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

HEI

Task Description

The micro-tasking method is becoming more and more popular. Companies like Amazon develop micro-tasking web applications where people can earn money by doing micro-tasks for others. The method is used for tasks that involve both use of technology and a large number of people. By using the micro-tasking methodology, this thesis aims to study how people solves micro-tasks within geospatial data imports, which is a very complex and large process.

This study will have an emphasis on the data validation and conflict handling part of the import. These parts are complicated to do fully automatic through scripts. By varying the number of objects to solve at a time, adding rewards on some tasks, among other factors, the study will hopefully find a significant approach to prefer when using the micro-tasking method within geospatial data. 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? Do rewards have an impact on how the tasks are solved?

This thesis will also explore the micro-tasking methods usage potential within geospatial data. Can other organizations doing a process that needs humans to interfere take advantage of this method? An example is OpenStreetMap, who has taken good advantage of the method both in mapping and import projects.

Specific tasks:

- Study related literature
- Do a micro-tasking survey
- Examine how many elements are optimal when creating geospatial micro-tasks

Abstract

This paper propose a method for extracting buildings in satellite photos. The proposed network makes use of a digital surface model and multispectral satellite data. It

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

1 | Introduction

To the authors best knowledge, little research has been done on micro-tasking geospatial data. This thesis aim is to study how well geospatial data tasks is solved through micro-tasking. The quality of the completed tasks when doing micro-tasks it is important. In this study the quality is measured through the number of correctly chosen elements in each task. The resulting data will distinguish between experienced and inexperienced participants. A micro-task should be small enough so that all individuals can complete the task, independent on their background and experience.

Salk et al. (2016) looked at how local knowledge and professional background, impact the volunteer performance in a Land-Cover classification task. The paper concluded that there was no difference in how well the participants did, and their background.

A task can traditionally be divided by time, place, person, object, and skill [(Meier, 2013b), p. 13]. A task can be created by identifying the time it will require, the place where it must be done, the people who need to do the task, the object on which the work is done and finally, the skill needed for the task. Today we have technology that can create and move tasks around based on the four first categories. Technology can allocate tasks based on the deadline and time the task requires. It can also establish communication between any team of people dependent on which people the task requires. One thing that technology can't do is change the skill of individual workers, though it can only connect people with different skills to work on the same tasks [(Meier, 2013b), p. 14]. Crowdsourcing moves beyond this and looks at the skills of individual workers, the problem that needs to be solved and combines the best skills of workers to solve the problem. The division of labor by skill has more economic impact than the other four categories [(Meier, 2013b), p. 15]. 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).

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 time, and the usage so far can be evaluated as successful. This thesis also aims to find out if inexperienced individuals also manage to solve tasks on maps that involves geospatial data. The study also aims to determine if the number of elements in each micro-task has an impact on how well individuals solve the micro-tasks. The quality of the work, the number of correctly solved tasks and time, is measured. The thesis uses a survey hosted through a web-application to gather participant data. The data is then used to answer this thesis hypothesizes. The next chapter will give a thorough introduction to micro-tasking and hopefully make it clearer what this thesis aim is. Chapter 3 will explain the survey and chapter 4 will contain the statistics,

both hypothesis, theory, and results.

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).

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 can require much time and high costs. By exploiting both machines and people through the appropriate platform, the cost can decrease and the quality increase. The author will argue in this chapter that combining machines and people is often a better and faster solution than a fully-automatic or fully-manual approach and implementing such an approach into a micro-tasking platform can be a good solution.

2.1 Why not just use machine learning?

Machine learning give computers the ability to learn without being explicitly programmed. It involves computer intelligence, but the computer do not know the answers up front (Stanford University, 2017). Machine-learning algorithms have enormous problems when the contextual information is lacking. Without a pre-set of rules, a machine has trouble solving the problem. Machines do not have creativity, which is required to solve complex problems (Holzinger et al., 2016). According to the company "Mighty AI", humans cannot be removed from Artificial Intelligence training loops. They believe that humans will continue to play a crucial role in creating training data for the algorithms.

It is suggested that machine learning accuracy should follow the Pareto 80:20 principle. Getting 80 % accuracy can be fairly easy to accomplish, but the last 20 % should to be handled by human input (Biewald, 2015) (Oppenheimer, 2017). Human input can be to label the original training dataset or help correct inaccurate predictions outputted from the algorithm (Oppenheimer, 2017). A machine learning company called "developmentSEED" use a micro-tasking solution for cleaning their machine learning output data. They are using humans to get a more accurate output data faster. Skynet Scrubber, a GUI web application solution was developed to get the human input easier and faster (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). reCaptcha is a good example on how to exploit human cognitive abilities. Humans are through this solution identifying themselves as a human, and at the same time digitizing old books. In older books where the pages are bleached and yellow, machines struggle to understand what the words say. 750 000 000 individuals has helped digitize books through reCaptcha. "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 humans and a human computation systems contains both humans and computers working together to solve difficult problems (Schulze et al., 2012). The author argue 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 often a solution that uses humans cleverly by exploiting the human brain's cognitive abilities can sometimes create much faster and better results than a software. One of the pioneers of crowdsourcing, Luis von Ahn, wanted to find a cheep and effective 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 images though 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 don't have a proper caption associated with them and this makes it difficult to create search engines for images for instance. A fast and cheap method of labeling images is by using humans cleverly, humans can very easily see if the image contains a dog or cat for instance. 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 didn't even have to pay them doing it. Another game that was created by Luis Von Ahn is called "Peekaboom", this is also using a gamification approach. Here the players would locate objects in images. Such information is very useful in computer vision research for instance (von Ahn, 2008). 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.

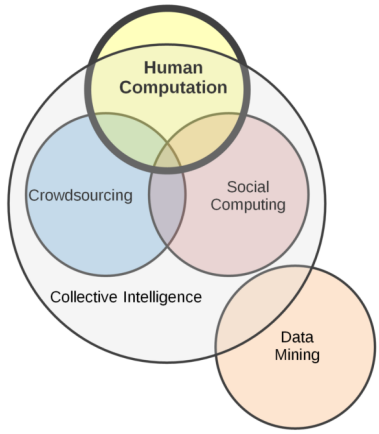


Figure 2.1: Collective intelligence (Quinn and Bederson, 2011)

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 effective one need 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 in this chaper that micro-tasking can be another solution to 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 (section 2.2) replaces computers with humans, crowdsourcing replaces traditional human workers with members of the public (Quinn and Bederson, 2011). EYeka (2015) state that 85 % of the top global brands use crowdsourcing for various purposes. Crowdsourcing has become a widespread approach to dealing with machine-based computations where we leverage the human intelligence (Gadiraju et al., 2015). Crowdsourcing is an increasingly important concept, where the concept is the completion of large projects by combining small distributed contributions from the public (Salk et al., 2016).

When the scope 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.

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 a number of simple tasks, that can be completed in parallel in a relatively short period of time (from seconds to minutes), 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 problems like: (1) completing surveys, (2) translating text between two languages, (3) matching pictures of people, (4) summarizing text (Bernstein et al., 2015), etc.

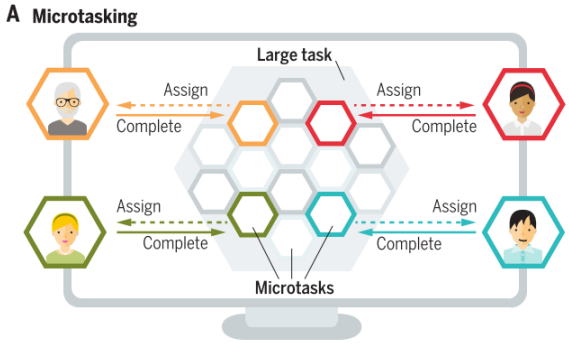


Figure 2.2: Micro-tasking (Michelucci and Dickinson, 2016)

Micro-tasking and human computation are closely related. In the "Handbook of Human Computation", micro-tasking is present in the *Human Computation for Disaster Response* chapter [(Meier, 2013b), p. 95-105], as well as in several other chapters. In the *Human Computation for Disaster Response* chapter they 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. With the help of CrowdFlower micro-tasking platform, 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 relevant tweets were uploaded to the CrowdCrafting micro-tasking platform were volunteers both tagged and geo-tagged each photo and video 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 2.3. This 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, since the algorithm detected them (Meier, 2014).

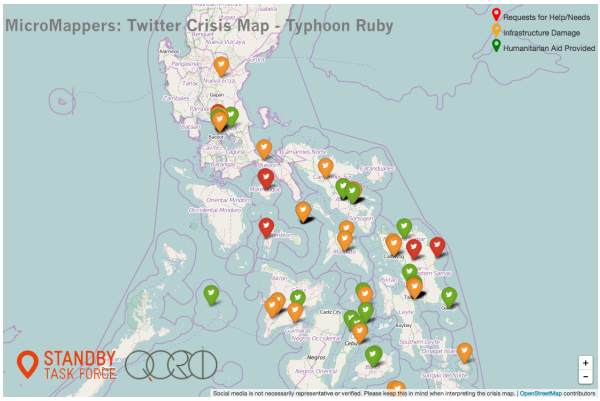


Figure 2.3: Typhoon Ruby Crisis map (Meier, 2014)

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 or sub-problems are then solved by multiple workers independently, often in return for a reward (Sarasua et al., 2012). Thanks to micro-tasking platforms as Amazon's Mechanical Turk (MTurk; www.mturk.com), 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 (www.crowdflower.com), 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.

