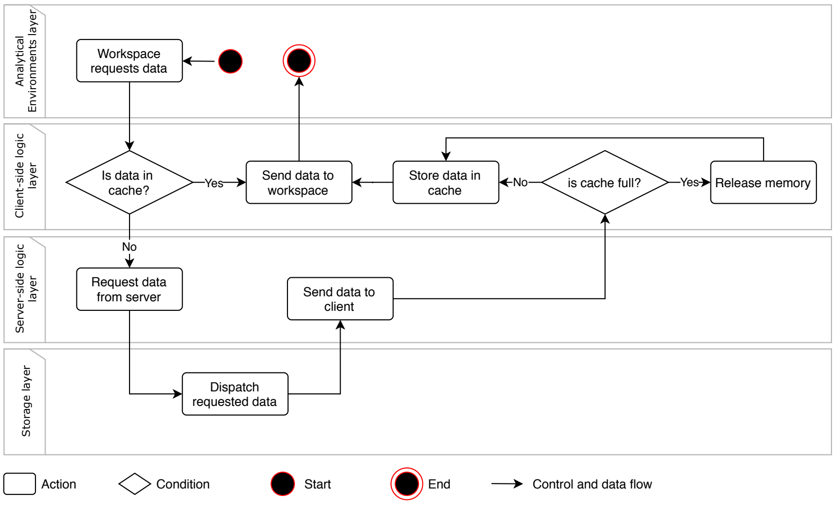


Figure 3.17 Cache memory workflow

by adding spare load balancers that can take over when an active load balancer fails.

The introduction of extra hardware components increases the system's complexity, which involves higher deployment and maintenance costs. For this reason, the performance and availability requirements should be properly analyzed before deciding on the number of replicas for each hardware component. An alternative to reduce the deployment and maintenance costs is cloud technology, which also provides effective

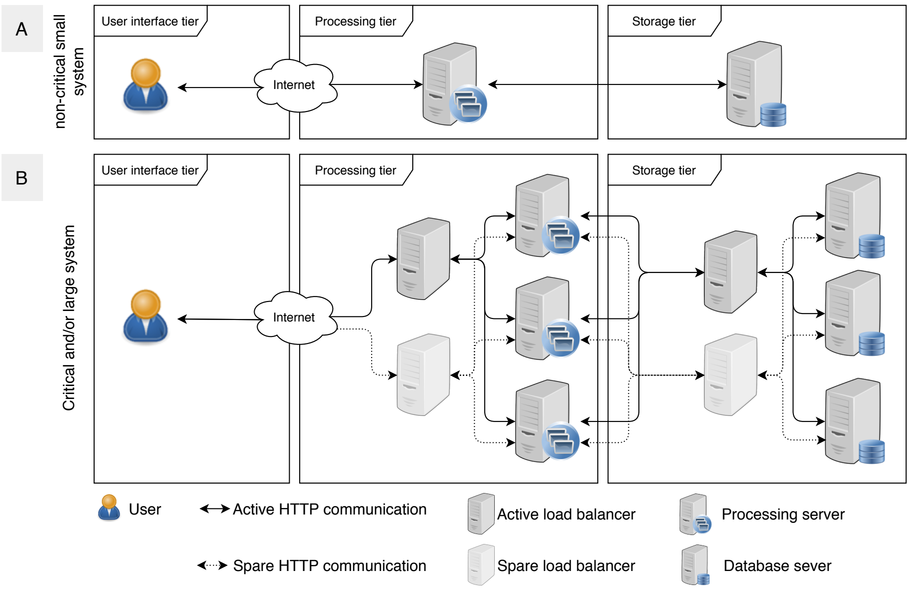


Figure 3.18 A) Infrastructure for a non-critical small GVA system; B) Infrastructure for a critical and/or large GVA system.

mechanisms to scale up and down the system's processing and storage capacity when needed. Additionally, having multiple servers in a tier requires to consider synchronization between those servers.

3.3.5 Scenarios

A relevant scenario for a system based on this reference architecture is that the analyst needs to move back and forth between individual and collaborative work during the analysis effort. These two working modes are enabled by the analytical environments as described before. Figure 3.19 shows a finite state machine representing the actions that can be performed on each working mode and the transitions between them. While working individually, the analyst's contributions remain private until he or she decides to make them available to others. This might require reconciling any conflicting updates, particularly for contributions created in offline individual work. Additionally, the analyst can import contributions that are publicly available to work on them privately. In the collaborative workspace, the analyst can participate in synchronous and asynchronous collaborative activities, and all the generated contributions are immediately publicly available. Given that a device might implement either or both types of analytic environments, as mentioned before, switching between workspaces might imply switching the device.

Another relevant scenario is that the analyst needs to work with the system in online and offline modes. Figure 3.20 shows a finite state machine representing the transitions between these modes. The online

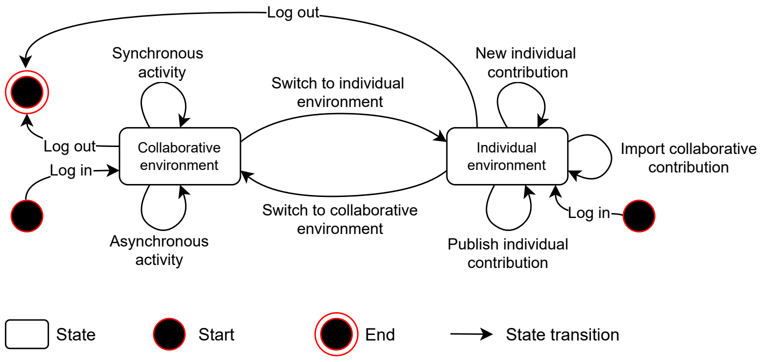

Log out
Synchronous activity
Switch to individual environment
New individual contribution
Log out
Collaborative environment
Log in
Individual environment
Import collaborative contribution
Switch to collaborative environment
Log in
Asynchronous activity
Publish individual contribution
State
Start
End
State transition

Figure 3.19 The analysis effort can combine individual and collaborative work, with each working mode enabling different actions and the possibility to move back and forth between them.

mode enables the analyst to perform individual and collaborative work using the workflow shown in Figure 3.19. While in online mode, the analyst can prepare a device to go offline by pulling the necessary data from the server to the local storage. Once the data is in the analyst's device, it can be disconnected from the server and be used to perform individual analysis. All the contributions are stored in the device's memory during offline work, and therefore, only available to their author. Later, when the device is reconnected to the server, the analyst can decide whether to reconcile or discard the contributions.

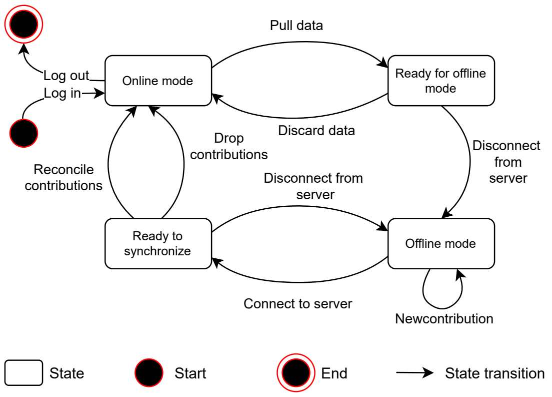


Figure 3.20 The system enables the user to move back and forth between online and offline working modes.

A third relevant scenario is the synchronization across sites and devices. In this scenario, the interface synchronization component plays two roles: first, it submits the contribution created in the user interface to the shared data storage; and second, it updates the user interface when an update is received. Figure 3.21 shows the events on a synchronous collaboration scenario, from the creation of a contribution until it is displayed on other analysts' interfaces.

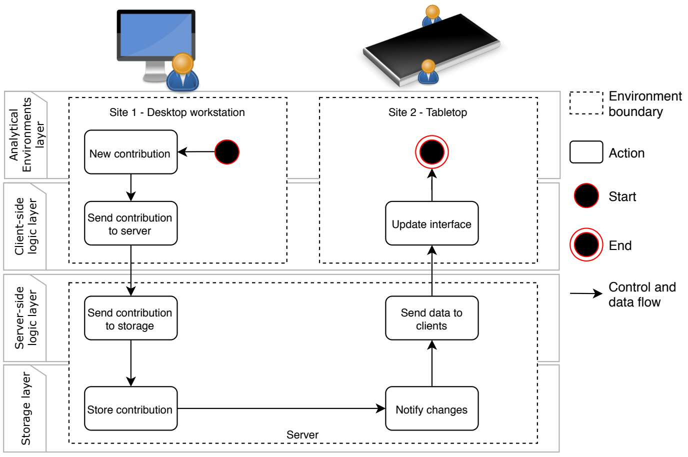


Figure 3.21 Workflow of the propagation of contributions across sites/devices.

3.3.6 Towards a microservices-based architecture

The architecture described in this chapter leads to what is known as a monolith software system, this can be defined as a "software application composed of modules that are not independent of the application to which they belong" [29, p. 1], which means that the modules of the software system are not independently executable. While the combination of client-server and layered patterns enables defining a well-organized architecture that is easy to develop and deploy, it might become hard to scale and maintain as it grows. For example, the time to compile and deploy increases because the whole system is involved in those processes.

Microservices can be used to solve the previously mentioned issues. The microservices pattern provides a system with high flexibility by allowing services to be developed in different programming languages [29, 146], and scalability by allowing to add as many services as needed that can be designed, developed, deployed, and replicated as individual units [29, 130]. However, this pattern comes with the following trade-off: regardless of the number and complexity of the services to be included, the design and development effort and the required infrastructure are more complex than the one for a monolithic software system [29].

The need to support continuous deployments with as little disruption as possible to the analysis process emerged during the analysis of time-critical analysis scenario. The microservices architecture was designed particularly for such scenarios, and in this regard, Dragoni et al. [29] state:

> "Microservices are the first architecture developed in the post-continuous delivery era and essentially microservices are

> meant to be used with continuous delivery and continuous integration, making each stage of delivery pipeline automatic. By using automated continuous delivery pipelines and modern container tools, it is possible to deploy an updated version of a service to production in a matter of seconds, which proves to be very beneficial in rapidly changing business environments" (p. 7).

To map the proposed software reference architecture into a microservices architecture, the functionality provided by the storage and server-side logic layers is divided into small independent units (i.e., the microservices). A unit has a single responsibility; therefore, it implements a small portion of the system's logic and manages a small portion of the system's data. For example, a *Security* microservice offers functionality such as authentication and authorization and manages data related to users, roles, and permissions. The functionality of the analytical tools is divided into different units. For example, one microservice can implement the logic and storage for a snapshot tool and another for storytelling. The same logic applies to processing functionality. Multiple applications can consume these microservices to expose their functionality to the user; therefore, those applications implement the analytical environments and client-side logic layers. This structure is illustrated in Figure 3.22.

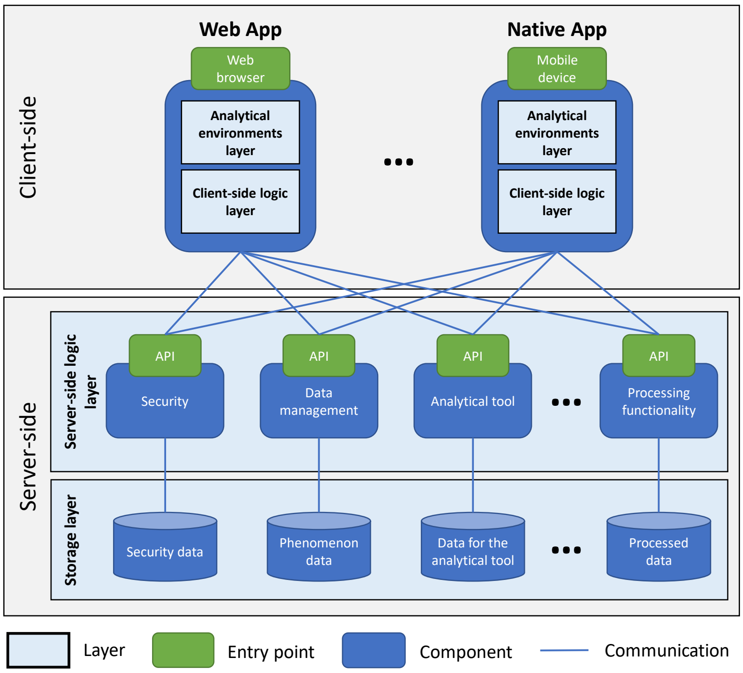


Figure 3.22 Microservices-based software reference architecture

The structure presented in Figure 3.22 is the common approach to microservices. However, this has received some criticism because the user interface is monolithic (sometimes called a *frontend monolith*) [99, 68]; therefore, as it grows, it becomes harder to maintain. An alternative is *micro frontends*, which is “An architectural style where independently deliverable frontend applications are composed into a greater whole” [68]. Like any other pattern, it comes with drawbacks, such as duplication of dependencies due to the independence of the interface fragments and more complex management due to the many small pieces of code to be designed, developed, deployed, and maintained [68]. Figure 3.23 illustrates the concept of micro frontends. An important advantage of this approach is that the development teams can take responsibility end-to-end in designing, developing, and deploying a small portion of the system.

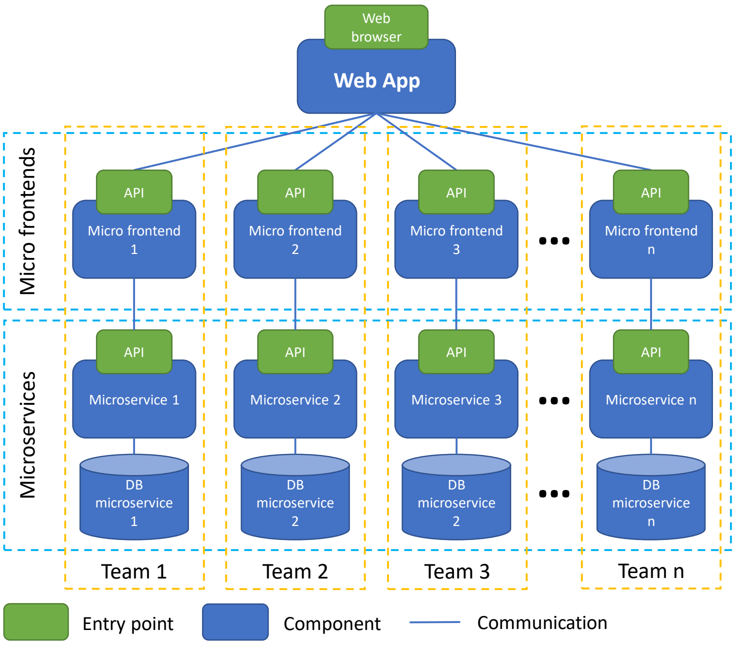


Figure 3.23 The micro frontends pattern. Based on [99]

3.4 Chapter summary

A software architecture is an abstract high-level description of the fundamental structures of a system, their relationships, and properties of both. It provides a reference model for the design, development, and implementation of a software system. Given that unprecedented requirements are uncommon while designing an architecture, reusable solutions for known design problems have been created, called software architectural patterns. Three of those patterns were discussed in this chapter: client-server, layered, and microservices. Software architectures can have different goals and scopes. The one presented in this chapter is a software reference architecture, which means that it is not designed for a specific system, but as guidance for designing architectures for specific systems.

The design criteria for the architecture presented in this chapter is based on the analysis of the research challenges presented in Chapter 2, which are the lack of support for hybrid collaboration scenarios, cross-device collaboration, and time-critical and long-term analysis. Based on the analysis of those challenges, the three architectural patterns mentioned in the previous paragraph were chosen as building blocks for the software reference architecture.

The general structure of the system is divided into client and server sides, where the client-side implements the user interface, and the server-side provides storage and processing power. The system is further divided into five layers: analytical environments, client-side and server-side logic, storage, and security. The rationale for dividing the system into layers is to provide a well-organized structure that facilitates a software system's design, development, implementation, and maintenance. Finally, the architecture can be modified to offer high scalability and flexibility by using the microservices pattern. The trade-off is that a system becomes more complex and expensive to implement.

Spatiotemporal Analysis Space 4

This chapter proposes a novel approach for long-term distributed asynchronous collaborative analysis in GVA environments, called Spatiotemporal Analysis Space (STAS). Thus addressing the third research objective, "Design an approach for collaborative analysis in geovisual analytics environments." The motivation to design this approach is to support long-term analysis processes such as the analysis of pest population dynamics. This application domain is of great importance because a better understanding of the pest population dynamics enables the design of eco-friendly and cost-effective pest control strategies, with direct positive impacts on food security and biodiversity. Additionally, the approach is relevant for other applications that require long-term analysis efforts, such as criminal activity, food production, and monitoring of flora and fauna.

The chapter is organized as follows: Section 4.1 describes the design criteria for the approach; Section 4.2 describes the STAS approach; Section 4.3 describes a mapping of the approach's functionality into the software reference architecture proposed in Chapter 3; and Section 4.4 describes an implementation of the approach.

4.1 Design criteria

The proposed approach was designed to support long-term analysis processes because it is one of the challenges identified in Chapter 2, and the application case of this research (i.e., monitoring and control of the OFF in Andalusia, Spain) requires it. The stakeholders of the application case provided three design requirements for an approach to analyze pest management data, they are: first, to provide a mechanism that enables the analysis of data sets with large spatial and temporal extents; second, to provide a mechanism to support long-term analysis efforts; and third, to provide a mechanism that can rely on multiple collaboration techniques.

Several groundbreaking advances in geospatial technologies such as small GPS-enabled devices, high-resolution remote sensors, and linking of geo-web services have led to more and larger geo-data sets [24, 22, 149]. Some examples are: 150 years of U.S. census data [56]; the OECD regional database from 1960 to the present, containing around 50 indicators for 1700 sub-regions in the 34 OECD countries [71]; the hurricane dataset of UNISYS from 1851 to the present, covering the Atlantic, Indian and Pacific oceans [160]; imagery datasets such as the Landsat archive with global coverage since the early 1970s to the present [108]; the OpenStreetMap dataset which contains a diversity of physical features for the whole world and is regularly updated [114, 113]; or the varied user-generated geographic content available through public application programming interfaces (API) of social media platforms and dedicated citizen science projects. In such data sets, features of interest occur in dispersed locations and times; therefore, these can be defined as data subsets with identifiable boundaries in space and time. The proposed approach enables the creation of artifacts to define those data subsets that constitute features of interest (See Section 4.2.1.1) and enables collaborative analysis of them (See Section 4.2.1.2).

Pest management efforts can extend for several years, during which data is continuously being produced and analyzed; therefore, it is a long-term data analysis scenario. As described in Chapter 2, the support for long-term analysis scenario is of key importance for (G)VA [41, 79, 2, 156]. This analysis scenario is not exclusive of pest management; other applications can benefit or even require a long-term analysis effort, such as criminal activity, food production, and monitoring of flora and fauna. The proposed approach enables the analyst to identify and create artifacts to document, represent and communicate features of interest, which are relevant events within the application domain, for example: in pest management, population outbreaks and collapses; in criminal activity, hotspots of crimes; and in the monitoring of flora and fauna, the occurrence of a species out of its known living environment. Additionally, relationships can be created between related features of interest, which aims to promote the reuse of previously generated findings and to ensure that those remain discoverable in the long term. These links can be created manually (i.e., through human input) or automatically (i.e., from computer inference), as described in Section 4.2.2.

The stakeholders of the application case pointed out that a system should provide tools that are known or easy to learn for the target users, such that the system can be easily adopted. Diverse techniques can support collaborative analysis in GVA environments, for example: annotation, discussion board, instant messaging, interaction history, snapshot, and storytelling [41]. The approach does not depend on a specific collaboration technique; this provides flexibility and facilitates its adoption in diverse domains. Therefore, an implementation of this approach can offer a combination of techniques that are adequate to support collaborative work in the target analysis effort. These tools are available when working with a feature of interest (i.e., data subset), which is described in Section 4.2.1.

4.2 Approach's description

The STAS approach proposes a means to perform long-term distributed asynchronous collaborative analysis of spatiotemporal data. It is not tailored for a specific GVA system; therefore, diverse systems can implement it, including variations to fit within the system's application domain. This section describes the general conception of the approach, and Section 4.4 describes an implementation in the context of the application case of this research.

The design of STAS is based on the principles of Distributed Cognition (DC), which provides a framework in which cognition is conceived as a social process, involving many human actors (i.e., the analysts) as thinking entities, and artifacts as means for knowledge exchange and shared memory [152, 61]. Examples for artifacts are the representation for features of interest, relationships (or links) between them, and contributions to make sense of those features of interest, such as snapshots and messages. The creation of those artifacts does not require analysts to synchronize in space or time; therefore, collaboration occurs in a distributed and asynchronous manner. Additionally, DC recognizes that cognition can be distributed over time, and therefore, provides theoretical support for long-term cognitive systems [61]. In this context, the creation of artifacts for features of interest and contributions enables externalization and communication of analytical findings, and the links between features of interest enable those contributions to remain easily discoverable as the analysis effort advances, which facilitates the long-term analysis processes.

The main assumption in the design of STAS is that in data sets with large spatial and temporal extents, features of interest such as patterns (e.g., pest population outbreak) and outliers (e.g., unrealistic high pest population abundance) occur in diverse locations and times. In this context, the central concept of STAS is the *analysis space*, which is a container for a feature of interest and analytical contributions to make sense of it. To define an analysis space, an analyst provides a spatiotemporal boundary for the data subset representing the feature of interest and a description. The analysis space aims to focus the analysts' attention on a feature of interest to elicit sensible contributions and generate meaningful knowledge.

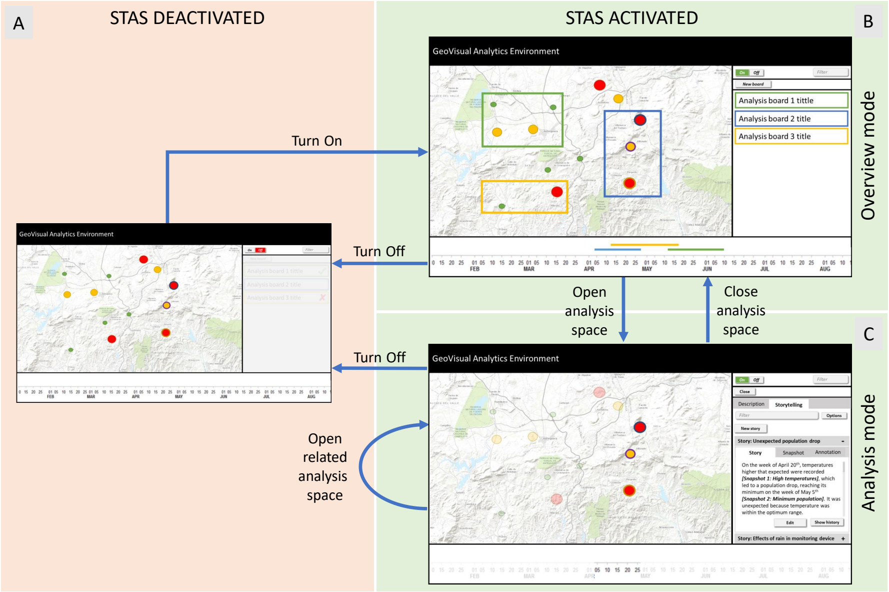
The analysis spaces can be linked to one another, whose purpose is twofold: to provide relevant information for the analysis spaces, to promote building knowledge upon previous contributions; and to enable navigation based on the identified features of interest.

Let us elaborate on those links in the context of pest management. Regarding the provision of relevant information, during data exploration, an analyst may observe a sudden increase or decrease of the target species abundance, which might be relevant for the analysis effort and lead to creating an analysis space. Based on the data subset which defines the feature of interest and its description, a processing engine that implements the automatic identification of related analysis spaces (described in Section 4.2.2.2) may provide potentially relevant information even before any contribution occurs in the newly created analysis space. The potential of this approach to identify and offer relevant information to the analyst improves as more features of interest are identified and analyzed. Regarding the exploration based on features of interest, the analyst might be interested in exploring the occurrence of pest outbreaks. By filtering the analysis spaces based on keywords related to them, the analyst can get an overview of the spatiotemporal distribution of the outbreak occurrences, explore an analysis space (i.e., an outbreak occurrence), and use the relationships to move between the related analysis spaces. The creation of links and how their relevance is determined is explained in Section 4.2.2.

The analysis spaces are represented by overlaying their spatiotemporal boundaries in the visual components of the GVA environment. For example, in a space-time cube combining the spatial and temporal extents, or in a map component for the spatial extent and a timeline component for the temporal extent. This visual representation may disrupt the typical analysis workflow of the host environment. To address this, STAS may be activated and deactivated as per user convenience. In this sense, the STAS can be in three states: *deactivated*, in which case, it does not display any information on the user interface and thus does not affect the typical analysis workflow. The other two states correspond to the working modes: overview and analysis. In the *overview mode*, a list of analysis spaces and their spatiotemporal boundaries are displayed. In the *analysis mode*, a feature of interest (i.e., data subset) is highlighted, and collaboration tools are available to analyze it. Section 4.2.1 describes the working modes.

Figure 4.1 illustrates the three states and the events that trigger a change from one to another. This illustration depicts a simple user interface to highlight the information displayed in the map and timeline components. An actual interface may include several visualization components, and in the analysis mode, the feature of interest would be highlighted in all of them. Additionally, while the illustration uses a very simple dataset for simplicity, the approach was designed to work with complex datasets, which is also the reason to organize the analysis effort around features of interest. Finally, in contrast with the illustrated examples, a feature of interest might have complex boundaries.

The workflow of STAS is: explore the data set, identify features of interest, create analysis spaces and links between them, and analyze the features of interest collaboratively. This workflow was designed based on the visual information-seeking mantra “overview first, zoom and filter, then details on demand” [144, p. 336], and the visual analytics mantra “analyze first, show the important, zoom, filter and analyze further, details on demand” [80, p. 82]. This workflow combines individual and collaborative analysis activities, which is common in practice [69]. Given that both types of activities occur in the same GVA environment, the workflow offers seamless integration of individual and collaborative



On the week of April 20th, temperatures higher that expected were recorded ***[Snapshot 1: High temperatures]***, which led to a population drop, reaching its minimum on the week of May 5th ***[Snapshot 2: Minimum population]***. It was unexpected because temperature was within the optimum range.

Figure 4.1 A) When the STAS is deactivated, no information related to it is displayed in the user interface. B) In the overview mode, a list of analysis spaces is shown together with their spatiotemporal boundaries. C) In the analysis mode, the feature of interest is highlighted in all the visualization components, and analytical tools are available to analyze the data.

activities.

4.2.1 Working modes

The following sections explain the specific characteristics and role of each working mode (i.e., overview and analysis). For the sake of clarity, the working modes are illustrated using sketches of simple interfaces.

4.2.1.1 Overview mode

The *overview mode* enables the analyst to explore the whole data set and the existing analysis spaces. As mentioned before, the STAS approach is a means to perform long-term collaborative analysis, which can be implemented in diverse GVA systems. Therefore, depending on the application domain of the system, there is a broad range of well-established exploratory tools that can be available in this mode. These can include many visualization products such as maps and charts, tools to select, filter, and brush the data, options to configure the visual representation schema, and multiple linked views. Additionally, in this mode, the analyst can identify features of interest and create analysis spaces for them.

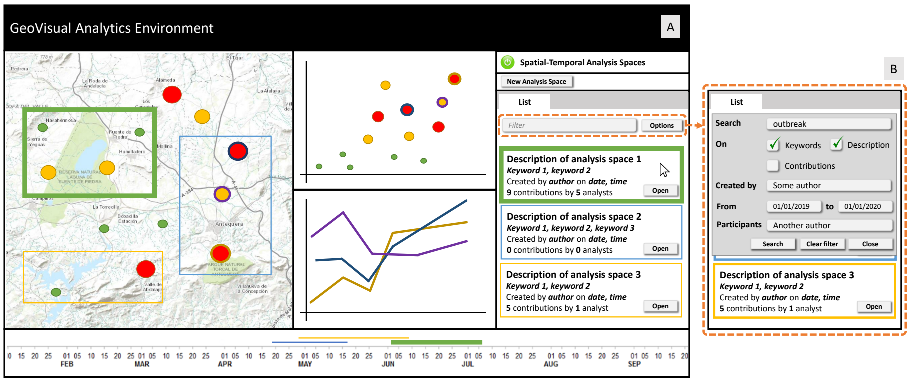
The overview mode displays a list of the existing analysis spaces, which are visually represented using the basic graphic variables (i.e., size, value, grain/texture, color, orientation, shape) [14] to distinguish between them[1]. Despite that a large variety of unique visual representations can be created by combining the graphic variables, due to the limitations of human perception, as the number of analysis spaces increase, it becomes harder for the analyst to distinguish between them based on their visual representation. The list displays for each analysis space: description, keywords, author, creation timestamp, and the number of contributions and contributors. Additionally, the spatiotemporal boundaries are displayed using the same visual representation from the list, which facilitates relating the analysis spaces' thematic, spatial, and temporal dimensions. This system feature improves the analyst's awareness regarding the progress of the analysis effort by providing an overview of the locations and times at which features of interest existed and how much attention they have received, measured by the number of contributions and contributors.

In the overview mode, if the mouse cursor is placed over the analysis space's thematic, spatial, or temporal representation, it is highlighted in all the views (i.e., coordinated linked views). There are several options to highlight an analysis space. For example: increase the line thickness of its representation, and reduce that of the others; or apply partial transparency to all the data outside the boundaries of the selected analysis space (depending on the amount of data being displayed, this option might be computationally expensive). Given that the analysis spaces can overlap or even be nested (as described later in this section), several analysis spaces might be highlighted simultaneously. Figure 4.2A illustrates the list and an example of a highlighted analysis space.

The analysis spaces aim to organize the long-term analysis process. However, as their number grows, it can become harder for an analyst to find a specific one. To address this, the analysis spaces can be automatically filtered to show only those within the spatial and temporal extents displayed on the visualization components. Additionally, they can be filtered out thematically, in which case, a search criterion defined by the analyst is matched with the description, keywords, author, creation timestamp, and/or participants of the analysis spaces, as shown in Figure 4.2B.

If a feature of interest is identified during data exploration, a new analysis space can be created to analyze it collaboratively. An analyst can define an analysis space by providing: the spatiotemporal boundary for the feature of interest (which can be selected from components where the spatial and temporal dimensions are represented); a description and keywords to explain why it is considered a feature of interest; and a visual representation to distinguish it from others. Additionally, the author and creation timestamp are recorded in the definition of the analysis space.

[1] For simplicity, the examples in this thesis only use the color variable.



Description of analysis space 1
Keyword 1, keyword 2
Created by ***author*** on ***date, time***
9 contributions by **5** analysts

Description of analysis space 2
Keyword 1, keyword 2, keyword 3
Created by ***author*** on ***date, time***
0 contributions by **0** analysts

Description of analysis space 3
Keyword 1, keyword 2
Created by ***author*** on ***date, time***
5 contributions by **1** analyst

Search: outbreak
On: ☑ Keywords ☑ Description □ Contributions
Created by: Some author
From: 01/01/2019 to 01/01/2020
Participants: Another author

Description of analysis space 3
Keyword 1, keyword 2
Created by ***author*** on ***date, time***
5 contributions by **1** analyst

Figure 4.2 A) The overview mode uses a visual scheme to facilitate relating the thematic, spatial and temporal dimensions of the analysis spaces. In the illustration, the analysis space represented in green color is highlighted because the mouse cursor is over it on the list of analysis spaces. B) Example of filtering options for the list of analysis spaces.

The spatiotemporal boundary is an artifact that can be intersected with the dataset to produce a data subset (i.e., data that defines the feature of interest); therefore, how it is defined and represented depends on the type of data for the application domain. For example, suppose the dataset is a data cube, in which two dimensions represent space, and the other one represents time. In that case, the spatiotemporal boundary can be a 3-dimensional object such as a cube or cylinder. In this regard, the illustrations presented in this thesis are inspired by its application case. All the examples use a rectangular boundary for the spatial dimension; however, it can be as complex as needed to define the feature of interest. For example, if the feature of interest is the occurrence of pest outbreaks happening simultaneously, but in geographically separated areas, the spatial boundary can be a multi-polygon. Regarding the temporal boundary, all the examples use a time-span with one starting and one ending point in a time continuum; however, the temporal boundary can include several time-spans, if necessary. For example, to represent a cause-effect phenomenon, such as optimum weather conditions for a pest development followed by an outbreak, where the two events might be several weeks apart. The STAS can work with complex spatiotemporal boundaries. However, it might not be used frequently because the analyst would be required to know in advance of a spatial or temporal pattern in the phenomena, which will probably emerge as a result of the analysis effort and will be documented by the analysis spaces and their relationships.

