UTRECHT UNIVERSITY

Department of Information and Computing Science

Applied Data Science master thesis

Analyzing GIS sub-workflow patterns: bridging natural language and workflows

First examiner:
Simon Scheider

Candidate:
Letícia Marçal Russo

Second examiner:
Judith Verstegen

Daily supervisor:
Eric Top

July 11, 2023

Abstract

The use of Geographic Information Systems (GIS) has expanded across various areas and has become even more important with the growth of big data in recent decades. However, its accessibility can be limited, because geo-analytical tools are spread out across multiple software programs and also scattered within a single software environment. To address this issue, one possible solution is to develop a geo-analytical Question-Answering (QA) system. Unlike traditional QA systems that work with simple query task, a geoGIS QA system requires a more complex transformation task. In order to achieve this, the answers provided by a GeoQA system should be in a workflow format, while the user's questions are expressed in natural language. In this context, our research focuses on bridging the gap between workflows and natural language by introducing the concept of sub-questions, which describe an underlying task, along with their corresponding sub-workflows. For this study, we generated data in the form of (sub-)questions, (sub-)workflows, and parse trees. These generated data were then subjected to analysis using two methods. The first method involves measuring the similarity between sub-questions and between parse trees to understand the extent to which GIS tools impact changes in sentences. The second method is based on word subtraction, which aims to identify specific question fragments associated with GIS tools. The methodology has demonstrated its ability to find patterns in natural language sentences and connect them to GIS tools, which are part of workflow structures. These patterns can potentially contribute to the development of a geoQA system, taking us a step closer towards its realization.

Contents

1. Introduction

The significance of Geographical Information System (GIS) nowadays extends beyond its initial application in the geoscience, encompassing various other domains such as engineering (Ramesh, 2021), healthcare (Khashoggi & Murad, 2020), business (Azaz, 2011), and tourism (Stankov et al., 2012). With the exponential growth of Big Data in recent decades (Hilbert, 2016), the importance of spatial analysis becomes even more prominent. However, the utilization of geo-analytical tools poses a challenge due to their dispersion across multiple software programs, as well as their scattered implementation within a single software environment. This issue can demand considerable effort from individuals who are already familiar with GIS, while also serving as an obstacle to those who lack the necessary skills to work with it.

What if one could ask a geo-analytical question and receive a direct answer regarding the appropriate data and analysis tools to employ? The recent progress in Question-Answering (QA) systems and language models, such as ChatGPT, presents a promising solution to the aforementioned problem.

QA systems are typically trained using machine learning algorithms on a large data set of questions and their corresponding answers. Once available for use, these systems go through a series of steps (Bouziane et al., 2015). First, the system employs Natural Language Processing (NLP) to comprehend the questions submitted by the user. Through techniques like word embeddings, it extracts information and context from the question. Following the question understanding, the system searches for relevant answers within its knowledge base (Kwok et al., 2001). It employs algorithms to determine the quality and relevance of the retrieved answers and selects the most appropriate one (Radev et al., 2002). The chosen answer is then formatted and presented to the user. Note here that this entire process operates in natural language, meaning users ask questions in natural language

and receive responses in natural language.

However, when it comes to a potential QA for spatial questions, it should not be seen as a simple query task over known data sets, but rather as a transformation task (Scheider et al., 2021). This challenge is referred to as geo-analytical QA (GeoQA) by Scheider et al., 2017. This means that answers can not be filtered directly through queries; instead, they require transformations before they can be extracted. In this mater, GeoQA differs from conventional QA systems that rely on pre-existing answers. In GIS, the focus is on creating workflows that can address analytical questions that currently lack known answers.

To illustrate the argument, consider the question: *'What is the population of the Netherlands in 2023?'*. The response to this question is already known and can be accessed from a knowledge base, which states that it is approximately 17,850,000 people. However, when it comes to a geo-analytical questions like *'What is the number of people within 1,000 meters of the A2 highway in the Netherlands?'*, there is no direct answer available. In Figure 1.1, one can see that a sequence of GIS tools must be applied consecutively to produce the desired outputs in a transformation process.

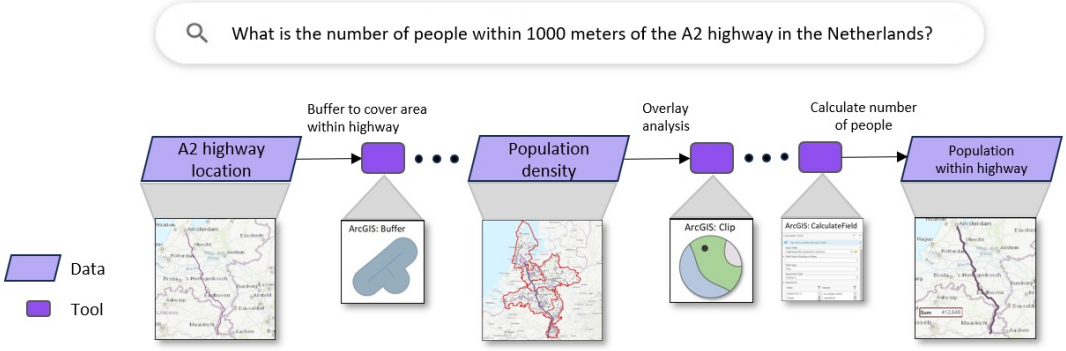

Figure 1.1: The chart presented above provides a condensed representation of the steps required to answer the spatial question: *'How many people are within 1000 meters of the A2 highway in the Netherlands?'* Various GIS tools such as Buffer, Clip, and CalculateField are employed in the process. The transformation of the outcome is depicted in the squared maps located below the parallelograms.

An ongoing academic initiative at Utrecht University named QuAnGIS project is dedicated to developing a solution for this challenge (Scheider et

al., 2020). QuAnGIS is an indirect Question-Answering system that automatically translates geo-analytical questions into GIS workflows that generate maps as answers. It is based on interpreting questions as requests for conceptual transformations (Xu et al., 2022). For further understanding on how QuAnGIS works, access the demo video [1].

Yet, it is important to highlight that while answers in GeoQA should be provided in a workflow format, the user continues to use natural language for their questions. Because of that, the issue we aim to tackle in this research is to establish a bridge between natural language and workflows. Up to the extent of our knowledge, this particular approach is currently missing in the literature. Therefore, we intend to bridge this gap as an initial step towards enabling a GeoQA system in the future.

In this context, we aim to answer the following research question:

To what extent can question patterns be extracted which correspond to the addition of tools and tool combinations in GIS workflows?

We will also address two sub-questions:

1. To what extent can natural language interpretations of GIS tools be used to construct questions phrases over GIS workflows?
2. Which phrases in natural language questions about GIS tasks are associated with different GIS tools and tool combinations?

