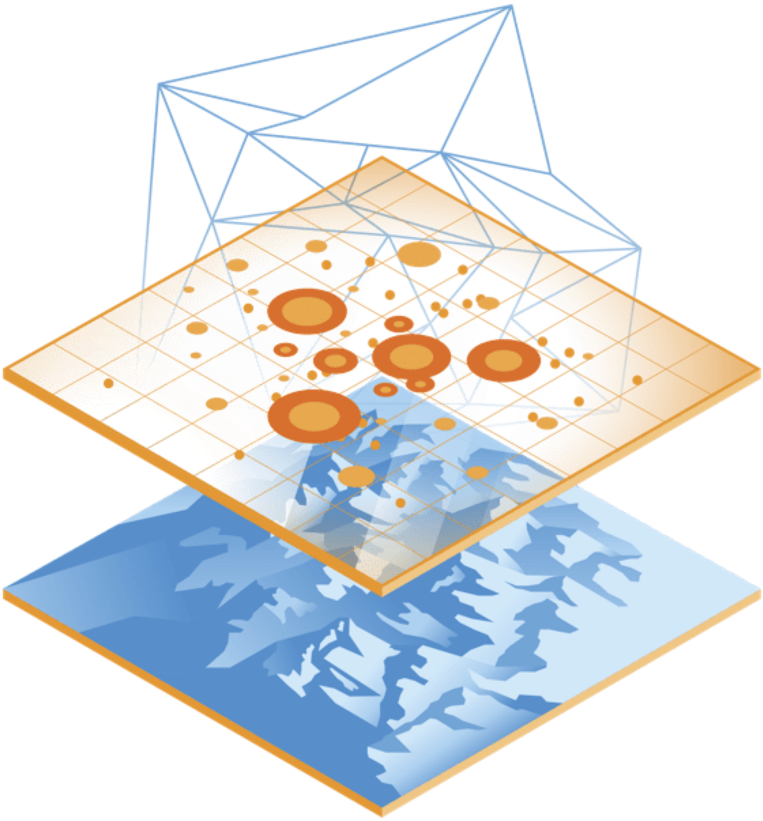
Core concepts in spatial analysis

Exploring for behavioural evidence for cognitive core concepts in spatial analysis



Bachelor thesis - Utrecht University – Faculty of Geosciences July 2023

Student:

Casper A. Prinsen

c.a.prinsen@students.uu.nl

1456474

Supervisor:

Enkhbold Nyamsuren

e.nyamsuren@uu.nl

Contents

Cover image: *ArcGIS Platform Also Provides Access to Spatial Analysis Tools*, 2021

List of tables

List of Figures

1. Introduction

1769 kilometres south of mainland South Africa are two small uninhabited islands called the Prince Edwards Islands. As the islands are free from human activities they are of major interest to researchers from a range of different disciplines. Biologists study local flora and fauna, meteorologists have an active presence monitoring weather patterns and cosmologists monitor solar activity and extraterrestrial radiation (Makoni, 2022). Despite their value to science, spatial data on the islands was not complete which hampered scientific research. This underlines a specific problem, namely that the absence of complete and accurate spatial data limits the capacity for research in other disciplines. In the case of the Prince Edward Islands, this problem was solved with the creation of an open-access geospatial database (Rudolph., Hedding, De Bruyn & Nel, 2022). The author of the article stated: ‘The one data type that all of the sciences on the Prince Edward Islands have as a common need is geospatial data’. (Makoni, 2022).

The example stated above underlines the importance of spatial data for transdisciplinary purposes which is also argued by Kuhn (2012). In his article, he mentions that for major challenges of humanity, such as biodiversity, climate change, energy, and more – spatial information is essential. Spatial information connects different disciplines and also the individuals active within those disciplines. With the proliferation of spatial information, and increasing ease of access as illustrated above, there are increasing opportunities for the advancement of interdisciplinary collaborations. Aside from the growing amounts of spatial information, the rise of Geographic Information Sciences (GIS) in general should be addressed. The possibilities for doing spatial analysis have become far easier as the availability and accessibility of GIS applications have increased (Nivala, Brewster & Sarjakoski, 2008). An interesting example is the increasingly streamlined appearance of programs such as ArcGIS and QGIS. Useful tools such as the model builder allow GIS users to create workflows combining different datasets and analytical operations in a single model

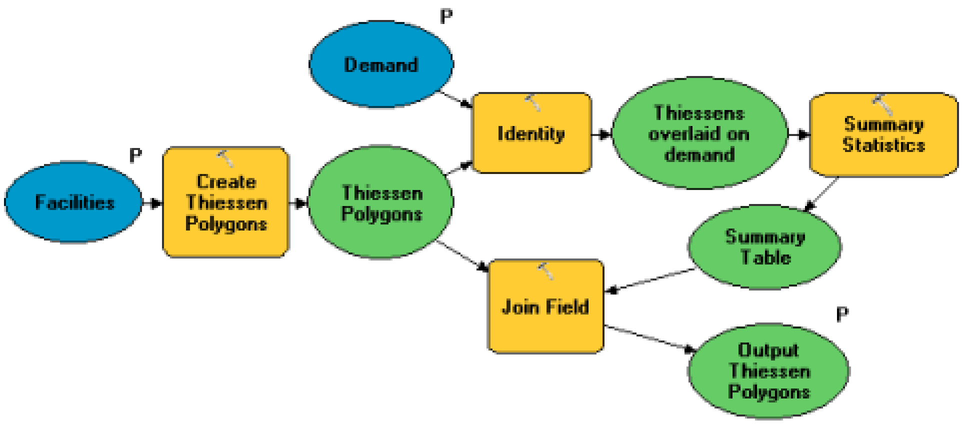


Figure 1: An example of a workflow created in the ArcGIS model builder showing data and the tools used on that data.

(see Figure 1). The creation of such workflows requires a certain level of expertise as understanding the functionalities and parameters of tools is essential (Kruiger, Meerlo, Lamprecht, Nyamsuren & Scheider, 2020). Datasets first need to be interpreted before deciding which tool is useful to reach the envisioned results. Performing appropriate spatial analytical research requires not only an effective GI system but also intelligent decision-making by the GI user. Because of the increasing transdisciplinarity, there is a need to convey expert GIS knowledge to those less familiar in the field. This would maximize the value of GIS technology as GIS can lead to improved insights. To illustrate this with the example of the Prince Edward Islands, a biologist might require spatial data in their research. Instead of intensive studying on how to use and transform the spatial data so it suits their needs, there might be other techniques to effectively convey the knowledge needed for them to reach their goal. Similarly, such techniques could be used in classroom settings to more efficiently convey expert GIS knowledge. First, this would improve decision-making: Efficient and effective conveyance of GIS knowledge can help users make more informed decisions. Secondly, this would help in reducing errors and improve efficiency: without a solid understanding of GIS principles, users may make errors in data collection, analysis and visualization.

There have been several ways of addressing these challenges. A well-known approach are the core concepts of spatial information by Kuhn (2012). Kuhn proposed core concepts that are intended to create a conceptual view of spatial information so it would be meaningful to non-specialists. The concepts are proposed as human cognitive constructions used to interpret data models and algorithms and are therefore meant to be understandable concepts and support broader use of spatial information in society. The current core concepts as by Kuhn & Ballatore (2015) are *location, field, object, network, event, granularity,* and *accuracy.* The concepts will be further explained in the related works and theoretical framework chapters. Core concepts are one of the major theories for spatial information and can be useful as a framework for interpreting maps and guiding geospatial analysis. Kuhn and Ballatore (2015) and later Scheider, Nyamsuren, Kruiger & Xu (2020b) described core concepts as cognitive lenses that can simplify understanding spatial data and guide the processing thereof. Representation of a core concept can take many forms. Realization by a GIS user what core concept is represented can imply which analytical tasks are possible on the given dataset (Scheider, Meerlo, Kasalica & Lamprecht, 2020a; Kruiger et al., 2020). For example, a dataset created by Rudolph et al. (2022) for the Prince Edward Islands is an elevation raster which can be described as belonging to the *field* concept. The reclassify tool is a possible analytical tool that can be applied to a *field* dataset of this geometry type. This tool would not apply to an *object* dataset of similar geometry.

However, empirical evidence of whether core concepts are being used as mental tools during analytical tasks has, apart from a few isolated studies, not been sufficiently found. Social geography student Azani (2022) has done a bachelor's thesis on whether GIS mentally differentiates between core concepts when interpreting maps. Further study was also done by Nyamsuren et al. (2022) where evidence was found that mental analytical skills exist and are used for distinguishing between maps. This is only proven for users with some GIS proficiency and the paper closes with statements that the study should be replicated to verify if this is also the case for users with no proficiency and whether mental tools are also used during spatial analysis.

The research objective of this study is to find empirical evidence of whether GIS users make use of core concepts as mental tools during geospatial analysis. Following this objective is the main research question:

> *To what extent do users apply mental tools that correspond to the core concepts of spatial information when applying analytical tools on visualized datasets?*

The use of core concepts as mental tools to distinguish between datasets has, as stated, already been subject to some research. Their presence as mental tools during spatial analysis has not. To explain core concepts in the context of spatial analysis sub-question *SQ1* was devised (see below). Datasets first need to be interpreted prior to performing any kind of spatial analysis. This leads to *SQ2*. As spatial analysis relies on the source data, and evidence for core concepts when interpreting maps exists, a possible correlation might suggest their presence as mental tools during analysis (*SQ3).* Finally, the level of expertise a user has might influence to what extend mental tools are used. This concern leads to *SQ4.*

- ⇒ *SQ1: What are the core concepts of spatial information in the context of spatial analysis?*
- ⇒ *SQ2: To what extent can users effectively distinguish visualized datasets based on different core concepts?*
- ⇒ *SQ3: To what extent does a correlation exist between effectively differentiating visualized datasets using core concepts and the utilization of analytical tools?*
- ⇒ *SQ4: To what extent is the ability to apply the correct analytical tools dependent on a user's level of expertise?*

This research will be useful for further studies as research on the existence of empirical evidence on whether mental tools based on the core concepts are limited. More verification on the existence of the cognitive use of core concepts can lead to a more widespread understanding of the concepts and therefore a larger transdisciplinary understanding of spatial information.

This thesis will first briefly discuss related work that is of interest for this research. The purpose is to describe and establish a background of the most important concepts in the theoretical framework. Thereafter, the methodology will be described. Outcomings of the survey as described in the methodology will then be described in the results chapter followed by a discussion and conclusion.

2. Related works

The core concepts of spatial information are first described by Kuhn (2012). Kuhn proposed a set of ten core concepts that are intended to be meaningful to scientists who are not specialists in the field of spatial information. The goal of the research was to establish a conceptual view of spatial information which would contribute to the transdisciplinarity of geographic information science. This manifested in ten core concepts Kuhn has divided spatial concepts on one side and information concepts on the other where spatial concepts serve to reason about space and information concepts to reason about spatial information. Important is to note is that information concepts can still possibly be spatial.

Location is the first core concept as spatial information is always linked to location. Kuhn (2012) explains that a location is a relation and not a property. With a relation, it is meant that nothing has an intrinsic location. Location changes based on the context through which it is seen. *Neighbourhood* is the second concept and describes the relation between two locations. In his paper, Kuhn explains this concept with the first law of geography according to Waldo Tobler (1970): ‘Everything is related to everything else, but near things are more related than distant things’. The third concept is *field*. Fields describe phenomena that have a scalar or vector attribute everywhere in a space of interest. *Objects* together with fields form the two fundamental ways of structuring spatial information according to Kuhn. Objects describe individuals that have an identity and are bounded in space. The fifth concept is *network*. Networks describe relations between objects. Kuhn describes network as one of the most recognized and most applied concepts. Questions about change can be answered using the sixth concept *event.* The six concepts mentioned above are the spatial concept. The following four concepts are information concepts starting with *granularity.* Granularity informs about the precision of spatial data, meaning that it characterizes the size of spatial, temporal, and thematic units. The second information concept is *accuracy* which is in regard to the correctness of spatial data. The third is *meaning* answers questions about how to interpret terms used in spatial information. The final core concept is *value* and answers questions about the role that spatial information plays in society. In research done by Kuhn (2015) revisions were made to the original ten core concepts. The amount of core concepts was reduced to seven and were differently structured as seen in Table 1. Spatial concepts also changed to core content concepts and information concepts to core quality concepts.

Table 1: The revised core concepts of spatial information.

Core Content Concepts					Core Quality Concepts
Location	Field	Object	Network	Event	Granularity
					Accuracy

The *location* concept is the most fundamental concept of spatial information (Kuhn, 2015). It describes the position of something which can either be a concrete location or somewhere relative to something. The concept *field* says something about the value of an attribute. An example of a geographic field dataset could be data on noise levels as mentioned in the introduction. A phenomenon (noise levels) can be described by a property with a single value (dB). Other examples could be temperatures or nitrogen concentrations. The concept of *object* answers questions about the properties and relations of objects. To provide some examples: a microphone measuring noise would fall under the object concept. The same goes for a thermometer for temperatures and a nitrogen gauge for nitrogen concentrations. Other possible objects include buildings or lakes. All objects have an identity. This means that properties and relations to other objects are tracked. Objects are also always bounded, meaning that infinite sizes are impossible as objects have boundaries. Important to note that boundaries may not always be clear. Kuhn & Ballatore (2015) describe the *field* and *object* concepts as the two fundamental ways in understanding spatial information. The fourth core content concept is *network*. Network answers questions about the information on connections between objects. An example of what network concept might include could be the shortest path between two locations/objects on a map. The final core content concept is *event*. Event answers questions about what has happened, is happening, or may happen. This concept relies heavily on its temporal aspect. All events are temporally bounded and have an identity the same as object. An important aspect of events as mentioned by Kuhn & Ballatore is that the relation between events and the other core content concepts is that field, object, and networks act as participants in an event. Then there are also two core quality concepts of *granularity* and *accuracy.* Granularity answers questions about the detail of in spatial information. It is a concept that assessed the quality of the data. For example, when creating a map of a neighbourhood it would be nonsensical to include all vegetation in the greatest detail. More likely areas with green are visualised as green parcels. Finally, accuracy answers questions about whether information describes something correctly. Together with granularity, they are indicators for the quality of data. In this research, Kuhn & Ballatore also explained core spatial computations that related to the core content concepts. The operators explained in his paper are basic forms of spatial analysis and therefore interesting to take into account in this research.