The boundaries of the analysis spaces can overlap or even be nested. There are different causes for this, such as differences in the analysts' definition for the features of interest or the scale at which those features are defined. When this happens, links are automatically created and tagged with the type of relationship, such as 'overlapped' and 'nested,' which aims to facilitate identifying potentially relevant information to analyze the features of interest. Details on the relationships between analysis spaces are provided in Section 4.2.2.

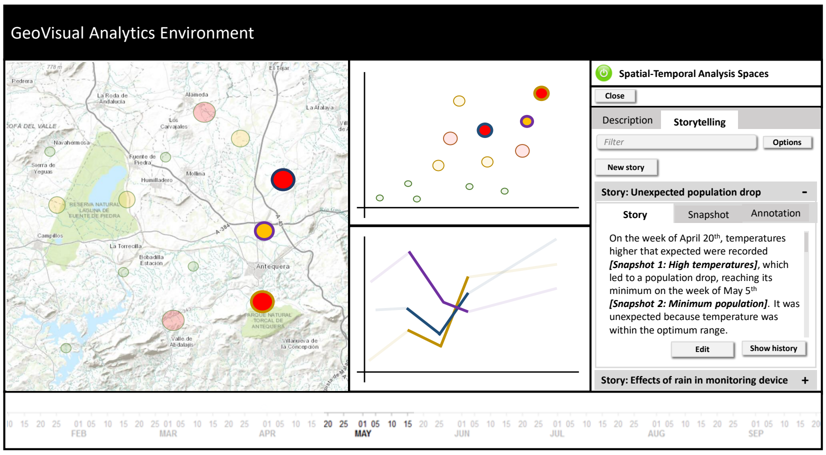
By default, the analysis spaces are available to all the analysts, which ensures awareness of others' work and avoids duplicated efforts. However, there might be cases where it is convenient to restrict analysts' participation in particular analysis spaces. To address this, the STAS approach can rely on a control mechanism to grant and revoke access permissions to the analysis spaces based on the users and roles of the system. Restricting participation might generate fragmentation of the known information, which affects the collaborative analysis process. Therefore, it should be considered only when legitimate reasons exist to restrict participation. In such cases, it might be convenient at least to let non-authorized user to know about the existence of the restricted analysis spaces, which might avoid duplicated efforts.

4.2.1.2 Analysis mode

The *analysis mode* enables collaborative analysis of the features of interest. To this end, the data subset is highlighted by applying partial transparency to all the data outside of the analysis space boundary. It aims to focus the analysts' attention on the feature of interest while keeping it within its spatiotemporal context. Additionally, this mode offers tools such as annotation, discussion board, instant messaging, interaction history, snapshot, and storytelling, enabling externalization and communication of findings among analysts to support collaborative analysis of the feature of interest.

To illustrate the analysis mode, Figure 4.3 depicts an environment that offers a combination of storytelling, snapshot, and annotation, which in practice is a common combination of tools [41]. This combination allows the analysts to describe the analysis process through a 'story' [72] and support their claims using snapshots to document findings and annotations to highlight and describe specific aspects of the data [162]. An advantage of storytelling is that once a conclusion is reached, the story can be communicated immediately to a broader audience. In this sense, Authors agree that results presented with storytelling are more effective, engaging, and easy to understand for specialists and laypersons [44, 148, 151]

A data set is commonly evolving either because new data is added or existing data is updated. The latter situation might change a feature of interest and/or invalidate contributions. Therefore, when changes occur to an analysis space's underlying data, the system should request the analysts to assess whether the changes affect the feature of interest and/or the contributions and, if so, take action to correct them, which aims to maintain the validity of the analysis. Detecting changes on the data can be addressed by computing and storing hash values for the analysis spaces. In cryptography, hashing is used to map data of arbitrary size to fixed-sized values, which can be used to confirm that


GeoVisual Analytics Environment
Spatial-Temporal Analysis Spaces
Close
Description
Storytelling
Filter
Options
New story
Story: Unexpected population drop
-
Story
Snapshot
Annotation
Edit
Show history
Story: Effects of rain in monitoring device
+
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
Antequera

On the week of April 20th, temperatures higher that expected were recorded ***[Snapshot 1: High temperatures]***, which led to a population drop, reaching its minimum on the week of May 5th ***[Snapshot 2: Minimum population]***. It was unexpected because temperature was within the optimum range.

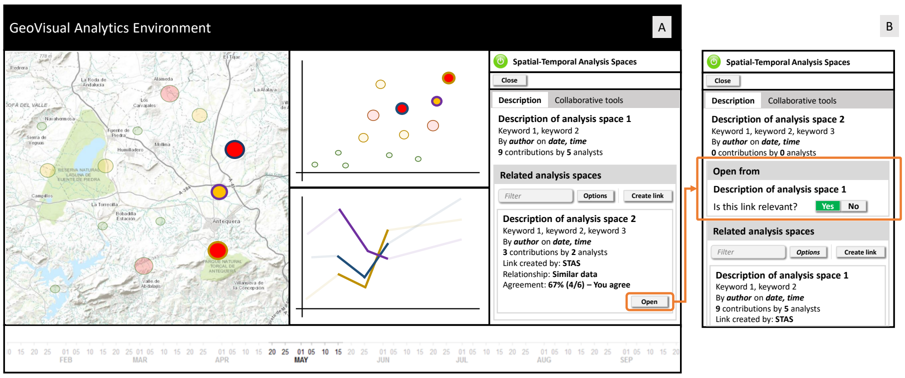
Figure 4.3 The STAS approach can offer diverse collaboration techniques. This example illustrates a combination of storytelling, snapshot and annotation.

data is unchanged [42, 23]. STAS can detect changes by computing the hash value for the data in an analysis space and comparing it with the stored value; if they are different, the data has changed.

Additionally, in the analysis mode, a list of related analysis spaces is available. This list is built from the links between analysis spaces. For each related analysis space, it includes: the general information of the analysis space, indicates whether the link was created manually or automatically (See Section 4.2.2), the type of relationship (i.e., the link's tag, also described in Section 4.2.2), and the level of agreement among analysts regarding whether the link is relevant or not, as illustrated in Figure 4.4A. This list offers the same search functionality as the one in overview mode, and it also allows to search by type of relationship. The level of agreement is based on analysts' votes in favor or against the link's relevance; in the example, four out of six votes (i.e., analysts) are in favor of the link's relevance. The option to vote is available when an analysis space is open from the list of related ones, as shown in Figure 4.4B. The level of agreement can indicate how likely it is that two linked analysis spaces provide relevant information to one another.

4.2.2 Identifying related analysis spaces

While analyzing a feature of interest, an analyst can benefit from information such as where and when the same type of feature had occurred, under which conditions it happened, and the previous analytical contributions to make sense of it. For example, if the feature of interest is a pest outbreak, the analyst can benefit from information about other occurrences of such events. In a more general sense, two analysis spaces can provide relevant information to make sense of each other based on different types of relationships. For example, relevant knowledge can



Spatial-Temporal Analysis Spaces

Description of analysis space 1
Keyword 1, keyword 2
By ***author*** on ***date, time***
9 contributions by **5** analysts

Related analysis spaces

Description of analysis space 2
Keyword 1, keyword 2, keyword 3
By ***author*** on ***date, time***
3 contributions by **2** analysts
Link created by: **STAS**
Relationship: **Similar data**
Agreement: **67% (4/6) – You agree**

Spatial-Temporal Analysis Spaces

Description of analysis space 2
Keyword 1, keyword 2, keyword 3
By ***author*** on ***date, time***
0 contributions by **0** analysts

Open from

Description of analysis space 1
Is this link relevant? Yes No

Related analysis spaces

Description of analysis space 1
Keyword 1, keyword 2
By ***author*** on ***date, time***
9 contributions by **5** analysts
Link created by: **STAS**

Figure 4.4 A) The analysis mode includes a list of related analysis spaces, which allows to navigate between them. B) When an analysis space is opened from this list, the analyst can indicate whether or not the link is relevant.

be obtained from analyzing the conditions that lead to opposite events, such as pest population outbreaks and collapses. In the former example, the relationship could be called 'same type of event,' and in the latter 'opposite type of event.'

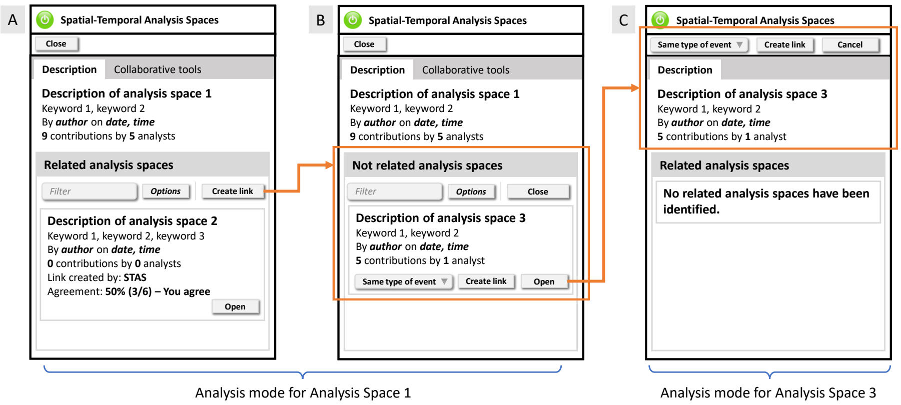
The STAS approach allows creating relationships between analysis spaces, enabling the analyst to easily find relevant information within the system and facilitate building knowledge upon previous analytical contributions. As mentioned in the previous paragraph, there can be different types of relationships. Therefore, there is a need for a mechanism to convey the semantics of those relationships. To this end, the approach relies on labels that define the type of relationship. While there might be several domain-specific labels, some can be of general application, such as: overlapped, nested, similar data, similar description, similar contributions, same type of event, opposite type of event, and cause-effect.

The links between analysis spaces are created either manually by an analyst who thinks that two analysis spaces contain related information as described in Section 4.2.2.1, or automatically by the system as described in Section 4.2.2.2.

4.2.2.1 Manual identification

Analysts can manually create links between analysis spaces. The rationale for this option is that VA relies strongly on human perception and cognition because of its potential to analyze and solve highly complicated problems in an intuitive way [59, 47]. In this context, analysts have expertise and knowledge that allow them to identify analysis spaces that can provide relevant information to understand each other.

To manually create links between analysis spaces, the approach includes in the analysis mode an option to show all the analysis spaces that are not related to the one currently open, as illustrated in Figures 4.5A and 4.5B. This list offers the same search functionality as the one in overview mode. From this list, the analyst can either create the link directly or open the analysis space in view-only mode, and decide whether to link it or not, as illustrated in Figure 4.5C. If an analyst creates a relationship, it is reasonable to assume that he or she agrees with its relevance; therefore, by default, the analyst’s vote is in favor of the relationship’s relevance.



Description of analysis space 1
Keyword 1, keyword 2
By ***author*** on ***date, time***
9 contributions by **5** analysts

Description of analysis space 2
Keyword 1, keyword 2, keyword 3
By ***author*** on ***date, time***
0 contributions by **0** analysts
Link created by: **STAS**
Agreement: **50% (3/6) – You agree**

Description of analysis space 3
Keyword 1, keyword 2
By ***author*** on ***date, time***
5 contributions by **1** analyst

No related analysis spaces have been identified.

Figure 4.5 Creation of links between analysis spaces by an analyst.

4.2.2.2 Automatic identification

The links between analysis spaces can also be created automatically by STAS. For this aim, implementations may rely on approaches such as similarity of textual content and/or spatiotemporal data similarity.

Regarding the similarity of textual content, two common types of metrics are semantic similarity and semantic relatedness. These measures have been subject to intensive research efforts and are central to many Natural Language Processing applications [49]. Although the two are sometimes confused with each other, they are distinct: while semantic similarity measures how similar in meaning two pieces of text are, semantic relatedness can measure diverse types of relationships between them [40]. For example, if we compare the words ‘pest’ and ‘monitoring,’ a semantic similarity measure will indicate that they are different in meaning. However, a semantic relatedness measure (tailored for this purpose) will indicate that they are related because monitoring is an activity of pest management. An implementation of the approach can use the textual content of the analysis spaces, including description, keywords, and contributions from the collaboration tools, to compare the semantic similarity/relatedness of two analysis spaces and automatically link them.

The semantic measures rely on data sources such as text corpora and knowledge models [49]. The former consists of unstructured or semi-structured texts such as plain text and dictionaries and evidence extraction can be based on co-occurrence of terms. The latter consists of structured sources such as ontologies, where the terms and relationships are represented explicitly.

There are several approaches available for semantic analysis, such as Word2Vec [100, 101] and Bidirectional Encoder Representation from Transformers (BERT) [28]. The Word2vec approach, as its name might suggest, maps words (from a large corpus of text) to vectors (word embeddings) that capture the words' semantics and relationships. Those vectors can be used to compare similarity between words or texts using conventional semantic measures (e.g., cosine similarity) [7]. Some words can have different meanings depending on the context; for example, 'bank' can refer to a financial institution or the land along a river. The Word2vec approach cannot capture diverse contexts for a single word. Additionally, it does not support out-of-vocabulary words [45]. BERT addresses both limitations of Word2vec. However, it is computationally intensive, and at present, it is hard to implement in production systems. The reason is that BERT's vectors are dynamic; in other words, they are recomputed to create contextualized vectors [73]. These approaches are adequate to implement the automatic creation of relationships based on semantic analysis.

The STAS approach can also rely on spatiotemporal data similarity (also known as geospatial data matching) to determine whether two analysis spaces contain similar features of interest. This approach relies on metrics that compare several spatiotemporal characteristics of the data sets to determine how similar they are [169, 98]. To calculate the similarity between analysis spaces, their spatiotemporal boundaries define the data subsets to be compared. Given that the type, format, and characteristics of the data are application-dependent, the functionality to create links automatically is also application-dependent. Therefore, its implementation will change from one system to another.

4.3 Mapping the STAS approach to the software architecture

The design of the STAS approach is generic and independent from any software architecture. Therefore, it can be integrated into existing and new GVA environments based on different architectures. In this section, the approach is mapped to the software reference architecture proposed in Chapter 3, which aims to illustrate a possible distribution of the approach's functionality into architectural components.

The three states of the STAS approach (i.e., deactivated, overview, and analysis) can be mapped to the workspaces of the software architecture. Deactivated and overview are mapped to the individual workspace and analysis to the collaborative workspace. Figure 4.6 illustrates this mapping and the events that can occur in each state. The STAS states naturally map to the architecture workspaces because both are designed under the premise that actual analysis processes benefit from a combination of individual and collaborative work [69].

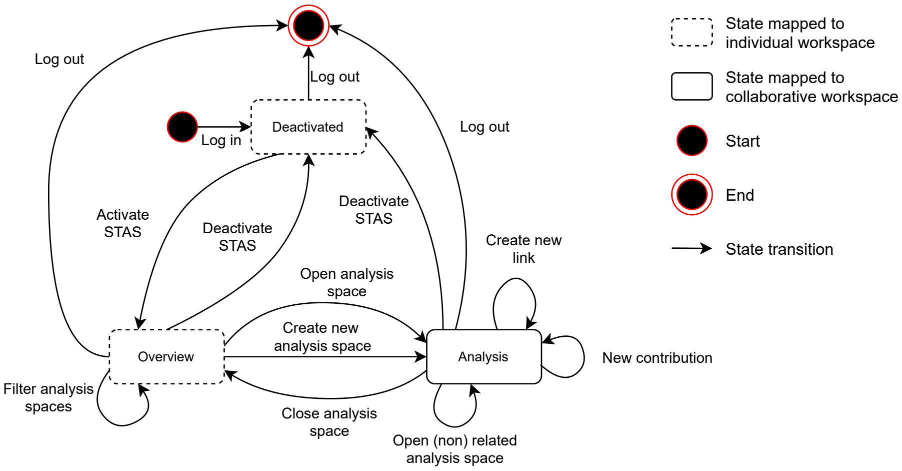


Figure 4.6 Finite state machine representing the transitions between the STAS states, and the mapping of those states to the software architecture workspaces.

The software architecture divides the system into client and server components and further into five layers: analytical environments, client-side and server-side logic, storage, and security. Figure 4.7 shows the distribution of the STAS functionality over those layers.

The STAS interface has components in both the individual and the collaborative workspaces of the analytical environments layer. The overview mode is integrated into the individual workspace and allows the analyst to explore the whole dataset and visualize the existing analysis spaces. The analysis mode is integrated into the collaborative workspace, which enables to explore a feature of interest and perform analysis through diverse collaboration techniques.

The function to filter the analysis spaces can be located in the client-side and/or server-side logic layers. It depends on whether this functionality would affect only the data that is loaded on the client-side or the whole dataset. As mentioned in Chapter 3, it is not practical or even feasible in most cases to load all the data. Consequently, an implementation of the approach might assign this functionality to either or both of the logic layers, depending on the specific user requirements. While working over the whole dataset might be more useful, it is also more computationally demanding; therefore, a reasonable solution is to implement both and let the analyst use them as needed.

The client and server sides need to exchange data to create, read, update, and delete the analysis spaces. This functionality includes the *Data exchange* component, which takes care of sending requests to the server-side and of handling the received responses, thus providing the link between the visual interface and the server-side. Additionally, it includes

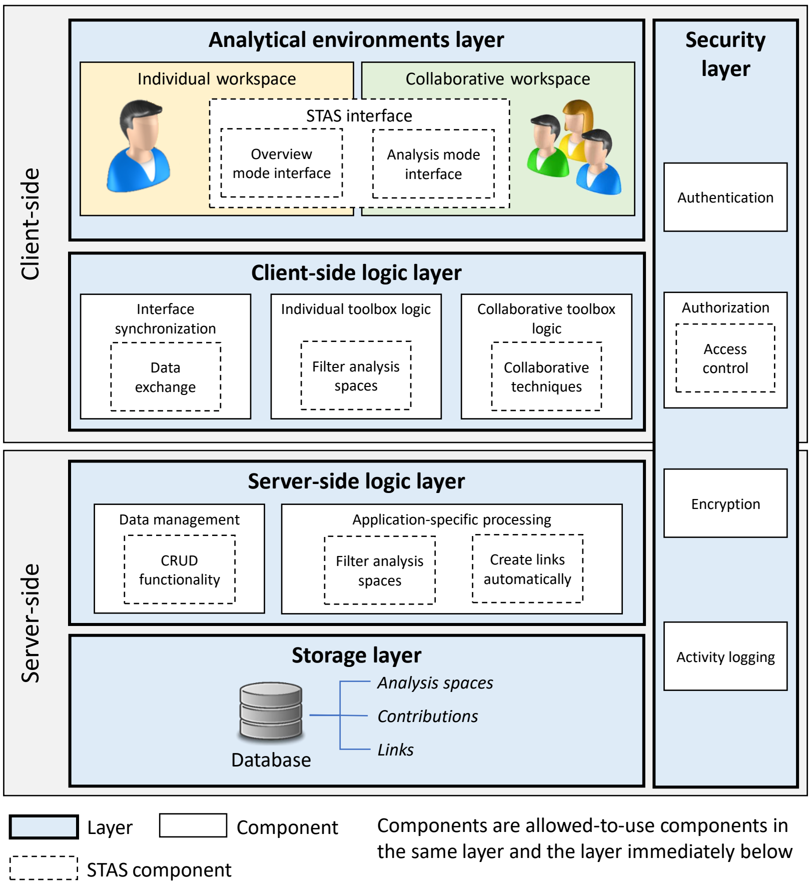


Figure 4.7 Mapping of the STAS approach to the software architecture proposed in Chapter 3.

the *CRUD*[2] *functionality* component, which manages the persistent data, thus linking the client-side with the storage layer.

The client and server sides also need to exchange data about the contributions generated by the diverse collaboration techniques implemented in the analysis mode. The interface for the collaboration techniques is implemented by the *Analysis mode interface* component, and the logic is implemented in the *Collaboration techniques* component, with the latter being in charge of the data exchange between the client and server sides. The logic to automatically create links between the analysis spaces can be computationally expensive and requires working with all the system's data. Therefore, it is located on the application-specific processing component of the server-side logic layer. The *Create links automatically* component may implement functionality to identify related analysis spaces based on text and/or geodata similarity, as discussed before. This

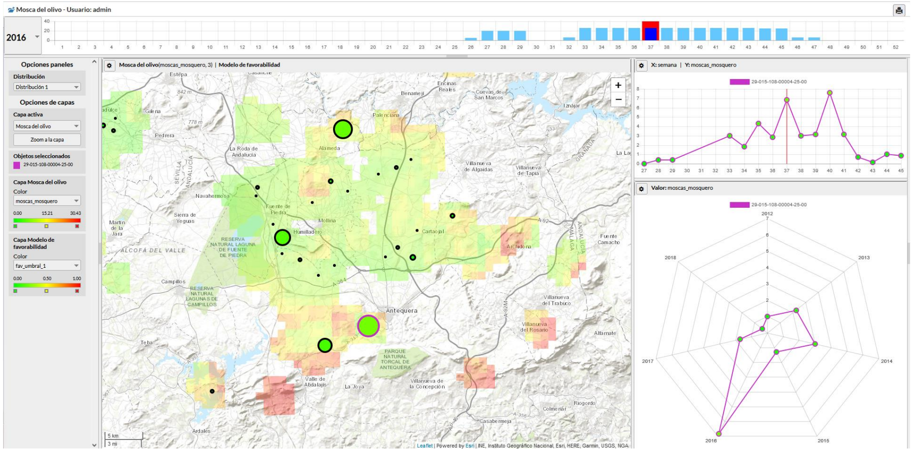
[2] CRUD is a common term in informatics, meaning Create, Read, Update, and Delete.

component works in the background, and it is triggered by changes to an analysis space's underlying data, and when an analysis space or a contribution is created or edited.

Finally, the storage layer provides persistent storage for the analysis spaces, the links between them, and the contributions created with the collaboration techniques.

4.4 An implementation of the STAS approach

As a proof of concept, the STAS approach was implemented in a web-based GVA prototype. The prototype was developed to analyze pest management data of the OFF in Andalusia, Spain. The prototype works with data created by the monitoring and control activities, represented as proportional circles, and with data from statistical models, represented as a regular grid. Stakeholders' requirements guided the design and development of this prototype; details are provided in Section 5.3. Figure 4.8 shows the interface of the prototype when the STAS approach is not in use.


Mosca del olivo - Usuario: admin

2016

Opciones paneles

Distribución: Distribución 1

Opciones de capas

Capa activa: Mosca del olivo

Zoom a la capa

Objetos seleccionados: 29-015-108-00004-25-00

Capa Mosca del olivo

Color: moscas_mosquero

0.00 15.21 30.43

Capa Modelo de favorabilidad

Color: fav_umbral_1

0.00 0.50 1.00

Mosca del olivo(moscas_mosquero, 3) | Modelo de favorabilidad

X: semana | Y: moscas_mosquero

Valor: moscas_mosquero

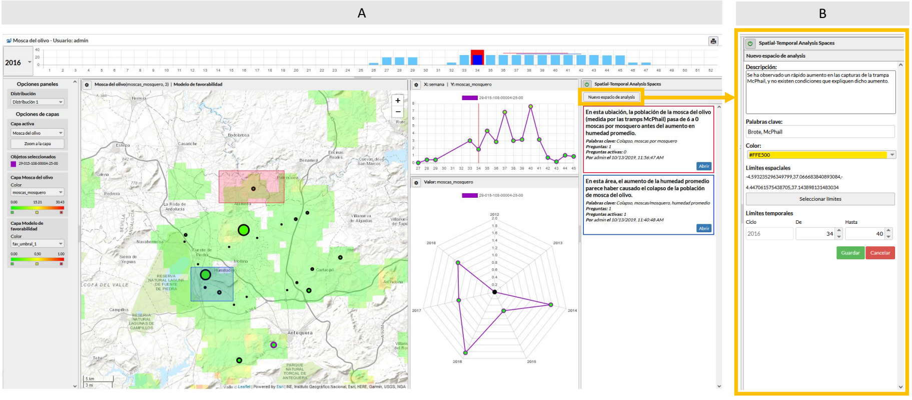
Figure 4.8 The prototype's interface when the STAS approach is deactivated and hidden.

Before implementing the approach in the prototype, it was discussed with the stakeholders. During these discussions, they requested to keep the implementation as simple as possible. The provided requirements led to the implementation of a simplified version of the approach. Specifically, the requests can be summarized as follows:

1. To use rectangular spatial boundary, which should be defined by clicking in two locations of a map.
2. To use a continuous temporal boundary, which should be specified by indicating a starting and ending week.

3. The relationships between analysis spaces should be created by matching keywords, such that it will be clear why two analysis spaces are related.
4. The collaboration tool to analyze the features of interest should be a question-based forum.
5. All analysis spaces, questions, and answers should be accessible to all the system users.

During testing, the stakeholders indicated that it is easy to relate the thematic, spatial, and temporal extents of the analysis spaces in the overview mode because of the color-coding. However, they highlighted that it is not easy to decide the boundaries to create an analysis space because some features of interest do not have well-defined boundaries. Figure 4.9A shows the interface in overview mode, and Figure 4.9B shows the form to create a new analysis space.



En esta ubiación, la población de la mosca del olivo (medida por las tramps McPhail) pasa de 6 a 0 moscas por mosquero antes del aumento en humedad promedio.

Palabras clave: Colapso, moscas por mosquero
Preguntas: 1
Preguntas activas: 0
Por admin *el* 10/13/2019, 11:56:47 AM

En esta área, el aumento de la humedad promedio parece haber causado el colapso de la población de mosca del olivo.

Palabras clave: Colapso, moscas/mosquero, humedad promedio
Preguntas: 1
Preguntas activas: 1
Por admin *el* 10/13/2019, 11:40:48 AM

Descripción: Se ha observado un rápido aumento en las capturas de la trampa McPhail, y no existen condiciones que expliquen dicho aumento.

Palabras clave: Brote, McPhail

Color: #FFE500

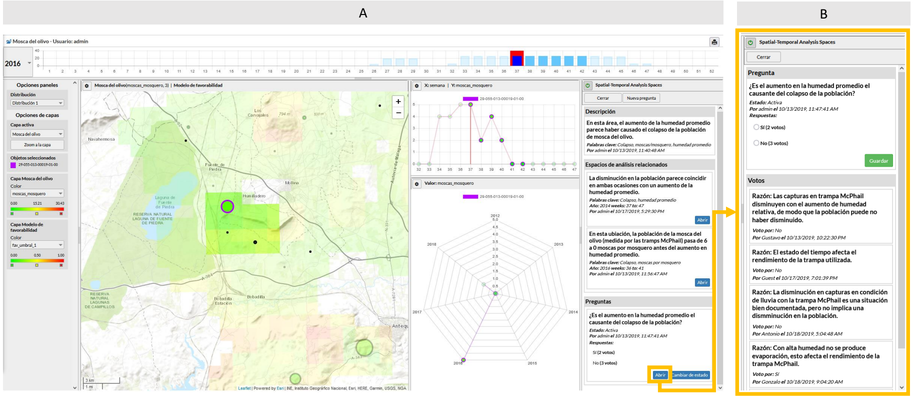
Límites espaciales
-4.593235296349799,37.066683840893084,-
4.447061575438705,37.143898131483034

Límites temporales

Ciclo	De	Hasta
2016	34	40

Figure 4.9 A) Prototype's interface in overview mode. B) Form to create new analysis spaces.

In the analysis mode, stakeholders mentioned that the reduced opacity of the data outside an analysis space makes it easy to understand which data (i.e., the feature of interest) is being analyzed. At the same time, it is clear what is happening outside the analysis space boundaries. However, they noticed that the user has no control over the amount (i.e., percentage) of opacity. Additionally, stakeholders mentioned that moving between related analysis spaces is an easy way to find relevant information. In this regard, they particularly appreciated that they could move between analysis spaces even if those are located in different years, which in their opinion, ensures that analysis spaces in previous years will remain relevant in the analysis process. On the downside, it was noticed that no option exists to 'jump back' when moving between related analysis spaces. Finally, the question-based mechanism was perceived as intuitive but somehow restrictive, emphasizing that a general-purpose discussion forum could be more flexible. Figure 4.10A shows the interface in analysis mode, and Figure 4.10B shows a discussion topic inside an analysis space.



Spatial-Temporal Analysis Spaces

Descripción

En esta área, el aumento de la humedad promedio parece haber causado el colapso de la población de mosca del olivo.

Palabras clave: Colapso, moscas/mosquero, humedad promedio

Por admin *el* 10/13/2019, 11:40:48 AM

Espacios de análisis relacionados

La disminución en la población parece coincidir en ambas ocasiones con un aumento de la humedad promedio.

Palabras clave: Colapso, humedad promedio

Año: 2014 *weeks:* 37 *to:* 47

Por admin *el* 10/17/2019, 5:29:30 PM

En esta ubiación, la población de la mosca del olivo (medida por las tramps McPhail) pasa de 6 a 0 moscas por mosquero antes del aumento en humedad promedio.

Palabras clave: Colapso, moscas por mosquero

Año: 2016 *weeks:* 36 *to:* 41

Por admin *el* 10/13/2019, 11:56:47 AM

Preguntas

¿Es el aumento en la humedad promedio el causante del colapso de la población?

Estado: Activa

Por admin *el* 10/13/2019, 11:47:41 AM

Respuestas:

Sí (2 votos)

No (3 votos)

Pregunta

¿Es el aumento en la humedad promedio el causante del colapso de la población?

Estado: Activa

Por admin *el* 10/13/2019, 11:47:41 AM

Respuestas:

○ Sí (2 votos)

○ No (3 votos)

Votos

Razón: Las capturas en trampa McPhail disminuyen con el aumento de humedad relativa, de modo que la población puede no haber disminuido.

Voto por: No

Por Gustavo *el* 10/13/2019, 10:22:30 PM

Razón: El estado del tiempo afecta el rendimiento de la trampa utilizada.

Voto por: No

Por Guest *el* 10/17/2019, 7:01:39 PM

Razón: La disminución en capturas en condición de lluvia con la trampa McPhail es una situación bien documentada, pero no implica una dismminución en la población.

Voto por: No

Por Antonio *el* 10/18/2019, 5:04:48 AM

Razón: Con alta humedad no se produce evaporación, esto afecta el rendimiento de la trampa McPhail.

Voto por: Sí

Por Gonzalo *el* 10/18/2019, 9:04:20 AM

Figure 4.10 A) Prototype's interface in analysis mode. B) Discussion topic inside an analysis space.

Further details on the prototype's design and development are provided in Section 5.3. Additionally, the evaluation of the prototype is described in Section 5.4.

4.5 Chapter summary

This chapter proposes the Spatiotemporal Analysis Space (STAS) approach, which is an approach for long-term distributed asynchronous collaborative analysis in GVA environments. The central concept of the approach is the analysis space, which is a container for a feature of interest and the analytical contributions to make sense of it. The approach enables analysts to explore a data set with large spatial and temporal extents, identify features of interest, and analyze them collaboratively. It offers a simple workflow composed of two working modes: overview and analysis. The former provides functionality to explore the whole data set and the existing analysis spaces, identify features of interest, and create new analysis spaces. The latter highlights the feature of interest and provides collaboration tools to analyze it. The approach allows creating relationships between analysis spaces to facilitate finding relevant information within the system and promote knowledge building from previous contributions. The links can be created manually through human input; or automatically from computer inference, using approaches such as textual-content similarity and/or geodata similarity.
This chapter also proposes a mapping of the STAS approach functionality into the software reference architecture proposed in Chapter 3. Briefly, it describes an implementation in the context of the application case of this research, which is the monitoring and control of the OFF in Andalusia, Spain.

