Question-Answer patterns in GIS

Semantic analysis of geo-analytical questions in Human Geography

Master thesis Human Geography

Presented to the School of Geosciences

Written by

Romay Evers

August 2020

Utrecht University

Student Number: 6922937

Email address: r.evers3@students.uu.nl

Supervisor: Dr. Simon Scheider

Second assessor: Pieter Hooimeijer

Abstract

This research aims to find common patterns in the answering of geo-analytical questions by creating scenarios and conducting an exploratory data analysis on the corresponding workflows and the GIS tools used in those workflows. The analyses were based on 27 geo-analytical questions with corresponding workflows, all in line with the Human Geography scenario 'livability and health of residents in Amsterdam'. Several exploratory analyses were performed in order to find patterns in geo-analytical questions and workflows in parallel. By building upon the findings in this thesis, an universal basis for asking spatial questions in a systematic manner and automatically retrieving the resources necessary to answer them will be one step closer. This thesis ends with theoretical and practical implications for the next steps in this line of expertise.

Key words: geo-analytical Question-Answering, QuAnGIS, livability, workflow, GIS tool, Natural Language Processing.

Table of contents

Acknowledgments

Throughout the process of writing my thesis, I have received a great deal of support, feedback and supervision, and therefore, I would like to thank the people who were a part of this thesis.

First, I would like to thank my supervisor Simon Scheider, who introduced me to the topic and whose level of GIS expertise helped me in formulating the research questions and exploring this field of Human Geography. I am honored to have been a part of his QuAnGIS project.

I would also like to acknowledge two other researchers in the QuAnGIS project; Enkhbold Nyamsuren and Haiqi Xu. They guided me, together with Simon Scheider, through the process and provided me with useful feedback, by imparting their knowledge and technical know-how.

I would also like to thank Pieter Hooimeijer in advance, my second assessor, for reading and grading my thesis with a different point of view.

In addition, a special thanks goes to Geke Meessen, who supported me and offered her professional opinion about important decisions in the final phase of my studies.

A special word of gratitude goes to my parents and sister for the encouragement and safe place during this time, which helped me in completion of this thesis.

Finally, I am grateful for my friends, who were of great support throughout the process. They helped me in deliberating over my problems and findings. Not only did they help me with my thesis, they were a happy distraction too.

1 Introduction

In today's age, which is characterized by big data, computer and information science, data scientists of various disciplines have made geographic information a key objective in order to embed their analysis in a spatiotemporal context (Yin, Zhang, Goldberg, & Prasad, 2019). Spatiotemporal context is used in data analysis when data is collected across both space and time. However, while the demand and variety of data sources and software increases, it becomes more difficult to comprehend and make use of all the tools and data available to answer geo-analytical questions. A common obstacle for geospatial analysts is when a specific tool functionality is needed, but not available in another tool. This obstacle forces them to reformulate their questions to make them compatible with another tool or other datasets (Yin et al., 2019). Up to this day, it is not straightforward which data types would be needed to ease certain geo-analytical tasks (Scheider, Meerlo, Kasalica, & Lamprecht, n.d.). Therefore, understanding the syntactic and semantic structure of geographic question-answering systems is a fundamental step towards geo-analytical question-answering machines (Xu et al., 2020) Furthermore, geo-analytical question-answering systems have become a popular topic in the science of Geographic Information System (GIS) (Xu et al., 2020). GIS is defined as "a powerful set of tools for collecting, storing, retrieving at will, transforming, and displaying spatial data from the real world for a particular set of purposes" (Burrough, McDonnell & Lloyd, 2015, pp. 3). Unfortunately, geo-analytical technology and tools are not able to answer geo-analytical questions. That is because the user interface of existing GIS software is not able to support this directly (yet) (Gao & Goodchild, 2013). The increasing amount of available data makes the search of data more difficult and selecting the fitting tool to analyze the data can be a demanding process. Therefore, a basic understanding and training of GIS is needed to link the available data to the fitting tools in order to answer the geo-analytical question. Natural Language Processing (NLP) is also an important method, which can help in interpreting the common patterns among the geo-analytical questions and their corresponding answers. NLP is, as defined by Canbek and Mutlu (2016, pp. 594), the "analysis of linguistic data, most commonly in the form of textual data such as documents or publications, using computational methods". In the field of NLP and information science, the definition of question-answering is: "the methods, processes, and systems which allow users to ask questions in the form of natural

language sentences and receive one or more answers, often in the form of sentences" (Laurent, Séguéla & Nègre (2006) in Mai, Yan, Janowicz, & Zhu, 2019, pp. 2).

Suppose someone is interested in the percentage of people smoking per neighborhood in Amsterdam. The GIS question would be: 'What is the percentage of people smoking per neighborhood in Amsterdam?'. Since the answer to this question is probably not directly available on the web, it needs to be answered by making use of workflows. That is why this question is a geo-analytical question, rather than a geographic question, due to the fact that this question is meant to be answered with so-called GIS workflows. As defined by ESRI (2020b), a workflow is a "set of tasks carried out in a certain order to achieve a goal". A workflow is the bridge between the available data and the goal the user specified. Gil et al. (2007) examined the challenges of scientific workflows, but they concluded that workflows should serve as reliable representations of complex computations and that workflows should be the number one entity in describing data and linking that data, especially in the field of computer science. A GIS workflow is a common tool used in software, such as ArcGIS Pro. The essential idea of GIS concerns designing workflows as answers to geo-analytical questions of which the answer is yet unknown (Scheider et al., 2020). The following example will help illustrate this important GIS tool, and why it is needed to answer the geo-analytical question.

Figure 1

Workflow to answer the question: 'What is the percentage of people smoking per neighborhood in Amsterdam?'

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix xml: <http://www.w3.org/XML/1998/namespace>.
@prefix wf: <http://geographicknowledge.de/vocab/Workflow.rdf#>.
@prefix tools: <http://geographicknowledge.de/vocab/GISTools.rdf#>.
@prefix arcpro: <https://pro.arcgis.com/en/pro-app/tool-reference/>.
@prefix RIVM: <statline.rivm.nl/#/RIVM/nl/dataset/>.
@prefix cbs: <https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/>.

# @author Romay Evers

#Workflow the percentage of people smoking per neighborhood in Amsterdam
# Workflow metadata (result and data sources)

_:wf1 a wf:Workflow;
    rdfs:comment "What is the percentage of people smoking per neighborhood in Amsterdam?"@en;
    wf:source cbs:wijk-en-buurtkaart-2019; #neighborhood
    wf:source RIVM:50052NED%2Ftable%3Fts%3D1590482338197;
    wf:edge _:wf1_1, _:wf1_2, _:wf1_3.


_:wf1_1 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/data-management/select-layer-by-attribute.htm> ;
    wf:input cbs:wijk-en-buurtkaart-2019 ;
    wf:output _:neighborhoodsAmsterdam2019_0.

_:wf1_2 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/data-management/copy-features.htm> ;
    wf:input _:neighborhoodsAmsterdam2019_0 ;
    wf:output _:neighborhoodsAmsterdam2019.