With the emergence of the core content data types (CCD) ontology, multiple research has been done on the influence of the core concepts on geo-analytical question answering and workflows. For example, the paper by Scheider et al. (2020b) on geo-analytical question answering with GIS tries to address the scientific challenges of question answering using the core concepts of spatial information. They affirm their goals as ultimately asking a spatial question and instantly finding the right data and analysis. This would make spatial analysis far more accessible for non-expert GIS users and they argue the importance of core concepts for reaching this goal. First, they provide the semantic constraints that capture the analytic potential of tools. Secondly, the concepts are important for the formulation and interpretation of spatial questions. And third, workflows can be constructed by exploiting constraints. Further study done by Scheider et al. (2020a) extends the research by answering the question of what semantic types would be needed to capture the variety of which a core concept can be represented and the implications for geospatial analysis. An important section in this research is about geometric layer types for representing core concepts. They state that in analysis the division between vector and raster geometry types is the most prevalent way in which we distinguish between layers. For this they provide two reasons: dividing between vector and raster might often be irrelevant as there exist trivial translations between the two formats. Second, it is not always relevant whether a dataset is expressed using vector or raster format. In their research, they suggest, that instead of focusing on geometric type, distinguishing between geometric properties is more useful. This includes layers that are tessellation or not, point, line, and region datasets. A simplified overview is shown in Figure 2. These are the sub-concepts to the core concepts of spatial information.

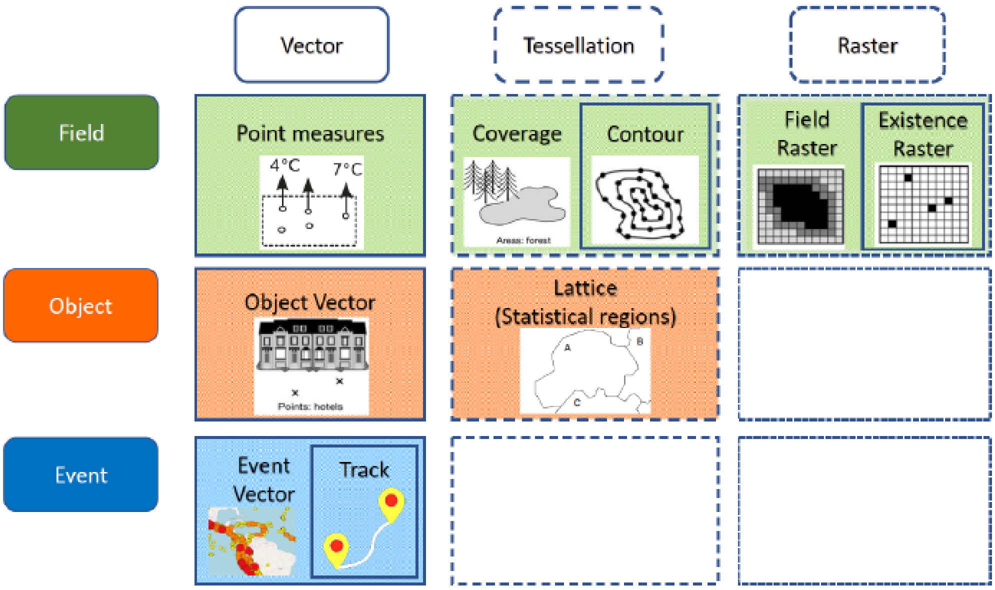


Figure 2: Data types based on the combination of geometric type and core concepts of spatial information. Tessellation overlaps with both vector and raster and are therefore not mutually exclusive (Scheider et al., 2020)

Except from a few isolated studies, little research can be found that verifies the existence of the core concepts as mental tools. The quantitative research by Nyamsuren et al. (2022) tries to find empirical evidence whether the core concepts are cognitive means for interpreting and

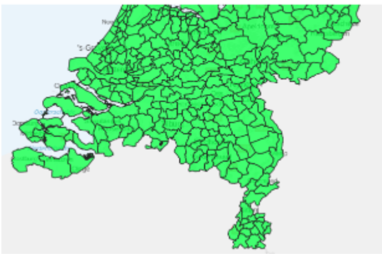
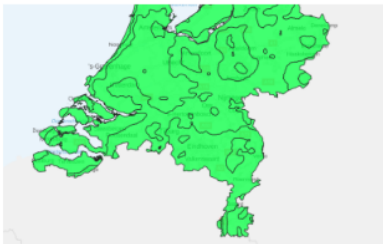
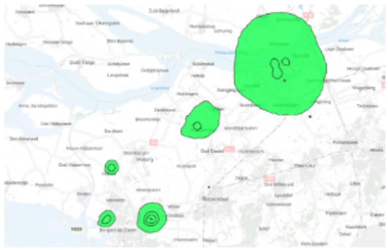
		
Topic: ATES-based heating storage and recovery capacity per municipality; **Geometry:** Vector Tessellation; **Attributes:** Amount of heating in joules (e.g., 350, 1650, 3600)	**Topic:** Windspeed at 100 meter altitude; **Geometry:** Vector Tessellation; **Attribute:** Windspeed (meters per second) range (e.g., 5.5 - 6, 6 - 6.5, 6.5 - 7).	**Topic:** Health impact scores (GES-score) calculated from industry odor emissions; **Geometry:** Vector Tessellation; **Attribute:** GES-score on a scale from 0 to 8 (e.g., 0, 1, 2, 3)

Figure 3: Example contrast question as was shown in the study by Nyamsuren et al. (2020). The leftmost map is heating capacity as tessellated objects. The other two maps visualize wind-speed and GES-score as contour maps.

using maps. In their research they have used the core concepts of field and object and linked these concepts with different geometric types, namely points, lines, lattices and raster. A survey was set up where respondents would see three different types of maps where two of the maps would visualize the same core concept and one would be divergent. The respondent had to choose which of the three maps was the divergent entry (Figure 3). The results of this research provided evidence that supports the hypothesis that GIS users use core concepts as mental analytical tools when interpreting maps. Similar research was done by Azani (2022) for a bachelor's thesis where the research question was whether GIS users mentally differentiate between core concepts. Results are similar to those found by Nyamsuren et al. (2022) showing that GIS users can distinguish between maps using core concepts. In the study done by Azani respondents had to choose one out of three maps that visualized a different core concept. This was done using only visualized point datasets. This was different compared to the study by Nyamsuren et al. (2022) where lines and polygons were also included. Azani chose to only use points because of time and resource constraints. Azani further underlines that visual encoding may have influenced the results. Nyamsuren et al. explains that in the setup of their survey, all maps were given similar symbology to counter this visual encoding.

Other research searching for behavioural evidence regarding the use of core concepts as mental tools does not exist. Aside from the studies by Nyamsuren et al. and Azani there is little empirical evidence for the existence of core concepts as mental tools when distinguishing between maps. The presence of core concepts as mental tools during analysis has not been researched which is what this study aims to accomplish.

3. Theoretical framework

The goal of this study is to find if core concepts are used to interpret which geoanalytical tool can be applied to a visualized dataset. For a user to decide which tools might, or might not be, applicable to reach their analytical goal, they would first need to interpret the visualized dataset (Scheider & Meerlo, 2020). Therefore this study focuses on both dataset- and tool interpretation. Because several analytical tools are included in the study, participants likely need to have some sort of GIS training. Without prior knowledge about geospatial analysis, it would difficult for the participant to first interpret the data source and then envision a meaningful outcome. The study focuses only on the core content concepts of *object* and *field* and only datasets of three different types of geometry are used. From all six core concepts *object* and *field* are the easiest to distinguish between as they are often mutually exclusive. To give an example, the two points datasets in Figure 1 are both similar in their type of geometry but relate to different core concepts. The point object dataset describes swimming pools which are distinct spatial entities that exist as a physical place. The point field dataset describes a measurement that was taken. The measurement is visualized as a point on the location where the measurement was done but only exists as an attribute of that point.

In this study, there was opted for using only three types of geometry. These are points, regions, and raster/squared lattices as shown in Figure 1..

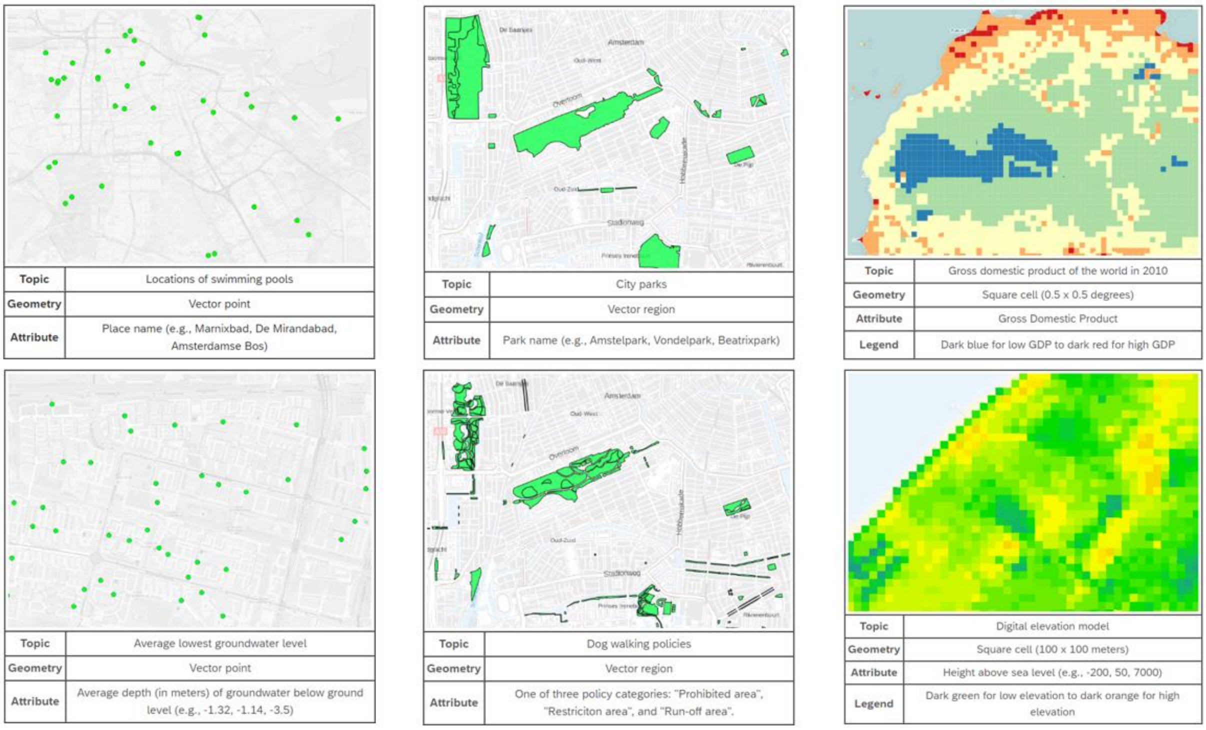


Figure 4: A collection of the three different geometry types that are used for each core concept. The top row shows three visualized object datasets that are used in the study. The bottom row shows three visualized field datasets. The columns are ordered by geometric characteristics. These are points, regions, and raster/squared lattice from left to right.

Other geometries such as lines and tessellated polygons, which are used in the study by Nyamsuren et al. (2022), have been omitted. To include these geometries where for each geometry type there would also be a need for *object* and *field* datasets, would be out of the scope of this thesis research.

The core concepts can also be interpreted for analytical tools after the data source has been interpreted. Scheider & Meerlo (2020) give the example of an analyst interpreting a data source as related to the concept of *field* and therefore search for tools that can handle a *field* representation. An example would be the IDW interpolation tool that is meaningful to calculate possible measurements on unknown locations which you could for instance apply on the ‘average lowest groundwater level’ dataset in Figure 4. Applying this tool to the ‘locations of swimming pools’ dataset would yield no meaningful results. Conversely, some tools are only meaningfully applicable to object data sources such as an Euclidian distance operation. Calculating the distance to the nearest swimming pool is a meaningful operation that yields a useful outcome. Applying the same tool on a *field* dataset such as the groundwater would give an outcome, namely the distance to the location where the measurement was taken, but this is not a meaningful analytical goal. Therefore geoanalytical tools, the same as data sources, can be interpreted using core concepts and also distinguished between.

To find whether a participant can distinguish between core concepts when interpreting a visualized dataset, the same contrast model was adopted as was used in the study by Nyamsuren et al. (2022) The participant is shown a set of three visualized datasets of similar geometry (see Figure 5). Besides the map, other information is also given such as the topic, geometry, and attribute of the dataset. One of the three datasets is different in terms of the related core concept. The participant is asked to identify which of the three maps is different

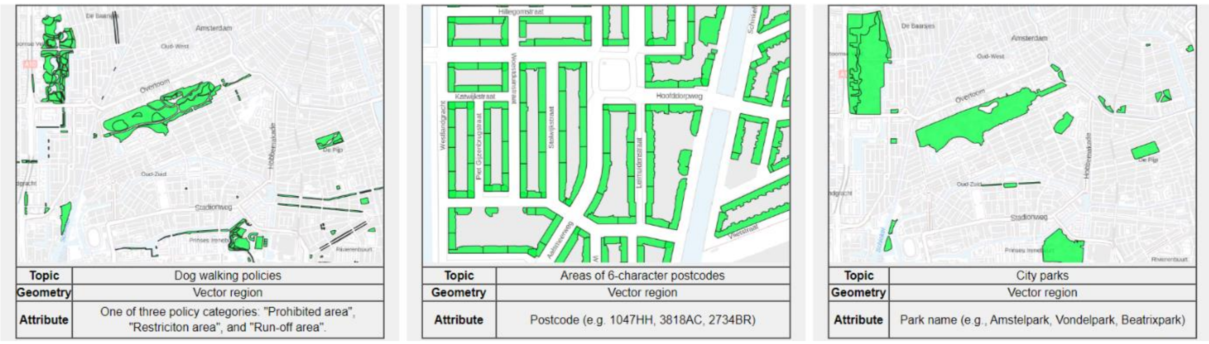


Figure 5: An example of a triplet question as they appear in Q1. The image on the left is a region field dataset. The other two maps are both region object datasets

in terms of operations that can be meaningfully applied to them. With meaningful it is meant that the results should give a conceptually valid output. During these questions, participants will be asked this question for two out of the three geometry types, where for each type the odd map is in one case a *field* dataset and in the other an *object* dataset. For a participant to distinguish the odd map it is hypothesized that they need to rely on some sort of spatial cognitive concept. Because participants are asked to identify different core concepts for different geometries it can be assumed that an accuracy, which is significantly above the chance level, would mean that the participant effectively uses core concepts as a strategy for determining the different dataset. With significantly above chance level it is meant that the participant has a higher accuracy than using random choice strategies. participant has a higher

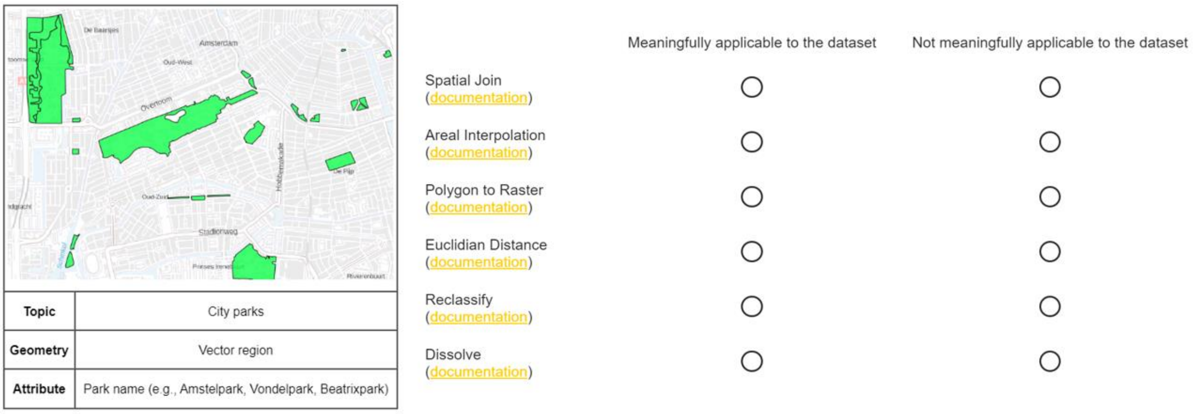


Figure 6: An example of a question as they appear in Q2. Six tools are shown for each dataset which are either meaningfully applicable or not meaningfully applicable.

accuracy than using random choice strategies.