In the upcoming chapters, we will start by explaining the process of data generation (chapter 2). During this explanation, we will introduce the concept of sub-questions and their corresponding sub-workflows that were specifically developed for this study. In chapter 3, we will describe two methods utilized to identify patterns in the sub-questions and establish connections with GIS tools within workflows. The first method involves measuring similarity between questions and also between parse trees, while the second method is based on word subtraction. These methods aim to determine the extent to which GIS tools impact changes in sentences and how

[1] https://video.uu.nl/videos/quangis-demo/

these tools can be associated with specific question fragments, respectively. Subsequently, we will delve into a discussion and provide suggestions for future work (chapter 5), before concluding the article (chapter 6).

2. Data

The data utilized in this study were generated through a three-stage process. Initially, workflows were reproduced by experts from the QuAnGIS project [1], using ArcGIS Pro tutorials [2] as a reference. Subsequently, a methodology was devised to generate sub-workflows and corresponding sub-questions for each of the aforementioned workflows. Finally, parse trees were generated to depict the sub-questions created during the preceding process. In this chapter, we provide an example by describing one of these workflows. In section 2.4, we will quickly report how the data was prepared for further analysis.

2.1 Workflows

A workflow is a series of steps required to generate a map in GIS that is suitable for answering the question of an analysis. For easier comprehension, we will refer to the outcome of the workflow as a map. However, it is important to note that from a technical standpoint, this outcome is actually a data set.

The workflows were annotated as directed acyclic graphs (DAG), meaning that the edges have specific direction and, in this case, converge to the root node. The root is the only final node of the workflow and represents the final output, which is the map in GIS. Besides the root, the nodes can be of two types, namely actions and artefacts. An action node represents an application of a GIS-tool to data, and the artefact depicts a GIS-data set. The workflows annotation are available in this GitHub repository [3].

To illustrate our arguments, we will use one of these workflows and its

[1] https://questionbasedanalysis.com/
[2] https://learn.arcgis.com/en/gallery/
[3] https://github.com/quangis/QuAnGIS_workflow_annotation/tree/main

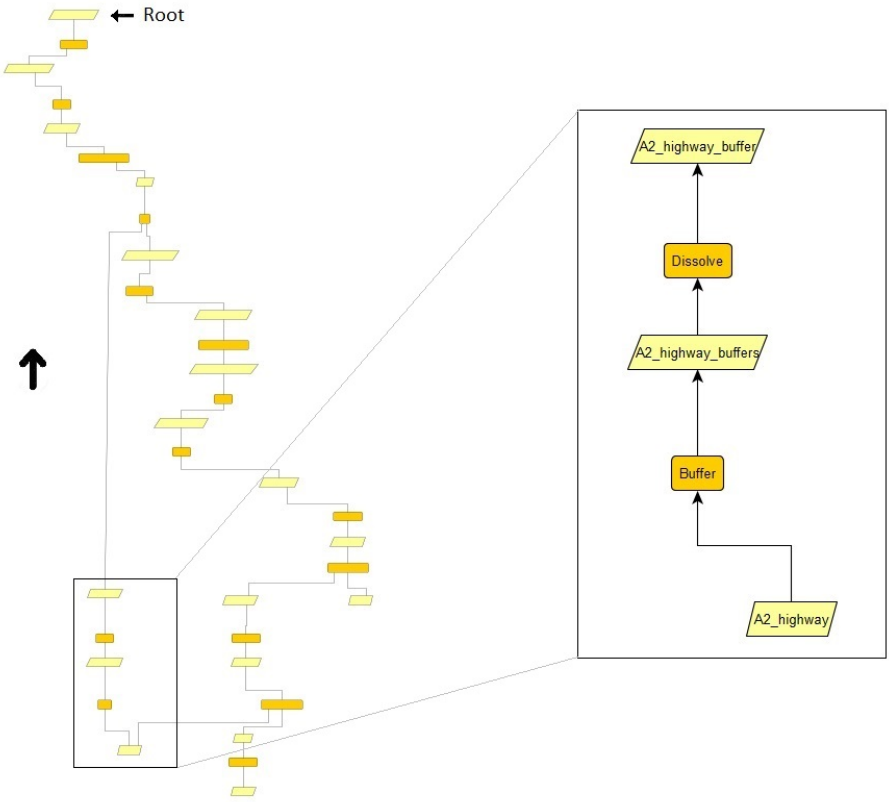

Figure 2.1: The workflow starts with artefact nodes that, after several steps, converge to the root node, which contains the final output.

corresponding task as example along the article. The Dutch government has to make public policy decisions and therefore needs to find out how many people are living within 1,000 meters from the A2, a Dutch highway. The workflow starts at the bottom with artefact nodes and after a set of steps it converges to the root node at the top (see Figure 2.1), which represents the map with the number of people around the highway. When we zoom in, one can see that the first step consists of an artifact node that represents data about the location of the A2 highway. It is followed by an action node that depicts the buffer tool in GIS, which generates an artefact as output, the *A2_highway_buffers*. Note that artefacts are represented by a parallelogram, and actions, by a rounded rectangle.

This workflow is just one of many possibilities to achieve the desired output. Additionally, the subsequent analysis is based on six distinct workflows, each outlining a step-by-step procedure for six different tasks. These workflows were deliberately chosen to have diverse characteristics and cover

a wide range of application areas in GIS, aiming to identify a variety of patterns. For information about the objective task of each of the six workflows, please refer to the appendix.

Finally, remember that the goal is that these workflows will function as answers in a future GeoQA system. In our example, the workflow in Figure 2.1 would be the answer for the question: *'What is the number of people within 1000 meters of the A2 highway in the Netherlands?'*

2.2 Sub-workflows and sub-questions

In order to achieve a suitable GeoQA system, it is crucial to not only focus on the workflows but also to delve into question understanding. This leads us to the second phase of data generation, where a specific method was developed for generating task questions, sub-questions, and sub-workflows specifically tailored for this study.

A task question is a question whose answer is the map generated by the entire workflow solving this task. It is expressed in natural language and relates to the root node. Correspondingly, sub-questions are inquiries that are answered by corresponding sub-workflows and align with an underlying task. Considering that the workflow direction is from bottom to top, where the root node is located, all the nodes under the sub-question that are in the same branch form a sub-workflow.

In Figure 2.2, one can see the sub-questions added to the workflow of Figure 2.1. The purple octagons contain the sub-questions and all the nodes below it form its corresponding sub-workflow.