Many human computation systems use crowdsourcing platforms to recruit workers (Schulze et al., 2012). In machine-learning algorithms, combining human computation abilities and crowdsourcing with fast machine learning algorithms can be proven to be a success so far. An example is the team "Tomnod" [1] 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 created polygons where there was likely to be a swimming pool inside. Then the users participating

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

in finding the swimming pools were only shown polygons the algorithm had created, minimizing the search area, shown in figure 2.4. This approach was used in order to reduce the number of user votes required to classify the pools but yet obtain a sufficient confidence (Kostas, 2016). This is a good example of micro-tasking. Instead of serving the users with satellite photos of huge areas, the photos was divided into polygons created by an algorithm. The "Tomnod" tesm 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 2.4: Polygons created by an algorithm that maybe contain pools (Nikki, 2016)

Most cases of micro-tasking usage exploit the large volume capabilities machines have and the cognitive capabilities of humans (Difallah et al., 2016). Micro-tasking has also been used to process queries. In the Franklin et al. (2011) paper they extended a traditional query engine with a small number of operations that requires human input by generating and submitting requests. They used a micro-tasking platform to get the crowd to answer queries that cannot otherwise be answered. There are especially two cases where human input is needed: a) when the data is unknown or incomplete, b) when there is a need for subjective comparisons. They used auto-generated user interfaces and new query operators that can obtain human input via the interfaces. The Franklin et al. (2011) paper demonstrated that human input can be leveraged to dramatically extend the range of SQL-based query processing. People are good at comparing items, such as how well an image represents a particular concept. Humans are also good at finding relevant information with the help of search engines etc (Franklin et al., 2011). By utilizing these qualities in humans, like they did in (Franklin et al., 2011) paper by developing a micro-tasking based implementation of query operations, a huge cost and time sparing potential can be utilized.

One of the advantages of micro-tasking platforms like "Tomnod" and "CrowdFlower", that is mentioned by Meier (2013b) (p. 99), is the built-in quality control mechanisms that ensure a 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.

2.4.1 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), evidence indicates that 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 compensated with tiny rewards every time they complete a micro-task. The workers are then rewarded often, which is motivating.

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 2.4.2.1), 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).

2.4.2 Micro-tasking platforms

2.4.2.1 Amazon's Mechanical Turk

Amazon's Mechanical Turk (MTurk) is one of the biggest (if not the biggest) micro-tasking platform today. 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 for it. 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).

2.4.2.2 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. The tool sorts the area covered by the satellite images 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 don't map the same areas. This is an very effective approach to coordinate the crowd participating in the mapping. The tasking manager is mainly used by the *Humanitarian OpenStreetMap Team* (HOT). This platform do not have a rewarding system og a gamification approach.

It is solely based on volunteer contributors. The page shows which areas need to be mapped and which areas need the mapping validated by others.

There are also other tools in OpenStreetMap. Tofix etcetc.

2.4.2.3 CrowdFlower

CrowdFlower is a company that wants to help businesses take advantage of crowdsourcing and/or human computation. They act as an intermediary for these businesses (Quinn and Bederson, 2011). CrowdFlower receives tasks from businesses wanting to crowdsource their work or problems. A project done for eBay exploited CrowdFlower's large online contributor pool and completed the tasks given to them from eBay five times faster than a traditional outsourcing team. eBay got a solution that was optimized for both quality and cost. CrowdFlower works 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 judgement when they're not confident on the technology and the human work can make the algorithms smarter (Ha, 2016). Founder of CrowdFlower says that "self-driving cars have gotten pretty good at recognizing many of the objects they encounter on the street, [..] 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).

2.5 Building imports using micro-tasking

In OpenStreetMap, at least to building imports was successfully completed using micro-tasking. The tasking manager platform (2.4.2.2) was used to organize the import.

2.5.1 Challenges

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 operations added to a micro-tasking platform consider the talents and limitations of human workers (Franklin et al., 2011) and this is what this thesis try to examine. What are the limitations of human workers when dealing with maps and geospatial data. It has been shown that crowds can be "programmed" to execute classical algorithms such as Quicksort, but such use of available resources is neither performant nor cost-efficient (Franklin et al., 2011).

New software developed by researchers at Facebook can score 97.25 percent on the same challenge, regardless of variations in lighting or whether the person in the picture is directly facing the camera."

3 | Methodology and experiment

Gadiraju et al. (2015) categorize typically crowdsourced tasks into six top-level classes. Interesting classes within geospatial data is *Verification and validation*, *Interpretation and analysis* and *Content creation*. There are examples of all three task classes in geospatial crowdsourcing. During imports of large datasets into OpenStreetMap, crowdsourcing is used to validate the new data. 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. In a machine learning process, they are starting to use micro-tasks to both validate the created data and also create test data sets to the algorithms. *Interpretation and analysis* tasks rely on the individual to use their interpretation skills during task completion. This is the task class used in this thesis during the survey. Is it safe to assume that individuals, both experienced and inexperienced, can interpret and analyse geospatial data presented to the on both map and in tables correctly? This is the main goal of the survey.

3.1 Survey

This thesis will try to determine questions regarding micro-tasks containing geospatial data. Little research is done on how well inexperienced individuals solve micro-tasks when they involve map interaction and geospatial data. To the authors best knowledge, little, if any, research has been done on micro-tasks involving map interaction and geospatial interpretation and analysis.

The survey will be administered through a Self-Administered Questionnaire, there the questions will be located in a web-application developed and implemented by the author. When doing questionnaire designed survey, a key factor is to keep the questions short, simple, and clearly worded.

Our first sample application asks users to identify which object does not belong in a collection of items. This kind of task requires both image- and semantic-classification capability

3.2 Experiment

The survey is a part of this thesis. The survey is used to answer different hypothesis around geospatial micro-tasks. To be able to answer the hypothesizes three tasks containing the same two questions was developed. The questions will represent two different micro-tasks involving geospatial data, while the three tasks will vary the

number of elements the participant will use to answer the questions with. The participant will always answer the two question's on six elements, but the tasks vary how many element's need to be handled at the same time. The variation of a number of elements in the tasks is to hopefully find out if or how much the number of elements in a micro-task effect how well people solves the task.

When selecting the number of elements in the three different 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 negative effects on task completion, also the cognitive load that is imposed by a learning task is much higher for novices that for more advanced students (Leppink et al., 2014).

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 maps and geospatial data. The last task will only contain one element as a minimum cognitive load task. This can help answer how many elements a human can process when doing micro-tasks containing geospatial data. The goal is to determine a preferred number of elements within a micro-task to use when developing micro-tasks so that they are most efficient and accurate.

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 this survey. Kiefer et al. (2016) argues that high cognitive load may lead to less effective map reading and spatial orientation, as well as decreased spatial learning. Since polygon comparison doesn't demand high cognitive load, the task should at least not be too demanding on the one element task 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 survey can help determine if geospatial micro-tasks are too demanding on inexperienced individuals.

The survey will then contain three tasks, each task contains six elements but the tasks vary how many elements the participant need to handle at the same time. One task will serve the participant with one and one element, the task that demands the smallest cognitive load. The other task will serve the participant with three and three elements at the same time. This number is just under the limit of how much information humans can process. The last task will serve the participant with all the elements at

the same time. This number exceeds the human capacity when processing information according to Leppink et al. (2014).

There are two variable types used in this survey, dependent- and independent variables. The dependent variables are: time spent on each question and each task, the number of correctly chosen elements in both questions and also how difficult the participant though the task was. The independent variables are: number of elements in the task, experienced or inexperienced participant, gender, age and if the participant knows micro-tasking.

3.3 Determining the building shapes

Remote sensing is a tool or technique for extracting information about objects or geographic areas. All remote sensing images are subject to some form of geometric distortions. The distortions depend on how the data are obtained (Toutin, 2004). In Norway, most remote sensing images are analysed manually. This is also the case in OpenStreetMap. When using remotely sensed images to create for instance building footprints, it's important to be aware of the distortions in these images.

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 large diversity in the geometries (Fan et al., 2014), and this is probably not the only city with this kind of diversity. To evaluate the quality of the building footprints in OSM, the Fan et al. (2014) paper used four criterion's, completeness, semantic accuracy, position accuracy and shape accuracy.

In the creation of the elements and conflicts used in the first question in the survey, the quality criterion's shape- and position accuracy were emphasized. The first question asks the participant to select the shape that fits the marked building on the map best. The goal is to create shapes that matches realistic cases that occur for instance in OSM.