To determine whether a participant also uses core concepts when interpreting the participant is shown a dataset that it had previously seen during the dataset interpretation questions. For each dataset, six geoanalytical tools are shown (see Figure 6). The tools chosen for this study are from the ArcGIS software by ESRI, being one of the most well-known GIS software on the market and with similar tools existing in other software. In total nine tools were chosen for this study where for each geometry/core-concept combination a uniquely applicable tool exists. Alongside these tools, several dummies were also added to the pool of tools.

With each dataset, the participant has to decide for six shown tools whether a specific tool is meaningfully applicable to the dataset. At least two of the shown tools are only applicable to either an *object* or *field* dataset. From these answers, it can be concluded whether a participant uses mental tools when interpreting analytical tools. Besides these two tools, two other tools are shown which could be either applicable, not applicable or both (see Table 2). If only the four tools were shown with each dataset for each datatype, both object and field datasets with the same geometry would have four similar tools which could hint at underlying differences per geometry type. Participants could then use other strategies to decide on applicable tools. Therefore, besides the four tools shown in Table 2, two dummy tools are also added which might not apply at all to the respective geometry type. Which tool is applicable to which datatype is shown in Table 3. This structure was chosen over showing all nine tools for each dataset as it would make the survey very lengthy and possibly lead to participants deciding to quit the survey. Because of this design two different analyses can be carried out, where in the first analysis all tools are included.

Table 2: Tools that were always included in the question for each geometry type

	Points	Region	Raster/Squared lattice
Tools	Euclidian Distance	Euclidian Distance	Zonal Statistics
	Create Thiessen Polygon	Dissolve	Reclassify
	IDW Interpolation	Spatial Join	Areal Interpolation
	Spatial Join	Polygon to Raster	Polygon to Raster

Table 3: Meaningfully applicable tools per datatype

	Point Object	Point Field	Region Object	Region Field	Squared Lattice Object	Raster Field
Euclidian Distance	X		X	X		
IDW Interpolation		X				
Create Thiessen Polygons	X	X				
Zonal Statistics						X
Reclassify						X
Areal Interpolation					X	
Polygon to Raster			X	X	X	
Dissolve				X	X	
Spatial Join	X		X			

The questions relating to determining whether a participant can effectively distinguish between maps related to different core concepts and when interpreting tools are similarly phrased. As previously stated to interpret and apply tools, the data source first needs to be interpreted. When a participant obtains high accuracy scores when interpreting the datasets they are therefore making effective use of mental skills to differentiate between core concepts which supposedly means that they will obtain high accuracy scores when interpreting tools. It is therefore hypothesized that there is a potential correlation between scoring a high accuracy when distinguishing datasets based on core concepts and applying analytical tools on the same datasets.

Other factors could also influence whether a participant gives the correct answers in both sections of the survey. A participant with five years of experience in the field of GIS is most likely more familiar with performing geospatial analysis than someone who has just finished an introductory course. For this reason, the participant with more experience will likely have higher accuracies in the second section of the survey. This could be because they have used the tools used in the survey numerous times or are already acquainted with the core concepts. Even when a user is not acquainted with the core concepts it could also be the case that when someone is more proficient with GIS they are more successful in this study because they are more efficient in applying mental tools related to the core concepts. It is therefore important to additionally take the level of expertise into account during the analysis.

4. Methodology

The goal of this research is to find to what extent users apply core concepts as mental skills when performing geoanalytical tasks. To figure out this extent the relation between distinguishing between core concepts and performing analytical tasks needs to be studied. Quantitative methods are best suited for finding relations as this research aims to produce generalizable knowledge about the application of core concepts as mental tools during geoanalysis.

4.1 Datasets

The datasets that were used to create the maps in the survey are shown in Table 4. The table shows the source of the data and a corresponding dataset ID which is used to identify datasets in the following sections of this thesis. The datasets are a selection of the datasets used in the study by Nyamsuren et al. (2022). Finding new datasets for this research would be too time-consuming for the duration of writing this thesis. Also because the datasets were already collected and visualized as maps, building on this provided setup guarantees uniformity between datasets therefore eliminating any inconsistencies that might arise.

4.2 Survey Design in Qualtrics

For the research design, there was chosen to distribute participants across three different blocks. This structure was chosen over a single block because of the three different geometry types that are included. If a participant would need to interpret and analyse all the datasets they would have to do so for 12 different datasets. To keep the participants engaged in the survey but still include all the datasets a design was chosen where in each block the participants saw only the datasets for two different geometry types. As for each geometry type, there are four datasets, the participant is asked to interpret and analyse eight datasets. With four datasets, four different questions can be formulated with each having a different odd dataset during a contrast map question. Table 5 illustrates how these questions are distributed across the three blocks. This distribution makes sure that for each dataset there is a question where it is the odd dataset.

Table 4: Datasets, corresponding Dataset ID, source, and provided description

Dataset	Dataset ID	Source	Description
Locations of swimming pools	PO1	Maps Amsterdam	Geometry: Vector point Attribute: Place name (e.g., Marnixbad, De Mirandabad, Amsterdamse Bos
Nitrogen dioxide (NO2) sensors	PO2	Maps Amsterdam	Geometry: Vector point Attribute: Unique sensor code (e.g., CP27, TK10, T2)
Nitrogen dioxide (NO2) sensing	PF1	Nationaal georegister	Geometry: Vector point Attribute: Amount (micrograms per cubic meter) of NO2 (e.g., 31.394, 39.22, 40.311)
Average lowest groundwater level	PF2	Maps Amsterdam	Geometry: Vector point Attribute: Average depth (in meters) of groundwater below ground level (e.g., -1.32, -1.14, -3.5)
City Parks	ReO1	Maps Amsterdam	Geometry: Vector region Attribute: Park name (e.g., Amstelpark, Vondelpark, Beatrixpark)
Areas of 6-character postcodes	ReO2	Maps Amsterdam	Geometry: Vector region Attribute: Postcode (e.g. 1047HH, 3818AC, 2734BR)
Function mix for built areas	ReF1	Maps Amsterdam	Geometry: Vector region Attribute: One of seven mix categories: "living only", "working only", "facility only", "living and facility", "living and working", "working and facility", "living, working and facility".
Dog walking policies	ReF2	Maps Amsterdam	Geometry: Vector region Attribute: One of three policy categories: "Prohibited area", "Restriction area", and "Run-off area".
Gross Domestic Product of the world 2010	SLO1	NIES	Geometry: square cell (0.5 x 0.5 degrees) Attribute: Gross Domestic Product Legend: Dark blue for low GDP to dark red for high GDP
Population at residences place	SLO2	PDOK	Geometry: Square cell (1 x 1 kilometres) Attribute: Number of inhabitants (e.g. 2, 87, 1621) Legend: Darker colours indicate higher numbers
Digital elevation model	RaF1	PDOK	Geometry: Square cell (100 x 100 meters) Attribute: Height above sea level (e.g. -200, 50, 7000) Legend: Dark green for low elevation to dark orange for high elevation
Areas with potential heat stress during summer days	RaF2	Nationaal georegister	Geometry: Square cell (granularity not known) Attribute: Sensitivity score (a value between 0 and 10) Legend: The red areas are very sensitive to heat stress, the blue areas hardly

Before progressing with the questions in the survey the participant was asked for consent in participating in the survey. If the participant agreed to partake in the survey an information window would show describing the contrast questions (Q1). In Q1 the participant would see four of the contrast questions (Figure 5). With a contrast map question, three maps are shown of similar geometry type. The participant is asked which of the three maps is different in terms of operations that can be meaningfully applied to them. When the participant has answered all four questions another information window would show explaining the second part of the survey (Q2). In Q2 the participant would see all the datasets that were previously shown during the contrast map questions (Figure 6). With each dataset, six tools are shown with two clickable options. The participant is asked to select for each tool whether the specific tool can be either meaningfully applied or not meaningfully applied to the dataset that is shown. The participant would see eight different datasets with every six tools. In total the

Table 5: Table with the distribution of datasets in Q1

Question ID	Odd dataset	Contrast dataset 1	Contrast dataset 2
B1_Q1_1	SLO1	RaF1	RaF2
B1_Q1_2	RaF1	SLO1	SLO2
B1_Q1_3	PO1	PF1	PF2
B1_Q1_4	PF1	PO1	PO2
B2_Q1_1	SLO2	RaF1	RaF2
B2_Q1_2	RaF2	SLO1	SLO2
B2_Q1_3	ReO2	ReF1	ReF2
B2_Q1_4	ReF2	ReO1	ReO2
B3_Q1_1	ReO1	ReF1	ReF2
B3_Q1_2	ReF1	ReO1	ReO2
B3_Q1_3	PO2	PF1	PF2
B3_Q1_4	PF2	PO1	PO2

participant has to decide 48 times whether a tool is meaningfully- or not meaningfully applicable.

Before progressing with the questions in the survey the participant was asked for consent in participating in the survey. If the participant agreed to partake in the survey an information window would show describing the contrast questions (Q1). In Q1 the participant would see four of the contrast questions (Figure 5). With a contrast map question, three maps are shown of similar geometry type. The participant is asked which of the three maps is different in terms of operations that can be meaningfully applied to them. When the participant has answered all four questions another information window would show explaining the second part of the survey (Q2). In Q2 the participant would see all the datasets that were previously shown during the contrast map questions (Figure 6). With each dataset, six tools are shown with two clickable options. The participant is asked to select for each tool whether the specific tool can be either meaningfully applied or not meaningfully applied to the dataset that is shown. The participant would see eight different datasets with every six tools. In total the participant has to decide 48 times whether a tool is meaningfully- or not meaningfully applicable.

At the end of the survey, the participants are asked to categorize their level of expertise with GIS. The four options that can be chosen are: ‘Laymen: never used GIS, Beginner: can use basic GIS functions, Trained: formally trained by a GIS course, Expert: used GIS for 5 years or more’. They were also asked for their familiarity with the core concepts as proposed by Kuhn and for their familiarity with each of the tools. Data gathered about familiarity with tools was not used in the final analysis and only existed in the survey because of early design intentions.

The survey was distributed using a variety of methods. Firstly among peers through two WhatsApp groups. One was a group of students who followed the Geo-Information minor provided by the VU, the other consisted of students from the human geography and planning Utrecht bachelor program. An E-mail was also sent out to the entire department of the program. Furthermore, the survey was distributed among employees of the municipality of Amersfoort at the GIS department. Additionally, an anonymous link was posted on the ESRI forums and several GIS communities on Reddit such as r/gis, r/ArcGIS, and r/cartography.

4.3 Statistical analysis

The data that was gathered using the survey described above would be subject to several statistical tests. First, regarding the second sub-question whether participants can effectively distinguish between visualized datasets applying an exact binomial test should indicate whether respondents are significantly better than randomly guessing. This test is useful for this specific case as the percentage of success when randomly guessing is known for both Q1 and Q2 (33% and 50% respectively). With the exact binomial test, we can test if the observed results significantly differ from the random chance strategy. The test is also fitting because it is expect that the analysis will be carried out over a small sample size.

Second, regarding the third sub-question whether a correlation exists between effectively distinguishing datasets and correctly applying analytical tools a Pearson's R correlation test is used.

Finally regarding the fourth sub-question about the influence of the users' level of expertise on effectively applying analytical tools on visualized datasets a Kruskal-Wallis test is used to compare the samples of two levels of expertise. The samples are created by combining two levels of expertise. Layman and Beginner expertise levels are aggregated in a sample called beginner. This sample has not been formally trained by GIS. The second sample is created by aggregating Skilled and Expert expertise levels. Participants in this sample have been officially trained by a course and/or have over 5 years or more of experience. Applying a Kruskal-Wallis test is fitting for research as it does not require groups to be normally distributed which will probably be the case for this research.

The statistical tests were carried out using the programming language R and are displayed in the results section along with general descriptive statistics.

5. Results

In total 74 participants agreed to the terms on the opening page. Out of these 74 participants, only 30 completely answered the survey. Most of the 44 other participants decided to quit before answering the first question. Because the survey requires a certain level of understanding regarding GIS, participants most likely left the survey immediately after reading the first question as they would not have fully comprehended it.