The sub-questions were generated using the following rules:

(1) **A new sub-question and its corresponding sub-workflow was created whenever the semantic meaning of the output changed, and therefore the question's goal, with the application of one or more GIS tools**. The sub-questions were generated following the natural flow of actions that a GIS user would follow, that is, starting with the first data set until reaching the final map. In the workflow, it begins at the bottom until it attains the root

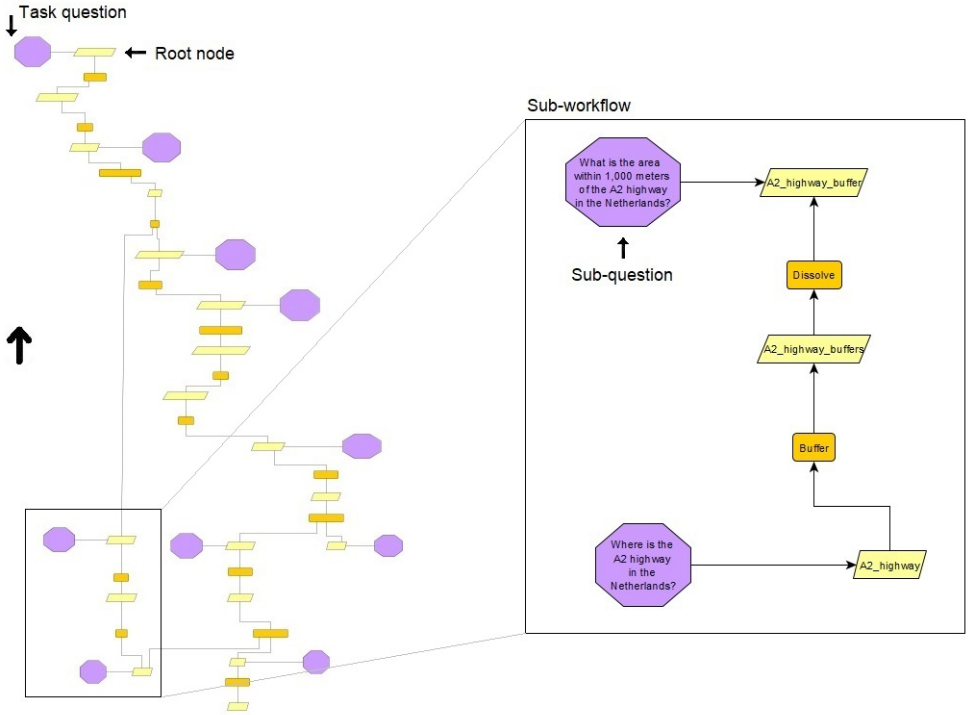

Figure 2.2: A sub-workflow and its correspondent sub-questions.

node.

In Figure 2.2, the first node *A2_highway*, which represents the data set with the location of the A2, corresponds to the question *'Where is the A2 highway in the Netherlands?'* After applying the buffer and dissolve tools, the output is the *A2_highway_buffer*, which equates to the question *'What is the area within 1,000 meters of the A2 highway in the Netherlands?'*

(2) **The sub-questions were elaborated in the QuanGIS blockly-based natural language interface** [4] . The interface is built on Google Blockly library and incorporates four distinct types of blocks. These blocks serve as fundamental components for constructing geo-analytical questions and need to be logically connected. The green blocks represent question words like "where," "what," and "which," along with their corresponding auxiliary words such as "is," "are," and "do not." Noun phrases representing geographic concepts can be formed using the blue blocks, which support three different syntactic structures. The purple blocks represent relation-

[4]https://haiqixu.github.io/

ships and qualities that impose restrictions on the geographic concepts in geo-analytical questions. Lastly, the yellow blocks are designed for specifying spatial extent and can accept any place names in the text box. They also have an optional temporal extent field that allows for the inclusion of month and year information (Xu et al., 2022).

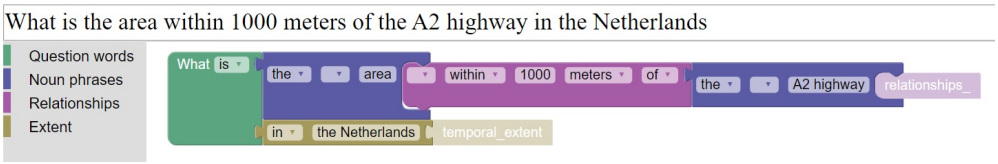

Figure 2.3: The QuanGIS blockly-based natural language interface is composed of four blocks.

In this case, the interface is a controlled natural language, functioning as a semantic and syntactic constraint (Schwitter, 2010). This ensures that sentences adhere to predefined structures, thereby mitigating the human bias introduced during sentence formulation, which arises from the various possibilities of formulating a question.

In Figure 2.3, one can see an example of how the question *'What is the area within 1,000 meters of the A2 highway'* was generated in the blockly-based interface. Each block represents a syntactic structure such as question word (What is), noun phrase (the area), relationship (within 1000 meters of), noun phrase (the A2 highway) and spatial extent (in the Netherlands).

(3) **The sub-questions use non-technical language: this means that they do not contain terms or names of tools used in GIS such as buffer and clip**. In the example above, the buffer concept was phrased as *'the area within 1,000 meters of'*. Although this is not imperative for the analysis in this study, the idea is that the questions should be developed in a way that reflects how users would ask them. Therefore, the questions should not include the names of specific GIS tools, as users might not be familiar with the necessary actions to obtain the desired output.

(4) **The sub-questions were formulated in such a way as to keep them concise and as short as possible, while still conserving their previous concepts when more than one step in the workflow was apparent in the out-**

put. In the following example *'What is the area of each population center within 1,000 meters of the A2 highway in the Netherlands?'*, two steps were needed to get to this sub-workflow: (1) the area of each population center that is inside (2) the 1,000 meters buffer around the highway. These two outputs are, therefore, expressed in the sub-question.

The same rules were applied to the formulation of a task question, except for the fact that the task question relates to the whole workflow. A set of 55 questions was formulated, spanning across the six workflows. From here on, the term *questions* will be used as synonym of task question and sub-question for ease of mention.

2.3 Parse trees

Once the sub-questions and task questions for the six workflows were elaborated, each of them was subsequently transformed into parse trees of conceptual transformations needed to answer a geographic question. These trees, which provide hierarchical representations of sentence structure, enable subsequent syntactic and semantic analysis of the sentences. The methodology for obtaining the parse trees was proposed by Xu et al., 2023.

It is important to note that the generation of parse trees is just one component of the questioning parsing process proposed by the previously mentioned study. This approach examines questions by employing the core concepts of spatial information and their functional roles within a context-free grammar. To achieve this, the process begins by identifying the core concepts in questions (Scheider et al., 2020), then applies the concept transformation model, and utilizes functional roles to determine the ordering of concept transformations.

To illustrate, the parse tree provided below depicts the corresponding structure for the sub-question *'Where is the A2 highway in the Netherlands?'*:

(start (measure (location where is (coreC object 0))) (extent extent)

That is the result we obtain by substituting the core concepts and extents of the parse trees:

(start (measure (location where is (the A2 highway))) (in the Netherlands)

The top-level element is the 'start', indicating the beginning of the parse tree. The first major component is the measure, represented by the sub tree *(measure (location where is (coreC object 0)))*. The *measure* component is a functional role and yields the question's goal and the inputs for GIS workflows (Xu et al., 2023). Within the measure sub tree, we have another sub tree: *(location where is (coreC object 0))*, representing where a spatial phenomenon is. Finally, we have the *(coreC object 0)* sub tree, which provides additional details about the object being referred to. *Location* and *object* are core concepts according to (Kuhn & Ballatore, 2015). The second major component is the *extent*, represented by the sub tree *(extent extent)*. In this case, *extent* defines the spatial boundary for *measure* and it is also a functional role. Note that the parse trees capture conceptual transformations based on recognizing concepts and their ordering using functional roles.