Shape accuracy evaluates how well the layer matches the building with reference to an aerial image. Fan et al. (2014) mentions three main reasons to why building footprints are simplified in OSM. First reason is because of the difficulties following building details when looking from a bird's-eye view. Second reason is the limited resolution on the Bing aerial image used during digitalization. The last reason that is mentioned is that the volunteers in OSM don't have the patience to digitalize a complicated footprint exactly as it is. Drawing two layers 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 judgements by only using an aerial image as reference.

Position accuracy evaluates how well the coordinate value of a building relates to the reality on the ground. The correct layer will be drawn on the corresponding ground coordinates, while the conflicting layer will not match the ground. Fan et al. (2014) tested the accuracy of buildings in OSM, and concluded with an 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 don't 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.

3.4 Web application

This thesis used an online web-based survey to conduct the experiment. An online survey avoid 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).

In a self selected sample, which is some the case here, there is potentially a bias in the sample (Ben and Plaisant, 2009).

3.4.1 Technology

React
Django
Postgis
AWS

3.4.2 Architecture

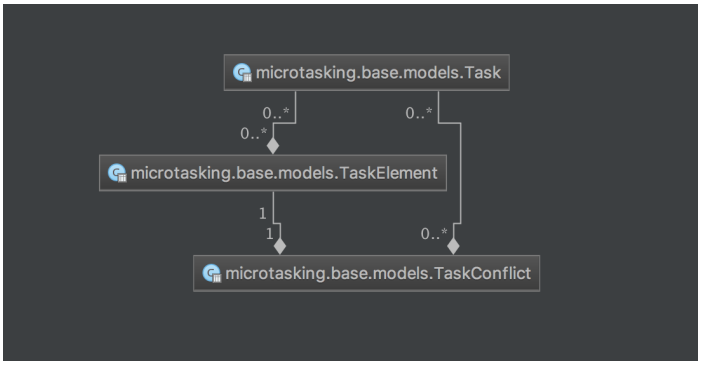


Figure 3.1: UML diagram, Task: Task Element and Task Conflict

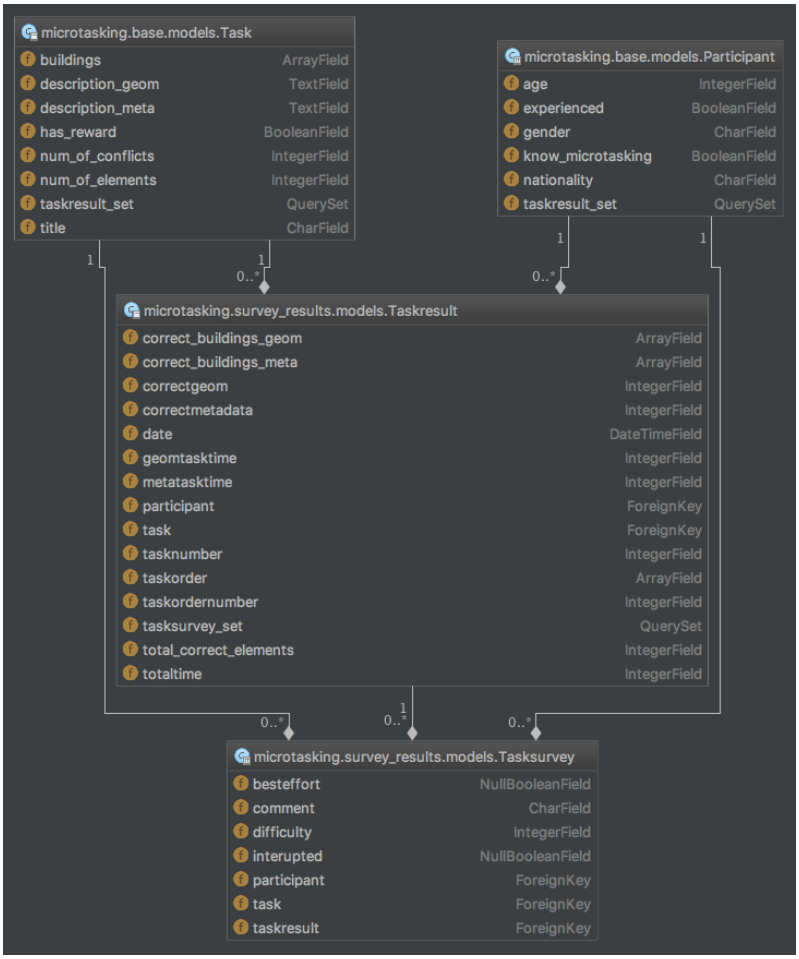


Figure 3.2: UML diagram, Survey result: Task result and Task survey

3.4.3 Graphical Interface

3.5 Pilot test

It is important to pilot test the survey prior to actual use (Ben and Plaisant, 2009). A pilot test provides an opportunity to validate the wording of the tasks, do the participants understand the tasks? It also helps understand the time necessary for completing the survey, which should be communicated to the participants in prior to the survey (Schade, 2015). The pilot-test will be conducted with a small sample of users. Results from the pilot test are in this thesis used to do improvements to the actual survey, to the web application hosting the survey and to find errors or weaknesses in the database models.

After the pilot test, the usability was measured. The standard ISO 9241-11 suggests that measures of usability should cover effectiveness, efficiency and satisfaction (ISO, 1998). Measuring these three classes of metric can vary widely and makes it difficult to make comparisons of usability between different systems. "[...] just because a particular design feature has proved to be very useful in making one system usable does not necessarily mean that it will do so for another system" (Brooke, 1996). Usability in this thesis will be 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 attitude scales (Brooke, 1996). SUS was developed to be quick and simple, 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 don't 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).

3.5.1 Execution of the pilot test

The pilot test was conducted with a total of eight participants, five experienced and three non-experienced participants aged from 22 to 64 years. It started with a brief information about this study and the survey. They were told to talk out load during the survey, no help or guidance was given to the participants. The author observed the participants while they conducted the survey. The author took notes and watched if the participants understood the questions in the survey correctly. After the survey a *System Usability Scale* questionnaire was answered by the participants. At 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 web application and to determine which improvements to be done.

3.5.2 Results from the pilot test

- Did someone knew micro-tasking? Can we see something here?

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 good indication that the web-application is user friendly.

All participants thought that the instruction movie was confusing. It was short, the instructions went too fast and it lacked voice descriptions. The movie needed major improvements, an important discovery. The purpose of the movie is to give the participant an introduction to how to answer the two questions. It gives important instructions, especially for participants that are not used to working with maps on a web page.

Overall feedback on question one was that it was difficult to understand which building was which and also when a building layer was selected or not. The lack of labels on the buildings was done on purpose to get the task as much as possible realistic. The process of selecting the best building layer needed improvements, it had to be clearer that selection was done by clicking on the layer on the map, not by using the layer control as some thought. This problem was added to the movie with voice description, describing in detail how a layer was selected. The design on the question one page was also improved by adding color to the text telling the participant which layer they had selected.

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 movie. This request was fulfilled in the new movie. The task progress bar was also removed, during the eight pilot tests the author didn't notice that any of the participants looked at the task progress bar. The progress bar was thought of as an extra help to inform how many elements was left in the task. Only the survey progress bar on the top right was found necessary.

The pilot test data was used to test some of the hypothesis to find errors or weaknesses in the databasemodel. The data was extracted with the help of Django QuerySets and saved in csv files. Some preliminary results can be seen in section 3.5.3. There were a few errors and weaknesses found during the statistical tests. Changes to the database models was necessary, and the changes done are listed under:

1. Add foreign key from TaskResult model to TaskSurvey
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. Difficulty field in Tasksurvey model was changed from Char field to Integer field

The additional fields will mainly help with creating plots to better interpret the data and to more easily visualize the different results.

3.5.3 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 3.3 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 1 represents the task with one elements, task 2 the task with three elements and task 3 the task with six elements.