_:wf1_3 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/data-management/add-join.htm> ;
    wf:input _:neighborhoodsAmsterdam2019 ;
    wf:input RIVM:50052NED%2Ftable%3Fts%3D1590482338197 ;
    wf:output _:percentagesmokersneighborhoods.
```

The corresponding workflow to answer the question consists of three ArcGIS Pro tools: select-layer-by-attribute (WF1_1), copy-features (WF1_2) and add-join (WF1_3). The data sources used in this question are RIVM, which is the National Institute for Public Health and the Environment and CBS, which is a Dutch governmental institution and roughly translates to Statistics Netherlands. The URI's for the tools can be found on the ArcGIS Pro website. As illustrated, a total of three tools were needed to answer this geo-analytical question. The CBS data was used to import the different neighborhoods in The Netherlands and the percentage of people smoking per neighborhood was retrieved from RIVM. The tool select-layer-by-attribute creates and returns a new layer of the input (CBS neighborhoods of The Netherlands), with the given criteria. In this case, a new layer was created by only selecting the neighborhoods in Amsterdam. The tool copy-features copies features from a layer to a new feature class. In this case, the tool creates a new feature class from the newly created layer of the neighborhoods in Amsterdam. The final tool add-join joins two layers based on a common field. In this workflow, it adds the two layers – neighborhoods in Amsterdam and people smoking per neighborhood – based on the common field: neighborhoods in Amsterdam. By doing this, the smoking numbers will automatically be added to the neighborhoods. By following this workflow, a thematic map can be created in ArcGIS Pro, which illustrates the percentage of people smoking per neighborhood in Amsterdam. So, by making use of workflows, the available datasets and tools are matched in order to answer the geo-analytical question, since the answer to this question is not readily available on the web.

The missing piece of the puzzle is understanding how geo-analytical recourses can be captured with questions they answer. The QuAnGIS project aims to understand this. The QuAnGIS project is a research project with a duration of five years at Utrecht University. The project launched in early 2019 and is funded by the European Research Commission. The project works on developing a theory about 'interrogative spatial concepts', which are needed to turn geo-analytical questions into a machine-readable form. By analyzing question-answering systems, Semantic Web technology and GIS workflows, the project aims to automatize the geographic question-answering systems. Up to this day, both tools and data sources are still divided by technicalities specific to only one software environment (Scheider, Ballatore, & Lemmens, 2019). It would be beneficial if the QuAnGIS project would eventually develop a theory in which data analysists and researchers would be able to ask questions in familiar, generic and no platform-specific terms in order to obtain the tools and data necessary to answer these questions (Scheider et al., 2019). Successfully developing this theory would mean a tremendous breakthrough in the field of Geographic Information Systems, because it

would mean that geo-analytical questions can directly be matched with GIS tools and data on the Web (Universiteit Utrecht, 2018).

1.1 Purpose

This thesis will be a part of the bigger picture, the QuAnGIS project. It will be an exploratory research on finding common patterns in the answering of geo-analytical questions by creating scenarios and analyzing the corresponding workflows and the GIS tools used in those workflows. This research will be a starting point in filling the gap of automatic retrieval of answers to geo-analytical questions. In the course of this project, it will investigate the structure of corresponding research questions asked in Human Geography. Several knowledge gaps will be answered.

- What are the kinds of workflows (which GIS tools are used)?
- Which standard data sources are relevant?
- Which GIS tools are commonly used together?

By making an attempt at answering these questions, the first steps will be taken in creating a universal basis which will enable Human Geographers to ask geo-analytical questions and automatically retrieve the workflows needed to answer them. The aim is to help prepare an empirical basis for a Semantic Web based question-answering system and therefore, this thesis will be a fundamental part of the overall goal of the QuAnGIS project. Several methods of analysis will be explored in order to find common patterns in geo-analytical questions and workflow answers in parallel. This will be done on an experimental basis, in order to find the most useful methods of analysis. The focus of this thesis will be on Human Geography scenarios, namely the livability and health of residents in Amsterdam. The geo-analytical questions formulated and answered will relate to this scenario in some way. One more, current topic will be included as well; the corona virus, which might influence the livability and health of residents in Amsterdam. Therefore, this thesis distinguishes itself by focusing on one scenario and will only be the tip of the iceberg in the evolving and relatively new field of geo-analytical question-answering.

1.2 Research questions

In accordance with the previous literature and the QuAnGIS project, several research questions are proposed.

Q1: "Which types of GIS workflows answer which kinds of geographic questions?"
The first question will provide insight into the different types of GIS workflows that exist in this scenario and what kind of geographic questions they answer. By answering this question, the first patterns in geo-analytical questions and workflows in parallel may be discovered.

Q2: "Which part of a question is answered by which part of a GIS workflow?"
The purpose of the second question is to discover parallels between parts of a question and parts of a GIS workflow. A geo-analytical question consists of several parts (section 3.6.5) and by analyzing the workflows that answer the specific parts of a question, patterns may be discovered.

Q3: “Do similar workflows answer similar questions?”
The third question will form a guideline in exploring the findings found in answering question one and two. The patterns in parallels between both (parts of a) GIS workflow(s) and (parts of a) question(s) will be analyzed in order to answer the final question. By making use of the literature and an ontology (section 2.1 and 3.4), the several patterns existing will be explained, and thereby, the first step will be taken in creating a universal basis for automatically answering geo-analytical questions.

2 Literature

2.1 Geo-analytical QA problem

Question-answering research and products have increased rapidly in recent years, in both academy and industry (Mai et al., 2019). Several efforts have been made with question-answering by making use of fact retrieval (geographic question-answering). However, geo-analytical question-answering (questions meant to be answered with workflows) remained almost untouched, although those types of questions contribute substantially to daily communication (Mai et al., 2019), partly due to the rise in the use of smart assistants, such as Siri and Alexa (Canbek & Mutlu, 2016). Although question-answering systems designed for answering complex geographic questions are highly demanded, they are not quite available yet (Chen et al., 2013). This leads to a technology gap, which is yet to be filled. That is why geographic question-answering systems have been given more attention in Geographic Information Science (Mai et al., 2019). Human posed geographic questions differ fundamentally from computational geographic questions in several ways; many questions are highly subjective and dependent on their context (Mai et al., 2019) and human-posed questions are often vague and uncertain at the conceptual level (Bennett, Mallenby, & Third, 2008). Furthermore, geo-analytical questions asked by professionals in (Human) Geography are even more complex and have to be answered by more deliberate answers. This is partly due to varying complexity of the posed questions. Questions can be ranked on their level of complexity, from simple questions to very complex questions, and this is determined by their characteristics (Xu et al., 2020). Location questions are the simplest form of questions. An example is: “Where is Rome?”. More complex forms of questions include (from less complex to very complex): condition, routing, pattern modeling, trend modeling, and what-if modeling questions (Xu et al., 2020). The challenge lies in the several building blocks of the questions. It may even be necessary to decompose a question in simpler parts in order to fully be able to answer the geo-analytical question in workflows. Therefore, a question-based GIS should be conceived as a transformation task, rather than a query task (as in ordinary QA). An example to help illustrate this is provided next, based on the examples provided by Scheider et al. (2020); consider the following questions: “How far is it from Amsterdam to Rome?” and “How much is Tom exposed to green space while running through Amsterdam?”. The first question might be directly answered from a knowledge base or database, while the second cannot be answered by a standard database and the reason is twofold; first, the answer is not readily available, because it depends on the analytical parameter (Tom’s particular run). Second, it is not obvious

how an answer to this question can be generated, even if we presume Tom's particular run would be available in a knowledge base somewhere. In order to attempt at answering this question, a choice has to be made; which combination of data sets and GIS tools would give a valid answer? Which set of transformations should be performed in order to get a valid answer to the question-based GIS? The main difference between these two types of questions is that the answer to the geo-analytical question does not lie in a knowledge base but is rather a reference to a geo-analytical tool (Scheider et al., 2020). According to a recent study (Kruiger, Kasalica, Meerlo, Lamprecht & Scheider, n.d.), creating useful workflows to use in analysis with GIS requires a lot of expertise, including background knowledge of data sources, their semantics and particular qualities and formats, and knowledge of GIS functions.

In the literature, this challenge is called the geo-analytical QA problem (Scheider, Ostermann, & Adams, 2017), which is part of a more general venture of indirect question-answering. Indirect in this sense means that it is not possible to directly filter out questions by queries, but the answers need to undergo several transformations first, and therefore, workflows are needed to answer these kinds of indirect, geo-analytical questions. As mentioned before, answering geo-analytical questions requires more than what current question-answering technology has to offer. That is because the answers are basically unknown, and because of that, they need to be given as workflows instead of queries. This calls for creativity in finding an answer. In order to guide this necessary creativity, core concepts of spatial information could form a guideline, not only for understanding how these geo-analytical questions and especially the answers are composed, but for also for being certain if geodata is fit for its purpose, namely; answering the geo-analytical question (Scheider et al., 2020). The core concepts of spatial information were proposed by Kuhn (2012), and they serve as a framework to study the data and make sense of the environment. Kuhn (2012) introduced these core concepts of spatial information as a simple and computational interface for Geographic Information Systems (Scheider et al., 2016) and they include terms such as Location, Field and Concept.

Scheider et al. (2016) researched the different ways of interpreting geodata, in terms of core concepts and how geodata can be made explicit in a semantic type system. They provided a "systematic investigation of data type representations of core concepts of spatial information" (pp. 3). Their purpose was not to (re-) design a novel abstract interface for GIS – like the original purpose of Kuhn's core concepts (2012), mentioned above – but rather to improve the (re-) usability and accessibility of existing tools and data sources. Important to note is that core concepts and date types are rather different. Core concepts are an interpretation of an analyst, while data types are data artefacts that represent concepts in an indirect matter (Scheider et al.,

2016). Scheider et al. (2016) successfully proposed an ontology design pattern, which helps answer geo-analytical questions; CCDT. CCDT is an abbreviation of *core concept data types* ontology, which is a semantic type system that can be used to constrain GIS functions for workflow synthesis (Kruiger et al., n.d.). The ontology captures the several ways core concepts can be represented in geographic data models on different levels of measurement (Kruiger et al., n.d). Scheider et al. (2016) proposed this CCDT ontology, because creating GIS workflows requires more than just fitting existing data types to inputs and outputs (Scheider et al., 2019). Scheider et al. (2016) showed that the developed CCDT ontology is able to express diverse kinds of geo-analytical questions. More importantly, the geo-analytical questions posed by researchers can be answered in the form of valid, automatically construable GIS workflows by making use of the ontology. Therefore, the CCDT ontology is a helpful tool in developing and answering geo-analytical questions because it is able to capture the variety of the questions and its implications for geographic analysis and the different forms of geodata.

2.2 Previous work

Some other tools to help understand and analyze geo-analytical questions and the corresponding answers exist as well. To understand the basic structures of the workflows, several techniques can be used, from both statistics and NLP. Like mentioned before, a GIS application for answering geo-analytical questions has gained a lot of interest in the GIS domain (Yin et al., 2019). Previous studies have sought to introduce some interaction methods for GIS, but a QA-based system made to simplify spatiotemporal analyses do not have a large body of literature to support it yet. However, several attempts at other related work have been made. Some examples in the field of Artificial Intelligence are a human-centered multimodal GIS interface design in order to support emergency management situations (Rauschert, Agrawal, Sharma, Fuhrmann, Brewer, & MacEachren, 2002) and 3D visualization of spatial data , for which virtual reality was used (Huang, Jiang, & Li, 2001). These are examples of the more practical side of the quest for a universal approach to geographic question-answering. In the domain of computer science, the quest is more focused on the theoretical side, on how to train a better QA system (Yin et al., 2019). The focus of this domain has been primarily on investigating the ontology to spatial analysis, such as the ontology developed by Scheider et al. (2016) and creating QA datasets for training purposes (Yin et al. 2019). So, QA has a long history in Artificial Intelligence and Computational Linguistics (Scheider et al., 2020). The first systems research dates back to the year 1970. Back then, these systems did not have much

impact, but they are still relevant today because they touched upon several techniques, such as linguistic grammars and semantic frames for parsing questions (Scheider et al., 2016). Over the years, the field of QA systems has evolved, due to the rise of the World Wide Web (WWW). The WWW made large query and answer sets available (Lin, 2002) and thereby improved the general Information Retrieval systems. By building these QA systems, research may make use of a structured database of knowledge or information (explained in the previous section). On top of that, QA datasets, also called question corpora, have been developed in the past years. These question corpora can help analyze question structures from both a syntactic and semantic viewpoint (Yin, 2019). Current research has been focusing on the collection and analysis of Web queries, and on questions with directly retrievable answers for QA systems, like geographic questions (Xu et al., 2020). However, a geo-analytical question corpus did not exist yet, which made it impossible to analyze why and how geo-analytical questions differ from their simpler form; geographic questions. This year, Xu et al. (2020) proposed a novel geo-analytical question corpus 'GeoAnQu', the first geo-analytical QA database, containing 429 geo-analytical questions, from two different sources: scientific articles and textbooks. By building upon this corpus, newly created geo-analytical questions can be added to the now existing corpus. Two helpful tools in extracting semantic information from the questions are encoding and parsing, by differentiating between the intent of the question and the description and criteria of the intent (Xu et al., 2020). The intent, in the field of geo-analytical QA, can be described as the definition of a valid answer to the question. The criteria and descriptions determine the restrictions in order to yield that valid answer. Question corpora, especially GeoAnQu, are useful tools for both investigating geo-analytical questions asked in (Human) Geography and for training parsers of QA systems (Xu et al., 2020). So, by adding to the current corpus, both the investigation and parsing of geo-analytical questions can be improved, and therefore, the overall geo-analytical QA problem will be closer to a universal approach.

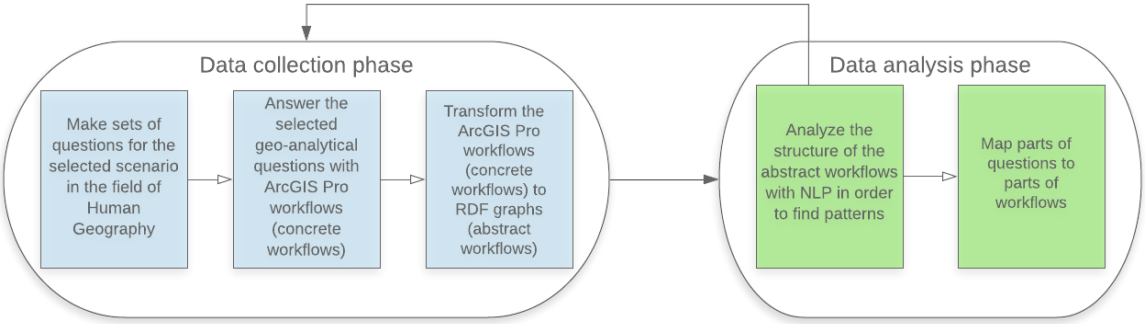
3 Method

In this section, the methodologies for the data analysis and collection of data in this research are described. First, a brief overview of the research is provided, in the form of a flowchart (figure 1). Second, the two types of workflows, concrete and abstract, are described. This is the data collection part. Third, the methods used for the analysis are described, from Levenshtein's distance analysis to TF-IDF. This is the data analysis part.

3.1 Workplan data collection & data analysis

Figure 2

Workplan showing the steps in the data collection and data analysis phase



A workplan consisting of five steps in two phases is provided in figure 2, in order to stepwise explain the plan for this research. The first phase was the data collection, which consists of three steps. First, the scenario had to be selected. The chosen scenario will be described in the next section (3.2). In line with that scenario, the geo-analytical questions were selected in order to contribute to the central theme in the scenario. Second, the selected questions were answered with workflows created in ArcGIS Pro. Those workflows also go by the name; concrete workflows. Several terms will be described in section 3.3. Followed by that, the concrete workflows were transformed into so-called abstract workflows, which are workflows in the form of RDF graphs. This term will be explained in further detail in section 3.3. The RDF graphs were the final data, and the data collection was thereby complete.

The next phase was the data analysis and had the purpose of analyzing the created abstract workflows, the RDF data. Several methods of analysis were performed, which will be explained in detail in section 3.6. In the primary stage of the research, the analyses were performed on a smaller dataset in order to validate the methods used for the analysis. After successfully performing the first analyses, the analyses were performed again, on a bigger

dataset (a total of 27 abstract workflows). The final step in the data analysis phase was to make sense of the results obtained in the first step. This was done by mapping the parts of questions to parts of workflows. This final step provided the research with an empirical basis for future research in the field of Human Geography and other fields.

Several terms used in the workplan will be defined in the next sections in order to fully understand the several steps. Since this thesis is part of the QuAnGIS project, the methodology was partly dependent on the whole project. Nonetheless, the proposed methods and data used will be provided in the next sections.

3.2 Scenario

The scenario of this thesis is the livability and health of residents in the city of Amsterdam. This focus is derived from a minor course in GIS: GIMinor GeoAnVis (Geographic Analysis and Visualization) (Vrije Universiteit, 2020). The students in this course were assigned a case study about the livability of Amsterdam. Their goal was to create a livability atlas of the city and thereby focusing on different aspects. Two of those livability atlases will be used in this thesis, namely the students' reports of group eight and fifteen. The selection was based on the level of completeness of the reports. Both groups had the most complete reports, provided with the links to the needed data and the creation of concrete workflows. Group eight focused its research on the livability of Amsterdam for students, while group fifteen focused its research on the livability of the elderly during a heatwave. The quality of the students' workflows made were verified by the course instructors, and therefore, the assumption was made that the correctness is assured.

Livability has been an important topic in Human Geography, especially in Urban Geography (Pacione, 2009). Urban designers have made several efforts in improving the livability of a city by improving the built environment and social, economic and natural factors, such as safety levels, health, exposure to noise and access to nature and green, among other things (Southworth, 2016). The components of the concept livability are diverse and complex, and the students in the GeoAnVis course tried to grasp a small aspect of the livability of cities. Their focus was Amsterdam, the capital city of the Netherlands.

3.3 Data and software availability

This research required several data sources, software environments and the assistance of the members of the QuAnGIS project (Simon Scheider, Enkhbold Nyamsuren and Haiqi Xu).

Several people contributed to the data sample; the students who wrote the GeoAnVis reports, which were two groups (eight and fifteen), both consisting of three students. The first nineteen workflows were based on their reports. Group eight's report was named 'Facilitating the perfect student life', and group fifteen's report was named 'Mapping the livability of the elderly during heat waves'. To extend the data sample, eight more, newly created workflows were added. In total, the data sample consists of 27 workflows. In this section, several data sources, terms and software environments will be explained.

3.3.1 ArcGIS Pro

ArcGIS Pro is a desktop GIS application, which can be used to explore, visualize and analyze data. Thematic maps, concrete workflows (ModelBuilder) and charts can be created with ArcGIS Pro (ESRI, 2020a). In this research, ArcGIS pro was used to create concrete workflows.

3.3.2 ModelBuilder

ModelBuilder is a visual programming language for building geoprocessing workflows (ESRI, 2020d). ModelBuilder creates concrete workflows based on data and tools used in GIS. A concrete workflow is executed in a GIS environment and is formed by importing selected (geo)data, resources and addressing certain additional tools to it (Wang, Ge, Rizos, & Babu, 2004). So, a concrete workflow can be created with ModelBuilder, the built-in application of ArcGIS Pro.

3.3.3 GitHub

GitHub is a code host, which became the largest in recent years (Gousios, Vasilescu, Serebrenik, & Zaidman, 2014). GitHub supports GIT (distributed version control), pull-based development and social coding (Gousios, 2014). In this research, GitHub was used to share the students' reports, share the created workflows and correctly store the available data. The reports were made available by the members of the QuAnGIS project, containing the atlases with the corresponding concrete workflows (https://github.com/simonscheider/QuAnGIS). All reports, atlases, created concrete and abstract workflows can be found in the GitHub repository GuAnGIS (https://github.com/simonscheider/QuAnGIS). The code for the analyses was also provided by the members of the QuAnGIS project (https://github.com/quangis/WorkflowSimilarityAnalysis).

3.3.4 RDF

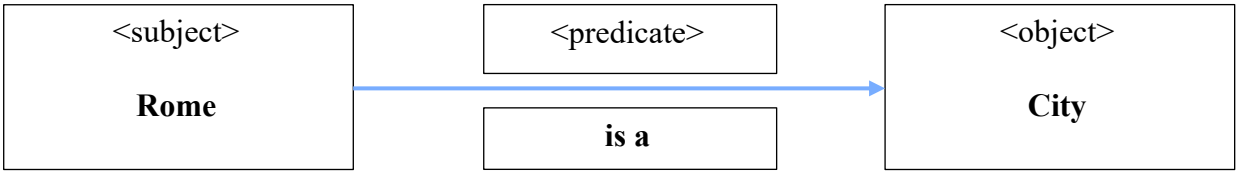
The next generation of the World Wide Web is often denoted as 'Semantic Web'. Information on the Semantic Web is not only intended for human readers anymore, but it serves another goal as well; processing information by machines, enabling intelligent information services and semantically empowered search engines (Decker et al., 2000). The Semantic Web is a Web of data, and it is important to have a huge amount of data available and stored in a standard format (W3C, 2015a). Not only the access to that data reachable and manageable by Semantic Web tools should be made available, but the relationships among data as well (W3C, 2015b). Linked data is the term used for this collection of interrelated datasets on the Semantic Web.

In order to properly achieve and create linked data, technologies should be available for a universal format. The World Wide Web Consortium (W3C) recommends the Resource Description Framework (RDF) in achieving this (Lassila & Swick (1999) in Decker et al., 2000). RDF is a foundation for processing metadata, and it provides "interoperability between applications between that exchange machine-understandable information on the Web" (Lassila & Swick, 1999 pp. 1), in other words: RDF is a framework for expressing information about resources available on the Web (W3C, 2014) .

In RDF, data is visualized as so-called triples. They were given the name triples, because an RDF statement, expressing a relationship between two resources, consists of three elements: 'subject', 'predicate' and 'object' (Antoniou & Van Harmelen, 2004).; Lassila & Swick, 1999). In a triple, the subject represents a resource, the object represents either a resource or a literal, whereas the predicate represents the nature of the relationship between the subject and object. Important to note is that the predicate is phrased directionally, from subject to object. Resources in RDF are defined by a Unique Resource Identifier (URI), which can be a Uniform Resource Locator (URL) but does not necessarily have to be so. By making use of RDF in the Semantic Web, all data, workflows, and methods of analysis can be described and stored by making use of one standard. Figure 3 illustrates an example of a triple statement used in RDF.

Figure 3

Triple statement of 'Rome is a city'



Ontologies have been developed to guide and structure all this information available on the Semantic Web. Ontologies can be used in defining the several concepts and relationships, which describe and represent an area of interest (W3C, 2015b). The CCDT ontology (Scheider et al., 2020) is an example of an ontology developed for answering geo-analytical questions.

The workflows created in RDF will be called abstract workflows in this thesis, since they are abstractions of concrete workflows. An abstract workflow is defined in terms of ontological concepts and is written in the form of triplets. RDF can be represented by different syntaxes, such as XML syntax, N-triples, and JSON. However, the preferred syntax is Turtle, which will be used in this thesis too. The Turtle syntax provides a human readable instantiation (Chiarcos, & Fäth, 2017). Turtle represents RDF statements, which are labeled 'edges' in the graph, as triples (see bold text in figure 4). A triple consists of a source node ('subject'), an edge ('property'/'relation') and a target ('object'). The Turtle triple is concluded with a dot (Chiarcos, & Fäth, 2017).

Listing 1

Example of Turtle triple statement

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix xml: <http://www.w3.org/XML/1998/namespace>.
@prefix wf: <http://geographicknowledge.de/vocab/Workflow.rdf#>.
@prefix tools: <http://geographicknowledge.de/vocab/GISTools.rdf#>.
@prefix arcpro: <https://pro.arcgis.com/en/pro-app/tool-reference/>.
@prefix pdok: <https://www.pdok.nl/introductie/-/article/>.
@prefix maps: <https://maps.amsterdam.nl/>.

# @author Romay Evers

#Workflow the average distance to hospitals per PC4 area in Amsterdam
# Workflow metadata (result and data sources)

_:wf1 a wf:Workflow;
      rdfs:comment "What is the average distance to hospitals per PC4 area in
Amsterdam?"@en;
      wf:source maps:open_geodata%2F%3Fk%3D192%2F;  #PC4 areas
      wf:source maps:functiekaart; #hospitals
      wf:edge _:wf1_1, _:wf1_2, _:wf1_3, _:wf1_4, _:wf1_5, _:wf1_6.

_:wf1_1 tools:implements <https://pro.arcgis.com/en/pro-app/tool-
reference/analysis/select.htm> ;
      wf:input maps:functiekaart ;
      wf:output _:hospitalsfeature0.
```