2.4 Data preparation

Lastly, the data generated in the above-mentioned processes were prepared in a suitable format -a data set- to apply the subsequent analysis. The nodes of the directed acyclic graphs (DAG) in GraphML file format were extracted so that a column of the data set contains all the 55 questions of the six workflows used in this study and an extra attribute to describe which workflow that question belongs to. The third column contains all the artefact and ac-

tion nodes of the sub-workflow corresponding to each question. Finally, the parse trees were extracted and stored in the last column of the data set.

3. Method

In order to analyze the patterns of the sub-workflows and understand how they relate to natural language, we will apply Natural Language Processing (NLP) techniques. The aim is to comprehend to what extent the sub-questions change when the sub-workflow is pruned and how this relates to corresponding GIS tools.

3.1 Description of the method used

To attain the aforementioned objective, two types of analysis were employed. The first type entails the application of methods to measure the similarity between two questions and the similarity between parse trees. The second type involves working with word subtraction between two questions. The resulting similarity measures, as well as the subtracted words, were then associated with GIS tools or combinations of tools that corresponded to the workflow subtraction.

Here, it is appropriate to provide an explanation for the process of conducting the analysis. Consider Z as the entire workflow, which includes the root node along with all the artefact and action nodes. By breaking down the workflow into sub-workflows starting from the root node and focusing on nodes with sub-questions, we can make pairwise comparisons whenever a sub-workflow is pruned. These comparisons can be made between questions, parse trees, and workflow structures. See Figure 3.1 and remember that the purple octagons represent sub-questions and serve as the starting points for generating new sub-workflows.

When performing the cutting process from the root to the bottom of Z, the subsequent octagon will be considered as a cutting point, resulting in the generation of sub-workflow A that encompasses all nodes below the corresponding sub-question. The subsequent octagon, situated further down in

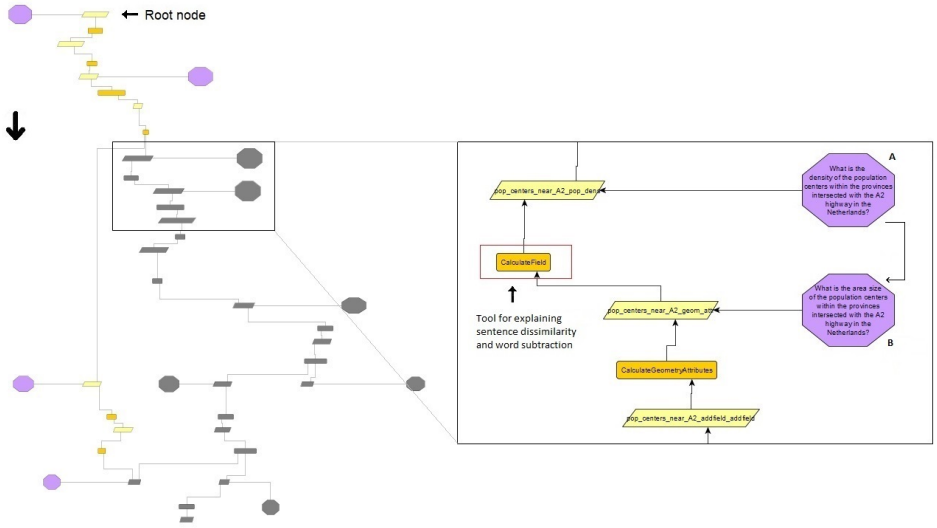

Figure 3.1: When we remove the root node from a workflow, new sub-workflows are generated whenever there is a sub-question, indicated by purple octagons in this directed acyclic graph. In the graph, there are highlighted examples of sub-workflows and a pairwise comparison between sub-workflow A and B, shown in gray. Here, sub-workflow B is a subset of sub-workflow A. In both types of analysis, namely word subtraction and similarity, the resulting output would be connected to the GIS tool *CalculateField*.

the same direction, will be referred to as sub-workflow B. Consequently, the analysis involves subtracting sub-workflow B from sub-workflow A. It is important to note that pairwise comparisons will be conducted when sub-workflow A contains sub-workflow B, expressed as:

$$A \subseteq B$$

Following this explanation of how the comparison process functions, the subsequent subsections will provide further details on similarity and word subtraction methods, along with some examples to illustrate them. Furthermore, the implementation of the methods can be found on this github [1].

[1] https://github.com/leticiaamarcal/thesis_QuAnGIS/tree/main/Code

3.2 Similarity

The concept of similarity, within this context, allows for the measurement of likeness between two questions based on different criteria, such as syntactic structure and semantic relationships. By employing different approaches and comparing their outputs, we can identify patterns that emerge when sub-questions change and establish connections to GIS tools. This process allows us to obtain insights into the extent to which specific GIS tools, either individually or in combination, produce significant or negligible changes in the questions. Below, we present a detailed description of the two similarity methods that have been applied with distinct objectives:

(1) **SBERT and cosine similarity**: The Sentence-BERT (SBERT) is "a modification of the pretrained BERT network that use siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity" (Reimers & Gurevych, 2019). SBERT followed by cosine similarity were applied in the questions and, therefore, are suitable for this analysis since sentences are being compared. The goal of this analysis is to understand how similar two sentences are semantically. It is important to note that although SBERT indirectly takes into account certain aspects of syntax by comprehending the contextual meaning of words, its main purpose is to capture the semantic similarity between sentences.

In the first step of this method, SBERT transforms the input sentences into embeddings, which are vector representations able to encode the contextual and semantic meaning of the phrases. After this process, cosine similarity technique is applied to measures the the cosine of the angle between two vectors. It captures the similarity in direction rather than magnitude, where two similar vectors are expected to have a small angle between them (Xia et al., 2015). The similarity measure ranges from 0 to 1, where 0 indicates total dissimilarity and 1 indicates total similarity.

As mentioned in chapter 1, the question comprehension is a crucial aspect of a Question-Answering system. In the scenario of a GeoQA, examining the similarity between questions can aid in this stage of the process

and also later on, in the answer retrival, as we are associating this measure to GIS tools, which will be part of the answer. Additionally, since SBERT has been trained on a vast corpus of textual data, the controlled natural language used in this study to elaborate the questions may not pose a limitation, as SBERT can identify alternative question phrasings that convey the same meaning.

(2) **Tree Edit Distance**: The similarity of parse trees is determined by calculating the minimum number of operations needed to transform one tree into another (Bille, 2005). This measurement takes into consideration three edit operations: insertion, deletion, and relabeling. In our work, we utilized the Zhang-Shasha algorithm, which was originally introduced by Kaizhong Zhang and Dennis Shasha in 1989 (Zhang & Shasha, 1989). To implement this algorithm, we employed the *zss* package in Python and applied a function that assigns equal costs to all edit operations. These operations enable the algorithm to account for changes in node labels, addition or removal of nodes, and modifications to the tree structure. The results of TED were scaled to range from 0 to 1, where a value of 0 indicates identical trees, while a value of 1 indicates maximum dissimilarity.