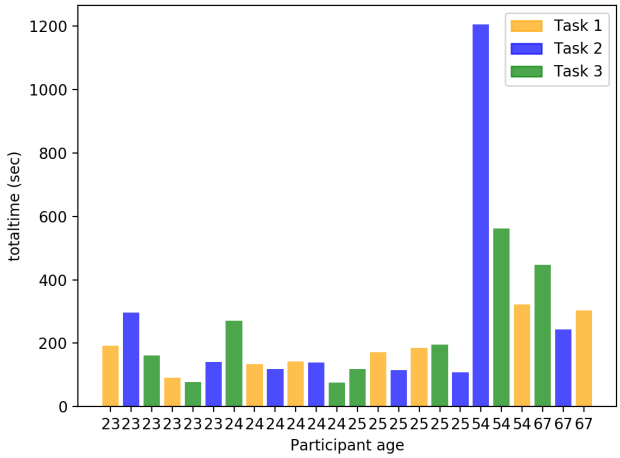
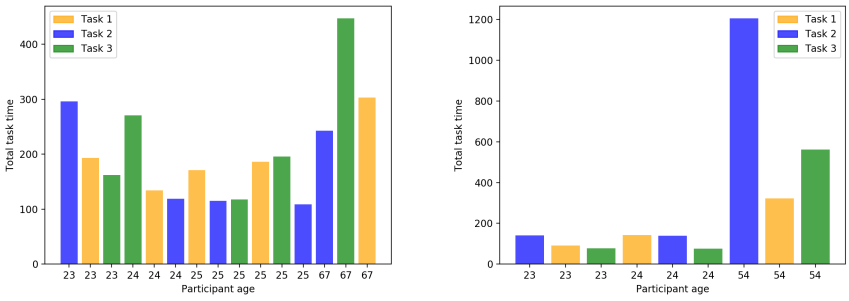


Figure 3.3: Total time - all participants ordered by age

Splitting the data into two groups, experienced and inexperienced, and then mapping totaltime and order by age plots can be seen in figures 3.4a and 3.4b. Experienced participant had a mean time of 3 minutes and 24 seconds (204 sec) and inexperienced participants had a mean time of 5 minutes and 4 seconds (304 sec), almost 2 minutes difference. It is clear to see that the 54 year old participant's total time on task 2 is dramatically higher than the rest. By looking at figure 3.4b, two of the inexperienced participants spent less than 200 seconds on all their tasks.



(a) Experienced: Mean 204 sec

(b) Inexperienced: Mean 304 sec

Figure 3.4: Total time - ordered by age, splitted into experienced and inexperienced

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 layers and meta data rows was 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 in the last task. $\frac{7}{8}$ answered yes on the question. This information was important. If every participant does a better job at the last task the result will not be as useful. Even though the task order varies. Reading the data in figure 3.3, $\frac{6}{8}$ participants spent less time on the last task, even though the task order varied. This matches the number of participants who remembered the buildings and meta information from the previous tasks.

Total number of correctly chosen elements in each task is shown in figure 3.5.

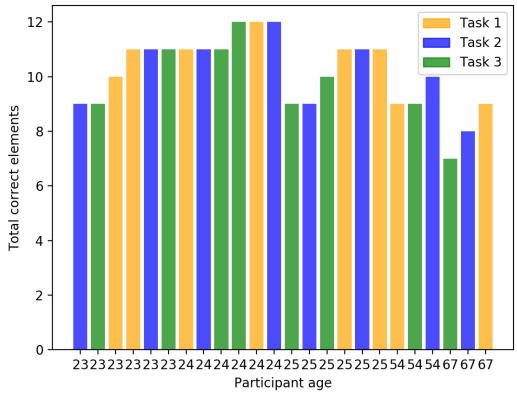
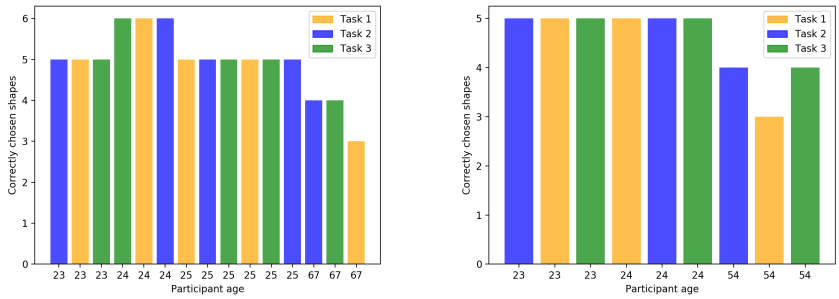


Figure 3.5: Total correct, both questions - ordered by age



(a) Experienced

(b) Inexperienced

Figure 3.6: Number of correctly chosen building shapes - splitted by experience

3.6 Determining the sample size

The sample size is influenced by a number of factors, including the purpose of the study, population size, the risk of selecting a "bad" sample and the allowable sampling error (Israel, 1992). In this survey there are three possible ways of determining the sample size.

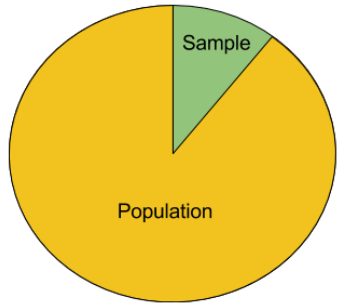


Figure 3.7: Population vs. sample

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 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, a precision of 0.05, confidence level of 95% and a size of population greater than 100'000, the necessary sample size is 400. If the precision is changed to 0.1, the sample size necessary increases to 100 (Israel, 1992). The numbers found in the table reflects the number of obtained responses. The last approach is to use formulas to calculate the sample size. The formulas requires 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 matches this study.

It's important to mention that the quality of the sample is as important as it's size. The more variable the sampled data is, the larger the sample size is required (Israel, 1992). It's 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

4.1 Sample data from Survey

Independence of observations. This is mostly a study design issue and, as such, you will need to determine whether you believe it is possible that your observations are not independent based on your study design (e.g., group work/families/etc). A lack of independence of cases has been stated as the most serious assumption to fail. Often, there is little you can do that offers a good solution to this problem.

Designed the survey so that the observations should be random and independent
- Random order on the tests
- Random color on the layers
- Random which order the layers was drawn on the map
- Random which order the metadata was written in the table

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 visualize 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 survey 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, see figure 4.1, and kurtois, see figure 4.2, 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).

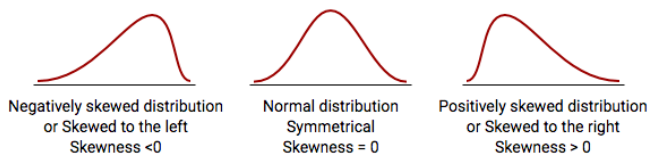


Figure 4.1: Skew (MedCalc Software bvba, 2017)

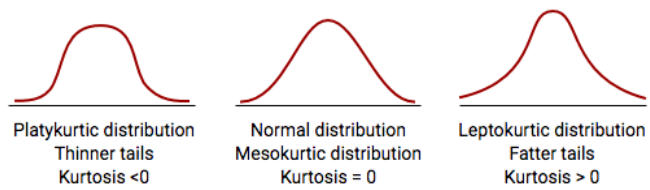


Figure 4.2: Kurtois (MedCalc Software bvba, 2017)

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).

4.2.2 Bionomal disstibution

Used for discrete variables. It used the probability of getting x successes and $n - x$ failures in n trials. Each success comes with a probability p and each failure with probability $q = 1 - p$ [(Walpole et al., 2012), p. 145]. The sample mean $\overline{x}$ and variance of $\overline{x}$ of the bionomial distribution is: $\overline{x} = n * p$ and $\sigma^2 = n * p * q$. The probability p has to be the same on every trial - NOT TRUE HERE.

In the survey, the number of correctly chosen elements is recorded. Here x = the number of correct elements. x is a random variable who has the binomial distribution. The following null and alternative hypothesis can be used:

H_0: All elements are correctly chosen in each task
H_A: Not all element are correctly chosen in each task

$p = \frac{\overline{x}}{n}$, if $p*n >= 5$ and $n*(1-p) >= 5$ then we can use the normal distribution.

4.2.3 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 (Ha) usually represents the question to be answered or the theory to be tested, while the null hypothesis (H_0) nullifies or opposes Ha (Walpole et al., 2012). The sample collected in the study is used to test which statement is most likely (technically it's 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 first step, after identifying the hypothesis, 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).

You should not report the result as "significant difference", but instead report it as "statistically significant difference". This is because your decision as to whether the result is significant or not should not be based solely on your statistical test. Therefore, to indicate to readers that this "significance" is a statistical one, include this is your sentence (Lund Research Ltd, 2013c).

4.2.3.1 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 on two means can be written as:

$$H_0\text{: } \mu_1 \text{ - } \mu_2 = 0 \text{ or } \mu_1 = \mu_2$$
$$H_A\text{:} \mu_1 \text{ - } \mu_2 \neq 0 \text{ or } \mu_1 \text{ - } \mu_2 > 0 \text{ or } \mu_1 > \mu_1$$

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 - $\mu_2 > 0$), and the 4.3 equation is used on hypothesis where the test is to check if the mean is lower than zero (μ_1 - $\mu_2 < 0$). 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.

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

Hypothesis - Two sample t-test

H_0**:** Experienced and inexperienced spent the same amount of time on the tasks
H_A**:** Total task time differs between them

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

H_0: Total number of correct elements between experienced and inexperienced are equal
H_A: There is a difference in number of correct elements between them

H_0: Experienced no not have more total correct elements then inexperienced
H_A: Experienced participants have a higher number of correct elements

Before solving the hypothesis the conditions needs to be testet. More on this later.

4.2.3.2 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 are 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 subsection 4.2.3.1.

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 like this:

$$H_0\text{: } \mu_1 = \mu_2 = ... = \mu_k$$
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:

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

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 detemine which groups means are significantly different [(Walpole et al., 2012), p.526].

In a one-way *ANOVA* test there should be one dependent variable and minimum three independent groups, which is an relevant approach considering the data produced from this thesis survey. There are at least two dependent variables in the survey data, task time and number of correctly chosen elements. The survey result can be divided into three groups, one element task, three elements task and six elements task. Each entry in the sample should only be assigned to one group. Relevant hypothesis from the study that can be used in an one-way *ANOVA* analusis is shown under.