Turtle has some benefits in working with the syntax. First, abbreviations are possible, in this thesis abbreviations can be found in the form of prefixes (see the pink text in listing 1). Second,

whitespaces are not relevant in Turtle. Third, after a '#', an explanatory text can be written (see gray text in listing 1).

An example of both a concrete and abstract workflow will be provided in the next section, thereby illustrating and explaining the different elements in more detail.

3.3.5 Python

Python is a computer language, which is general-purpose and open source (Lutz, 2001). In this research, its purpose is to serialize the RDF datasets, which is the first step in the data analysis.

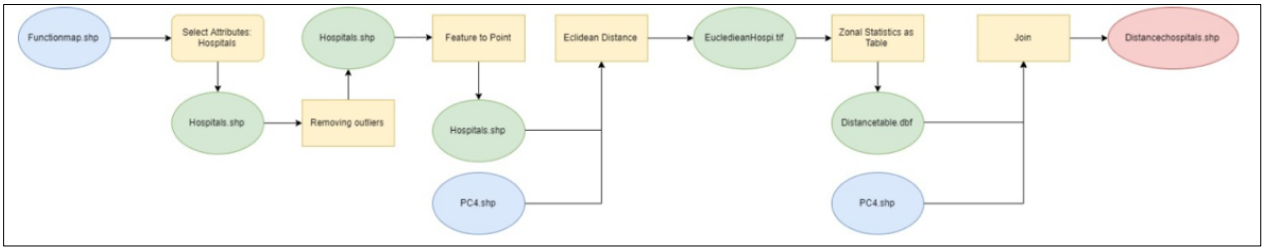
3.3.6 Rstudio

Rstudio is a relatively new Integrated Development Environment (IDE) specially designed for the programming language R (Allaire, 2012). It enables users to combine the various components of R. In this research, Rstudio is used to perform the analyses on the output created with the use of Python.

3.4 Data collection

The first nineteen abstract workflows created for this thesis were based on the scenarios created by students' groups eight and fifteen. A text editor was used to write down the corresponding abstract workflows. Below, both the concrete and abstract workflow are shown, in order to explain the methodology more clearly.

Figure 4
Concrete workflow



Listing 2

Abstract workflow

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix xml: <http://www.w3.org/XML/1998/namespace>.
@prefix wf: <http://geographicknowledge.de/vocab/Workflow.rdf#>.
@prefix tools: <http://geographicknowledge.de/vocab/GISTools.rdf#>.
@prefix arcpro: <https://pro.arcgis.com/en/pro-app/tool-reference/>.
@prefix pdok: <https://www.pdok.nl/introductie/-/article/>.
@prefix maps: <https://maps.amsterdam.nl/>.

# @author Romay Evers

#Workflow the average distance to hospitals per PC4 area in Amsterdam
# Workflow metadata (result and data sources)

_:wf1 a wf:Workflow;
      rdfs:comment "What is the average distance to hospitals per PC4 area in
Amsterdam?"@en;
      wf:source maps:open_geodata%2F%3Fk%3D192%2F;  #PC4 areas
      wf:source maps:functiekaart; #hospitals
      wf:edge _:wf1_1, _:wf1_2, _:wf1_3, _:wf1_4, _:wf1_5, _:wf1_6.

_:wf1_1 tools:implements <https://pro.arcgis.com/en/pro-app/tool-
reference/analysis/select.htm> ;
      wf:input maps:functiekaart ;
      wf:output _:hospitalsfeature0.

_:wf1_2 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/data-
management/copy-features.htm> ;
      wf:input _:hospitalsfeature0 ;
      wf:output _:hospitalsfeature.

_:wf1_3 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/data-
management/feature-to-point.htm> ;
      wf:input _:hospitalsfeature ;
      wf:output _:hospitals.

_:wf1_4 tools:implements <https://pro.arcgis.com/en/pro-app/tool-
reference/spatial-analyst/euclidean-distance.htm> ;
      wf:input _:hospitals ;
      wf:input maps:open_geodata%2F%3Fk%3D192%2F;
      wf:output _:euclhospitals.

_:wf1_5 tools:implements <https://pro.arcgis.com/en/pro-app/tool-
reference/spatial-analyst/zonal-statistics-as-table.htm>;
      wf:input _:euclhospitals ;
      wf:output _:eucltablehospitals.

_:wf1_6 tools:implements <https://pro.arcgis.com/en/pro-app/tool-
reference/geoanalytics-desktop/join-features.htm> ;
      wf:input _:eucltablehospitals ;
      wf:input maps:open_geodata%2F%3Fk%3D192%2F ;
      wf:output _:distancehospitals.
```

Figure 4 shows the concrete workflow created by the students in group 15, in order to answer the geo-analytical question: "What is the average distance to hospitals per PC4 area in Amsterdam?". A PC4 area is an area identified by a four-digit postal code, such as 1064 or 1057. All the questions can be categorized among different the different types (QuAnGIS, n.d.). This example is a 'what' question type, since the questions starts with 'what'. Other types of questions are 'how' and 'which'. Most of the other types of questions can be rephrased into 'what' questions, which was done in this research too.

The model consists of different elements, building blocks (ESRI, 2020c). The blue building block in a model is a Data variable, more specifically; Input data. Variables in a model are elements that hold a value or a reference to data. In the example above, Functionmap.shp and PC4.shp are input data. The yellow building blocks are geoprocessing tools added to the model. They perform various operations on geographic data or tabular data. Among other tools, Euclidean Distance and Zonal Statistics as Table are tools in the example above. The green building blocks represent a Derived or Output data variable, which is new data created by a tool in the model. Hospitals.shp and Distancetable.dbf are two of the five derived or output data variables in the example above. The next step was to transform a concrete workflow into its abstract form defined with ontological concepts.

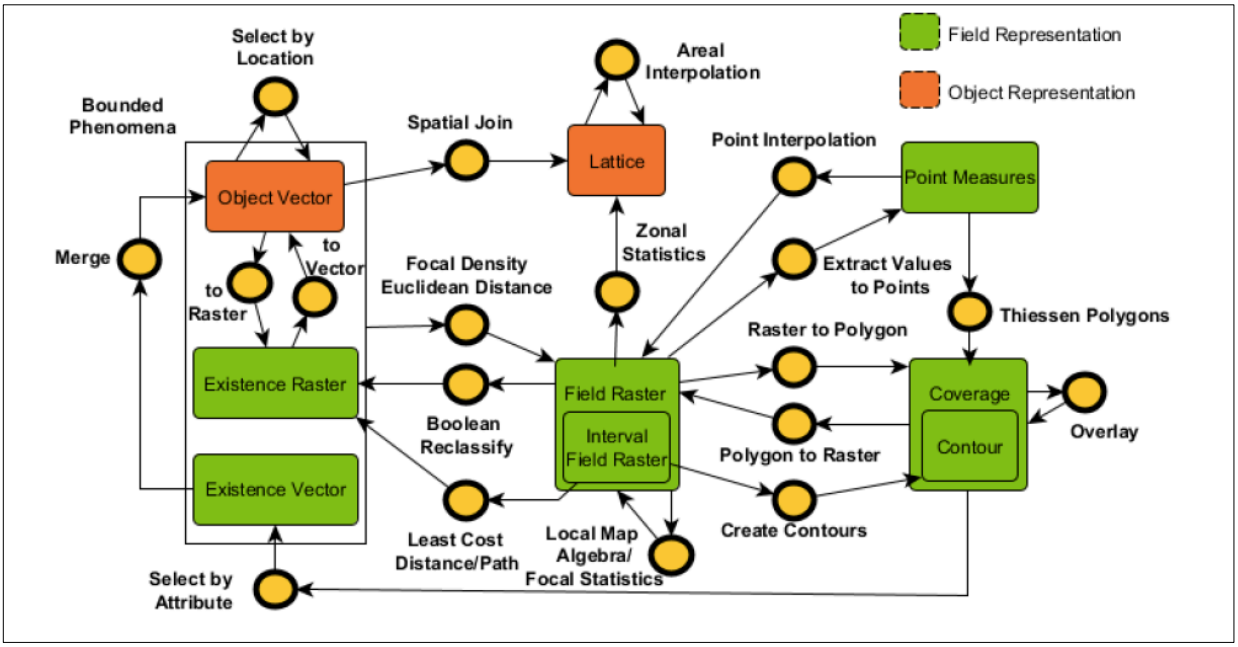
Listing 2 shows the abstract workflow that corresponds to the concrete workflow displayed in Figure 4. In this context, 'input' is equal to the blue and green building blocks in the model; input data. 'Output' is equal to the green building blocks; derived or output data variable. The link to the tool on the pro.argis.com website is equal to the yellow building blocks in the model; geoprocessing tools. In figure 4, WF1_2 was originally not in the concrete workflow created by group 15. However, it was added in the abstract workflow (listing 2) in the final data, to make the data more complete.

In order to get a better understanding of the CCDT ontology, introduced in the previous chapter, figure 5 was used as a guideline for interpreting the several workflows and the core concepts data types used in those workflows (Scheider et al., 2016). The computational diagram offers a schematic overview of the signatures and types of geo-analytical operations, which are applicable to abstract data types. The green and orange boxes denote abstract data types, and the yellow circles denote operation types, so-called geoprocessing tools. The CCDT ontology helps in providing explanation in why some patterns in questions and workflows may occur and can be used as a framework. All in all, a total 27 abstract workflows were created, of which nineteen were based on the students' reports. Eight workflows were added to that sample of

nineteen workflows, in the field of livability. Those eight workflows were created by the author of this thesis and they were reviewed by the QuAnGIS members, to assure correctness.

Figure 5

Scheider et al. (2016, pp. 18): "Computational diagram of types of geoanalytical operations applicable to abstract data types."



3.5 Validity, reliability and generalizability

The definition of validity in a quantitative study is as follows: the extent to which a concept or construct is accurately measured (Heale & Twycross, 2015). Reliability relates to the consistency of a measure, the accuracy of an instrument (Heale & Twycross, 2015). Considering validity and reliability of the data collection is an important part of conducting research. However, assessing the level of validity and reliability in this particular research is difficult, because the used data is not purely quantitative, since the workflows used to answer the geo-analytical questions are open to interpretation and might differ among researchers. This causes the validity and reliability hard to asses, because there are several workflows suitable to answer the same geo-analytical question. Therefore, the term generalizability is used to assess the quality of the used method (Polit & Beck, 2010). The aim of this exploratory research is to help form an empirical basis to perform future analyses on. The goal is to be able to generalize question-answering methods of analysis across different kinds of questions and corresponding workflows. Therefore, the generalizability of the approach is central and the used method in this thesis will be central to analyze new kinds of questions in the same field, or other fields.

3.6 Data analysis

An exploratory research was done in order to find out if it was possible to conduct analyses on the data. The results of this exploratory data analysis (EDA) showed that it is possible to analyze the data, and analyses were performed on the 27 abstract workflows. All the analyses were performed in RStudio, and the script was written by another researcher part of the QuAnGIS research group. During the EDA, multiple methods to analyze the data have been used, and the results of this EDA decided on which data analysis techniques were used in the end. The results of those techniques will be explained in the next sections.

3.6.1 Serialization

The first necessary step in making the data analysis work, was to perform the serialization of the abstract workflows into a character string format. The reason for this is that it will be easier to apply statistical analyses on serialized workflows rather than workflows in RDF graph format. So, serialization is an essential step in making the collected data interpretable. After serialization, the data can be stored and transferred across domains and other computers (Chiarcos, & Fäth, 2017).

The serialization matches each tool in the workflow to an ASCII character, so the serialization transforms a workflow in the form of an RDF graph into a string of letters. The tool to character matching can change dynamically, due to rerunning the Python script. Every time the Python script is rerun, the R script should be rerun too, due to the dynamic change of the tool to character matching. Because the abstract workflows are RDF graphs, a few rules needed to be applied. This was done in order to make the RDF workflow into a single string.

Creating a single string of the workflow is necessary in order to be able to compare the different workflows. The rules applied in this thesis are as followed: the first step in the serialization is the dividing of the workflow into a sequence of execution cycles. Each execution cycle executes a minimum number of one operation, but can execute more than one at once. So, each execution cycle contains one or more operations. Second, an operation should be executed as soon as possible (the earliest cycle). That is the when all input data used in that operation becomes available in the workflow. Third, the term ‘parallel operations’ applies, which means: if it possible to execute multiple operations at the same cycle, then the operations in that cycle are executed in the order of the ASCII character corresponding to the tool of that operation. Fourth, because the same tool can be executed in a single workflow multiple times, each execution is regarded as a distinct operation.

Listing 3
Example serialization

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix xml: <http://www.w3.org/XML/1998/namespace>.
@prefix wf: <http://geographicknowledge.de/vocab/Workflow.rdf#>.
@prefix tools: <http://geographicknowledge.de/vocab/GISTools.rdf#>.
@prefix arcpro: <https://pro.arcgis.com/en/pro-app/tool-reference/>.
@prefix pdok: <https://www.pdok.nl/introductie/-/article/>.
@prefix maps: <https://maps.amsterdam.nl/>.

# @author Romay Evers

#Workflow the average distance to hospitals per PC4 area in Amsterdam
# Workflow metadata (result and data sources)

_:wf1 a wf:Workflow;
        rdfs:comment "What is the average distance to hospitals per PC4 area in Amsterdam?"@en;
        wf:source maps:open_geodata%2F%3Fk%3D192%2F;       #PC4 areas
        wf:source maps:functiekaart; #hospitals
        wf:edge _:wf1_1, _:wf1_2, _:wf1_3, _:wf1_4, _:wf1_5, _:wf1_6.

_:wf1_1 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/analysis/select.htm> ; Tool ASCII
character: D
        wf:input maps:functiekaart ;
        wf:output _:hospitalsfeature0.

----------Cycle 1

_:wf1_2 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/data-management/copy-
features.htm> ; Tool ASCII character: 1
        wf:input _:hospitalsfeature0 ;
        wf:output _:hospitalsfeature.

----------Cycle 2

_:wf1_3 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/data-management/feature-to-
point.htm> ; Tool ASCII character: E
        wf:input _:hospitalsfeature ;
        wf:output _:hospitals.

----------Cycle 3

_:wf1_4 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/euclidean-
distance.htm> ; Tool ASCII character: 9
        wf:input _:hospitals ;
        wf:input maps:open_geodata%2F%3Fk%3D192%2F;
        wf:output _:euclhospitals.

----------Cycle 4

_:wf1_5 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/zonal-statistics-
as-table.htm>; Tool ASCII character: 6
        wf:input _:euclhospitals ;
        wf:output _:eucltablehospitals.

----------Cycle 5

_:wf1_6 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/geoanalytics-desktop/join-
features.htm> ; Tool ASCII character: 3
        wf:input _:eucltablehospitals ;
        wf:input maps:open_geodata%2F%3Fk%3D192%2F ;
        wf:output _:distancehospitals.

----------Cycle 6
```