It is important to note that each node in the tree contains a label, while the edges represent the relationships between the nodes. In our case, the parse trees utilize core concepts and functional roles as labels. This implies that the method also measures semantic similarities but at a higher level, which may result in certain semantic changes going unnoticed. For instance, considering the example provided in section 2.3, *'in the Netherlands'* is substituted by the functional role *'(extent extent)'*, whereas *'the A2 highway'* is referred to as *'(coreC object 0)'*, representing a core concept. This means that if we only replace *'in the Netherlands'* with *'in Brazil'* in a sentence, TED would not detect any change in the parse trees. Consequently, comparing the results obtained from SBERT and TED can lead to interesting analysis since TED primarily reveals structural changes in sentences, represented as parse tress, while SBERT can provide deeper insights into semantic alterations.

To illustrate, when comparing the sub-questions *'What areas are within*

1000 meters of the A2 highway in the Netherlands?' and *'Where is the A2 highway in the Netherlands?'*, the analysis yields a SBERT similarity of 0.85 and a Tree Edit Distance of 0.48. These numbers are associated with the GIS tool combination of Buffer and Dissolve. More comprehensive details regarding the analysis of these values will be presented in chapter 5.

3.3 Word subtraction

In the word subtraction analysis, we will compare two questions by identifying the differences between their sets of words. Specifically, we will subtract words from the question of workflow A by words from the sub-question of sub-workflow B. Consequently, the resulting output will consist of the words from question A that do not correspond with the words in question B. In this case, the overarching workflow question will be subtracted by the sub-question of the nested workflow. The objective here is to understand which words in the question are a result of specific GIS tools or their combination.

To provide a concrete illustration of the subtraction process, consider the following example: when subtracting the sub-question *'What is the area size of population centers within the provinces intersected with the A2 highway in the Netherlands?'* from the sub-question *'What is the density of population within provinces intersected with the A2 highway for each population center in the Netherlands?'*, the resulting words are *'density for each center'*. It is worth noting that in this case, words are only subtracted if they are identical. Therefore, *'centers'* and 'center' are not subtracted from each other. The words obtained from this subtraction example are associated with the *CalculateField* GIS tool.

4. Results

In this chapter, we will provide an overview of the results. The tables resulting from the similarity analyses and word subtraction will be partially displayed to prevent excessive information overload. We will selectively display the key columns of the output table. Remember that SBERT and cosine were employed to assess similarity in the questions. Conversely, the Tree Edit Distance was specifically utilized for the parse trees. The parse trees will not be displayed due to lack of space. Regarding word subtraction, it is worth reiterating that it was applied pairwise to the questions. The complete tables can be accessed in this github [1]. Moreover, the interpretation of the results will be further expanded in subsections 4.1.1 and 4.2.1.

4.1 Similarity results

The result table includes pairwise comparisons of questions and corresponding measures for SBERT and Tree Edit Distance. The results are linked to either a single tool or a combination of tools. This distinction arises from the fact that in some cases, the path between two sub-questions in the corresponding sub-workflow involves only one tool, while in other cases, it involves multiple tools.

[1] https://github.com/leticiaamarcal/thesis_QuAnGIS/tree/main/Data

question 1	question 2	SBERT	TED	GIS tools
What are the rooftop cells with slope lower than 45 degrees in Glover Park, Washington, D.C.	What is the solar radiation in KWh/m2 for each rooftop cell in Glover Park, Washington, D.C.	0.821	0.455	Slope + Con
What is the percentage of the rural population within 2 kilometers of the all-season roads in Shikoku, Japan	What is the proportion of population within 2 kilometers of the all-season roads for each rural district in Shikoku, Japan	0.989	0.035	SummaryStatistics + AddField + CalculateField + SummaryStatistics + AddFields + CalculateField + JoinField + CalculateField
What are the rooftop cells with slope lower than 45 degrees and with solar radiation higher than 8000 kWh/m² in Glover Park, Washington, D.C.	What are the slopes in Glover Park, Washington, D.C.	0.755	0.568	RasterCalculator + Con + Con
What are the rooftop cells with slope lower than 10 degrees, with solar radiation higher than 8000 kWh/m², with aspect higher than 22.5 degrees, and with aspect lower than 337.5 degrees in Glover Park, Washington, D.C.	What is the aspect for each rooftop cell in Glover Park, Washington, D.C.	0.774	0.714	RasterCalculator + Slope + Con + Con + Con + Con
What is the solar radiation in KWh/m2 for each rooftop cell in Glover Park, Washington, D.C.	What is the solar radiation in Wh/m2 for each rooftop cell in Glover Park, Washington, D.C.	0.997	0	RasterCalculator

Table 4.1: Measures for similarity for SBERT and Tree Edit Distance

4.1.1 Interpretation

Below, we will suggest some questions that can be answered using the result table. These inquiries can facilitate the interpretation of the results and indirectly contribute to constructing the answer for the research question and sub-questions of this study.

(1) **What is the relationship between SBERT and Tree Edit Distance in the context of measuring similarity?**

By examining the relationship between SBERT and TED, we observe a negative correlation between them. See Figure 4.1. The Pearson correlation coefficient of -0.62 indicates a large effect, as described by Field, 2013. The negative relationship arises because a TED value of zero signifies complete similarity, whereas an SBERT and cosine value of zero indicates complete dissimilarity.

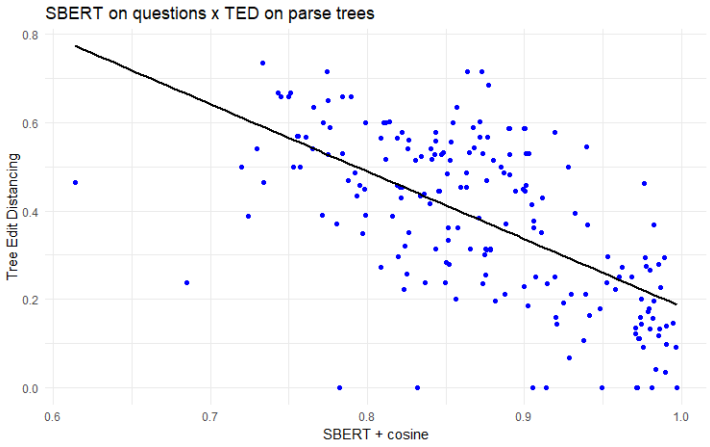

Figure 4.1: SBERT and TED measurements demonstrate a strong correlation. When examining a scatter plot, one can effectively compare situations where TED values exceed SBERT values and vice versa by taking into consideration the linear model represented by the black line.

The blue points can represent the comparison of two (sub-)questions and their parse trees. Alternatively, they can also symbolize the GIS tools that generate the dissimilarities between the two sentences when utilized within the workflow. If we consider each point as a tool or a combination of tools, we can see that, overall, tools that generate dissimilarities according to SBERT also do so according to TED. In other words, the core concep-

tual interpretation of the question, as determined by the parse tree, and the word vector interpretation are correlated. Assuming that the core concepts are also represented in GIS tools, we can infer there is also a correlation between tools and question phrases. Additionally, this demonstrates that word vector methods such as SBERT can be used to associate questions with core concepts.