5.1 Descriptive statistics

Because of randomization when assigning participants to blocks and participants choosing to leave the survey after agreeing to the terms, there is an uneven distribution of participants across blocks. As expertise level was not taken into account before distribution there is also an uneven spread of expertise levels. Block 1 has 11 participants (3 beginners, 8 skilled), block 2 has 7 participants (4 beginners, 3 skilled), and block 3 has 12 participants (2 beginners, 10 skilled). As stated in the methodology section the blocks only exist because of the relatively large amount of datasets and are not subject to statistical tests but do influence finding significant results for the Q1 section.

Figure 7 shows the percentage of correct answers for each question in Q1 for all respondents regardless of expertise levels. B1_Q1_1 has the lowest accuracy score with 27% (*SD* = 13.4), which is lower than the 33% random guess strategy. Figure 7 shows that B1_Q1_1 and B1_Q1_2 have much lower accuracy scores than other questions. They are contrast map questions making use of the raster and squared lattice geometry types.

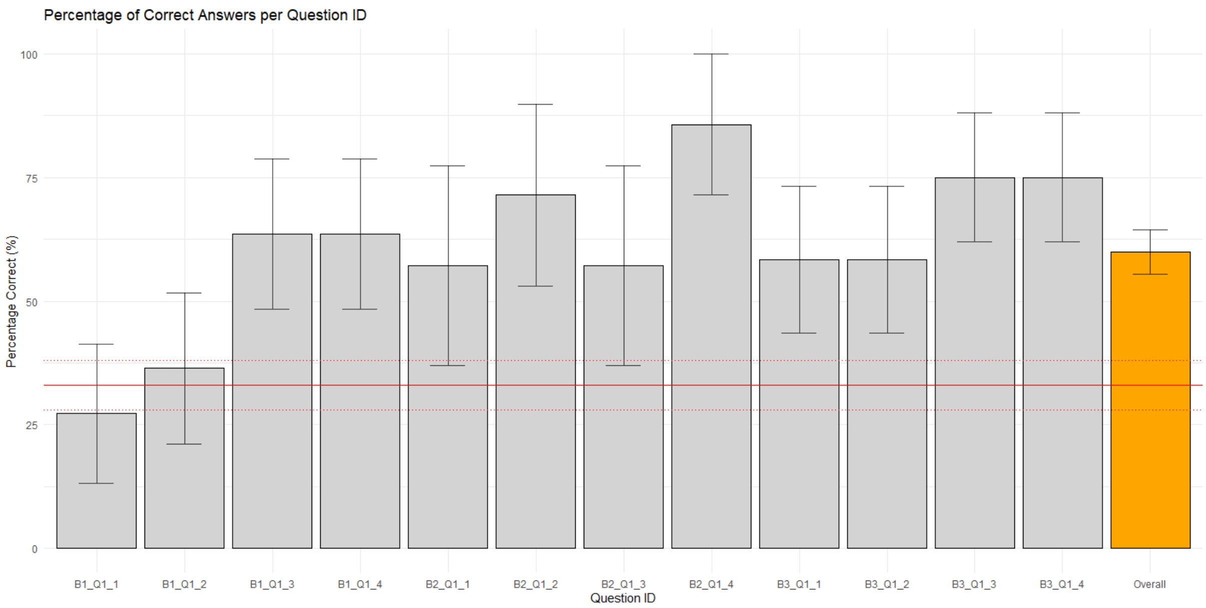


Figure 7: Bar graph of the percentage of correct answers for each question in Q1. Standard error bars are included as well as the accuracy across all Q1 questions. The red line depicts the random chance strategy of 33%.

The influence of geometry will be further discussed in the discussion section. B2_Q1_4 has the highest accuracy score with 86% (*SD* = 13.2) (see Appendix A for specific question details). Calculating the percentage of correct answers for all questions gives an accuracy score of 60% (*SD* = 4.5). This is again higher than the percentage that is expected when participants use a random guess strategy.

Figure 8 shows the percentage of correct answers for the Q2 section of the survey. Except for the squared lattice geometry type datasets, the mean accuracies are all above the random guess strategy of 50%. Participants obtained the highest accuracy score for selecting the correct meaningfully applicable tools for the PO2 dataset with 74.6% (*SD* = 3.7). SLO1 has the lowest accuracy score with 36.1% (*SD* = 4.6). Overall participants had an accuracy score of 61.0% (*SD* = 1.3)

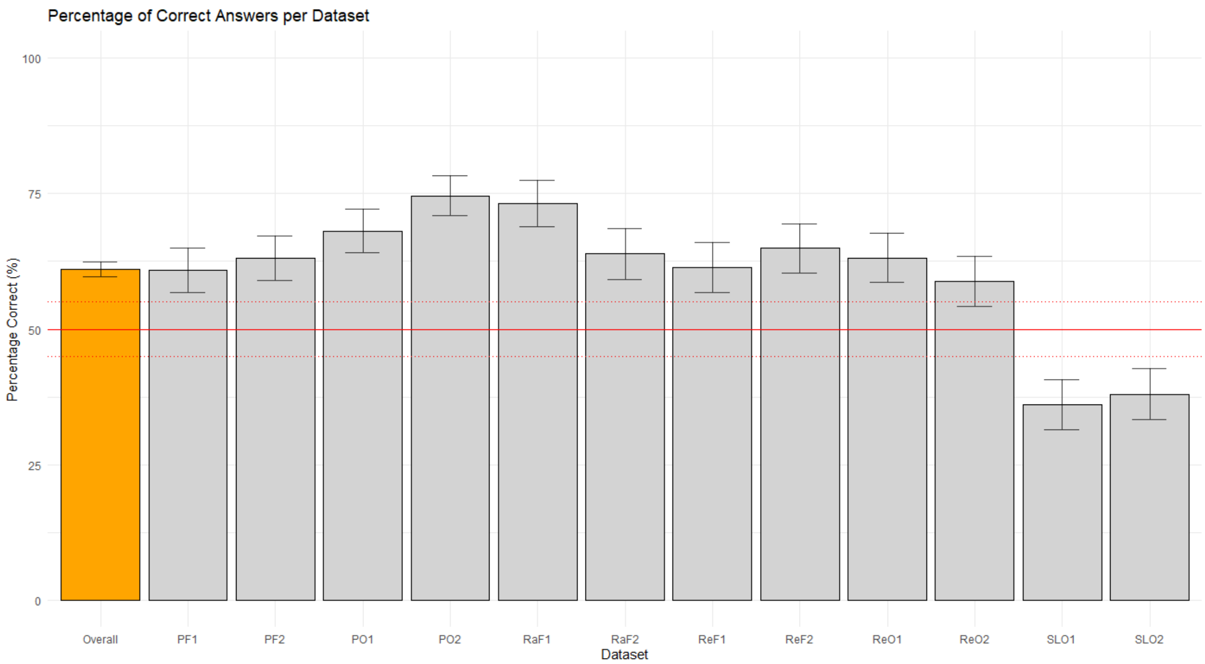


Figure 8: Bar graph of the percentage of correct answers for each dataset shown in Q2. Standard error bars are included as well as the overall accuracy scores of all datasets. The red line depicts the random chance strategy of 50%.

5.2 Exact binomial tests

Using exact binomial tests the accuracies of both Q1 and Q2 grouped by expertise class are tested against a 33% and 50% chance rate respectively. The result for the beginner class in Q1 is 50.0% (N = 9, p = 0.03). The result for the skilled class is 64.3% (N = 21, p < .001). Both results are significant, (the skilled class is extremely significant while the beginner class is only less than 0.05) therefore the null hypothesis is rejected. Participants in both expertise classes score significantly better than when making use of the random guess strategy. Tables 6 - 8 provide more detailed tests for each specific question grouped by expertise levels for each block. Because of low sample sizes, a column showing the combined classes of beginner and skilled expertise levels is also added.

Table 6: Exact binomial test for questions in block 1 grouped by user level of expertise. Green cells indicate significant results with darker green being more significant.

	Beginner (N = 3) Correct (*p-value*)	Skilled (N = 8) Correct (*p-value*)	Total (N = 11) Correct (*p-value*)
B1_Q1_1	0 (.555)	3 (.724)	3 (1.000)
B1_Q1_2	0 (.555)	4 (.452)	4 (.759)
B1_Q1_3	0 (.555)	7 (.003)	7 (.049)
B1_Q1_4	1 (1.000)	6 (.019)	7 (.049)

Table 7: Exact binomial test for questions in block 2 grouped by user level of expertise. Green cells indicate significant results with darker green being more significant.

	Beginner (*N* = 4) Correct (*p-value*)	Skilled (*N* = 3) Correct (*p-value*)	Total (*N* = 7) Correct (*p-value*)
B2_Q1_1	3 (.109)	1 (1.000)	4 (.228)
B2_Q1_2	3 (.109)	2 (.255)	5 (0.04)
B2_Q1_3	2 (.603)	2 (.255)	4 (.228)
B2_Q1_4	4 (.012)	2 (.255)	6 (.006)

Table 8: Exact binomial test for questions in block 3 grouped by user level of expertise. Green cells indicate significant results with darker green being more significant.

	Beginner (*N* = 2) Correct (*p-value*)	Skilled (*N* = 10) Correct (*p-value*)	Total (*N* = 12) Correct (*p-value*)
B3_Q1_1	1 (.551)	6 (.092)	7 (.071)
B3_Q1_2	1 (.551)	6 (.092)	7 (.071)
B3_Q1_3	1 (.551)	8 (.003)	9 (.004)
B3_Q1_4	2 (.109)	7 (.019)	9 (.004)

Table 9: Results of exact binomial tests for each dataset grouped by user level of expertise against the 50% random guess strategy. Green depicts significant effective strategies used by participants. Red depicts significant inefficient strategies used by participants. Darker colour indicate a high level of significance.

	Beginner (*N* = 9) Correct (*p-value*)	Skilled (*N* = 21) Correct (*p-value*)	Total (*N* = 30) Correct (*p-value*)
PO1	19 (.200)	75 (<.001)	94 (<.001)
PO2	22 (.016)	81 (<.001)	103 (<.001)
PF1	12 (.362)	72 (<.001)	84 (.013)
PF2	13 (.585)	74 (<.001)	87 (.003)
ReO1	22 (.243)	50 (.017)	72 (.007)
ReO2	21 (.405)	46 (.141)	67 (.075)
ReF1	24 (.065)	46 (.141)	70 (.019)
ReF2	23 (.132)	51 (.009)	74 (.006)
SLO1	15 (.088)	24 (.036)	39 (<.001)
SLO2	18 (.441)	23 (.019)	41 (.016)
RaF1	26 (.164)	53 (<.001)	79 (<.001)
RaF2	26 (.164)	43 (.017)	69 (.005)

The results for the exact binomial test for the beginner class in Q2 against a success rate of 50% is 55.8% ($N = 21$, $p = .018$). For the skilled class, the result is 63.3% ($N = 21$, $p < .001$). The results of the exact binomial test show that both expertise classes are significantly different than the random guess success rate. Table 9 shows the exact binomial test for each unique dataset grouped by expertise classes. For most datasets participants make use of effective strategies to apply meaningful tools. For the squared lattice datasets, SLO1 and SLO2 participants make use of ineffective strategies, and have accuracies significantly below the random guess strategy.

5.3 Pearson's r correlation

To determine whether GIS users make use of mental tools during geospatial analysis there should be a correlation between effectively distinguishing datasets based on their corresponding core concept and applying the appropriate analytical tool. Therefore a Pearson correlation coefficient was computed to assess the relationship between the two. There is a moderate positive correlation between the two variables, $r(28) = .47$, $p = .008$. Results are significant for the $p < .01$ level.

5.4 Kruskal-Wallis test

The Kruskal-Wallis test was conducted for both Q1 and Q2 to compare the number of correct answers between the Beginner and Skilled groups. The test revealed a significant difference in the number of correct answers across the groups for Q1: $H(1) = 6.003$, $p = .014$. For Q2 the test also revealed a significant difference: $H(1) = 5.062$, $p = .024$. Post-hoc analysis using Dunn's test was conducted to determine specific pairwise differences in the number of correct answers between the Beginner and Skilled groups.

The analysis revealed a significant difference in the number of correct answers between the Beginner and Skilled groups for Q1 ($p = .014$, *adjusted* $p = .014$) and Q2 ($p = .024$, *adjusted* $p = .024$). The pairwise comparisons showed that the Skilled group had a significantly higher number of correct answers compared to the Beginner group ($p < .05$).

6. Discussion

Previous studies done by Nyamsuren et al. (2022) and Azani (2022) found that two strategies are most commonly used when distinguishing between maps. One strategy is based on visual cues, the other is based on the attributes of those datasets. Distinguishing between datasets based on attributes requires a more thoughtful approach and is therefore considered to be a more advanced technique. The results also prove this as skilled users have significantly higher accuracy scores than beginners. Important to note that beginners were still more effective than applying a random guess strategy suggesting that even those with no or little prior knowledge in GIS can still effectively distinguish maps and apply tools to some extent.

Because of the small sample sizes, it was difficult to gain accurate results, mainly for the section regarding distinguishing between maps. Figure 9 shows the same results as Figure 7 but this time grouped by user level of expertise. It could also be deducted from the binomial test but the visualized data paints a clearer picture. Beginners had very low accuracy scores for the first questions while for others every beginner was correct. The graph is misleading because of the distribution of participants across the blocks. Increasing the number of participants would be useful for this study as including the graphs grouped by expertise would support the research questions more clearly. Because the distribution of participants across blocks was less of an issue for Q2, Figure 10 depicting the percentage of correct answers grouped by user level of expertise is more interesting. There are several datasets that have large discrepancies between the percentage of correct answers. Beginners struggled with the point field datasets compared to the skilled users.