Hypothesis - One-way ANOVA

H_0: Mean task time is not different between the three tasks
H_A: Mean task time is different between at least two of the tasks
Variable = time, group = tasks

H_0: Total number of correct elements between the three tasks are equal
H_A: Total number of correct elements between at least two of the tasks are not equal
Variable = Number of correct elements, group = tasks

The hypothesis written above will be tested in section 4.3.4.3.

4.2.3.3 Wilcoxon Rank-Sum test

The Wilcoxon Rank-Sum test is an appropriate alternative to the two-sample t-test (see subsection 4.2.3.1) when the normality assuptions do not hold, but the samples are still independend and have a continous distribution (Walpole et al., 2012). Since this method is nonparametric (or distribution-free) it do not require the assumption of normality.

The hypothesis for Wilcoxon Rank-Sum Test is:

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

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 preffered (The Scipy community, 2017).

4.2.3.4 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 not normally distributed samples. 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.

4.2.3.5 Kruskal-Wallis test

The Kruskal-Wallis test is a nonparametric alternative to one-way *ANOVA* (see subsection 4.2.3.2) (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 4.2.3.3), 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:

(4.6)

4.3 Survey results

- All participants ordered by age
- All participants ordered by age, excluded by task 4
- All results in one task, ordered by age
- Average time per micro-task
- Is there a difference in task order number 1, 2, 3? time and correct
- Is there a difference in task number 1, 2, 3? time and correct

Can use it to explain the data

4.3.1 Gathered data

The gathered data will be analyse on the two variables: 1) total time used to complete each task and 2) number of correctly chosen elements per task. Total time and number of correct elements adds time and correctly chosen elements on question one and question two together. Sample mean $\overline{x}$, standard deviation of $\overline{x}$, standard error ($\frac{standarddeviation}{\sqrt{samplesize}}$) of $\overline{x}$, minimum in sample and maximum in sample are listed in the tables. In these tables, results from the training task is removed. Only results from the three tasks is used. Maximum possible correct elements per task is twelve. There are six elements in question one and six elements in question two, and the number of correctly chosen elements in each task is added together, maximum twelve correctly chosen elements.

The tables in this section, (4.3.1), are task results from all participants and all three tasks, excluding the training task. Task results with total time longer than 2160 seconds are filtered out. This is to remove 4 outliers that spend more than twice the approximated time (average time on the survey was 1080 seconds in the pilot test). These 4 participants also answered that they were disturbed during the test.

The samples are in the first subsection (4.3.1.1) divided into experienced and inexperienced, but the three tasks are not separated in this sample. In subsection 4.3.1.2 the sample are separated into the three tasks, all participants are kept in the sample. Section 4.3.1.3 and 4.3.1.4 separates the samples in the three tasks and also in experienced and inexperienced participants.

Removed all participants that said they was distracted. 26 task results was removed,

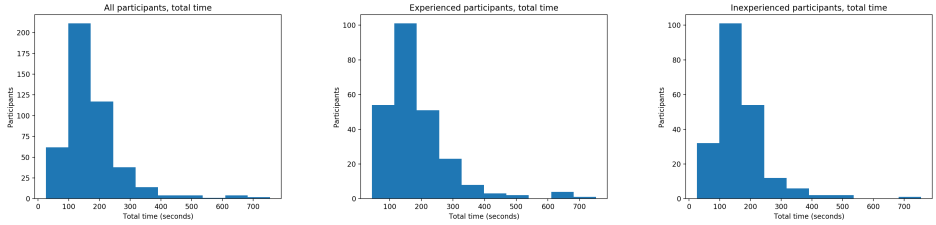
10 inexperienced and 18 experienced results.

4.3.1.1 All, experienced and inexperienced participants

The mean, standard deviation, minimum and maximum values are listen i table 4.1 and 4.2. I.

Total time per task (seconds)	All	Experienced	Inexperienced
Number of observations	429	229	200
Sample mean $\overline{x}$	170.32	177.65	161.94
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 (*4 entries per volunteer)*



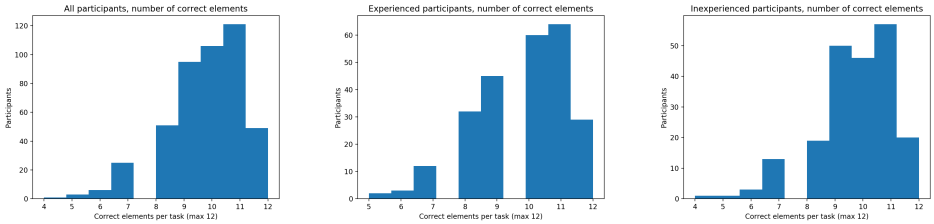
(a) All (b) Experienced (c) Inexperienced

Figure 4.3: Total time per task divided in all-, experienced- and inexperienced-participants

The histogram show that finishing a task within 100 seconds has a higher probability when the individual is experienced. Finishing the task in about 200 seconds is equally likely for both experienced and inexperienced individuals.

Correct elements per task	All	Experienced	Inexperienced
Number of observations	429	229	200
Sample mean $\overline{x}$	9.82	9.81	9.83
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 (*4 entries per volunteer)*



(a) All (b) Experienced (c) Inexperienced

Figure 4.4: Correctly chosen elements per task divided in all-, experienced- and inexperienced-participants

The histogram shows that for experienced individuals there are a higher probability of getting 12 correct than for inexperienced individuals.

4.3.1.2 All participants, divided in task 1, task 2 and task 3

In table 4.3 and 4.4 mean, standard deviation, minimum and maximum is listed for the three different task the participants did in the survey. Task 1 is the task that served the participants with one and one elements. Task 2 is the task that served the participants with three and three elements, and task 3 gave all six elements at the same time.

Total time per task (seconds)	One element task	Three elements task	Six elements task
Number of observations	146	142	141
Sample mean $\overline{x}$	166.38	172.25	172.48
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 into task 1, task 2 and task 3

Correct elements per task	One element task	Three elements task	Six elements task
Number of observations	146	142	141
Sample mean $\overline{x}$	10.19	9.71	9.55
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 into task 1, task 2 and task 3

4.3.1.3 Experienced participants, divided in task 1, task 2 and task 3

Dividing task 1, task 2 and task 3 results into experienced and inexperienced. Table 4.5 are data gathered about experienced participants total time per task. Table 4.6 are data gathered about experienced participants number of correctly chosen elements per task.

Total time per task	One element task	Three elements task	Six elements task
Number of observations	81	84	82
Sample mean $\overline{x}$	187.74	174.65	191.93
Standard deviation of $\overline{x}$	122.08	84.30	115.18
Standard error of $\overline{x}$	13.64	9.20	12.79
Minimum in sample	57.00	52.00	44.00
Maximum in sample	657.00	492.00	752.00

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

Correct elements per task	One element task	Three elements task	Six elements task
Number of observations	81	84	82
Sample mean $\overline{x}$	10.23	9.69	9.51
Standard deviation of $\overline{x}$	1.30	1.62	1.46
Standard error of $\overline{x}$	0.14	0.18	0.16
Minimum in sample	7.00	5.00	5.00
Maximum in sample	12.00	12.00	12.00

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

4.3.1.4 Inexperienced participants, divided in task 1, task 2 and task 3

Table 4.7 are mean time, standard deviation, minimum time and maximum time spent on each task for inexperienced participants. Number of correctly chosen elements per task for inexperienced participants is shown in table 4.8.

Total time per task (seconds)	One element task	Three elements task	Six elements task
Number of observations	71	69	70
Sample mean $\overline{x}$	159.28	174.42	169.50
Standard deviation of $\overline{x}$	68.24	106.69	83.17
Standard error of $\overline{x}$	8.10	12.84	9.94
Minimum in sample	47.00	26.00	38.00
Maximum in sample	487.00	755.00	529.00

Table 4.7: Inexperienced participant's time spent per task, divided by task

Correct elements per task	One element task	Three elements task	Six elements task
Number of observations	71	69	70
Sample mean $\overline{x}$	9.99	9.65	9.60
Standard deviation of $\overline{x}$	1.50	1.45	1.56
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 participant's number of correct elements per task, divided by task

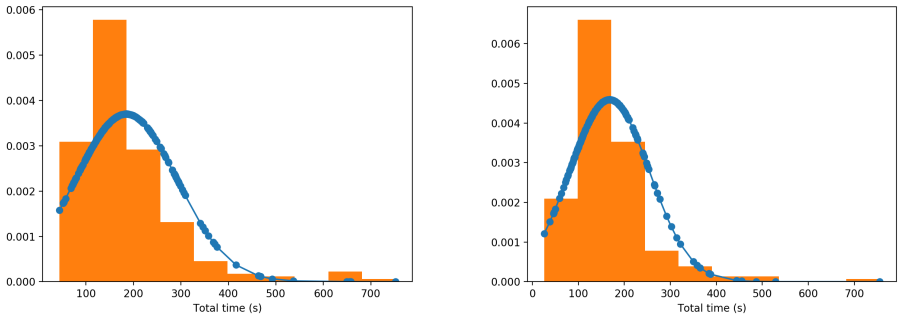
4.3.2 Normality tests

To check if a two-sample t-test (subsection 4.2.3.1) and *ANOVA*-test (subsection 4.2.3.2) can be used, the samples need to be tested if they are normally distributed or not. Both tests assume normally distributed samples. The normality section 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 tests. 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

4.3.2.1 Experienced and inexperienced participants - total time samples

The histograms 4.5a and 4.5b are positively skewed (see figure 4.1). This gives an indication that sample 1 and 2 are not normally distributed. Samples involving time measurements are rarely normally distributed. This is because the sample will always be skewed since it is impossible to have negative time. There will always be a limit to how fast a participant can finish the tasks.