Listing 3 shows an example of a serialization of the workflow corresponding to the question: "What is the average distance to hospitals per PC4 area in Amsterdam?". The resulting string of characters in this case is 'D1E963'. To annotate the cycles as well, the string can be written as follows; 'D|1|E|9|6|3|', resulting in six operations in six cycles.

Listing 4 shows the first cycle in a serialization, executing two operations. The workflow corresponds to the question: "What is the mean distance to a subway station per neighborhood in Amsterdam?". WF1_1 and WF1_5 are executed at the same time; within the same cycle. The resulting string of characters in this example is '091416'. To annotate the cycles as well, the string is written as '09|1|4|1|6|'. The serialization of this workflow results in a string with six operations in five cycles.

Listing 4

Example of two operations within one cycle

```
_:wf1_1 tools:implements <https://pro.arcgis.com/en/pro-app/tool-reference/data-
management/select-layer-by-attribute.htm> ;  tool ASCII character : 0
       wf:input pdok:cbs-gebiedsindelingen ;
       wf:output _:amsterdam0.

_:wf1_5 tools:implements <https://pro.arcgis.com/en/pro-app/tool-
reference/spatial-analyst/euclidean-distance.htm> ;  tool ASCII character: 9
       wf:input maps:trammetro ;
       wf:output _:outputbackdirectionraster ;
       wf:output _:outputdirectionraster ;
       wf:output _:euclideandistancesubwaystation.

----------Cycle 1
```

By performing the serialization and applying the above-mentioned rules, it is possible to create a comparable string of ASCII characters, corresponding to the workflow (an RDF graph). Two or more similar workflows will become two or more similar orders of cycles, and similar order of operations within those cycles, due to the serialization and that causes two or more strings to become comparable. String serialization can therefore be used to compare the workflows and is the first step in the data analysis part.

3.6.2 Levenshtein's distance analysis

The next step was to perform the Levenshtein's distance analysis. Quantifying the similarity between two strings has been an important topic in science, especially in the field of text retrieval and NLP (Yujian, & Bo, 2007. The Levenshtein's distance is often called edit distance, which refers to the method: it compares strings by various edit operations, including

deletion, insertion and substitution. The Levenshtein's distance is the minimum edit distance: the number of single-character edit operations required to change one string into another. The Levenshtein's distance is a common metric for measuring the difference between two strings. The higher the number, the more different the two strings are. For example, the Levenshtein's distance for the strings 'Rome' and 'Romay' has a value of two, since one substitution ('e' → 'a') and one insertion ('y') is needed to transform Rome into Romay.

One problem that could be encountered is when strings with different lengths are not compared correctly. For example, when analyzing a pair of strings with a dissimilarity of one character, that one dissimilarity may be less critical to a string containing six characters than a string containing just two characters. The solution to this problem is normalization. By applying normalization, the dissimilarity values are divided by the length of the longest string in the pair. This makes the pair of strings better to compare. By comparing the normalized dissimilarity values of the strings in the pair, it can be analyzed if similar geo-analytical are answered by similar workflows. This can be used in discovering patterns in questions and workflows in parallel.

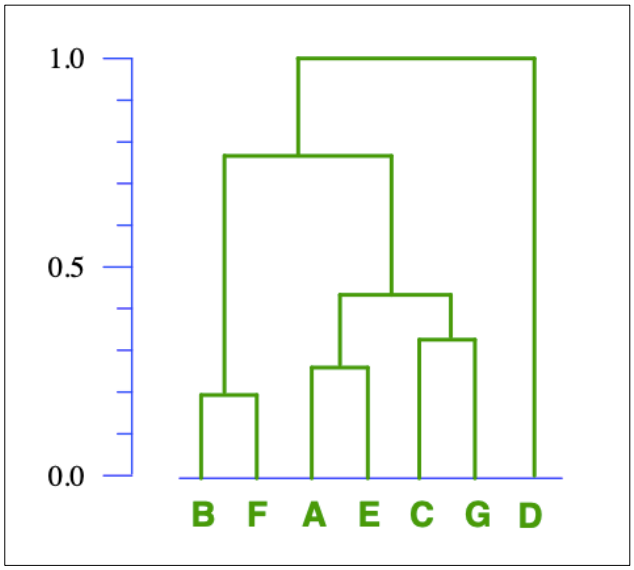
The Levenshtein's distance analysis was, after normalization, very abstract, and therefore, conclusions could not be drawn. However, the created dissimilarity matrices were used in the hierarchical clustering analysis in the next section.

3.6.3 Cluster analysis

Clustering is concerned with creating groups where objects in the same cluster are similar to each other and dissimilar to objects inside another cluster. Since the first publications, hierarchical clustering has been a popular method in science, since it is a good way to represent clusters of empirically measured relations of similarity (Johnson, 1967). According to Karypis, Han, and Kumar (1999, p.p. 1), "clustering in data mining is a discovery process that groups a set of data such that the intracluster similarity is maximized and the intercluster similarity is minimized". This means the optimal clustering should have minimal dissimilarity values for strings within a cluster and maximal dissimilarity values for the created clusters itself. Possible similar characteristics of the underlying data can be explained by these newly discovered clusters (Karypis, Han, and Kumar, 1999). The underlying data are the character strings corresponding to the workflows in this case.

Figure 6

Basic example of a dendrogram (Stanford, n.d.)



A stepwise cluster algorithm can be applied to textual data of *n* objects, which merges two objects at each step, which are the two objects which have the least dissimilarity (so the two which are most similar). The algorithm produces a nested sequence of clusters. At the top of the nested sequence, a single all-inclusive cluster can be found (all strings in a single cluster), and at the bottom, single point clusters can be found (each string is its own cluster) (Karypis, Han, & Kumar, 1999). Thus, the result of a hierarchical cluster analysis is a dendrogram, with *n-1* nodes. An example of a dendrogram is provided in figure 6, with seven clusters at the bottom and six nodes. The dissimilarity values vary from 0.0 at the bottom (completely similar) to 1.0 at the top (completely dissimilar). To illustrate, B and F are more similar to each other than B and A are. Three pairs of the samples are quite close (B&F, A&E and C&G), while sample D separates itself from the other samples. This dendrogram is a good way to visualize and interpret the results from the dissimilarity matrix.

The cluster analysis in this thesis was performed on the normalized dissimilarity values calculated in the previous step.

3.6.4 N-grams

Several other analyses were performed, following the string serialization and hierarchical clustering. One useful analysis is a N-gram. N-gram based techniques have gained popularity in modern NLP and its applications (Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2012). N-grams are sequences of elements as they appear in some form of text, in

this context; as they appear in the different character strings of the corresponding workflows. So, a N-gram can give insights into which sequences of operations are more common than others, by showing the most frequent operation sequences among all workflows (frequency ≥ 2). This could for example help in researching which tools are used together and which tool is followed by another tool more than once.

3.6.5 Co-occurrence matrix

Another analysis which was explored in this research is the co-occurrence matrix, which is a method to store the count of how often a particular pair of strings occur together. In the context of this research, the pair of things are measure *r* and a tool *s*. In GIS analysis, the notions 'measure', 'support' and 'extent' exist, and those several elements were explored in another article in the QuAnGIS project (n.d.). In this context, the measure in a geo-analytical "specifies the quality of immediate interest" (QuAnGIS, n.d. pp. 2). In a simple example question "What is the population of elderly people in Amsterdam?", the measure is 'population'. The measure in a question can be extracted by making use of parser, based on regular expressions, by automatically identifying syntactic patterns, which correspond to those elements; measure, support and extent (QuAnGIS, n.d.). The measure in a question can also be rephrased. For example, the measure 'population' can be rephrased as 'population count', in which 'count' becomes the new measure (QuAnGIS, n.d.). The support in a question refers to that part that further specifies the measure (QuAnGIS, n.d.). In the example 'population count of elderly', the support would be 'people'. The extent in a question refers to the part that spatially or temporally frames the measure (QuAnGIS, n.d.). In this example, the extent would be 'in Amsterdam'.

In this thesis, the measure in a geo-analytical question is compared to the tools used in the workflows to answer that same geo-analytical question. The more a certain measure co-occurs with one or more tools, the more specific that one (or more) tool(s) might be to that measure. For example, suppose the co-occurrence of measure *r* and tool *s* is three. That means that tool *s* was used in three workflows estimating measure *r*. However, the true values of the co-occurrence might not be easy to compare. That is why the co-occurrence values are often normalized. By normalizing the values, the value of co-occurrence of measure *r* and tool *s* is divided by the sum of the co-occurrence values of all tools co-occurring with measure *r*. The results of this normalization show a co-occurrence with a value between zero and one, and can better determine how specific tool *s* might be to measure *r*. The more the value is to zero, the

less they occur together and the more the value is to one, the more they occur together. To add to that; important to note is that a co-occurrence matrix does not account for how many times tool *s* was used in each workflow of measure *r*.

3.6.6 TF-IDF

Another analysis which was done is TF-IDF. It is quite an experimental approach of applying information retrieval and data mining in the field of geoscience. However, TF-IDF is one of the most used term weighing systems in today's information retrieval and data science and has gained popularity in the past years (Aizawa, 2003). It is a commonly used analysis, since TF-IDF has been considered an empirical method with many possible variations (Aizawa, 2003). So, the results in the next section should be considered with caution, due to the many variations and applications of the method.

TF-IDF is an abbreviation of 'term frequency – inverse document frequency'. This TF-IDF weight is a statistical measure used to determine the importance of a word to a document. TF-IDF calculates the value for each word in a document by calculating a weight. The weight is composed by two terms (Jones, 1972; Ramos, 2003; Wu, Luk, Wong, & Kwok, 2008);

- The term frequency (TF): as the term might indicate, this part measures how often a term occurs in a document. Since almost every document differs in length, the term frequency is normalized; the term frequency is divided by the length of the document (the total number of terms in the document). In the context of this thesis, the term frequency is how often tool *s* in a workflow occurs in geo-analytical question containing measure *r*.
- Inverse document frequency (IDF): this part measures a term's importance to a document. When calculating TF, it is assumed all terms are equally important in a document. However, terms like 'the', 'if' and 'this' may have a high TF, but may not be that important to a document. Therefore, it is necessary to weigh down the more frequent terms (occurring frequently in more documents), while increase the weight of the rare terms (occurring frequently in just that specific document). The IDF calculates if a term is common or rare among all documents. This is done by taking the total number of documents and dividing that number by the number of the documents containing that term. To inverse the document frequency, the algorithm is taken as final step in calculating the IDF. In the context of this thesis, the TF-IDF weight is calculated

by taking the total number of unique measures and divide that by the number of measures co-occurring with the particular tool.