5 Monitoring and control of the Olive Fruit Fly in Southern Spain

The content of this chapter is published on: *García-Chapeton, G.A., Toxopeus, A.G., Olivero, J., Ostermann, F.O., de By, R.A. Combining favorability modeling with collaborative geovisual analysis to improve agricultural pest management. Transactions in GIS. 2021; 25: 985- 1008.* *https://doi.org/10.1111/tgis.12714*

This chapter describes a case study performed to test the proposed software reference architecture and the collaboration approach. For this aim, a GVA prototype was developed based on the architecture, and it implements the STAS approach. It aims to enable stakeholders such as field technicians, authorities, researchers, and landowners to analyze the dynamics of a pest. The prototype was used to analyze monitoring data of the Olive Fruit Fly (OFF) in Andalusia, Spain, and the outputs of two statistical models developed as part of the case study. Therefore, this chapter addresses the fourth research objective, "Implement the software reference architecture and the collaborative analysis approach in a prototype for the monitoring and control of the Olive Fruit Fly and evaluate its usability and utility." Specifically, the case study assessed the potential of the favorability function to estimate locations and times at which a combination of conditions favors the OFF to exceed the acceptable abundance levels and requires the use of control measures such as pesticides. Discussions with the stakeholders indicate that the produced models and the prototype constitute valuable pest management tools.

The chapter is structured as follows: Section 5.1 sets the context of the case study by introducing the study area and the target species; Section 5.2 describes the modeling approach and results of the statistical models; Section 5.3 describes the design and development of a GVA prototype based on the software architecture and the collaboration approach proposed in Chapters 3 and 4, respectively; finally, Section 5.4 describes the testing of the prototype and the obtained results.

5.1 The Olive Fruit Fly in Andalusia, Spain

The olive is the fruit of the olive tree (Olea europaea), which is a species of small tree in the Oleaceae family. Its cultivation dates back to some 6,000 years ago in the Mediterranean region [163]. To date, the Mediterranean is still the most important olive producing region worldwide [163], with Spain as the world leader in olive production. In 2017, Spain alone produced about 35% of the world's total production. Figure 5.1 shows the evolution from 1980 to 2017 in the production of olive worldwide and the Spanish contribution to it. From the various pathologies that can affect olive trees, the most serious one is the OFF [93].

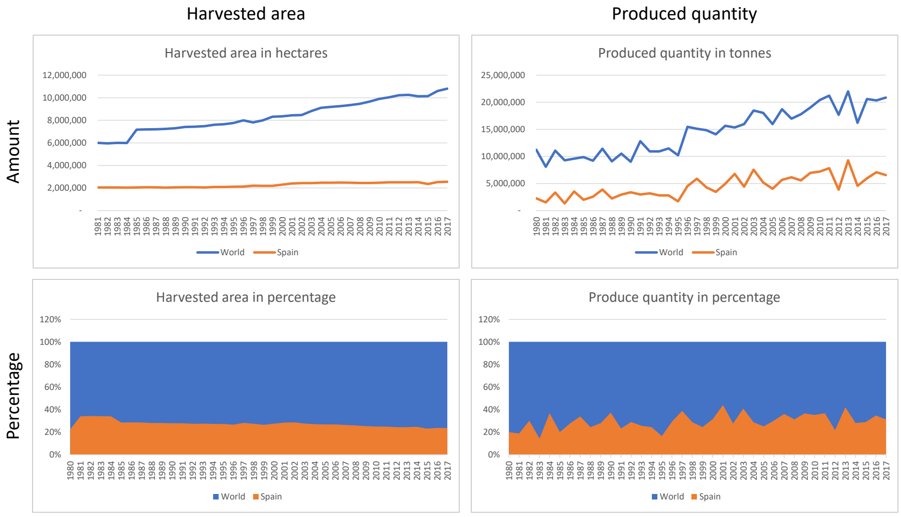


Figure 5.1 Olive production from 1980 to 2017. Data source: FAOSTAT [37].

5.1.1 Olive Fruit Fly

The OFF is considered a major pest in olive-growing regions worldwide. It is currently present in southern Europe, North Africa, the Middle East, and some areas of the United States and Mexico [107]. It has a high reproductive potential, and depending on the local conditions, there can be between three to five generations per year [119]. Additionally, it has high mobility with reported flying distances of up to 4 km to find olive tree hosts [129].

The damage caused by the OFF affects the quantity and quality of the produced table olives and olive oil [107]. A single female fly can lay up to 500 eggs (in its lifetime), usually one egg per olive fruit [172]. The damage is caused by the oviposition stings[1] and the OFF larvae who feed inside the olives, resulting in destroyed pericarp and the entry of

[1] An oviposition sting is a hole in the olive fruit where the OFF egg was laid.

secondary infection by bacteria and fungi that rot the fruit [172]. The oil from affected olives shows a higher acidity level [111], which reduces its commercial value. Economic losses due to OFF infestations have been reported up to 100% for table olives and 80% for olive oil [129, 172], because the former are not sellable, and the latter can only be used to produce low-quality oils.
The OFF population development is greatly influenced by the seasonal development of its primary host, the cultivated olive tree [172] and climatic factors, especially temperature and relative humidity [77]. Under optimal temperature conditions (20° to 30°C), a complete generation cycle takes about 30 to 35 days [129, 172]. Additionally, temperature affects the activity of adult flies; the species is not very active below 15°C and above 35°C [172]. Finally, high relative humidity favors ovarian maturation, egg production, and longevity of the OFF [18].

5.1.2 The study area

The study area is located in the center of the region of Andalusia, in southern Spain (see Figure 5.2). It covers an area of approximately 7,000 km^2, with terrain elevation ranging between approximately 0 and 1,400 meters above sea level. The area includes a total of 1,210 olive growing parcels with an approximated total area of 40 km^2. The main olive variety in the region is hojiblanca, which can be used for table olives and oil production. The harvest time defines the olives' destination. Table olives are harvested very soon after summer, between September and October, when fruits have already got their maximum size and are still green-colored and robust. Olives for oil production are harvested later, between November and January, when fruits have naturally turned black, and their pulp is becoming soft. The production of olives is an important source of income in the local and regional economy [35].

5.2 Modelling the Olive Fruit Fly dynamics

In agriculture, pest management is of key importance to ensure that damage to crops and stored products remains below an acceptable economic threshold. To this end, detailed monitoring data enables producers to decide when and where to perform pest control actions, but data collection is expensive and time-consuming. To overcome this limitation, statistical models provide a means to understand the factors driving the population dynamics and estimate species abundance or presence at non-monitored locations or periods.

5.2.1 Favorability function

The favorability function provides a measurement of the degree to which a set of conditions favors the occurrence of an event, regardless of the event prevalence [1, 126]. In this case study, the model output measures

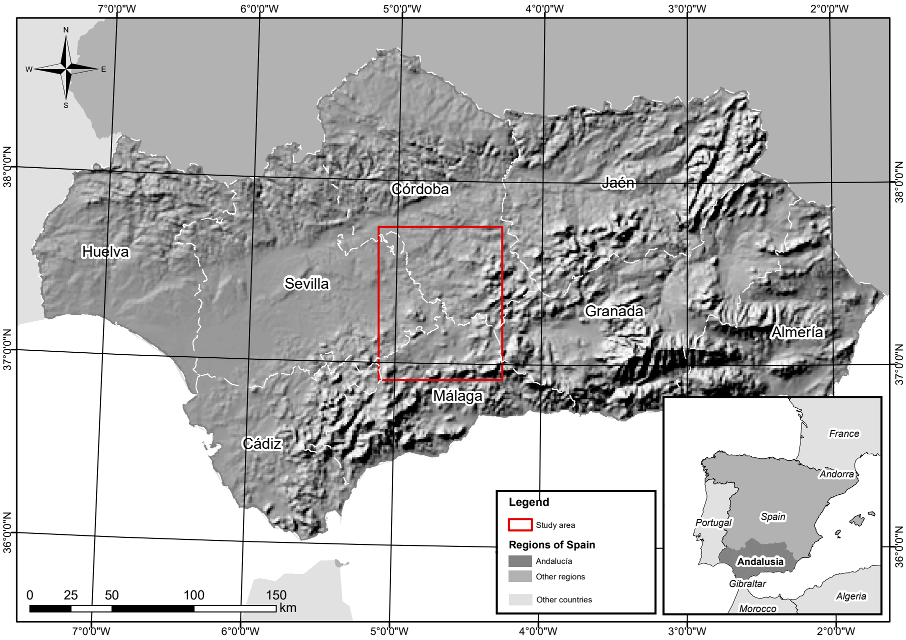


Figure 5.2 Study area in Andalusia, Spain

how favorable the topographic, environmental, and weather conditions are for the OFF to exceed the acceptable abundance thresholds. Favorability values range between 0 and 1 and are defined by the equation:

$$F = \frac{e^{y}}{\frac{n_1}{n_0} + e^{y}} \tag{5.1}$$

Here, n_1 and n_0 are the number of positive (i.e., the event occurs) and negative (i.e., the event does not occur) samples, respectively, and y is a regression equation of the form:

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n \tag{5.2}$$

Where, α is a constant and $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the n predictor variables $x_1, x_2, \ldots, x_n$. This y can be yielded by logistic regression:

$$P = \frac{e^{y}}{1 + e^{y}} \tag{5.3}$$

favorability values can, however, also be obtained from any method capable of producing probability estimates (P) using the equation:

$$F = \frac{\frac{P}{1-P}}{\frac{n_1}{n_0} + \frac{P}{1-P}} \tag{5.4}$$

Because favorability values are leveled to the event prevalence in the dataset, the value 0.5 indicates a combination of conditions (characterized by the predictor variables) that neither increase nor decrease the probability of the event's occurrence with respect to its prevalence, while values under (over) it represent conditions that are detrimental (favorable) for the occurrence of the event [1, 126]. Some successful applications of the favorability function are downscaling a species distribution model [112], assessment of the vulnerability of a native species due to an invasive species [135], and using favorability values as a proxy for species density [106]. Additionally, the favorability function was also applied successfully in the context of spatiotemporal modeling to assess the effect of deforestation in Ebola virus disease outbreaks [110].

5.2.2 Data sources and data preparation

The Integrated Production Associations (APIs, by Spanish acronym) "Antequera" and "La Camorra" provided a data set of the monitoring and control of the OFF in the study area for the years from 2012 to 2018 (inclusive). This data set was obtained following the protocol established by the Junta de Andalucía (i.e., the regional administration) for the monitoring of olive crops.[2] This data set includes information about weekly OFF abundance and fertility, damage to olives, application of chemical treatments (i.e., control measurements), and phenology of the olive trees. Every measurement is georeferenced by a parcel identifier, for which coordinates are available and timestamped with the date of the field observation.

The data set was produced for practical pest management. Therefore, it does not necessarily follow the highest scientific standards for data collection. This is a common scenario in many studies, where data was produced under a legal and regulatory framework. Such data sets result from several years of work and a significant expenditure, and therefore, an asset for its stakeholders. Changing the data collection protocol is a political and administrative effort, and therefore, out of the researcher's control; moreover, changing it also means that old data will become incomparable. This is the case of pest management in Andalusia, which motivated this case study to look for scientifically sound solutions to take advantage of the existing data.

The APIs use two types of monitoring devices to measure the OFF abundance: plastic McPhail flycatchers and yellow chromotropic sticky traps. This combination of devices measures the general population changes and sexual activity of the OFF. The McPhail flycatchers capture flies attracted by the yellow color of the trap and a liquid feeding lure. This device provides information about the general size of the population and attracts in similar proportion males and females [111]. The weather conditions the efficacy of this device because the feeding lure requires

[2] Available on `https://www.juntadeandalucia.es/agriculturaypesca/portal/export/sites/default/comun/galerias/galeriaDescargas/minisites/raif/manuales_de_campo/ProtocolosCampos_Olivar.pdf`

evaporation to work. The yellow chromotropic sticky traps capture principally males attracted by the yellow color of the trap and a pheromone. The information provided by this device is directly related to the sexual activity of the OFF population [111]. The captured flies are inspected to determine the percentage of female flies and female flies with eggs. Additionally, olives are sampled to determine fruit damage; the provided measurements include percentages of stung olives, olives with alive forms, olives with exit holes, and olives with parasitized flies. This monitoring strategy is described in the Andalusian Integrated Production regulation for Olives.[3]

The abundance measurements are reported as "flies per trapping device per day," which means the measurements are the average numbers captured by several devices during several days. The protocol and regulation established by the Junta de Andalucía define that a monitoring point is representative for an area of 300 hectares (i.e., 3 Km^2), and it should include three devices of each type (i.e., six devices in total per monitoring point), with monitoring visits every seven days to count the trapped individuals and clean the trapping devices. For example, a measurement of one fly per flycatcher per day means that the field technician found 21 flies captured by the three flycatchers in seven days. The monitoring protocol defines four thresholds for decision-making about the application of chemical treatment. Two of the thresholds are for the first application of a chemical treatment, one for table olives and the other for olives for oil production. The other two thresholds are for the subsequent treatment applications, one for table olives and the other for olives for oil production. The provided information was used to define dependent variables for two favorability models using the thresholds for the first application. The events were defined as "the observation exceeded the threshold." The observations were labeled as 1 (i.e., positive) if the value exceeds the threshold and 0 otherwise.

To rule out any influence of differing data recording practices between the two APIs, it was decided to continue the modeling process only with the data from "Antequera," because it is the largest association and contributes 75% of the data. Table 5.1 describes the thresholds, the number of observations for each threshold, and the number and percentage of positives and negatives.

Additionally, a set of predictors was selected as potential explanatory factors for the occurrence of the previously defined events. These were selected on the basis of a literature review and interviews with experts. They are seven expert stakeholders of the OFF management in the study area, and the information was obtained in face-to-face meetings, in which they were asked to describe the behavior of the OFF, and the behavioral drivers. The factors fall into the following four categories: human intervention and topographic, environmental, and weather conditions. The data for location (i.e., X and Y coordinates), human intervention (i.e., application of chemical treatment), and phenology of olive tree came

[3] Available on https://www.juntadeandalucia.es/boja/2008/83/d2.pdf

Table 5.1 Treatment thresholds used to define the dependent variables for the statistical models.

Name	Description	# of valid observations	Positives	% Positives	Negatives	% Negatives
Threshold 1 *table olives*	(flies per flycatcher per day >= 1) AND (percentage of female flies with eggs >= 50%)	1701	595	35%	1106	65%
Threshold 2 *olives for oil production*	(flies per flycatcher per day >= 1) AND (percentage of female flies with eggs >= 60%) AND (percentage of stung olives > 0 %)	1701	354	21%	1347	79%

from the data set provided by the APIs. All remaining data was obtained from the publicly available official sources at the Centro Nacional de Información Geográfica (CNIG)[4], the Red de Información Ambiental de Andalucía (REDIAM)[5], and the Red de Información Agroclimática de Andalucía (RIA)[6]. For a list of the potential explanatory factors and their source, see Table 5.2.

Data aggregation and interpolation methods were used to prepare data layers for the predictors at a spatial resolution of 1 km^2, and where applicable, at a temporal resolution of 1 week (i.e., some predictors are static in time, see Table 5.2). In other words, there is only one data layer for each of the static predictors such as altitude, slope, and distance to in-land water, and 350 data layers for each of the time-varying predictors such as phenophase, average temperature, and radiation (one for each week of the study period, running from January 1st, 2012 to September 16th, 2018). Later, the location and timestamp of the measurements were used to extract data from those layers and create vectors of the predictors for each field measurement. In this step, variables that represent the conditions of the N previous weeks (i.e., time-lagged predictors) can be included. This procedure is illustrated in Figure 5.3. For the modelling process, it was decided to use a time lag of five weeks, because under optimal conditions, a complete OFF generation cycle takes about 30 to 35 days [129, 172].

Additionally, on this step, the 24 derived variables defined in Table 5.3

[4] https://www.cnig.es

[5] https://www.juntadeandalucia.es/medioambiente/site/rediam/

[6] https://www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/

Table 5.2 Potential predictors for the favorability models for the Olive Fruit Fly.

No	Category	Predictor	Temporal variation	Source
1	Topographic	X	No	APIs
2		Y		
3		Altitude		CNIG
4		Altitude average*		
5		Altitude difference*		
6		Slope		
7		Slope average*		
8		Slope difference*		
9		Exposition to south		
10		Exposition to west		
11	Environmental	Distance to in-land water		REDIAM
12		Distance to sea		
13		Distance to wild olives		
14		Distance to roads		
15		Distance to urban centers		
16		Phenophase	Yes	APIs
17	Weather	Minimum temperature		RIA
18		Average temperature		
19		Maximum temperature		
20		Precipitation		
21		Accumulated precipitation**		
22		Minimum humidity		
23		Average humidity		
24		Maximum humidity		
25		Radiation		
26		Evapotranspiration		
27		Accumulated evapotranspiration**		
28		Wind direction		
29		Wind speed		
30	Human intervention	Chemical treatment		APIs

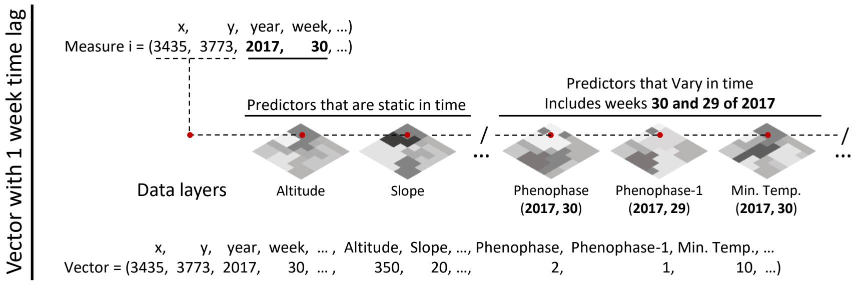
* Variables were interpolated over a grid with cells of 1 Km^2, the values for these variables were calculated from the eight neighboring cells of the corresponding predictor.

** Accumulated precipitation/evapotranspiration from the last September 1st (start of hydrological year for the region).

were created. Figure 5.4 illustrates the procedure to create these variables.

5.2.3 Modeling process and validation

The modeling process started with a set of 129 candidate variables: 15 static independent variables and 15 dynamic independent variables for which the values of the week of the measurement and the previous five weeks (i.e., six variables for each predictor, and 90 variables in total) were used, and 24 derived variables as defined in Table 5.3. After removing any predictor with a constant value, the predictors with high multicollinearity were identified and removed. For the latter, variables were iteratively removed until the remaining ones had a Variance Inflation


Vector with 1 week time lag
x, y, year, week, ...)
Measure i = (3435, 3773, 2017, 30, ...)
Predictors that are static in time
Predictors that Vary in time
Includes weeks 30 and 29 of 2017
Data layers
Altitude
Slope
Phenophase (2017, 30)
Phenophase-1 (2017, 29)
Min. Temp. (2017, 30)
x, y, year, week, ... , Altitude, Slope, ..., Phenophase, Phenophase-1, Min. Temp., ...
Vector = (3435, 3773, 2017, 30, ... , 350, 20, ..., 2, 1, 10, ...)

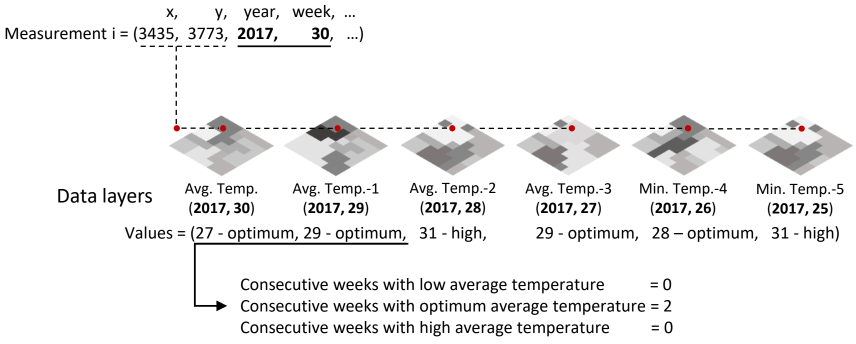
Notes: 1) Min. Temp. = Minimum temperature. 2) Phenophase-1 = Phenophase of 1 week before the measurement.

Figure 5.3 Using location and timestamp to prepare vectors with one-week time lag of the predictors for each measurement.

Table 5.3 List of derived variables

Derived variable	No variables	Base variable
Consecutive weeks with [low, optimum, high] minimum temperature	3	Minimum temperature
Consecutive weeks with [low, optimum, high] average temperature	3	Average temperature
Consecutive weeks with [low, optimum, high] maximum temperature	3	Maximum temperature
Consecutive weeks with [low, optimum, high] minimum humidity	3	Minimum humidity
Consecutive weeks with [low, optimum, high] average humidity	3	Average humidity
Consecutive weeks with [low, optimum, high] maximum humidity	3	Maximum humidity
Consecutive weeks [with, without] precipitation	2	Precipitation
Amount of precipitation in weeks	1	Precipitation
Consecutive weeks [with, without] chemical treatment	2	Chemical treatment
Note: Variables of the type "consecutive weeks with" are computed starting from the week of the measurements and moving backwards to a maximum of five previous weeks.		

Factor (VIF) of less than 10 [94, 105]. The VIF measures the correlation between variables, which can be used to detect and remove redundant predictors. Removing redundant predictors is important because adding highly correlated variables increases the model's complexity but contributes little to its accuracy. The high number of predictor variables might cause type-I errors. To reduce the False Discovery Rate (FDR), the procedure proposed by Benjamini and Hochberg [13] was used to keep only those predictors that are significant when tested on q = 0.05. This is crucial because as the number of performed hypothesis tests increases, the probability of obtaining false positives also increases; the FDR is the ratio of false positives to total positives; therefore, the controlling


x, y, year, week, …
Measurement i = (3435, 3773, 2017, 30, …)
Data layers
Avg. Temp. (2017, 30)
Avg. Temp.-1 (2017, 29)
Avg. Temp.-2 (2017, 28)
Avg. Temp.-3 (2017, 27)
Min. Temp.-4 (2017, 26)
Min. Temp.-5 (2017, 25)
Values = (27 - optimum, 29 - optimum, 31 - high, 29 - optimum, 28 – optimum, 31 - high)
Consecutive weeks with low average temperature = 0
Consecutive weeks with optimum average temperature = 2
Consecutive weeks with high average temperature = 0

Notes: 1) Avg. Temp. = Average temperature. 2) Avg. Temp.-N = Average temperature of N weeks before.

Figure 5.4 Preparing derived variables for the data vectors

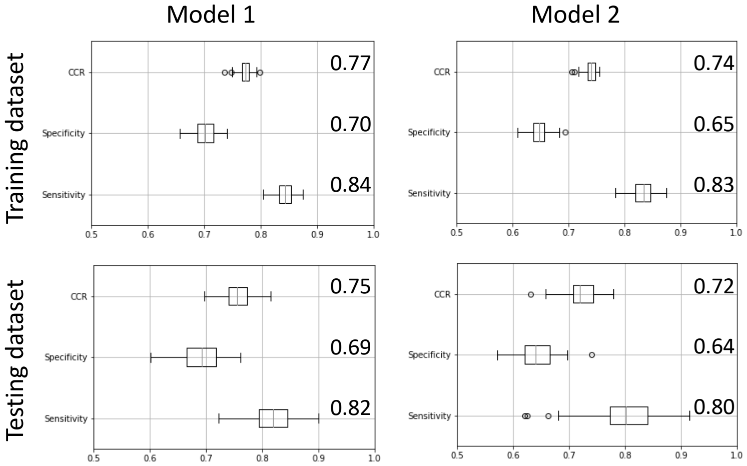
procedures aim to limit the tolerance for that ratio. Finally, a linear combination of variables was selected using forward-backward stepwise logistic regression based on statistical significance. The importance of each variable within the model was assessed using the Wald test. This test measures whether there is a significant difference in the model's accuracy with and without a predictor; therefore, it provides evidence to decide if a predictor should be included or not.

A cross-validation test was performed to assess whether the modeling process results in the overfitting of the models. The repeated hold-out method was applied to each dependent variable. Later, boxplots (for sensitivity, specificity, and correct classification rate – based on a favorability threshold of 0.5) were generated to compare the models' classification performance on the training and testing data sets. A total of 100 tests were run for each dependent variable. On each test, the data set was split into training and testing data sets. For this aim, a random sample with substitution was selected, including 20% of the observations as testing data set, and the remaining 80% of the observations as training data set. The training data set was used to generate a model using the procedure described at the beginning of this section and compute classification metrics for this data set. Later, the model was used to compute classification metrics for the testing data set. Finally, the results of the tests were used to produce the boxplots.

Once tested that the modeling procedure was not generating overfitted models, two models (i.e., one for each dependent variable) were produced using all the observations. The models' quality was assessed based on classification and discrimination capacity. For the classification capacity, the metrics were sensitivity, specificity, correct classification rate, and Cohen's kappa, all of them based on a favorability threshold of 0.5. Above this threshold value is where the conditions measured by the predictors favor the occurrence of the event. For the discrimination capacity (i.e., capacity to separate positive and negative instances), the area under the ROC curve (AUC) was used. The ROC curve is a graphical summary of the sensitivity and specificity values for different thresholds ranging between 0 and 1. The AUC is the value of a given threshold, in our case 0.5.

5.2.4 Modelling results

The cross-validation test showed that the modeling process is not producing overfitted models. Figure 5.5 shows that, on average, the performance of the models drops by about 2% for the test data sets. A significant drop in performance would indicate that a model is overfitted and hence generalizes poorly.



Notes: 1) CCR stands for Correct Classification Rate; 2) number beside CCR, specificity and sensitivity is the average for that specific boxplot.

Figure 5.5 Results from the cross-validation test

The modeling process resulted in the selection of 21 variables for the model based on threshold 1 (from here on referred to as Model 1) and 10 variables for the model based on threshold 2 (from here on referred to as Model 2). Tables 5.4 and 5.5 show the selected variables, coefficients (β), standard error (SE), Wald test value (Wald), significance (P), and VIF for Model 1 and 2, respectively.

Model 1 succeeds to classify correctly 85% of the cases in which the threshold is exceeded (sensitivity), and 72% of the cases in which it did not (specificity), which represents a correct classification rate of 77%. In comparison, Model 2 also achieves a sensitivity of 85%, but only a specificity of 67%, for a correct classification rate of 71%. The Cohen's Kappa for Model 1 is 0.53, and for Model 2, it is 0.37. According to Landis and Koch [86], this is moderately good ($0.41 < K < 0.6$) for Model 1, and fair ($0.21 < K < 0.40$) for Model 2. Finally, the discrimination capacity (AUC) for Model 1 is 0.84, while for Model 2, it is 0.81. According to

Table 5.4 Variables included in Model 1. Number in parentheses in a variable name indicates the number of weeks before the measurement.

No	Variable	β	SE	Wald	P	VIF
0	Const	-1.084	0.079	-13.665	< 0.001	—
1	Accumulated precipitation (1)	-0.809	0.116	-6.978	< 0.001	3.324
2	Minimum temperature (2)	0.897	0.155	5.804	< 0.001	5.404
3	Wind speed (3)	-0.620	0.103	-6.034	< 0.001	1.892
4	Accumulated evapotranspiration	-0.407	0.123	-3.303	0.001	2.804
5	Accumulated evapotranspiration (2)	-0.235	0.120	-1.973	0.049	4.663
6	Consecutive weeks with medium maximum humidity	0.661	0.099	6.678	< 0.001	3.015
7	Maximum humidity (3)	0.384	0.098	3.904	< 0.001	2.525
8	Distance to wild olives	-0.334	0.079	-4.231	< 0.001	1.280
9	Wind speed (4)	-0.222	0.094	-2.364	0.018	1.834
10	Maximum humidity (1)	0.244	0.115	2.135	0.033	3.072
11	Altitude difference	0.501	0.111	4.517	< 0.001	3.097
12	Consecutive weeks with medium minimum humidity	-0.446	0.105	-4.271	< 0.001	3.091
13	Distance to in-land water	-0.286	0.069	-4.178	< 0.001	1.170
14	Slope difference	-0.275	0.105	-2.632	0.009	2.809
15	Accumulated evapotranspiration (3)	-0.523	0.129	-4.054	< 0.001	4.791
16	Consecutive weeks with low maximum temperature	0.546	0.119	4.574	< 0.001	2.650
17	Minimum temperature (4)	0.634	0.135	4.689	< 0.001	3.375
18	Consecutive weeks with precipitation	0.318	0.082	3.903	< 0.001	1.788
19	Maximum humidity (5)	0.291	0.099	2.945	0.003	2.716
20	Consecutive weeks with low minimum humidity	-0.346	0.105	-3.308	0.001	2.787
21	Phenophase (1)	-0.216	0.096	-2.263	0.024	2.390
Abbreviations: β – coefficients' value; SE - standard error; Wald - Wald test value; P – statistical significance; and VIF – Variance Inflation Factor;						

Hosmer and Lemeshow [63], this is excellent (0.8 <= AUC < 0.9) for both models.

5.2.5 favorability maps

The models were used to produce favorability maps for the monitoring seasons 2012 to 2018. These maps were produced at a spatial resolution of 1 km^2, specifically for the areas where olive crops are located (i.e., in 1,162 pixels of 1 km^2), and a temporal resolution of 1 week. A file for each week of the study period was prepared to create the maps, containing 1,162 rows with the locations of interest and timestamps (which are constant within each file because each file corresponds to a specific week of a year). Later, the coordinates and timestamps on each file were used to produce vectors of the predictors following the same procedure as with the field measurements. Once the vectors were prepared, the models were used to produce the favorability values and

Table 5.5 Variables included in Model 2. Number in parentheses in a variable name indicates the number of weeks before the measurement.

No	Variable	β	SE	Wald	P	VIF
0	Const	-1.993	0.110	-18.073	< 0.001	—
1	Accumulated precipitation (1)	-0.546	0.135	-4.040	< 0.001	2.586
2	Altitude difference	0.307	0.069	4.450	< 0.001	1.180
3	Accumulated evapotranspiration	-0.665	0.170	-3.914	< 0.001	2.499
4	Y	-0.361	0.086	-4.204	< 0.001	1.212
5	Accumulated evapotranspiration (2)	-0.573	0.107	-5.331	< 0.001	3.013
6	Minimum temperature (2)	0.690	0.117	5.894	< 0.001	2.666
7	Wind speed (4)	-0.333	0.097	-3.434	0.001	1.320
8	Consecutive weeks with precipitation	0.202	0.066	3.062	0.002	1.241
9	Maximum humidity (4)	-0.190	0.082	-2.329	0.020	1.584
10	Consecutive weeks with high maximum temperature	-0.309	0.136	-2.274	0.023	2.309
Abbreviations: β – coefficients' value; SE - standard error; Wald - Wald test value; P – statistical significance; and VIF – Variance Inflation Factor;						

the weekly maps.

The maps show how favorable or detrimental the conditions were at locations and in times of interest for the OFF development. Figure 5.6 shows examples of the generated maps together with spatial and temporal summaries of the results. The former shows the spatial distribution of the monitoring points and the percentages of correctly classified observations for each, and the latter shows the number of correctly classified observations per week. The gap between weeks 28 and 34 is due to the absence of field observations, usually, because the species was inactive in previous weeks. The visual comparison of the summaries shows that the models' accuracy is not uniform, neither in space nor in time. In the maps, values closer to one are visualized in red tones because they represent favorable conditions for the pest development, therefore, a negative situation for the olive producers. Interestingly, the accuracy drops around week 37 in both models, which I cannot yet explain.