(a) Sample 1 - Experienced

(b) Sample 2 - Inexperienced

Figure 4.5: Histograms with normal distribution fit with samples containing total time to complete each task

An D'Agostino and Pearson normality test (4.2.1) confirmed the visual assessment conclusion with an significance level of 5% (0.05). Both samples are not normally distributed with a confidence level of 95%.

D'Agostino and Pearson normality test
Significance level: 5%

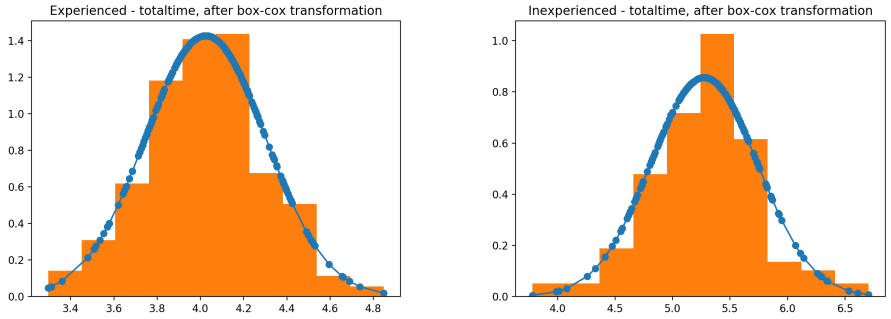
Sample 1: Experienced, total time per task
P-value: $3.874 * 10^{-22}$
The p-value is lower than the significance level (0.05), the null hypothesis is rejected and H_1 accepted.

Sample 2: Inexperienced, total time per task
P-value: $2.574 * 10^{-21}$
The p-value is lower than the significance level (0.05), the null hypothesis is rejected and H_1 accepted.

In both sample 1 and 2, the p-value was significantly lower than the significance level of 0.05. Data transformations are commonly used tools to improve normality of a sample distributions, but there are many types of data transformations. Osborne (2010) claim that almost all analyses, even non-parametric tests, benefit from improving the normality of the samples, especially when the normality test is significantly denied. Common traditional transformations are square root, inverse or converting to logarithmic scales (Osborne, 2010).

A Box-Cox power transformation is used in this thesis. This transformation can be used on positive data and the data used in this thesis will never negative. Box-Cox takes the idea of having a range of power transformations (square root $x^{\frac{1}{2}}$, inverse x^{-1} etc.) 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 transformation to be maximally effective in moving each sampled data towards normality. This is the reason why this thesis will use the Box-Cox transformation to hopefully achieve normally distributed samples.

The transformed data is shown in histogram 4.6a and 4.6b. A visual inspection gives a good indication that the transformed data is normally distributed.



(a) Sample 1 - Experienced participants

(b) Sample 2 - Inexperienced participants

Figure 4.6: Histograms with normal distribution fit after Box-Cox transformation

D'Agostino and Pearson normality test is then completed on the transformed data. This test confirms the visual inspection, both sample 1 and sample 2 are normally distributed after the Box-Cox transformation with an confidence level of 95%. Calculated p-value is larger than the significance level (0.05).

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

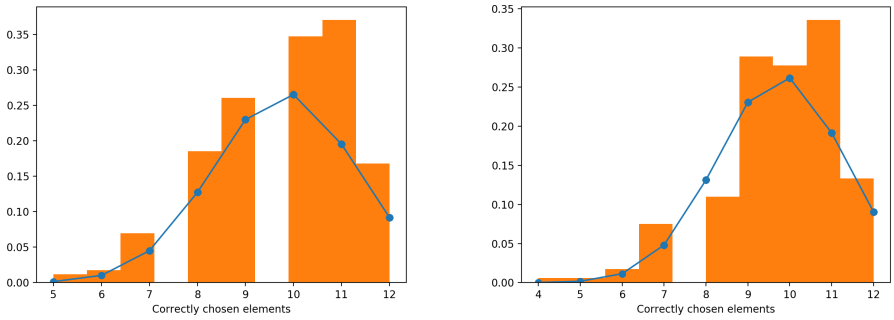
Sample 1: Experienced, total time per task
P-value: 0.849
The p-value is higher than the significance level (0.05), the null hypothesis is accepted.

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 are normally distributed is now accepted and can be used in parametric methods as the two sample t-test and *ANOVA* test.

4.3.2.2 Experienced and inexperienced participants - number of correctly chosen elements samples

Visual inspection of histogram 4.7a and 4.7b gives a good indication that sample 3 and 4 are not normally distributed. Both are negatively skewed (see figure 4.1).



(a) Sample 3 - Experienced

(b) Sample 4 - Inexperienced

Figure 4.7: Histograms with normal distribution fit with samples containing the number of correctly chosen elements in each task

D'Agostino and Pearson normality test confirm ou visual interpretation. Both samples accept the alternative hypothesis with p-values lower than the significant level 0.05.

D'Agostino and Pearson normality test
Significance level: 5%

Sample 1: Experienced, correct elements per task
P-value: 0.00443
The p-value is lower than the significance level (0.05), the null hypothesis is rejected and H_1 accepted.

Sample 2: Inexperienced, correct elements per task
P-value: 0.00013
The p-value is lower than the significance level (0.05), the null hypothesis is rejected and H_1 accepted.

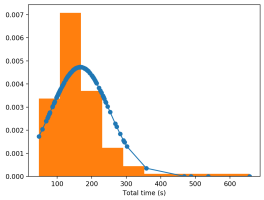
Sample 3 and 4 was then Box-Cox transformed. After transformation a new D'Agostino and Pearson normality test was done. Both samples also failed this test. Sample 3 and 4 are not normally distributed and need to be tested with non-parametric methods.

4.3.2.3 All participants - Task 1, Task 2 and Task 3 - total time per task

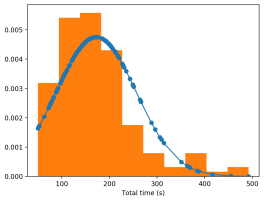
In this section the data is separated in three samples, each sample containing one of the tasks. The participants had to do three different tasks in the survey. Task 1 is the one element task, task 2 is the three elements task and task 3 is the six elements task. The three samples are named sample 5, 6 and 7. These samples will be used to test weather there are an significant difference between the three tasks when looking at the total time variable. The total time variable tells us how much time each participants spent on each of the three tasks. In this section sample 5, 6 and 7 will be normality tested.

Visual analysis of the three histograms in figure 4.8a, 4.8b and 4.8c show a positive skewness as the time histograms in section 4.3.2.1. This gives an indication that the three samples are not normally distributed.

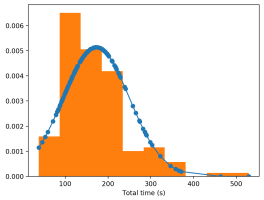
The D'Agostino and Pearson normality test agreed with the visual analysis. P-values for all samples are smaller than the significance level 0.05, and the null hypothesis is rejected. The samples are not normally distributed with a confidence level of 95%.



(a) Sample 5 - Task 1



(b) Sample 6 - Task 2



(c) Sample 7 - Task 3

Figure 4.8: Histogram with normal distribution fit - sample with total time per task

D'Agostino and Pearson normality test
Significance level: 5%

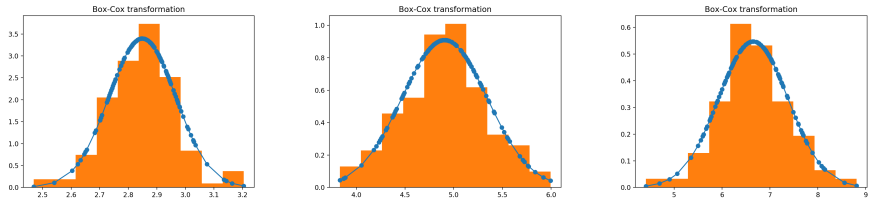
Sample 5: All, total time on task 1
P-value: $2.39 * 10^{-24}$
The p-value is lower than the significance level (0.05), the null hypothesis is rejected and H_1 accepted.

