NTNU - NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET
Faculty of Engeneering Science and Technology
Department of Civil and Transport Engineering
TBA4925 - Master Thesis

Optimizing the micro-tasking workflow and exploring it's usage potential within geospatial data

Anne Sofie Strand Erichsen
Trondheim, June 2017

DAIM page

Background

Micro-tasking is an Internet phenomenon that has increased in popularity over the last years. The method is used for solving large tasks that can be divided into many smaller ones (micro-tasks). This involves both the use of computers and a large number of people. Importing geospatial data over large areas can be considered as a large task. This is a task that can be divided into smaller ones, for example by dividing into smaller areas.

Task Description

This Master thesis will have emphasis on the data validation and conflict handling part of the import of geospatial data. These processes are too complicated to do fully automatic through scripts, and the thesis investigates how micro-tasking can be relevant approach to the problem. Specific tasks:

- Study related literature
- Make a web-based experiment in order to answer the following questions:
  - What are the number of objects optimal within a task to get it completed as quickly as possible?
  - Does the quality of the work vary between the different tasks given?
  - Do amateurs manage to do the tasks?
- Explore the micro-tasking methods usage potential within geospatial data
- Can other process that requires human interaction make advantage of this method?

Administrative/guidance

The work on the Master Thesis starts on January 18th, 2017

The thesis report as described above shall be submitted digitally in DAIM at the latest at June 18th, 2017

External supervisor Atle Frenvik Sveen, Norkart AS

Supervisors at NTNU and professor in charge: Terje Midtbø

Abstract

This paper proposes to extend the usage of the micro-tasking method to also geospatial tasks. We see a huge potential in exploiting geospatial tasks with the micro-tasking method. Two research hypotheses are answered: 1) Inexperienced workers cannot solve micro-tasks containing geospatial data as good as experienced workers, 2) The fewer elements in a micro-task, the better the worker solves the task.

An online web experiment was developed and implemented to gather data about how individuals solve geospatial micro-tasks. Statistical analysis is conducted on the gather data to answer the research hypotheses. The experiment registered the individual's background to see if the quality of the solved micro-tasks differs between experienced and inexperienced participants. The tasks varied the number of elements that had to be handled at the same time to complete the micro-task. This approach was used to determine if the number has an influence on the quality of the solved micro-tasks.

Statistical analysis found significant evidence of inexperienced participants finishing the micro-tasks faster. The quality of the completed micro-tasks did not differ when comparing experts and non-experts. When comparing the three different tasks in the experiment, the task containing the fewest elements had statistically better quality than the two other. There was also a difference in time spent completing the three tasks, but not enough to be significant.

Sammendrag

Sammendrag på norsk

Preface

This paper is a master thesis written for the Department of Civil and Transport Engineering at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway. It is a part of the study program Engineering and ICT - Geomatics, and was written in the spring of 2017.

I would like to thank my supervisor Terje Midtbø for his help and feedback, and also Atle Frenvik Sveen for his support and help every time I needed it.

Trondhiem, 2017-06-16?
Anne Sofie Strand Erichsen

Contents

List of Figures

List of Tables

1 Introduction

Having access to information is not the same as learning and understanding the information. Today, the amount of information available is enormous. Computers have the ability to learn, but is dependent on well-developed training data to learn. Training data has to be information sat in contextual meaning. Understanding and giving contextual meaning to information is the strengths of the human brain. The human brain has the ability to create new ideas and concepts from unstructured information (Ross and Jamilly, 2016). Therefore, computers need the human brain to learn, as humans need computers for their speed and accuracy.

Albert Einstein illustrates this perfectly: "Computers are incredibly fast, accurate, but stupid. Humans are incredibly slow, inaccurate, but brilliant. Together they may be powerful beyond imagination" (Holzinger, 2013).

This thesis will argue that a good approach for combining the strengths of both computers and humans together is by using micro-tasking. Micro-tasks are the smallest, simplest types of tasks that should demand little time to complete. We believe that micro-tasking has an unexplored potential within geospatial tasks. Geospatial tasks are tasks involving data with a geographic position. The tasks can be everything from creating geospatial data to validating or analyzing it.

To the authors best knowledge, little research has been done on micro-tasks involving geospatial data. To be able to exploit the micro-tasking method together with geospatial data, it is important to study how well humans solves these kinds of tasks. Micro-tasks are often published on a micro-tasking platform, where the tasks and people are connected. It is important to know if inexperienced individuals are capable of solving the tasks. If only experienced people can solve geospatial micro-tasks, a platform who can distinguish between people with geospatial knowledge needs to be used. We do not have any knowledge if that kind of platform exists.

In Remote-Sensing, humans perform land-cover classification tasks (Salk et al., 2016). At least two papers have studied whether there are significant differences in quality of the information contributed by experts and non-experts. Salk et al. (2016) concluded that there was a little first-order relationship between professional background and the task accuracy. When comparing specialists with non-specialists, the non-specialists performed slightly better on images near home. See et al. (2013) concluded that there was little difference between experts and non-experts, and also found that the non-experts improved more than experts over time. Overall the non-experts were as reliable in what they identified as the experts. These two studies show that the person's background is not necessarily important. See et al. (2013) concluded that with proper targeted training material the differences between experts and non-experts could potentially decrease.

This thesis aim is to study if micro-tasking can successfully be expanded to involving maps and geospatial data. The OpenStreetMap community has used the method some years, and the usage so far can be evaluated as successful (Erichsen, 2016). This thesis will use knowledge from the OSM community's usage of micro-tasking to examine if people that do not have experience of working with geospatial data manage to solve geospatial tasks involving an interactive map and also metadata interpretation. The thesis also aims to determine if the number of elements in each micro-task has an impact on how well individuals solves the task. The quality of each completed task, when using the micro-task method, has to be checked and approved. By varying the number of elements present in each micro-task, this thesis can examine if the quality varies between the tasks.

There is potentially a lot of work creating micro-tasks. The large task needs to be appropriately broken down to micro-tasks that are easy, enjoyable, and fast to solve. This breakdown requires design skills and proper tutorials and examples for new workers (Schulze et al., 2012). Guidelines can be used to avoid putting too much emphasis on the preparations. It can support how to utilize the preparation resources best. This thesis can be utilized as a guideline on how to break down large geospatial tasks.

This thesis will answer the research hypothesis:

- Inexperienced workers cannot solve micro-tasks containing geospatial data as good as experienced workers
- The fewer elements in a micro-task, the better the worker solves the task

An online web experiment was developed to answer the research hypotheses. The experiment contains three tasks varying the number of elements given to the participant. Each task has the same two questions representing two micro-tasks. One question involves map interaction and the other an interpretation task containing metadata.

The next chapter will give a thorough introduction to micro-tasking and hopefully make it clearer what this thesis aim is. Chapter three will introduce the experiment, the web application hosting the experiment and results from the experiments pilot test. Chapter four will present the individuals participating in the experiment, a description of the statistical theory used in the analysis, and give a summary of the gathered data. The last section in chapter four will contain the analyses of hypotheses which in sum will answer the research hypotheses. Chapter five contains the discussion, where the results are summed and discussed. At last, the thesis will give a conclusion and proposal for future work.

2 Background

Creating and maintaining real-world knowledge bases in a classical work environment demands a high cost, and is a cost that is often unnecessary [(Meier, 2013b), p. 134]. Alternative approaches are to rely on the knowledge of open crowds, volunteer contributions, or services like micro-tasking platforms where there are people ready to work on the tasks given to them [(Meier, 2013b), p. 134].

Today, geospatial data is more available than ever. Governments are releasing more and more data and the OpenStreetMap database is still growing. While general data availability is increasing, the quality of the data is not necessarily perfect and manual pre-processing is often necessary before using it (Difallah et al., 2015). Pre-processing of the data are time consuming and expensive. By exploiting both machines and people through the appropriate platform and approach, the cost can decrease and the quality increase. We will argue that combining machines and people is often a better and faster solution than a fully-automatic or fully-manual approach. Implementing this kind of approach into a micro-tasking platform can be a good solution.

2.1 Why do we need humans?

Machine learning gives computers the ability to learn without being explicitly programmed. It involves computer intelligence, but the computers do not know the answers up front (Stanford University, 2017). Machine-learning algorithms have enormous problems when contextual information is missing. Without a pre-set of rules, a machine has trouble solving the problem. Machines do not have creativity, which is required to answer complex problems (Holzinger et al., 2016). According to the company Mighty AI[1], humans cannot be removed from Artificial Intelligence training loops. Machine-learning approaches still require a huge amount of training data to work on new domains Schulze et al. (2012). Mighty AI believe that humans will continue to play a crucial role in creating training data for the algorithms (Gutzwiller, 2017).

It is suggested by Biewald (2015) and Oppenheimer (2017) that machine learning accuracy should follow the Pareto 80:20 principle. Getting 80 % accuracy can be reasonably easy to accomplish, but the last 20 % should be handled by human input (Biewald, 2015) (Oppenheimer, 2017). Important human input is providing training data and to route tasks when the algorithm is unsure of its answer [(Oppenheimer, 2017), (Nakhuda, 2016)]. The machine learning company developmentSEED[2] use a micro-tasking solution to clean their machine learning output data. They are using humans to get a faster, more accurate output data by developing Skynet Scrubber, a

[1] Mighty AI generates high-quality AI training data.

[2] DevelopmentSEED is a creative engineering team solving complex problems with open software and open data.

GUI web application solution to get human input quicker and easier (Their algorithm is called Skynet). In their blog, Derek Lieu writes: "Skynet gets more capable every day, but the output is still not perfect [..] We built Skynet Scrub so we could start using Skynet data sooner".

Holzinger (2016) claim that most people from the machine learning community are concentrating on *automatic* machine learning by bringing the humans out of the process. When humans are out of the loop, the training data sets can be uncertain and incomplete, and the resulting algorithm can be questionable (Holzinger, 2016). By bringing humans back in the process, especially in domains where the data sets are questionable, for instance in the health domain, one enables what neither a human or a computer can do on their own (Holzinger, 2016). It is today possible to build hybrid human-machine systems that combine both the scalability of computers and the yet unmatched cognitive abilities of the human brain (Difallah et al., 2016). "Computers are bad at finding patterns unless we have a well-understood problem" quote Stephen Cohen, co-founder of Palantir Technologies, only humans can understand and frame a new problem. Palantir Technologies believe in augmenting human intelligence, not replacing it. As Holzinger (2013) say, "[...] the problem-solving knowledge is located in the human mind and - not in machines, " and this is something we must acknowledge according to Holzinger.

2.2 Human computation

Human computing is, at its most general level, computation performed by human beings and a human computation system contains both humans and computers working together to solve difficult problems (Schulze et al., 2012). The author argues that utilizing the human processing power is still important. Humans are necessary even though our computers are becoming more and more complex. Traditional approaches to solving problems are to focus on improving the software, but as the reader will see in this thesis, a solution that uses humans cleverly by exploiting the human brain's cognitive abilities can create much faster and better results than software. One of the pioneers of crowdsourcing, Luis von Ahn, wanted to find a cheap and efficient way to label images (von Ahn, 2008). The solution was to exploit the use of a game-like approach in a non-game context to motivate individuals to label the pictures through a game. This approach is called gamification (Huotari and Hamari, 2017). The game was called "The ESP game" and solved the problem of labeling images with words. Most images do not have a proper caption associated with them, and this makes it difficult to create search engines for images. A fast and cheap method of labeling images is by using humans cleverly, and humans can very easily see if the picture contains, i.e., a dog or a cat. Through "The ESP game" humans where labeling images without even knowing it, they only played a fun game. Within a few months, the game collected more than 40 million image labels (von Ahn, 2008), and they did not even have to pay them doing it. Human computation is one of the major areas where the gamification approach has been employed (Morschheuser et al., 2016). Each human performs a small part of a massive computation task.

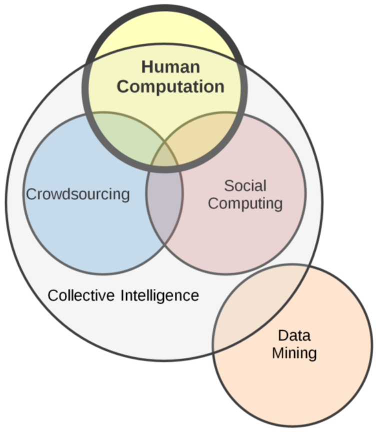


Figure 2.1: Collective intelligence (Quinn and Bederson, 2011)

Human computation, a term introduced by Luis von Ahn, refers to, according to Quinn and Bederson (2011), a distributed system that combine the strengths of humans and computers to accomplish tasks that neither can do alone. To make human computation in crowdsourcing compelling one needs to know how the results can be optimally acquired from humans and how the results can be integrated into productive environments without having to change established workflows and practices [(Meier, 2013b), p. 134]. Gamification can be one solution on how to make human computation in crowdsourcing effective (Wang et al., 2017). The author will argue that micro-tasking can be another solution for effective crowdsourcing using humans cognitive abilities.

2.3 Crowdsourcing

The first time the term "crowdsourcing" appeared was in Wired magazine article by Jeff Howe (Howe, 2006). Whereas human computing (2.2) replaces computers with humans, crowdsourcing replaces traditional human workers with members of the public (Quinn and Bederson, 2011). Crowdsourcing companies came onto the scene with the goal of helping businesses solve simple problems with massive scale (Webster, 2016). EYeka (2015) state that 85 % of the top global brands use crowdsourcing for various purposes. Crowdsourcing is an increasingly important concept (Salk et al., 2016) and has become a widespread approach to dealing with machine-based computations where we leverage the human intelligence (Gadiraju et al., 2015). Crowdsourcing is a way of refactoring work in a way that exploits the worker's flexibility and gets the right skills to the right part of the problem. To get the right skills to the right part of the problem it needs to be partitioned into smaller parts. Having smaller parts will make it easier to distribute the problem. The distribution can be done through micro-tasking, also called "smart crowdsourcing" by Patrick Meier (Meier, 2013a).

When the field of a crowdsourced project is explicitly geographical, it is often called *volunteered geographical information* (VGI). According to Salk et al. (2016), the best known VGI project is OpenStreetMap (OSM). OSM is an open-source mapping project, where volunteers contribute with their local knowledge and mapping abilities, called "the Wikipedia of maps" (Palen et al., 2015). Wikipedia can be said as beeing the best known and most successful example of crowdsourcing on a global scale. Chilton et al. (2009) suggested that the OSM project would have a similar impact on open and free geodata as Wikipedia had on 'fact finding'.

2.4 Micro-tasking

The simplest type of tasks are called micro-tasks and is illustrated in figure 2.2. Micro-tasks should not require any special training, and a task should be completed within a couple of minutes (Ipeirotis and G., 2010). Problems that are suitable for solving through micro-tasking are those that are easy to distribute into many simple tasks, which can be completed in parallel in a relatively short period of time, without requiring specific skills (Sarasua et al., 2012). Research has also demonstrated that micro-tasking is effective for far more complex problems when using sophisticated workflow management techniques. Micro-tasking can then be applied to a broader range of challenges like: (1) completing surveys, (2) translating text between two languages, (3) matching pictures of people, (4) summarizing text (Bernstein et al., 2015).

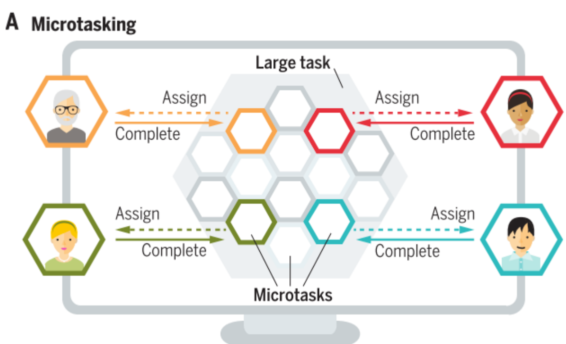


Figure 2.2: Micro-tasking (Michelucci and Dickinson, 2016)

2.4.1 Micro-tasking platforms

There exist multiple micro-tasking platforms. The platforms serve as a place where the crowd (workers) willing to perform small tasks is connected to work providers (Difallah et al., 2015). Tasks conducted by the crowd has to pass a quality control mechanism to be approved. The quality mechanisms vary between the platforms, but a common measure, noted by the author, is that at least three independent workers need to agree on the task before approved, this was also observed by James McAndrew[3]. Three popular and different platforms are presented in this section.

2.4.1.1 Amazon's Mechanical Turk

Amazon's Mechanical Turk (MTurk) is a very popular micro-tasking platform created in 2005, but still in use today (Difallah et al., 2016). MTurk acts as an online labor marketplace (Sarasua et al., 2012). It provides the infrastructure, connectivity and payment mechanisms so that hundreds of thousands of people can perform micro-tasks on the Internet and get paid per completed task. MTurk is used for many different tasks that are easier for people than computers. It contains simple tasks such as labeling or segmenting images or tagging content, to more complex tasks such as translating or even editing text (Franklin et al., 2011). In the marketplace, employers are known as requesters, and they post tasks, called *human intelligence tasks* (HIT's). The HIT's are then picked up by online users, *crowd workers*, who complete the tasks in exchange for a small payment (a few cents per HIT) (Ipeirotis and G., 2010). The workers self select the HIT's, a fundamental difference from traditional employment models (Schulze et al., 2012).

2.4.1.2 Tasking manager

The Tasking Manager tool is OpenStreetMap's micro-tasking platform. It was created in the aftermath of the Haiti earthquake in 2010 (Palen et al., 2015). The tool is used to coordinate satellite image tracing projects and sorts the area covered by the satellite image into grids so that multiple people can map the same area at the same time. Each person works at one grid each, this way they avoid mapping the same areas. Organizing the areas into grids is a very effective approach to coordinate a mapping job. The tasking manager is mainly used by the *Humanitarian OpenStreetMap Team* (HOT). This platform does not have a rewarding system or a gamification approach. It is solely based on volunteer contributors. This platform shows that it is not necessary to have a game or rewards for a successful platform.

There are also other micro-tasking tools in OpenStreetMap. MapRoulette and To-Fix are examples of such tools. Both tools are listed as error detection tools on the OSM quality assurance wiki page. MapRoulette uses a gamification approach, while To-Fix does not. The tools break common errors in the data into micro-tasks so that multiple individuals can work on the tasks simultaneously.