The higher the TF-IDF value, the more specific a word is to a document. In the context of this thesis, the higher the TF-IDF value for tool *s* and measure *r*, the more specific tool *s* is to the measure *r,* relative to the other measures. To conclude, TF-IDF is a rather experimental approach, because it is an abstract tool in NLP and because it is not common in geoscience, but it could be a helpful analysis to capture the specificness of a tool.

4 Results

4.1 Tool to character ASCII matching

The tool to ASCII character matching results are shown in table 1. Each tool URI was matched to an ASCII character. The results changed dynamically throughout the process, due to rerunning the Python script, but the results in this section are final. In total, 23 tools in the workflows were used to answer 27 geo-analytical questions.

Table 1

Tool to character ASCII matching

Tool URI	ASCII character
https://pro.arcgis.com/en/pro-app/tool-reference/data-management/select-layer-by-attribute.htm	0
https://pro.arcgis.com/en/pro-app/tool-reference/data-management/copy-features.htm	1
https://pro.arcgis.com/en/pro-app/tool-reference/data-management/add-join.htm	2
https://pro.arcgis.com/en/pro-app/tool-reference/geoanalytics-desktop/join-features.htm	3
https://pro.arcgis.com/en/pro-app/tool-reference/data-management/select-layer-by-location.htm	4
https://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/point-density.htm	5
https://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/zonal-statistics-as-table.htm	6
https://pro.arcgis.com/en/pro-app/tool-reference/data-management/calculate-field.htm	7
https://pro.arcgis.com/en/pro-app/tool-reference/data-management/add-field.htm	8
https://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/euclidean-distance.htm	9
https://pro.arcgis.com/en/pro-app/tool-reference/analysis/spatial-join.htm	A
https://pro.arcgis.com/en/pro-app/tool-reference/analysis/clip.htm	B
https://pro.arcgis.com/en/pro-app/tool-reference/data-management/calculate-geometry-attributes.htm	C
https://pro.arcgis.com/en/pro-app/tool-reference/analysis/select.htm	D
https://pro.arcgis.com/en/pro-app/tool-reference/data-management/feature-to-point.htm	E
https://pro.arcgis.com/en/pro-app/tool-reference/analysis/union.htm	F
https://pro.arcgis.com/en/pro-app/tool-reference/data-management/dissolve.htm	G
https://pro.arcgis.com/en/pro-app/tool-reference/analysis/summarize-within.htm	H
https://pro.arcgis.com/en/pro-app/tool-reference/conversion/point-to-raster.htm	I
https://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/reclassify.htm	J
https://pro.arcgis.com/en/pro-app/tool-reference/data-management/join-field.htm	K
https://pro.arcgis.com/en/pro-app/help/analysis/image-analyst/overview-of-image-classification.htm	L
https://pro.arcgis.com/en/pro-app/tool-reference/geoanalytics-desktop/calculate-field.htm	M

4.2 Summary statistics

Each workflow has a unique Qid (question id), which is shown in logical order. Each workflow was translated into a string of letters. The Serial column shows this string, the so-called Serial. Summary statistics of the analyses calculated with the use of an Rscript are provided in table 2 below.

Table 2
Geo-analytical questions and their corresponding Qid, serialization & summary statistics

Question	Qid	Serial	Measure	Op-Count	Cycle-count
What is the percentage of people meeting the physical activity guideline per neighborhood in Amsterdam?	WF-1	0|1|2|	Percentage	3	3
What is the WOZ-waarde per square meter per neighborhood in Amsterdam?	WF-2	0|1|4|1|3|	WOZ-waarde	5	5
What is the number of sport facilities in a one kilometer radius per neighborhood in Amsterdam?	WF-3	05|1|4|1|6|2|	Number	7	6
What is the percentage of reported cases of covid-19, relative to amount of residents in the municipality, per municipality in the Netherlands?	WF-4	2|8|7|	Percentage	3	3
What is the amount of reported cases of covid-19 per municipality in the Netherlands?	WF-5	2|	Amount	1	1
What is the average distance to parks and green per PC4 area in Amsterdam?	WF-6	9|6|3|	Distance	3	3
What is the percentage of people with obesity per neighborhood in Amsterdam?	WF-7	0|1|2|	Percentage	3	3
What is the number of elderly people for each neighborhood in Amsterdam?	WF-8	0|1|4|1|A|	Number	5	5
What is the average urban heat island effect per PC4 area in Amsterdam?	WF-9	B|6|3|	Effect	3	3
What is the percentage of people suffering from a long-term illness or condition (physical health) per neighborhood in Amsterdam?	WF-10	0|1|2|	Percentage	3	3
What is the mean distance to a tram station per neighborhood in Amsterdam?	WF-11	09|16|4|1|3|	Distance	7	5
What is the average tree density per PC4 area in Amsterdam?	WF-12	A|A|A|A|8|C|8|	Density	7	7
What is the average distance to residential care complexes per PC4 area in Amsterdam?	WF-13	D|1|E|9|6|3|	Distance	6	6
What is the average energy label per PC4 area in Amsterdam?	WF-14	3|	Label	1	1
What is the average distance to hospitals per PC4 area in Amsterdam?	WF-15	D|1|E|9|6|3|	Distance	6	6
What is the average income per PC4 area in Amsterdam?	WF-16	F|G|	Income	2	2
What is the average year of construction per PC4 area in Amsterdam?	WF-17	H|	Year	1	1
What is the Safety Index per district in Amsterdam?	WF-18	0|1|4|1|3|	Index	5	5
What is the average green roof density per PC4 area in Amsterdam?	WF-19	H|8|8|	Density	3	3
What is the percentage of people smoking per neighborhood in Amsterdam?	WF-20	0|1|2|	Percentage	3	3
What is the percentage of people with a high risk of anxiety or depression (mental health) per neighborhood in Amsterdam?	WF-21	0|1|2|	Percentage	3	3
What is the variety of sport facilities per neighborhood in Amsterdam?	WF-22	0I|1|4|1|6|2|	Variety	7	6
What is the mean distance to a subway station per neighborhood in Amsterdam?	WF-23	09|1|4|1|6|	Distance	6	5
What is the percentage of population between 16 and 24 years per neighborhood in Amsterdam?	WF-24	0|1|4|1|A|	Percentage	5	5
What is the average wall plant density per PC4 area in Amsterdam?	WF-25	H|8|8|	Density	3	3
What is the average percentage of people experiencing severe loneliness in the PC4 areas in Amsterdam?	WF-26	F|G|	Percentage	2	2
What is the average percentage of water area in the PC4 areas in Amsterdam?	WF-27	L|J|6|K|8|M|	Percentage	6	6

The table shows several columns, which need some further explanation. *Qid* is the workflow/question id, linking an id to each one of the 27 geo-analytical questions. *Serial* refers to the serialization of a corresponding workflow, annotating cycles too, by displaying the serial as 1|2|3|4| instead of 1234 for example. The serials are the data used for further analysis. *Measure* corresponds to the measure used in the question sentence. This will be used in both the tool-measure co-occurrence matrix and TF-IDF analyses. *OpCount* is the number of operations in the corresponding workflow, whereas *cycleCount* is the number of cycles in the corresponding workflow. For only four out of twenty-seven workflows the *cycleCount* differs

from the *OpCount* (WF-3, WF-11, WF-22, WF-23) , so most of the cycles contain a single operation.

4.3 Results analyses

In the following sections, the results of the analyses are provided. Several analyses were performed in order to help answer the research questions posed in the introduction. First, the dissimilarity matrix was calculated, in order to provide data for the hierarchical clustering that followed. Second, a N-gram analysis was done, in order to show the most frequent operation segments and patterns in used ArcGIS Pro tools. Third, a tool-measure co-occurrence matrix was calculated, to compare pairs of measures and tools. Finally, TF-IDF analysis was performed, in order to explore the specificness of a certain tool to a measure. This section will provide raw results, with some further explanation. The results and answers to the research questions will be explained and described in further detail in section 5; discussion and conclusions.

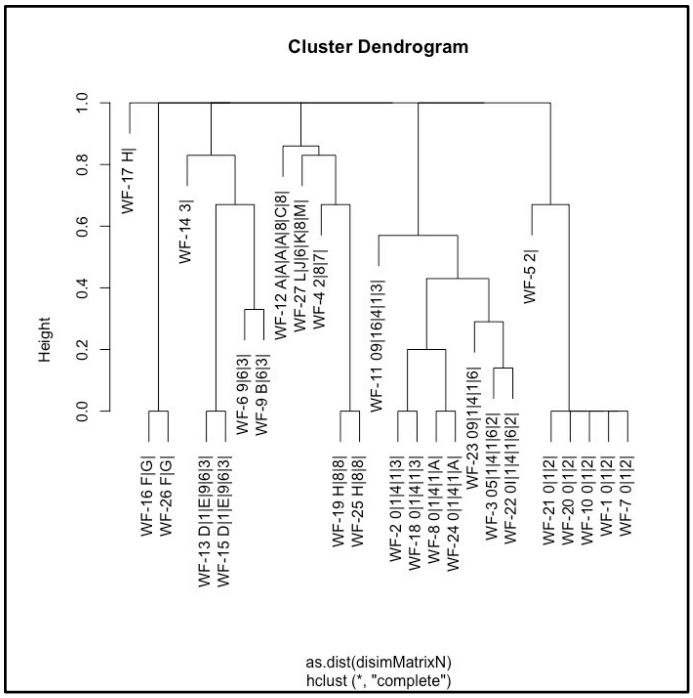
4.3.1 Workflow dissimilarity matrix & hierarchical clustering

The results of the Levenshtein's distance analysis are displayed in Appendix A; table 5. The numbers show how many single-character edits are required to change one string of ASCII characters into the other. The higher this number, the more different the two strings are. The interpretation of the results does not directly give an answer on the similarity between two strings, because it does not show how many characters each string contains. To resolve the issue of comparing strings with different lengths, the analysis was performed a second time, but with normalized dissimilarity values. The normalization was done based on the length of the longest string in the pair. Due to this normalization, the comparability of the results went up, and thus the results became better interpretable (Appendix B; table 6). Hierarchical clustering was then applied on the dissimilarity matrix of the workflows.

The tree, or so-called dendrogram, is a compact way of visualizing the dissimilarity matrix (Figure 7). Interpretation of the data is made easier this way. The node names are written as follows: "[workflow id] [char string]". For example; in "WF-21 0|1|2|", 'WF-21' is the workflow id and 0|1|2| is the corresponding string serialization

Figure 7

Hierarchical clustering on the workflows based on the dissimilarity matrices



The hierarchical clustering results shows which workflows are similar. This applies to WF-16 and WF-26 (F|G|), WF-13 and WF-15 (D|1|E|9|6|3|), WF-19 and WF-25 (H|8|8|), WF-2 and WF-18 (0|1|4|1|3|), WF-8 and WF-24 (0|1|4|1|A|), and WF-21, WF-20, WF-10, WF-1 and WF-7 (0|1|2|). An overview of these workflows is given in table 3, together with the corresponding geo-analytical questions.

Table 3

Results hierarchical clustering with corresponding questions (dissimilarity 0.0)

Workflow id	Serialization	Corresponding geo-analytical question
WF-16	F|G|	What is the average income per PC4 area in Amsterdam?
WF-26		What is the average percentage of people experiencing severe loneliness in the PC4 areas in Amsterdam?
WF-13	D|1|E|9|6|3|	What is the average distance to residential care complexes per PC4 area in Amsterdam?
WF-15		What is the average distance to hospitals per PC4 area in Amsterdam?
WF-19	H|8|8|	What is the average green roof density per PC4 area in Amsterdam?
WF-25		What is the average wall plant density per PC4 area in Amsterdam?
WF-2	0|1|4|1|3|	What is the WOZ-waarde per square meter per neighborhood in Amsterdam?
WF-18		What is the Safety Index per district in Amsterdam?
WF-8	0|1|4|1|A|	What is the number of elderly people for each neighborhood in Amsterdam?
WF-24		What is the percentage of population between 16 and 24 years per neighborhood in Amsterdam?
WF-21	0|1|2|	What is the percentage of people with a high risk of anxiety or depression (mental health) per neighborhood in Amsterdam?
WF-20		What is the percentage of people smoking per neighborhood in Amsterdam?
WF-10		What is the percentage of people suffering from a long-term illness or condition (physical health) per neighborhood in Amsterdam?
WF-1		What is the percentage of people meeting the physical activity guideline per neighborhood in Amsterdam?
WF-7		What is the percentage of people with obesity per neighborhood in Amsterdam?

The questions corresponding to the workflows with a dissimilarity of 0.0 are merely the same as well, which would be expected based on the similar structure of the questions. For example; the questions of WF-19 and WF-25 are similar, apart from the words in the middle: 'green roof' and 'wall plant', so it is expectable these geo-analytical questions are answered with the same workflow. However, similar geo-analytical questions can result in different workflows. For example; WF-6 and both WF-13 and WF-15 calculate average distances, but the workflows vary. This is probably due to the differences in the available data. All three workflows end in the same operation segment (9|6|3|), but the input data used for WF-6 needed some changes before it could be used. That caused the final workflows to differ from the other two with the same measure. Several clusters can be formed; a 'distance' cluster, 'percentage' cluster, 'density' cluster and a 'population' cluster. In the conclusion and discussion, this aspect will be discussed in further detail and some possible explanations will be provided why these clusters were formed.