Sample 6: All, total time on task 2
P-value: $2.57 * 10^{-9}$
The p-value is lower than the significance level (0.05), the null hypothesis is rejected and H_1 accepted.

Sample 7: All, total time on task 3
P-value: $1.71 * 10^{-11}$
The p-value is lower than the significance level (0.05), the null hypothesis is rejected and H_1 accepted.

The samples are then Box-Cox transformed. The histograms after the transformation is shown in figure 4.9a, 4.9b and 4.9c. A visual analysis says that these histograms looks approximately normally distributed. The histograms looks like they have a skewness of approximately zero.

The D'Agostino and Pearson normality test confirms the visual conclusion. The data is normally distributed after the Box-Cox transformation with a confidence interval of 95%. The p-values of all three samples are higher than the significance level (0.05).



(a) Sample 5 - Task 1 (b) Sample 6 - task 2 (c) Sample 7 - task 3

Figure 4.9: Histogram with normal distribution fit after Box-Cox transformation, sample with total time per task

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

Sample 5: All, total time on task 1
P-value: 0.164
The p-value is higher than the significance level (0.05), the null hypothesis is accepted.

Sample 6: All, total time on task 2
P-value: 0.982
The p-value is higher than the significance level (0.05), the null hypothesis is accepted.

Sample 7: All, total time on task 3
P-value: 0.354
The p-value is higher than the significance level (0.05), the null hypothesis is accepted.

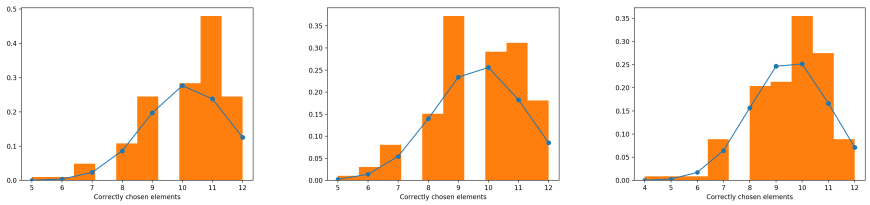
Sample 5, 6 and 7 are normally distributed after the transformation and the assumptions of parametric tests are met.

4.3.2.4 All participants - Task 1, Task 2 and Task 3 - correct elements per task

In this section the data is also separated in three samples, each sample containing one of the tasks. Task 1 is the one element task, task 2 is the three elements task and task 3 is the six elements task. The three samples are named sample 8, 9 and 10. These samples will be used to test weather there are an significant difference between the three tasks when looking at the number of correctly chosen elements variable. This variable tells us how many correct elements each participants chose on each of the

three tasks. Total number of elements is 12, 6 in each question, so the maximum correct is 12. First sample 8, 9 and 10 will be normality tested.

Visual analysis of the three histograms in figure 4.10a, 4.10b and 4.10c show a negative skewness, just like the histograms in section 4.3.2.2. This give an indication that the three samples are not normally distributed.



(a) Sample 8 -Task 1 (b) Sample 9 - Task 2 (c) Sample 10 - Task 3

Figure 4.10: Histogram with normal distribution fit showing samples with number of correct elements per task

D'Agostino and Pearson normality test confirms our visual inspection of the histograms in two of three samples. Sample 9 actually 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 normality test with a confidence interval of 95%.

D'Agostino and Pearson normality test
Significance level: 5%

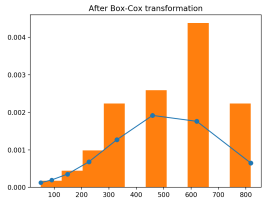
Sample 8: All, correct elements in task 1
P-value: 0.00022
The p-value is lower than the significance level (0.05), the null hypothesis is rejected and H_1 accepted.

Sample 9: All, correct elements in task 2
P-value: 0.099
The p-value is higher than the significance level (0.05), the null hypothesis is accepted.

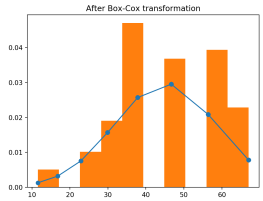
Sample 10: All, correct elements in task 3
P-value: 0.0047
The p-value is lower than the significance level (0.05), the null hypothesis is rejected and H_1 accepted.

To try to get all the samples approved in the normality test the Box-Cox transfor-

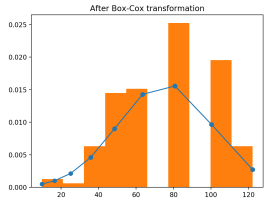
mation is applied to all three samples. The transformation changes the data, at to correctly compare the samples, sample 9 also has to be transformed.



(a) Sample 8 - Task 1



(b) Sample 9 - Task 2



(c) Sample 10 - Task 3

Figure 4.11: Histogram with normal distribution fit showing samples with number of correct elements per task - after Box-Cox transformation

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

Sample 8: All, correct elements in task 1
P-value: 0.00269
The p-value is lower than the significance level (0.05), the null hypothesis is rejected and H_1 accepted.

Sample 9: All, correct elements in task 2
P-value: 0.0752
The p-value is higher than the significance level (0.05), the null hypothesis is accepted.

Sample 10: All, correct elements in task 3
P-value: 0.2104
The p-value is higher than the significance level (0.05), the null hypothesis is accepted.

4.3.2.5 Normality test summary

	Sample	Normally distributed	Normally distributed after Box-Cox
Total time			
Experienced	1	No	**Yes**
Inexperienced	2	No	**Yes**
Correct elements			
Experienced	3	No	No
Inexperienced	4	No	No
Total time			
Task 1	5	No	**Yes**
Task 2	6	No	**Yes**
Task 3	7	No	**Yes**
Correct elements			
Task 1	8	No	No
Task 2	9	**Yes**	**Yes**
Task 3	10	No	**Yes**

Table 4.9: Summary of normality tests done in section 4.3.2

4.3.3 Levene's test

As mentioned in section 4.2.1 and 4.2.3.2, the two sample t-test and *ANOVA* assumes that the samples come from a population with equal variances. This assumption will be examined with levene's test. The hypothesis in this test is:

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

4.3.3.1 Sample 1 and 2 - experienced and inexperienced

Since sample 1 and 2 was normal distributed accepted after the Box-Cox transformation, will the Levene's test use the transformed data in the analysis.

Levene's test, sample 1 and 2
Significance level: 5%

P-value: $1.258 * 10^{-25}$

$P - value$ is smaller than the significance level ($1.258 * 10^{-25} < 0.05$) and the null hypothesis is rejected and H_1 accepted.

We cannot assume that sample 1 and sample 2 comes from populations with equal variances.

4.3.3.2 Sample 5, 6 and 7 - task 1, 2 and 3

Sample 5, 6 and 7 was normal distributed accepted after the Box-Cox transformation. Before using the samples in a *ANOVA test* they need to be equal variance tested.

Levene's test, sample 1 and 2
Significance level: 5%

P-value: 0.636

$P - value$ is larger than the significance level ($0.636 > 0.05$) and the null hypothesis is accepted.

4.3.4 Hypothesis testing

4.3.4.1 Two sample t-test results

Test differences between experienced and inexperienced participants
Sample 1 and 2

First hypothesis tested with the two sample t-test is:

H_0: Equal task time between participants
H_A: Unequal task time between participants

This test is covered by sample 1 and sample 2 from section 4.3.2.1. Sample mean and standard deviation for both samples is listed in table 4.1. Sample 1 is experienced participants and sample 2 inexperienced participants. $\overline{x}_1$ equals the mean total time for experienced-, and $\overline{x}_2$ the mean total time for inexperienced participants, the hypothesis can be written as:

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

Since we cannot assume equal variances in the two samples, 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 T-statistic calculated is smaller than the critical value. We then

conclude with that there is a significant difference between the means of the two population samples wit a confidence interval of 95%.

Two sample 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: 1.960

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_1 accepted.

Test if experienced or inexperienced participants finish the task fastest

The hypothesis tested is:

H_0: Equal total task time between all participants
H_A: Experience participants has a lower total time

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

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

This test gives the same T-statistics as the previous test, but the critical value is changed since it is an two sample, one way t-test. Since we cannot assume equal variances in the two samples a Welch's test is performed. The test results are shown under. The $T - statistic$ is still smaller than the critical value (-64.654 < 1.645). Our test is to check if sample 1 mean is significantly larger than sample 2 mean, then we use the test written in equation 4.2. Our $T - statistic$ is not larger than the critical value, therefore 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 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.