2.4.1.3 CrowdFlower

CrowdFlower is a company that wants to help businesses take advantage of crowd-sourcing and human computation. They act as an intermediary for these companies

[3] James McAndrew noted it in a mail correspondence with the author

(Quinn and Bederson, 2011). CrowdFlower receives tasks from businesses wanting to crowdsource their work or problems. CrowdFlower operates with a variety of services to get connected with workers (i.e., MTurk) (Quinn and Bederson, 2011).

What's special with CorwdFlower is their close ties with AI technology and a crowdsourced workforce. Their costumers are allowed to perform tasks with algorithms and machine learning, but bring in human judgment when they are not confident in the technology, and the human work can make the algorithms smarter (Ha, 2016). The founder of CrowdFlower says that "self-driving cars have gotten pretty good at recognizing many of the objects they encounter on the street, [..] (but) they can still struggle with tricky things like "a person in a Halloween costume dressed as a stationary object, or a pole with a person painted on it," which is where CrowdFlower comes in." (Ha, 2016).

2.4.2 Micro-tasking workforce

It is said that crowdsourcing is radically changing the nature of work (Deng et al., 2016). Traditional workers are restricted to offices and arranged office hours. With crowdsourcing, through for instance micro-tasking platforms, the workers can choose when to work, and even better: which jobs to perform. This appears very attractive, but is it only on the surface?

According to Deng et al. (2016), crowdsourcing is radically changing people's perspectives on how to manage their work-life balance. Compared to "traditional" work tasks, the micro-tasks are simple and fast to finish (within a couple of minutes). The worker is also often motivated by tiny rewards every time they complete a micro-task.

Individuals who perform micro-tasks for micropayment is called *crowd workers* by (Deng et al., 2016). A study done on workers in the micro-tasking platform MTurk (section 2.4.1.1), says that the workers are representative for the general Internet user population, but are generally younger and have lower incomes and smaller families (Ipeirotis and G., 2010). Workers select which tasks to solve themselves, a task selection process is visualised in figure 2.3. Micro-tasks can also be assigned to workers, shown in figure 2.2. This can be if the task requires specific skills, age, language (Schulze et al., 2012). Most workers look for tasks that utilize their knowledge, skills, and abilities in the best possible way (Schulze et al., 2012).

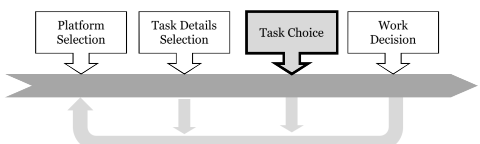


Figure 2.3: Micro-tasking workers task selection process (Schulze et al., 2012)

2.4.3 Micro-tasking usage

Micro-tasking and human computation have close ties. In the "Handbook of Human Computation," micro-tasking is strongly present in the *Human Computation for Disaster Response* chapter [(Meier, 2013b), p. 95-105], as well as in several other parts of the book. In the disaster response chapter, the authors give an overview of how human computation methods, such as paid micro-tasks, could be used to help in major disasters. In 2012, Philippines was struck by a typhoon called Ruby, devastating large regions. Through CrowdFlower the workers collected over 20 000 tweets related to the typhoon and identified the tweets containing links to either photos or video footage from the damaged areas. The photos or videos in the relevant tweets were tagged and geo-tagged by volunteers if they portrayed evidence of damage. Within 12 hours a dataset of 100 georeferenced images and videos were collected. It resulted in a very detailed crisis map shown in figure 2.4. This map was the first official crisis-map based solely on social media content [(Meier, 2013b), p. 101]. In the aftermath of this crisis, an algorithm was developed to automatically detect tweets that link to photos and videos, which freed more time for the volunteers to georeference and tag more images and videos portraying evidence of damage (Meier, 2014).

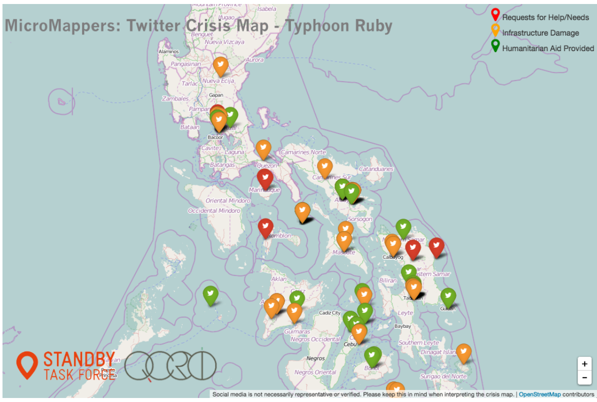


Figure 2.4: Typhoon Ruby Crisis map (Meier, 2014)

Micro-task crowdsourcing refers to a problem-solving model in which a problem or task is outsourced to a distributed group of people by splitting the task or problem into smaller sub-tasks or sub-problems. The sub-tasks is solved by multiple workers independently, often in return for a reward (Sarasua et al., 2012). Thanks to micro-tasking platforms as MTurk, it is possible to build a hybrid human-machine system that combines the scalability of computers with the yet unmatched cognitive abilities of the human brain (Difallah et al., 2016). Gadiraju et al. (2015) findings when analyzing data from MTurk, indicate rapid growth in micro-task crowdsourcing. With the establishment of micro-task crowdsourcing platforms as MTurk and CrowdFlower, micro-tasking is much more accessible. Micro-tasking practitioners are actively turning towards paid crowdsourcing to solve data-centric tasks that require human input

(Gadiraju et al., 2015). Most cases of micro-tasking combine human computation abilities with crowdsourcing.

Human computation systems often use crowdsourcing platforms to recruit workers (Schulze et al., 2012). Companies developing machine-learning algorithms has seen an advantage in combining human computation abilities and crowdsourcing with fast machine learning algorithms. An example is the team Tomnod[4] who did a project in Australia where they combined human computation, crowdsourcing, and machine learning to locate swimming pools (Kostas, 2016). The machine learning algorithm classified polygons where there was likely to be a swimming pool inside. The crowd who participated in finding the swimming pools were only presented with the classified polygons, which minimized the search area and then also the time required finishing the job (Kostas, 2016). Examples of classified polygons are shown in figure 2.5. The Tomnod team divided the task into smaller tasks, they micro-tasked the work. The resulting dataset was then used to train a swimming pool detecting convolutional neural network (Nikki, 2016).

Figure 2.5: Classified polygons created by the algorithm (Nikki, 2016)

New York Public Library uses micro-tasking to train computers to recognize building shapes and other data on digitized insurance atlases. The micro-tasks is used to check the computer's work and also to capture information the computer missed. Individuals contributing checks and fixes building footprints drawn by the computer. The individuals also enter addresses and classify the building footprints using colors. To ensure accuracy the same footprints are shown to several people. At least three different individuals check the same footprint, and 75% or more must agree on the footprint for the answer to be approved.

Most cases of micro-tasking usage exploit the large volume capabilities machines have and the cognitive capabilities of humans (Difallah et al., 2016). One of the advantages of micro-tasking platforms like MTurk, Tomnod and CrowdFlower, mentioned by Meier (2013b) (p. 99), is the built-in quality control mechanisms that ensure a

[4]Tomnod is a team of volunteers who work together to identify important objects and interesting places in satellite images; www.tomnod.com

relatively high quality of output data. They set a review constraint, for instance in a project where they tagged satellite imagery of Somalia each unique image was reviewed by at least three different volunteers and only when all three agreed on type and location it was approved.

2.4.4 Building imports in OpenStreetMap using micro-tasking

In OpenStreetMap (OSM), at least two large building imports have been successfully achieved using micro-tasking (Erichsen, 2016). The first was an import in New York, the second in Los Angeles. Both import teams divided the building dataset into smaller parts to lower the complexity. They used the same python script to create the micro-tasks containing small chunks of building data. Having small chunks of building data made it possible to review and import the data manually into the OSM database (Barth, 2014a). In New York, they imported one million buildings partitioned into 5258 micro-tasks (Barth, 2014a). In Los Angeles, their dataset contained three million buildings (Sambale, 2016). We have not found how many micro-tasks that were created in total. In both projects, all buildings needed to be quality checked and merged correctly with existing data in OSM. This validation process is done manually in OSM (community, 2017a). Two massive imports jobs to coordinate. Both building imports used the tasking manager platform (2.4.1.2) to organize and distribute the micro-tasks [(Barth, 2014b), (community, 2017b)]. The number of buildings in each micro-task varied because the building dataset was partitioned based on already existing subregions which had various building density (Erichsen, 2016).

A challenge during the New York building import was the underestimated complexity of the import job. It started as a community import where every user in OSM could contribute. Due to the underestimated complexity of the review and upload tasks, and time spent training and supporting new individuals, the team loosely formed a group around the import (Barth, 2014a). The company Mapbox participated in the import with experienced team members. The Mapbox members outpaced the local volunteers by a huge factor (Barth, 2014a). Barth (2014a) writes that the tasks was not hard but demanded a certain learning curve which meant time someone had to spend teaching new volunteers. Having few experts doing the micro-tasks was evaluated as more effective than involving the whole OSM community. The Los Angeles building import allowed everyone to work on the micro-tasks. They developed the Tasking Manager 2, adding new features to fit their needs. Over 100 volunteers contributed to the import job, working on the micro-tasks posted on the Tasking Manager 2 platform (Sambale, 2016). February 2017, all three million buildings were successfully imported into OSM (contributors, 2017). When examining micro-tasks in both import projects, the overall micro-task in Los Angeles contained fewer buildings than in New York's micro-tasks. Erichsen (2016) claim that the micro-tasking method is the best-known method when importing large datasets into OSM. This gives a good indication of how useful and functional this method is.

2.4.5 Micro-tasking limitations

Getting enough people to use the micro-tasking platforms is crucial for its success. Most of the platforms mentioned in this chapter give payments to the workers. Another option is to make the platform as a game, which is also shown in this chapter. Creating a micro-tasking platform without payments or gamification factors the page is likely to have a short life, even though the tasking manager, supported by HOT, is an exception to this rule.

A problem when combining machines and humans is that machines can do their operations in real-time, while humans are unpredictable, they can come and go as they wish. This creates a gap where the micro-tasking platforms cannot guarantee on the task completion time (Difallah et al., 2016).

The human computation abilities can also be overestimated. During the classification of swimming pools in Australia, the Tomnod team faced some unexpected challenges. As described in section 2.4, they used the crowd to classify if a polygon contained a swimming pool or not, an algorithm had pointed out the polygons first. When reviewing a random sample from the result, they found an indication that 26% of polygons that contained a pool were identified as not containing pools by the crowd (Kostas, 2016). Further studies also showed that the guilty part was the crowd, the algorithm had correctly detected polygons containing pools. In a case where the algorithm was 85% confident that the polygon contained a pool, only one voted 'yes', six voted 'no, this polygon do not contain a pool'. The solution was to combine the human verdict with the machine's prediction. This example shows that it is important to use the right combination of humans and machines. Tasks that at first seems simple to do for humans, may be more challenging that expected. Basic object detection using machine learning perform very well when used together with human operations.

It is important that the tasks added to a micro-tasking platform consider the talents and limitations of human workers (Franklin et al., 2011). By using knowledge provided from OpenStreetMap of the usage of the micro-tasking, in Erichsen (2016), this thesis will examine and hopefully reveal the limitations and talents of human workers when dealing with geospatial micro-tasks.

3 Development of the Experiment

This chapter will introduce the developed experiment and the implemented web-application hosting the experiment. The experiment generates the data for the analyses, so it is important to implemented and executed the experiment correctly. A pilot-test of the web application is used to secure the usability and discover weaknesses in the implementation. How the pilot-test was executed and some preliminary results, is also presented in this chapter. At last, there is a section about how to determine the sample size.

3.1 Experiment

The developed experiment contains three different tasks that each participant needs to complete. In each task, they answer the same two questions. The three tasks will vary the number of elements the participant has to use to complete each of the two questions. The questions represent two different micro-tasks involving geospatial data.

3.1.1 The two micro-tasks

Gadiraju et al. (2015) categorize the top-level crowdsourced tasks, after analyzing platforms as MTurk and CrowdFlower. It resulted in three relevant classes within geospatial data: *1)Verification and validation*, *2)Interpretation and analysis* and *3)Content creation*. There are examples of all three task classes in geospatial crowdsourcing. In humanitarian OpenSteetMap, they map areas during a crisis to support the help organizations through crowdsourcing, creating valuable content to the workers in the field (2.4.1.2). In a machine learning process, they are starting to use micro-tasks to both validate the created data and also create data sets to train the algorithms (2.1). *Interpretation and analysis* tasks rely on the individual to use their interpretation skills during task completion. The micro-tasks given in the building imports from section 2.4.4 was primarily interpretation and analysis tasks (Erichsen, 2016). The experiment conducted in this thesis use Gadiraju et al. (2015) three classes and the building imports described in Erichsen (2016) as guidance when the two micro-tasks given to the participants was developed.

The first question asks the participant to click on the color that fits the shape of the marked building(s) on the map best. Here the participant is given two footprint layers covering a building and needs to determine which of the footprints that fit the building shape shown on the base map best. This task is highly relevant during building imports and data validation if one has two overlapping geometry layers. Question one displayed on the web application is presented in figure 3.1. In this example, the participants have three buildings to select before the micro-task is completed.

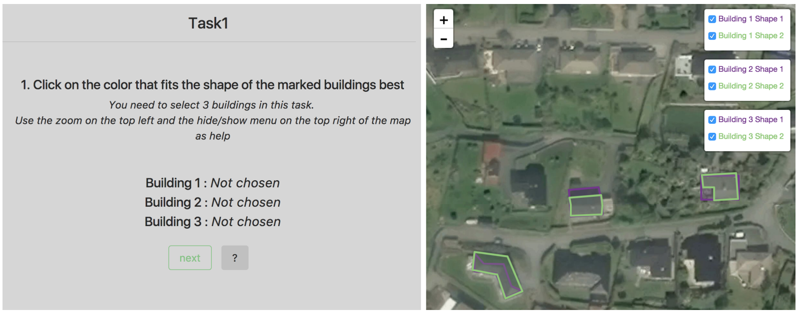
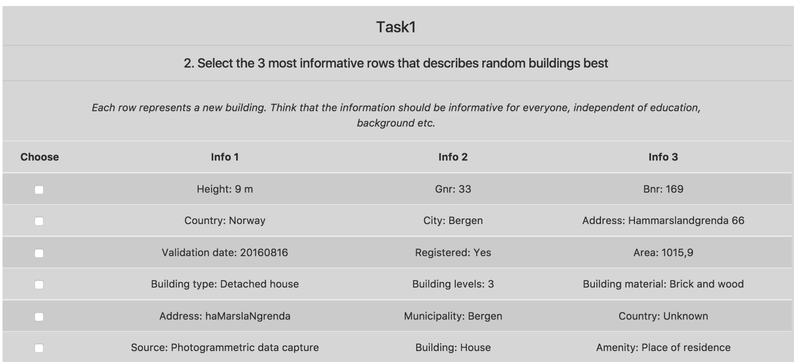


Figure 3.1: Question one as it is displayed in the web application during task B

The second question asks the participant to select the 1/3/6 most informative row(s) that describes random buildings best. Question two is an interpretation task where the participant needs to interpret the information written in the table to decide which row(s) gives the most informative information about an arbitrary building. This is hard for computers do to. What the most informative information is will vary between buildings and what information is present. In figure 3.2 question two is displayed in the web application. In this example, the participant has to choose three rows from the table before the micro-task is completed.


Task1

2. Select the 3 most informative rows that describes random buildings best

Each row represents a new building. Think that the information should be informative for everyone, independent of education, background etc.

Choose	Info 1	Info 2	Info 3
	Height: 9 m	Gnr: 33	Bnr: 169
	Country: Norway	City: Bergen	Address: Hammarslandgrenda 66
	Validation date: 20160816	Registered: Yes	Area: 1015,9
	Building type: Detached house	Building levels: 3	Building material: Brick and wood
	Address: haMarslaNgrenda	Municipality: Bergen	Country: Unknown
	Source: Photogrammetric data capture	Building: House	Amenity: Place of residence

Figure 3.2: Question two as it is displayed in the web application during task B

3.1.2 The three experiment tasks

The experiment consists of three different tasks, in addition to a training task. When determining the number of elements in the three tasks the author decided to base this on cognitive load theory. Cognitive load theory refers to the total amount of mental effort being used in the working memory. Working memory is determined by the number of information elements that need to be processed simultaneously within a certain amount of time (Barrouillet et al., 2007). A heavy cognitive load can have disadvantageous effects on task completion. The cognitive load that is imposed by a task is much higher for beginners than for more advanced students (Leppink et al., 2014).

The survey will contain three tasks, each task contains six elements, but the task varies how many elements the participant need to answer at the same time. One task will serve the participant with one and one element, called task A, demanding the smallest cognitive load. The other task will serve the participant with three and three elements at the same time, called task B. This number is just under the limit of how much information humans can process. The last task will serve the participant with all six elements at the same time, called task C. This number exceeds the human capacity when processing information according to Leppink et al. (2014). An illustration of the tasks is shown in figure 3.3.

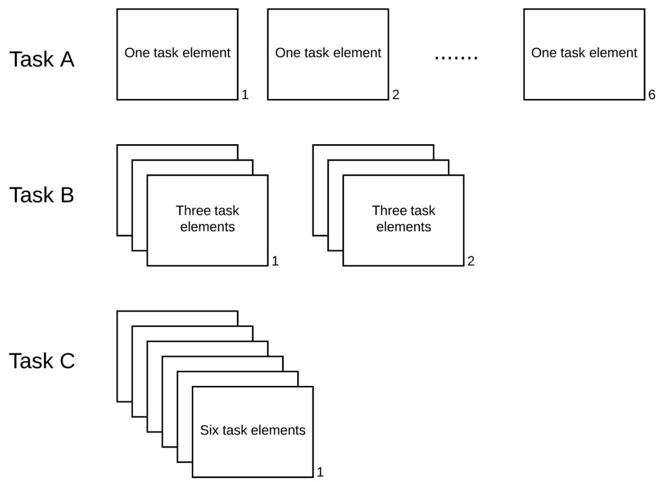


Figure 3.3: Illustration of the three tasks given in the experiment

It is stated that the working memory has a limited capacity of seven plus or minus two elements (or chunks) of information when merely holding information and even fewer (ca four) when processing information (Leppink et al., 2014). By choosing three elements in one task and six elements in the other task, this paper can determine if the theories about the limited capacity of the human brain also apply to micro-tasks involving geospatial data. One task will only contain one element as a minimum cognitive load task. The three tasks can help answer how many elements a human can process at the same time without impacting the quality of the result. The goal is to determine a preferred number of elements to include in a micro-task. This information can be useful when developing micro-tasks to achieve adequate task progress as well as accurate results.