By individually analyzing each point, we can identify which tools exhibit higher dissimilarity in terms of SBERT and lower dissimilarity in terms of TED, and vice versa. It is important to remember that SBERT is focused on measuring semantic changes, while TED captures syntactic changes in sentences and also semantic changes at a core concept level. Within this context, we can identify patterns in how GIS tools modify sentences and utilize these patterns in the mechanisms of a geo-analytical QA system.

(2) **Are there specific tools or combinations of tools that can significantly alter the semantics of a question while other tools have minimal impact on the question's semantics?**

Certain tools or combinations of tools indeed have the potential to significantly alter the semantic meaning of a sentence, while others may only bring about subtle changes. An example of a tool combination is *Slope* and *Con*, which yields a larger semantic change, with a SBERT score of 0.821 and a TED score of 0.455. These tools transform the question *'What is the solar radiation in KWh/m2 for each rooftop cell in Glover Park, Washington, D.C?'* into *'What are the rooftop cells with slope lower than 45 degrees in Glover Park, Washington, D.C.?'* On the other hand, multiple tools can work together to generate subtle semantic changes. For instance, the combination of *SummaryStatistics*, *AddField*, *CalculateField*, *SummaryStatistics*, *AddFields*, *CalculateField*, *JoinField*, and *CalculateField* achieves a SBERT similarity score of 0.989 and a TED score of 0.035. This combination alters the question *'What is the proportion of population within 2 kilometers of the all-season roads for each rural district in Shikoku, Japan?'* into *'What is the percentage of the rural population within 2 kilometers of the all-season roads in Shikoku, Japan?'*

(3) **Are there GIS tools or combination of that make the similarity mea-**

sured by SBERT and by TED vary in opposite directions?

Some GIS tools or combinations of tools can increase the similarity measured by SBERT while decreasing the similarity measured by TED, and vice versa. One example is the extra addition of the tool combination *Slope*, *Con* and *Con* in the workflow. When comparing the sentences *'What are the rooftop cells with slope lower than 45 degrees and with solar radiation higher than 8000 kWh/m² in Glover Park, Washington, D.C.'* and *'What are the slopes in Glover Park, Washington, D.C.'*, it yields a SBERT of 0.755 and a TED of 0.568. This similarity result corresponds to the GIS tools combination *RasterCalculator*, *Con* and *Con*. However, when comparing the similarity between the sentences *'What are the rooftop cells with slope lower than 10 degrees, with solar radiation higher than 8000 kWh/m², with aspect higher than 22.5 degrees, and with aspect lower than 337.5 degrees in Glover Park, Washington, D.C.'* and *'What is the aspect for each rooftop cell in Glover Park, Washington, D.C.'*, we observe higher similarity according to SBERT (0.774), but lower similarity according to TED (0.714) compared to the previous example. Remember here that a SBERT close to 1 represents a perfect similarity, while the total similarity measured by TED is equal to 0. The similarity analyzed in the second example relates to the tool combination of *RasterCalculator*, *Slope*, *Con*, *Con*, *Con* and *Con*. This suggests that the addition of *Slope*, *Con* and *Con* in the workflow increases the similarity based on vector representation but induces a decrease when considering the core conceptual parse tree representation.

(4) **Which GIS tools can induce semantic changes in a question while preserving its syntax intact?**

The *RasterCalculator* GIS tool can make slight semantic modifications to a question while keeping its syntax unchanged. SBERT captured the semantic change, scoring 0.997, while the TED score remained at zero, indicating no discernible change. By utilizing *RasterCalculator*, the only tool required, the sub-question *'What is the solar radiation in Wh/m2 for each rooftop cell in Glover Park, Washington, D.C.?'* was transformed into *'What is the solar radiation in KWh/m2 for each rooftop cell in Glover Park, Washington, D.C.?'*. The alteration specifically pertains to the measurement unit

for solar radiation, changing it from *Wh/m2* to *KWh/m2*. From a syntactical standpoint, both sub-questions retain the same sentence structure. They share an identical parse tree: `(start what is (measure (coreC conamount 0 era_) (support support) (extent extent))`. It is worth noting that the parse tree employs core concepts, indicating that the unit measurement in both sentences refers to the same concept. In this scenario, the TED score would only detect any syntactic changes, which, in this particular case, are nonexistent.

The scatter plot depicted in Figure 4.2 demonstrates the observed pattern shared by other GIS tools or their combinations, which are represented by the highlighted blue color.

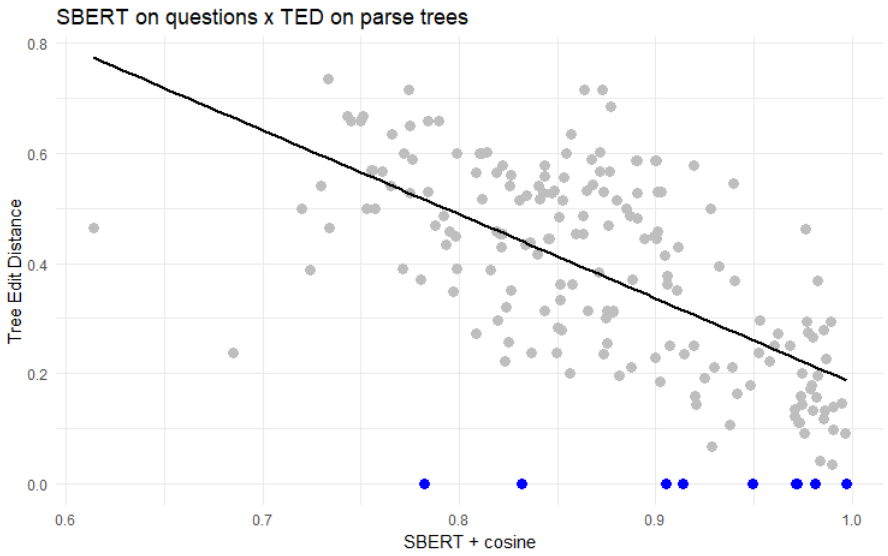

Figure 4.2: The blue points in the scatter plot highlight cases in which GIS tools promote a semantic change in the questions but the syntactic structure remains unchanged.

4.2 Word subtraction results

The table generated after subtracting words consists of GIS tools along with their corresponding sentence fragments. The questions themselves are not displayed, but it is important to remember that the *words* colum represents the result of subtracting two questions. For instance, in the case of the buffer tool, the words *'What areas within 2 kilometers of'* correspond to the subtrac-

tion of the question *'What are the areas within 2 kilometers of the all-season roads in Shikoku, Japan?'* from the question *'Where are the all-season roads in Shikoku, Japan?'*. See table 4.2.