4.3.2 N-gram analysis

Table 4 provides an overview of the most frequent operation segments (N-grams). Redundant items were removed. For example, the N-gram of both 1|4| and 1|4|1| were calculated, but only the longest operation segment is displayed in table 4. The left column shows the N-gram and the right column shows the N-gram's frequency among all workflows. The most common pattern repeating in workflows is 0|1| (select-layer-by-attribute, copy-features), followed by 4|1| (select-layer-by-location, copy-features). The two most common patterns were used together as well, repeating four times in total (0|1|4|1) and it is the part of a workflow which calculates the geodata for 'each neighborhood in Amsterdam', just like 1|4|1|. This pattern occurred in five workflows with the same geodata as input, but with varying measures in the corresponding questions. Thus, that pattern is dependent on the input used in the workflow, and not specific to the measure in the questions. The pattern 0|1|2| (select-layer-by-attribute, copy-features, add-join) repeated 5 times in total, all in the workflows with a dissimilarity of 0.0 (WF-21, WF-20, WF-10, WF-1 and WF-7) and the measure 'percentage'. This shows that some single functionality can be found among more tools, instead of one specific one. Tools such as select-layer-by-attribute and copy-features are software specific for ArcGIS Pro and that helps explain why they can be found in several different workflows. The pattern 9|6|3| (Euclidian-distance, zonal-statistics-as-table, join-features) repeated three times among all workflows. Those three

workflows answer questions with the same measure; ‘distance’. So, this pattern points towards a functionality needed in those workflows and might be measure specific, instead of software specific or input specific.

Table 4
N-gram analysis of workflow serialization

N-gram	Frequency
0|1|	9
4|1|	8
1|4|1|	7
0|1|4|1|	4
0|1|2|	5
6|3|	4
1|4|1|6|	3
4|1|3|	3
4|1|6|	3
9|6|3|	3

4.3.3 Tool-measure co-occurrence matrix

Another exploratory analysis which was performed was the co-occurrence matrix (Appendix C; table 7). A co-occurrence matrix normalized by the total number of workflows per measure was performed as well (Appendix D; table 8). By normalizing the values, the value of co-occurrence of measure *r* and tool *s* is divided by the sum of the co-occurrence values of all tools co-occurring with measure *r*. This makes the results more reliable and more comparable. To improve comparability and interpretation of the results, plots of the co-occurrence matrices are provided (figure 8 and 9). The plots are a compact visualization of the co-occurrence matrices. The lighter the blue in the plot, the more times tool *s* was used with measure *r*.

As was concluded in the previous section, the measure ‘percentage’ can be answered with varying workflows, and co-occurrence matrix proves that the measure ‘percentage’ can be answered with varying tools as well. The most common tools used to answer a ‘percentage’ question are select-layer-by-attribute, copy-features, add-join, but this has more to do with the ‘neighborhood in Amsterdam’ part than the actual measure (‘percentage’). This pattern (0|1|4|1|) co-occurs with several tools, and that shows again that pattern is not specific to a measure, but rather to the software and input. The measures ‘amount’ and ‘year’ only occur together with one tool, so the normalized values are not reliable. An interesting result is the tools used to answer a question with the measure ‘distance’. The matrices show that the tools

zonal-statistics-as-table and euclidean-distance are the most common tools in combination with this measure, as could be seen in the previous analysis too. The co-occurrence values show that these tools indeed occur more than once with the measure ‘distance’ and that they do not occur often with other measures.

Figure 8

Tool-measure co-occurrence

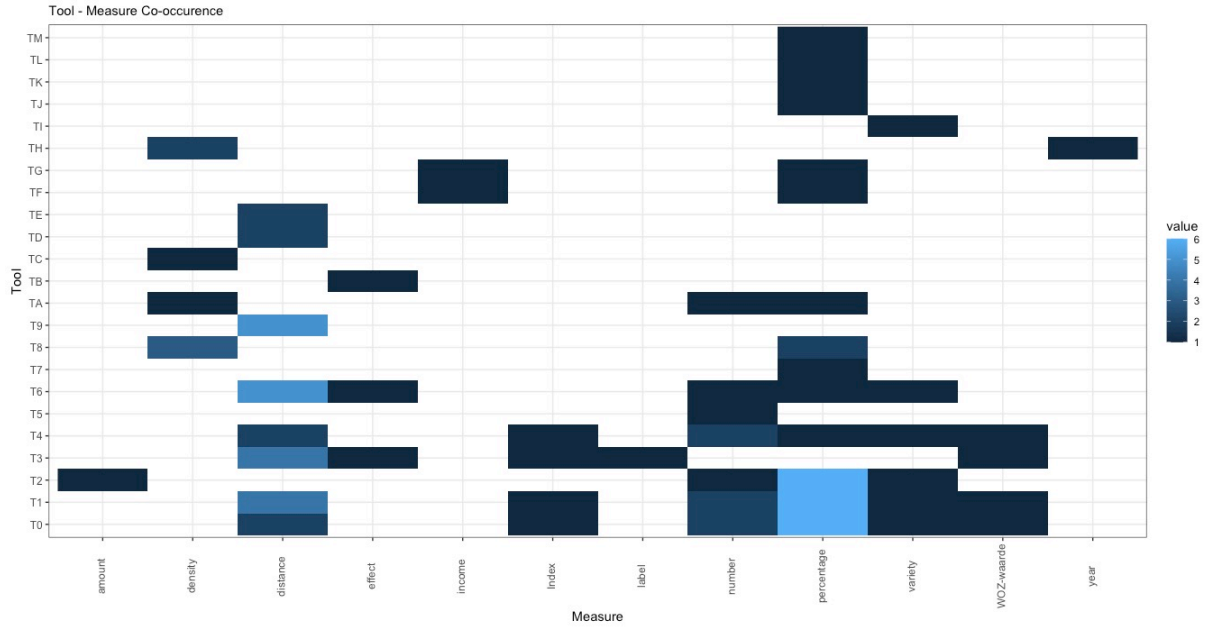
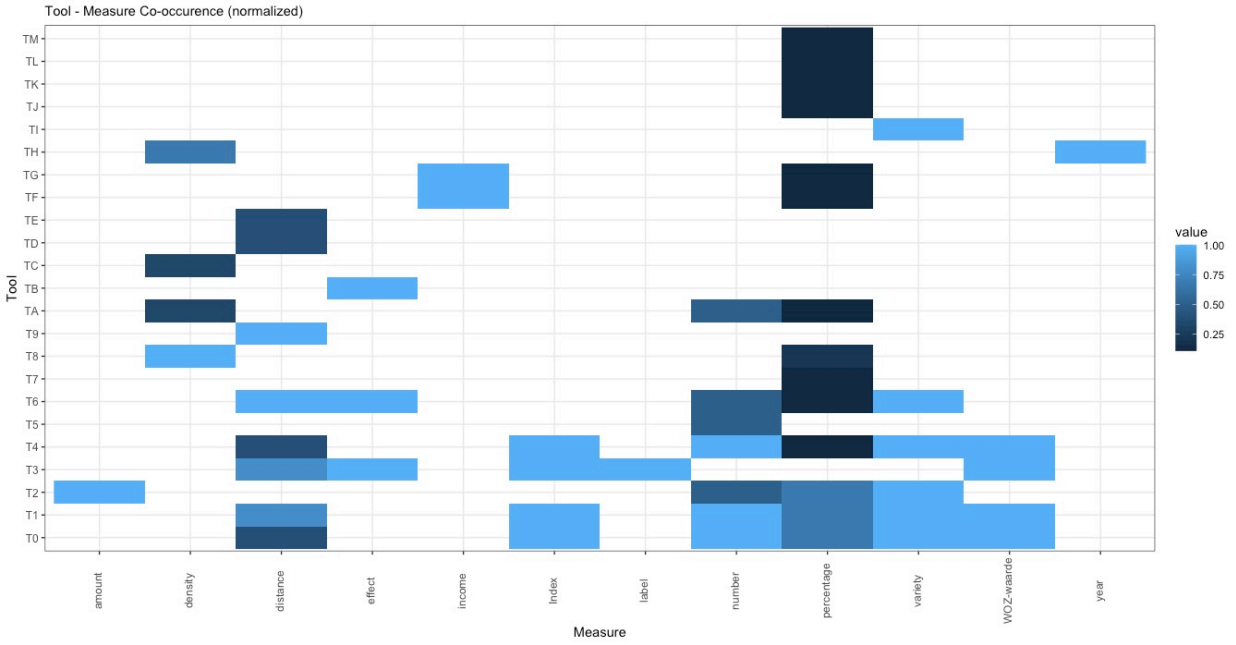


Figure 9

Tool-measure co-occurrence (normalized)

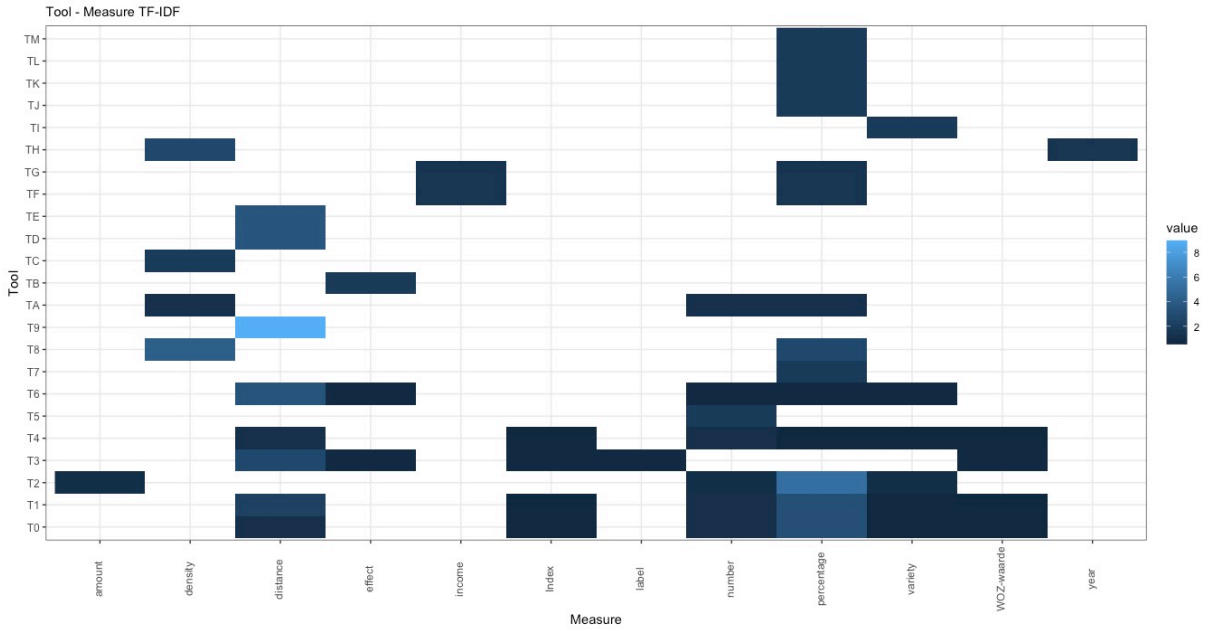


4.3.4 TF-IDF

The higher the TF-IDF value, the more specific tool *s* is to the measure *r*, relative to the other measures. The lighter the blue in the plot, the more specific tool *s* is to measure *r*. As shown in the plot (figure 10), the tool euclidean-distance is quite specific to the measure 'distance', as was suspected in the previous sections. As for the measure 'density', the tools add-field, spatial-join and calculate-geometry-attributes, the co-occurrence was relatively high, but according to the TF-IDF matrix, the tools are not that specific to the tool.

Figure 10

Tool-measure TF-IDF



The same applies to the measures 'index', 'number', 'variety' and 'WOZ-waarde'. The co-occurrence (normalized) of these measures with the tools select-layer-by-attribute, copy-features, add-join, join-features and select-layer-by-location were relatively high, but the TF-IDF values show that the tools are not very specific to those measures. So, the previous findings are validated again; these tools calculate data not specific to the corresponding measure, but rather specific to the input data or the software.

An explanation and the discussion of the results will be provided in the next section.

5 Discussion and conclusions

This study sought to explore how Human Geographers could ask spatial questions in a systematic manner and automatically retrieve the resources necessary to answer them. The data used for this research was collected by creating scenarios and workflows to answer the posed geo-analytical questions. A part of the scenarios was selected based on students' reports. In total, the analyses were based on twenty-seven workflows, of which nineteen were derived from students' reports and eight were self-created workflows. Both abstract and concrete workflows were created. In order to be able to perform analyses on the data, a serialization was performed, to transform the workflows into strings of ASCII characters. After that, several ADE were done; Levenshtein's distance analysis, hierarchical clustering, N-gram analysis, tool-measure co-occurrence and TF-IDF.

To answer the research questions; "Which types of GIS workflows answer which kinds of geographic questions?", "Which part of a question is answered by which part of a GIS workflow?", "Do similar workflows answer similar questions?", a small note has to be made. Since this thesis is conducts a relatively small, exploratory research, it was not able to completely capture the potential of this field of research. It does form an interesting basis to perform future analyses on. Thus, important to note is that the results found in this research are only a small part of the solution to the research questions, and the solution is only the beginning in this field of research.

In total, twenty-three tools were used to answer the twenty-seven geo-analytical questions, containing twelve measures. All questions could be categorized as 'what' questions and they were all in line with the Human Geography scenario: the livability of residents in Amsterdam. In order to answer the research questions, some methods of analysis were explored. All analyses were of value to the overall research. Especially when combining the several analyses, the research questions can be partly answered. Like mentioned before, this research only focused on one scenario and therefore, the scope is limited. However, the answers to the research questions in this scenario may provide a basis for future and more deliberate research. In the next section, the research questions will be answered by combining the obtained results and literature.

5.1 Research questions answers

5.1.1 Question 1

"Which types of GIS workflows answer which kinds of geographic questions?"