The "magical" number of 4 has been demonstrated to limit much of human information processing (Mandler, 2013). It is said that polygon comparison demand medium cognitive load (Kiefer et al., 2016), which is what the participants do in the first question in the experiment. Kiefer et al. (2016) argues that high cognitive load may lead to less efficient map reading and spatial orientation, as well as decreased spatial learning. Since polygon comparison demand medium cognitive load, question one should at least not be too demanding on the one- and the three elements task. A worry is that the inexperienced participants will have a bigger struggle than the experienced participants. The extraneous cognitive load imposed high for the inexperienced when solving problems because their lack of prior knowledge of how to solve that type of problem forces them to resort to weak problem-solving strategies (Leppink et al., 2014). By dividing the participants into experienced and inexperienced categories, the results from the experiment can help determine if geospatial micro-tasks are too demanding on inexperienced individuals.

3.1.3 Determining the building footprints used in question one

According to Fan et al. (2014), there was over 77 million buildings in the OpenStreetMap (OSM) database in 2013. A study of the geometries of building footprints in the city Munich reveal a huge diversity in the geometries (Fan et al., 2014), and this is probably not the only city with this kind of diversity. The Fan et al. (2014) paper used four criterion's, completeness, semantic accuracy, position accuracy and shape accuracy to evaluate the quality of the building footprints in OSM. In the creation of the two footprints representing the buildings used in question one, the quality criterion's shape- and position accuracy was emphasized. The goal is to create shapes that match realistic cases that occur for instance in OSM.

Shape accuracy evaluates how well the layer matches the building in an aerial image. Fan et al. (2014) mentions two main reasons to why building footprints are simplified in OSM. The first reason is the difficulties following building details when looking from a bird's eye view. The second reason is the limited resolution of the Bing aerial image used in OSM during digitalization. In question one two footprints is drawn with one of them matching the building shape better than the other. The participant has to use an aerial image to determine which layer fits the building best. This will test if the participants manage to make correct shape judgments by only using an aerial image as a reference.

(a) Official building footprint (b) Drawing a footprint (c) Both building footprints

Figure 3.4: Creating of footprint layers used in question one

Position accuracy evaluates how well the coordinate value of a building relates to the reality on the ground. Fan et al. (2014) tested the accuracy of buildings in OSM and concluded with a mean offset of 4.13 m. The low positional accuracy of OSM building footprints data is caused by the limited resolution of Bing map images. By combining shape- and position accuracy in some of the cases used in question one this study can also determine if participants manage to evaluate both factors. In this study, the participants do not have available information about what the true ground coordinates are. Therefore position accuracy will be examined by shifting one of the layers. The correct positional accuracy will be at the building in the aerial image.

3.2 Web application

This thesis used an online web-based survey to conduct the experiment. An online survey avoids the cost and effort of printing, distributing, and collecting paper forms. Many people prefer to answer a brief survey displayed on a screen instead of filling in and returning a printed form (Ben and Plaisant, 2009). The participants do not have to share the same geographic location as the researcher.

An online web environment also makes it easier to use interactive maps. An interactive map is necessary to answer question one. It is not possible to have an experiment involving interactive maps on a piece of paper. The web is the obvious way of implementing interactive maps. Making it online, available via URL, makes the distribution faster and easier.

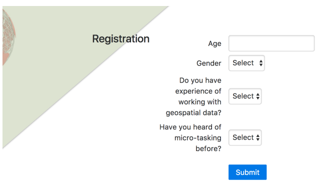
A common web programming language is JavaScript, together with the library React[1] creates the client side of this thesis application. The client communicates with a server that fetches the task elements from and saves the task results in a PostGIS database. The server is written in Python with the framework Django[2]. The PostGIS database contains the task elements, and also the task results gathered from the participants.

[1] React is a open-source JavaScript library for building user interfaces. In React, the displayed data can change without reloading the page. It's main goal is to be fast, simple and scalable.

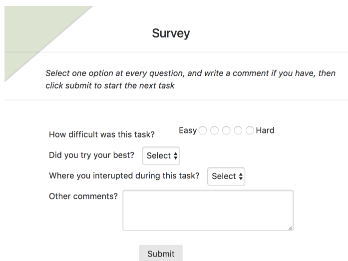
[2] Django is a open-source web framework written in Python. It's primary goal is to ease the creation of complex, database-driven websites.

3.2.1 React application

The React application was created to serve the experiment to all participants. It contains all the steps of the experiment. First, the participant registers, giving information about age, gender and answers yes or no on the following two questions: 1) "Do you have experience of working with geospatial data?", 2) "Have you heard of micro-tasking before?". Next, the participant is given an introduction page with a detailed introduction video describing how to answer the two questions, on how to interact with the map and building layers. A training task comes after the instruction video, in the training task the participant solves both questions just like the normal tasks, only it contains different building footprints and only two elements in each task to not replicate the other tasks. After the training task, the participant starts with the experiment containing the three tasks, with a short survey after each task. The survey asks the participant to rate the task difficulty between one and five, and if the participant tried his/hers best or was interrupted during the task. The participant can also write a comment. The interfaces for the two questions used in each task is shown in figure 3.1 and 3.2 in section 3.1.1. The registration form and survey interface is shown in figure 3.5.



(a) Registration site



(b) After each task survey site

Figure 3.5: Interface for the client application

To ensure random and independent observations multiple measurements were implemented during the development of the React application. The main measurements to ensure random, independent observations was:

- Random order on the three tasks
- Random buildings in the tasks
- Random color on the building footprints
- The building footprints was drawn on the map in a random order
- Random order on where the information rows was written in the table

When distributing an online experiment, the result can be inconsistent since the researcher is not present to either control the participant or the environment surrounding the participant. The researcher implemented functions securing the completeness of the data. The buttons navigating to the next question and task was disabled until enough building footprints or rows were selected. The participant could not submit their task-result before answering both questions correctly. The submit button on the register form and after each task survey only submitted the answers if all fields were answered. If a field missed a message appeared asking the participant to fill out the entire form before submitting. All registered results were complete in the database thanks to these measurements.

3.2.2 Data acquisition

In this study, the independent variables are: 1) experienced or inexperienced participant, 2) number of elements in the task, 3) age and 4) gender of the participant. Independent variables are factors we think might influence the results of the study [(Kitchin and Tate, 2000), p. 49]. When a participant registers at the start of the survey the independent variables is generated. Participants who answer yes to the question "Do you have experience of working with geospatial data?" are registered as experienced. The independent variables are believed to influence the dependent variables. Dependent variables are factors the study is interested in explaining [(Kitchin and Tate, 2000), p. 49]. In this study, the dependent variables are: 1) time spent on each question and task, 2) the number of correctly chosen elements in each question and task, and 3) how difficult the participant though the task was.

The React application generates and saves the task results. In both questions, the time and number of correct elements are registered. The React application has a timer that registers time elapsed on both questions and adds the time measurements together to get the total time spent on the task. The time only reflects how long the participants spent solving the two questions. Time spent loading new layers, moving to the next question, etc., is not included. Which building footprints and rows the participant selects is also registered, and before saving the result counts the number of correctly chosen elements in each question and then adds the number together to get the total number of correctly chosen elements in the task. A participant can maximum have twelve correct elements, six from each question. Total time and number of correct elements are the two primary dependent variables, and they create the basis of the statistical analyses together with the independent variables mentioned above. Participants task results are saved after each task and contain the following information:

- Task number (which task)
- Task order number (which order)
- Time spent on question one
- Time spent on question two
- Total time spent on the task
- Correct elements in question one

- Correct elements in question two
- Total correct elements in the task

After each task the participants answers a short survey. This information was used to remove task results where the participants was interrupted. The difficulty question can be used to determine if one of the three tasks is preferred by the participants.

3.3 Pilot test

Testing the experiment before actual use is highly recommended (Ben and Plaisant, 2009). A pilot test provides an opportunity to validate the wording of the tasks. It also helps understand the time necessary for completing the survey, which should be communicated to the participants (Schade, 2015). The pilot test was carried out on a small sample of users. Results from the pilot test were in this thesis used to make improvements to the actual survey, to the react application and to find errors or weaknesses in the database models.

After the pilot test, the usability was measured. Usability in this thesis was measured with the *System Usability Scale*(SUS) because it gives a subjective measure of usability. The *System Usability Scale* questionnaire consists of ten statements where the participants rate their agreement on a five-point scale (Ben and Plaisant, 2009). Subjective measure of usability is usually obtained through the use of questionnaire and scales (Brooke, 1996). SUS was developed to be quick and straightforward, but also reliable enough to be able to compare performance changes between versions (Brooke, 1996). It is also easy to administer the participants through the usability test, and it can be used on small sample sizes and still give reliable results (Affairs, 2013).

The usability is important to measure. If the participants do not understand how the web application works, they will probably not do the survey since they then have to invest time in understanding what to do. It is also important to get enough participants to do the whole survey and not quit halfway in frustration of not understanding it properly. The *System Usability Scale* can effectively differentiate between usable and unusable systems (Affairs, 2013).

3.3.1 Execution of the pilot test

The pilot test was conducted with a total of eight participants, five experienced and three inexperienced participants aged from 22 to 64 years. The test started with a brief information about this study and the experiment. They were told to talk out load during the test, no help or guidance was given to the participants. The participants was observed while they conducted the survey. The author took notes and watched if the participants understood the questions and tasks correctly. After the survey a *System Usability Scale* questionnaire was answered by the participants. In the end, the participants were asked to give general feedback on the web application. The

SUS score and the feedback were then used to determine the usability of the React application and to determine which improvements to be done.

3.3.2 Results from the pilot test

The average SUS score was 84.64 out of 100. Anything above 68 is considered above average (Affairs, 2013). When adding the SUS score to an adjective rating will an score of 85.5 or higher be described as excellent (Bangor et al., 2009). A score of 84.64 is then described as good/excellent. This result gives a strong indication that the React application is user-friendly.

All participants thought that the instruction video was confusing. It was short, the instructions went too fast, and it missed voice descriptions. The instructions needed major improvements, an important discovery. The purpose of the video is to give the participant an introduction to how to answer the two micro-tasks. It should include important instructions, particularly useful for participants not used to working with interactive maps.

Overall feedback on question one was that it is hard to understand which building was which because of missing labels, and also to know when a building footprint was selected or not. The lack of labels on the buildings was done on purpose to get the task as much as possible realistic to traditional GIS programs (i.e., QGIS). The process of selecting the best fitting building footprint needed improvements. It had to be clearer that one had to click on the layer on the map to choose a footprint, not by using the layer control as some thought. This part was added to the video with voice description, describing in detail how a footprint was selected. Improvements to the design of the question one page was made by adding the selected footprint's color to the text telling the participant which layer they had chosen, giving a better visual feedback.

Another feedback from one of the participants was that both question views had too much information and long sentences. The participant advised to shorten the sentences and to move some of the information to the video.

The test data was used to examine some of the hypothesis to find errors or weaknesses in the database model. The data was extracted from the PostGIS database and saved in CSV files. Some preliminary results can be seen in section 3.3.3. There were a few errors and weaknesses found during the statistical tests. Changes to the database models were necessary, and the implemented changes are listed under:

1. Add foreign key from TaskResult model to TaskSurvey model
2. Added four other fields in TaskResult model
    - Total correct elements
    - Task order
    - Task number
    - List of correctly chosen building numbers in both questions

3. The difficulty field in Tasksurvey model was changed from Char field to Integer field

3.3.3 Preliminary results

The pilot-test data was not normally distributed, and doing statistical analysis on data from eight participants didn't seem relevant. The data was mapped in char plots, visualizing some trends.

The two oldest participants spent almost twice as much time on the test than the younger. Maybe it was too much cognitive load on them. Learning a new application and at the same time understanding how to do the survey and answer the questions given to them. One of them were experienced and the other inexperienced, so this is a surprising result. Figure 3.6 show the task results from all participants ordered by age. There are three entries per participant, so three and three bars are results from the same participant. Task A represents the task with one elements, task B the task with three elements and task C the task with six elements. It is clear to see that the 54 year old participant's total time on task B is dramatically higher than the rest.

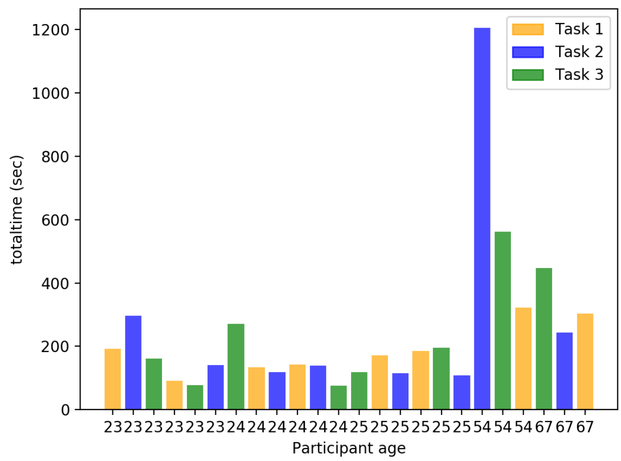


Figure 3.6: Total time - all participants ordered by age

The average time spent on the survey was 18 minutes. The two oldest participants used on average 33 minutes, while the rest of the participants spent on average 13 minutes to complete the survey.

In the pilot test, the same building footprints (question one) and information rows (question two) were used in all three tasks. At the end of the pilot test, the author asked the participants if they remembered the buildings and meta information from the previous tasks. $\frac{7}{8}$ answered yes on the question. This information was valuable.

If every participant does a better job at the last task, because they remember the elements from earlier tasks, the result will not be as useful. Reading the data in figure 3.6, $\frac{6}{8}$ participants spent less time on the last task, even though the task order varied. This almost matches the number of participants who remembered the previous elements in the last task. This finding made the author create three different task element groups. The three element groups were randomly assigned to each task, to avoid the thee building groups to influence the result. The risk of participants remembering previous task elements disappeared with this decision. Each task will contain new building footprints and meta information.

3.4 Determining the sample size

The sample size is influenced by various factors, including the purpose of the study, population size, the risk of selecting a "bad" sample and the allowable sampling error (Israel, 1992).

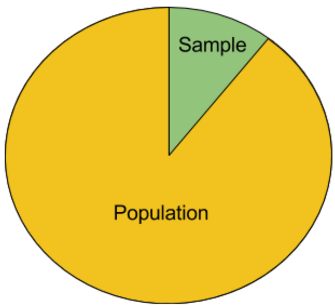


Figure 3.7: Population vs. sample

A sample is a collection of observations and is the subset of a population, illustrated in figure 3.7. The population size in this survey is not easily determined. A population is the collection of individuals of a particular type (Walpole et al., 2012). All individuals with access to a computer and the internet interested in contributing to micro-tasks can be one description of the population. It is important that the sampled population and the target population is similar to one another.

There are three possible ways of determining the sample size in this study. The first option is to use a sample size from a similar study. The risk is to repeat errors that were made in determining the sample for another study. The second option is to rely on published tables, depending on precision, confidence levels, and variability. According to Israel (1992) table 1, an accuracy of 0.05, confidence interval of 95% and a size of population greater than 100'000, the necessary sample size is 400. If the accuracy is changed to 0.1, the sample size necessary increases to 100 (Israel, 1992). The numbers found in the table must reflect the number of obtained responses. The last approach is to use formulas to calculate the sample size. The formulas require the standard deviation and how much variance to expect in the response (Smith, 2013)(Israel, 1992). Israel (1992) mentions that the table gives a useful guide for determining the sample size and that formulas are used if the study has a different combination of precision and confidence. This study will use the table result since the combinations match this study.

It is important to mention that the quality of the sample is as important as the size. The more variable the sampled data is, the larger the sample size is required (Israel, 1992). It is also desirable to choose a random sample, which means that the observations are made independently and random. The main purpose of using a random sample is to obtain correct information about the unknown population parameters (Walpole et al., 2012).

4 | Result

This chapter will present the result of the experiment conducted. It will summarize who the participants were, the gathered data and then present the statistical results from the analysis performed on the gathered data. The researcher used eight days to collect enough participants completing the experiment. Only 38% of the registered participants completed the experiment. All task results saved in the database was valid and could be used in the analyses. The analyses were calculated in Python, using statistical packages as SciPy, Numpy, and Panda. The data was extracted from the database using Django Queryset and saved in CSV files. On the authors GitHub, in the repository *thesis-statisticmethods*, the implemented statistical methods are available[1].

4.1 Participants

Benefits of using an online based experiment are that it has potential to reach a huge number of people, the problem is how to reach out to the people to make them aware of the existence of the application. Using mailing threads and sites the researcher had available was this studies solution. The participant's contribution to the survey was contacted through email and the organization Geoforum's website and Facebook page. All students at *Civil and engineering* was emailed, as well as a mailing list reaching out to the Norwegian OpenStreetMap community. The web application was also published at the Geoforum website (www.geoforum.no) and Facebook page. Geoforum is a Norwegian association for individuals and companies working in the field of geomatics. The participants receiving the email or looking at the Geoforum site could click on the web application URL and access the experiment from there.

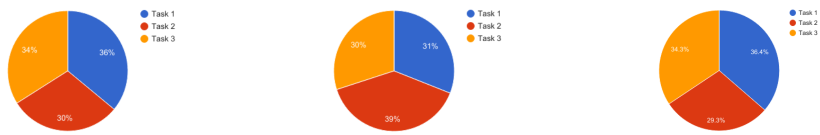
There was in total 461 task results in the database after the gathering period, where almost all participants contributed with three task results (the training task was removed). 402 participants registered on the website during the data gathering period and only 38% of the registered participants completed the survey. This number was surprisingly low, but time to complete the whole experiment can probably explain why so few completed. It probably lasted to long for participants to have the patience to finish. 152 participants completed all three tasks. Results from participants not completing all three tasks are also included in the dataset. Including these result is not a problem. The three tasks were given to each participant in a random order, and the tasks are also independent, containing different buildings and rows.