GIS tools	**words**
Buffer	'What areas within 2 kilometers of'
Buffer + Dissolve	'What areas are within 1000 meters of'
Buffer + Clip	'areas are within 2 kilometers of all-season roads'
Clip + Buffer + Clip	'areas within 2 kilometers of all-season roads for each district'
Con	'10' 'and solar radiation higher 8000 kWh/m²'
SelectLayerByAttribute + ExportFeatures	'Utrecht' 'Which are and with lower than 200,000 euros'

Table 4.2: Word subtraction results associated with GIS tools

4.2.1 Interpretation

To facilitate the analysis of the word subtraction result table, we will employ the same question format used in sub section 4.1.1.

(1) **Which components of a sentence can a particular GIS tool contribute to constructing?**

For example, the GIS tool *Buffer* can be used independently to construct part of a sentence such as *'What areas within 2 kilometers of'*. When combined with the *Dissolve* tool, it can produce similar outcomes, such as *'What areas are within 1000 meters of'*. Both sentence fragments maintain the same structure, but differ only in the unit of measurement: kilometers and meters. This analysis demonstrates that when the *Dissolve* tool is applied after the *Buffer* tool, it does not introduce any significant changes to the question. On the other hand, when *Buffer* is combined with the *Clip* tool, it adds extra information: *'areas are within 2 kilometers of all-season roads'*. In this case, we gain knowledge about the specific reference point for the buffer, which is

the all-season roads.

The answers to these two questions utilize specific GIS tools as examples to demonstrate how the analysis can be conducted. However, it is important to note that these examples should be expanded upon in a broader analysis. Since our data set is small, we had a limited number of examples for each tool. Consequently, gathering more data is essential to identify consistent patterns in the application of these tools. More data would enable us to uncover reliable and recurring patterns in the usage of these tools.

(2) **Could we use tool combination results to analyze which part of the sentence a single tool is responsible for?**

If there is a specific tool that is not applied independently in the available workflows, we can utilize tool combinations to understand the role that particular tool plays in the resulting phrase of the tool combination.

When the tools *Clip*, *Buffer*, and *Clip* are applied in this order, they relate to the phrase *'areas within 2 kilometers of all-season roads for each district'*. In the case of the *Buffer* and *Clip* combination, they generate the phrase *'areas within 2 kilometers of all-season roads'*. This allows us to infer that the additional use of the *Clip* tool in the first example is responsible for the fragment *'for each district'*. Considering the example relating to the *Buffer* tool available in the table, the *Buffer* tool could be responsible for the fragment *'areas within 2 kilometers of'*. Lastly, the second use of the *Clip* tool is likely generating the sentence fragment *'all-season roads'*. The *Clip* tool is used to *'cut out a piece of one data set using one or more features in another data set as a cookie cutter.'*[2] This suggests that the *Clip* tool likely generates a piece of information from a data set.

(3) **Are there any contradicting combinations of GIS tool and phrases?**

Some GIS tools or combinations of tools can be associated with very different phrases. For instance, the ArcGIS Pro tool known as *Con* can be associated with the phrases *'10'* and *'solar radiation higher than 8000 kWh/m²'*. This tool "performs a conditional if/else evaluation on each of the input cells of

[2]https://pro.arcgis.com/en/pro-app/latest/tool-reference/analysis/clip.htm

an input raster" [3].

Another example is the combination of *Select Layer By Attribute* and *Export Features* tools, which can be related to phrases such as *'Utrecht'* and *'Which are and with lower than 200,000 euros'*. The *Export Features* tool converts a feature class or feature layer into a feature class [4], while the *Select Layer By Attribute* tool "adds, updates, or removes a selection based on an attribute query" [5].

Certain tools or combinations of tools exhibit broader applicability, as exemplified by the aforementioned ones, while others have a more consistent application, like *Buffer*, which always "generates polygons around input features at a specified distance" [6]. This broader scope of application appears to connect the tool with different phrases, thereby making it more challenging to identify patterns.

[3]https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/con-.htm
[4]https://pro.arcgis.com/en/pro-app/latest/tool-reference/conversion/export-features.htm
[5]https://pro.arcgis.com/en/pro-app/latest/tool-reference/data-management/select-layer-by-attribute.htm
[6]https://pro.arcgis.com/en/pro-app/latest/tool-reference/analysis/buffer.htm

5. Discussion

It is crucial to keep in mind that the primary objective of this study is to propose a methodology for connecting natural language and workflows, rather than making direct assertions about specific GIS tools and their relationship to particular aspects of the question. The analysis is based on a limited data set of six workflows. As such, drawing conclusive statements about the tools themselves would be premature. In this chapter, we will address the limitations of our study and provide recommendations for future work that can contribute to the development of a robust GeoQA system.

5.1 Limitations and future work

For the next step, we propose scaling up the utilization of the methodology introduced in this study by leveraging larger data sets. With larger data, it becomes easier to identify consistent patterns across GIS tools and natural language questions. Additionally, it is important to expand the number of tools studied in this research, as the current study is limited to only six workflows.

Still in relation to the improvement of the methodology proposed in this study, we suggest to enhance the performance of SBERT by fine-tuning it with a larger labeled data set, allowing the model to better fit the data. In the context of the Tree Edit Distance algorithm, it is possible to develop a cost design that is more suitable for the given task. Furthermore, future research could also explore the subtraction of parse trees. This analysis would enable the identification of the tools responsible for adding specific core concepts. Hence, it is appropriate to evaluate both the models and the methodology in this regard.

In relation to word subtraction, as previously stated, only exact matching words were subtracted. However, this rule can be reconsidered when

conducting analysis on a larger data set where more robust patterns are discovered. An alternative approach to this rule could involve applying stemming or lemmatization to the questions, thereby increasing the flexibility of the subtraction process. For instance, in the provided example in section 3.3, *'centers'* and *'center'* would be considered as the same word and, therefore, subtracted from each other.

Once the methodology has been expanded and a large labeled data set is available, we propose employing deep learning techniques to establish connections between phrases and GIS tools. One suitable approach for this task could involve utilizing probabilistic models such as Attention-Based Sequence-to-Sequence Models (Chorowski et al., 2015). By implementing it, it would enable the development of a GeoQA system that utilizes the model to generate answers in a workflow format based on natural language questions.

Finally, future work could also concentrate on analyzing the data set nodes rather than focusing on the GIS tool nodes as done in this research. In our analysis, we observed that certain tools, such as *Buffer*, result in modifications to the questions without requiring the use of new data sets. Conversely, in other cases, it appears that most question changes are attributed to the addition of new data sets in the workflow. This hypothesis can be examined in a future study, albeit with a different approach from the one proposed in this study. It is worth noting that unlike GIS tools, the names of data sets are not repeated across workflows.

6. Conclusion

The primary goal of this research is to establish a connection between natural language and workflows with the aim of enabling a GeoQA system in the future. In this context, our objective was to address the following research question: *'To what extent can question patterns be extracted which correspond to the addition of tools and tool combinations in GIS workflows?'* We have developed a methodology for elaborating sub-questions, in natural language, that pertain to the underlying tasks along the workflow, thereby creating sub-workflows. The sub-questions enabled us to compare them with each other and connect the results of these comparisons to specific GIS tools, thus establishing connections between workflows and natural language. Two methods, namely similarity measures and word subtraction, were employed to analyze patterns within the sub-workflows and their corresponding sub-questions. Through these methods, standard behaviors of GIS tools and their impact on the (sub-)questions were identified, although some outliers and contradicting results were found.