5.3 Collaborative geovisual analytics prototype

A prototype was developed as proof-of-concept for the software architecture described in Chapter 3, and the collaboration approach (i.e., STAS) described in Chapter 4. The prototype aims to enable stakeholders such as field technicians, authorities, researchers, and landowners to analyze a pest's dynamics and support decision-making on how to monitor and control it. Figure 5.7 shows a simplified overview of the components of the system. In the system, field technicians provide monitoring and control data for the species under study, which is combined with other relevant data sets and processed using application-specific processing

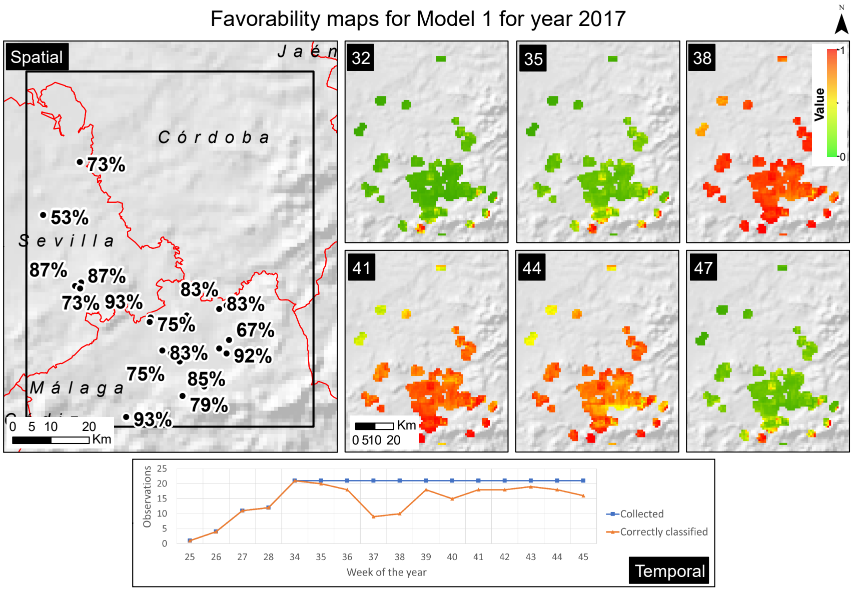


Figure 5.6 Spatial) shows the spatial distribution of the monitoring points and the percentage of correctly classified observations for 2017. Temporal) shows the temporal distribution of the field observations and correctly classified ones per week for 2017. 32-47) shows the spatial and temporal variation of the favorability for the OFF population to exceed threshold 1 for weeks 32 to 47 (with steps of 3 weeks), of 2017.

functionality (i.e., the favorability models described in Section 5.2). The monitoring data and models' outputs are available through a visual interface that enables collaborative analysis among stakeholders.

Due to direct requests from most of the stakeholders, the prototype was designed as a web-based application. Therefore, the stakeholders do not need to install specific software, and the prototype is operating-system-independent. Additionally, stakeholders emphasized the need for the following specific functionality:

1. **Hierarchical access to the data:** the data should be accessible following the hierarchy of cycle (i.e., year), observation period (i.e., week of the year), and monitoring location. Depending on the species under study, this hierarchy may change, but the concept of cyclic observation periods is likely to be universal for pest population analyses, although potentially with different time boundaries and granularity. This is due to the seasonality of crops, which influence the pest species development cycles.

2. **Spatial distribution of the pest:** analysts need to visualize the spatial distribution of the pest per week. This visualization enables assessing the effect of different parameters (i.e., human intervention and topographic, environmental, and weather conditions) in

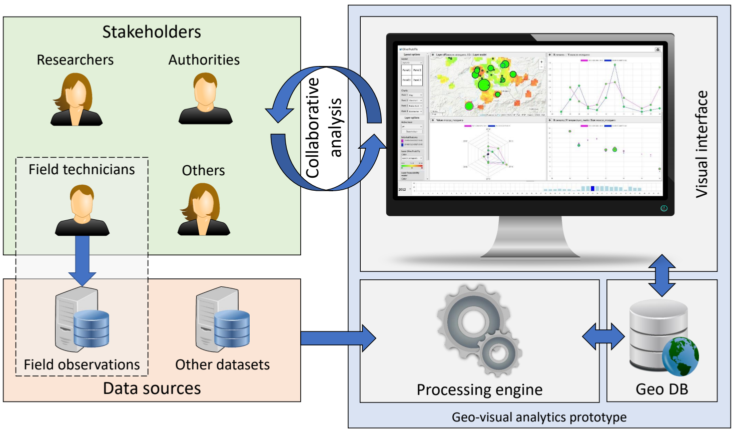


Figure 5.7 The GVA system aims to support stakeholders to understand the development of a pest, and support decision-making regarding control measurements and assess their effectiveness.

the pest population's spatial distribution and assessing the changes between observation periods (i.e., weeks).

3. **Temporal evolution of observation sites:** analysts need to visualize the evolution of a monitoring site over a cycle and compare it with other locations. This feature allows them to compare different locations and analyze how different topographic, environmental, and weather conditions affect the temporal dynamics of the species.

4. **Comparison over cycles:** analysts need to compare specific observation periods (i.e., week of the year) over different cycles (i.e., years), which allows them to identify variations in the population dynamic due to differences in weather conditions or human intervention between years. This feature also enables analysts to assess inter-annual periodic behaviors.

5. **Use of statistical models:** because data collection is expensive and time-consuming, the stakeholders emphasized that to support pest management properly, a system should be capable of accommodating processing capabilities to model the pest's behavior.

These case-specific requirements do not affect the software architecture, but they affect the design of the database structure and the user interface. The prototype is designed as two software applications: a CGVA environment and a data processing application. These applications include components that are distributed into five running environments: web-browser, web-server, processing-server, database-server, and file-server. Figure 5.8 shows the architecture layers used by each application and the mapping of those layers into the running environments.

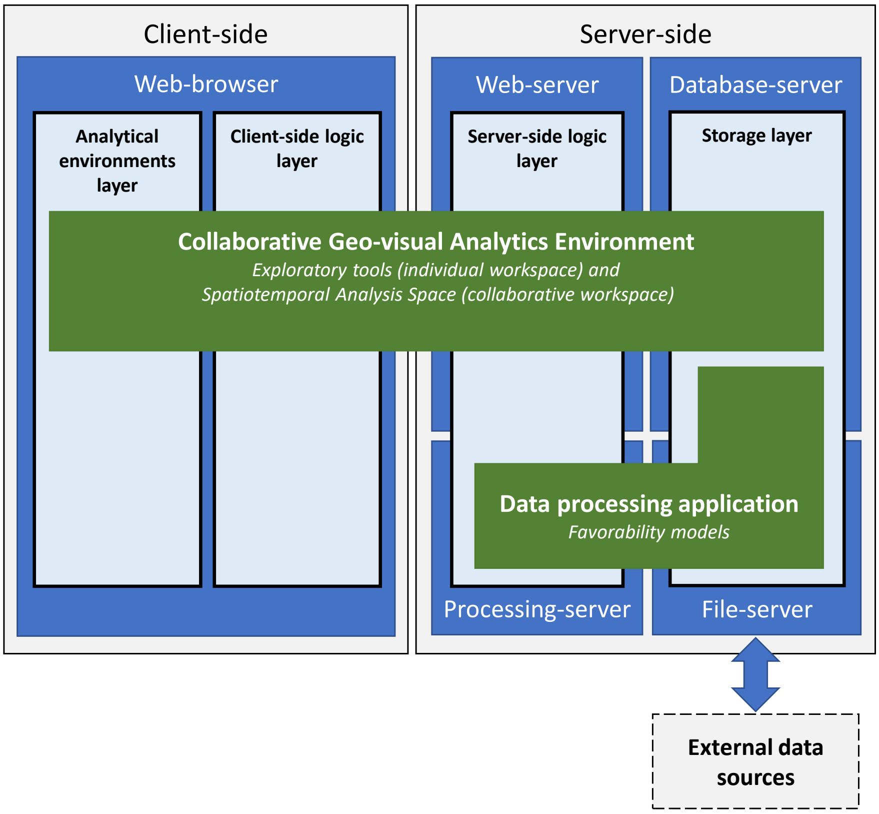

Client-side
Server-side
Web-browser
Web-server
Database-server
Analytical environments layer
Client-side logic layer
Server-side logic layer
Storage layer
Collaborative Geo-visual Analytics Environment
Exploratory tools (individual workspace) and
Spatiotemporal Analysis Space (collaborative workspace)
Data processing application
Favorability models
Processing-server
File-server
External data sources

Figure 5.8 The prototype includes two software applications: a collaborative geovisual analytics environment, and a data processing application.

5.3.1 Collaborative geovisual analytics environment

The software architecture defines individual and collaborative workspaces (See Chapter 3). Those workspaces can be developed as independent interfaces or be integrated, depending on the users' requirements. The prototype's design integrates the individual and collaborative workspaces into a single visual interface, enabling a seamless combination of individual and collaborative analysis. The individual analysis tools are for data exploration. These enable the analyst to select the data to be visualized, the visual style to represent it, and the type of visualization. The collaborative tools are a simplified version of the STAS approach, enabling the analyst to create analysis spaces and discuss the feature of interest with a questions-based forum. Figure 5.9 shows a simplified class diagram of the prototype's client-side, which includes classes to represent the data visually, and which enables user interaction (i.e., analytical environments layer), as well as local storage, processing capabilities, and communication between the client-side and the server-side (i.e., client-side logic layer), and classes to implement the STAS approach.

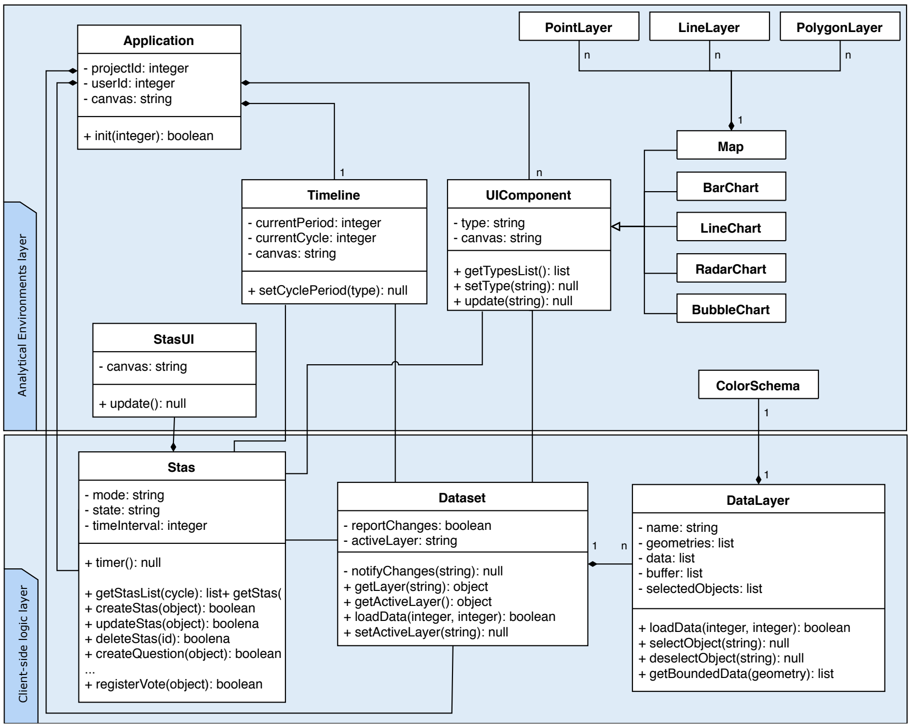


Figure 5.9 Simplified class diagram for the prototype's client side.

The *Dataset* class provides temporal client-side storage, which reduces the data exchange with the server-side, and enables a responsive interaction of the user with the data. This class is responsible for providing data to all the visualization components. Due to the requirement of hierarchical access to the data, this class was designed to load the data in one-year chunks. The *Dataset* class triggers the call to load data, but the data is loaded and stored by the *DataLayer* class. A *Dataset* object can hold several *DataLayer* objects, which enables the prototype to work with multiple data layers simultaneously. For the application case, two layers are needed: OFF field observations and favorability model outputs. To reduce the network traffic even further, once the data for a year is loaded, the *DataLayer* objects keep it in a data buffer, such that when data for a year is required, the object checks the buffer before loading data from the database. The buffer can be configured to hold data for a specific number of years, ensuring that the buffer does not get too heavy and negatively impacts the prototype's performance. The *ColorSchema* class allows defining the colors to represent each data layer on the visualization components.

The *Timeline* class controls the first two levels of the hierarchy to access the data: year and week of the year, which corresponds to monitoring cycle and period, respectively. When a user interacts with an object of this class, the changes are notified to a linked *Dataset* object, which prepares the data accordingly and notifies the changes to other components, which triggers the update procedure on them. If the STAS is activated, the *Timeline* shows the temporal extent of the analysis spaces available on the year being displayed on it.

The *UIComponent* class is a container for the data visualization classes and includes the *Map*, *BarChart*, *LineChart*, *RadarChart*, and *BubbleChart* classes (see Figure 5.10). The *LineChart* class is designed to fulfill the requirement of visualizing the evolution of monitoring sites through the year. It plots a selected variable for one or more monitoring sites against the weeks of the year. The *BarChart* and *RadarChart* classes are designed to address the requirement of comparing monitoring sites over different years. They plot a variable for one or more monitoring sites for the same week across different years. The *Map* class is designed to show the spatial distribution of the displayed variable for a selected week of a year. Additionally, it controls the last level of the hierarchy to access the data, the location. A user can select/deselect locations from the map view. The changes in selected elements trigger the update procedures of the *Dataset* class, which later propagates to other *UIComponent* classes to show detailed data for the selected locations.

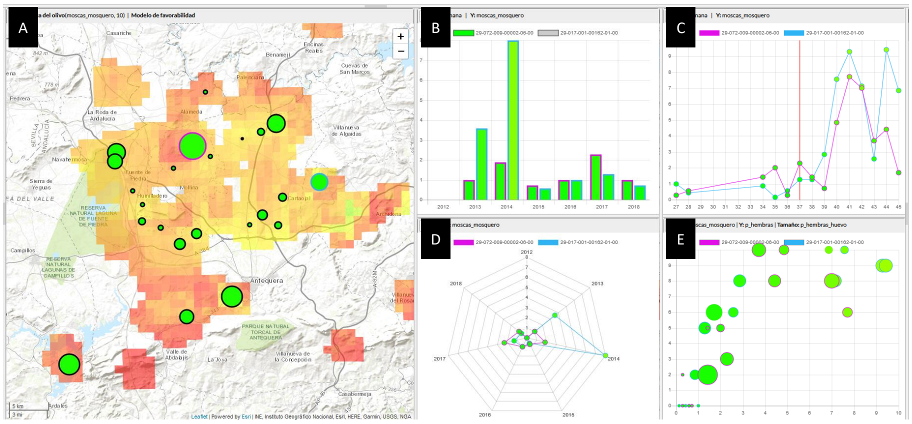


Figure 5.10 Examples of the visualization components. A) Map, B) Bar chart, C) Line chart, D) Radar chart, and E) Bubble chart.

The STAS implemented in the prototype is a simplified version of the design described in Chapter 4. It allows defining analysis spaces for features of interest and posting and answering questions inside them. It is designed as two classes: *Stas* and *StasUI*. The former contains the logic to communicate with the *server-side*, to synchronize the interface with the database, and to create, read, update, and delete analysis spaces, questions and answers. The latter controls the user interface. The prototype uses a simple synchronization method between the user interface and the database: a timer in the *Stas* class triggers a method to query the database in a fixed time interval, and a method in the *StasUI* class updates the user interface to represent the changes in the database.

Depending on the state of the STAS (i.e., overview or analysis), the query could retrieve the list of analysis spaces for a year, detailed data for an analysis space, or detailed data for a question. Regarding the links between analysis spaces, the prototype uses a simple word-matching algorithm between the keywords of the analysis spaces.

The *Application* class glues together all the other classes. It creates the instances of the classes and the appropriate links between those instances. The sequence to instantiate the classes is based on the dependencies of each class. The first class to be instantiated is *Dataset* because all the coordination between classes happens through it. Later, the *Stas*, *Timeline* and *UIComponent* classes are instantiated, in that order. When all the instances are ready, the method *loadData* on *Dataset* is called to load the data for the most recent available monitoring period. When the data is loaded, it triggers the update methods on the classes that visualize data.

Given that the prototype is designed following the layered architecture pattern, only the classes in the *client-side logic layer* can communicate with the *server-side*, specifically with the *server-side logic layer*. The *server-side logic layer* for the analytic environment is designed as a webserver, which has methods to process the requests from the client-side. Those methods aim to create, read, update, and delete objects in the database.

The *Storage layer* for the analytic environment is designed as a relational database. Figure 5.11 shows the database's structure. The database is designed to manage multiple analysis projects with several data layers each, and each layer can be used in several projects. The *project*, *layerInProject*, and *layer* schemas were designed for this aim. The *layer* schema only contains the metadata of the data layers. The geometries and attributes are stored in tables based on the generic schemas labeled as *[layer]_g* and *[layer]_a*, respectively. The data for the *STAS* is organized in the *stas*, *question*, *answer*, and *vote* schemas. This structure allows for several analysis spaces for each project, each of which can have several questions and answers to them.

5.3.2 Data processing application

This application implements the statistical models described in Section 5.2. The user can analyze the models' outputs through the analytic environment; therefore, this application has no dedicated user interface. It was designed as a series of methods that extract and prepare data, train the favorability models, and use them to produce maps. Figure 5.12 shows a schematic view of the methods and their interactions with the internal data storage and external data sources.

While the layers for the static variables were manually produced and stored in the file server, the production of layers for the time-varying variables was automatized as follows. The application starts by extracting observations of the monitoring and control of the OFF from databases provided by the APIs in the study area and climatic records from an

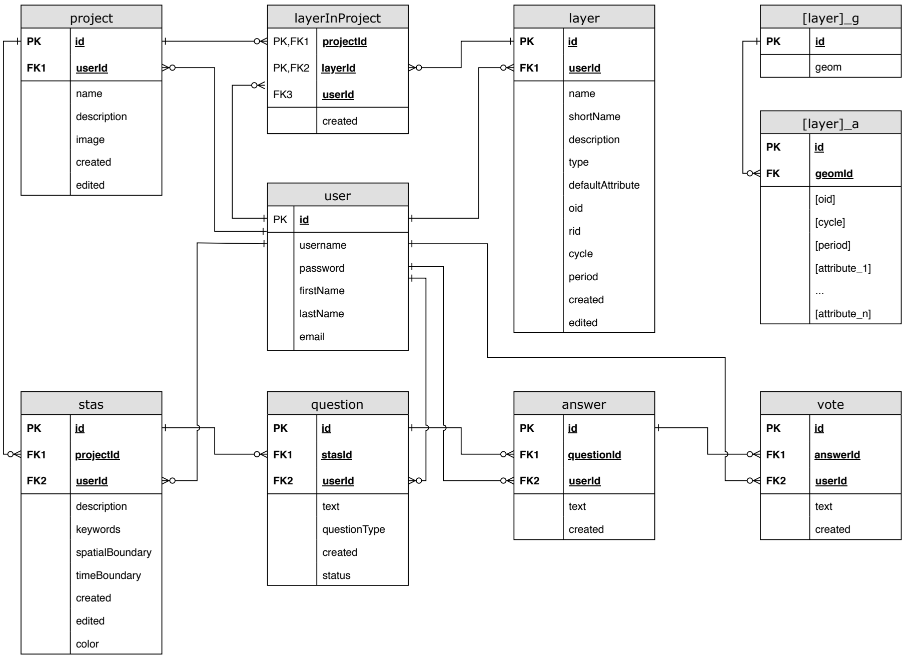


Figure 5.11 Simplified database schema for the analytic environment.

official data source. Those observations are stored in a spatial database using a convenient structure for further processing. Later, weekly layers for each time-varying predictor are produced using interpolation methods such as Inverse Distance Weighting (IDW) and Kriging. Data is extracted from the layers for the static and dynamic variables to produce data vectors for the OFF observations and the locations where olive crops are located. The former is used to calibrate the models, and the latter to produce the favorability maps.

5.3.3 Development

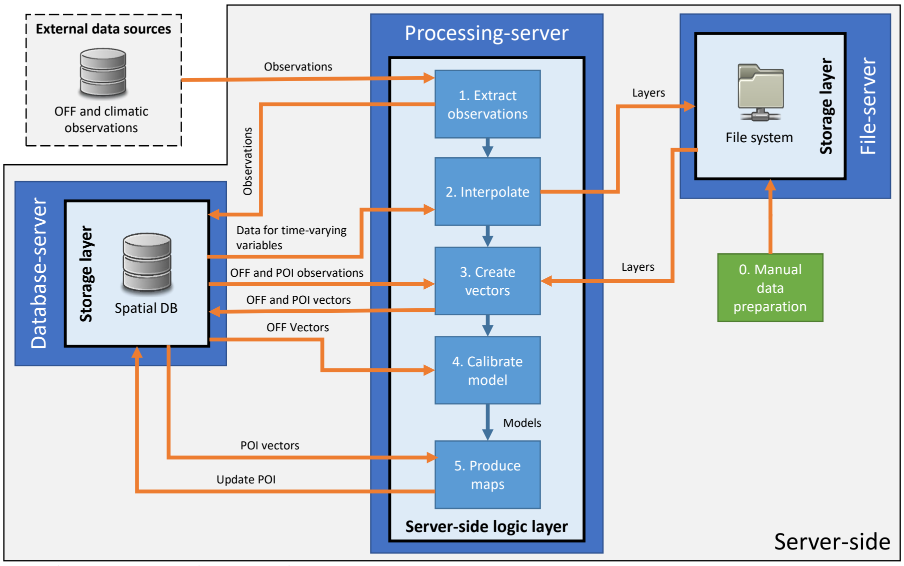
The client-side of the prototype was developed with HTML, CSS, and JavaScript. For the general structure of the user interface and its controls (e.g., buttons, dropdown lists, and text boxes), the JQWidgets library[7] is used. The map view is developed based on the Leaflet library[8], and the other visualization components are based on the ChartJS library[9]. The JQuery library [10] is used to manipulate the Document Object Model (commonly referred to as DOM), allowing to update the contents of the interface at runtime. The client-side logic layer is completely developed with JavaScript and uses the AJAX method of JQuery to communicate

[7] https://www.jqwidgets.com

[8] https://leafletjs.com/

[9] https://www.chartjs.org/

[10] https://jquery.com/



Notes: 1) OFF: Olive Fruit Fly; 2) POI: Points of interest, locations where olives are grown.

Figure 5.12 Application's workflow for the favorability models

with the server-side. The server-side logic layer for the analytic environment uses the Django Web Framework [11]. Finally, the database (i.e., storage layer) uses PostgreSQL[12] with its spatial extension PostGIS[13].

The data processing application was developed as a series of Python[14] scripts. This application is integrated with the analytic environment through the spatial database. It includes tables that store the geometries for the monitoring locations (i.e., points) and for the areas of interest (i.e., olive crops), which are 1km x 1km squares (i.e., 1km^2 polygons). Additionally, the database has tables that store the values for the different variables and models' outputs, for the monitoring locations and the areas of interest, for each week of the study period. These tables are based on the generic data structures labeled as *[layer]_g* and *[layer]_a* in Figure 5.11.

Figure 5.13 shows the interface of the prototype displaying the monitoring data and favorability map produced by Model 1 for week 36 of 2017. The user can choose the year and week of the year to be visualized from the timeline located in the upper part of the interface. The timeline displays the distribution of the available data through the (selected) year; thus, there are always 1,162 locations with data for the models. However, for the field observations layer, this changes from week to week. The prototype enables the user to select locations of interest, assess their evolution through the year (using the line chart), and compare the values

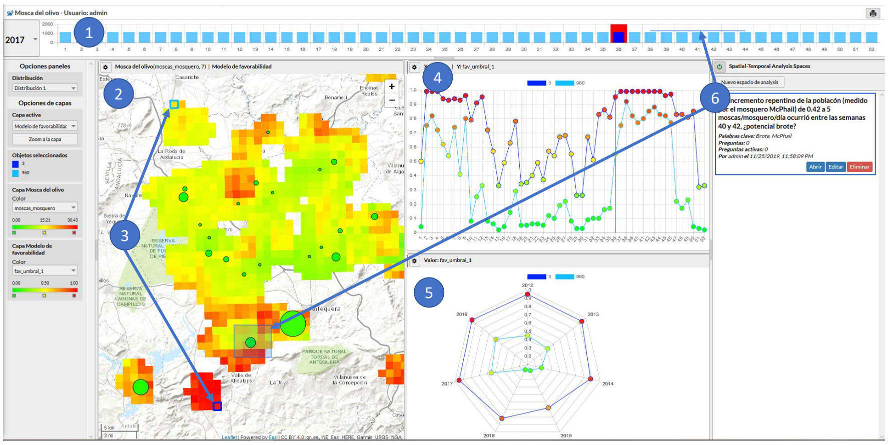
[11] `https://www.djangoproject.com/`

[12] `https://www.postgresql.org/`

[13] `https://postgis.net/`

[14] `https://www.python.org`

of the locations for the same week across different years (using the bar and radar charts). Additionally, selecting multiple locations of interest allows the user to compare between locations. All five charts (i.e., map, line, radar, bar, and bubble) include settings to select the variables to be displayed, the size of symbols, and the thickness of lines. The STAS main interface is located on the right side of the screen. As shown in the figure, when it is in overview mode, it displays a list of available analysis spaces (for the year being displayed) and their spatial and temporal extensions over the map view and timeline, respectively.


Mosca del olivo - Usuario: admin

2017

Opciones paneles

Distribución: Distribución 1

Opciones de capas

Capa activa: Modelo de favorabilidad

Objetos seleccionados: 3, 960

Capa Mosca del olivo — Color: moscas_mosquero — 0.00, 15.21, 30.43

Capa Modelo de favorabilidad — Color: fav_umbral_1 — 0.00, 0.50, 1.00

Mosca del olivo(moscas_mosquero, 7) | Modelo de favorabilidad

Y: fav_umbral_1

Valor: fav_umbral_1

Spatial-Temporal Analysis Spaces

Incremento repentino de la poblacón (medido por el mosquero McPhail) de 0.42 a 5 moscas/mosquero/día ocurrió entre las semanas 40 y 42, ¿potencial brote?

Palabras clave: Brote, McPhail
Preguntas: 0
Preguntas activas: 0
Por admin *el* 11/25/2019, 11:58:09 PM

Figure 5.13 CGVA prototype interface. 1) The timeline displays the distribution of available data over the selected year (here: 2017). 2) The map view displays the spatial distribution of a selected variable for a given week of the year (here: week 36). 3) Two locations have been selected, with charts 4 and 5 displaying detailed information. 4) the line chart shows the evolution of a variable for locations of interest through the year. 5) the radar chart shows the values of a variable for the same week in different years. 6) an analysis space in the Spatiotemporal Analysis Space, with its temporal and spatial extents displayed in the timeline and map views.

5.4 Prototype evaluation

A user evaluation was conducted to assess the prototype's usability and its utility in the monitoring and control of pests. The evaluation was done with the participation of seven stakeholders in the monitoring and control of the OFF in Andalusia, Spain. Despite the ideal number of participants still being under discussion in the user-experience research community, we consider seven participants sufficient, since literature suggests between five and ten users for a problem-discovery tests [145, 91]. The test aimed to identify design flaws and assess how well it could support the monitoring and control of pests in the opinion of stakeholders. The prototype's performance was not part of the test.

5.4.1 Evaluation setup

The prototype was tested by seven stakeholders: one authority representative, one researcher, and five field technicians. Their expertise regarding spatiotemporal analysis and Geographic Information Systems (GIS) varies from one very skilled user (i.e., the researcher) to others with no expertise at all. They all have domain and local knowledge about the monitoring and control of the OFF and were involved to some degree throughout the prototype's development. Regarding technical skills, they all use computers and mobile devices to perform their daily tasks; therefore, they can be considered technology-literate at the operation level. All participants were trained to use the prototype and then were asked to use the prototype to explore the monitoring data and the models' outputs and perform activities that required the exploration and collaboration tools.

The test consists of an evaluation form divided into four sections. The first section aimed to train the participants to use the prototype. It asked them to watch a series of videos explaining the different features of the prototype and how to use them, try each of the tools, and rate several statements about those features on a five-step scale ranging from 'strongly disagree' to 'strongly agree' (i.e., a Likert scale). The second section aims to assess the ease of performing tasks with the prototype once the user is trained and asks the user to perform a total of four tasks, which require using the individual and collaborative (i.e., STAS) analysis tools. Once each task has been completed, the user is asked to assess how easy it was to perform it. The third section uses the System Usability Scale (SUS) [16] to evaluate the general usability of the prototype. The SUS uses the Likert scale described earlier, and for consistency, this scale is used throughout the four sections. Finally, the fourth section aims to assess how useful the prototype might be to support the monitoring and control of pests. The evaluation form is included in Annex B.[15]

The SUS is a short questionnaire that consists of ten statements about the system under evaluation. In the questionnaire, the odd-numbered statements are positive (regarding the system's usability), and the even-numbered ones are negative. This alternation of the statements aims to avoid biased responses [17]. A test user can grade the statements with the values 'Strongly disagree,' 'Disagree,' 'Neutral,' 'Agree,' and 'Strongly agree,' which correspond to the values from 1 to 5 in the listed order. To aggregate the individual values: the contribution of odd-numbered statements are calculated as the user's score (from 1 to 5) minus 1; and for the even-numbered statements, it is calculated as 5 minus the user's score (from 1 to 5) [16]. Given this scoring schema, for example, a 'Strongly agree' response to an odd-numbered statement or a 'Strongly disagree' to an even-numbered one contributes 4 points to the score, and a 'Strongly disagree' response to an odd-numbered statement or a 'Strongly agree' to an even-numbered one, contributes

[15] The evaluation form is written in Spanish because all the stakeholders are native speakers of it.

0 points. Therefore, the best possible result is obtained when the test user fully agrees with the odd-numbered statements and fully disagrees with the even-numbered ones. When all the individual contributions are added up, the score ranges between 0 and 40. Later, this score is multiplied by 2.5 such that the score ranges from 0 to 100 [16]. The following equations can express the scoring of the SUS:

$$OS = Q_1 + Q_3 + Q_5 + Q_7 + Q_9 - 5 \tag{5.5}$$

$$ES = 25 - Q_2 - Q_4 - Q_6 - Q_8 - Q_{10} \tag{5.6}$$

$$SUS = (OS + ES) * 2.5 \tag{5.7}$$

Where Q_n are the scores for each statement, OS is the score for odd-numbered statements, ES is the score for even-numbered statements, and SUS is the total score for a test user, which ranges from 0 to 100. The statements in Section 3 are numbered from 28 to 37, therefore, what is considered is the order of these statements. A simple average can aggregate the scores for multiple users. The resulting scores can be interpreted based on different grading rankings, Figure 5.14 provides two examples. Those grading rankings were designed based on empirical experimentation [9, 10].

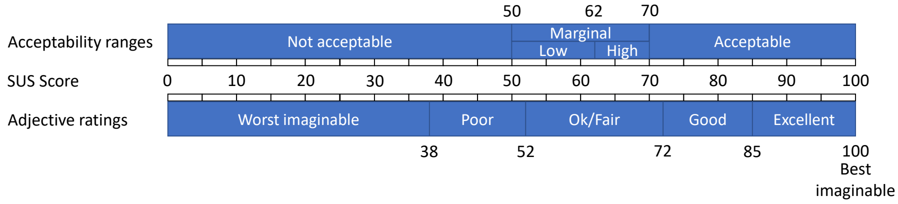


Figure 5.14 Grading rankings for SUS scores. Based on [9].

There is no alternation of the statements for the other questionnaire sections (i.e., sections 1, 2, and 4). They are all positive; therefore, the "strongly agree" responses are considered the best. Additionally, the scores of these sections are not aggregated to produce an overall score as with SUS. Finally, after the test, discussions with the participants were held to understand better their opinions regarding the prototype's usability and utility.

5.4.2 Evaluation results

In this section, the results of the prototype evaluation are presented and analyzed in an aggregated manner. The individual responses for each test user are available in Annex C.