The mean participant age was 31.5 years and the median 25 years. The youngest participant was 19 and the oldest 58 years. 33% of the participants was female and 66% male. The average male was 33.5 years old and the average female 31.2 years

[1] https://github.com/annesofie/thesis-statisticmethods

old. 19% of the participants that completed the survey said they had heard of micro-tasking before. 53% of the participants stated that they had experience of working with geospatial data. The distribution between experienced and inexperienced participants was approximately even. A very pleasing distribution.

Random and independent observations are essential, and a random task order was used to ensure this. Analyzing the task results and the order the tasks was presented in gives a pleasing result. The distribution of how many times the three tasks (Task 1, 2 and 3) was first, second and last task in the analyzed data is approximately evenly distributed. This is shown in figure 4.1.



(a) First task (b) Second task (c) Third task

Figure 4.1: The distribution of the task order in the analysed data

4.2 Statistics theory

This section will give an introduction to the statistics used in this thesis. The thesis will examine the data with parametric methods but also with non-parametric methods if the assumption of a normally distributed samples fails. A nonparametric method is much more efficient than the parametric procedure when the set of data used in the test deviates significantly from the normal distribution (Walpole et al., 2012). There are also some disadvantages using nonparametric methods. The methods will be less efficient, and to acchieve the same power as the corresponding parametric method a larger sample size is required. If parametric and nonparametric tests are both valid on the same set of data, the parametric test should be used (Walpole et al., 2012).

4.2.1 Normal testing

The sampling distribution of a statistic depend on the distribution of the population, the size of the samples, and the method of choosing the samples (Walpole et al., 2012). Sampling distibution describes the variability of sample averages around the population mean μ. All parametric statistics assumes normally distributed, independent observations. Parametric tests are preffered in statistics because it got more statistical power than nonparametric tests (Frost, 2015). The power of a test is the probability of correctly rejecting a false null hypothesis, which in this case is the ability to detect if the sample comes from a non-normal distribution. To determine if a sample is normally distrubuted there exists both visual methods and normality tests to assess the samples normality. A visual inspection of the sample's distribution is usually unreliable and does not guarantee that the distribution is normal (Pearson et al., 2006). Presenting the data visually gives the reader an opportunity to judge the distribution themselves. In this thesis histograms are used to visually analyse the data for normality.

Normality tests compare the scores in the sample to a normally distributed set of scores with the same mean and standard deviation (Ghasemi and Zahediasl, 2012). There are multiple normality tests, and deciding which test to use is not easy. This study needs a test that doesn't require every value to be unique, a test that can handle ties (identical observations). The application used to collect the samples in this study do not guarantee unique values.

The D'Agostino-Pearson omnibus test stand out as the best choice. This test first computes the skewness, figure 4.2, and kurtois, figure 4.3, to quantify how far from the normal distribution the sample is from the terms of assymetry and shape. Then it calculates how far each of these values differs from the value expected with a normal distribution (Pearson et al., 2006). It works well even if all values are not unique (Motulsky, 2013). The test also works well on both short- and long-tailed distributions (Yap and Sim, 2011).

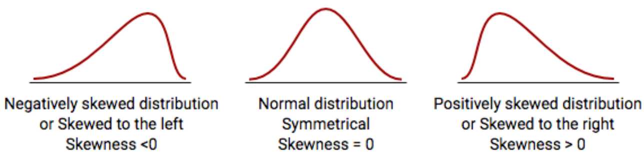


Figure 4.2: Skew (MedCalc Software bvba, 2017)

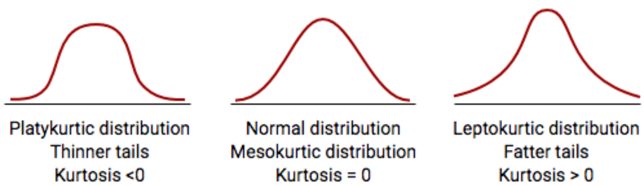


Figure 4.3: Kurtois (MedCalc Software bvba, 2017)

The D'Agostino-Pearson test uses the following hypothesis:

H_0: The data follows the normal distribution
H_A: The data do not follow the normal distribution

For small sample sizes, normality tests have little power to reject the null hypothesis, therefore small sample sizes most often pass normality tests. For large sample sizes, significant results would be derived even in the case of a small deviation from normality (Pearson et al., 2006). When the null hypothesis cannot be rejected, then there are two possible cases. First case is to accept the null hypothesis or the second case is that the sample size is not large enough to either accept or reject the null hypothesis (The Pennsylvania State University, 2017). An acceptance of the null hypothesis implies that the evidence was insufficient, the result does not necessary accept H_0, but fails to reject H_0 (Walpole et al., 2012). Therefore will both visual analysis and D'Agostino-Pearson test assess the normality assumption.

4.2.2 Hypothesis testing

The null- and alternative hypothesis are statements regarding a difference or an effect that occur in the population of the study. The alternative hypothesis (H_A) usually represents the question to be answered or the theory to be tested, while the null hypothesis (H_0) nullifies or opposes H_A (Walpole et al., 2012). The sample collected in the study is used to examine which statement is most likely (technically it is testing the evidence against the null hypothesis). When the hypothesis is identified, both null and alternative, the next step is to find evidence and develop a strategy for or against the null hypothesis (Lund Research Ltd, 2013a).

The next step is to determine the level of statistical significance, often expressed as the *p-value*. A statistical test will result in the probability (*the p-value*) of observing your sample results given that the null hypothesis is true. A significance level widely used in academic research is 0.05 or 0.01 (Walpole et al., 2012). 0.05 significance level will be used in this thesis analyses.

The result should not be reported as "significantly different," but instead report it as "statistically significantly different." This is because the statistical decision as to whether the result is significant should not be based solely on the statistical test. To indicate to readers that the result is a statistical one, include statistically in the conclusion sentence (Lund Research Ltd, 2013c).

4.2.2.1 Two-sample t-test

When estimating the difference between two means a two-sample t-test is used (Walpole et al., 2012). A two sampled test assumes two independent, random samples from distributions with means [μ_1 , μ_2] and variances [σ_1^2, σ_2^2]. The hypothesis tested on two means can be written as

H_0: Sample means are equal
H_A: Sample means differ

The two-sample t-test is used to estimate if differences between two means are significant. In a two-sample, two-sided, t-test (μ_1 - $\mu_2 \neq 0$) the null hypothesis is rejected when [(Walpole et al., 2012), p. 345]:

$$|T| > t_{\frac{\alpha}{2},v} \tag{4.1}$$

In a two-sample, one-sided, t-test the null hypothesis is rejected when [(Walpole et al., 2012), p. 350]:

$$T > t_{\frac{\alpha}{2},v} \tag{4.2}$$

$$T < -t_{\frac{\alpha}{2},v} \tag{4.3}$$

Equation 4.2 is used on one sample test where the alternative test is to check if the mean is greater than zero (μ_1 - μ_2 > 0), and the 4.3 equation is used on hypothesis where the test is to check if the mean is lower than zero (μ_1 - μ_2 < 0). T is the calculated statistical value and t is the critical value with the given significance level (α) and degree of freedom (v). The critical value is found in the table of Critical values for t-distribution.

Before doing tests on the two means, the Levene's Test is used to test if the samples are from populations with equal variances. It tests the hypothesis:

H_0: Input samples are from populations with equal variances
H_A: Input samples are from populations that do not have equal variances

If we can assume equal variances in the two samples and the samples are normal distributed, a two-sampled t-test may be used.

Hypothesis in this study that can be tested with a two-sampled t-test (if the conditions mentioned above are valid) is listed in figure 4.4.

1
H_0**:** No difference in time spent on the tasks between the participants
H_A**:** Time spent differs between experienced and inexperienced participants

2
H_0**:** Experienced participants do not finish the tasks faster than inexperienced
H_A**:** Experienced participants finish the tasks faster

3
H_0: Equal number of correct elements between experienced and inexperienced
H_A: There is a difference in number of correct elements between the two groups

4
H_0: Experienced participants do not have more correct elements
H_A: Experienced participants have a higher number of correct elements

Figure 4.4: Two-sample t-test hypothesis

4.2.2.2 Analysis-of-Variance

Analysis-of-Variance ($ANOVA$) is according to Walpole et al. (2012) a very common procedure used for testing population means. Where a two sample t-test is restricted to consider no more than two population parameters, $ANOVA$ can test multiple population parameters. A part of the goal of $ANOVA$ is to determine if the differences among the means of two or more samples are what we would expect due to random variation alone, or due to variation beyond merely random effects. $ANOVA$ assumes normally distributed, independent, samples with equal variance. The equal variance assumtion will be tested with Levene's Test also mentioned in section 4.2.2.1.

One-way $ANOVA$ tests the null hypothesis that two or more groups have the same population mean given that the mean is measured on the same factor or variable in all groups (Lund Research Ltd, 2013c). The hypothesis test can be written as:

H_0: $\mu_1 = \mu_2 = ... = \mu_k$
H_A: At least two of the means are different

μ equals the group mean and k represents the number of groups. It is important to check that each group are normally distributed (Lund Research Ltd, 2013c). The weakness of one-way $ANOVA$ is that it cannot tell which specific groups were significally different from each other if H_0 is rejected. To be able to determine which group a *post hoc test* is used. The null hypothesis is accepted if:

$$F(obtained) < f_{\alpha,v_1,v_2}(criticalvalue) \tag{4.4}$$

If the alternative hypothesis is accepted a *post hoc test* is used. A post hoc test makes paired comparisons to determine which groups differs. This thesis will use Tykey's test to determine which group means are significantly different [(Walpole et al., 2012), p.526]. Hypotheses in this study that are tested in an one-way *ANOVA* analysis (if the conditions mentioned above are valid) is listed in figure 4.5.

1

H_0: Task time do not differ between the three tasks
H_A: Task time differ between at least two of the tasks
Variable = time, group = tasks

2

H_0: Correct elements in each of the three tasks do not differ
H_A: Correct elements between at least two of the tasks differs
Variable = Number of correct elements, group = tasks

Figure 4.5: One-way ANOVA hypothesis

The hypothesis written above is tested in section 4.3.4.3.

4.2.2.3 Wilcoxon Rank-Sum test

The Wilcoxon Rank-Sum test is an appropriate alternative to the two-sample t-test (see subsection 4.2.2.1) when the normality assumptions do not hold, but the samples are still independent and have a continuous distribution (Walpole et al., 2012). Since this method is non-parametric (or distribution-free) it do not require the assumption of normality.

The hypothesis for Wilcoxon Rank-Sum Test is:

$$H_0\text{: } \tilde{\mu}_1 = \tilde{\mu}_2$$
$$H_A\text{: } \tilde{\mu}_1 > \tilde{\mu}_2 \text{ or } \tilde{\mu}_1 < \tilde{\mu}_2 \text{ or } \tilde{\mu}_1 \neq \tilde{\mu}_2$$

The alternative hypothesis depends on what the test should determine. If the sample with mean $\tilde{\mu}_1$ is greater than, smaller than or unequal to the sample with mean $\tilde{\mu}_2$. First select a random sample from each population with means $\tilde{\mu}_1$ and $\tilde{\mu}_2$. If the sample sizes are different, let n_1 be the number of observations in the smallest sample and n_2 for the largest sample. Then $\tilde{\mu}_1$ will be the mean for the smallest sample. If there are ties (identical observations) in the sample a Mann-Whitey U test is preferred (The Scipy community, 2017).

4.2.2.4 Mann-Whitey U test

The Mann-Whitney U test is used to compare differences between two independent groups. This test can be used to conclude whether two populations differ. It can for instance test if there are differences in medians between groups (Lund Research Ltd, 2013b). In contrast to the t-test, it compares the median scores of two samples instead of the mean score. The test is non-parametric and can therefore be used on samples that are not normally distributed. The test assumes that the samples come from populations with equal variances. When comparing two sample medians the two independent variables (i.e experienced and inexperienced participants) has to have a similar shape. It can test the hypothesis:

H_0: The two populations are equal
H_A: The two populations are not equal

The null hypothesis is rejected if (LaMorte, 2017):

$$U <= criticalvalue \tag{4.5}$$

The critical value is found in the table of Critical Values for U and depends on the sample sizes, n1 and n2, and the significant level α. U is the statistical value calculated.

4.2.2.5 Kruskal-Wallis test

The Kruskal-Wallis test is a nonparametric alternative to one-way $ANOVA$ (see subsection 4.2.2.2) (Walpole et al., 2012). This test should be used if the assumption of normal distribution failed. As mentioned in this sections introduction, a nonparametric method does not assume normality. This test is an generalization of the rank-sum test when there are more than 2 samples.

Kruskal-Wallis is used to test equality of means in one-way $ANOVA$, so the hypothesis for the Kruskal-Wallis test is:

H_0: $\mu_1 = \mu_2 = ... = \mu_k$
H_A: Minimum two of the μ_k's are different

Here μ_k is the rank mean for the group k. As in Wilcoxon Rank-Sum test (subsection 4.2.2.3), the number of observations in the smallest sample is assigned to n_1, the second smallest to n_2 and the largest sample is assigned to n_k.

The null hypothesis is accepted if:

$$H - value < h_\alpha \tag{4.6}$$

4.3 Survey results

This section will first introduce the gathered data and divide the data into samples. Key-values for each sample is listed i tables. Then the section will test each sample for the normal distribution assumption using D'Agostino and Pearson test and equal variance assumption using Levene's test. When the assumptions in each sample is tested, they are used to answer hypothesis leading to an answer on the research hypothesis listed in the introduction. After each subsection a table containing a summary of the test results is presented. This will hopefully make it easier to get an overview of which analyses is completed in the subsection.

4.3.1 Gathered data

The gathered data is analyzed on the two dependent variables: 1) total time used to complete each task and 2) number of correctly chosen elements per task. Both variables sum the participants time spent and correct elements on question one and question two together. The gathered data is also divided by the two independent variable pairs: 1) experienced- and inexperienced participants and 2) the three survey tasks. Combining the dependent- and independent variables create the foundation of the 22 different samples in this thesis. The samples are listed in the tables in this section. Sample mean $\overline{x}$, sample median, standard deviation of $\overline{x}$, standard error of $\overline{x}$, minimum and maximum in each sample is included. The sample ID is also listed in the tables and is referred to in the analysis of the data, to easier distinguish which sample is used in which analysis. In all the samples, results from the training task and from participants that were disturbed during the task are removed. We examine four different divisions of the gathered data in the following section.

4.3.1.1 All, experienced and inexperienced participants

Table 4.1 and 4.2 are samples containing task results from all, experienced and inexperienced participants. The result is grouped by the two dependent variables, total time and the number of correctly chosen elements.

Total time per task (seconds) Sample ID	All	Experienced 1	Inexperienced 2
Number of observations	429	229	200
Sample mean $\overline{x}$	170.32	177.65	161.94
Sample median	155.00	158.00	154.00
Standard deviation of $\overline{x}$	82.19	88.24	73.99
Standard error of $\overline{x}$	3.98	5.83	5.23
Minimum in sample	38.00	52.00	38.00
Maximum in sample	657.00	657.00	529.00

Table 4.1: Total time spent on each task

Correct elements per task Sample ID	All	Experienced 3	Inexperienced 4
Number of observations	429	229	200
Sample mean $\overline{x}$	9.82	9.81	9.83
Sample median	10.00	10.00	10.00
Standard deviation of $\overline{x}$	1.52	1.53	1.51
Standard error of $\overline{x}$	0.07	0.10	0.11
Minimum in sample	4.00	5.00	4.00
Maximum in sample	12.00	12.00	12.00

Table 4.2: Number of correctly chosen elements per task

4.3.1.2 All participants, divided by task

In table 4.3 and 4.4 the task results is divided by the three different tasks, grouped by the dependent variables. Task A is the task that served the participants with one and one elements. Task B is the task that served the participants with three and three elements, and task C gave all six elements at the same time.

Total time per task (seconds) Sample ID	Task A 5	Task B 6	Task C 7
Number of observations	146	142	141
Sample mean $\overline{x}$	166.38	172.25	172.48
Sample median	150.00	155.50	157.00
Standard deviation of $\overline{x}$	84.57	84.21	77.95
Standard error of $\overline{x}$	7.00	7.07	6.56
Minimum in sample	47	50	38
Maximum in sample	657	492	529

Table 4.3: Total time divided by task

Correct elements per task Sample ID	Task A 8	Task B 9	Task C 10
Number of observations	146	142	141
Sample mean $\overline{x}$	10.19	9.71	9.55
Sample median	11.00	10.00	10.00
Standard deviation of $\overline{x}$	1.43	1.53	1.52
Standard error of $\overline{x}$	0.12	0.13	0.13
Minimum in sample	5.00	5.00	4.00
Maximum in sample	12.00	12.00	12.00

Table 4.4: Number of correctly chosen elements divided by task

4.3.1.3 Experienced participants, divided by task

In the tables in this section, only task results from experienced participants are included, and the result is also divided by the three tasks, grouped by the dependent variables.

Total time per task Sample ID	Task A 11	Task B 12	Task C 13
Number of observations	77	80	77
Sample mean $\overline{x}$	173.04	176.70	181.06
Sample median	158.00	156.00	165.00
Standard deviation of $\overline{x}$	96.76	86.13	79.70
Standard error of $\overline{x}$	11.03	9.63	9.08
Minimum in sample	57.00	52.00	53.00
Maximum in sample	657.00	492.00	463.00

Table 4.5: Experienced total time per task, divided by task

Correct elements per task Sample ID	Task A 14	Task B 15	Task C 16
Number of observations	77	80	77
Sample mean $\overline{x}$	10.29	9.66	9.48
Sample median	11.00	10.00	10.00
Standard deviation of $\overline{x}$	1.32	1.64	1.47
Standard error of $\overline{x}$	0.15	0.18	0.17
Minimum in sample	7.00	5.00	5.00
Maximum in sample	12.00	12.00	12.00

Table 4.6: Experienced number of correct elements per task, divided by task

4.3.1.4 Inexperienced participants, divided by task

In this section, the task results from only inexperienced participants are included, and the result is also divided into the three survey tasks. Table 4.7 is the total time variable and 4.8 the number of correctly chosen elements.