To address the first sub-question, *'To what extent can natural language interpretations of GIS tools be used to construct questions phrases over GIS workflows?'*, we employed similarity measures as a means of investigation. Specifically, we utilized SBERT with cosine similarity to assess the pairwise similarity between (sub-)questions. Additionally, we applied the Tree Edit Distance method to analyze the parse tree representation of the questions, providing insights into the semantics in a core concept level and structural similarities between them. The results obtained from both methods were then linked to specific GIS tools, enabling us to identify patterns that illustrate how the employment of these tools can impact the semantic and syntax of the sentences. These patterns can be used later as data to feed a probabilistic model, enabling the construction of question phrases based on GIS workflows.

Finally, we have investigated the second sub-question, *'Which phrases in natural language questions about GIS tasks are associated with different GIS tools and tool combinations?'*, by means of the word subtraction method. We pairwise subtracted two (sub-)questions, A and B, and linked the non-matching words of sentence A to the corresponding GIS tools. It is important to note that these tools are related to those present in workflow A but not in B, with the understanding that workflow B is nested within workflow A. In other words, these tools are responsible for transforming question B into question A. This method, therefore, allowed us to find out which phrases in natural language questions are associated with different GIS tools and tool combinations.

To conclude, the methodology presented in this research demonstrates its effectiveness in connecting natural language with workflows. As a suggestion for future work, we recommend scaling the application of this methodology by utilizing larger data sets. This expansion will enable us to make more reliable assertions regarding specific GIS tools. A further step would involve training a probabilistic model capable of generating answers in a workflow format using natural language as questions, thereby paving the way for a geo-analytical question-answering system.

Acknowledgments

I would like to express my gratitude to my supervisor Simon Scheider for his patience, extensive knowledge sharing, inspiration, and above all, his kindness. I would also like to extend a heartfelt thank you to the QuAnGIS team. Firstly, Eric Top, my daily supervisor, who provided invaluable feedback and was always there to offer assistance whenever needed. Secondly, I want to express my gratitude to Haiqi Xu for her invaluable support during the crucial phase of data generation. And lastly, I would like to thank Niels Steenbergen for his insights into coding.

I cannot overlook the pleasure I experienced in sharing my daily routine during this thesis with Anqi Jiang and Wiktoria Libera. Having you both around made the journey much smoother.

Lastly, I would like to emphasize how much I have learned through this project and express my deep appreciation for the privilege of receiving such a high-quality education. I sincerely hope that my work can contribute, even if in a small and indirect manner, to creating a better and more equitable society.

Appendix

A. Appendix: Workflows and task questions

Workflow	Task	Task question
population [1]	Analyze how many people are living within 1,000 meters of the Dutch A2 highway	What is the number of people within 1000 meters of the A2 highway in the Netherlands?
access [2]	Estimate access to all-season roads in rural areas of Japan	What is the percentage of the rural population within 2 kilometers of the all-season roads in Shikoku, Japan?
hospital [3]	Find one or more facilities that are closest to an incident based on travel time	Which hospital is closest to the incident within a 2-minute drive in San Francisco, USA?
malaria [4]	Calculate malaria incidence rate in Democratic Republic of the Congo and understand where prevention and aid are most needed	Which administrative regions have the highest malaria incidence rate in the Democratic Republic of the Congo from 2000 to 2015?
neighborhoods [5]	Find all the neighborhoods in the municipality of Utrecht that are a suitable living area for families with young children (aged <12 years)	Which Utrecht's neighborhoods are within 100 meters from a school and with housing price lower than 200,000 euros in the Netherlands?
solar [6]	Determine the amount of solar radiation received by each rooftop in the Glover Park neighborhood of Washington, D.C. throughout the year	What is the total annual amount of electric power production in MWh for each building's usable area in Glover Park, Washington, DC?

[1] The population workflow was reproduced based on a tutorial provided in a course at Utrecht University

[2] https://learn.arcgis.com/en/projects/estimate-access-to-infrastructure/

[3] https://pro.arcgis.com/en/pro-app/latest/help/analysis/networks/closest-facility-tutorial.htm

[4] https://learn.arcgis.com/en/projects/monitor-malaria-epidemics/

[5] The neighborhoods workflow was reproduced based on a tutorial provided in a course at Utrecht University

[6] https://learn.arcgis.com/en/projects/estimate-solar-power-potential/

Bibliography

Ramesh, G. (2021). Importance and applications of gis in engineering. *Indian Journal of Structure Engineering (IJSE) Volume-1 Issue-1*, 4–8.

Khashoggi, B. F., & Murad, A. (2020). Issues of healthcare planning and gis: A review. *ISPRS International Journal of Geo-Information*, *9*(6), 352.

Azaz, L. (2011). The use of geographic information systems (gis) in business. *Int. Conf. Humanit*, 299–303.

Stankov, U., Durdev, B., Markovic, V., & Arsenovic, D. (2012). Understanding the importance of gis among students of tourism management. *Geographia Technica*, *2*, 68–74.

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review*, *34*(1), 135–174.

Bouziane, A., Bouchiha, D., Doumi, N., & Malki, M. (2015). Question answering systems: Survey and trends. *Procedia Computer Science*, *73*, 366–375.

Kwok, C. C., Etzioni, O., & Weld, D. S. (2001). Scaling question answering to the web. *Proceedings of the 10th international conference on World Wide Web*, 150–161.

Radev, D., Fan, W., Qi, H., Wu, H., & Grewal, A. (2002). Probabilistic question answering on the web. *Proceedings of the 11th international conference on World Wide Web*, 408–419.

Scheider, S., Nyamsuren, E., Kruiger, H., & Xu, H. (2021). Geo-analytical question-answering with gis. *International Journal of Digital Earth*, *14*(1), 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.

Scheider, S., Meerlo, R., Kasalica, V., & Lamprecht, A.-L. (2020). Ontology of core concept data types for answering geo-analytical questions. *Journal of Spatial Information Science*, (20), 167–201.

Xu, H., Nyamsuren, E., & Scheider, S. (2022). Assemble geo-analytical questions through a blockly-based natural language interface. *AGILE: GI-Science Series*, *3*, 69.

Schwitter, R. (2010). Controlled natural languages for knowledge representation. *Coling 2010: Posters*, 1113–1121.

Xu, H., Nyamsuren, E., Scheider, S., & Top, E. (2023). A grammar for interpreting geo-analytical questions as concept transformations. *International Journal of Geographical Information Science*, *37*(2), 276–306.

Kuhn, W., & Ballatore, A. (2015). *Designing a language for spatial computing*. Springer.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information sciences*, *307*, 39–52.

Bille, P. (2005). A survey on tree edit distance and related problems. *Theoretical computer science*, *337*(1-3), 217–239.

Zhang, K., & Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, *18*(6), 1245–1262.

Field, A. (2013). *Discovering statistics using ibm spss statistics*. sage.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in neural information processing systems*, *28*.