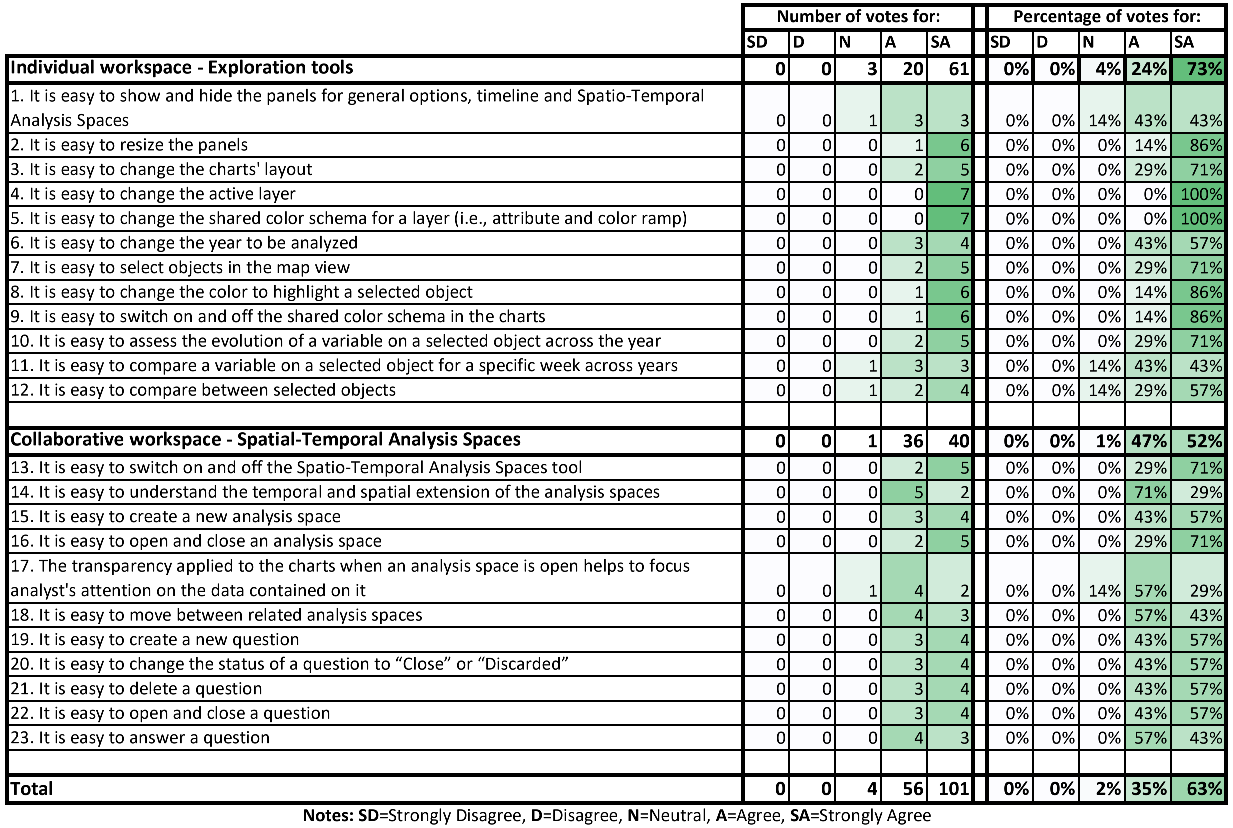
The responses for the first section of the questionnaire show that the prototype has a gentle learning curve. Participants watched a series of videos explaining the prototype features, tried them, and assessed the ease of using those features. This section is almost entirely composed of statements with the form "it is easy to..." (i.e., one statement does not follow this pattern). Therefore, a good result would be a tendency to the 'strongly agree'-side of the scale. Table 5.6 shows the answers for the first section, and it uses the heatmap technique to highlight the trend in the responses. It is clear from the table that all the statements in this section received a highly positive response from the test users. In general, the 'Strongly agree' alone accounts for 63% of the total answers, and combined with the 'Agree,' they account for 98%.

Comparing the responses to the statements regarding the individual workspace (i.e., statements 1 to 12) and the ones about the collaborative workspace (i.e., statements 13 to 23) show a slightly different tendency. While both are almost entirely within 'Agree' and 'Strongly agree,' the statements for the individual workspace were rated as 'Strongly agree' in 73% of the responses, and for the collaborative workspace, it was 52%. Post evaluation discussions with stakeholders suggested two reasons for this: first, it is not straightforward to decide the spatial and temporal boundaries for an analysis space because the features of interest do not necessarily have well-defined boundaries; and second, the sections for related analysis spaces and questions do not have a prominent separation, which results confusing.

Once the users were familiar with the prototype, they were asked to perform the four tasks from Section 2. While the users did not have any serious difficulty completing the tasks, it was clear that using the prototype on their own was harder than following the training videos. In this section, all the statements are "It was easy to complete the task," therefore, as with Section 1, a good result would be a tendency to the 'strongly agree'-side of the scale. The majority of responses were for the 'Agree,' which represents 68% of the total responses, and together with the 'Strongly agree,' they represent 89%, while the rest (i.e., 11%) of the responses were for the 'Neutral.' The responses for Section 2 are available in Table 5.7.

Section 3 is the SUS, which aims to assess the general usability of the prototype. As mentioned before, the statements in this section are alternated between positive and negative; therefore, a good result would be a tendency to 'Strongly agree' with positive statements and 'Strongly disagree' with negative ones. Table 5.8 shows the individual scores for each statement, the total score for each test user, and the summary of responses for Section 3. In total, two users rated the prototype as 'Ok/Fair ' with scores of 67.5 and 70, three rated it as 'Good' with scores of 72.5, 75, and 80, and two rated it as 'Excellent' with scores of 85 and 90. The total average score is 77.14, which corresponds to the rate of 'Good' [9]. This score means that while the prototype is certainly usable, there is room for improvement.

Section 4 focuses on how useful the prototype might be in the monitoring and control of pests. All the statements in this section are positive; therefore, a good result would be a tendency to the ‘Strongly agree’-side of the scale. The responses indicate that users considered the prototype to be a valuable tool for the monitoring and control of pests, as shown in Table 5.9. In the post-evaluation discussion, all the users mentioned that they would like the prototype to become a production system, and that it could also be applicable to the other pests and diseases they work with. Additionally, they suggested potential extensions such as highlighting locations and times where a treatment threshold is reached and producing predictions based on historical data.



	Number of votes for:					Percentage of votes for:				
	SD	D	N	A	SA	SD	D	N	A	SA
Individual workspace - Exploration tools	**0**	**0**	**3**	**20**	**61**	**0%**	**0%**	**4%**	**24%**	**73%**
1. It is easy to show and hide the panels for general options, timeline and Spatio-Temporal Analysis Spaces	0	0	1	3	3	0%	0%	14%	43%	43%
2. It is easy to resize the panels	0	0	0	1	6	0%	0%	0%	14%	86%
3. It is easy to change the charts' layout	0	0	0	2	5	0%	0%	0%	29%	71%
4. It is easy to change the active layer	0	0	0	0	7	0%	0%	0%	0%	100%
5. It is easy to change the shared color schema for a layer (i.e., attribute and color ramp)	0	0	0	0	7	0%	0%	0%	0%	100%
6. It is easy to change the year to be analyzed	0	0	0	3	4	0%	0%	0%	43%	57%
7. It is easy to select objects in the map view	0	0	0	2	5	0%	0%	0%	29%	71%
8. It is easy to change the color to highlight a selected object	0	0	0	1	6	0%	0%	0%	14%	86%
9. It is easy to switch on and off the shared color schema in the charts	0	0	0	1	6	0%	0%	0%	14%	86%
10. It is easy to assess the evolution of a variable on a selected object across the year	0	0	0	2	5	0%	0%	0%	29%	71%
11. It is easy to compare a variable on a selected object for a specific week across years	0	0	1	3	3	0%	0%	14%	43%	43%
12. It is easy to compare between selected objects	0	0	1	2	4	0%	0%	14%	29%	57%
Collaborative workspace - Spatial-Temporal Analysis Spaces	**0**	**0**	**1**	**36**	**40**	**0%**	**0%**	**1%**	**47%**	**52%**
13. It is easy to switch on and off the Spatio-Temporal Analysis Spaces tool	0	0	0	2	5	0%	0%	0%	29%	71%
14. It is easy to understand the temporal and spatial extension of the analysis spaces	0	0	0	5	2	0%	0%	0%	71%	29%
15. It is easy to create a new analysis space	0	0	0	3	4	0%	0%	0%	43%	57%
16. It is easy to open and close an analysis space	0	0	0	2	5	0%	0%	0%	29%	71%
17. The transparency applied to the charts when an analysis space is open helps to focus analyst's attention on the data contained on it	0	0	1	4	2	0%	0%	14%	57%	29%
18. It is easy to move between related analysis spaces	0	0	0	4	3	0%	0%	0%	57%	43%
19. It is easy to create a new question	0	0	0	3	4	0%	0%	0%	43%	57%
20. It is easy to change the status of a question to “Close” or “Discarded”	0	0	0	3	4	0%	0%	0%	43%	57%
21. It is easy to delete a question	0	0	0	3	4	0%	0%	0%	43%	57%
22. It is easy to open and close a question	0	0	0	3	4	0%	0%	0%	43%	57%
23. It is easy to answer a question	0	0	0	4	3	0%	0%	0%	57%	43%
Total	**0**	**0**	**4**	**56**	**101**	**0%**	**0%**	**2%**	**35%**	**63%**

Notes: SD=Strongly Disagree, **D**=Disagree, **N**=Neutral, **A**=Agree, **SA**=Strongly Agree

Table 5.6 Aggregated responses for Section 1 of the prototype evaluation. This section aims to train the analyst to use the prototype, and to assess the prototype’s learning curve.

	Number of votes for:						Percentage of votes for:				
	SD	D	N	A	SA		SD	D	N	A	SA
Task 1: create an analysis space 24. It was easy to complete the task	0	0	2	4	1		0%	0%	29%	57%	14%
Task 2: create a question 25. It was easy to complete the task	0	0	0	4	3		0%	0%	0%	57%	43%
Task 3: answer a question 26. It was easy to complete the task	0	0	1	3	2		0%	0%	17%	50%	33%
Task 4: change a question's status 27. It was easy to complete the task	0	0	0	4	2		0%	0%	0%	67%	33%
Total votes	**0**	**0**	**3**	**15**	**8**		**0%**	**0%**	**11%**	**58%**	**31%**

Notes: SD=Strongly Disagree, **D**=Disagree, **N**=Neutral, **A**=Agree, **SA**=Strongly Agree

Table 5.7 Aggregated responses for Section 2 of the prototype evaluation. This section aims to assess the ease to perform tasks with the prototype for a trained user.

	SUS scores							Number of votes for:					Percentage of votes for:				
	TU 1	TU 2	TU 3	TU 4	TU 5	TU 6	TU 7	SD	D	N	A	SA	SD	D	N	A	SA
28. I think that I would like to use this system frequently	4	4	4	4	4	5	5	0	0	0	5	2	0%	0%	0%	71%	29%
29. I found the system unnecessarily complex	2	2	2	2	2	4	2	0	6	0	1	0	0%	86%	0%	14%	0%
30. I thought the system was easy to use	4	4	4	4	3	4	5	0	0	1	5	1	0%	0%	14%	71%	14%
31. I think that I would need the support of a technical person to be able to use this system	2	3	2	2	3	3	2	0	4	3	0	0	0%	57%	43%	0%	0%
32. I found the various functions in this system were well integrated	4	4	4	5	4	4	5	0	0	0	5	2	0%	0%	0%	71%	29%
33. I thought there was too much inconsistency in this	1	2	2	2	2	1	1	3	4	0	0	0	43%	57%	0%	0%	0%
34. I would imagine that most people would learn to use this system very quickly	4	4	4	4	3	4	4	0	0	1	6	0	0%	0%	14%	86%	0%
35. I found the system very cumbersome to use	2	2	2	1	2	2	2	1	6	0	0	0	14%	86%	0%	0%	0%
36. I felt very confident using the system	5	4	4	5	4	4	5	0	0	0	4	3	0%	0%	0%	57%	43%
37. I needed to learn a lot of things before I could get going with this system	2	3	2	1	2	2	1	2	4	1	0	0	29%	57%	14%	0%	0%
	80	**70**	**75**	**85**	**67.5**	**72.5**	**90**	**6**	**24**	**6**	**26**	**8**	**9%**	**34%**	**9%**	**37%**	**11%**
	Good	**Ok/Fair**	**Good**	**Excellent**	**Ok/Fair**	**Good**	**Excellent**										

Notes: TU=Test User, **SD**=Strongly Disagree, **D**=Disagree, **N**=Neutral, **A**=Agree, **SA**=Strongly Agree

Table 5.8 Aggregated responses for the SUS (i.e., Section 3) of the prototype evaluation. This section aims to assess the general usability of the system.

	Number of votes for:						Percentage of votes for:				
	SD	D	N	A	SA		SD	D	N	A	SA
38. The system helps to explore and understand the population dynamics of a pest	0	0	1	2	4		0%	0%	14%	29%	57%
39. The system may support decision-making regarding the application of control actions	0	0	0	5	2		0%	0%	0%	71%	29%
40. The system may help to assess the effects of the control actions	0	0	0	3	4		0%	0%	0%	43%	57%
41. The system may help to understand the long-term changes in the population dynamics of a pest	0	0	2	2	3		0%	0%	29%	29%	43%
42. The system may help to improve communication among stakeholders of the pest management	0	0	1	3	3		0%	0%	14%	43%	43%
43. The system may help to build a knowledge base with domain and local knowledge regarding a pest	0	0	0	5	2		0%	0%	0%	71%	29%
Total votes	**0**	**0**	**4**	**20**	**18**		**0%**	**0%**	**10%**	**48%**	**43%**

Notes: SD=Strongly Disagree, **D**=Disagree, **N**=Neutral, **A**=Agree, **SA**=Strongly Agree

Table 5.9 Aggregated responses for Section 4 of the prototype evaluation. This section aims to assess how useful the prototype might be in the monitoring and control of the pest.

5.5 Chapter summary

A CGVA prototype was designed and developed as a proof-of-concept for the software reference architecture proposed in Chapter 3 and the collaboration approach (i.e., STAS) described in Chapter 4. This prototype was designed and developed in the context of a case study about the monitoring and control of the OFF in Andalusia, Spain.

Regarding the case study, pest management is of key importance to ensure food security. To this end, detailed monitoring data enables producers to decide when and where to perform pest control actions, but data collection is expensive and time-consuming. To overcome this limitation, statistical models can be used to produce data for non-monitored locations and times. Among the statistical models, the favorability model allows assessing how a combination of conditions (e.g., topographic, environmental, and weather) prevents or favors a species from becoming a pest. As part of the case study, two favorability models were developed based on economic thresholds for the OFF. Later, the monitoring data and the models' outputs were integrated into the prototype's visual interface to enable collaborative visual analysis by the case study stakeholders.

The prototype and models were tested with the participation of seven stakeholders, and the results show that these are valuable tools for the OFF management. In the post-evaluation discussions, all the participants mentioned that they would like to see the prototype becoming a production system to support their monitoring and control activities.

Discussion and conclusions 6

In Chapter 1, the need to support collaborative analysis in GVA is identified, the main reasons being the abundance of geodata and the complexity of current analytical problems. Several advancements in Information and Communication Technologies (ICT) and geospatial technologies have led to an unprecedented abundance of geodata, which presents an opportunity to better understand natural and artificial processes. Analyzing such data sets requires the combination of human analytical skills and computers' storage and processing power. GVA can enable this synergy. However, addressing the complexity of current analytical problems requires collaboration among analysts from different backgrounds, such as domain experts, data analysts, scientists, and laypersons. In this context, the starting point for this thesis is the limited research regarding the support for collaborative analysis in GVA [2, 20, 33, 54, 66, 79, 156]. The execution of this research produced four important results, which are presented in Chapters 2 to 5:

- Chapter 2: a literature review of the support for collaborative analysis in GVA systems.
- Chapter 3: a software reference architecture for CGVA systems.
- Chapter 4: an approach for long-term distributed asynchronous collaborative analysis in GVA systems.
- Chapter 5: a case study to evaluate the software reference architecture and the collaboration approach.

In this chapter, I reflect on the activities and results presented in those chapters, highlight their contributions to science, draw conclusions for the thesis, and propose future research directions.

6.1 Discussion

6.1.1 Support for collaborative analysis in geovisual analytics

The first research objective, "Review the state-of-the-art of collaborative geovisual analytics, and propose a research agenda," was addressed with a systematic literature review, following the guidelines proposed by Kitchenham and Charters [83]. The review focuses on the systems' characteristics regarding the technological platforms in which they are deployed and their support of collaboration scenarios and implemented collaboration techniques. This focus provides various insights relevant to the research field, which are discussed later in this section. The reason to adopt a particular focus is due to the multi-disciplinary nature of GVA, which leaves a comprehensive state-of-the-art of all related literature beyond the scope of this research. In this respect, other perspectives might be adopted in future research, leading to an improved understanding of the support for collaborative analysis in GVA.

The GVA systems identified in the literature review present a variety of characteristics regarding the supported collaboration scenarios, implemented collaboration techniques, and types of deployment. The analysis of those characteristics revealed three developments that suggest that GVA environments aim to reach a broader audience. First, the most common collaboration scenario is asynchronous distributed, which promotes participation by removing the constraints on time and location to contribute. Since analysis efforts can benefit from diverse expertise and domain and local knowledge, removing this constraint is important because it enables participation across geographic locations and domains. In this regard, the study by Benbunan-Fich, Hiltz, and Turof [12] found that asynchronous collaboration, when compared to a face-to-face scenario, resulted in higher-quality outcomes because participants have time to generate and reflect on new ideas and can contribute regardless of their location. The second development concerns the increasing use of cloud technology, a common trend in analytics systems [161] and which improves the scalability of and distributed access to the system. Cloud technology offers a flexible and straightforward process to scale up or down the system's storage and processing capacity. This characteristic enables it to cope with the system's workload fluctuations and improves its responsiveness and availability. Finally, the third development is increasing support for multiple devices, eliminating the need for specialized hardware and promoting participation from diverse stakeholders. This is particularly relevant for the growing field of citizen science, in which a participant is likely to contribute from low-end devices.

The review identifies six collaboration techniques: annotation, discussion board, instant messaging, interaction history, snapshot, and storytelling. The most common collaboration techniques are snapshot, storytelling, and annotation, which often co-occur and complement each other. The combination of these techniques offers a flexible working environment that allows analysts to combine individual and collaborative analysis and produces self-explanatory results that can be immediately communicated. The other techniques are uncommon due to either lack of flexibility, difficulty to integrate them into the system interface, or implementation requirements. The identified techniques support the whole data analysis process, including identification of features of interest, generation of hypothesis, provision of evidence, and communication of analytic results. However, features are missing to aid the synthesis of analytical contributions, which is important to support knowledge generation; also missing are features to summarize the level of agreement about evidence and conclusions, which would provide certainty when results are communicated.

The literature review identifies three research challenges. These are the lack of support for: hybrid collaborative scenarios, cross-device collaboration, and time-critical and long-term analysis. In this regard, Neumayr et al. [109] observed that in real-world analysis scenarios, analysts use different systems in parallel, which allow them to move back and forth between analysis scenarios and devices. In other words, analysts fulfill their need for hybrid collaboration scenarios and cross-device collaboration by combining multiple systems. They argue that current theoretical frameworks do not have sufficient descriptive power to capture the true nature of real-world collaboration. Based on their observations, it is apparent that the absence of an adequate theoretical framework has prevented the development of integrated systems offering support for those analysis scenarios.

The outcomes of this stage constitute a valuable contribution for the research community for three reasons: first, the review summarizes the progress to support collaborative analysis in GVA made since 2005, defining three main developments and six collaboration techniques; second, it provides a research agenda in the form of three research challenges and proposes specific strategies to address them; and third, it sets a landmark to measure the progress regarding the research agenda in the future.

6.1.2 Designing a software reference architecture for collaborative geovisual analytics

A system's software architecture is an abstract high-level description of the structures needed to reason about the system. An architecture comprises of elements such as: classes, processes, devices, and protocols; their relationships such as 'shares data with,' 'provides services to,' and 'executes on;' and the properties of both elements and relationships that together form a software system [11, 104]. In other words, an architecture describes the components that form a software system, the interfaces they use to communicate, and the distribution of functionality over those components, which helps to understand the system structure and behavior and provides guidance for the system development.

Software architectures can have different goals and scopes. A concrete software architecture aims to guide the design, development, test, deployment, maintenance, and extension of a single software system [4]. In comparison, a software reference architecture serves as an inspiration or standardization tool for the design of multiple concrete architectures to be implemented in multiple systems [4]. Due to the importance of software architecture, the second research objective of this thesis was to "Design a software reference architecture for collaborative geovisual analytics systems."

Real-world analysis processes are complex and commonly include participants with diverse backgrounds, expertise, and interests, combining individual and collaborative work across different interaction scenarios [109, 69]. For this reason, GVA systems need to provide a flexible workflow that adapts to the needs of the analysis effort. The three research challenges identified in Chapter 2 reflect the need for such flexibility. The software architecture design is based on the analysis of those challenges, which ensures that it provides a flexible workflow. Additionally, the literature review identifies six collaboration techniques which the architecture has to be capable to accommodate for a flexible analysis workflow. In this regard, the architecture layers conveniently organize the code to implement a technique by separating it into three responsibilities: interface, logic, and data persistence.

Given that analysts may interact in a synchronous, asynchronous, and multi-synchronous manner, a fundamental element in the architecture is a flexible synchronization component capable of supporting all those types of interactions. This component is responsible for sending and receiving contributions to and from the shared data storage component, enabling the contributions propagation. In synchronous and asynchronous collaboration, the contributions are immediately integrated into the shared database, making it a straightforward process. However, in the multi-synchronous collaboration, contributions are integrated in a delayed manner, which might produce conflicting updates. For example, when an annotation is added offline to an object that was deleted in the shared database, or modifications are done offline to a story (using storytelling) that was already concluded in the shared database. Addressing the issue of version conflicts between the local and shared data storage is not trivial. However, different approaches can be adopted: versioning conflicts can be avoided by creating a lock on the data and any analysis artifacts that are being used for offline analysis, thus providing exclusive usage rights to the analyst working offline [74, 143]. On the downside, this disturbs the analysis process. Following the same line of thinking, but in a less restrictive manner, the system could warn about potential conflicts, for example, when an analyst is working with data and/or artifacts for which local copies exist. Creating awareness of potential conflicts might promote communication and coordination to reduce conflicting updates [39]. In many cases, conflicting updates cannot be avoided, leading to the need for reconciliation procedures. Depending on the type of conflicts, the procedure might range from manual (i.e., user-driven) to automatic (e.g., rule-based) [64]. Regardless of the adopted approach, it will affect the analysis workflow. Therefore, it should be selected considering the application domain, supported analysis scenarios and collaboration techniques, and system requirements.

To promote the participation of diverse stakeholders in GVA systems, these should not require specialized devices. However, the system might need to store and process large data sets, which might require high-end devices. These are conflicting needs, which the proposed architecture address by separating the system into client- and server-sides. Each has a specific role; the client-side implements the user interface, and the server-side the storage and processing functionality. This separation enables the system to take advantage of the unique characteristics of diverse types of devices by implementing specialized interfaces for each type. This characteristic also removes the technological barrier to work with the system. Therefore, I expect it to promote broad participation in collaborative analysis efforts, bringing more varied knowledge and expertise and increasing the potential for completeness and accuracy of analysis results. Additionally, this enables different device types to play specific roles within the system. For example, desktop/laptop workstations can be used for individual and distributed collaborative work, touch tables for co-located collaboration, and mobile devices for off-line in-field individual analysis. The server-side can be designed based on the storage and processing demands for the analysis effort and can be scaled up or down when needed without affecting the client-side. Finally, given that the storage is centralized on the server-side, this architectural design simplifies the management and security of the data.

Regardless of the device in use, the analyst would expect the system to be responsive. Failing to fulfill this expectation might discourage participation. In the architecture, this is addressed by separating the system into client- and server-sides. The server-side is in charge of data storage and processing, thus reducing the need for storage and processing capacity on the client-side. Responsiveness is further addressed by adding a local storage component to the client-side, which acts as a cache memory, reducing the need to contact the server and speeding up data access. The client-side local storage is also important for multi-synchronous collaborations. It allows to store the required data into a device before going offline and stores the contributions locally until they can be synchronized with the server-side.

The proposed software reference architecture for this research is based on the client-server and layered patterns. Many architectural patterns can be used as the building blocks for a software architecture. Based on the design criteria described in Section 3.2, three architectural patterns were identified as the building blocks for the software architecture proposed in Section 3.3: client-server, layered, and micro-services. For the last few years, micro-services architectural pattern have seen a sharp rise in usage [76]. However, this pattern has disadvantages as any other pattern, and hence it is a suitable choice for some systems, but not for all [130, 146, 29]. The micro-services pattern offers excellent advantages in flexibility and scalability of the system. However, it comes with significant trade-offs in the form of upfront complexity and economic cost for (software and hardware) infrastructure and development time, which might not be acceptable for small to middle size projects [130, 146]. For this reason, the proposed architecture is based only on the client-server and layered patterns. However, given that there are scenarios in which an architecture based on micro-services is appropriate, and the previously mentioned trade-offs might be affordable, I also briefly described a modified version of the architecture that uses micro-services. To the best of my knowledge, to date, there is no GVA system based on the micro-services architectural pattern. The lack of micro-services-based

GVA systems is a research gap. Addressing it will help to understand better how the micro-services pattern characteristics affect the design, development, deployment, and operation of GVA systems. Some guiding research questions might be: what hardware and software infrastructure is required to build a core GVA system based on the micro-services architectural pattern? What are the necessary micro-services to build a core GVA system? How easy is it to adapt a core system to diverse contexts and application domains? Does the micro-services pattern provide adequate support for collaborative features?.

6.1.3 An approach for collaborative analysis in geovisual analytics environments

To address the third research objective, "Design an approach for collaborative analysis in geovisual analytics environments," I analyzed the research challenges from the literature review (See Chapter 2) and user requirements provided by the stakeholders of the case study in Southern Spain. This led to identify three design criteria: first, to provide a mechanism that enables the analysis of data sets with large spatial and temporal extents; second, to provide a mechanism to support long-term analysis efforts; and third, to provide a mechanism that can rely on multiple collaboration techniques. After designing the approach, its components were mapped to the software reference architecture proposed in Chapter 3, which purpose was to assess whether the approach could be implemented in a system based on the proposed architecture.

One of the starting points for this research is the abundance of geodata [24, 22, 149]. In this context, the discussions with the stakeholders of the case study highlighted that pest management efforts might collect data for several years over a large production area (e.g., Figure 5.1 shows that in 2017, the total cultivated olive area in Spain was around 2,500,000 ha). Further, it was discussed that this is also common in other applications. For example, the monitoring of species for conservation purposes, such as tracking of migratory birds which may fly from one continent to another; the monitoring of vehicles (e.g., cargo or fishing ships, and airplanes) traveling long distances and facing changing and potentially hazardous conditions; and intelligence analysis, for example, to prevent crime in a country or a city. These potential applications led to the first design criterion "to provide a mechanism that enables the analysis of data sets with large spatial and temporal extents."

The second design criterion, "to provide a mechanism to support long-term analysis efforts," follows from the potential applications mentioned in the previous paragraph. In this context, data collection and analysis may span several years and is a long-term analysis process. This design criterion directly relates to the third research challenge from the literature review, which is the lack of support for "time-critical and long-term analysis." Throughout the analysis process, diverse contributions and results are generated over several years, and human memory can not keep track of them all. Therefore, to provide effective support for a long-term analysis effort, the approach should facilitate the identification of previous relevant analytical contributions and their continued use in analysis to build knowledge incrementally.

Building the approach around a single collaboration technique might have led to a lack of flexibility and limited applicability. For this reason, the third design criterion was "to provide a mechanism that can rely on multiple collaboration techniques." Given that collaboration techniques can be used in diverse interaction scenarios and across devices, this design criterion relates to the first and second research challenges from the literature review, which are the lack of support for "hybrid collaborative scenarios" and "cross-device collaboration."

The Spatiotemporal Analysis Space (STAS) is an approach for long-term distributed asynchronous collaborative analysis in GVA environments and is designed to fulfill the three criteria mentioned above. To enable the analysis of data sets with large spatial and temporal extents, the approach's premise is that in such data sets, features of interest occur in different locations and times; hence these features can be defined as data subsets. The central concept of the STAS is the *analysis space*, which is a container for a feature of interest and the analytical contributions (e.g., snapshots and annotations) to make sense of it. An analysis space is defined by a spatiotemporal boundary containing the data subset, and by a description of the reason to consider it a feature of interest. The analysis space provides a well-defined context for the analysis of the feature of interest, including thematic (e.g., a species outbreak), spatial (e.g., within a given land extent), and temporal (e.g., weeks 30 to 35 of 2018) characterizations. Providing such context focuses the analysts' attention on the feature of interest, elicits relevant contributions, and produces knowledge about the features of interest and further about the phenomenon under study. Therefore, the analysis of the data set is partitioned into the analysis of several relevant data subsets.

To provide the STAS approach with flexibility, an analysis space can offer diverse collaboration techniques such as annotations, discussion board, instant messaging, interaction history, snapshot, storytelling, and combinations of several techniques. Additionally, the contributions created with any technique remain inside the analysis space, which helps to maintain the GVA environment organized. To this end, the STAS offers two working modes: overview and analysis. In the former, the analyst can explore the whole data set and visualize the description and spatiotemporal distribution of the features of interest. In the latter, a data subset is highlighted to focus the analyst's attention on a feature of interest, and the collaboration techniques are available to make sense of it.

The STAS approach enables incremental knowledge building. To this end, analysts can create links between analysis spaces, which promotes knowledge building upon previous contributions. This feature prevents analysis spaces and their contributions from being forgotten when the analysis effort advances and supports building on previously generated knowledge. These links can represent different types of relationships between the analysis spaces. To this end, the approach relies on labels to convey the meaning of the relationship, e.g., same type of event, opposite type of event, and similar data. The links can be created manually or automatically. In the former, the relationship is created by human input with the rationale that VA strongly relies on human perception and cognition to solve analytical problems in an intuitive manner [59, 47]. In the latter, the relationship is created from computer inference by a processing engine measuring similarity between analysis spaces based on methods such as textual content similarity and spatiotemporal data similarity. Given that analysts may have different criteria, it is important to measure the consensus on the relevance of the links between analysis spaces, in other words, consensus about whether two analysis spaces are relevant to each other. To enable this, STAS allows the analysts to vote in favor or against the relevance of the links.

6.1.4 From theory to practice: applied collaborative geovisual analytics

A case study was conducted with the support of stakeholders of the OFF management in Andalusia, Spain. This activity aimed to address the fourth research objective "Implement the software architecture and collaboration approach in a prototype for the monitoring and control of the Olive Fruit Fly and evaluate its usability and utility." There are three important outcomes from this exercise to be discussed in the following paragraphs: first, the software architecture can accommodate real-world requirements; second, the mapping of the STAS approach into the software architecture can be realized in a GVA system; and third, CGVA has potential as a tool for daily use in pest management.

To assess whether the software architecture is applicable to real-world scenarios, I collected and analyzed user requirements from the case study stakeholders. The requirements include different visualization products, interaction techniques, and configuration options, which are common in GVA systems. Additionally, stakeholders emphasized that two requirements are of key importance to support their pest management activities properly: first, to access the data hierarchically, and second, the capability to integrate statistical models.

The first requirement is that data access should follow a three-level hierarchy: year, week of the year, and monitoring location. The rationale is that pest management efforts are organized into campaigns that correspond to a crop production cycle. During the campaign, relevant variables that characterize the species' behavior are recorded at regular time intervals in relevant locations. Depending on the species under study, this hierarchy may change, but the concept of cyclic observation periods is widespread for pest population analyses.

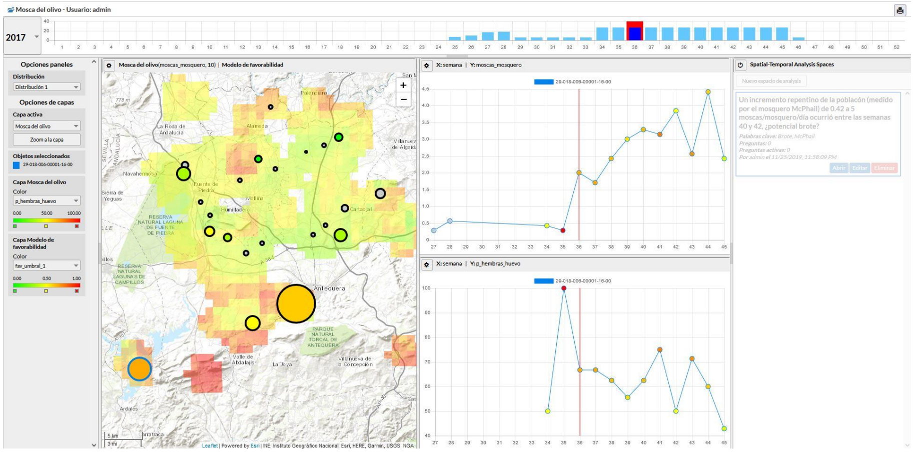
Regarding the second requirement, pest monitoring data enables producers to decide when and where to perform control actions. However, the collection of monitoring data is expensive and time-consuming. Thus, statistical models play a key role in the study of pest populations by

providing a means to understand the factors driving the population dynamics and to estimate pest abundance/presence in non-monitored locations/areas. For this reason, stakeholders emphasized that the system should include processing capabilities to model the pest behavior.
Additionally, regarding the software and hardware components of the prototype, stakeholders expressed their preference for a web-based system that they would like to access from desktop/laptop workstations and tablets. Finally, regarding the collaboration features, a simplified version of the STAS approach was implemented. By stakeholders' request, the collaboration technique in the STAS analysis mode is a question-based forum.