Total time per task (seconds)	Task A	Task B	Task C
Sample ID	17	18	19
Number of observations	71	64	65
Sample mean $\overline{x}$	158.30	165.69	162.23
Sample median	148.00	154.50	154.00
Standard deviation of $\overline{x}$	67.57	80.93	74.53
Standard error of $\overline{x}$	8.02	10.12	9.24
Minimum in sample	47.00	50.00	38.00
Maximum in sample	487.00	455.00	529.00

Table 4.7: Inexperienced total time per task, divided by task

Correct elements per task	Task A	Task B	Task C
Sample ID	20	21	22
Number of observations	71	64	65
Sample mean $\overline{x}$	10.07	9.78	9.61
Sample median	10.00	10.00	10.00
Standard deviation of $\overline{x}$	1.54	1.38	1.57
Standard error of $\overline{x}$	0.18	0.17	0.19
Minimum in sample	5.00	6.00	4.00
Maximum in sample	12.00	12.00	12.00

Table 4.8: Inexperienced number of correct elements per task, divided by task

4.3.2 Normality tests

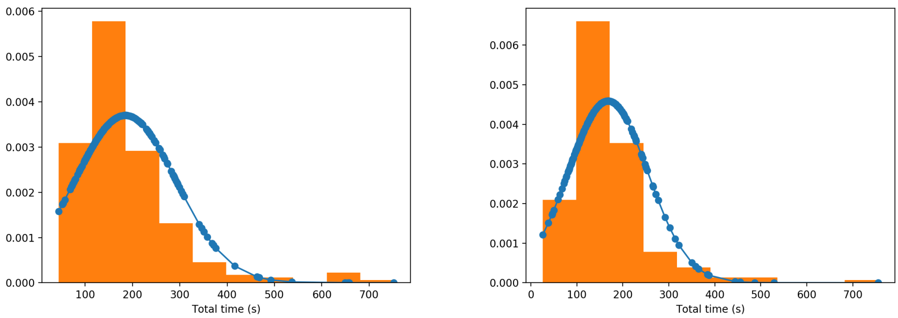
The samples need to be tested to see if they follow a normal distribution. This test is important for determining if a parametric or a non-parametric test should be used when including the different samples in the hypothesis tests. The section about normal testing (4.2.1) concluded that the D'Agostino and Person normality test should be used in this thesis. A visual interpretation of histograms will also be a part of the normality test. D'Agostino-Pearson uses the following hypothesis:

$$H_0\text{: The data follows the normal distribution}$$
$$H_A\text{: The data does not follow the normal distribution}$$

4.3.2.1 Experienced and inexperienced participants, total time variable

This section will test if sample 1 and 2 (table 4.1) follow a normal distribution. Sample 1 and 2 will be used to determine if there are a significant difference in time spent on the tasks between experienced and inexperienced participants.

A visual interpretation of histogram 4.6a and 4.6b gives an indication that sample 1 and 2 does not follow the normal distribution. Both histograms are positively skewed (figure 4.2). Samples involving time measurements are rarely normally distributed. This is because the samples will always be skewed since it is impossible to have negative time and there will always be a limit to how fast a participant can finish the task.



(a) Sample 1 - Experienced

(b) Sample 2 - Inexperienced

Figure 4.6: Histograms with normal distribution fit with samples containing total time to complete each task

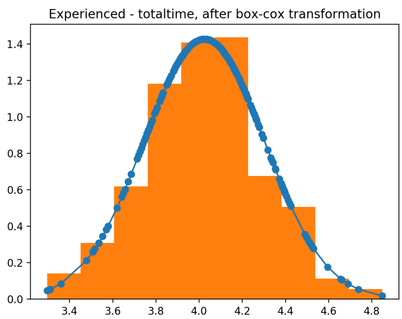
D'Agostino and Pearson tests confirmed the visual interpretation with a significance level of 5%. Both samples obtained p-values lower than the significance level, and the null hypothesis is rejected. Sample 1 and 2 do not follow the normal distribution with a confidence level of 95%.

> *D'Agostino and Pearson normality test*
> Significance level: 5%
>
> Sample 1
> P-value: $3.874 * 10^{-22}$
> The p-value is lower than the significance level (0.05), the null hypothesis is <u>rejected</u> and H_A accepted.
>
> Sample 2
> P-value: $2.574 * 10^{-21}$
> The p-value is lower than the significance level (0.05), the null hypothesis is <u>rejected</u> and H_A accepted.

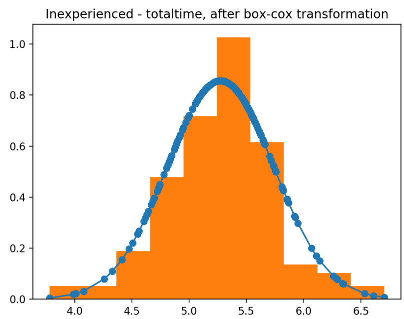
In both sample 1 and 2, the p-value was significantly lower than the significance level of 5%. Data transformations are commonly used tools to improve the normality of a samples distributions, but there are many types of data transformations. Osborne (2010) claim that almost all tests, even non-parametric tests, benefit from improving the normality of the samples, especially when the normality test is significantly denied. Typical traditional transformations are square root, inverse or converting to logarithmic scales (Osborne, 2010).

A Box-Cox power transformation (Box-Cox) is used in this thesis. This transformation can only be used on positive data. The data gathered in this thesis will never be below zero, so this is not a concern. Box-Cox takes the idea of having a range of power transformations (i.e., square root $x^{\frac{1}{2}}$, inverse x^{-1}) available to improve the effectiveness of normalizing and variance equalizing for both positively- and negatively-skewed variables (Osborne, 2010). This transformation will always use the appropriate conversion to be maximally effective in moving each sampled data towards normality. This is the reason why this thesis will use the Box-Cox transformation.

Sample 1 and 2 after a Box-Cox power transformation is shown in histogram 4.7a and 4.7b. A visual inspection gives a good indication that the transformed data follows a normal distribution after the transformation. The skewness looks approximately zero.



(a) Sample 1 - Experienced participants



(b) Sample 2 - Inexperienced participants

Figure 4.7: Histograms with normal distribution fit after Box-Cox power transformation

The transformed data is then applied to the D'Agostino and Pearson test. This test confirms the visual analysis, both sample 1 and sample 2 follows a normal distribution after the Box-Cox with a confidence level of 95%. The calculated p-value is larger than the significance level of 5%.

D'Agostino and Pearson normality test
(After Box-Cox transformation)
Significance level: 5%

Sample 1: Experienced, total time per task
P-value: 0.849
The p-value is higher than the significance level (0.05), the null hypothesis is accepted.

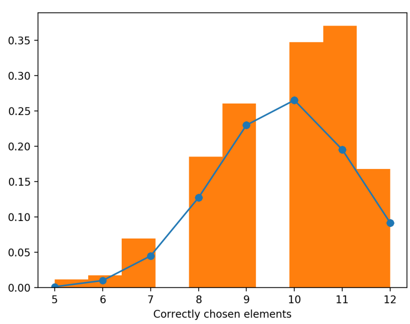
Sample 2: Inexperienced, total time per task
P-value: 0.0623
The p-value is higher than the significance level (0.05), the null hypothesis is accepted.

The assumption that sample 1 and sample 2 follows a normal distribution is now accepted and the transformed data can be used in parametric methods.

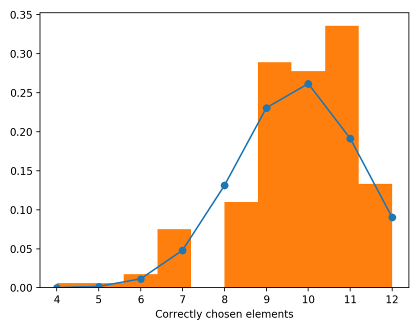
4.3.2.2 Experienced and inexperienced participants, number of correctly chosen elements variable

This section will test if sample 3 and 4 (table 4.2) are drawn from a normal distribution. These samples will be used to test if there are any difference in the number of correctly chosen elements per task between experienced and inexperienced participants.

A visual analysis of the samples histogram 4.8a and 4.8b gives a good indication that sample 3 and 4 are not drawn from a normal distribution. Both are clearly negatively skewed.



(a) Sample 3 - Experienced



(b) Sample 4 - Inexperienced

Figure 4.8: Histograms with normal distribution fit with samples containing the number of correctly chosen elements

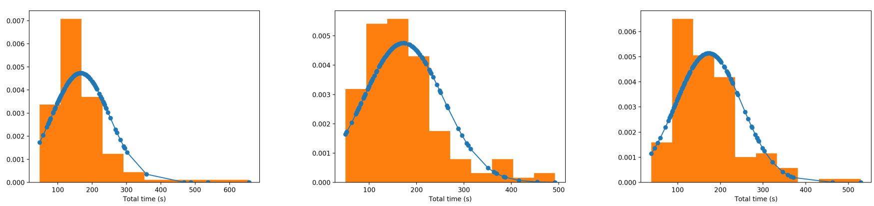
D'Agostino and Pearson normality test confirm our visual analysis. Both samples accept the alternative hypothesis with p-values (0.00443, 0.00013) lower than the significance level (0.05). The null hypothesis is rejected and H_A accepted for sample 3 and 4.

Sample 3 and 4 is Box-Cox power transformed because the null hypothesis was rejected. After the transformation, a new D'Agostino and Pearson normality test was performed. Both samples also failed this test and the alternative hypothesis (H_A) is accepted. Hypothesis including sample 3 and 4 need to be tested with non-parametric methods.

4.3.2.3 All participants divided by task, total time variable

In this section sample 5, 6, and 7 (table 4.3) is normal distribution tested. These samples will be used to test whether there is a significant difference between the three tasks when considering the total time variable.

A visual analysis of the three histograms in figure 4.9a, 4.9b and 4.9c show a positive skewness, just like the histograms in figure 4.6. This gives an indication that the three samples do not follow the normal distribution.



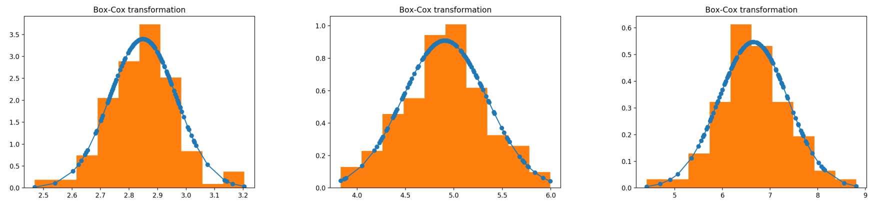
(a) Sample 5 - Task A (b) Sample 6 - Task B (c) Sample 7 - Task C

Figure 4.9: Histogram with normal distribution fit - sample with total time per task

The D'Agostino and Pearson normality test agreed with the visual analysis. Obtained p-values for all three samples ($2.39 * 10^{-24}$, $2.57 * 10^{-9}$, and $1.71 * 10^{-11}$) are smaller than the significance level (0.05), and the null hypothesis is rejected. The samples do not follow the normal distribution with a confidence interval of 95%.

Because the null hypothesis was rejected, the samples are Box-Cox power transformed. Histograms of each sample after the transformation is shown in figure 4.10a, 4.10b and 4.10c. A visual analysis of the histograms gives a good indication that the transformed data is approximately normal distributed. The histograms have a skewness of approximately zero.

The D'Agostino and Pearson normality test confirms the visual interpretation. The p-values (0.164, 0.982, and 0.354) of all three samples are higher than the significance



(a) Sample 5 - Task A (b) Sample 6 - Task B (c) Sample 7 - Task C

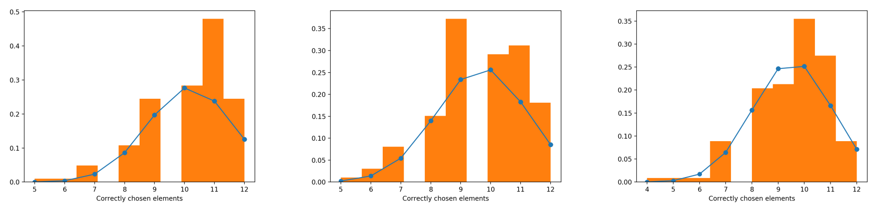
Figure 4.10: Histogram with normal distribution fit after Box-Cox transformation, total time variable

level (0.05), and the null hypothesis is accepted. Sample 5, 6 and 7 follows the normal distribution after the transformation, and parametric methods can be used with these samples.

4.3.2.4 All participants divided by task, correct element variable

This section will examine sample 8, 9 and 10 (table 4.4) for the normal distribution assumption. These samples will be used to test whether there is a significant difference between the three tasks when looking at the number of correctly chosen elements variable.

A visual analysis of the three histograms in figure 4.11a, 4.11b and 4.11c show a negative skewness, just like the histograms in section 4.3.2.2. This give an indication that the three samples are not drawn from a normal distribution.



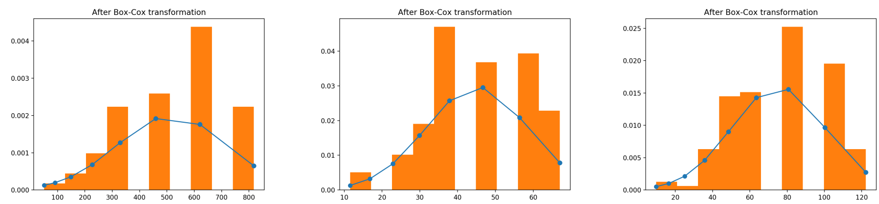
(a) Sample 8 -Task A (b) Sample 9 - Task B (c) Sample 10 - Task C

Figure 4.11: Histogram with normal distribution fit showing samples with number of correct elements per task

D'Agostino and Pearson normality test confirms our visual analysis of the histograms in two of three samples. Sample 9 passes the normality test, even though the p-value (0.099) is close to the significance level (0.05). Sample 8 and sample 10 do not pass the normal assumption test. Both samples obtained a p-value (0.00022 and 0.0047) smaller than the significance level. The null hypothesis is rejected for sample 8 and

10, and the alternative hypothesis is accepted. The null hypothesis is accepted for sample 9.

A Box-Cox power transformation is applied to all three samples. A transformation changes the data, and to correctly compare the results sample 9 has to be transformed even though it follows a normal distribution. The transformed data is shown in histogram 4.12a, 4.12b, and 4.12c. All three are negatively skewed, sample 9 and 10 less than sample 8. The conclusion is not obvious in these histograms.



(a) Sample 8 - Task A (b) Sample 9 - Task B (c) Sample 10 - Task C

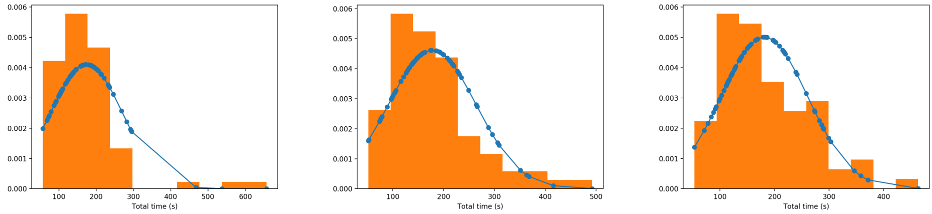
Figure 4.12: Histogram with normal distribution fit after Box-Cox

D'Agostino and Pearson normality test accepts the null hypothesis on sample 9 and 10 and rejects it on sample 8. Sample 9 and 10 has p-values (0.0752 and 0.2104) higher than the significance level, while computed p-value for sample 8 (0.0027) is significantly lower. When using these three samples in hypothesis tests, a non-parametric method should be used. This is because sample 8 do not follow the normal distribution.

4.3.2.5 Experienced participants divided by task, total time variable

In this section, sample 11, 12, and 13, shown in table 4.5, is normal distribution tested. The three samples will be used to test whether there is a significant difference between the total time results for the three tasks when considering only experienced participants.

A visual interpretation of the histograms in figure4.13 show that all three samples are positively skewed (figure 4.2). Skew gives a fairly strong evidence that the samples are not normally distributed.



(a) Sample 11 - Task A

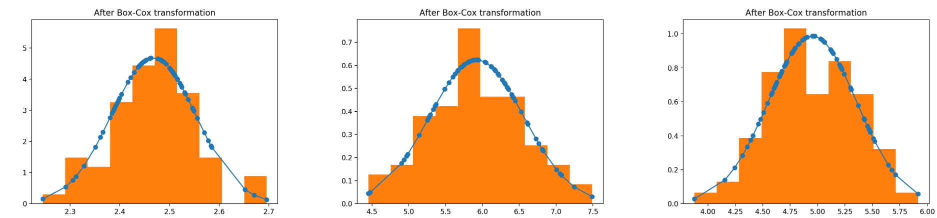
(b) Sample 12 - Task B

(c) Sample 13 - Task C

Figure 4.13: Histograms with normal distribution fit with samples containing total time to complete each task

D'Agostino and Pearson normality test confirms our visual interpretation of the three histograms. All three p-values ($1.229 * 10^{-14}$, $2.678 * 10^{-5}$ and 0.000884) are lower than the significance level (5%). Sample 11, 12 and 13 do not pass the normality assumption with a confidence interval of 95%.

A Box-Cox power transformation is applied to all three samples since the null hypothesis was rejected. Histograms with normal distribution fit containing the transformed data is shown in figure 4.14. Visually, the histograms look like they follow a normal distribution with minimal skewness.



(a) Sample 11 - Task A

(b) Sample 12 - Task B

(c) Sample 13 - Task C

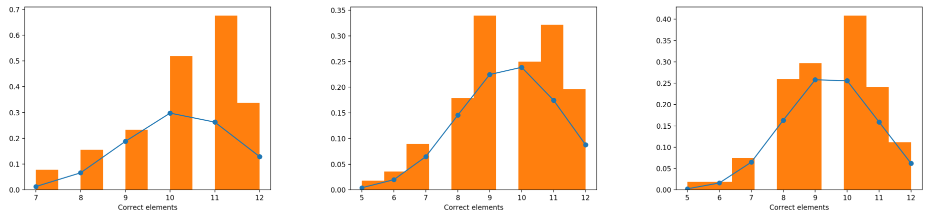
Figure 4.14: Histograms with normal distribution fit containing Box-Cox transformed data

The Box-Cox transformed data is tested with D'Agostino and Pearson normality test. All three samples obtained p-values (0.694, 0.955 and 0.887) larger than the significance level (0.05). Within a confidence interval of 95%, the test concludes that sample 11, 12 and 13 is normally distributed. These samples can be used in parametric methods.