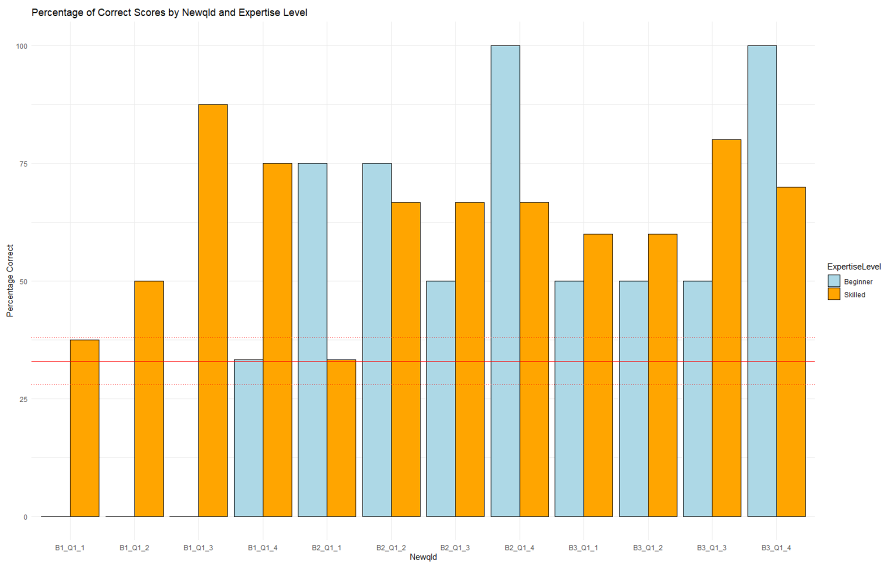


Figure 9: Bar graph of percentage of correct answers for question in Q1 grouped by user level of expertise. The red line depicts the random guess strategy of 33%

The SLO1 and SLO2 datasets were difficult for both beginner and skilled users to analyse, both scoring significantly below the random guess strategy threshold. Possibly participant obtain these low accuracies as it is harder for them to distinguish between the squared lattice object datatype with the raster field. This was also found when participants were asked to distinguish datasets based on these geometry types. The influence of geometry types on the use of core concepts as mental tools was not further studies in this research.

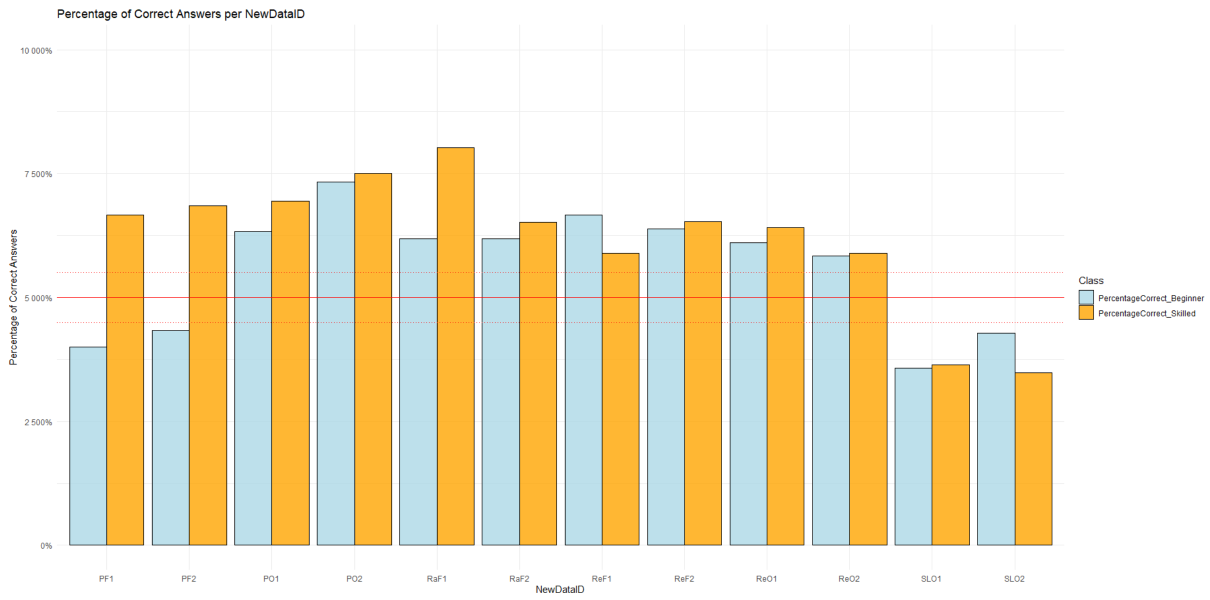


Figure 10: Bar graph of percentage of correct answers per dataset (Q2) grouped by user level of expertise. The red line depicts the random guess strategy of 50%.

7. Conclusion

The results show that participants score significantly higher than randomly guessing correct answers when distinguishing between maps that relate to different core concepts and when applying analytical tools on those datasets. This supports the second sub-question and reconfirms the findings of the study by Nyamsuren et al. (2022). Pearson's r test suggests that higher accuracy scores when distinguishing between maps significantly correlate to higher accuracy when choosing which tools are meaningfully applicable to the visualized dataset (supports the third sub-question). The user level of expertise was found to influence both effective map interpretation and analysis. Significant differences were found between the two levels of expertise, with beginners being less effective then skilled users (support the fourth sub-question).

Overall the results support the existence of core concepts as mental tools during geospatial analysis. Even participants with little or no training performed significantly higher then randomly guessing which might suggest they are using effective strategies. Because of small sample sizes this can not be significantly concluded and follow up research with more participants is recommended.

References

ArcGIS Platform also provides access to spatial analysis tools. (2021, March 10). Esri. https://www.esri.com/about/newsroom/arcwatch/esri-introduces-arcgis-platform/

Azani, A. (2022). *Exploring behavioural evidence of cognitive core concepts in spatial analysis* [Bachelor's thesis]. Utrecht University.

Janelle, D. G., & Goodchild, M. F. (2011). Concepts, Principles, Tools, and Challenges in Spatially Integrated Social Science. *SAGE Publications, Inc. eBooks*, 26–45. https://doi.org/10.4135/9781446201046.n2

Kruiger, J. F., Kasalica, V., Meerlo, R. J., Lamprecht, A., Nyamsuren, E., & Scheider, S. (2021). Loose programming of GIS workflows with geo-analytical concepts. *Transactions in Gis*, *25*(1), 424–449. https://doi.org/10.1111/tgis.12692

Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, *26*(12), 2267–2276. https://doi.org/10.1080/13658816.2012.722637

Kuhn, W., & Ballatore, A. (2015). Designing a Language for Spatial Computing. *Lecture Notes in Geoinformation and Cartography*, 309–326. https://doi.org/10.1007/978-3-319-16787-9_18

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic Information Science and Systems*. Wiley.

Makoni, M., & Makoni, M. (2022). Geospatial Database for the Prince Edward Islands. *Eos*. https://eos.org/articles/geospatial-database-for-the-prince-edward-islands

Nivala, A., Brewster, S., & Sarjakoski, T. (2008). Usability Evaluation of Web Mapping Sites. *Cartographic Journal*, *45*(2), 129–138. https://doi.org/10.1179/174327708x305120

Nyamsuren, E., Top, E. J., Xu, H., Steenbergen, N., & Scheider, S. (2022). Empirical Evidence for Concepts of Spatial Information as Cognitive Means for interpreting and using Maps. *15th International Conference on Spatial Information Theory (COSIT 2022)*, *240*, 7:1-7:14. https://doi.org/10.4230/LIPIcs.COSIT.2022.7

Rudolph, E. M., Hedding, D. W., De Bruyn, N., & Nel, W. (2022). An open access geospatial database for the sub-Antarctic Prince Edward Islands. *South African Journal of Science*, *118*(9/10). https://doi.org/10.17159/sajs.2022/12302

Scheider, S., Meerlo, R. J., Kasalica, V., & Lamprecht, A. (2020a). Ontology of core concept data types for answering geo-analytical questions. *Journal of Spatial Information Science*, *20*. https://doi.org/10.5311/josis.2020.20.555

Scheider, S., Nyamsuren, E., Kruiger, H., & Xu, H. (2020b). Geo-analytical question-answering with GIS. *International Journal of Digital Earth*, *14*(1), 1–14. https://doi.org/10.1080/17538947.2020.1738568

Ye, H., Brown, M. E., & Harding, J. A. (2014). GIS for All: Exploring the Barriers and Opportunities for Underexploited GIS Applications. *OSGeo Journal*, *13*(1), 19–28. https://doi.org/10.5446/1554

Appendix A: Survey

Utrecht University

Block_Introduction

Welcome to the survey for my bachelor thesis research project. For this bachelor thesis I am conduc evidence on whether Geographic Information System (GIS) users apply cognitive tools during geos

This survey consists of two sections. The first section consists of 4 questions. The second sections estimated to take around 7 - 12 minutes.

If you have any questions about the survey, feel free to send an e-mail.

Thank you in advance for your participation.

By submitting this form you agree to the following: All answers will be collected for use in the researc your personal information and your answers closely. All the data will be stored safely.

- Yes, I consent in participating in the study
- No, I do not consent in participating in the study

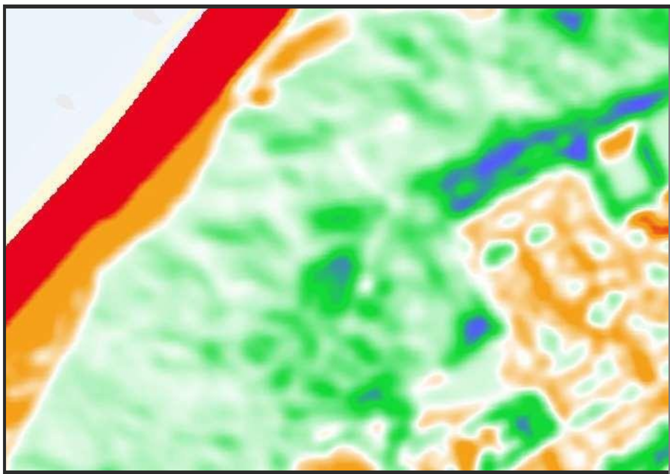
B1_Q1

In the following questions, you are asked to consider three datasets. For each question, one of the t other two in terms of the analytical operations that can be meaningfully applied to them. With meani should give a conceptually valid output. Your task is to select the dataset that is most different from t

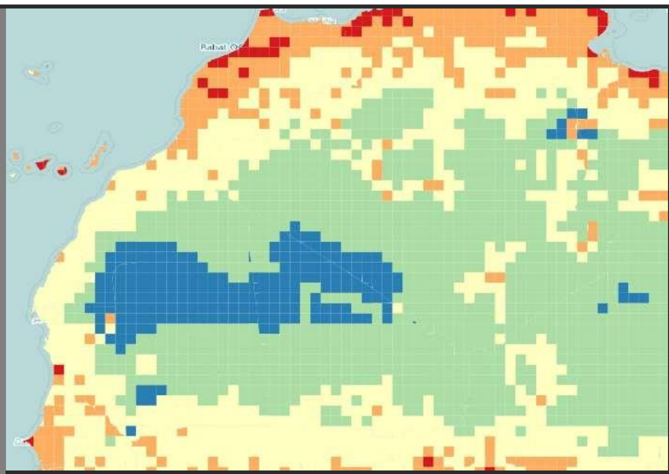
For each dataset, a **map** and information about the **topic**, the **geometry**, the **attribute,** and optiona

The **topic** describes what the dataset represents.
The **geometry** describes the shape and relative arrangement of the data.
The **attribute** describes the dataset's key attribute and possible values.
The **legend** details how the visualizations should be interpreted. Legends are only provided for cert

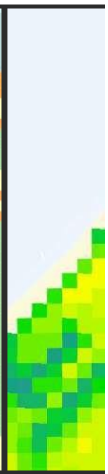
Which of the three datasets below is different from the other two in terms of operations that can be r



Topic	Areas with potential heat stress during summer days
Geometry	Square cell (granularity is not known)
Attribute	Sensitivity score (a value between 0 and 10)
Legend	The red areas are very sensitive to heat stress, the blue areas hardly

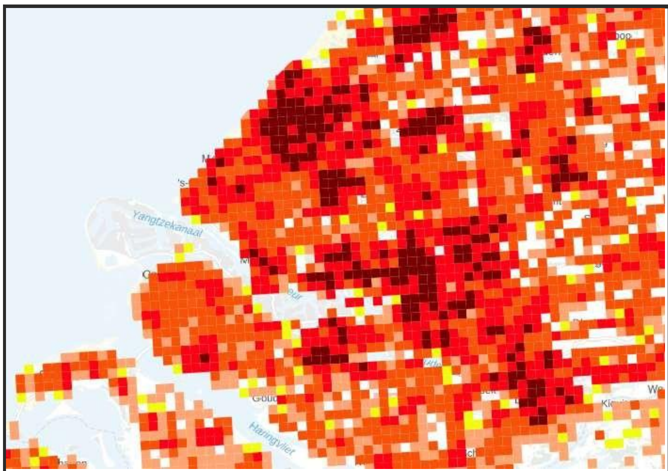


Topic	Gross domestic product of the world in 2010
Geometry	Square cell (0.5 x 0.5 degrees)
Attribute	Gross Domestic Product
Legend	Dark blue for low GDP to dark red for high GDP

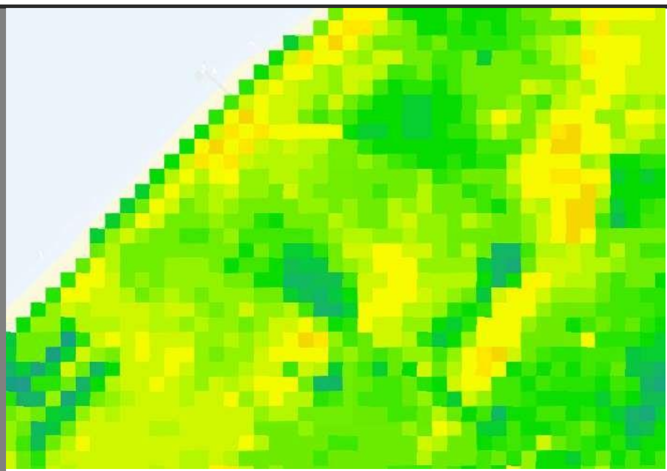


Topic
Geometry
Attribute
Legend

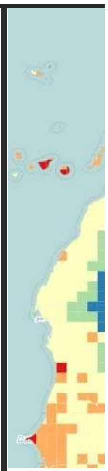
Which of the three datasets below is different from the other two in terms of operations that can be r



Topic	Population at residences place
Geometry	Square cell (1 x 1 kilometer)
Attribute	Number of inhabitants (e.g., 2, 87, 1621)
Legend	Darker colors indicate higher numbers