The software architecture was adequate to design and develop a web-based GVA prototype that fulfills all the stakeholders' requirements, providing evidence of the architecture's applicability. During the prototype's testing, the stakeholders were able to work independently with the prototype very soon after the initial training, showing that the prototype has a gentle learning curve. I observed the stakeholders experimenting with the various tools of the prototype to visualize known dynamics on the monitoring data and later using the resulting visualizations to explain how the prototype enables them to analyze the pest dynamics. All the stakeholders produced a visualization (five entirely on their own, and two with some assistance) to display the locations and times at which a particular threshold (i.e., [flies per flycatcher per day >= 1] AND [percentage of female flies with eggs >= 50%]) was exceeded. This threshold is of key importance for their pest management activities. In Figure 6.1, the interface displays this information, and a location is selected to assess how the two variables involved change through the year.

Additionally, I observed the stakeholders selecting diverse locations and comparing the favorability values (i.e., models' outputs) with different predictors, especially temperature, precipitation, and humidity. For example, Figure 6.2 suggests that favorability values are influenced by precipitation. From this exercise, they pointed out that while the selected predictors for the models are correct and the general patterns make sense with the known dynamics of the OFF population, in their opinion, the models seem to overestimate favorable conditions for pest development. In this regard, based on their domain and local knowledge, they identified diverse locations and times at which the topographic, environmental, and weather conditions were not as good for pest development as the models' outputs suggested. Additionally, by assessing the models' classification errors, it was observed that 80% of errors are false positives, which supports the stakeholders' opinion. These results suggest that visualization tools enable stakeholders to analyze the monitoring data and processing outputs and use them to explain their findings to others. Therefore, the prototype fulfills its objective of enabling collaborative analysis of the pest dynamics.

To enable the prototype to support collaborative analysis, it implements the STAS approach. The implemented version is a simplification of the description provided in Chapter 4 because the prototype was designed



Mosca del olivo - Usuario: admin

2017

Opciones paneles

Distribución: Distribución 1

Opciones de capas

Capa activa: Mosca del olivo

Zoom a la capa

Objetos seleccionados: 29-018-006-00001-16-00

Capa Mosca del olivo

Color: p_hembras_huevo

Capa Modelo de favorabilidad

Color: fav_umbral_1

Mosca del olivo(moscas_mosquero, 10) | Modelo de favorabilidad

X: semana | Y: moscas_mosquero

X: semana | Y: p_hembras_huevo

Spatial-Temporal Analysis Spaces

Nuevo espacio de analysis

Un incremento repentino de la poblacón (medido por el mosquero McPhail) de 0.42 a 5 moscas/mosquero/día ocurrió entre las semanas 40 y 42, ¿potencial brote?

Palabras clave: Brote, McPhail

Preguntas: 0

Preguntas activas: 0

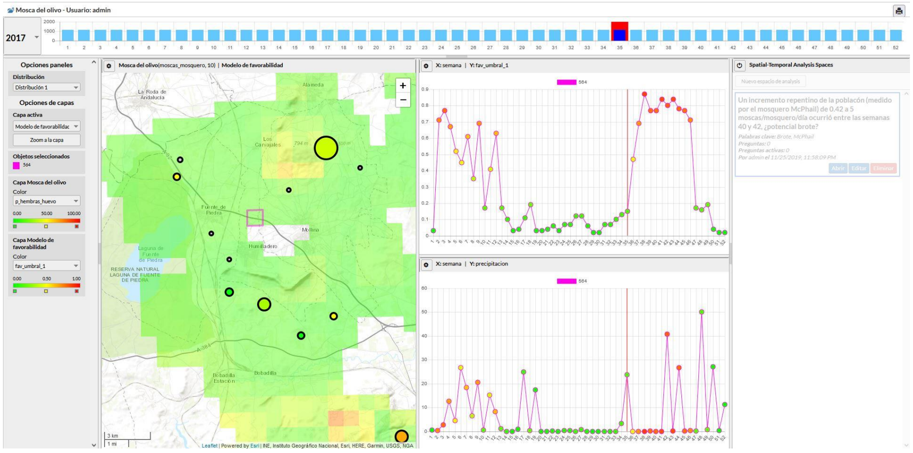
Por admin el 11/25/2019, 11:58:09 PM

Abrir Editar Eliminar

Figure 6.1 Stakeholders used the prototype to visualize the location and times at which the threshold was exceeded. For this aim, they configured the map to show the number of flies per flycatcher per day as the size of the circles, and the color ramp represents the percentage of female flies with eggs. Additionally, two line-charts were configured in most cases to show the evolution of the two variables through the year. Above: the number of flies per flycatcher per day, below: the percentage of female flies with eggs.

and developed based on stakeholders' requirements. Nevertheless, the implemented version allowed to assess the main functionality of the approach, which includes the analysis spaces, links between them, and collaborative analysis of features of interest. The prototype's evaluation showed that it was easy for the participants to understand and use the implemented features of the STAS approach. However, a long-term evaluation was out of scope due to time constraints. Such long-term evaluation of the STAS approach might lead to an improved design. In this respect, three potentially useful changes were identified during the prototype's testing. First, stakeholders highlighted that it is not straightforward in some cases to decide on the spatial and temporal boundaries for an analysis space because some features of interest do not necessarily have well-defined boundaries. This observation suggests that, in some cases, it might be necessary to use fuzzy boundaries to define an analysis space. Second, in the current design, an analyst can jump between related analysis spaces, and stakeholders mentioned that it would be useful to have an easy way to 'jump back.' In this context, it is important to design a more flexible mechanism to navigate the related analysis spaces, which could improve the potential to take advantage of previous contributions. Third, the prototype only offers a question-based forum. In an implementation with multiple collaboration techniques, it will be necessary to design a mechanism to synthesize the contributions from those multiple techniques.

The evaluation results suggest that CGVA has the potential to become a valuable tool in pest management. It enables stakeholders without



Mosca del olivo - Usuario: admin

2017

Opciones paneles

Distribución: Distribución 1

Opciones de capas

Capa activa: Modelo de favorabilidad

Objetos seleccionados: 564

Capa Mosca del olivo

Color: p_hembras_huevo

Capa Modelo de favorabilidad

Color: fav_umbral_1

Mosca del olivo(moscas_mosquero, 10) | Modelo de favorabilidad

X: semana | Y: fav_umbral_1

X: semana | Y: precipitacion

Spatial-Temporal Analysis Spaces

Un incremento repentino de la poblacón (medido por el mosquero McPhail) de 0.42 a 5 moscas/mosquero/día ocurrió entre las semanas 40 y 42, ¿potencial brote?

Palabras clave: Brote, McPhail

Preguntas: 0

Preguntas activas: 0

Por admin el 11/25/2019, 11:58:09 PM

Figure 6.2 Stakeholders used two line-charts to visualize the effect of the various predictors on the favorability values. In this example the favorability values are shown in the upper chart, and weekly precipitation is shown in the bottom one.

knowledge in GIS to work with complex geodata sets, which has the potential to improve their understanding of the spatial dynamics of a pest and further support the design of eco-friendly and cost-effective control strategies. The analysis activities that stakeholders performed with the OFF monitoring data were to visualize and explain: known dynamics of the species, relationships between diverse variables, and spatiotemporal characteristics of the species behavior. The prototype developed in this research enables interactive visualization of the OFF data, but it does not provide on-demand processing capabilities. However, on-demand processing is a common feature in GVA. In this sense, processing capabilities can facilitate the identification of subtle relationships that might be complex to identify only through visual analysis. Therefore, it would be interesting to include on-demand processing capabilities in a long-term evaluation exercise. GVA also enables stakeholders to work with the outputs of advanced processing methods and to use their domain and local knowledge to provide feedback to improve and validate such methods. Additionally, by enabling collaboration among stakeholders, varied knowledge and expertise enrich the analysis process, which can lead to a better understanding of the species dynamics and more effective decision-making regarding its monitoring and control. In the post-evaluation discussions with the stakeholders, they directly asked if there was a plan to continue with the research because they were willing to see the prototype becoming a production system supporting their pest management activities.

Finally, the experience during the case study suggests that most of the user requirements apply to many pest management efforts. Therefore, the prototype might be applicable to other application cases within this domain. To demonstrate this, future research should be conducted to assess the prototype's design against the requirements of diverse pest management efforts, which can also provide further evidence about the potential of CGVA to become an effective tool to support pest management.

6.2 Conclusions

The results of the literature review lead to three conclusions. First, CGVA is becoming more accessible to a broader audience, which is suggested by three developments: the most common collaborative scenario is asynchronous distributed, which promotes participation by removing the constraints on time and location to contribute; the increasing use of cloud technology, which is a common trend in analytics systems, and which improves the scalability of and distributed access to the system; and an increasing in the support for multiple devices, which eliminates the need for specialized hardware. Second, the collaboration techniques implemented in GVA systems support the entire process of data analysis. However, the techniques do not have a well-defined role in the analysis process; and there is a lack of features to synthesize the analytical contributions and represent the level of agreement regarding evidence and conclusions. Third, technology is evolving at an ever-increasing fast pace and is becoming more pervasive. For this reason, the concept of ubiquitous analytics may become a reality in the mid-term. To materialize this concept, it is necessary to address the three research challenges identified in Chapter 2, which are the lack of support for: hybrid collaborative scenarios, cross-device collaboration, and time-critical and long-term analysis. In a general sense, the literature review results lead to the conclusion that GVA is moving towards effective support of multi-disciplinary and cross-domain collaborative analysis.

The process of designing a software reference architecture and its implementation in a prototype lead to two conclusions: first, the proposed architecture constitutes a valuable tool for GVA researchers and practitioners. From a scientific standpoint, the architecture provides a framework to plan and execute research on the support for collaborative analysis in GVA, which may focus on diverse aspects such as the general structure of the system, collaborative analysis workflows, collaboration techniques, or technological platforms. For practitioners, it provides a reference model that can aid in analyzing, designing, developing, and evaluating GVA systems that enable multi-disciplinary collaborative analysis in diverse application domains. Second, given that GVA is a rapidly-evolving field, it is necessary to design and develop software systems that can evolve fast. For this reason, in the near future, GVA systems will likely be developed based on the micro-services architectural pattern. Currently, the upfront complexity and cost of implementing a micro-services-based architecture is a barrier to adopt it. However, as with any other technology, its maturity will lower this barrier, facilitating its adoption in diverse fields, including GVA.

The design of the STAS approach and its implementation in a prototype lead to the following conclusions: first, there are several application domains in which the analysis efforts can benefit or require to be addressed as long-term processes. Although it was not possible to perform a long-term testing of the prototype due to time constraints, the obtained results suggest that the approach is easy to understand and use for collaborative analysis. Second, the concept of analysis space is easy to understand and helps to focus the analyst's attention on a feature of interest, which has the potential to increase the completeness and accuracy of analysis results. However, in the proposed design, features of interest are assumed to have well-defined boundaries, which in the opinion of the testing users does not necessarily hold true. Third, the relationships between analysis spaces are effective as a mechanism to find relevant previous contributions and to promote their continued use in the analysis effort. Nevertheless, a more flexible navigation mechanism is needed to materialize the potential to take advantage of those previous contributions.

The case study provided an opportunity to realize the proposed software reference architecture and the collaboration approach (i.e., STAS) in a CGVA system prototype. The design, development, and testing of the prototype to analyze the OFF dynamics in Andalusia, Spain, provided evidence to conclude that the architecture and the collaboration approach are applicable to real-world analysis processes. The case study also provided evidence to conclude that CGVA has the potential to become an important tool in pest management. It enabled pest management stakeholders with and without experience working with GIS to analyze monitoring data and processing outputs, which can support the design of eco-friendly and cost-effective pest management strategies.

Bibliography

[1] P. Acevedo and R. Real. Favourability: concept, distinctive characteristics and potential usefulness. *Naturwissenschaften*, 99(7):515–522, 2012.

[2] G. Andrienko, N. Andrienko, P. Jankowski, D. Keim, M. Kraak, A. MacEachren, and S. Wrobel. Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science*, 21(8):839–857, 2007.

[3] G. Andrienko, N. Andrienko, D. Keim, A. M. MacEachren, and S. Wrobel. Challenging problems of geospatial visual analytics. *Journal of Visual Languages & Computing*, 22(4):251–256, 2011.

[4] S. Angelov, P. Grefen, and D. Greefhorst. A framework for analysis and design of software reference architectures. *Information and Software Technology*, 54(4):417–431, 2012.

[5] M. M. Arimoto and E. F. Barbosa. A systematic review of methods for developing open educational resources. In *20th International Conference on Computers in Education, ICCE 2012.*, pages 1–8, Singapore, 2012.

[6] P. Avgeriou and U. Zdun. Architectural patterns revisited — a pattern language. In *10th European Conference on Pattern Languages of Programs*, volume 81, pages 431–470, 2005.

[7] K. Babić, S. Martinšić-Ipšić, A. Meštrović, and F. Guerra. Short texts semantic similarity based on word embeddings. In *Central European Conference on Information and Intelligent Systems*, pages 27–33, Varaždin, Croatia, 2019.

[8] S. K. Badam, E. Fisher, and N. Elmqvist. Munin: A peer-to-peer middleware for ubiquitous analytics and visualization spaces. *IEEE Transactions on Visualization and Computer Graphics*, 21(2):215–228, 2015.

[9] A. Bangor, P. Kortum, and J. Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

[10] A. Bangor, P. T. Kortum, and J. T. Miller. An empirical evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, 24(6):574–594, 2008.

[11] L. Bass, P. Clements, and R. Kazman. *Software Architecture in Practice*. The SEI Series in Software Engineering. Pearson Education, Inc., United States of America, 3rd. edition, 2013.

[12] R. Benbunan-Fich, S. R. Hiltz, and M. Turoff. A comparative content analysis of face-to-face vs. asynchronous group decision making. *Decision Support Systems*, 34(4):457–469, 2003.

[13] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[14] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, Madison, WI, 1983.

[15] J. M. Bowers and S. D. Benford, editors. *Studies in Computer Supported Cooperative Work: Theory, Practice and Design*. North-Holland Publishing Co., NLD, 1990.

[16] J. Brooke. SUS — a quick and dirty usability scale. In P. Jordan, B. Thomas, I. McClelland, and B. Weerdmeester, editors, *Usability Evaluation In Industry*, chapter 21. CRC Press, London, 1986.

[17] J. Brooke. SUS: a retrospective. *Journal of usability studies*, 8(2):29–40, 2013.

[18] G. Broufas, M. Pappas, and D. Koveos. Effect of relative humidity on longevity, ovarian maturation, and egg production in the Olive Fruit Fly (Diptera: Tephritidae). *Annals of the Entomological Society of America*, 102:70–75, 2009.

[19] G. Cai and B. Yu. Spatial annotation technology for public deliberation. *Transactions in GIS*, 13:123–146, 2009.

[20] A. Çöltekin, S. Bleisch, G. Andrienko, and J. Dykes. Persistent challenges in geovisualization — a community perspective. *International Journal of Cartography*, 3(sup1):115–139, 2017.

[21] S. Chatterjee, T. Abhichandani, L. Haiqing, B. TuIu, and B. Jongbok. Instant messaging and presence technologies for college campuses. *IEEE Network*, 19(3):4–13, 2005.

[22] M. Chen, S. Mao, and Y. Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.

[23] L. Chi and X. Zhu. Hashing Techniques: A Survey and Taxonomy. *ACM Computing Surveys*, 50(1):37, 2017.

[24] J. Choi and Y. Tausczik. Characteristics of collaboration in the emerging practice of open data analysis. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, page 835–846. Association for Computing Machinery, 2017.

[25] P. Clements, D. Garlan, R. Little, R. Nord, and J. Stafford. *Documenting software architectures: views and beyond*. Proceedings of the 25th International Conference on Software Engineering. IEEE Computer Society, Portland, Oregon, 2003.

[26] R. De Amicis, G. Conti, S. Piffer, and B. Simões. *Geospatial visual analytics*. Springer, Dordrecht, 2009.

[27] J. Desjardins. The evolution of instant messaging. `http://www.visualcapitalist.com/evolution-instant-messaging/`, 2016. Accessed: 04/04/2017.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[29] N. Dragoni, S. Giallorenzo, A. L. Lafuente, M. Mazzara, F. Montesi, R. Mustafin, and L. Safina. Microservices: Yesterday, today, and tomorrow. In M. Mazzara and B. Meyer, editors, *Present and Ulterior Software Engineering*, pages 195–216. Springer International Publishing, 2017.

[30] R. Eccles, T. Kapler, R. Harper, and W. Wright. Stories in GeoTime. *Information Visualization*, 7(1):3–17, 2008.

[31] S. G. Eick, M. A. Eick, J. Fugitt, B. Horst, M. Khailo, and R. A. Lankenau. Thin client visualization. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, 2007.

[32] M. Elias, M.-A. Aufaure, and A. Bezerianos. Storytelling in visual analytics tools for business intelligence. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, editors, *IFIP Conference on Human-Computer Interaction*, pages 280–297, Berlin, Heidelberg, 2013. Springer.

[33] N. Elmqvist. Visualization reloaded: Redefining the scientific agenda for visualization research. In *Proceedings of HCI Korea*, HCIK'15, pages 132–137, South Korea, 2014. Hanbit Media, Inc.

[34] N. Elmqvist and P. Irani. Ubiquitous Analytics: Interacting with big data anywhere, anytime. *Computer*, 46(4):86–89, 2013.

[35] Extenda. Estudio del sector del aceite de oliva de Andalucía. Report, Agencia Andaluza de Promoción Exterior — extenda, 2017.

[36] FAO. Impact of climate change, pests and diseases on food security and poverty reduction. Report, FAO, 2005.

[37] FAO. FAOSTAT: crops. http://www.fao.org/faostat/en/#data/QC, n.d. Accessed: 30-11-2017.

[38] FAO. Integrated Pest Management. http://www.fao.org/agriculture/crops/core-themes/theme/pests/ipm/en/, n.d. Accessed: 30-11-2017.

[39] T. Fechner, D. Wilhelm, and C. Kray. *Ethermap: Real-Time Collaborative Map Editing*, page 3583–3592. Association for Computing Machinery, New York, NY, USA, 2015.

[40] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, San Francisco, CA, United States, 2007. Morgan Kaufmann Publishers Inc.

[41] G. A. García-Chapeton, F. O. Ostermann, R. A. de By, and M.-J. Kraak. Enabling collaborative geovisual analytics: Systems, techniques, and research challenges. *Transactions in GIS*, 22(3):640–663, 2018.

[42] D. Garn. An introduction to hashing and checksums in Linux. https://www.redhat.com/sysadmin/hashing-checksums, 2021. Accessed: 05/03/2021.

[43] G. Gilioli, S. Pasquali, and E. Marchesini. A modelling framework for pest population dynamics and management: An application to the Grape Berry Moth. *Ecological Modelling*, 320:348–357, 2016.

[44] S. Grainger, F. Mao, and W. Buytaert. Environmental data visualisation for non-scientific contexts: Literature review and design framework. *Environmental Modelling & Software*, 85:299–318, 2016.

[45] L. Gupta. Differences Between Word2Vec and BERT. https://medium.com/swlh/differences-between-word2vec-and-bert-c08a3326b5d1, 2020. Accessed: 25/05/2021.

[46] M. Haklay. Citizen science and volunteered geographic information: Overview and typology of participation. In D. Sui, S. Elwood, and M. Goodchild, editors, *Crowdsourcing Geographic Knowledge*, chapter 7, pages 105–122. Springer, Netherlands, 2013.

[47] E. Hand. Citizen science: People power. *Nature*, 466(7307):685-687, 2010.

[48] F. Hardisty. Geojabber: enabling geo-collaborative visual analysis. *Cartography and Geographic Information Science*, 36(3):267–280, 2009.

[49] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain. *Semantic Similarity from Natural Language and Ontology Analysis*, volume 8 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, 2015.

[50] J. Heard. Geoanalytics. Report TR-11-03, Renaissance Computing Institute, Chapel Hill, North Carolina 27517, 2011.

[51] J. Heard, S. Thakur, J. Losego, and K. Galluppi. Big Board: Teleconferencing over maps for shared situational awareness. *Computer Supported Cooperative Work (CSCW)*, 23(1):51–74, 2013.

[52] J. Heer and M. Agrawala. Design considerations for collaborative visual analytics. *Information Visualization*, 7(1):49–62, 2008.

[53] J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE transactions on visualization and computer graphics*, 14(6):1189–1196, 2008.

[54] J. Heer and B. Shneiderman. Interactive dynamics for visual analysis. *Queue*, 10(2):30–55, 2012.

[55] J. Heer, F. van Ham, S. Carpendale, C. Weaver, and P. Isenberg. Creation and collaboration: Engaging new audiences for information visualization. In A. Kerren, J. Stasko, J.-D. Fekete, and C. North, editors, *Information Visualization*, volume 4950 of *Lecture Notes in Computer Science*, chapter 5, pages 92–133. Springer, Berlin Heidelberg, 2008.

[56] J. Heer, F. B. Viégas, and M. Wattenberg. Voyagers and voyeurs: Supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'07, New York, 2007. ACM.

[57] D. Herzfeld and K. Sargent. Chapter 1: Integrated pest management. In N. Goodman, editor, *Private Pesticide Applicator Safety Education Manual*, chapter 1. University of Minnesota Extension, Minnesota, USA, 19th edition, 2017.

[58] A. J. Hey, S. Tansley, and K. M. Tolle. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft Research, Redmond, WA, 2009.

[59] Q. Ho. *Architecture and Applications of a Geovisual Analytics Framework*. PhD thesis, Linköping University, Sweden, 2013.

[60] Q. Ho, P. Lundblad, T. Astrom, and M. Jern. A web-enabled visualization toolkit for geovisual analytics. *Information Visualization*, 11(1):22–42, 2012.

[61] J. Hollan, E. Hutchins, and D. Kirsh. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(2):174–196, 2000.

[62] S. Hopfer and A. M. MacEachren. Leveraging the potential of geospatial annotations for collaboration: a communication theory perspective. *International Journal of Geographical Information Science*, 21(8):921–934, 2007.

[63] D. W. Hosmer and S. Lemeshow. *Applied logistic regression*. Wiley series in probability and statistics. John Wiley & Sons, Inc., second edition edition, 2000.

[64] M. S. Hossain, M. Masud, G. Muhammad, M. Rawashdeh, and M. Mehedi Hassan. Automated and user involved data synchronization in collaborative e-health environments. *Computers in Human Behavior*, 30:485–490, 2014.

[65] IBM. IBM Watson Analytics. `https://www.ibm.com/analytics/watson-analytics/us-en/`, n.d. Accessed: 2017-02-15.

[66] P. Isenberg, N. Elmqvist, J. Scholtz, D. Cernea, K.-L. Ma, and H. Hagen. Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization*, 10(4):310–326, 2011.

[67] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. vispubdata.org: A metadata collection about IEEE visualization (VIS) publications. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, 2017.

[68] C. Jackson. Micro Frontends. `https://martinfowler.com/articles/micro-frontends.html`, 2019. Accessed: 05/03/2021.

[69] D. H. Jeong, S. Y. Ji, E. A. Suma, B. Yu, and R. Chang. Designing a collaborative visual analytics system to support users' continuous analytical processes. *Human-centric Computing and Information Sciences*, 5(1), 2015.

[70] M. Jern. Collaborative explorative data analysis applied in HTML. In Y. Luo, editor, *Cooperative Design, Visualization, and Engineering*, volume 5220 of *Lecture Notes in Computer Science*, pages 36–44. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[71] M. Jern. Collaborative web-enabled geoanalytics applied to OECD regional data. In Y. Luo, editor, *Cooperative Design, Visualization, and Engineering*, volume 5738 of *Lecture Notes in Computer Science*, chapter 5, pages 32–43. Springer Berlin Heidelberg, 2009.

[72] M. Jern. Explore, collaborate and publish official statistics for measuring regional progress. In Y. Luo, editor, *Cooperative Design, Visualization, and Engineering*, volume 6240 of *Lecture Notes in Computer Science*, pages 189–198. Springer Berlin Heidelberg, 2010.

[73] S. Jeske. Google BERT Update and What You Should Know. `https://blog.marketmuse.com/google-bert-update/`, 2019. Accessed: 25/05/2021.

[74] C. Jing, Y. Zhu, J. Fu, and M. Dong. A lightweight collaborative gis data editing approach to support urban planning. *Sustainability*, 11(16):4437, 2019.

[75] R. Johansen. *GroupWare: Computer Support for Business Teams*. The Free Press, 1988.

[76] C. T. Joseph and K. Chandrasekaran. Straddling the crevasse: A review of microservice software architecture foundations and recent advancements. *Software: Practice and Experience*, 49(10):1448–1484, 2019.

[77] R. Kalamatianos, K. Kermanidis, M. Avlonitis, and I. Karydis. Environmental impact on predicting Olive Fruit Fly population using trap measurements. In L. Iliadis and I. Maglogiannis, editors, *Artificial Intelligence Applications and Innovations*, pages 180–190. Springer International Publishing, 2016.

[78] M. Kassab, M. Mazzara, J. Lee, and G. Succi. Software architectural patterns in practice: an empirical study. *Innovations in Systems and Software Engineering*, 14(4):263–271, 2018.

[79] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering the Information Age: Solving Problems with Visual Analytics*. Eurographics Association, Goslar, Germany, 2010.

[80] D. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. *Visual Analytics: Scope and Challenges*, volume 4404 of *Lecture Notes in Computer Science*, book section 6, pages 76–90. Springer Berlin Heidelberg, 2008.

[81] T. Kijewski-Correa, N. Smith, A. Taflanidis, A. Kennedy, C. Liu, M. Krusche, and C. Vardeman. CyberEye: Development of integrated cyber-infrastructure to support rapid hurricane risk assessment. *Journal of Wind Engineering and Industrial Aerodynamics*, 133:211–224, 2014.

[82] K. Kim, W. Javed, C. Williams, N. Elmqvist, and P. Irani. Hugin: A framework for awareness and coordination in mixed-presence collaborative information visualization. In *ACM International Conference on Interactive Tabletops and Surfaces*, ITS '10, page 231–240, New York, NY, USA, 2010. Association for Computing Machinery.

[83] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, Keele University, Keele, UK, 2007.

[84] M. Koch, G. Schwabe, and R. O. Briggs. CSCW and Social Computing. *Business & Information Systems Engineering*, 57(3):149–153, 2015.

[85] P. B. Kruchten. The 4+1 view model of architecture. *IEEE software*, 12(6):42–50, 1995.

[86] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[87] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. Technical Report 949, META Group, 2001.

[88] P. Lundblad. *Applied Geovisual Analytics and Storytelling.* PhD thesis, Linköping University, Sweden, 2013.

[89] K. Luther, S. Counts, K. B. Stecher, A. Hoff, and P. Johns. Pathfinder: An online collaboration environment for citizen scientists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 239–248, New York, NY, USA, 2009. ACM.

[90] A. M. MacEachren. Cartography and GIS: facilitating collaboration. *Progress in Human Geography*, 24(3):445–456, 2000.

[91] R. Macefield. How to specify the participant group size for usability studies: a practitioner's guide. *Journal of Usability Studies*, 5(1):34–45, 2009.

[92] N. Mahyar. *Supporting Sensemaking during Collocated Collaborative Visual Analytics.* PhD thesis, University of Victoria, Canada, 2014.

[93] D. Marchini, R. Petacchi, and S. Marchi. *Bactrocera Oleae* reproductive biology: New evidence on wintering wild populations in olive groves of Tuscany (Italy). *Bulletin of Insectology*, 70:121–128, 2017.

[94] D. W. Marquaridt. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):591–612, 1970.

[95] T. Marrinan, J. Leigh, L. Renambot, A. Forbes, S. Jones, and A. E. Johnson. Mixed presence collaboration using scalable visualizations in heterogeneous display spaces. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 2236–2245, New York, 2017. ACM.

[96] T. Marrinan, L. Renambot, J. Leigh, A. Forbes, S. Jones, and A. E. Johnson. Synchronized mixed presence data-conferencing using large-scale shared displays. In *Proceedings of the 2016 ACM on Interactive Surfaces and Spaces*, ISS '16, pages 355–360, New York, 2016. ACM.

[97] A. Mathisen. *Collaborative Visual Analytics: Leveraging Mixed Expertise in Data Analysis.* PhD thesis, Aarhus University, Denmark, 2019.

[98] J. McIntosh and M. Yuan. Assessing similarity of geographic processes and events. *Transactions in GIS*, 9(2):223–245, 2005.

[99] Microsoft. Creating composite UI based on microservices. `https://docs.microsoft.com/en-us/dotnet/architecture/microservices/architect-microservice-container-applications/microservice-based-composite-ui-shape-layout`, 2021. Accessed: 05/03/2021.

[100] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, 2013.

[101] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

[102] R. Miller. IBM's new Watson Analytics want to bring big data to the mases. https://techcrunch.com/2014/09/16/ibms-new-watson-analytics-wants-to-bring-data-to-the -masses/, 2014. Accessed: 21/08/2017.

[103] P. Molli, H. Skaf-Molli, G. Oster, and S. Jourdain. SAMS: synchronous, asynchronous, multi-synchronous environments. In *The 7th International Conference on Computer Supported Cooperative Work in Design*, pages 80–84, 2002.

[104] R. T. Monroe, A. Kompanek, R. Melton, and D. Garlan. Architectural styles, design patterns, and objects. *IEEE Software*, 14(1):43–52, 1997.

[105] D. Montgomery and E. Peck. *Introduction to linear regression analysis*. Wiley, 1982.

[106] A.-R. Muñoz, A. Jiménez-Valverde, A. L. Márquez, M. Moleón, and R. Real. Environmental favourability as a cost-efficient tool to estimate carrying capacity. *Diversity and Distributions*, 21(12):1388–1400, 2015.

[107] F. Nardi, A. Carapelli, R. Dallai, G. K. Roderick, and F. Frati. Population structure and colonization history of the Olive Fly, *Bactrocera oleae* (Diptera, Tephritidae). *Molecular Ecology*, 14(9):2729–38, 2005.

[108] NASA. About: Landsat Then and Now. `https://landsat.gsfc.nasa.gov/about`. Accessed: 09/05/2021.

[109] T. Neumayr, H.-C. Jetter, M. Augstein, J. Friedl, and T. Luger. Domino: A descriptive framework for hybrid collaboration and coupling styles in partially distributed teams. *Proceedings of the ACM Human-Computer Interaction*, 2, 2018.

[110] J. Olivero, J. E. Fa, R. Real, A. L. Márquez, M. A. Farfán, J. M. Vargas, D. Gaveau, M. A. Salim, D. Park, J. Suter, S. King, S. A. Leendertz, D. Sheil, and R. Nasi. Recent loss of closed forests is associated with ebola virus disease outbreaks. *Scientific Reports*, 7(1):14291, 2017.

[111] J. Olivero, E. J. García, M. E. Wong, and J. P. Ros. Ensayo de eficacia de diferentes combinaciones soporte-atrayente para el trampeo de "Bactrocera oleae" (Gmel.), Mosca del Olivo. *Boletín de Sanidad Vegetal - Plagas*, 30(2):439–450, 2004.

[112] J. Olivero, A. G. Toxopeus, A. K. Skidmore, and R. Real. Testing the efficacy of downscaling in species distribution modelling: a comparison between MaxEnt and favourability function models. *Animal biodiversity and conservation*, 39(1):99–114, 2016.

[113] OpenStreetMap Wiki. Main Page — OpenStreetMap Wiki, . `https://wiki.openstreetmap.org/wiki/Main_Page`, 2020. Accessed: 09/05/2021.