4.3.2.6 Experienced participants divided by task, correct elements variable

This section will test if sample 14, 15, and 16, shown in table 4.6, follows the normal distribution. The three samples will be used to test whether there is a significant difference in the number of correct elements between the three tasks when comparing only experienced participants.

A visual interpretation of the histograms in 4.15 show that all three samples are slightly negatively skewed. Sample 15 (4.15a) and sample 16 (4.15b) has less skew than sample 14 (4.15c).



(a) Sample 14 - Task A (b) Sample 15 - Task B (c) Sample 16 - Task C

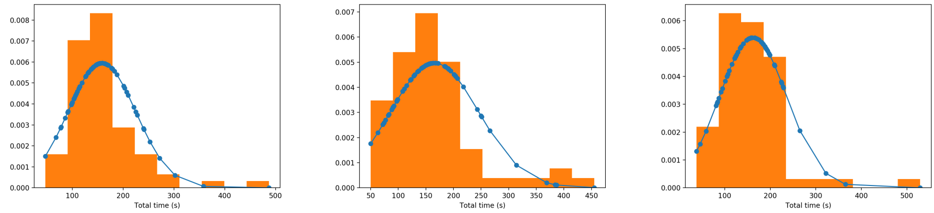
Figure 4.15: Histogram with normal distribution fit showing samples with number of correct elements results for experienced participants

D'Agostino and Pearson normality test confirms our visual interpretation of the three histograms. All three p-values (0.0588, 0.2067 and 0.2975) are higher than the significant level (0.05). Notice that sample 14 has a lower p-value than the two other samples. This sample is not as significant as the two other samples. Sample 14, 15 and 16 pass the normality test with a confidence interval of 95%. The null hypothesis is accepted. These samples can be used in parametric methods.

4.3.2.7 Inexperienced participants divided by task, total time variable

This section will test if sample 17, 18 and 19, table 4.7, follows the normal distribution. These samples will be used to test whether there is a significant difference in total time between the three tasks when looking at only inexperienced participants.

A visual analysis of the histograms 4.16a, 4.16b and 4.16c show a positive skew. The skew is less than the histograms in figure 4.13, but is most likely too large for the samples to be normally distributed.

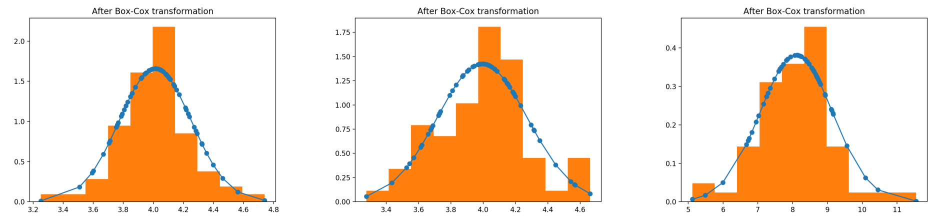


(a) Sample 17 - Task A (b) Sample 18 - Task B (c) Sample 19 - Task C

Figure 4.16: Histogram with normal distribution fit

The D'Agostino and Pearson normality test agrees with the visual analysis. The three obtained p-values ($1.586 * 10^{-11}$, $1.773 * 10^{-6}$ and $2.312 * 10^{-11}$) are all significantly lower than the significance level of 0.05. Sample 17, 18, and 19 do not follow a normal distribution with a confidence interval of 95%. The null hypothesis is rejected.

The samples are Box-Cox power transformed. The histograms after transformation (4.17a, 4.17b, and 4.17c) are visually evaluated to be normal distributed.



(a) Sample 17 - Task A (b) Sample 18 - Task B (c) Sample 19 - Task C

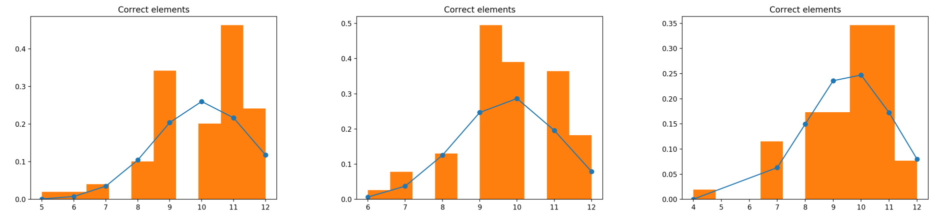
Figure 4.17: Histogram with normal distribution fit after Box-Cox

A new D'Agostino and Pearson normality test on the transformed data confirms that all three samples are drawn from a normal distribution with a significance level of 5%. The obtained p-values (0.139, 0.909 and 0.067) is larger than 0.05, and the null hypothesis is accepted. These samples can be used in parametric methods

4.3.2.8 Inexperienced participants divided by task, correct elements variable

This section will test if sample 20, 21, and 22, table 4.8, follows the normal distribution. These samples contain the number of correctly chosen elements in each of the three tasks from only inexperienced participants. The samples will be used to test if inexperienced participants do better in one of the tasks.

A visual interpretation of histogram 4.18a, 4.18b and 4.18c show a negative skew, similar to the histograms containing results from only experienced participants (4.15).

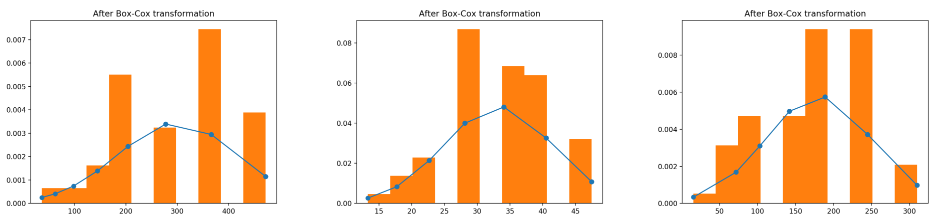


(a) Sample 20 - Task A (b) Sample 21 - Task B (c) Sample 22 - Task C

Figure 4.18: Histogram with normal distribution fit

The D'Agostino and Pearson normality test rejects the null hypothesis on sample 20 and 22 and accepts the null hypothesis on sample 21. Sample 20 and 22 obtained p-values (0.007 and 0.004) lower than 0.05 and sample 21 obtained a p-value (0.523) higher than 0.05. The test concludes that sample 21 follows the normal distribution, and sample 20 and 22 does not with a significant level of 5%.

A Box-Cox power transformation is applied to all three samples. The transformation changes the data, and to correctly compare the data, sample 21 also has to be transformed even though the original data was followed the normal distribution. The transformed samples are shown in histogram 4.19a, 4.19b and 4.19c. All three histograms are less skewed than the original histograms (4.18).



(a) Sample 20 - Task A (b) Sample 21 - Task B (c) Sample 22 - Task C

Figure 4.19: Histogram with normal distribution fit after Box-Cox transformation

The D'Agostino and Pearson normality method is executed on the transformed samples. The three obtained p-values (0.061, 0.714 and 0.311) are higher than 0.05. The null hypothesis is accepted. Sample 20, 21 and 22 follows the normal distribution with a significant level of 5% and can be used in parametric methods.

4.3.2.9 Normality test summary

Table 4.9: Summary of normality tests done in section 4.3.2

	Sample ID	Normally distributed	Normally distributed after Box-Cox power transformation
Total time			
Experienced	1	No	**Yes**
Inexperienced	2	No	**Yes**
Correct elements			
Experienced	3	No	No
Inexperienced	4	No	No
Total time			
Task A	5	No	**Yes**
Task B	6	No	**Yes**
Task C	7	No	**Yes**
Correct elements			
Task A	8	No	No
Task B	9	**Yes**	**Yes**
Task C	10	No	**Yes**
Total time, experienced participants			
Task A	11	No	**Yes**
Task B	12	No	**Yes**
Task C	13	No	**Yes**
Correct elements, experienced participants			
Task A	14	**Yes**	*not tested*
Task B	15	**Yes**	*not tested*
Task C	16	**Yes**	*not tested*
Total time, inexperienced participants			
Task A	17	No	**Yes**
Task B	18	No	**Yes**
Task C	19	No	**Yes**
Correct elements, inexperienced participants			
Task A	20	No	**Yes**
Task B	21	**Yes**	**Yes**
Task C	22	No	**Yes**

4.3.3 Levene's test of Equality of Variance

As mentioned in section 4.2.2.1, 4.2.2.2, and 4.2.2.4, the two sample t-test, one-way *ANOVA* and Mann-Whitey U test assumes that the samples come from populations with equal variances. This assumption will be examined with Levene's test. The hypothesis tested is:

$$H_0\text{: Input samples are from populations with equal variances}$$
$$H_A\text{: Input samples are from populations that do not have equal variances}$$

The null hypothesis is accepted if the obtained p-value is higher than the significance level. Table 4.10 contains the summary of the Levene's test performed on all sample pairs. All sample pairs accepted the null hypothesis except sample 1 and 2, who obtained a p-value lower than the significance level. The test used a significance level of 5% on all the tests.

Table 4.10: Summary of Levene's tests

Participants		Obtained p-value	Samples come from populations with equal variances
All			
	Total time		
	Sample 1 and 2	0.030	**No**
	Correct elements		
	Sample 3 and 4	0.823	Yes
	Total time, divided by task		
	Sample 5, 6, and 7	0.636	Yes
	Correct elements, divided by task		
	Sample 8, 9, and 10	0.805	Yes
Experienced			
	Total time, divided by task		
	Sample 11, 12, and 13	0.972	Yes
	Correct elements, divided by task		
	Sample 14, 15, and 16	0.724	Yes
Inexperienced			
	Total time, divided by task		
	Sample 17, 18, and 19	0.499	Yes
	Correct elements, divided by task		
	Sample 20, 21, and 22	0.626	Yes

4.3.4 Hypothesis testing

This section will test the hypothesis listed in figure 4.4 and 4.5 to be able to answer the two research hypotheses written in the introduction. The order of the tests will be the same as the numbering of the samples. Which statistic method, from the theory section (4.2.2), that is used to answer the hypothesis tests is determined by the results of the normality test (4.3.2) and equal variance test (4.3.3).

4.3.4.1 Differences in total time between experienced and inexperienced participants

This section will test if there are any difference in total time spent on the tasks between experienced and inexperienced participants. This test is covered by sample 1 and sample 2 in table 4.1. Sample 1 is experienced and sample 2 inexperienced participants. Both samples was normally distributed after a Box-Cox transformation (4.3.2.1). A two-sample t-test will be used to answer this hypothesis since the normality assumption is valid. The hypothesis tested in this section is number one in figure 4.4:

H_0: Equal task time between experienced and inexperienced participants
H_A: Unequal task time between experienced and inexperienced participants

If $\overline{x}_1$ equals the mean time for experienced, and $\overline{x}_2$ the mean time for inexperienced participants, the hypothesis can be written as:

$$H_0\colon \overline{x}_1 = \overline{x}_2$$
$$H_A\colon \overline{x}_1 \neq \overline{x}_2$$

Since we cannot assume equal variances in the two samples (Table 4.10), this test will use the Welch's t-test for unequal variances [(Walpole et al., 2012), p. 345]. Equation 4.1 is still valid. The obtained values from the test are shown in the below box. The obtained T-statistic is smaller than the critical value. The t-test then conclude that there is a significant difference between the means of the two population samples with a confidence interval of 95%.

Two sample, two-way t-test
Sample 1 and 2

Degree of freedom (v): 447
Significance level (α): 0.05
Critical value: 1.960

$T - statistic$: -60.442
Using equation 4.1, the absolute value of the $T - statistic$ is larger than the critical value ($|60.442| > 1.960$) and the null hypothesis is rejected and H_A accepted.

Test if experienced or inexperienced participants finish the task fastest

Because there was a statistical significant difference between time spent on each task between the participants, this section will test which group finished the task fastest. The second hypothesis tested in this section is number two in figure 4.4:

H_0: Experienced do not finish the tasks faster
H_A: Experience participants finish the tasks faster

With sample 1 being experienced participants and sample 2 inexperienced participants we get the hypothesis:

$$H_0\colon \overline{x}_1 = \overline{x}_2$$
$$H_A\colon \overline{x}_1 < \overline{x}_2$$

This test gives the same T-statistics as the previous test, but the critical value is changed since this test used in the second hypothesis is a two sample, one-way t-test. The Welch's test is used since we cannot assume equal variances in the two samples. Obtained $T-statistic$ is still smaller than the critical value (-64.654 < 1.645). Our test is to check if the mean value of sample 1 is significantly larger than the mean value of sample 2. We use the comparison test written in equation 4.2, section 4.2.2.1. Our $T-statistic$ is not larger than the critical value, and we need to accept the null hypothesis. There is no evidence that experienced participants use less time on the tasks than the inexperienced.

Two sample, one-way t-test
Sample 1 and 2
Significance level: 5%

$T-statistic$: -60.442
Degree of freedom (v): 447
Significance level (α): 0.05
Critical value: 1.645

T-statistic is smaller than the critical value ($-60.442 < 1.645$) and the null hypothesis is accepted.

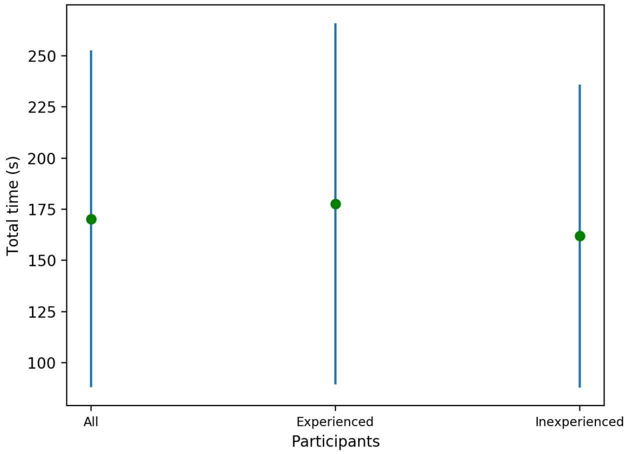


Figure 4.20: Sample 1 and 2 - mean (green dot) and standard deviation (blue line)

Since we know that there is a statistically significant difference between the two sample means, we conclude that the inexperienced participants finished the task faster than the experienced participants. The time difference can also be seen in plot 4.20. Inexperienced participants finished the tasks in average 16 seconds faster than experienced participants.

4.3.4.2 Difference between experienced and inexperienced participants in total correct elements

This section will test if there is a difference between experienced- and inexperienced participants when looking at the number of correctly chosen elements. Sample 3 and 4, table 4.2, is the correct samples to use in this test. Neither samples followed the normal distribution (4.3.2.2), and we need to use a non-parametric method. Both samples have ties (identical observations), and, as mentioned in section 4.2.2.3, the Mann-Whitey U test is then preferred. From histogram 4.8a and 4.8b we see that the samples are identical in some cases. Mann-Whitey U test should, therefore, be used to compare the population medians. The hypothesis to be tested is:

$$H_0\colon\ median_3 = median_4$$
$$H_A\colon\ median_3 \neq median_4$$

Using equation 4.5 in section 4.2.2.4 and the obtained $U-statistic$, we conclude that there is not enough evidence to reject the null hypothesis with a confidence interval of 95%. The $U - statistic$ is larger than the critical value, and the null hypothesis is accepted.

Two sample t-test
Sample 3 and 4
Significance level: 5%

$U - statistic$: 17012
Significance level (α): 0.05
Sample size, n1: 229
Sample size, n2: 200
Critical-value: 127

$U - statistic$ is larger than the critical value ($17012 > 127$) and the null hypothesis is accepted.

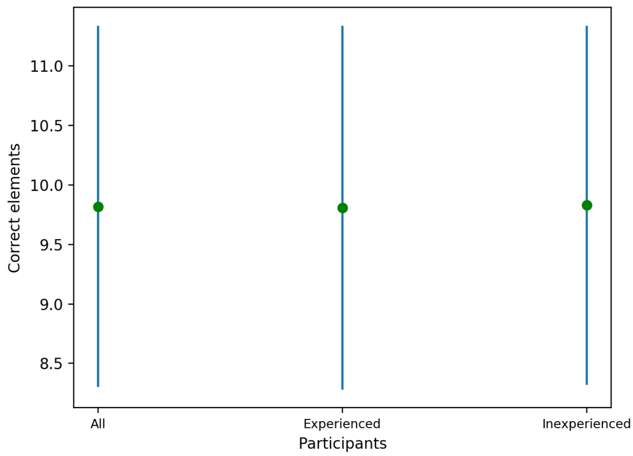


Figure 4.21: Sample 3 and 4 - mean (green dot) and standard deviation (blue line)

Results from this section show that there is not enough evidence to conclude that there is any difference between experienced and inexperienced participants when looking at the number of correctly chosen elements per task. We conclude that experienced and inexperienced participants did equally well on the task. This can also be seen visually in figure 4.21. Mean values shown in the figure is approximately equal between the participants.

4.3.4.3 Test if total time differs between the three tasks

This section will test if time spent varies between each of the three tasks which are hypothesis number one in figure 4.5. Sample 5, 6, and 7, table 4.3, is used in this test. The one-way *ANOVA* method will be applied to answer the hypothesis. All three samples come from populations with equal variances (4.10), the samples are also normally distributed after a Box-Cox transformation (4.3.2.3).

$$H_0\text{: } \overline{x}_5 = \overline{x}_6 = \overline{x}_7$$
$$H_A\text{: Total time differs between at least two of the tasks}$$

Using equation 4.4 in section 4.2.2.2 and results obtained from the calculations, the one-way *ANOVA* test rejects the null hypothesis. The obtained $f - value$ is lower than the critical value. With a confidence interval of 95% we claim that there is a difference between the mean value of the three tasks.

One-way ANOVA
Sample 5, 6 and 7

Significance level (α): 0.05
$v_1 = 2$, $v_2 = 426$
Critical-value: 3.00

$f - value$: 2123.308
$f - value$ is significantly higher than the critical value (2123.308 > 3.00) and the null hypothesis is rejected, H_A is accepted

When rejecting the null hypothesis, *Tukey's method* is used to make comparisons between task A, B, and C. This test did not find any statistically significant difference between the three tasks. Visual evaluation of figure 4.22 show that task A was completed slightly faster than the two other tasks.