Based on the dissimilarity matrices and the hierarchical clustering, several clusters were formed. The clusters were formed based on the (dis)similarities between the several strings of ASCII characters, corresponding to a workflow. The strings within those clusters were similar, and the corresponding workflows were compared. The workflows in the same cluster were used to answer questions with the same measure in the corresponding question. Therefore, four clusters based on the hierarchical clustering and the measures could be formed: a 'distance' cluster, 'percentage' cluster, 'density' cluster and a 'population' cluster (based on a dissimilarity of 0.0). The type of workflows corresponding to those clusters were as follows:

- distance cluster*: select* | *copy-features* | *feature-to-point* | Euclidean-distance | zonal-statistics-as-table | join-features | (D|1|E|9|6|3|)
- percentage cluster: *select-layer-by-attribute* | *copy-features* | *add-join* | (0|1|2|)
- density cluster*: summarize-within* | *add-field* | *add-field* | (H|8|8|)
- population cluster: *select-layer-by-attribute* | *copy-features* | *select-layer-by-location* | *copy-features* | *spatial-join* | (0|1|4|1|A|)

As mentioned in section 4, the mere part of these tools is either software specific or input specific (highlighted in *Italic*). For example, derived from the N-grams, the pattern 0|1|4|1| occurred four times among all workflows, but this pattern is software specific, and not measure specific. The 'density' clusters contain the same pattern, but another question containing the measure 'density' is answered with a different workflow (WF_12), so the tools are again input specific, rather than measure specific. Two other clusters were formed, but the measure in the corresponding workflows were not the same: income & percentage and WOZ-waarde & safety index. In order to make them the same, the questions would have had to be rephrased (QuAnGIS, n.d.), but because the questions were derived from students' reports, this was not possible. They did have the same geodata as input, so these patterns could again be explained by a input specificity. The only possible measure specific pattern occurred in the 'distance' cluster: D|1|E9|6|3|. The pattern 9|6|3| occurred in one other workflow, which brings the total of workflows with this pattern to three (WF_6, WF_13, WF_15). These workflows correspond to geo-analytical questions containing the measure 'distance'. This was proved by the co-occurrence matrix and TF-IDF as well. The tools Euclidean-distance and zonal-statistics-as-

table proved to be specific to the measure ‘distance’. To conclude, the 9|6|3| pattern of a workflow is a specific answer of a ‘distance’ kind of geo-analytical question.

5.1.2 Question 2

"Which parts of a question is answered by which parts of a GIS workflow?"

In the context of this scenario, several parts of a question can be distinguished: measure, support and extent (QuAnGIS, n.d.). By combining the results from the several analyses, it can be concluded that these different parts can be answered with different GIS workflows.

Measure parts of a question seem to be answered with specific workflows. One finding was the measure ‘distance’, which was answered with a specific tool pattern (9|6|3|). No other measure specific patterns were found.

Support parts of a question were answered by a big variety of workflows. The workflows seemed to be subject to the availability of geodata and resources. This caused the workflows to vary and the workflows were dependent on the availability and relevance of the input. For example, the ‘percentage’ cluster consists of five workflows, corresponding to five geo-analytical questions, all containing the same measure (‘percentage’) and extent (‘in Amsterdam’). However, the support part differs. So, no conclusions can be drawn based on the support part of a question.

Extent parts of a question were answered by similar workflows. The pattern 0|1|4|1| was used to answer the specific part ‘per neighborhood in Amsterdam’. However, in the questions containing ‘per PC4 area in Amsterdam’, the input was readily available, and no tools were needed to transform that geodata. So, the extent part of a question seems to be dependent on the relevance of the input data used for the extent part. If the input data is readily available, no part of a GIS workflow is needed to answer that part. However, if it necessary to transform the extent input data first, it is mostly done with the same pattern of tools in ArcGIS Pro. In the context of this research, this means that part of a workflows is either answered by a pattern of 0|1|4|1|, or a chunk of that pattern. So, the answer, in the form of a workflow, to an extent part of a question seems to be software specific.

5.1.3 Question 3

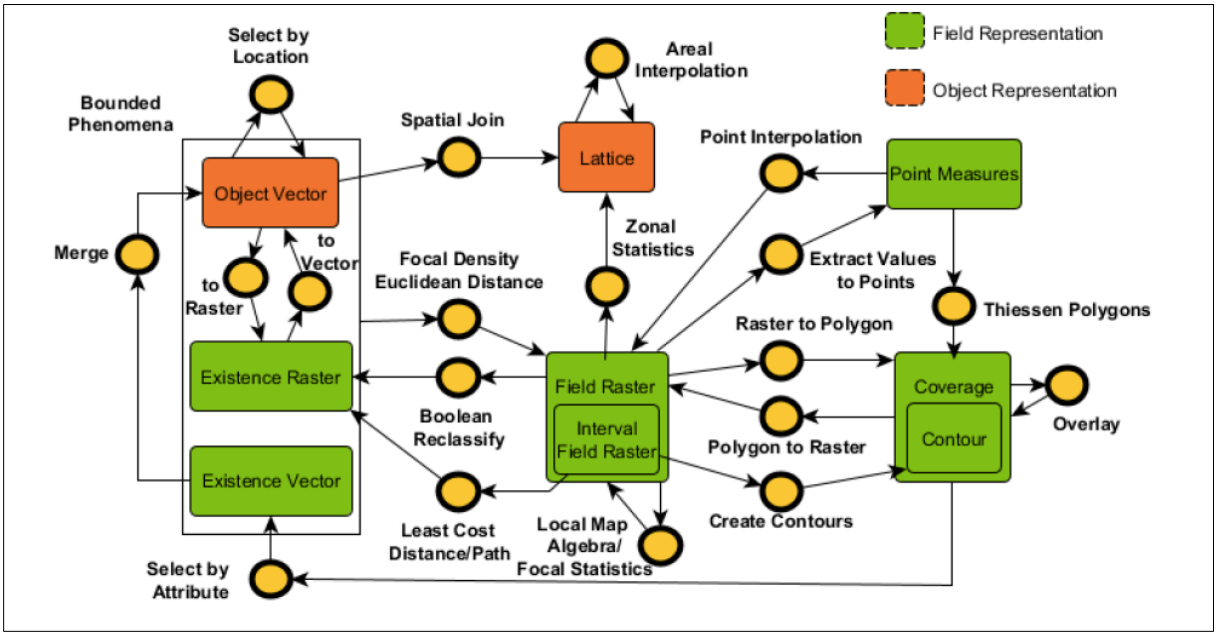
“Do similar workflows answer similar questions?”

For the most part, similar workflows do answer similar questions. However, this is not applicable to all workflows and questions. Several clusters of workflows were formed, and as

shown in the previous sections (5.1.1 and 5.1.2), they corresponded to questions with some similarity too, based on the measure, support or extent part of a question. Measure, input and software specificities were used to explain some of these similarities. The CCDT ontology (Scheider, 2016) may help in explaining the measure and input specificities. As can be seen in figure 11, several geo-analytical operations are applicable to several abstract data types. By following the computational diagram, several patterns can be discovered. This helps explain the specific pattern used to answer the questions containing the measure 'distance' (Euclidean-distance | zonal statistics).

Figure 11

Scheider et al. (2016, pp. 18): "Computational diagram of types of geoanalytical operations applicable to abstract data types."



So, by combining the results of this research and the literature, common patterns in geo-analytical questions and workflows answers in parallel were found, in the scenario of 'livability and health of residents in Amsterdam'. The explored methods of analysis were proved to be useful.

6 Limitations, directions for future research & implications

6.1 Limitations and directions for future research

This research has several limitations. First, the research was based on a small sample. Creating scenarios is a painstaking process and creating the corresponding workflows takes time. The limited timeframe for this thesis is the reason the sample consists of 27 workflows, and not more than that.

Second, the research was very technical. Specific knowledge was needed in order to perform the analyses. It might take time and patience to become familiar with these terms and several methods. This causes this research to be in a preliminary stage, since the timeframe did not allow the researcher to become fully equipped with the right knowledge in order to grasp the full potential of this research. Therefore, future research should base the analyses on this exploratory research. Future research could add more scenarios and more technical and complex analyses.

Third, some processes had some technical issues. The program 'RDF primer' could not be used, so 'Sublime text 3' had to be used instead. Sublime text 3 is program which cannot detect errors in the code itself, so a Turtle Validator had to be used as well. This caused some unnecessary steps in order to get the right and correct workflows. Optimally, future research could streamline this specific methodology, in order to get more accurate data, faster. This will smoothen the data collection.

Fourth, a substantial part of the scenarios and corresponding workflows were based on reports written by students and the scope of their research was Amsterdam. This makes that the scope of this research was Amsterdam too, consequently. That offers a limited scope, and future research could try to focus on not only Amsterdam, but more regions too. Different regions might cause other geo-analytical questions to pop up, leading to more diverse questions and workflows. This could improve the generalizability of the results and should add to the overall value of the research.

Fifth, the results found in this research were subject to the selection of certain databases, and the correctness of the questions and workflows. For example, several workflows calculated average distances, but were answered with varying workflows. This might be due to the unavailability of relevant datasets. Some datasets had to be transformed first, and others were ready to use. It might also be due to the overall correctness of the questions and workflows. The assumption was made that the workflows were corrected by the course instructors. Future

research could base their analyses on self-created workflows, to assure quality and relevance. A geo-analytical question can be answered with varying workflows. Not all options were explored, but they could and should be in future research. The workflows used to calculate one specific measure should be consistent and all possible workflows to answer one specific question should be explored. Varying questions could be rephrased in terms of their measure, to be able to compare corresponding workflows better. Future research could focus more on the selection of the scenarios and the varying workflows to answer a geo-analytical question, instead of the analyses.

Sixth, the access to people and data was compromised due to the corona virus. Utrecht University was closed during the period of writing this thesis. Everyone was forced to work at home, instead of working together at the University. This slowed down the pace of communicating and made it less easy to contact one another. The technical problems could have been solved faster, by visiting the University or have someone to take a look at it. However, this was not allowed, and the issues needed to be solved otherwise. For this research, MacOS was used, but future research should rather another device, such as Microsoft. MacOS did not support certain programs, and Microsoft would have. Future research could thus make better use of resources, especially when the University reopens. By making use of the optimal resources and circumstances, the research could be performed more accurately and smoothly.

Finally, future research could include more data and analyses. More scenarios could be added, and other analyses could be performed and included as well, such as regression analysis. The TF-IDF vectors can be used as input to the clustering as well. Other distance metrics could be used for the hierarchical clustering, since the distance metric was not stated in the Rscript. In future research, both Euclidean distance and cosine-similarity could be used as distance metrics. Another option to extend this research would be to incorporate more other data derived from the workflows. Future research could first decompose questions into measure, support and extent phrases. This might allow researchers to identify and quantify analytical goals. This way, the structure of geo-analytical questions can be analyzed in greater depth.

6.2 Theoretical and practical implications

This study extends the research on geo-analytical question-answering. First, an exploratory research was done on which types of spatial concepts were used in geo-analytical questions. This contributes to a better and more broad understanding of the spatial concepts and geodata. Second, the research explored what the kinds of answers could look like (which GIS tools).

This expands the literature by providing more answers to geo-analytical questions. This study also shed light on which data sources are relevant and that it matters which data sources are used. Overall, this study contributes to the literature by helping prepare an empirical basis for a Semantic Web based question-answering system and is a starting point in the QuAnGIS project.

In order to not only highlight the theoretical implications, a practical implication will be provided as well. By providing the basis for the Semantic Web based question-answering system, Human Geographers will eventually be able to ask spatial questions in a systematic manner and automatically retrieve the resources necessary to answer them. Eventually, the goal is to create a basis for people outside the field too. By providing the first steps in creating this basis, this thesis contributes to the quest of developing QA systems which will be able to automatically answer questions such as; “How much is Romay exposed to green space while biking from Amsterdam to Utrecht University?”.

7 References

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), 45-65.

Allaire, J. (2012). RStudio: integrated development environment for R. *Boston, MA*, *770*, 394.

Antoniou, G., & Van Harmelen, F. (2004). *A semantic web primer*. MIT press.

Bennett, B., Mallenby, D. & Third, A. (2008), An ontology for grounding vague geographic terms., in *FOIS,* Vol. 183, pp. 280–293.

Burrough, P. A., McDonnell, R. A., & Lloyd, C. D. (2015). *Principles of geographical information systems* (third edition). Oxford university press.

Canbek, N.G., Mutlu, M.E. (2016): On the track of artificial intelligence: Learning with intelligent personal assistants. *Journal of Human Sciences 13(1), 592{601*

Chen, W., Fosler-Lussier, E., Xiao, N., Raje, S., Ramnath, R., & Sui, D. (2013). A synergistic framework for geographic question answering. In *2013 IEEE Seventh International Conference on Semantic Computing* (pp. 94-99). IEEE.

Chiarcos, C., & Fäth, C. (2017). CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *International Conference on Language, Data and Knowledge* (pp. 74-88). Springer, Cham.

Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., & Horrocks, I. (2000). The semantic web: The roles of XML and RDF. *IEEE Internet computing*, *4*(5), 63-73.

[QuAnGIS: Deconstruction of geo-analytic questions in terms of measures, supports, and spatio-temporal extents]. (n.d.).

ESRI. (2020a). *About ArcGIS Pro.* Retrieved from https://pro.arcgis.com/en/pro-app/get-started/get-started.htm

ESRI. (2020b). *GIS dictionary*. Retrieved from https://support.esri.com/en/other-resources/gis-dictionary/term/5500cd6a-84a7-4b2b-9705-83bab17604f0

ESRI. (2020c). *ModelBuilder vocabulary.* Retrieved from https://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/modelbuilder/modelbuilder-vocabulary.htm

ESRI. (2020d). *What is ModelBuilder*? Retrieved from https://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/modelbuilder/what-is-modelbuilder-.htm

Gao, S., & Goodchild, M. F. (2013, July). Asking spatial questions to identify GIS functionality. *In 2013 Fourth International Conference on Computing for Geospatial Research and Application* (pp. 106-110). IEEE.

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., ... & Myers, J. (2007). Examining the challenges of scientific workflows. *Computer,* 40(12), 24-32.

Golledge, R. G. (1995). Primitives of spatial knowledge. In T. L. Nyerges, D. M. Mark, R. Laurini, & M. J. Egenhofer (Eds.), *Cognitive aspects of human-computer interaction for geographic information systems* (pp. 29–44). Springer.

Gousios, G., Vasilescu, B., Serebrenik, A., & Zaidman, A. (2014). Lean GHTorrent: GitHub data on demand. In *Proceedings of the 11th working conference on mining software repositories* (pp. 384-387).

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-based nursing*, *18*(3), 66-67.

Huang, B., Jiang, B., & Li, H. (2001). An integration of GIS, virtual reality and the Internet for visualization, analysis and exploration of spatial data. *International Journal of Geographical Information Science*, *15*(5), 439-456.

IDLab Turtle Validator. (n.d.). Retrieved from http://ttl.summerofcode.be

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241-254.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, *32*(8), 68-75.

Kruiger, Kasalica, Meerlo, Lamprecht & Scheider (n.d.). Loose programming of GIS workflows with geo-analytical concepts.

Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science* 26, 12, 2267–2276.