[114] OpenStreetMap Wiki. Map features — OpenStreetMap Wiki, . `https://wiki.openstreetmap.org/wiki/Map_features`, 2021. Accessed: 09/05/2021.

[115] ORACLE. ORACLE Business Analytics. https://www.oracle.com/solutions/ business-analytics/index.html, 2017. Accessed: 2017-02-15.

[116] F. O. Ostermann and C. Granell. Advancing science with VGI: Reproducibility and replicability of recent studies using VGI. *Transactions in GIS*, 21(2):224–237, 2017.

[117] J. Paritsis and T. Veblen. Outbreak species. In R. Craig, J. Nagle, B. Pardy, O. Schmitz, and W. Smith, editors, *The Berkshire Encyclopedia of Sustainability*, volume Vol. 5: Ecosystem Management and Sustainability. Berkshire Publishing, Great Barrington, MA., 2012.

[118] M. Petronzio. A brief history of instant messaging. `http://mashable.com/2012/10/25/instant-messaging-history/`, 2012. Accessed: 2017-02-15.

[119] C. M. Pontikakos, T. A. Tsiligiridis, and M. E. Drougka. Location-aware system for Olive Fruit Fly spray control. *Computers and Electronics in Agriculture*, 70(2):355–368, 2010.

[120] N. Preguiça, J. L. Martins, H. Domingos, and S. Duarte. Integrating synchronous and asynchronous interactions in groupware applications. In H. Fuks, S. Lukosch, and A. Salgado, editors, *Groupware: Design, Implementation, and Use (CRIWG 2005)*, volume 3706, pages 89–104, Berlin, Heidelberg, 2005. Springer.

[121] P. Proulx, S. Tandon, A. Bodnar, D. Schroh, R. Harper, and W. Wright. Avian flu case study with nSpace and GeoTime. In *2006 IEEE Symposium On Visual Analytics Science and Technology*, pages 27–34. IEEE, 2007.

[122] QLikTech. QLik data analytics. `https://www.qlik.com/us/`, 2017. Accessed: 2017-02-15.

[123] M. Rafikov and J. M. Balthazar. Optimal pest control problem in population dynamics. *Computational & Applied Mathematics*, 24:65–81, 2005.

[124] C. Rahhal, H. Skaf-Molli, P. Molli, and S. Weiss. Multi-synchronous collaborative semantic wikis. In G. Vossen, D. D. E. Long, and J. X.

Yu, editors, *Web Information Systems Engineering - WISE 2009*, pages 115–129. Springer Berlin Heidelberg, 2009.

[125] L. Ramos, L. Silva, M. Y. Santos, and J. M. Pires. Detection of road accident accumulation zones with a visual analytics approach. *Procedia Computer Science*, 64:969–976, 2015.

[126] R. Real, A. M. Barbosa, and J. M. Vargas. Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, 13(2):237–245, 2006.

[127] D. Reinsel, J. Gantz, and J. Rydning. The digitization of the world: From edge to core. Technical Report US44413318, IDC, 2018.

[128] D. Ren, M. Brehmer, B. Lee, T. Höllerer, and E. K. Choe. ChartAccent: Annotation for data-driven storytelling. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pages 230–239, 2017.

[129] R. E. Rice. Bionomics of the Olive Fruit Fly Bactrocera (Dacus) Oleae. *Plant Protection Quarterly*, 10:1–5, 2000.

[130] M. Richards. *Software architecture patterns*. O'Reilly Media, Incorporated, 2015.

[131] C. Richardson. Pattern: Microservice Architecture. `https://microservices.io/patterns/microservices.html`, 2020. Accessed: 10/09/2020.

[132] C. Rinner. Argumentation Maps: GIS-based discussion support for on-line planning. *Environment and Planning B: Planning and Design*, 28(6):847–863, 2001.

[133] C. Rinner, C. Keßler, and S. Andrulis. The use of Web 2.0 concepts to support deliberation in spatial decision-making. *Computers, Environment and Urban Systems*, 32(5):386–395, 2008.

[134] A. Robinson and C. Weaver. Re-visualization: Interactive visualization of the process of visual analysis. In *Workshop on Visualization, Analytics & Spatial Decision Support at the 2006 GIScience conference*, 2006.

[135] D. Romero, J. C. Báez, F. Ferri-Yáñez, J. J. Bellido, and R. Real. Modelling favourability for invasive species encroachment to identify areas of native species vulnerability. *The Scientific World Journal*, 2014, 2014.

[136] R. E. Roth, K. Ross, and A. MacEachren. User-centered design for interactive maps: A case study in crime analysis. *ISPRS International Journal of Geo-Information*, 4(1):262, 2015.

[137] D. Sacha, A. Stoffel, F. Stoffel, K. Bum Chul, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1604–1613, 2014.

[138] SAP. Business Intelligence (BI) Solutions. http://go.sap.com/solution/platform-technology/analytics/business-intelligence-bi.html, n.d. Accessed: 2017-02-15.

[139] SAS. SAS Visual Analytics — business visualization: Dashboards, reporting and approachable analytics – all from one interface. USA, 2015.

[140] SAS. SAS Visual Analytics. http://www.sas.com/en_us/software/business-intelligence/visual-analytics.html, n.d. Accessed: 2017-02-15.

[141] S. Scheider, F. O. Ostermann, and B. Adams. Why good data analysts need to be critical synthesists. determining the role of semantics in data analysis. *Future Generation Computer Systems*, 72:11–22, 2017.

[142] A. Sharma, M. Kumar, and S. Agarwal. A complete survey on software architectural styles and patterns. *Procedia Computer Science*, 70:16–28, 2015.

[143] A. Shatte, J. Holdsworth, and I. Lee. Multi-synchronous collaboration between desktop and mobile users: A case study of report writing for emergency management. *CoRR*, abs/1606.00750, 2016.

[144] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[145] J. M. Six and R. Macefield. How to determine the right number of participants for usability studies. https://www.uxmatters.com/mt/archives/2016/01/how-to-determine-the-right-number-of-participants-for-usability-studies.php, 2016. Accessed: 15/11/2019.

[146] K. Software. Microservices: Patterns for enterprise agility and scalability. White paper, Keyhole Software, 2017.

[147] E. Steiger, J. P. de Albuquerque, and A. Zipf. An advanced systematic literature review on spatiotemporal analyses of Twitter data. *Transactions in GIS*, 19(6):809–834, 2015.

[148] D. Stodder. Visual analytics for making smarter decisions faster: applying self-service business intelligence technologies to data-driven objectives. Technical report, TDWI, USA, 2015.

[149] D. Sui, M. Goodchild, and S. Elwood. Volunteered geographic information, the exaflood, and the growing digital divide. In D. Sui, S. Elwood, and M. Goodchild, editors, *Crowdsourcing Geographic Knowledge*, chapter 1, pages 1–12. Springer, Netherlands, 2013.

[150] A. Sun. Enabling collaborative decision-making in watershed management using cloud-computing services. *Environmental Modelling & Software*, 41:93–97, 2013.

[151] G.-D. Sun, Y.-C. Wu, R.-H. Liang, and S.-X. Liu. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867, 2013.

[152] T. Susi and T. Ziemke. Social cognition, artefacts, and stigmergy: A comparative analysis of theoretical frameworks for the understanding of artefact-mediated collaborative activity. *Cognitive Systems Research*, 2(4):273–290, 2001.

[153] Tableau. Tableau. `http://www.tableau.com/`, 2017. Accessed: 2017-02-15.

[154] R. Tapia-McClung. Exploring the use of a spatio-temporal city dashboard to study criminal incidence: A case study for the mexican State of Aguascalientes. *Sustainability*, 12(6), 2020.

[155] R. Terra, M. T. Valente, K. Czarnecki, and R. S. Bigonha. Recommending refactorings to reverse software architecture erosion. In *2012 16th European Conference on Software Maintenance and Reengineering*, pages 335–340, 2012.

[156] J. Thomas and K. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005.

[157] TIBCO. TIBCO Spotfire. `http://spotfire.tibco.com/`, 2017. Accessed: 2017-02-15.

[158] B. M. Tomaszewski, A. C. Robinson, C. Weaver, M. Stryker, and A. M. MacEachren. Geovisual analytics and crisis management. In *Proceedings of the 4th International ISCRAM Conference*, pages 173–179. Delft, the Netherlands, 2007.

[159] Uncharted. GeoTime. `http://geotime.com/`, 2017. Accessed: 2017-02-15.

[160] UNISYS. `http://weather.unisys.com/hurricane/`, 2019. Accessed: 04/01/2016.

[161] D. Vesset. Cloud business analytics: A step closer to pervasive adoption of decision support services. White paper, IDC, Framingham, MA 01701 USA, 2016.

[162] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. ManyEyes: a site for visualization at internet scale. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1121–1128, 2007.

[163] P. Vossen. Olive oil: History, production, and characteristics of the world's classic oils. *HortScience*, 42(5):1093, 2007.

[164] A. Weinberger and F. Fischer. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1):71–95, 2006.

[165] W. Willett, J. Heer, J. Hellerstein, and M. Agrawala. CommentSpace: Structured support for collaborative visual analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 3131–3140, New York, NY, USA, 2011. ACM.

[166] D. T. Williams, N. Straw, M. Townsend, A. S. Wilkinson, and A. Mullins. Monitoring Oak Processionary Moth *Thaumetopoea Processionea L.* using pheromone traps: the influence of pheromone lure source, trap design and height above the ground on capture rates. *Agricultural and Forest Entomology*, 15(2):126–134, 2013.

[167] P. C. Wong and J. Thomas. Visual analytics. *Computer Graphics and Applications, IEEE*, 24(5):20–21, 2004.

[168] A. Wu, G. Convertino, C. Ganoe, J. M. Carroll, and X. Zhang. Supporting collaborative sense-making in emergency management through geo-visualization. *International Journal of Human-Computer Studies*, 71(1):4–23, 2013.

[169] E. M. A. Xavier, F. J. Ariza-López, and M. A. Ureña-Cámara. A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys*, 49(2):Article 39, 2016.

[170] Z. Yovcheva, C. van Elzakker, and B. Köbben. User requirements for geo-collaborative work with spatio-temporal data in a web-based virtual globe environment. *Applied Ergonomics*, 44(6):929–939, 2013.

[171] N. M. Yusoff and S. S. Salim. A systematic review of shared visualisation to achieve common ground. *Journal of Visual Languages & Computing*, 28:83–99, 2015.

[172] F. G. Zalom, R. A. Van Steenwyk, H. J. Burranckand, and M. W. Johnson. Olive Fruit Fly. *Pest Notes*, 2009.

[173] T. Zhang, J. Wang, C. Cui, Y. Li, W. He, Y. Lu, and Q. Qiao. Integrating geovisual analytics with machine learning for human mobility pattern discovery. *ISPRS International Journal of Geo-Information*, 8(10):434, 2019.

[174] P. Zikopoulos and C. Eaton. *Understanding Big Data: Analytics for enterprise class Hadoop and streaming data.* McGraw-Hill Osborne Media, 2011.

Search and selection of papers for systematic review A

In the following tables, the column “Duplicated” indicates if a paper was found in more than one database; the column “Removed” indicates if a duplicated paper was removed or kept in that database (every duplicated paper was kept in one of the databases); and the column “Selected” indicates if after manual screening the paper was kept, and further used to identify systems and techniques for the review.

A.1 ACM Digital Library

- Database: ACM Digital Library
- URL: https://dl.acm.org/
- Search procedure:
  1. From search in main page:
     - (collaborative OR cooperative) AND ("geovisual analytics" OR (geospatial AND "visual analytics") OR geoanalytics)
  2. Click on “Expand your search to the ACM Guide to Computing Literature”
  3. From the result select the papers that comply with the inclusion criteria

Table A.1 Search and selection results for ACM Digital Library

#	Paper	Duplicated	Removed	Selected
1	**Thin Client Visualization** Eick, S. G.; Eick, M. A.; Fugitt, J.; Horst, B.; Khailo, M. & Lankenau, R. A. (2007)	No	—	Yes

2	**Leveraging the potential of geospatial annotations for collaboration: a communication theory perspective** Hopfer, S. & MacEachren, A. M. (2007)	Yes	Yes	—
3	**Integrating InfoVis and GeoVis Components** Jern, M. & Franzen, J. (2007)	Yes	Yes	—
4	**nSpace and GeoTime: A VAST 2006 Case Study** Proulx, P.; Chien, L.; Harper, R.; Schroh, R.; Kapler, T.; Jonker, D. & Wright, W. (2007)	No	—	Yes
5	**Visual exploration and analysis of historic hotel visits** Weaver, C.; Fyfe, D.; Robinson, A.; Holdsworth, D.; Peuquet, D. & MacEachren, A. M. (2007)	No	—	No
6	**Collaborative Explorative Data Analysis Applied in HTML** Jern, M. (2008)	Yes	Yes	—
7	**Visual Analytics Presentation Tools Applied in HTML Documents** Jern, M.; Rogstadius, J.; Åström, T. & Ynnerman, A. (2008)	Yes	Yes	—
8	**Collaborative web-enabled geoanalytics applied to OECD regional data** Jern, M. (2009)	Yes	Yes	—
9	**Interactive Visualization of Weather and Ship Data** Lundblad, P.; Eurenius, O. & Heldring, T. (2009)	No	—	No
10	**Space, time and visual analytics** Andrienko, G.; Andrienko, N.; Demsar, U.; Dransch, D.; Dykes, J.; Fabrikant, S. I.; Jern, M.; Kraak, M-J.; Schumann, H. & Tominski, C. (2010)	No	—	Yes
11	**Explore, collaborate and publish official statistics for measuring regional progress** Jern, M. (2010)	Yes	Yes	—
12	**Swedish Road Weather Visualization** Lundblad, P.; Thoursie, J. & Jern, M. (2010)	No	—	No
13	**Omnidirectional 3D Visualization for the Analysis of a Large-Scale Corpus: Tripitaka Koreana** Kenderdine, S.; Lancaster, L.; Lan, H. & Gremmler, T. (2011)	No	—	No
14	**A web-enabled visualization toolkit for geovisual analytics** Ho, Q. V.; Lundblad, P.; Åström, T. & Jern, M. (2012)	Yes	Yes	—
15	**Geovisual Analytics and Storytelling Using HTML5** Lundblad, P. & Jern, M. (2013)	Yes	Yes	—

16	**Interactive visual summaries for detection and assessment of spatiotemporal patterns in geospatial time series** Kothur, P.; Sips, M.; Unger, A.; Kuhlmann, J. & Dransch, D. (2014)	No	—	No
17	**CyberGIS for data-intensive knowledge discovery** Wang, S.; Hu, H.; Lin, T.; Liu, Y.; Padmanabhan, A. & Soltani, K. (2015)	No	—	No
18	**nu-view: a visualization system for collaborative co-located analysis of geospatial disease data** Masoodian, M.; Luz, S. & Kavenga, D. (2016)	No	—	Yes
19	**Predicting the visualization intensity for interactive spatio-temporal visual analytics: a data-driven view-dependent approach** Li, J.; Zhang, T.; Liu, Q. & Yu, M. (2017)	No	—	No

A.2 GeoBase

- Database: GeoBase
- URL: https://www.engineeringvillage.com/
- Search procedure:
  1. From search in main page:
     - (collaborative OR cooperative) AND ("geovisual analytics" OR "geospatial visual analytics" OR geoanalytics)
     - Data -> Published: 2004 to 2017
  2. From the result select the papers that comply with the inclusion criteria

Table A.2 Search and selection results for GeoBase

#	Paper	Duplicated	Removed	Selected
1	**Leveraging the potential of geospatial annotations for collaboration: A communication theory perspective** Hopfer, S. & MacEachren, A.M. (2007)	Yes	No	Yes
2	**An XML-based infrastructure to enhance collaborative geographic visual analytics**	Yes	No	Yes

	Kramis, M.; Gabathuler, C.; Fabrikant, S. I. & Waldvogel, M. (2009)			
3	**Tropical cyclone trend analysis using enhanced parallel coordinates and statistical analytics** Steed, C. A.; Fitzpatrick, P. J.; Swan, J. E. & Jankun-Kelly, T. J. (2009)	No	—	No

A.3 IEEE Xplore

- Database: IEEE Xplore
- URL: http://ieeexplore.ieee.org/
- Search procedure:
    1. From advanced search:
        - collaborative OR cooperative in Metadata Only
        - "geovisual analytics" OR "geospatial visual analytics" OR geoanalytics in Metadata Only
        - Specify Year Range From 2004 To 2017
    2. From the result select the papers that comply with the inclusion criteria

Table A.3 Search and selection results for IEEE Xplore

#	**Paper**	**Duplicated**	**Removed**	**Selected**
1	**"GeoAnalytics" - Exploring spatio-temporal and multivariate data** Jern, M. & Franzen, J. (2006)	No	—	No
2	**Integrating InfoVis and GeoVis Components** Jern, M. & Franzen, J. (2007)	Yes	No	No
3	**The GAV Toolkit for Multiple Linked Views** Jern, M.; Johansson, S.; Johansson, J. & Franzen, J. (2007)	No	—	No
4	**Evacuation trace Mini Challenge award: Tool integration analysis of movements with Geospatial Visual Analytics Toolkit** Andrienko, N. & Andrienko, G. (2008)	No	—	No
5	**Exploratory 3D geovisual analytics** Ho, Q. V. & Jern, M. (2008)	No	—	No

6	**GeoAnalytics Tools Applied to Large Geospatial Datasets** Jern, M.; Åström, T. & Johansson, S. (2008)	No	—	No
7	**Visual Analytics Presentation Tools Applied in HTML Documents** Jern, M.; Rogstadius, J.; Åström, T. & Ynnerman, A. (2008)	Yes	No	Yes
8	**Time-based Geographical Mapping of Communicable Diseases** Cesario, M.; Jervis, M.; Luz, S.; Masoodian, M. & Rogers, B. (2012)	No	—	No
9	**A geospatial analytical system for mapping global medium-term earthquake probabilities** Zhan, F. B.; Cai, Z.; Zhu, Y. & Zhou, J. (2012)	No	—	No
10	**Geovisual Analytics and Storytelling Using HTML5** Lundblad, P. & Jern, M. (2013)	Yes	No	Yes

A.4 Science Direct

- Database: Science Direct
- URL: https://www.sciencedirect.com/
- Search procedure:
  1. From expert search:
     - (collaborative OR cooperative) AND ("geovisual analytics" OR "geospatial visual analytics" OR geoanalytics)
     - Year: 2004 to 2017
  2. From the result select the papers that comply with the inclusion criteria

Table A.4 Search and selection results for Science Direct

#	Paper	Duplicated	Removed	Selected
1	**Collaborative GIS for spatial decision support and visualization** Balram, S.; Dragicevic, S. & Feick, R. (2009)	No	—	No
2	**Geovisual evaluation of public participation in decision making: The grapevine**	Yes	No	No

	Aguirre, R. & Nyerges, T. (2011)			
3	**Analytical, visual and interactive concepts for geo-visual analytics** Schumann, H. & Tominski, C. (2011)	No	—	No
4	**Spatio-temporal evaluation matrices for geospatial data** Triglav, J.; Petrovič, D. & Stopar, B. (2011)	No	—	No
5	**Spatiotemporal crime analysis in U.S. law enforcement agencies: Current practices and unmet needs** Roth, R. E.; Ross, K. S.; Finch, B. G.; Luo, W. & MacEachren, A. M. (2013)	No	—	No
6	**Enabling collaborative decision-making in watershed management using cloud-computing services** Sun, A. (2013)	No	—	Yes
7	**A new Spatial OLAP approach for the analysis of Volunteered Geographic Information** Bimonte, S.; Boucelma, O.; Machabert, O. & Sellami, S. (2014)	No	—	No
8	**Geospatial information infrastructures to address spatial needs in health: Collaboration, challenges and opportunities** Granell, C.; Belmonte, O. & Díaz, L. (2014)	No	—	Yes
9	**CyberEye: Development of integrated cyberinfrastructure to support rapid hurricane risk assessment** Kijewski-Correa, T.; Smith, N.; Taflanidis, A.; Kennedy, A.; Liu, C.; Krusche, M. & Vardeman II, C. (2014)	No	—	Yes
10	**Road-based travel recommendation using geotagged images** Sun, Y.; Fan, H.; Bakillah, M. & Zipf, A. (2015)	No	—	No
11	**Environmental data visualisation for non-scientific contexts: Literature review and design framework** Grainger, S.; Mao, F. & Buytaert, W. (2016)	No	—	No
12	**Integrating geo web services for a user driven exploratory analysis** Moncrieff, S.; Turdukulov, U. & Gulland, E-K. (2016)	Yes	No	No

A.5 Scopus

- Database: Scopus

- URL: https://www.scopus.com/
- Search procedure:
  1. From Documents' search:
     - collaborative OR cooperative in All fields AND
     - "geovisual analytics" OR "geospatial visual analytics" OR geoanalytics in Article title, Abstract, Keywords
     - Year: 2004 to 2017
  2. Set the filter Language to English
  3. From the result select the papers that comply with the inclusion criteria

Table A.5 Search and selection results for Scopus

#	Paper	Duplicated	Removed	Selected
1	**Geovisual analytics for spatial decision support: Setting the research agenda** Andrienko, G.; Andrienko, N.; Jankowski, P.; Keim, D.; Kraak, M. J.; MacEachren, A. & Wrobel, S. (2007)	No	—	Yes
2	**Geovisual analytics and crisis management** Tomaszewski, B. M.; Robinson, A. C.; Weaver, C.; Stryker, M. & MacEachren, A. M. (2007)	No	—	Yes
3	**Collaborative explorative data analysis applied in HTML** Jern, M. (2008)	Yes	Yes	—
4	**Producing geo-historical context from implicit sources: A geovisual analytics approach** Tomaszewski, B. (2008)	No	—	No
5	**Collaborative web-enabled geoanalytics applied to OECD regional data** Jern, M. (2009)	Yes	Yes	—
6	**An XML-based infrastructure to enhance collaborative geographic visual analytics** Kramis, M.; Gabathuler, C.; Fabrikant, S. I. & Waldvogel, M. (2009)	Yes	Yes	—
7	**Shaping the display of the future: The effects of display size and curvature on user performance and insights** Shupp, L.; Andrews, C.; Dickey-Kurdziolek, M.; Yost, B. & North, C. (2009)	No	—	No
8	**Multi-level service infrastructure for Geovisual analytics in the context of territorial management**	No	—	No

	Conti, G.; De Amicis, R.; Piffer, S. & Simões, B. (2010)			
9	**Explore, collaborate and publish official statistics for measuring regional progress** Jern, M. (2010)	Yes	Yes	—
10	**Collaborative educational geoanalytics applied to large statistics temporal data** Jern, M. (2010)	No	—	Yes
11	**Geovisual analytics tools for communicating emergency and early warning** Jern, M.; Brezzi, M. & Lundblad, P. (2010)	Yes	Yes	—
12	**HEALTH GeoJunction: Place-time-concept browsing of health publications** MacEachren, A. M.; Stryker, M. S.; Turton, I. J. & Pezanowski, S. (2010)	No	—	No
13	**Challenges in data integration and interoperability in geovisual analytics** Turdukulov, U. D.; Blok, C. A.; Köbben, B. & Morales, J. (2010)	No	—	No
14	**Geovisual evaluation of public participation in decision making: The grapevine** Aguirre, R. & Nyerges, T. (2011)	Yes	Yes	—
15	**A web-enabled visualization toolkit for geovisual analytics – conference paper** Ho, Q. V.; Lundblad, P.; Aström, T. & Jern, M. (2011)	Yes	Yes	—
16	**Web GIS in practice IX: a demonstration of geospatial visual analytics using Microsoft Live Labs Pivot technology and WHO mortality data** Kamel Boulos, M. N.; Viangteeravat, T.; Anyanwu, M. N.; Ra Nagisetty, V. & Kuscu, E (2011)	Yes	No	Yes
17	**Cartography for everyone and everyone for cartography - why and how?** Meng, L. (2011)	No	—	No
18	**Visual storytelling - Knowledge and understanding in education** Stenliden, L. & Jern, M. (2011)	No	—	Yes
19	**A web-enabled visualization toolkit for geovisual analytics – journal paper** Ho, Q. V.; Lundblad, P.; Åström, T. & Jern, M. (2012)	Yes	No	Yes
20	**New approaches for an effective e-dissemination of statistics: The case of Noi Italia-100 statistics to understand the country we live in** De Martino, V.; Rossetti, S. & Rossi, D. (2013)	No	—	Yes
21	**Geovisual analytics and storytelling using HTML5** Lundblad, P. & Jern, M. (2013)	Yes	Yes	—

22	**Interactive maps: What we know and what we need to know** Roth, R. E. (2013)	No	—	No
23	**Flexible mixed reality and situated simulation as emerging forms of geovisualization** Lonergan, C. & Hedley, N. (2014)	No	—	No
24	**A geovisual analytics approach for mouse movement analysis** Tahir, A.; McArdle, G. & Bertolotto, M. (2014)	No	—	No
25	**Cartography** Meng, L. (2015)	Yes	Yes	—
26	**Interactivity and cartography: A contemporary perspective on user interface and user experience design from geospatial professionals** Roth, R. E. (2015)	No	—	No
27	**Interactive visual cluster detection in large geospatial datasets based on dynamic density volume visualization** Du, F.; Zhu, A. X. & Qi, F. (2016)	No	—	No
28	**Integrating geo web services for a user driven exploratory analysis** Moncrieff, S.; Turdukulov, U. & Gulland, E. K. (2016)	Yes	Yes	—
29	**Geovisual analytics and the science of interaction: An empirical interaction study (2016)** Roth, R. E. & MacEachren, A. M. (2016)	No	—	No
30	**Spatialization of user-generated content to uncover the multirelational world city network** Salvini, M. M. & Fabrikant, S. I. (2016)	No	—	No
31	**Visual synthesis of evolutionary emergency scenarios** Sebillo, M.; Tucci, M. & Vitiello, G. (2016)	No	—	No
32	**Enabling geovisual analytics of health data using a server-side approach** Turdukulov, U. & Moncrieff, S. (2016)	No	—	No

A.6 Springer link

- Database: Springer link
- URL: https://link.springer.com/
- Search procedure:
    1. From search on main page:

  - (collaborative OR cooperative) AND ("geovisual analytics" OR "geospatial visual analytics" OR geoanalytics)

2. Select date published: Between 2004 and 2017
3. From the result select the papers that comply with the inclusion criteria

Table A.6 Search and selection results for Springer Link

#	Paper	Duplicated	Removed	Selected
1	**GeoVISTA Studio: Reusability by Design** Gahegan, M.; Hardisty, F.; Demšar, U. & Takatsuka, M. (2008)	No	—	No
2	**Collaborative Explorative Data Analysis Applied in HTML** Jern, M. (2008)	Yes	Yes	—
3	**Visual Analytics: Definition, Process, and Challenges** Keim, D.; Andrienko, G.; Fekete, J-D.; Görg, C.; Kohlhammer, J. & Melançon, G. (2008)	No	—	Yes
4	**Interactive Visualization - A Survey** Brodbeck, D.; Mazza, R. & Lalanne, D. (2009)	No	—	No
5	**Skylineglobe: 3D Web Gis Solutions For Environmental Security and Crisis Management** Deiana, A. (2009)	No	—	No
6	**Soknos — An Interactive Visual Emergency Management Framework** Döweling, S.; Probst, F.; Ziegert, T. & Manske, K. (2009)	No	—	Yes
7	**Application of Virtual Worlds to Environmental Security** Gail, W. B. (2009)	No	—	No
8	**Collaborative Web-Enabled GeoAnalytics Applied to OECD Regional Data** Jern, M. (2009)	Yes	Yes	—
9	**Visual Analysis of Public Discourse on Environmental Issues** Kienreich, W. (2009)	No	—	No
10	**Visual Analytics for the Strategic Decision Making Process** Kohlhammer, J.; May, T. & Hoffmann, M. (2009)	No	—	No
11	**Remote and in Situ Sensing for Dike Monitoring: The Ijkdijk Experience**	No	—	No

	Pals, N.; De Vries, A.; De Jong, A. & Boertjes, E. (2009)			
12	**Exploring Environmental News Via Geospatial Interfaces and Virtual Globes** Scharl, A. (2009)	No	—	No
13	**Cooperative Decentralization: A New Way to Build an Added Value Chain With Shared Multi-Resolution Satellite and Aerial Imagery and Geoinformation** Villa, G. (2009)	No	—	No
14	**Explore, Collaborate and Publish Official Statistics for Measuring Regional Progress** Jern, M. (2010)	Yes	Yes	—
15	**Geovisual Analytics Tools for Communicating Emergency and Early Warning** Jern, M.; Brezzi, M. & Lundblad, P. (2010)	Yes	No	Yes
16	**Interactive Access and Processing of Multispectral Imagery: The User in the Loop** Simões, B.; Piffer, S.; Carriero, A.; Conti, G. & De Amicis, R. (2010)	No	—	No
17	**Web GIS in practice IX: a demonstration of geospatial visual analytics using Microsoft Live Labs Pivot technology and WHO mortality data** Kamel Boulos, M. N.; Viangteeravat, T.; Anyanwu, M. N.; Ra Nagisetty, V. & Kuscu, E. (2011)	Yes	Yes	—
18	**Geographic Information Science as a Common Cause for Interdisciplinary Research** Blaschke, T.; Strobl, J.; Schrott, L.; Marschallinger, R.; Neubauer, F.; Koch, A.; Beinat, E.; Heistracher, T.; Reich, S.; Leitner, M. & Donert, K. (2012)	No	—	No
19	**Visual Storytelling in Education Applied to Spatial-Temporal Multivariate Statistics Data** Lundblad, P. & Jern, M. (2012)	No	—	Yes
20	**Developing a multi-scale visualisation framework for use in climate change response (2012)** Pettit, C.; Bishop, I.; Sposito, V.; Aurambout, J-P. & Sheth, F.	No	—	No
21	**Visual Statistics Cockpits for Information Gathering in the Policy-Making Process** Burkhardt, D.; Nazemi, K.; Stab, C.; Steiger, M.; Kuijper, A. & Kohlhammer, J. (2013)	No	—	No
22	**Big Board: Teleconferencing Over Maps for Shared Situational Awareness** Heard, J.; Thakur, S.; Losego, J. & Galluppi, K. (2013)	No	—	Yes

23	**Visual Analytics in Environmental Research: A Survey on Challenges, Methods and Available Tools** Komenda, M. & Schwarz, D. (2013)	No	—	Yes
24	**Applying Geovisual Analytics to Volunteered Crime Data** Moore, A.; de Oliveira, M.; Caminha, C.; Furtado, V.; Basso, V. & Ayres, L. (2013)	No	—	No
25	**Visual Analytics and Information Retrieval** Santucci, G. (2013)	No	—	No
26	**Tendencies in Contemporary Cartography** Azócar Fernández, P. I. & Buchroithner, M. F. (2014)	No	—	No
27	**A Visual Analytics System for Supporting Rock Art Knowledge Discovery** Deufemia, V.; Indelli Pisano, V.; Paolino, L. & de Roberto, P. (2014)	No	—	No
28	**Visualisation: An Approach to Knowledge Building** Masala, E. & Pensa, S. (2014)	No	—	No
29	**An open source toolkit for identifying comparative space-time research questions** Ye, X.; She, B.; Wu, L.; Zhu, X. & Cheng, Y. (2014)	No	—	No
30	**Analysis and visualisation of movement: an interdisciplinary review** Demšar, U.; Buchin, K.; Cagnacci, F.; Safi, K.; Speckmann, B.; Van de Weghe, N.; Weiskopf, D. & Weibel, R. (2015)	No	—	No
31	**Geographic Visualization and MCDA** Malczewski, J. & Rinner, C. (2015)	No	—	No
32	**The Design of a Collaborative Social Network for Watershed Science** McGuire, M. P. & Roberge, M. C. (2015)	No	—	No
33	**Cartography** Meng, L. (2015)	Yes	No	No
34	**On the Support of Automated Analysis Chains on Enterprise Models** Ramos, A.; Sáenz, J. P.; Sánchez, M. & Villalobos, J. (2015)	No	—	No
35	**TweeProfiles3: Visualization of Spatio-Temporal Patterns on Twitter** Maia, A.; Cunha, T.; Soares, C. & Abreu, P. H. (2016)	No	—	No
36	**City Probe: The Crowdsourcing Platform Driven by Citizen-Based Sensing for Spatial Identification and Assessment** Shen, Y. T.; Shiu, Y. S. & Lu, P. (2016)	Yes	Yes	—