Tukey's test
Sample 5, 6 and 7
Significance level: 5%

Task A and Task B do not differ significantly
Task A and Task C do not differ significantly
Task B and Task C do not differ significantly

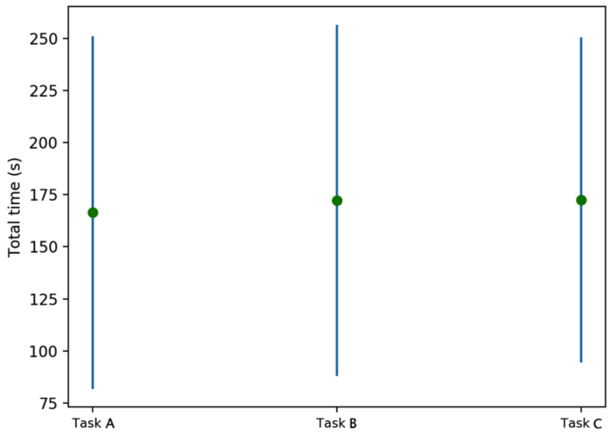


Figure 4.22: Sample 5, 6 and 7 - mean (green dot) and standard deviation (blue line)

Based on the results from this section we conclude that there is a statistically significant difference in total time between at least two of the tasks, but the difference is not enough for *Tukey's test* to notice a difference.

4.3.4.4 Test if the number of correct elements differs between the three tasks

This section will test if there is a difference in the number of correctly chosen elements between the three tasks which are hypothesis two in figure 4.5. The test will use sample 8, 9 and 10 from table 4.4. Since sample 8 are not normally distributed (4.3.2.4) a non-parametric test should be applied. The Kruskal-Wallis test is the non-parametric equivalent to one-way $ANOVA$ (4.2.2.5). The method tests equality of medians when the samples do not follow the normal distribution. The hypothesis tested is:

$$H_0\text{: } median_8 = median_9 = median_{10}$$
$$H_A\text{: Number of correctly chosen elements differ between at least two of the tasks}$$

Using equation 4.6 in section 4.2.2.5, the Kruskal-Wallis test rejects the null hypothesis. The obtained $H - value$ is smaller than the critical value. The p-value is approximately zero, and this gives a good indication that the result is significant. With a confidence interval of 95%, we claim that there is a difference between the median value of the three tasks.

Kruskal-Wallis test
Sample 8, 9 and 10

Significance level (α): 0.05
$v = 2$
Critical-value: 5.991

$P - value$: $3.967 * 10^{-72}$
$H - value$: 328.816
$H - value$ is significantly higher than the critical value (328.816 > 5.991) and the null hypothesis is rejected, H_A is accepted

Like in the one-way *ANOVA*, a *post hoc* test should be used to make paired comparisons to determine which groups differ. The *post hoc* test applied is Tukey's test. Results from Tukey's test resulted in a significant difference in the number of correctly chosen elements between task A and task B, and task A and task C. This can also visually be seen in figure 4.22. The participants had in average 0.8 more correct elements in task A compared to the two other tasks. Task A also has a smaller standard deviation than the other tasks.

Tukey's test
Sample 8, 9 and 10
Significance level: 5%

Task A and Task B differs significantly
Task A and Task C differs significantly
Task B and Task C does not differ significantly

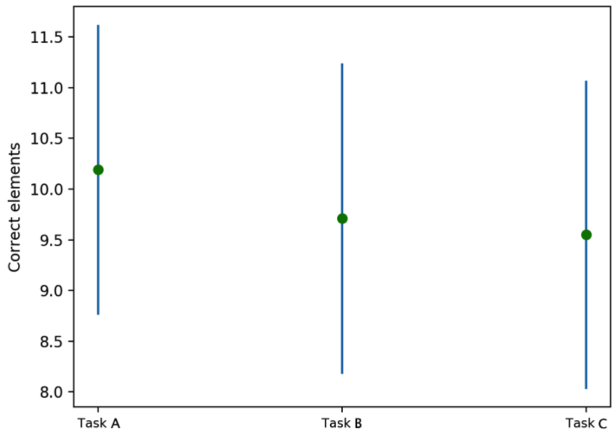


Figure 4.23: Sample 8, 9 and 10 - mean (green dot) and standard deviation (blue line)

The results in this section find statistically significant evidence that the participants had a higher number of correct elements in task A than in the two other tasks.

4.3.4.5 Test differenced in results from experienced participants

This section will answer two hypothesis about experienced participant's results divided by task. The first hypothesis will test if time spent on each of the three tasks differ and the second hypothesis will test if the number of correct elements in each of the three tasks differs when the samples only include results from experienced participants. Sample 11, 12 and 13, from table 4.5, will be used to answer the first hypothesis. All three samples are normally distributed after a Box-Cox power transformation (4.9) and also come from populations with equal variances (4.10). Sample 14, 15 and 16, from table 4.6, will be used on the second hypothesis. These three samples are also normally distributed (4.9) and come from populations with equal variances (4.10). The one-way *ANOVA* method will be used to test both hypotheses.

The first hypothesis is:

$$H_0\colon \overline{x}_{11} = \overline{x}_{12} = \overline{x}_{13}$$
$$H_A\colon \text{Total time differs between at least two of the tasks}$$

Using equation 4.4 from section 4.2.2.2, the one-way *ANOVA* test rejects the null hypothesis. The obtained $f - value$ (1216.919) is higher than the critical value (3.00). The calculated p-value is approximately zero, which gives a good indication that the result is significant. With a confidence interval of 95% the author claim that there is a time difference between the three tasks.

When the null hypothesis is rejected, a *post hoc* test is used to compare each task with each other. Tukey's *post hos* test did not find any significant difference between the three tasks with a significance level of 5%. Figure 4.24 show an approximately similar mean value in all three tasks. Task B has a lower mean and less standard deviation, but it is not statistically significantly different.

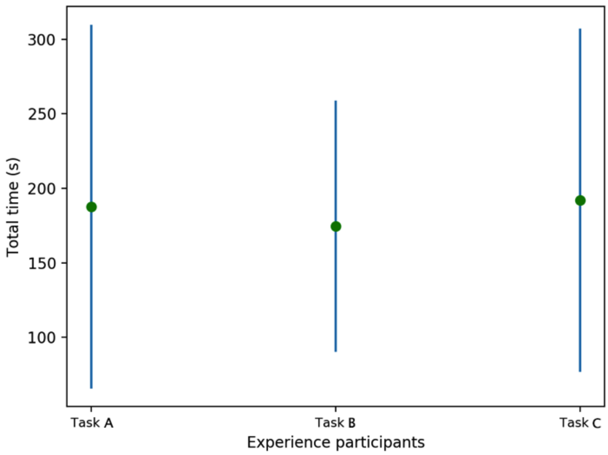


Figure 4.24: Sample 11, 12 and 13 - mean (green dot) and standard deviation (blue line)

The second hypothesis is:

$$H_0\colon\ \overline{x}_1 1 = \overline{x}_1 2 = \overline{x}_1 3$$

H_A: Number of correct elements in each task differs between at least two of the tasks

Using equation 4.4, the one-way $ANOVA$ test rejects the null hypothesis. The obtained $f - value$ (8.210) is higher than the critical value (3.00). The p-value is approximately zero, which gives a good indication that the result is significant. With a confidence interval of 95%, we claim that there is a difference between the mean value of at least two of the tasks.

Since the null hypothesis was rejected, a *post hoc* test should be used to make paired comparisons to determine which groups differ. Tukey's *post-hoc* test resulted in a significant difference in the number of correctly chosen elements between task A and task B, and task A and task C. Figure 4.25 show that task A has a higher mean value than the two other tasks. Task A also has a smaller standard deviation.

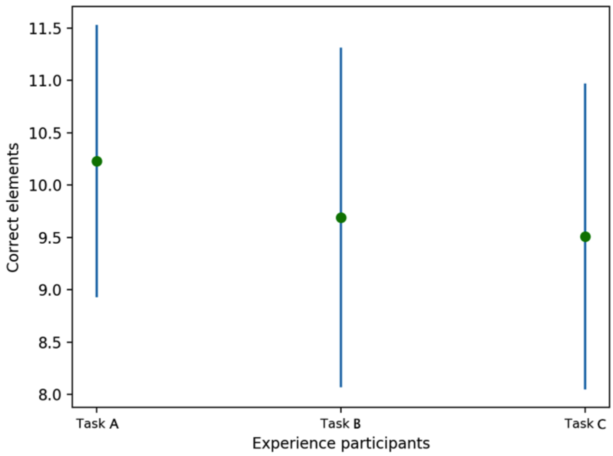


Figure 4.25: Sample 14, 15 and 16 - mean (green dot) and standard deviation (blue line)

With the results found in this section, we conclude that there is a statistically significant difference in the number of correctly chosen elements between the three tasks for experienced participants. They got the best result on task A. Time spent on each task also differs between at least two of the tasks, but the difference is not significant enough so that Tukey's test can determine a difference. Figure 4.24 show that task B has the lowest mean time value of the three tasks.

4.3.4.6 Test differenced in results from inexperienced participants

This section will answer the same hypothesis as the previous section, but using results from inexperienced participants. The first hypothesis will test if time spent on each task differ and the second hypothesis will test if the number of correct elements in each task differs. Sample 17, 18 and 19, from table 4.7, will be applied in the first hypothesis test. These samples are normally distributed (4.3.2.7) and come from populations with equal variances (4.10). Sample 20, 21 and 22, from table 4.8, will be used on the second hypothesis. All three samples are normally distributed (4.3.2.8) and also come from populations with equal variances (4.10). The one-way *ANOVA* will be used to test both hypotheses.

The first hypothesis is:

H_0: $\overline{x}_{17} = \overline{x}_{18} = \overline{x}_{19}$

H_A: Total time differ between at least two of the tasks

The one-way *ANOVA* test rejects the null hypothesis and accepts the alternative hypothesis (H_A) with a significance level of 5%. The obtained f-value from the test is higher than the critical value (905.34 > 3.00). Since the alternative hypothesis was accepted, Tukey's *post hoc* test is used to make compared comparisons between task A, task B and task C. The test does not find a significant difference when comparing each of the three tasks with a significance level of 5%. Figure 4.26 show that inexperienced participants spent more time on task B than the other tasks, but the difference is not significant according to Tukey's test.

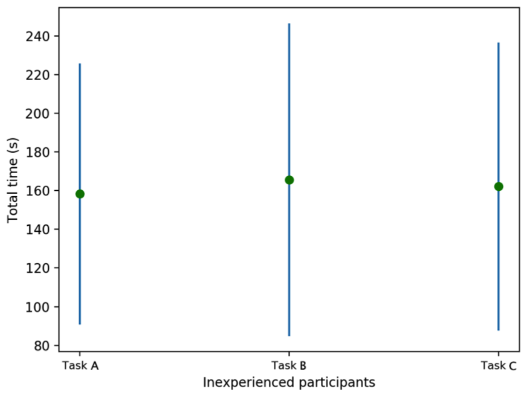


Figure 4.26: Mean (green dot) and standard deviation (blue line) for sample 17, 18 and 19

The seconds hypothesis is:

$$H_0\text{: } \overline{x}_{20} = \overline{x}_{21} = \overline{x}_{22}$$

H_A: Number of correct elements differ between at least two of the tasks

The one-way *ANOVA* test rejects the null hypothesis (H_0) and accepts the alternative hypothesis (H_A) with a significance level of 5%. The obtained f-value from the test is higher than the critical value (189.05 > 3.00). Since the null hypothesis was rejected, Tukey's *post hoc* test will be used to make comparisons between the three tasks. This test cannot find a significant difference when comparing the tasks with a significance level of 5%. Looking at figure 4.27, task A has a higher mean value than the two other tasks, task B also has a higher mean than task C. Even though there are differences in the number of correct elements between the tasks, it is not significant according to Tukey's test.

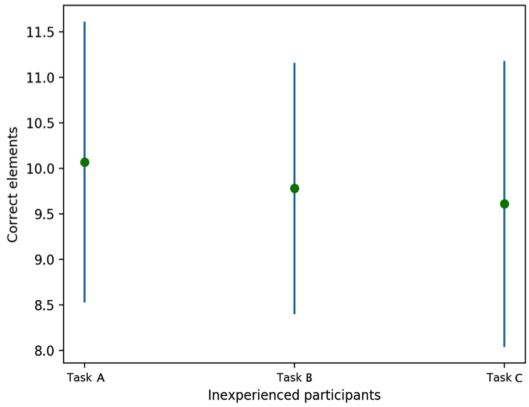


Figure 4.27: Mean (green dot) and standard deviation (blue line) for sample 20, 21 and 22

With the results found in this section, we conclude that there is a statistically significant difference in both total time spent on each task and the number of correctly chosen elements between at least two of the tasks. The differences are not significant enough so that Tukey's test can determine which task differs. Figure 4.26 and 4.27 show that task A has the lowest mean time value and the highest mean correct value.

4.3.4.7 Hypothesis test summary

Table 4.11: Summary of hypothesis tests done in section 4.3.4

Hypothesis (Dependent variable, Independent variable)	*Participants* Sample ID	Hypothesis is accepted
Total time, Experienced and Inexperienced	*All*	
There is a difference between experienced and inexperienced participants	1 and 2	**Yes**
Experienced participants finish the tasks faster than inexperienced	1 and 2	No
Inexperienced participants finish the tasks faster than experienced	1 and 2	**Yes**
Correct elements, Experienced and Inexperienced	*All*	
There is a difference between experienced and inexperienced participants	3 and 4	No
Total time, Task A, Task B and Task C	*All*	
Total time is different between at least two of the tasks	5, 6 and 7	**Yes**
Task A significantly differs from Task B	5 and 6	No
Task A significantly differs from Task C	5 and 7	No

Task B significantly differs from Task C	6 and 7	No
Correct elements, Task A, Task B, Task C	*All*	
The number of correctly chosen elements is different between at least two of the tasks	8, 9 and 10	**Yes**
Task A significantly differs from Task B	8, and 9	**Yes**
Task A significantly differs from Task C	8 and 10	**Yes**
Task B significantly differs from Task C	9 and 10	No
Total time, Task A, Task B, Task C	*Experienced*	
Total time is different between at least two of the tasks	11, 12 and 13	**Yes**
Task A significantly differs from Task B	11 and 12	No
Task A significantly differs from Task C	11 and 13	No
Task B significantly differs from Task C	12 and 13	No
Correct elements, Task A, Task B, Task C	*Experienced*	
Number of correct elements differs between at least two of the tasks	14, 15 and 16	**Yes**
Task A significantly differs from Task B	14 and 15	**Yes**
Task A significantly differs from Task C	14 and 16	**Yes**
Task B significantly differs from Task C	15 and 16	No
Total time, Task A, Task B, Task C	*Inexperienced*	
Total time is different between at least two of the tasks	17, 18 and 19	**Yes**
Task A significantly differs from Task B	17 and 18	No
Task A significantly differs from Task C	17 and 19	No
Task B significantly differs from Task C	18 and 19	No
Correct elements, Task A, Task B, Task C	*Inexperienced*	
Number of correct elements differs between at least two of the tasks	20, 21 and 22	**Yes**
Task A significantly differs from Task B	20 and 21	No
Task A significantly differs from Task C	20 and 22	No
Task B significantly differs from Task C	21 and 22	No

5 | Discussion

Micro-tasking is often used in projects that wish to exploit human computation and a huge crowd, through for instance crowdsourcing. However, there is little research, if any, on using the micro-tasking method with interactive maps and geospatial data. In OpenStreetMap, the method has become a dominant method, both in mapping jobs and imports jobs (Erichsen, 2016). The methods popularity in OSM confirms the potential of expanding micro-tasking to the geo-community. This thesis test if micro-tasking can or should expand to the mapping community, also outside OSM. The research hypothesis tested is 1) Inexperienced workers cannot solve micro-tasks containing geospatial data as good as experienced workers and 2) The fewer elements in a micro-task, the better the worker solves the task.

We evaluate the first research hypothesis as rejected. There was a statistically significant evidence that inexperienced participants finished the tasks faster than experienced participants, this is also shown in figure 4.20. The mean time difference was 16 seconds, a statistically significant, but a relatively small difference. One would expect that experienced participants is familiar with map interaction and visually analyzing geometries in base maps and also interpreting meta information. During the pilot test, the author noticed that the experienced participants used the map aids given in question one more frequently than inexperienced. The map aids were zoom, panning and a layer control to show/hide the building footprints. A possible explanation for the time difference is that the experienced participants spent more time using the map aids provided since they are more familiar with them and knew how to use them.

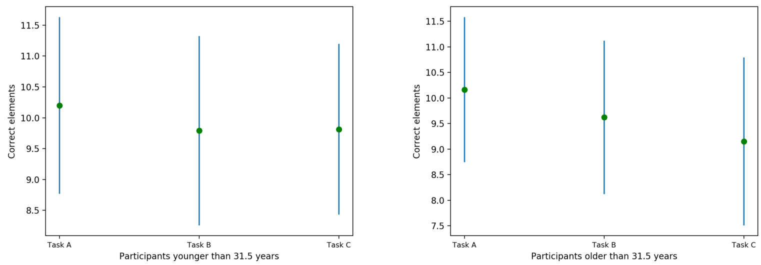
When analyzing the number of correctly chosen elements between experienced and inexperienced, the difference is not statistically significant. Figure 4.21 show a very similar mean value in the number of correct elements. Inexperienced participants had a mean of 9.83 correct elements, while experienced had a mean of 9.81 correct elements. One would expect that inexperienced participants had fewer correct elements since they spent less time on the tasks. Salk et al. (2016) results also had minor differenced between the professional and non-professional participants. They concluded that professional background had a limited first-order relationship with task accuracy. It can be argued that the design of the question interfaces, together with the introduction video and training task, was so easy to use that the professional background had no effect on the quality of the task results. See et al. (2013) concluded that with proper targeted training material the differences between experts and non-experts could decrease. One can also suspect the experienced not to follow the instructions video as carefully as inexperienced.

The splitting of participants into experienced and inexperienced, was based on the question "Do you have experience of working with geospatial data?". Other studies ask the participants for background information through a registration procedure. See et al. (2013) and Salk et al. (2016) considered people with a background in remote sensing/spatial science as experts, and people who were new to the discipline or had a self-declared limited background as non-experts. In this study, the participants are self-declared experts / non-experts. It is not possible to validate this information. In the pilot-test we knew the background of the participants, they answered yes and no on the experience question as the author anticipated.