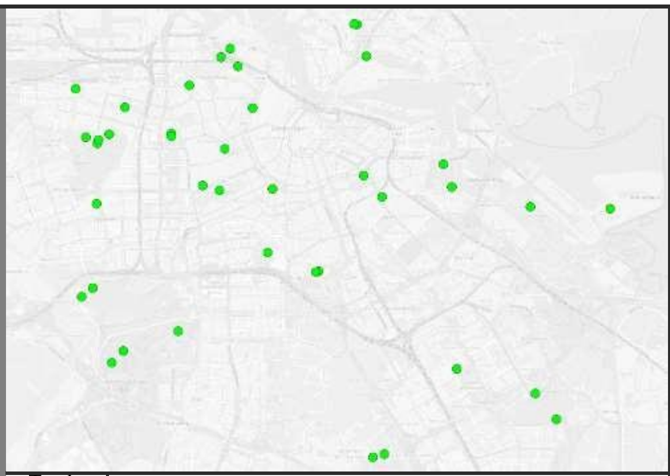
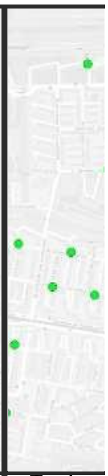


Topic	Digital elevation model
Geometry	Square cell (100 x 100 meters)
Attribute	Height above sea level (e.g., -200, 50 7000
Legend	Dark green for low elevation to dark orange for high elevation



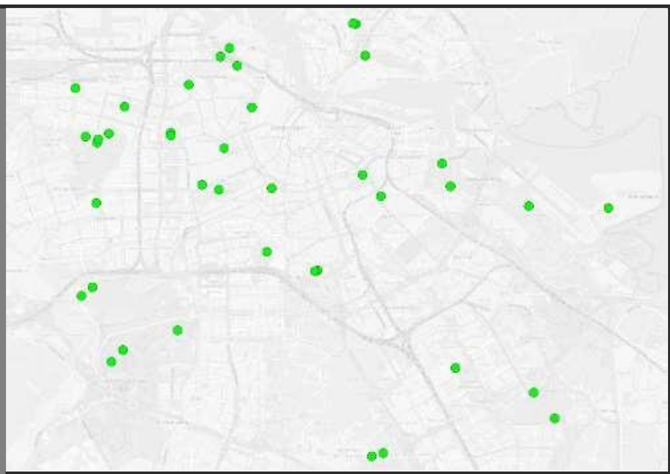
Topic
Geometr
Attribute
Legend

Which of the three datasets below is different from the other two in terms of operations that can be r

Topic	Nitrogen dioxide (NO2) sensing	**Topic**	Locations of swimming pools	**Topic**	
Geometry	Vector point	**Geometry**	Vector point	**Geometry**	
Attribute	Amount (micrograms per cubic meter) of NO2 (e.g., 31.394, 39.22, 40.311)	**Attribute**	Place name (e.g., Marnixbad, De Mirandabad, Amsterdamse Bos)	**Attribute**	

Which of the three datasets below is different from the other two in terms of operations that can be r


Topic	Nitrogen dioxide (NO2) sensing	**Topic**	Locations of swimming pools	**Topic**	
Geometry	Vector point	**Geometry**	Vector point	**Geomet**	
Attribute	Amount (micrograms per cubic meter) of NO2 (e.g., 31.394, 39.22, 40.311)	**Attribute**	Place name (e.g., Marnixbad, De Mirandabad, Amsterdamse Bos)	**Attribut**	

In each of the following questions, you are asked to consider a single dataset. For this dataset, plea that you think **can be meaningfully applied to the dataset**. With meaningful, the dataset can be us produce a conceptually valid output.

For each dataset, a **map** and information about the **topic**, the **geometry**, the **attribute,** and optiona

The **topic** describes what the dataset represents.
The **geometry** describes the shape and arrangement of the data.
The **attribute** describes the dataset's key attribute and possible values.
The **legend** details how the visualizations should be interpreted. Legends are only provided for certa

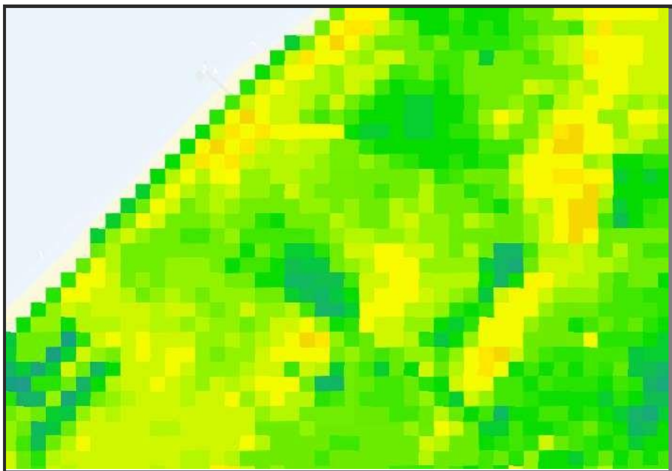
B2_Q1

In the following questions, you are asked to consider three datasets. For each question, one of the t
other two in terms of the analytical operations that can be meaningfully applied to them. With meani
should give a conceptually valid output. Your task is to select the dataset that is most different from t

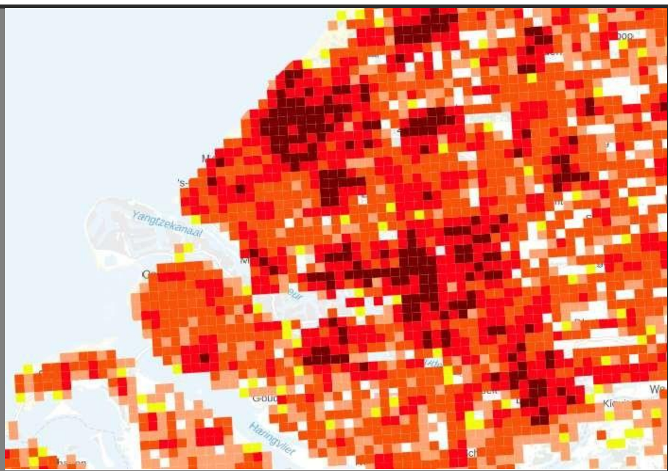
For each dataset, a **map** and information about the **topic**, the **geometry**, the **attribute,** and optiona

The **topic** describes what the dataset represents.
The **geometry** describes the shape and relative arrangement of the data.
The **attribute** describes the dataset's key attribute and possible values.
The **legend** details how the visualizations should be interpreted. Legends are only provided for certa

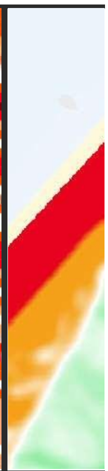
Which of the three datasets below is different from the other two in terms of operations that can be r



Topic	Digital elevation model
Geometry	Square cell (100 x 100 meters)
Attribute	Height above sea level (e.g., -200, 50 7000
Legend	Dark green for low elevation to dark orange for high elevation

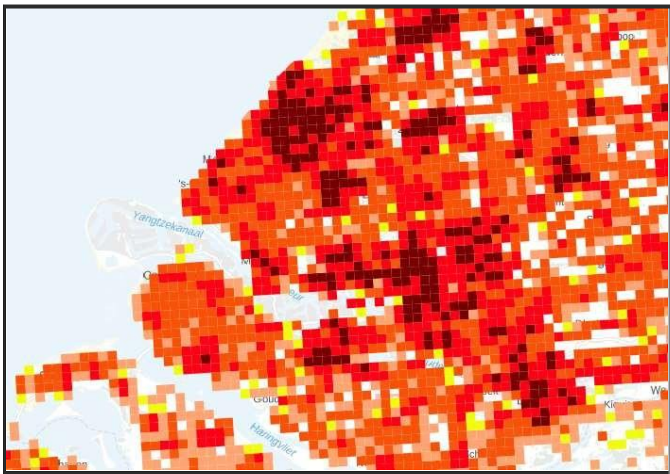


Topic	Population at residences place
Geometry	Square cell (1 x 1 kilometer)
Attribute	Number of inhabitants (e.g., 2, 87, 1621)
Legend	Darker colors indicate higher numbers

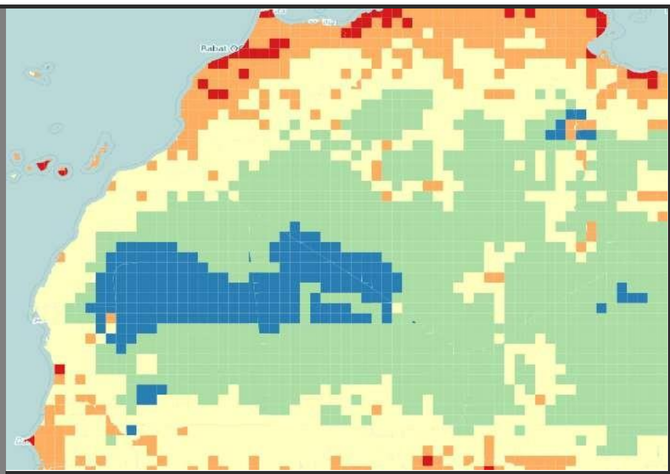


Topic	
Geometry	
Attribute	
Legend	

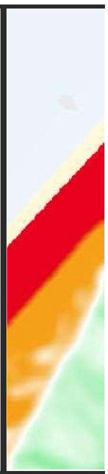
Which of the three datasets below is different from the other two in terms of operations that can be r



Topic	Population at residences place
Geometry	Square cell (1 x 1 kilometer)
Attribute	Number of inhabitants (e.g., 2, 87, 1621)
Legend	Darker colors indicate higher numbers



Topic	Gross domestic product of the world in 2010
Geometry	Square cell (0.5 x 0.5 degrees)
Attribute	Gross Domestic Product
Legend	Dark blue for low GDP to dark red for high GDP

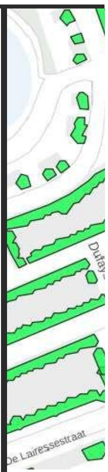


Topic	
Geometry	
Attribute	
Legend	

Which of the three datasets below is different from the other two in terms of operations that can be r

Topic	Areas of 6-character postcodes
Geometry	Vector region
Attribute	Postcode (e.g. 1047HH, 3818AC, 2734BR)

Topic	Dog walking policies
Geometry	Vector region
Attribute	One of three policy categories: "Prohibited area", "Restriciton area", and "Run-off area".



Topic	
Geometry	
Attribute	

Which of the three datasets below is different from the other two in terms of operations that can be r

Topic	Areas of 6-character postcodes
Geometry	Vector region
Attribute	Postcode (e.g. 1047HH, 3818AC, 2734BR)

Topic	Dog walking policies
Geometry	Vector region
Attribute	One of three policy categories: "Prohibited area", "Restriciton area", and "Run-off area".

Topic	
Geometry	
Attribute	

In each of the following questions, you are asked to consider a single dataset. For this dataset, plea that you think **can be meaningfully applied to the dataset**. With meaningful, the dataset can be us produce a conceptually valid output.

For each dataset, a **map** and information about the **topic**, the **geometry**, the **attribute,** and optiona

The **topic** describes what the dataset represents.
The **geometry** describes the shape and arrangement of the data.
The **attribute** describes the dataset's key attribute and possible values.
The **legend** details how the visualizations should be interpreted. Legends are only provided for certa

B3_Q1

In the following questions, you are asked to consider three datasets. For each question, one of the t other two in terms of the analytical operations that can be meaningfully applied to them. With meani should give a conceptually valid output. Your task is to select the dataset that is most different from t

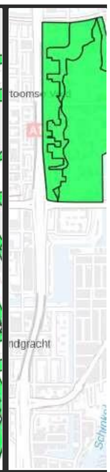
For each dataset, a **map** and information about the **topic**, the **geometry**, the **attribute,** and optiona

The **topic** describes what the dataset represents.
The **geometry** describes the shape and relative arrangement of the data.
The **attribute** describes the dataset's key attribute and possible values.
The **legend** details how the visualizations should be interpreted. Legends are only provided for certa

Which of the three datasets below is different from the other two in terms of operations that can be r

Topic	Vector region
Geometry	
Attribute	One of three policy categories: "Prohibited area", "Restriciton area", and "Run-off area".

Geometry	Vector region
Attribute	One of seven mix categories: "living only", "working only", "facility only", "living and facility", "living and working", "working and facility", "living, working and facility".



Topic	
Geometry	
Attribute	

Which of the three datasets below is different from the other two in terms of operations that can be r

Topic	Vector region
Geometry	
Attribute	Postcode (e.g. 1047HH, 3818AC, 2734BR)

Geometry	Vector region
Attribute	One of seven mix categories: "living only", "working only", "facility only", "living and facility", "living and working", "working and facility", "living, working and facility".

Topic	
Geometry	
Attribute	

Which of the three datasets below is different from the other two in terms of operations that can be r

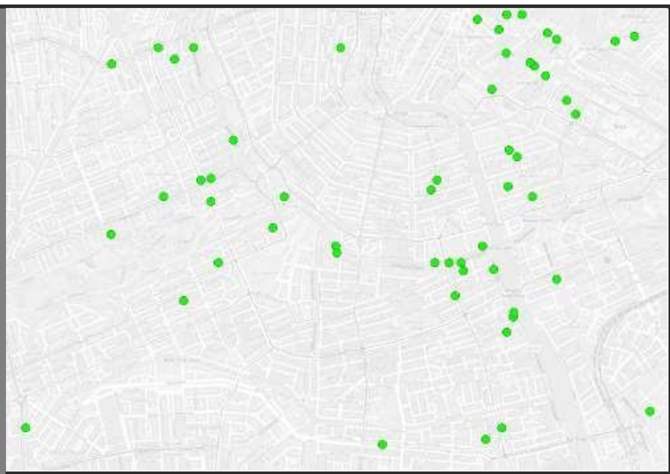
Topic	Nitrogen dioxide (NO2) sensing
Geometry	Vector point
Attribute	Amount (micrograms per cubic meter) of NO2 (e.g., 31.394, 39.22, 40.311)

Topic	Average lowest groundwater leve
Geometry	Vector point
Attribute	Average depth (in meters) of groundwater below ground level (e.g., -1.32, -1.14, -3.5)

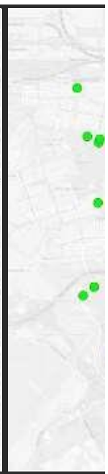
Topic	
Geomet	
Attribut	

Which of the three datasets below is different from the other two in terms of operations that can be r

Topic	Average lowest groundwater leve
Geometry	Vector point
Attribute	Average depth (in meters) of groundwater below ground level (e.g., -1.32, -1.14, -3.5)



Topic	Nitrogen dioxide (NO2) sensors
Geometry	Vector point
Attribute	Unique sensor code (e.g., CP27, TK10, T2)



Topic	
Geometry	
Attribute	

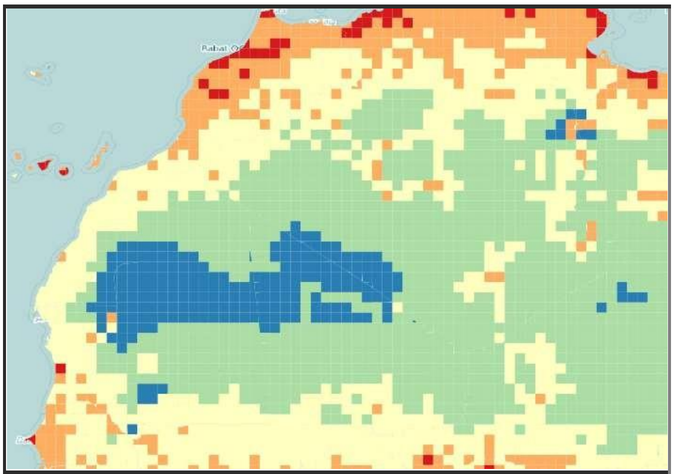
In each of the following questions, you are asked to consider a single dataset. For this dataset, plea that you think **can be meaningfully applied to the dataset**. With meaningful, the dataset can be us produce a conceptually valid output.