37	**Volunteered Geographic Information for Building Territorial Governance in Mexico City: The Case of The Roma Neighborhood** Martínez-Viveros, E.; Tapia-McClung, R.; Calvillo-Saldaña, Y. & López-Gonzaga, J. L. (2017)	No	—	No
38	**The Participatory Sensing Platform Driven by UGC for the Evaluation of Living Quality in the City** Shen, Y. T.; Shiu, Y. S.; Liu, W. K. & Lu, P. W. (2017)	No	—	No
39	**A survey of network anomaly visualization** Zhang, T.; Wang, X.; Li, Z.; Guo, F.; Ma, Y. & Chen, W. (2017)	No	—	No

A.7 Web of Science

- Database: Web of Science
- URL: http://login.webofknowledge.com
- Search procedure:
  1. From expert search:
     - (collaborative OR cooperative) AND ("geovisual analytics" OR "geospatial visual analytics" OR geoanalytics)
     - Year: 2004 to 2017
  2. From the result select the papers that comply with the inclusion criteria

Table A.7 Search and selection results for Web of Science

#	Paper	Duplicated	Removed	Selected
1	**Collaborative Explorative Data Analysis Applied in HTML** Jern, M. (2008)	Yes	No	Yes
2	**Collaborative Web-Enabled GeoAnalytics Applied to OECD Regional Data** Jern, M. (2009)	Yes	No	Yes
3	**An XML-based Infrastructure to Enhance Collaborative Geographic Visual Analytics (2009)** Kramis, M.; Gabathuler, C.; Fabrikant, S. I. & Waldvogel, M. (2009)	Yes	Yes	—

4	**Explore, Collaborate and Publish Official Statistics for Measuring Regional Progress** Jern, M. (2010)	Yes	No	Yes
5	**Collaborative educational geoanalytics applied to large statistics temporal data** Jern, M. (2010)	Yes	Yes	—
6	**A web-enabled visualization toolkit for geovisual analytics – Journal paper** Ho, Q. V.; Lundblad, P.; Astrom, T. & Jern, M. (2012)	Yes	Yes	No
7	**Information Visualization for Product Lifecycle Management (PLM) Data** Guo, C.; Chen, Y. V.; Miller, C. L.; Hartman, N. W.; Mueller, A. B. & Connolly, P. E. (2014)	No	—	No
8	**City Probe: The Crowdsourcing Platform Driven by Citizen-Based Sensing for Spatial Identification and Assessment** Shen, Y. T.; Shiu, Y. S. & Lu, P. (2016)	Yes	No	Yes
9	**Challenges and strategies for the visual exploration of complex environmental data** Helbig, C.; Dransch, D.; Boettinger, M.; Devey, C.; Haas, A.; Hlawitschka, M.; Kuenzer, C.; Rink, K.; Schafer-Neth, C.; Scheuermann, G.; Kwasnitschka, T. & Unger, A. (2017)	No	—	No

A.8 Papers included in the systematic review

Table A.8 List of selected papers to identify CGVA Systems and Collaboration Techniques

Source	#	Paper
ACM Digital Library	1	**Thin Client Visualization** Eick, S. G.; Eick, M. A.; Fugitt, J.; Horst, B.; Khailo, M. & Lankenau, R. A. (2007)
	2	**nSpace and GeoTime: A VAST 2006 Case Study** Proulx, P.; Chien, L.; Harper, R.; Schroh, R.; Kapler, T.; Jonker, D. & Wright, W. (2007)
	3	**Space, time and visual analytics** Andrienko, G.; Andrienko, N.; Demsar, U.; Dransch, D.; Dykes, J.; Fabrikant, S. I.; Jern, M.; Kraak, M-J.; Schumann, H. & Tominski, C. (2010)
	4	**nu-view: a visualization system for collaborative co-located analysis of geospatial disease data** Masoodian, M.; Luz, S. & Kavenga, D. (2016)

GeoBase	5	**Leveraging the potential of geospatial annotations for collaboration: A communication theory perspective** Hopfer, S. & MacEachren, A.M. (2007)
	6	**An XML-based infrastructure to enhance collaborative geographic visual analytics** Kramis, M.; Gabathuler, C.; Fabrikant, S. I. & Waldvogel, M. (2009)
IEEE Xplore	7	**Visual Analytics Presentation Tools Applied in HTML Documents** Jern, M.; Rogstadius, J.; Åström, T. & Ynnerman, A. (2008)
	8	**Geovisual Analytics and Storytelling Using HTML5** Lundblad, P. & Jern, M. (2013)
Science Direct	9	**Enabling collaborative decision-making in watershed management using cloud-computing services** Sun, A. (2013)
	10	**Geospatial information infrastructures to address spatial needs in health: Collaboration, challenges and opportunities** Granell, C.; Belmonte, O. & Díaz, L. (2014)
	11	**CyberEye: Development of integrated cyber-infrastructure to support rapid hurricane risk assessment** Kijewski-Correa, T.; Smith, N.; Taflanidis, A.; Kennedy, A.; Liu, C.; Krusche, M. & Vardeman II, C. (2014)
Scopus	12	**Geovisual analytics for spatial decision support: Setting the research agenda** Andrienko, G.; Andrienko, N.; Jankowski, P.; Keim, D.; Kraak, M. J.; MacEachren, A. & Wrobel, S. (2007)
	13	**Geovisual analytics and crisis management** Tomaszewski, B. M.; Robinson, A. C.; Weaver, C.; Stryker, M. & MacEachren, A. M. (2007)
	14	**Collaborative educational geoanalytics applied to large statistics temporal data** Jern, M. (2010)
	15	**Web GIS in practice IX: a demonstration of geospatial visual analytics using Microsoft Live Labs Pivot technology and WHO mortality data** Kamel Boulos, M. N.; Viangteeravat, T.; Anyanwu, M. N.; Ra Nagisetty, V. & Kuscu, E. (2011)
	16	**Visual storytelling - Knowledge and understanding in education** Stenliden, L. & Jern, M. (2011)
	17	**A web-enabled visualization toolkit for geovisual analytics** Ho, Q. V.; Lundblad, P.; Åström, T. & Jern, M. (2012)

	18	**New approaches for an effective e-dissemination of statistics: The case of Noi Italia-100 statistics to understand the country we live in** De Martino, V.; Rossetti, S. & Rossi, D. (2013)
Springer	19	**Visual Analytics: Definition, Process, and Challenges** Keim, D.; Andrienko, G.; Fekete, J-D.; Görg, C.; Kohlhammer, J. & Melançon, G. (2008)
	20	**Soknos — An Interactive Visual Emergency Management Framework** Döweling, S.; Probst, F.; Ziegert, T. & Manske, K. (2009)
	21	**Geovisual Analytics Tools for Communicating Emergency and Early Warning** Jern, M.; Brezzi, M. & Lundblad, P. (2010)
	22	**Visual Storytelling in Education Applied to Spatial-Temporal Multivariate Statistics Data** Lundblad, P. & Jern, M. (2012)
	23	**Big Board: Teleconferencing Over Maps for Shared Situational Awareness** Heard, J.; Thakur, S.; Losego, J. & Galluppi, K. (2013)
	24	**Visual Analytics in Environmental Research: A Survey on Challenges, Methods and Available Tools** Komenda, M. & Schwarz, D. (2013)
Web of Science	25	**Collaborative Explorative Data Analysis Applied in HTML** Jern, M. (2008)
	26	**Collaborative Web-Enabled GeoAnalytics Applied to OECD Regional Data** Jern, M. (2009)
	27	**Explore, Collaborate and Publish Official Statistics for Measuring Regional Progress** Jern, M. (2010)
	28	**City Probe: The Crowdsourcing Platform Driven by Citizen-Based Sensing for Spatial Identification and Assessment** Shen, Y. T.; Shiu, Y. S. & Lu, P. (2016)

B User evaluation form

In this annex, the reader can find the prototype evaluation form. The form is in Spanish because it is the mother tongue of all the participants.

Usuario: _____________ Contraseña: _____________

Ubicación: ______________________ Fecha: ___________ Hora inicio: _____ Hora fin: _____

Sistema operativo: _____________________ Navegador web: __________________________

Dispositivo: _____________________ Tamaño pantalla: _________ Resolución: ____________

Sección 1: Esta parte tiene como objetivo describir el uso del sistema, y evaluar su opinión con relación a la facilidad de utilizarlo.

Herramientas de exploración

1. Es fácil mostrar y ocultar los paneles para opciones generales, línea de tiempo y el panel de herramientas colaborativas

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

2. Es fácil cambiar el tamaño de los paneles que contienen los gráficos

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

3. Es fácil cambiar la disposición de los gráficos

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

4. Es fácil seleccionar el año y semana del año para visualizar

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

5. Es fácil seleccionar objetos en la vista de mapa

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

6. Es fácil cambiar el color para resaltar un objeto seleccionado

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

7. Es fácil cambiar la capa activa

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

8. Es fácil evaluar la evolución de una variable a lo largo del año en un objecto seleccionado

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

9. Es fácil comparar una variable para una semana específica a través de los años en un objecto seleccionado

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

10. Es fácil comparar entre objetos seleccionados

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

11. Es fácil cambiar el esquema de color compartido para una capa (atributo y rampa de color)

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

12. Es fácil activar y desactivar el esquema de color compartido en los gráficos

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

Spatial-Temporal Analysis Spaces

13. Es fácil activar y desactivar la herramienta Spatial-Temporal Analysis Spaces

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

14. Es fácil entender la extensión temporal y espacial de los espacios de análisis

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

15. Es fácil abrir y cerrar un espacio de análisis

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

16. La transparencia aplicada en los gráficos cuando se abre un espacio de análisis ayuda a enfocar la atención del usuario en dichos datos

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

17. Es fácil moverse entre espacios de análisis relacionados					
	Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo
18. Es fácil crear un nuevo espacio de análisis					
	Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo
19. Es fácil crear una nueva pregunta					
	Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo
20. Es fácil cambiar el estado de una pregunta a “Cerrada” o “Descartada”					
	Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo
21. Es fácil borrar una pregunta					
	Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo
22. Es fácil abrir y cerrar una pregunta					
	Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo
23. Es fácil responder una pregunta					
	Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

Observaciones:

____

____

____

____

____

____

Sección 2: Esta parte tiene como objetivo evaluar la facilidad para realizar una tarea, una vez que se conoce el sistema.

Tarea 1 – cree un espacio de análisis:

Hora inicio: _______

- Ir al año _______
- Busque un fenómeno de interés (utilice las herramientas de exploración para este fin, ejemplo, cambie los gráficos, variables, esquema compartido de color y seleccione objetos)
- Cree un espacio de análisis alrededor del fenómeno de interés

Hora fin: _______

24. Fue fácil completar la tarea

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

Tarea 2 – cree una pregunta:

Hora inicio: _______

- Ir al año _______
- Abrir un espacio de análisis
- Crear una nueva pregunta (con un mínimo de 3 opciones de respuesta)

Hora fin: _______

25. Fue fácil completar la tarea

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

Tarea 3 – contestar una respuesta

Hora inicio: _______

- Ir al año _______
- Abrir un espacio de análisis
- Abrir una pregunta
- ¿cuántas respuestas tiene la pregunta? ______
- Responder la pregunta

Hora fin: _______

26. Fue fácil completar la tarea

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

Tarea 4 – cambiar el estado de una pregunta

Hora inicio: _______

- Abra la pregunta que creó en la **Tarea 2**
- ¿Cuántas respuestas recibió? _____
- Cambie el estado de la pregunta a "Cerrado" o "Descartado"

Hora fin: _______

27. Fue fácil completar la tarea

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

Observaciones:

____

____

____

____

____

____

Sección 3: Esta parte tiene como objetivo evaluar su experiencia en general con el sistema.

28. Pienso que me gustaría utilizar este sistema con frecuencia

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

29. Encuentro el sistema innecesariamente complejo

□	□	□	□	□
Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

30. Pienso que el sistema es fácil de utilizar	□ Totalmente en desacuerdo	□ Desacuerdo	□ Neutral	□ De acuerdo	□ Totalmente de acuerdo
31. Pienso que necesitaría la ayuda de un técnico para poder utilizar este sistema	□ Totalmente en desacuerdo	□ Desacuerdo	□ Neutral	□ De acuerdo	□ Totalmente de acuerdo
32. Encuentro las funciones del sistema bien integradas	□ Totalmente en desacuerdo	□ Desacuerdo	□ Neutral	□ De acuerdo	□ Totalmente de acuerdo
33. Pienso que existe mucha inconsistencia en este sistema	□ Totalmente en desacuerdo	□ Desacuerdo	□ Neutral	□ De acuerdo	□ Totalmente de acuerdo
34. Me parece que la mayoría de las personas aprenderían a utilizar este sistema rápidamente	□ Totalmente en desacuerdo	□ Desacuerdo	□ Neutral	□ De acuerdo	□ Totalmente de acuerdo
35. Encuentro el sistema muy complicado de utilizar	□ Totalmente en desacuerdo	□ Desacuerdo	□ Neutral	□ De acuerdo	□ Totalmente de acuerdo
36. Me sentí cómodo utilizando el sistema	□ Totalmente en desacuerdo	□ Desacuerdo	□ Neutral	□ De acuerdo	□ Totalmente de acuerdo
37. Necesité aprender muchas cosas antes de poder utilizar el sistema	□ Totalmente en desacuerdo	□ Desacuerdo	□ Neutral	□ De acuerdo	□ Totalmente de acuerdo

Observaciones:

____

____

____

____

____

____

Sección 4: Esta parte tiene como objetivo evaluar la utilidad del sistema en el seguimiento y control de plagas.

38. El sistema puede ayudar a explorar y entender las dinámicas poblacionales de una plaga

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

39. El sistema puede ayudar en la toma de decisiones sobre las acciones de control para una plaga

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

40. El sistema puede ayudar a evaluar los efectos de las acciones de control

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

41. El sistema puede ayudar a entender los cambios de largo plazo en las dinámicas poblacionales de una plaga

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

42. El Sistema puede ayudar a mejorar la comunicación entre interesados en el manejo de la plaga

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

43. El Sistema puede ayudar a construir una base de conocimientos con conocimiento especifico de la plaga y del área de interés

Totalmente en desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo

Observaciones:

____

____

____

____

____

C Results of the user evaluation

In this annex, the reader can find the individual responses for the questionnaire. While the test was performed in Spanish, for reader's convenience, the statements in this annex were translated.

C.1 Section 1

Table C.1 Individual responses for Section 1 of the prototype evaluation.

	Votes casted by							Total votes for				
Statement	**Test user 1**	**Test user 2**	**Test user 3**	**Test user 4**	**Test user 5**	**Test user 6**	**Test user 7**	**1 - Strongly disagree**	**2 - Disagree**	**3 - Neutral**	**4 - Agree**	**5 - Strongly agree**
1. It is easy to show and hide the panels for general options, timeline and Spatio-Temporal Analysis Spaces	3	4	5	5	4	4	5	0	0	1	3	3
2. It is easy to resize the panels	5	5	5	5	4	5	5	0	0	0	1	6
3. It is easy to change the charts' layout	5	5	5	5	4	4	5	0	0	0	2	5
4. It is easy to change the active layer	5	5	5	5	5	5	5	0	0	0	0	7
5. It is easy to change the shared color schema for a layer (i.e., attribute and color ramp)	5	5	5	5	5	5	5	0	0	0	0	7
6. It is easy to change the year to be analyzed	4	4	5	5	4	5	5	0	0	0	3	4

7. It is easy to select objects in the map view	5	4	5	5	4	5	5	0	0	0	2	5
8. It is easy to change the color to highlight a selected object	5	4	5	5	5	5	5	0	0	0	1	6
9. It is easy to switch on and off the shared color schema in the charts	5	4	5	5	5	5	5	0	0	0	1	6
10. It is easy to assess the evolution of a variable on a selected object across the year	4	5	5	5	5	5	4	0	0	0	2	5
11. It is easy to compare a variable on a selected object for a specific week across years	4	4	4	5	3	5	5	0	0	1	3	3
12. It is easy to compare between selected objects	4	5	4	5	3	5	5	0	0	1	2	4
13. It is easy to switch on and off the Spatio-Temporal Analysis Spaces tool	5	4	4	5	5	5	5	0	0	0	2	5
14. It is easy to understand the temporal and spatial extensions of the analysis spacess	4	4	4	5	4	5	4	0	0	0	5	2
15. It is easy to create a new analysis space	5	4	4	5	4	5	5	0	0	0	3	4
16. It is easy to open and close an analysis space	5	4	5	5	4	5	5	0	0	0	2	5
17. The transparency applied to the charts when an analysis space is open helps to focus analyst's attention on the data contained on it	4	4	4	5	4	3	5	0	0	1	4	2
18. It is easy to move between related analysis spaces	5	4	4	5	4	4	5	0	0	0	4	3
19. It is easy to create a new question	5	4	4	5	4	5	5	0	0	0	3	4
20. It is easy to change the status of a question to "Close" or "Discarded"	5	4	4	5	4	5	5	0	0	0	3	4
21. It is easy to delete a question	5	4	4	5	4	5	5	0	0	0	3	4

22. It is easy to open and close a question	5	4	4	5	4	5	5	0	0	0	3	4
23. It is easy to answer a question	4	4	4	5	4	5	5	0	0	0	4	3
Total votes								0	0	4	56	101

C.2 Section 2

Table C.2 Individual responses for Section 2 of the prototype evaluation.

	Votes casted by							Total votes for				
Statement	**Test user 1**	**Test user 2**	**Test user 3**	**Test user 4**	**Test user 5**	**Test user 6**	**Test user 7**	**1 - Strongly disagree**	**2 - Disagree**	**3 - Neutral**	**4 - Agree**	**5 - Strongly agree**
24. It was easy to complete the task (Task 1: create an analysis space)	4	4	4	3	3	5	4	0	0	2	4	1
25. It was easy to complete the task (Task 2: create a question)	5	4	4	4	4	5	5	0	0	0	4	3
26. It was easy to complete the task (Task 3: answer a question)	3	*	4	4	5	4	5	0	0	1	3	2
27. It was easy to complete the task (Task 4: change a question's status)	4	*	4	4	4	5	5	0	0	0	4	2
Total votes								0	0	3	15	8
* The test user didn't completed this task because of work-related duties.												

C.3 Section 3 - SUS

Table C.3 Individual responses for Section 3 of the prototype evaluation.

	Votes casted by							Votes for				
Statement	**Test user 1**	**Test user 2**	**Test user 3**	**Test user 4**	**Test user 5**	**Test user 6**	**Test user 7**	**1 - Strongly disagree**	**2 - Disagree**	**3 - Neutral**	**4 - Agree**	**5 - Strongly agree**
28. I think that I would like to use this system frequently	4	4	4	4	4	5	5	0	0	0	5	2
29. I found the system unnecessarily complex	2	2	2	2	2	4	2	0	6	0	1	0
30. I thought the system was easy to use	4	4	4	4	3	4	5	0	0	1	5	1
31. I think that I would need the support of a technical person to be able to use this system	2	3	2	2	3	3	2	0	4	3	0	0
32. I found the various functions in this system were well integrated	4	4	4	5	4	4	5	0	0	0	5	2
33. I thought there was too much inconsistency in this system	1	2	2	2	2	1	1	3	4	0	0	0
34. I would imagine that most people would learn to use this system very quickly	4	4	4	4	3	4	4	0	0	0	6	1
35. I found the system very cumbersome to use	2	2	2	1	2	2	2	1	6	0	0	0
36. I felt very confident using the system	5	4	4	5	4	4	5	0	0	0	4	3
37. I needed to learn a lot of things before I could get going with this system	2	3	2	1	2	2	1	2	4	1	0	0
Total	80	70	75	85	67.5	72.5	90	6	24	6	26	8
	Good	Ok	Good	Excellent	Ok	Good	Excellent					

The general average of SUS scores for the prototype is 77.14, which means that based on test users' opinion the prototype's design is 'Good'[9].

C.4 Section 4

Table C.4 Individual responses for Section 4 of the prototype evaluation.

	Votes casted by							Total votes for				
Statement	**Test user 1**	**Test user 2**	**Test user 3**	**Test user 4**	**Test user 5**	**Test user 6**	**Test user 7**	**1 - Strongly disagree**	**2 - Disagree**	**3 - Neutral**	**4 - Agree**	**5 - Strongly agree**
38. The system helps to explore and understand the population dynamics of a pest	5	5	4	3	4	5	5	0	0	1	2	4
39. The system may support decision-making regarding the application of control actions	4	4	4	4	4	5	5	0	0	0	5	2
40. The system may help to assess the effects of the control actions	5	5	4	4	4	5	5	0	0	0	3	4
41. The system may help to understand the long-term changes in the population dynamics of a pest	3	5	4	3	4	5	5	0	0	2	2	3
42. The system may help to improve communication among stakeholders of the pest management	5	4	4	5	3	4	5	0	0	1	3	3

43. The system may help to build a knowledge base with domain and local knowledge regarding a pest	5	4	4	4	4	5	4	0	0	0	5	2
Total votes								0	0	4	20	18

Summary

Several advancements in information and communication technologies and geospatial technologies have led to an unprecedented abundance of geodata. This data abundance presents novel opportunities to improve our understanding of natural and artificial processes. However, it challenges analysts who need to make sense of such large, heterogeneous, and multivariate data sets to address complex analysis situations. Two main challenges are (i) the limited capacity of humans to work with large amounts of data, and (ii) the multi-disciplinarity and complexity of data sets that renders analysis by a single person infeasible.

Geovisual Analytics (GVA) enables a synergy of human analytical skills with computer storage and processing power, addressing the first challenge, and facilitates collaboration of multiple analysts through interactive user interfaces, addressing the second challenge. To date, research in this field has focused on developing data transformation algorithms, visualization techniques, and interaction methods. However, the support for collaborative analysis has received less attention. This research project addresses the challenge of supporting collaborative analysis in GVA systems and produced four main outcomes, which are described below.

First, a literature review investigated the state-of-the-art of collaborative analysis in GVA. Chapter 2 reports the results, which include thirteen GVA systems with functions to support collaborative analysis, six distinct collaboration techniques, and three research challenges. The review revealed that the most common collaboration scenario is asynchronous and distributed, because it removes the constraints of time and place for participation. Further, increasing support for multiple types of devices is enabled by cloud-based infrastructures, which improve the scalability of and accessibility to the system. Although the identified collaboration techniques support the whole data analysis process, functions are missing to synthesize analytical contributions and summarize the level of agreement regarding evidence and conclusions. The identified research challenges are the lack of support for (i) collaboration scenarios supporting collocated and distributed participants or synchronous and asynchronous interactions, (ii) collaboration across multiple types of devices, and (iii) time-critical and long-term analysis scenarios.

The second main outcome is a software reference architecture for collaborative geovisual analytics (CGVA) systems, proposed in Chapter 3. A

software architecture is an abstract high-level description of the fundamental structures of a system, their relationships, and properties of both. Software architectures can have different goals and scopes. Therefore, it is not designed for one specific system, but it is a template design for a group of similar systems. The architecture design criteria are based on the literature review. The proposed architecture uses the client-server and layered architectural patterns. An architectural pattern is a reusable solution with well-understood properties for a commonly occurring problem in software architecture design. In the proposed architecture, the client-server pattern allows assigning the processing and storage to the server side, enabling participants to work from low-end devices. Additionally, the layered pattern enables separation of concerns, meaning that software designers can address design problems - such as user interface, data processing, security, data storage, and the communication and coordination between the components that perform each function - in isolation.

The third outcome is the approach of Spatiotemporal Analysis Space (STAS), which proposes a philosophy for long-term distributed asynchronous collaborative analysis of spatiotemporal data, described in Chapter 4. STAS was specifically designed to support long-term analysis processes, which the application case of this research - agricultural pest management - requires. The main assumption in the STAS design is that in data sets with large spatial and temporal extents, events of interest such as patterns and outliers occur in diverse locations and times. The central concept of STAS is the analysis space, which is a container for a data subset that contains an event of interest and analytical contributions to make sense of it. The analysis space focuses the analyst's attention on an event of interest to elicit sensible contributions and generate meaningful knowledge. Additionally, the approach allows creating links between different analysis spaces to promote knowledge-building from previous contributions.

Lastly, a case study demonstrates the applicability and feasibility of the software reference architecture and the collaborative approach in a real-world scenario. Described in Chapter 5, the case study is the monitoring and control process of the Olive Fruit Fly (Bactrocera oleae) in olive groves in Andalusia, Spain. In this context, a CGVA prototype implementing the STAS approach was designed, developed, and tested with case study stakeholders. The stakeholders are one authority representative, one researcher, and five field technicians. They used the prototype to analyze monitoring and control data and the outputs of two statistical models. The results show that the architecture can accommodate the case-specific requirements and that the STAS approach enables collaborative analysis among the stakeholders. In the post-evaluation discussions, all the stakeholders mentioned that they would like to see the prototype becoming a production system to support their pest management activities.

To conclude, this thesis presents research that identifies remaining research challenges for CGVA, followed by the development of required software architecture and collaborative analytical approach (STAS), successfully implemented in a prototype in a real-world case study with stakeholder involvement.

Samenvatting

Vooruitgang in zowel informatie- en communicatietechnologie als in ruimtelijke datatechnologie heeft geleid tot een niet eerder vertoonde rijkdom aan ruimtelijke data. Deze rijkdom biedt kansen voor nieuwe gegevensprodukten maar leidt ook tot uitdagingen in de analyse ervan omdat de data typisch volumineus, heterogeen en multivariaat is. Twee belangrijke uitdagingen zijn (1) het beperkte vermogen van mensen om met volumineuze data te werken, en (2) het multidisciplinaire en complexe karakter van de data dat verwerking en analyse door slechts een persoon praktisch onmogelijk maakt.

GeoVisual Analytics (GVA) stelt ons in staat synergie te bereiken tussen analytische vaardigheden van de mens en opslag- en verwerkingscapaciteit van de machine, en dit adresseert de eerste uitdaging. GVA faciliteert de samenwerking tussen analisten door interactieve user interfaces, hetgeen de tweede uitdaging adresseert. Onderzoek in dit veld heeft zich vooral gericht op algoritmiek van datatransformatie, visualisatietechniek en interactieve werkmethoden. De ondersteuning van samenwerkende analisten heeft niet zoveel aandacht gekregen. Dit promotieonderzoek heeft zich daarop gericht, en heeft tot vier centrale resultaten geleid.

Ten eerste werd met een literatuurstudie the state-of-the-art van GVA in kaart gebracht. In Hoofdstuk 2 wordt een studie gepresenteerd, onder meer omvattend: dertien GVA-systemen die samenwerkingstechnieken aanbieden, zes samenwerkingstechnieken, en drie uitdagingen voor nader onderzoek. Deze studie toont aan dat de meest voorkomende techniek die van asynchroon-gedistribueerd is, omdat het the plaats/tijd-beperkingen wegneemt. Ook blijkt een toenemende ondersteuning van apparatuur, inclusief goedkope apparatuur, vanwege cloud-infrastruktuur, die schaalbaarheid en toegang tot het systeem vergroot. Hoewel de beschikbare samenwerkingstechnieken het hele data-analysepad dekken, zijn nog niet beschikbaar functies voor de synthese van analyseresultaten en voor het aggregeren van gelijkgestemdheid tussen analisten over vormen van bewijs en conclusies. Diverse onderzoeksuitdagingen worden benoemd: onvoldoende steun voor (1) hybride samenwerkingsscenario's, (2) samenwerking tussen apparaten van verschillende makelij, en (3) scenario's die een tijdkritisch of lange-termijn karakter hebben.

Het tweede resultaat is een referentiearchitectuur voor GVA systemen

die samenwerking tussen analisten ondersteunen. Deze is beschreven in Hoofdstuk 3. DE ontwerpcriteria voor deze architectuur zijn afgeleid uit de literatuur; in de basis bestaat deze uit een client-server filosofie en een gelaagde opbouw van systeemfuncties, die een systematisch benadering van ontwerp, ontwikkeling, realisatie en onderhoud van zo'n systeem mogelijk maken. Hoewel sommige scenario's van systeemgebruik voordeel zouden hebben van de beschikbaarheid van microservices om redenen van schaalbaarheid en flexibiliteit, zijn deze niet expliciet opgenomen in de referentiearchitectuur, en wel om een drietal redenen. De eerste reden is dat vooral tijdkritische analyses er voordeel van zouden hebben. Ten tweede zou de ermee gepaard gaande toegenomen systeemcomplexiteit onwenselijk zijn voor veel systemen. Ook geldt dat microservices een recent ontwerppatroon is waarvan de operationele karakteristieken nog niet helemaal duidelijk is.

Het derde resultaat behelst de Spatiotemporal Analysis Space (STAS) aanpak van lange-termijn, gedistribueerde, asynchrone samenwerking in ruimte-tijd analyse, zoals beschreven in Hoofdstuk 4. STAS is ontworpen om lange-termijnanalyses te ondersteunen, zoals wenselijk in de casus van dit promotieonderzoek — de bestrijding van plagen in de landbouw. Het ontwerp van STAS is gericht op de herkenning van patronen en afwijkingen in ruimte en tijd zoals gerepresenteerd in data met een grote voetafdruk in ruimte en tijd. Het belangrijke concept van STAS is de 'analysis space', een virtuele ruimte waarin data rondom een thema van belang en analytische bijdragen bij elkaar gebracht worden. Deze ruimte is zo ingericht dat de analist zich kan richten op het thema van belang en kan bijdragen in de discussie en kennisvorming. Daarbij komt de mogelijkheid om diverse ruimtes met elkaar te verbinden opdat kennisvorming ook profijt heeft van discussies die eerder hebben plaats gevonden.

Tenslotte wordt in Hoofdstuk 5 een casus beschreven die laat zien dat de referentiearchitectuur mogelijk en toepasbaar is in een echt scenario waarin analisten samenwerken. Deze casus behelst het monitoren en beheersen van het voorkomen van de Olijfvlieg (Bactrocera oleae) in Andalusië (Z Spanje), waarvoor een prototypesysteem werd ontwikkeld waarin de STAS-benadering wordt gebruikt. In deze casus werkten experts samen in het gebruik van systeem ter monitoring van de vlieg in olijfboomgaarden, en ontwikkelden zo twee statistische modellen. De casus laat zien dat casus-specifieke eisen succesvol kunnen worden geadresseerd en dat de STAS-benadering gebruikers in staat stelt samen analyses te ontwikkelen. In discussies ter afronding van de casus werd duidelijk dat de gebruikers pleitten voor de ingebruikname van het systeem ter ondersteuning van hun werk in plaagbestrijding.

In samenvatting beschrijft dit proefschrift een onderzoek dat braakliggend terrein in onderzoek naar GVA systemen die samenwerking ondersteunen identificeert. Het werk leidde tot de ontwikkeling van een software-architectuur en een benadering van samenwerking rond ruimte-tijd data, die is geïmplementeerd in een prototype dat is getoetst in een serieuze casus waarin gebruikers met expertise hun bijdragen leverden.

Biography

Gustavo García was born in Quetzaltenango, Guatemala, on March 31st, 1985. He graduated as Software Engineer in 2008 at the Universidad Mesoamericana Quetzaltenango. The same year, he started working as a lecturer in Software Engineering and Telecommunications Engineering at the same university. A year later, he was appointed by the national university of Guatemala, Universidad de San Carlos de Guatemala (USAC), as a lecturer and researcher in geoinformatics for Land Administration Engineering in the Centro Universitario de Occidente (CUNOC). Currently, he is still working for the national university of Guatemala. He also worked as a software developer for six years, including working for national and international enterprises.

In 2012, he graduated from the Master of Science in Human Resources Management at the CUNOC, USAC. The same year, he was granted a scholarship funded by NUFFIC to study at the Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, The Netherlands. In 2014, he graduated with honors (cum laude) from the Master of Science in Geo-information Science and Earth Observation, specializing in Geoinformatics. For his MSc thesis, he developed a specification language and a processing engine for producing animated maps. The same year, he was recruited as a PhD candidate by the Faculty ITC.