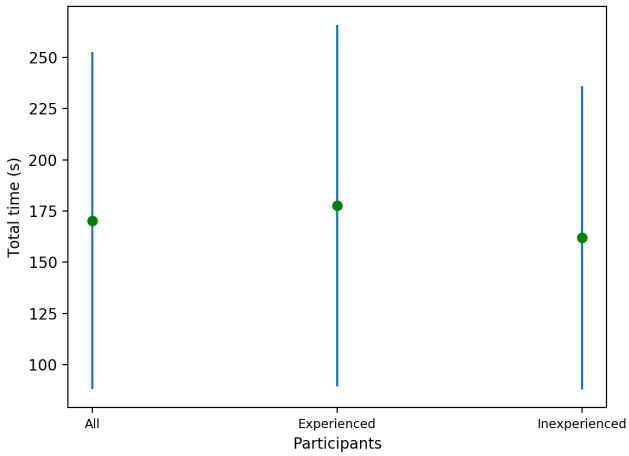


Figure 4.12: Error plot - mean (green dot) and standard deviation (blue line)

Since we know that there is a significant difference between the two sample means, the author concludes that the inexperienced participants finished the task faster than the experienced participants. This can also be seen in plot 4.12. Inexperienced participants finished the tasks 16 seconds faster than experienced participants.

4.3.4.2 Mann-Whitey U Test results

Test differences between experienced and inexperienced participants
Sample 3 and 4

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 is the correct samples to use in this test. Both samples are not normally distributed, we need to use a non-parametric method. The Mann-Whitey U test is the preferred test to use on these samples. As mentioned in section 4.2.3.3, the Mann-Whitey U test is preferred when the samples are ties (identical observations) in the data. From histogram 4.7a and 4.7b we see that the samples for both independent variables are similar. Mann-Whitey U test can therefore be used to compare the population medians. The hypothesis to be tested is:

$$H_A\text{: } median_3 = median_4$$
$$H_A\text{: } median_3 \neq median_4$$

The results from the test, the statistical value and finding the critical value in the Mann-Whitey U table is shown in the box under. Using equation 4.5 we conclude that there are not enough evidence to reject the null hypothesis with an confidence interval of 95%. The $U - statistic$ is larger than the critical value.

Two sample t-test, sample 1 and 2
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.

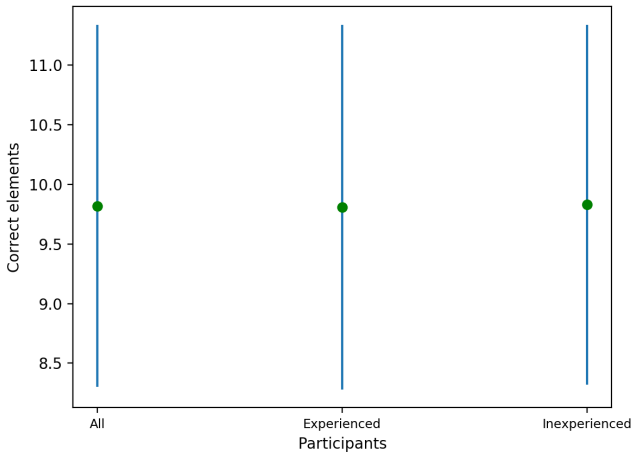


Figure 4.13: Error plot - mean (green dot) and standard deviation (blue line)

There is not enough evidence to conclude that there are any difference between experienced- and inexperienced participants when we use the dependent variable number of correctly chosen elements.

4.3.4.3 One-way *ANOVA* test results

The one-way *ANOVA* test will be used to test population means between the three task samples. This test is used when there is more than two samples being compared, see section 4.2.3.2. The assumption that sample 5, 6 and 7 come from populations with equal variances are met ()4.3.3.2). The hypothesis tested here is:

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

Using the one-way *ANOVA* test to answer the hypothesis. Using equation 4.4, the *ANOVA* test rejects the null hypothesis. P-value is approximately zero and this gives a good indication that the result is significant. With a confidence interval of 95% the author claim that there is a difference between the mean value of the three tasks.

One-way ANOVA, sample 5, 6 and 7
Significance level: 5%

$P - value$: $2.805 * 10^{-222}$
$f - value$: 2123.308
Significance level (α): 0.05
$v_1 = 2$, $v_2 = 426$ Critical-value: 3.00

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

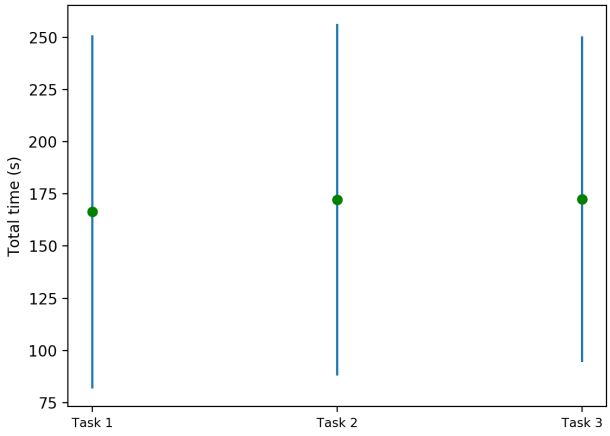


Figure 4.14: Error plot - mean (green dot) and standard deviation (blue line)

Using *Tukey's test* to make compared comparisons between task one, two and three since the null hypothesis. This test did not find any significant difference between the three tasks.

4.3.4.4 Kruskal-Wallis test results

The Kruskal-Wallis test is the non-parametric eqvivalent to one-way *ANOVA* (4.2.3.5). It is used to test equality of medians when the samples are not normally distributed. This method will test if there are any difference between task 1, 2 and 3 when looking at the dependent variable number of correctly chosen elements. The test will use sample 8, 9 and 10. Since sample 8 are not normally distributed a non-parametric test should be used. The hypothesis tested is:

$$H_0: median_8 = median_9 = median_10$$

H_A: Number of correctly chosen elements is different between at least two of the tasks

Kruskal-Wallis test, sample 8, 9 and 10
Significance level: 5%

$P - value$: $1.661 * 10^{-81}$
$H - value$: 372.004
Significance level (α): 0.05
$v_1 = 2$, $v_2 = 426$ Critical-value: 124.342

$H - value$ is significantly lower than the critical value (372.004 > 124.342) and the null hypothesis is rejected, H_A is accepted

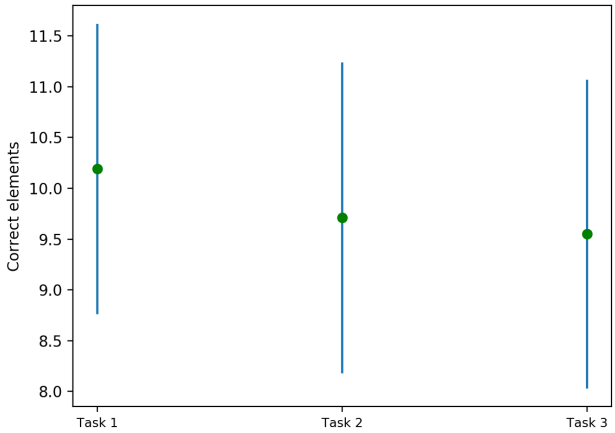


Figure 4.15: Error plot - mean (green dot) and standard deviation (blue line)

4.3.4.5 Hypothesis test summary

Hypothesis (Dependent variable, Independent variable)	Sample	Accepted
Total time, Experienced and Inexperienced		
There are a difference between experienced and inexperienced	1 and 2	**Yes**
Experienced finish the tasks faster than inexperienced	1 and 2	No
Correct elements, Experienced and Inexperienced		
There are a difference between experienced and inexperienced	3 and 4	No
Total time, Task 1, Task 2 and Task 3		
Total time is different between at least two of the tasks	5, 6 and 7	**Yes**
Correct elements, Task 1, Task 2, Task 3		
Task 1	8	No
Task 2	9	**Yes**
Task 3	10	No

Table 4.10: Summary of hypothesis tests done in section 4.3.4

5 | Proposed sections

5.1 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).

5.2 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 mi-

crotasks 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.

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: http://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.

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.

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: http://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.

Ha, A. (2016). CrowdFlower raises $10M to combine artificial intelligence with crowdsourced labor | TechCrunch. Date Acessed: 2017-05-08 URL: https://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.

ISO (1998). ISO 9241-11:1998(en), Ergonomic requirements for office work with visual display terminals (VDTs) — Part 11: Guidance on usability.

Israel, G. D. (1992). Determining Sample Size 1.

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

Kostas (2016). Using Crowdsourcing and Machine Learning to locate swimming pools in Australia · Tomnod. Date Accessed: 2017-05-04 URL: http://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: http://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.

Lund Research Ltd (2013b). Mann-Whitney U Test in SPSS Statistics | Setup, Procedure & Interpretation | Laerd Statistics. Date Accessed: 2017-05-12 URL: https://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: https://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: https://www.medcalc.org/manual/skewnesskurtosis.php.

Meier, P. (2013a). Digital Humanitarian Response: Moving from Crowdsourcing to Microtasking | iRevolutions. Date Accessed: 2017-05-04 URL: https://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: https://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.

Nikki (2016). Finding Swimming Pools in Australia using Deep Learning. Date Accessed: 2017-05-04 URL: http://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: http://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.

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.

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: https://www.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*.

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

Stanford University (2017). Machine Learning | Coursera.

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: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson.html.

Toutin, T. (2004). International Journal of Remote Sensing. *International Journal of Remote Sensing*, 25:10:1893–1924.

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.

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.