For each dataset, a **map** and information about the **topic**, the **geometry**, the **attribute,** and optiona

The **topic** describes what the dataset represents.
The **geometry** describes the shape and arrangement of the data.
The **attribute** describes the dataset's key attribute and possible values.
The **legend** details how the visualizations should be interpreted. Legends are only provided for certa

B1_Q2

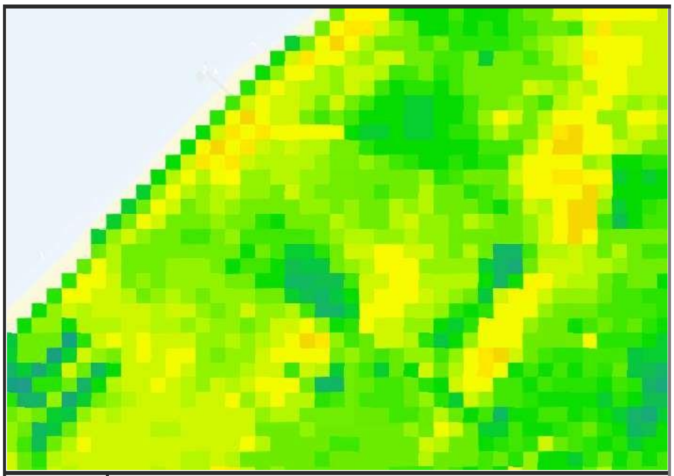
For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	Gross domestic product of the world in 2010
Geometry	Square cell (0.5 x 0.5 degrees)
Attribute	Gross Domestic Product
Legend	Dark blue for low GDP to dark red for high GDP

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Dissolve (documentation)	○	○
Reclassify (documentation)	○	○
Polygon to Raster (documentation)	○	○
Euclidian Distance (documentation)	○	○
Zonal Statistics (documentation)	○	○
Areal Interpolation (documentation)	○	○

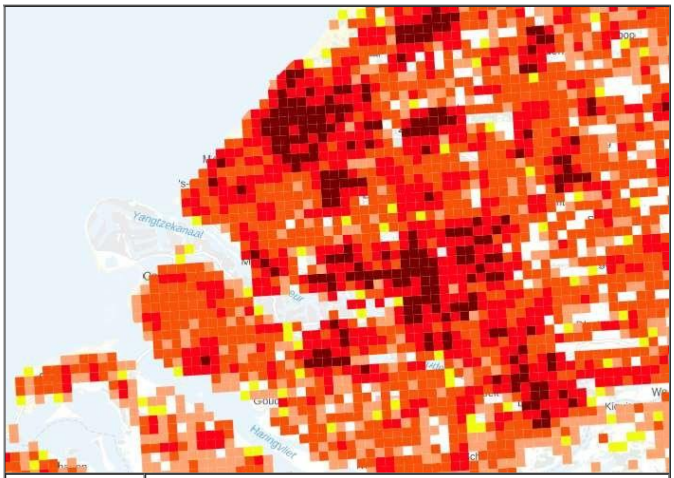
For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	Digital elevation model
Geometry	Square cell (100 x 100 meters)
Attribute	Height above sea level (e.g., -200, 50, 7000)
Legend	Dark green for low elevation to dark orange for high elevation

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Polygon to Raster (documentation)	○	○
Spatial Join (documentation)	○	○
Areal Interpolation (documentation)	○	○
Reclassify (documentation)	○	○
Create Thiessen Polygons (documentation)	○	○
Zonal Statistics (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

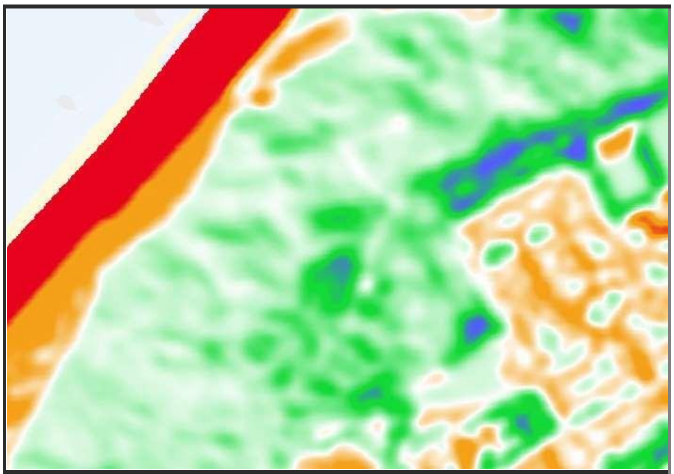


Topic	Population at residences place
Geometry	Square cell (1 x 1 kilometer)

Attribute	Number of inhabitants (e.g., 2, 87, 1621)
Legend	Darker colors indicate higher numbers

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Areal Interpolation (documentation)	○	○
Zonal Statistics (documentation)	○	○
Reclassify (documentation)	○	○
Polygon to Raster (documentation)	○	○
Euclidian Distance (documentation)	○	○
IDW Interpolation (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

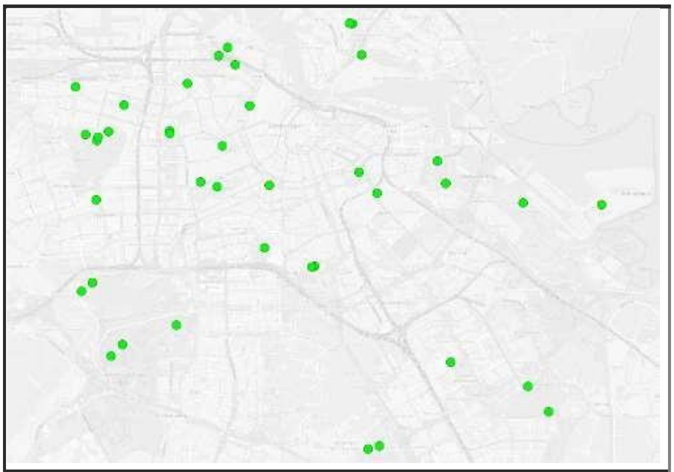


Topic	Areas with potential heat stress during summer days
Geometry	Square cell (granularity is not known)
Attribute	Sensitivity score (a value between 0 and 10)
Legend	The red areas are very sensitive to heat stress, the blue areas hardly

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Polygon to Raster (documentation)	○	○
Reclassify (documentation)	○	○
Areal Interpolation (documentation)	○	○
Create Thiessen Polygons (documentation)	○	○
IDW Interpolation (documentation)	○	○

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Zonal Statistics (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	Locations of swimming pools
Geometry	Vector point
Attribute	Place name (e.g., Marnixbad, De Mirandabad, Amsterdamse Bos)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Areal Interpolation (documentation)	○	○
Spatial Join (documentation)	○	○
IDW Interpolation (documentation)	○	○
Euclidian Distance (documentation)	○	○
Reclassify (documentation)	○	○
Create Thiessen Polygons (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

Topic	Average lowest groundwater level
Geometry	Vector point
Attribute	Average depth (in meters) of groundwater below ground level (e.g., -1.32, -1.14, -3.5)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Spatial Join (documentation)	○	○
Create Thiessen Polygons (documentation)	○	○
Zonal Statistics (documentation)	○	○
Reclassify (documentation)	○	○
IDW Interpolation (documentation)	○	○
Euclidian Distance (documentation)	○	○

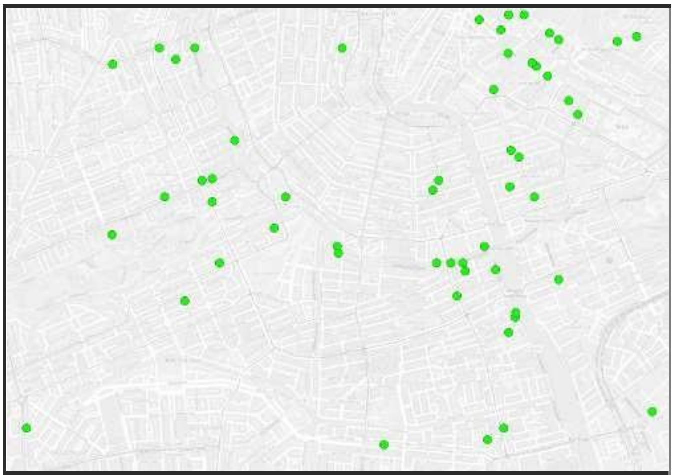
For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

Topic	Nitrogen dioxide (NO2) sensing
Geometry	Vector point

Attribute	Amount (micrograms per cubic meter) of NO2 (e.g ., 31.394, 39.22, 40.311)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Euclidian Distance (documentation)	○	○
Zonal Statistics (documentation)	○	○
IDW Interpolation (documentation)	○	○
Create Thiessen Polygons (documentation)	○	○
Spatial Join (documentation)	○	○
Dissolve (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



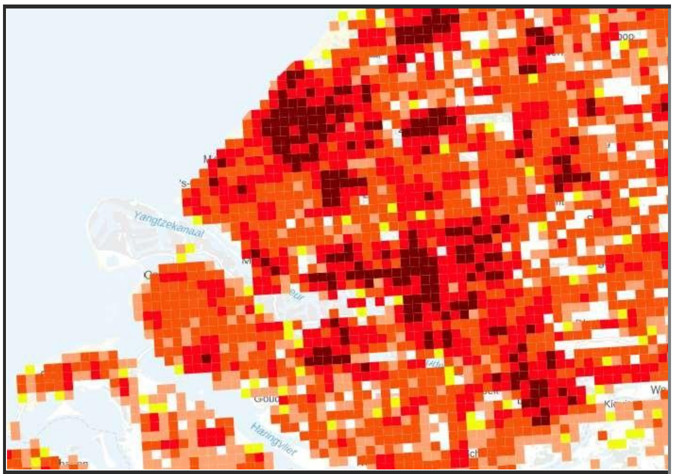
Topic	Nitrogen dioxide (NO2) sensors
Geometry	Vector point
Attribute	Unique sensor code (e.g., CP27, TK10, T2)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Create Thiessen Polygons (documentation)	○	○
IDW Interpolation (documentation)	○	○
Zonal Statistics (documentation)	○	○
Areal Interpolation (documentation)	○	○
Spatial Join (documentation)	○	○

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Euclidian Distance (documentation)	○	○

B2_Q2

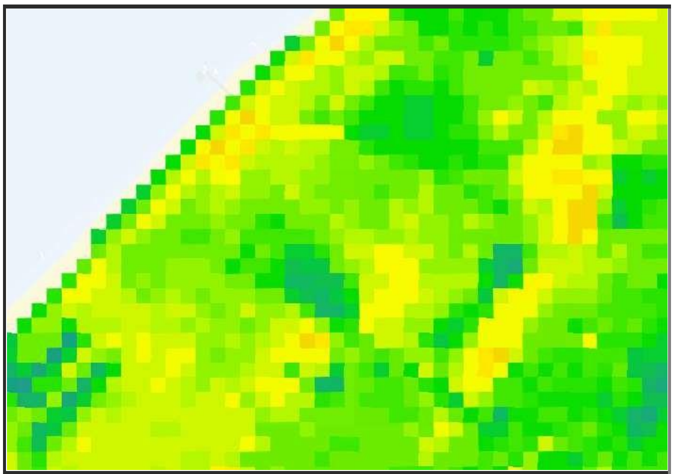
For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	Population at residences place
Geometry	Square cell (1 x 1 kilometer)
Attribute	Number of inhabitants (e.g., 2, 87, 1621)
Legend	Darker colors indicate higher numbers

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Spatial Join (documentation)	○	○
Polygon to Raster (documentation)	○	○
Zonal Statistics (documentation)	○	○
Reclassify (documentation)	○	○
Areal Interpolation (documentation)	○	○
Dissolve (documentation)	○	○

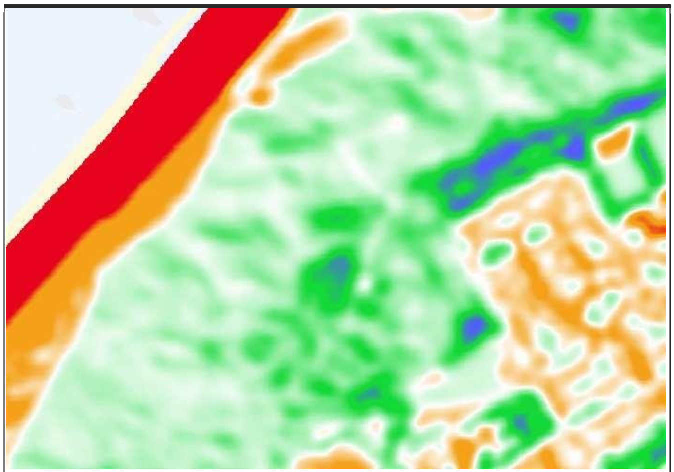
For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	Digital elevation model
Geometry	Square cell (100 x 100 meters)
Attribute	Height above sea level (e.g., -200, 50, 7000)
Legend	Dark green for low elevation to dark orange for high elevation

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Polygon to Raster (documentation)	○	○
Areal Interpolation (documentation)	○	○
Euclidian Distance (documentation)	○	○
Reclassify (documentation)	○	○
Dissolve (documentation)	○	○
Zonal Statistics (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