In the second research hypothesis, it is not as easy to come to a conclusion. There are minor differences between the three tasks. The statistical analysis concluded that there was no statistical difference between the tasks when considering the time variable. Time spent completing each task was approximately the same. Figure 4.24 show a slightly faster task completion on task A, but not a statistically significant difference according to *Tukey's test*.

The time variable does not reflect how much time the participant spent in front of the screen from task start to task end. It reflects how much time passed when solving the two questions. It can be argued that the participants spend more time on task A in total. Time spent switching to the next element in each question is not added to the time variable. In total the participant probably spent more time in front of the screen doing task A compared to task C. In task C all six elements are present, so the participants did not have to wait for the web application to switch to the next task element. In the Los Angeles building import, they used an approach wich combines task A and task C. The buildings were imported one by one, but in their solution, all buildings covering a selected area was visible on the map, but the map window highlighted one and one building. This approach eliminates the time spent fetching and switching to the next building footprint on the map.

Looking at the quality of the task results, task A is statistical significant better than the two other tasks according to *Tukey's test*, shown in figure 4.25. Using the figure and table 4.4, participants had in average one more correct element in task A compared to the two other tasks. Participants did worst on task C, but the difference is small. Figure 5.1 show mean task quality for the three different tasks, divided by the participant's average age (31.5 years). The youngest participants did better on all three tasks. Task A has the highest task quality in both groups. Looking at the participants above mean age, 5.1b, task C has poorer quality. Older participants struggle more on micro-tasks with six elements than three and one elements. Inexperienced participants had an average better quality on task C than the older, but this statement is not statistically tested.



(a) Younger than 31.5 years (b) Older than 31.5 years

Figure 5.1: Mean (green dot) and standard deviation (blue line) - correct elements per task, participants divided by mean age

The results from this thesis can be used as guidance on how to break down the large geospatial task. If quality is the key factor, the large task must be divided into one and one element. The downside of this approach is the time spent on not task specific operations. If the worker needs to click and wait for the next element to load, the workers will spend unnecessary time. If quality mechanisms already are implemented it is possible to break the task into larger chunks (containing, i.e., three or six elements). The number of mistakes will probably increase with the number of elements in the chunk, but if the quality mechanisms are well developed, the errors should be discovered. Benefits of dividing each task into chunks, containing more than one element, is that the worker will most likely spend less time in total doing the micro-tasks. We would not recommend chunks containing more than six elements unless one uses well-developed quality mechanisms.

6 | Conclusion

We conclude that it is not necessary to only use experienced individuals in geospatial micro-tasks, especially if the interface supporting the tasks are designed well and has a thorough training session. This is proven in the analysis from 4.3.4.1.

7 | Proposed sections

7.1 Future work

Create a survey to test how accurate both experienced and inexperienced participants digitize buildings from aerial images. Can use FKB as the correct polygon and compare it with the drawn polygon from participants.

Do a study with reward. Compare reward and not reward geo tasks. Do they solve the tasks better with reward? "A reward can be provided for merely participating in the task. The reward can also be provided as a prize for submitting the best solution or one of the best solutions. Thus, the reward can provide an incentive for members of the community to complete the task as well as to ensure the quality of the submissions."

The future in micro-tasking "belongs to hybrid methodologies that combine human computation with advanced computing" (Meier, 2013b).

When aiming towards wider adoption of crowdsourcing one have to be aware of the challenges of using it. It is important to remember that all tasks do not fit into the micro-tasking crowd worker model. Very complex tasks that can't be partitioned are not suitable for solving through micro-tasks.

Advanced computing techniques such as Artificial Intelligence and Machine Learning is needed to build approaches that combine the power of people with the speed and scalability of automated algorithms (Meier, 2013b).

7.2 Usage potential

Systems are exploiting the people's physical presence in an environment more, they are more location dependent. This can be particularly important when seeking to improve geospatial data quality [(Meier, 2013b), p. 323]. "For instance, UrbanMatch (Celino et al. 2012a) is a mobile location based game that uses player's familiarity with a city to link photos with points of interest in the city. Players are shown points of interest and known images from a trusted source (e.g. OpenStreetMap) and asked if photos from an untrusted source (e.g. Flickr) might also relate to the point of interest ".

(Meier, 2013b): "As the previous sections show there is a lot of potential for AR systems to use HC to provide content, and to support processing in other ways. However there has been little research to date combining AR and HC systems. In this section we review the first research efforts in this area. "

(Meier, 2013b) "Lastly, there is huge untapped potential in leveraging the "cognitive surplus" available in massively multiplayer online games to process humanitarian microtasks during disasters. The online game “League of Legends,” for example, has 32 million players every month and three million on any given day. Over 1 billion hours are spent playing League of Legends every month. Riot Games, the company behind League of Legends is even paying salaries to select League of Legend players. Now imagine if users of the game were given the option of completing microtasks in order to acquire additional virtual currency, which can buy better weapons, armor, etc. Imagine further if users were required to complete a microtask in order to pass to the next level of the game. Hundreds of millions of humanitarian microtasks could be embedded in massively multiplayer online games and instantaneously completed. Maybe the day will come when kids whose parents tell them to get off their computer game and do their homework will turn around and say: “Not now, Dad! I’m microtasking crisis information to help save lives in Haiti!” "

Machines are bad at tackling things they have never seen before. They need to learn from large amounts of passed data. Humans don’t need this. Humans can solve tasks we have never seen before. Tackling new/novel situations are humans much better than machines. Business strategies, marketing holes, this are tasks only humans can do.

Data Categorization, organize your data, no matter what the data is. Micro-tasking platforms can turn all the big data into rich data that is organized, streamlined, and useful. Micro-tasking let’s you organize your original data which again can be used to train machine learning models. According to CrowdFlower is human-curated training sets the best traning datasets to use.

Appendices

A Tets

Fbox

Some text esfljsf lskj lksdjflsk slk

Some text kduhaszkdh aszkdjhs zkjdfh skdj skd

dwkjdkwjdh wkjdhw kjdh wkjhd qwkjhd kwd qw .

text

dwkjdkwjdh wkjdhw kjdh wkjhd qwkjhd kwd qw .

Bibliography

Affairs, A. S. f. P. (2013). System Usability Scale (SUS).

Bangor, A., Kortum, P., and Miller, J. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies*, 4(3):114–123.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., and Camos, V. (2007). Time and Cognitive Load in Working Memory.

Barth, A. (2014a). OpenStreetMap | lxbarth sin dagbok | Importing 1 million New York City buildings and addresses. Date Accessed: 2017-05-23 URL: openstreetmap.org/user/lxbarth/diary/23588.

Barth, A. (2014b). Over 1 million New York City buildings and addresses imported to OpenStreetMap | Mapbox. Date Accessed: 2017-05-23 URL: www.mapbox.com/blog/nyc-buildings-openstreetmap/.

Ben, S. and Plaisant, C. (2009). *Designing the User Interface*. Pearson, fifth edition.

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. (2015). Soylent: A Word Processor with a Crowd Inside. *COMMUNICATIONS OF THE ACM*, 58(8):85–94.

Biewald, L. (2015). Why human-in-the-loop computing is the future of machine learning | Computerworld. Date Accessed: 2017-05-14 URL: www.computerworld.com/article/3004013/robotics/why-human-in-the-loop-computing-is-the-future-of-machine-learning.html.

Brooke, J. (1996). *SUS-A quick and dirty usability scale. "Usability Evaluation In Industry"*. Taylor & Francis.

Chilton, S., Building, F., and Burroughs, T. (2009). CROWDSOURCING IS RADICALLY CHANGING THE GEODATA LANDSCAPE : CASE STUDY OF OPENSTREETMAP.

community, O. (2017a). Import/Guidelines – OpenStreetMap Wiki. Date Accessed: 2017-05-23 URL: wiki.openstreetmap.org/w/index.php?title=Import/Guidelines&oldid=1447838.

community, O. (2017b). Los Angeles, California/Buildings Import – OpenStreetMap Wiki. Date Accessed: 2017-05-23 URL: wiki.openstreetmap.org/w/index.php?title=Los_Angeles,_California/Buildings_Import&oldid=

contributors, O. (2017). lacounty:bld_id | Keys | OpenStreetMap Taginfo. Date Accessed: 2017-05-31 URL: taginfo.openstreetmap.org/keys/lacounty:bld_id#overview.

Deng, X., Joshi, K. D., and Galliers, R. D. (2016). THE DUALITY OF EMPOWERMENT AND MARGINALIZATION IN MICROTASK CROWDSOURCING: GIVING VOICE TO THE LESS POWERFUL THROUGH VALUE SENSITIVE DESIGN 1. 40(2):279–300.

Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., and Cudré-Mauroux, P. (2015). The Dynamics of Micro-Task Crowdsourcing. *Proceedings of the 24th International Conference on World Wide Web - WWW '15*, pages 238–247.

Difallah, D. E., Demartini, G., and Cudré-Mauroux, P. (2016). Scheduling Human Intelligence Tasks in Multi-Tenant Crowd-Powered Systems. *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, pages 855–865.

Erichsen, A. S. S. (2016). Evaluation of the Micro-Tasking Method for Importing High-Detail Building Models to OpenStreetMap. *github.com/annesofie/Prosjektoppgave/blob/master/openstreetmap.pdf*.

EYeka (2015). The State of crowdsourcing 2015 - How the world's biggest brands and companies are opening up to consumer creativity. Technical report.

Fan, H., Zipf, A., Fu, Q., and Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, (28:4):700–719.

Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., and Xin, R. (2011). CrowdDB: answering queries with crowdsourcing. *Proceedings of the 2011 international conference on Management of data - SIGMOD '11*, pages 61–72.

Frost, J. (2015). Choosing Between a Nonparametric Test and a Parametric Test. Date accessed: 2017-04-23 URL: blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test.

Gadiraju, U., Demartini, G., Kawase, R., and Dietze, S. (2015). Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing. *IEEE Intelligent Systems*, 30(4):81–85.

Ghasemi, A. and Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2):486–9.

Gutzwiller, L. (2017). Why You Can't Remove Humans from AI Training Loops | Mighty AI. Date Accessed: 2017-05-23 URL: mty.ai/blog/why-you-cant-remove-humans-from-ai-training-loops/.

Ha, A. (2016). CrowdFlower raises $10M to combine artificial intelligence with crowdsourced labor | TechCrunch. Date Accessed: 2017-05-08 URL: techcrunch.com/2016/06/07/crowdflower-series-d/.

Holzinger, A. (2013). Human–Computer Interaction and Knowledge Discovery (HCI-KDD): What Is the Benefit of Bringing Those Two Fields to Work Together? *Springer Lecture Notes in Computer Science LNCS 8127*, pages 319–328.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

Holzinger, A., Plass, M., Holzinger, K., Crişan, G. C., Pintea, C.-M., and Palade, V. (2016). Towards interactive Machine Learning (iML): Applying Ant Colony Algorithms to Solve the Traveling Salesman Problem with the Human-in-the-Loop Approach. In *Availability, Reliability, and Security in Information Systems*, pages 81–95. Springer, Cham.

Howe, J. (2006). The Rise of Crowdsourcing. *Wired Magazine*, 14(06):1–5.

Huotari, K. and Hamari, J. (2017). A definition for gamification: anchoring gamification in the service marketing literature. *Electronic Markets*, 27(1):21–31.

Ipeirotis, P. G. and G., P. (2010). Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21.

Israel, G. D. (1992). Determining Sample Size 1. *Florida Cooperative Extension Service, University of Florida*, PEOD-6.

Kiefer, P., Giannopoulos, I., Duchowski, A., and Raubal, M. (2016). Measuring Cognitive Load for Map Tasks Through Pupil Diameter. pages 323–337. Springer, Cham.

Kitchin, R. and Tate, N. J. (2000). *Conducting Research into Human Geography*. Prentice Hall.

Kostas (2016). Using Crowdsourcing and Machine Learning to locate swimming pools in Australia · Tomnod. Date Accessed: 2017-05-04 URL: blog.tomnod.com/crowd-and-machine-combo.

LaMorte, W. W. (2017). Mann Whitney U Test (Wilcoxon Rank Sum Test). Date Accessed: 2017-05-12 URL: sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric4.html.

Leppink, J., Paas, F., Van Gog, T., Van Der Vleuten, C. P. M., and Van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, 30:32–42.

Lund Research Ltd (2013a). Hypothesis Testing - Significance levels and rejecting or accepting the null hypothesis. Date Accessed: 2017-04-20 URL: statistics.laerd.com/statistical-guides/hypothesis-testing-3.php.

Lund Research Ltd (2013b). Mann-Whitney U Test in SPSS Statistics | Setup, Procedure & Interpretation | Laerd Statistics. Date Accessed: 2017-05-12 URL: statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php.

Lund Research Ltd (2013c). One-way ANOVA - Its preference to multiple t-tests and the assumptions needed to run this test | Laerd Statistics. Date Accessed: 2017-04-25 URL: statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide-2.php.

Mandler, G. (2013). The Limit of Mental Structures, The Journal of General Psychology. pages 243–250.

MedCalc Software bvba (2017). Skewness and Kurtosis. Date Accessed: 2017-04-23 URL: medcalc.org/manual/skewnesskurtosis.php.

Meier, P. (2013a). Digital Humanitarian Response: Moving from Crowdsourcing to Microtasking | iRevolutions. Date Accessed: 2017-05-04 URL: irevolutions.org/2013/01/20/digital-humanitarian-micro-tasking/.

Meier, P. (2013b). Handbook of Human Computation. In *Handbook of Human Computation*. Springer New York, New York, NY.

Meier, P. (2014). Typhoon | iRevolutions. Date Accessed: 2017-05-07 URL: irevolutions.org/tag/typhoon/.

Michelucci, P. and Dickinson, J. L. (2016). The power of crowds. *Science*, 351(6268):32–33.

Morschheuser, B., Hamari, J., and Koivisto, J. (2016). Gamification in Crowdsourcing: A Review. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 4375–4384. IEEE.

Motulsky, H. (2013). Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking, 3rd edition. In *Intuitive Biostatistics*, chapter 24.

Nakhuda, D. (2016). Announcing General Availability of Mighty AI's Intelligent Crowdsourcing Platform | Mighty AI. Date Accessed: 2017-05-23 URL: mty.ai/blog/general-availability-intelligent-crowdsourcing-platform/.

Nikki (2016). Finding Swimming Pools in Australia using Deep Learning. Date Accessed: 2017-05-04 URL: blog.tomnod.com/finding-pools-with-deep-learning.

Oppenheimer, D. (2017). Machine Learning with Humans in the Loop - Algorithmia. Date Accessed: 2017-05-14 URL: blog.algorithmia.com/machine-learning-with-human-in-the-loop/.

Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, 15(12).

Palen, L., Soden, R., Anderson, T. J., and Barrenechea, M. (2015). Success & Scale in a Data-Producing Organization. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 4113–4122.

Pearson, A., Sözcükler, A., düzeltmeli Kolmogorov-Smirnov, L., Pearson ve Jarqua-Bera testleri Derya ÖZTUNA Atilla Halil ELHAN Ersöz TÜCCAR, A., and Öztuna, D. (2006). Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions. *Turk J Med Sci*, 36(3):171–176.

Quinn, A. J. and Bederson, B. B. (2011). Human computation. *Proceedings of the*

2011 annual conference on Human factors in computing systems - CHI '11, pages 1403–1412.

Ross, C. and Jamilly, M. (2016). Better Together how Computers can be Designed to Augment Human Ability. *ITNOW*, 58(2):44–47.

Salk, C., Sturn, T., See, L., and Fritz, S. (2016). Local Knowledge and Professional Background Have a Minimal Impact on Volunteer Citizen Science Performance in a Land-Cover Classification Task. *Remote Sensing*, 8(10):774.

Sambale, M. (2016). OpenStreetMap | manings sin dagbok | Building tools for LABuildings Import. Date Accessed: 2017-05-31 URL: openstreetmap.org/user/manings/diary/38969.

Sarasua, C., Simperl, E., and Noy, N. F. (2012). Crowdsourcing Ontology Alignment with Microtasks. pages 525–541.

Schade, A. (2015). Pilot Testing: Getting It Right (Before) the First Time. Date Accessed: 2017-04-19 URL: nngroup.com/articles/pilot-testing/.

Schulze, T., Krug, S., and Schader, M. (2012). Workers' Task Choice in Crowdsourcing and Human Computation Markets. *ICIS 2012 Proceedings*.

See, L., Comber, A., Salk, C., Fritz, S., van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F., and Obersteiner, M. (2013). Comparing the quality of crowdsourced data contributed by expert and non-experts. *PloS one*, 8(7).

Smith, S. (2013). Determining Sample Size: How to Ensure You Get the Correct Sample Size | Qualtrics. Date Accessed: 2017-04-19 URL: qualtrics.com/blog/determining-sample-size/.

Stanford University (2017). Machine Learning | Coursera. Date Accessed: 2017-05-14 URL: www.coursera.org/learn/machine-learning.

The Pennsylvania State University (2017). 7.5 - Power and Sample Size Determination for Testing a Population Mean | STAT 500.

The Scipy community (2017). scipy.stats.anderson — SciPy v0.19.0 Reference Guide. Date Accessed: 2017-04-21 URL: docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson.html.

von Ahn, L. (2008). Human Computation. In *2008 IEEE 24th International Conference on Data Engineering*, pages 1–2. IEEE.

Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. (2012). *Probability & Statistics*. Pearson Education, ninth edition.

Wang, X., Goh, D. H.-L., Lim, E.-P., Wei Liang Vu, A., and Chua, A. Y. K. (2017). Examining the Effectiveness of Gamification in Human Computation. *International Journal of Human–Computer Interaction*, pages 1–9.

Webster, E. (2016). Using Crowdsourcing for Complex Data Problems: It Can & Should Be Done! | Mighty AI. Date Accessed: 2017-

05-23 URL: mty.ai/blog/using-crowdsourcing-for-complex-data-problems-it-can-should-be-done/.

Yap, B. W. and Sim, C. H. (2011). Journal of Statistical Computation and Simulation Comparisons of various types of normality tests Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12):2141–2155.