[Lassila & Swick, 1999] O. Lassila, Ralph Swick (eds).: Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999, http://www.w3.org/TR/REC-rdf-syntax/

Laurent, D., Séguéla, P. & Nègre, S. (2006), QA better than IR? in *Proceedings of the Workshop on Multilingual Question Answering, Association for Computational Linguistics,* pp. 1–8.

Lin, J. J. (2002). "The Web as a Resource for Question Answering: Perspectives and Challenges." In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Canary Islands, Spain, 1–8.

Lutz, M. (2001). *Programming python*. " O'Reilly Media, Inc.".

Mai, G., Yan, B., Janowicz, K., Zhu, R. (2019): Relaxing unanswerable geographic questions

using a spatially explicit knowledge graph embedding model. In: *The Annual International Conference on Geographic Information Science. pp. 21{39. Springer International Publishing,* Cham.

Pacione, M. (2009). *Urban geography: A global perspective*. Routledge.

Polit, D. F., & Beck, C. T. (2010). Generalization in quantitative and qualitative research: Myths and strategies. *International journal of nursing studies*, *47*(11), 1451-1458.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 133-142).

Rauschert, I., Agrawal, P., Sharma, R., Fuhrmann, S., Brewer, I., & MacEachren, A. (2002). Designing a human-centered, multimodal GIS interface to support emergency management. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems* (pp. 119-124).

Scheider, S., Ballatore, A., & Lemmens, R. (2019). Finding and sharing GIS methods based on the questions they answer. *International journal of digital earth*, 12(5), 594-613.

Scheider, S., Meerlo, R., Kasalica, V., & Lamprecht, A. L. (2016.). Ontology of core concept data types for answering geo-analytical questions. *Journal of Spatial Information Science*.

Scheider, S., Nyamsuren, E., Kruiger, H., & Xu, H. (2020). Geo-analytical question-answering with GIS. *International Journal of Digital Earth*, 1-14.

Scheider, S., Ostermann, F. O., & Adams, B. (2017). "Why Good Data Analysts Need to Be Critical Synthesists. Determining the Role of Semantics in Data Analysis." *Future Generation Computer Systems* 72: 11–22.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2012). Syntactic dependency-based n-grams as classification features. In *Mexican International Conference on Artificial Intelligence* (pp. 1-11). Springer, Berlin, Heidelberg.

Southworth, M. (2016). Learning to make liveable cities. *Journal of Urban Design*, *21*(5), 570-573.

Stanford. (n.d.). *Hierarchical cluster analysis.* Retrieved from http://84.89.132.1/~michael/stanford/maeb7.pdf

Universiteit Utrecht. (2018, August 16). *Een Alexa voor geografische analyses.* Retrieved from https://www.uu.nl/nieuws/een-alexa-voor-geografische-analyses.

Vrije Universiteit. (2020). *Course: Geographic Analysis and Visualization.* Retrieved from https://studiegids.vu.nl/nl/Minor/2019-2020/national-geo-information/AB_1107

W3C. (2014). *RDF 1.1 Primer*. Retrieved from https://www.w3.org/TR/rdf11-primer/#section-triple

W3C. (2015a). *Linked data.* Retrieved from https://www.w3.org/standards/semanticweb/data

W3C. (2015b). *Vocabularies*. Retrieved from https://www.w3.org/standards/semanticweb/ontology

Wang, Y., Ge, L., Rizos, C., & Babu, R. (2004). Spatial data sharing on grid. *Geomatics Research*.

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, *26*(3), 1-37.

Xu, H., Hamzei, Ehsan, Nyamsuren, E., Winter, Stephan, Tomko, Martin & Scheider, S. (2020). Extracting interrogative intents and concepts from geo-analytic questions. *Proceedings of the 23rd AGILE conference on Geographic Information Science Springer.*

Yin, Z., Zhang, C., Goldberg, D. W., & Prasad, S. (2019). An NLP-based question answering framework for spatio-temporal analysis and visualization. In *Proceedings of the 2019 2nd International Conference on Geoinformatics and Data Analysis* (pp. 61-65).

Yujian, L., & Bo, L. (2007). A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, *29*(6), 1091-1095.

8 Appendix

8.1 Appendix A

Table 5

Levenshtein's distance analysis; workflow dissimilarity matrix

WF	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	0	3	4	3	2	3	0	3	3	0	5	7	5	3	5	3	3	3	3	0	0	4	4	3	3	3	6
2	3	0	3	5	5	4	3	1	4	3	2	7	4	4	4	5	5	0	5	3	3	3	2	1	5	5	6
3	4	3	0	7	6	6	4	3	6	4	4	7	5	7	5	7	7	3	7	4	4	1	2	3	7	7	7
4	3	5	7	0	2	3	3	5	3	3	7	6	6	3	6	3	3	5	2	3	3	7	6	5	2	3	5
5	2	5	6	2	0	3	2	5	3	2	7	7	6	1	6	2	1	5	3	2	2	6	6	5	3	2	6
6	3	4	6	3	3	0	3	5	1	3	4	7	3	2	3	3	3	4	3	3	3	6	5	5	3	3	5
7	0	3	4	3	2	3	0	3	3	0	5	7	5	3	5	3	3	3	3	0	0	4	4	3	3	3	6
8	3	1	3	5	5	5	3	0	5	3	3	7	5	5	5	5	5	1	5	3	3	3	2	0	5	5	6
9	3	4	6	3	3	1	3	5	0	3	5	7	4	2	4	3	3	4	3	3	3	6	6	5	3	3	5
10	0	3	4	3	2	3	0	3	3	0	5	7	5	3	5	3	3	3	3	0	0	4	4	3	3	3	6
11	5	2	4	7	7	4	5	3	5	5	0	7	5	6	5	7	7	2	7	5	5	4	2	3	7	7	6
12	7	7	7	6	7	7	7	7	7	7	7	0	7	7	7	7	7	7	5	7	7	7	7	7	5	7	6
13	5	4	5	6	6	3	5	5	4	5	5	7	0	5	0	6	6	4	6	5	5	5	5	5	6	6	6
14	3	4	7	3	1	2	3	5	2	3	6	7	5	0	5	2	1	4	3	3	3	7	6	5	3	2	6
15	5	4	5	6	6	3	5	5	4	5	5	7	0	5	0	6	6	4	6	5	5	5	5	5	6	6	6
16	3	5	7	3	2	3	3	5	3	3	7	7	6	2	6	0	2	5	3	3	3	7	6	5	3	0	6
17	3	5	7	3	1	3	3	5	3	3	7	7	6	1	6	2	0	5	2	3	3	7	6	5	2	2	6
18	3	0	3	5	5	4	3	1	4	3	2	7	4	4	4	5	5	0	5	3	3	3	2	1	5	5	6
19	3	5	7	2	3	3	3	5	3	3	7	5	6	3	6	3	2	5	0	3	3	7	6	5	0	3	5
20	0	3	4	3	2	3	0	3	3	0	5	7	5	3	5	3	3	3	3	0	0	4	4	3	3	3	6
21	0	3	4	3	2	3	0	3	3	0	5	7	5	3	5	3	3	3	3	0	0	4	4	3	3	3	6
22	4	3	1	7	6	6	4	3	6	4	4	7	5	7	5	7	7	3	7	4	4	0	2	3	7	7	7
23	4	2	2	6	6	5	4	2	6	4	2	7	5	6	5	6	6	2	6	4	4	2	0	2	6	6	6
24	3	1	3	5	5	5	3	0	5	3	3	7	5	5	5	5	5	1	5	3	3	3	2	0	5	5	6
25	3	5	7	2	3	3	3	5	3	3	7	5	6	3	6	3	2	5	0	3	3	7	6	5	0	3	5
26	3	5	7	3	2	3	3	5	3	3	7	7	6	2	6	0	2	5	3	3	3	7	6	5	3	0	6
27	6	6	7	5	6	5	6	6	5	6	6	6	6	6	6	6	6	6	5	6	6	7	6	6	5	6	0

8.2 Appendix B

Table 6

Levenshtein's distance analysis; workflow dissimilarity matrix (normalized)

WF	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	0	0.6	0.57	1	0.67	1	0	0.6	1	0	0.71	1	0.83	1	0.83	1	1	0.6	1	0	0	0.57	0.67	0.6	1	1	1
2	0.6	0	0.43	1	1	0.8	0.6	0.2	0.8	0.6	0.29	1	0.67	0.8	0.67	1	1	0	1	0.6	0.6	0.43	0.33	0.2	1	1	1
3	0.57	0.43	0	1	0.86	0.86	0.57	0.43	0.86	0.57	0.57	1	0.71	1	0.71	1	1	0.43	1	0.57	0.57	0.14	0.29	0.43	1	1	1
4	1	1	1	0	0.67	1	1	1	1	1	1	0.86	1	1	1	1	1	1	0.67	1	1	1	1	1	0.67	1	0.83
5	0.67	1	0.86	0.67	0	1	0.67	1	1	0.67	1	1	1	1	1	1	1	1	1	0.67	0.67	0.86	1	1	1	1	1
6	1	0.8	0.86	1	1	0	1	1	0.33	1	0.57	1	0.5	0.67	0.5	1	1	0.8	1	1	1	0.86	0.83	1	1	1	0.83
7	0	0.6	0.57	1	0.67	1	0	0.6	1	0	0.71	1	0.83	1	0.83	1	1	0.6	1	0	0	0.57	0.67	0.6	1	1	1
8	0.6	0.2	0.43	1	1	1	0.6	0	1	0.6	0.43	1	0.83	1	0.83	1	1	0.2	1	0.6	0.6	0.43	0.33	0	1	1	1
9	1	0.8	0.86	1	1	0.33	1	1	0	1	0.71	1	0.67	0.67	0.67	1	1	0.8	1	1	1	0.86	1	1	1	1	0.83
10	0	0.6	0.57	1	0.67	1	0	0.6	1	0	0.71	1	0.83	1	0.83	1	1	0.6	1	0	0	0.57	0.67	0.6	1	1	1
11	0.71	0.29	0.57	1	1	0.57	0.71	0.43	0.71	0.71	0	1	0.71	0.86	0.71	1	1	0.29	1	0.71	0.71	0.57	0.29	0.43	1	1	0.86
12	1	1	1	0.86	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0.71	1	1	1	1	1	0.71	1	0.86
13	0.83	0.67	0.71	1	1	0.5	0.83	0.83	0.67	0.83	0.71	1	0	0.83	0	1	1	0.67	1	0.83	0.83	0.71	0.83	0.83	1	1	1
14	1	0.8	1	1	1	0.67	1	1	0.67	1	0.86	1	0.83	0	0.83	1	1	0.8	1	1	1	1	1	1	1	1	1
15	0.83	0.67	0.71	1	1	0.5	0.83	0.83	0.67	0.83	0.71	1	0	0.83	0	1	1	0.67	1	0.83	0.83	0.71	0.83	0.83	1	1	1
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0.67	1	1	1	1	1	0.67	1	1
18	0.6	0	0.43	1	1	0.8	0.6	0.2	0.8	0.6	0.29	1	0.67	0.8	0.67	1	1	0	1	0.6	0.6	0.43	0.33	0.2	1	1	1
19	1	1	1	0.67	1	1	1	1	1	1	1	0.71	1	1	1	1	0.67	1	0	1	1	1	1	1	0	1	0.83
20	0	0.6	0.57	1	0.67	1	0	0.6	1	0	0.71	1	0.83	1	0.83	1	1	0.6	1	0	0	0.57	0.67	0.6	1	1	1
21	0	0.6	0.57	1	0.67	1	0	0.6	1	0	0.71	1	0.83	1	0.83	1	1	0.6	1	0	0	0.57	0.67	0.6	1	1	1
22	0.57	0.43	0.14	1	0.86	0.86	0.57	0.43	0.86	0.57	0.57	1	0.71	1	0.71	1	1	0.43	1	0.57	0.57	0	0.29	0.43	1	1	1
23	0.67	0.33	0.29	1	1	0.83	0.67	0.33	1	0.67	0.29	1	0.83	1	0.83	1	1	0.33	1	0.67	0.67	0.29	0	0.33	1	1	1
24	0.6	0.2	0.43	1	1	1	0.6	0	1	0.6	0.43	1	0.83	1	0.83	1	1	0.2	1	0.6	0.6	0.43	0.33	0	1	1	1
25	1	1	1	0.67	1	1	1	1	1	1	1	0.71	1	1	1	1	0.67	1	0	1	1	1	1	1	0	1	0.83
26	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1
27	1	1	1	0.83	1	0.83	1	1	0.83	1	0.86	0.86	1	1	1	1	1	1	0.83	1	1	1	1	1	0.83	1	0

8.3 Appendix C

Table 7
Tool-Measure co-occurrence matrix

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	TA	TB	TC	TD	TE	TF	TG	TH	TI	TJ	TK	TL	TM
amount	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
density	0	0	0	0	0	0	0	0	3	0	1	0	1	0	0	0	0	2	0	0	0	0	0
distance	2	4	0	4	2	0	5	0	0	5	0	0	0	2	2	0	0	0	0	0	0	0	0
effect	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
income	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Index	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
label	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
number	2	2	1	0	2	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
percentage	6	6	6	0	1	0	1	1	2	0	1	0	0	0	0	1	1	0	0	1	1	1	1
variety	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
WOZ-waarde	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
year	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

8.4 Appendix D

Table 8
Tool-Measure co-occurrence matrix (normalized)

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	TA	TB	TC	TD	TE	TF	TG	TH	TI	TJ	TK	TL	TM
amount	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
density	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
distance	0	1	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
effect	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
income	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Index	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
label	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
number	1	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
percentage	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
variety	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
WOZ-waarde	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
year	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

8.5 Appendix E

Table 9
TF-IDF matrix

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	TA	TB	TC	TD	TE	TF	TG	TH	TI	TJ	TK	TL	TM
amount	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
density	0	0	0	0	0	0	0	0	4	0	1	0	2	0	0	0	0	3	0	0	0	0	0
distance	1	2	0	3	1	0	3	0	0	9	0	0	0	4	4	0	0	0	0	0	0	0	0
effect	0	0	0	1	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
income	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Index	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
label	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
number	1	1	1	0	1	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
percentage	3	3	5	0	1	0	1	2	3	0	1	0	0	0	0	1	1	0	0	2	2	2	2
variety	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
WOZ-waarde	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
year	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0