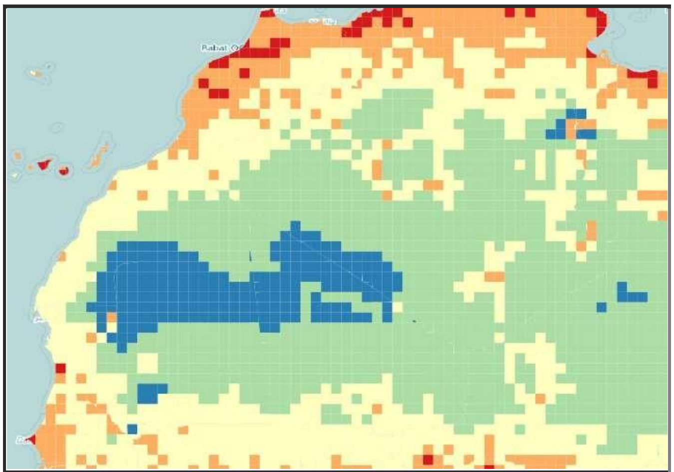


Topic	Areas with potential heat stress during summer days
Geometry	Square cell (granularity is not known)
Attribute	Sensitivity score (a value between 0 and 10)

Legend	The red areas are very sensitive to heat stress, the blue areas hardly

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Euclidian Distance (documentation)		
Reclassify (documentation)		
Zonal Statistics (documentation)		
Polygon to Raster (documentation)		
Areal Interpolation (documentation)		
Dissolve (documentation)		

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	Gross domestic product of the world in 2010
Geometry	Square cell (0.5 x 0.5 degrees)
Attribute	Gross Domestic Product
Legend	Dark blue for low GDP to dark red for high GDP

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Zonal Statistics (documentation)		
Reclassify (documentation)		
Areal Interpolation (documentation)		
Euclidian Distance (documentation)		
Polygon to Raster (documentation)		

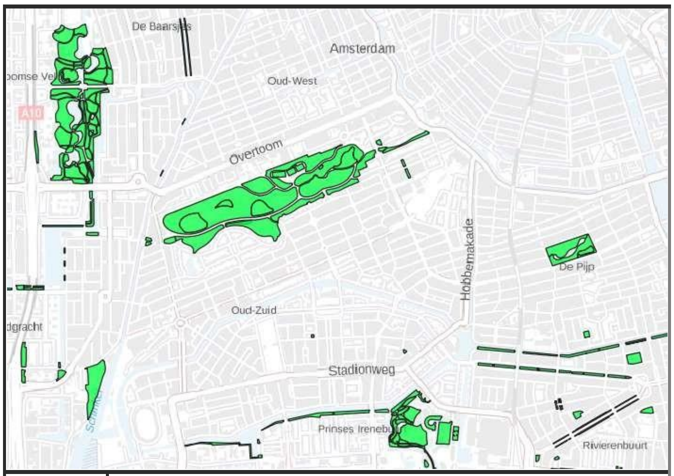
	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Spatial Join (documentation)	◯	◯

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

Topic	Areas of 6-character postcodes
Geometry	Vector region
Attribute	Postcode (e.g 1047HH, 3818AC, 2734BR)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Polygon to Raster (documentation)	◯	◯
IDW Interpolation (documentation)	◯	◯
Create Thiessen Polygons (documentation)	◯	◯
Spatial Join (documentation)	◯	◯
Dissolve (documentation)	◯	◯
Euclidian Distance (documentation)	◯	◯

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	Dog walking policies
Geometry	Vector region
Attribute	One of three policy categories: "Prohibited area", "Restriciton area", and "Run-off area"

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Spatial Join (documentation)	◯	◯
Dissolve (documentation)	◯	◯
Euclidian Distance (documentation)	◯	◯
Zonal Statistics (documentation)	◯	◯
IDW Interpolation (documentation)	◯	◯
Polygon to Raster (documentation)	◯	◯

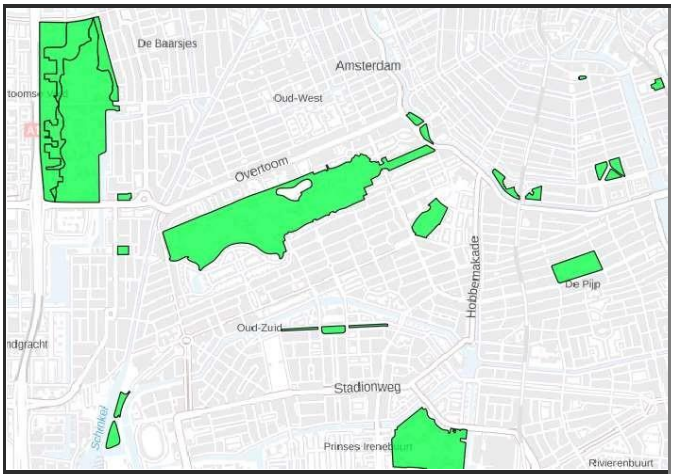
For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

Topic	Function mix for built areas
Geometry	Vector region
Attribute	One of seven mix categories: "living only", "working only", "facility only", "living and facility", "living and working" ,

	"working and facility", "living, working and facility".

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Euclidian Distance (documentation)	○	○
IDW Interpolation (documentation)	○	○
Polygon to Raster (documentation)	○	○
Spatial Join (documentation)	○	○
Dissolve (documentation)	○	○
Zonal Statistics (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

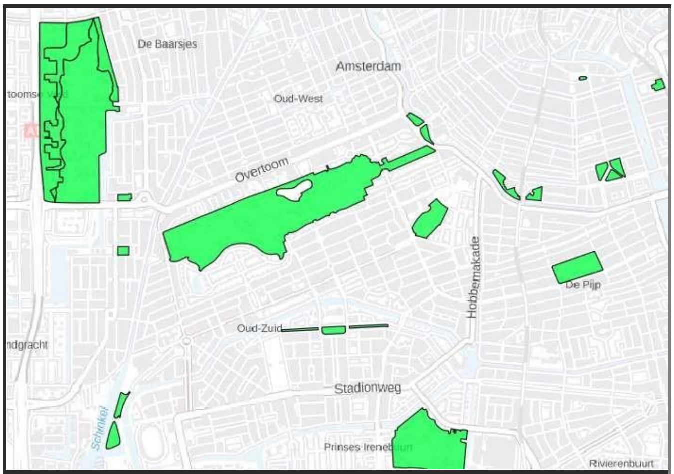


Topic	City parks
Geometry	Vector region
Attribute	Park name (e.g., Amstelpark, Vondelpark, Beatrixpark)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Zonal Statistics (documentation)	○	○
Spatial Join (documentation)	○	○
Polygon to Raster (documentation)	○	○
Euclidian Distance (documentation)	○	○
Dissolve (documentation)	○	○
Create Thiessen Polygons (documentation)	○	○

B3_Q2

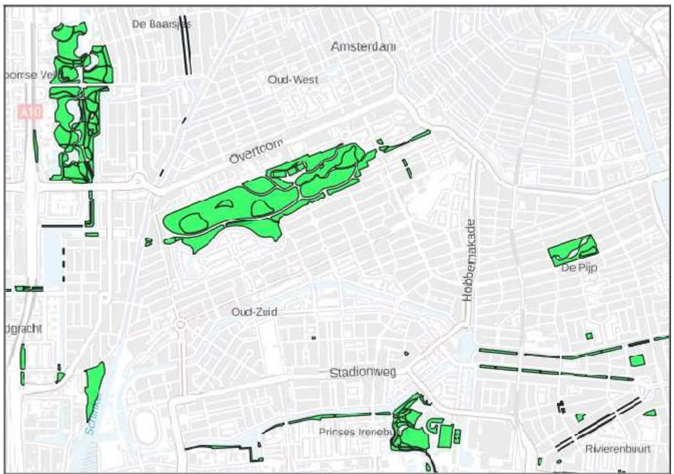
For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	City parks
Geometry	Vector region
Attribute	Park name (e.g., Amstelpark, Vondelpark, Beatrixpark)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Polygon to Raster (documentation)	○	○
Zonal Statistics (documentation)	○	○
Spatial Join (documentation)	○	○
Dissolve (documentation)	○	○
IDW Interpolation (documentation)	○	○
Euclidian Distance (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	Dog walking policies
Geometry	Vector region
Attribute	One of three policy categories: "Prohibited area", "Restriciton area", and "Run-off area"

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Areal Interpolation (documentation)	○	○
Spatial Join (documentation)	○	○
Polygon to Raster (documentation)	○	○
Euclidian Distance (documentation)	○	○
Dissolve (documentation)	○	○
Reclassify (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

Topic	Function mix for built areas
Geometry	Vector region
Attribute	One of seven mix categories: "living only" , "working only", "facility only", "living and facility", "living and working" , "working and facility", "living, working and facility"

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Polygon to Raster (documentation)	○	○
Create Thiessen Polygons (documentation)	○	○
Euclidian Distance (documentation)	○	○
Areal Interpolation (documentation)	○	○

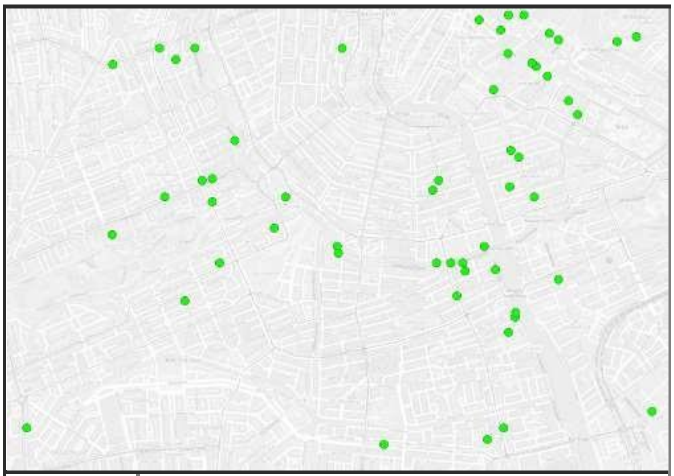
	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Spatial Join (documentation)	○	○
Dissolve (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

Topic	Areas of 6-character postcodes
Geometry	Vector region
Attribute	Postcode (e.g 1047HH, 3818AC, 2734BR)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Euclidian Distance (documentation)	○	○
Dissolve (documentation)	○	○
Spatial Join (documentation)	○	○
Polygon to Raster (documentation)	○	○
IDW Interpolation (documentation)	○	○
Areal Interpolation (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	Nitrogen dioxide (NO2) sensors
Geometry	Vector point
Attribute	Unique sensor code (e.g., CP27, TK10, T2)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Zonal Statistics (documentation)	○	○
Spatial Join (documentation)	○	○
Euclidian Distance (documentation)	○	○
IDW Interpolation (documentation)	○	○
Dissolve (documentation)	○	○
Create Thiessen Polygons (documentation)	○	○

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

Topic	Average lowest groundwater level
Geometry	Vector point
Attribute	Average depth (in meters) of groundwater below ground level (e.g., -1.32, -1.14, -3.5)

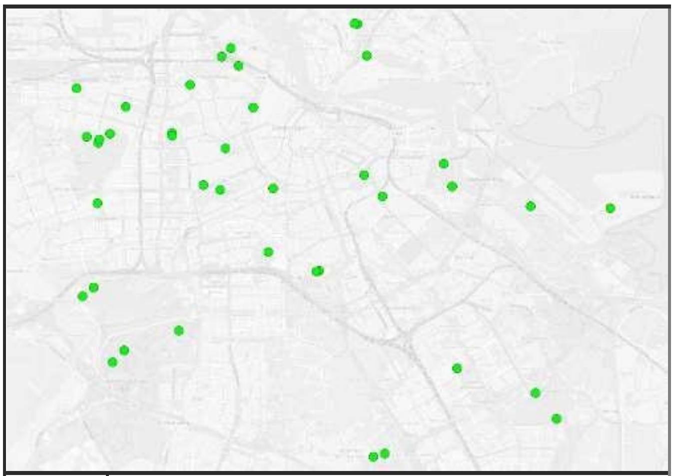
	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Create Thiessen Polygons (documentation)	◯	◯
IDW Interpolation (documentation)	◯	◯
Spatial Join (documentation)	◯	◯
Euclidian Distance (documentation)	◯	◯
Dissolve (documentation)	◯	◯
Zonal Statistics (documentation)	◯	◯

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.

Topic	Nitrogen dioxide (NO2) sensing
Geometry	Vector point
Attribute	Amount (micrograms per cubic meter) of NO2 (e.g ., 31.394, 39.22, 40.311)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
IDW Interpolation (documentation)	◯	◯
Dissolve (documentation)	◯	◯
Reclassify (documentation)	◯	◯
Create Thiessen Polygons (documentation)	◯	◯
Euclidian Distance (documentation)	◯	◯
Spatial Join (documentation)	◯	◯

For this dataset, please select those analytical operations that you think can be meaningfully applied to the dataset.



Topic	Locations of swimming pools
Geometry	Vector point
Attribute	Place name (e.g., Marnixbad, De Mirandabad, Amsterdamse Bos)

	Meaningfully applicable to the dataset	Not meaningfully applicable to the dataset
Euclidian Distance (documentation)	○	○
IDW Interpolation (documentation)	○	○
Spatial Join (documentation)	○	○
Areal Interpolation (documentation)	○	○
Create Thiessen Polygons (documentation)	○	○
Polygon to Raster (documentation)	○	○

B4

Categorize your expertise with Geographic Information Systems (GIS)

	Laymen: never used GIS	Beginner: can use basic GIS functions	Trained: formally trained by a GIS course	Expert: used GIS for 5 years or more
GIS Expertise	○	○	○	○

Categorize your familiarity with the presented analytical tools

	Never used	Used once	Used occasionally	Used extensively
Areal Interpolation (documentation)	○	○	○	○
Euclidian Distance (documentation)	○	○	○	○
Reclassify (documentation)	○	○	○	○
Dissolve (documentation)	○	○	○	○
IDW Interpolation (documentation)	○	○	○	○
Polygon to Raster (documentation)	○	○	○	○
Zonal Statistics (documentation)	○	○	○	○
Create Thiessen Polygons (documentation)	○	○	○	○
Spatial Join (documentation)	○	○	○	○

Categorize your familiarity with the core concepts (*field, object, event, network)* as proposed by Werner Kuhn (2012)

	Not at all familiar	Slightly familiar	Moderately familiar	Very familiar	Extremely familiar
Core concepts familiarty	○	○	○	○	○