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

1 | Introduction

Doing fully automated operations on geospatial data can be proven difficult to accomplish.

This thesis aim is to study if micro-tasks can successfully be expanded to involving maps and geospatial data. The OpenStreetMap community have used the method some time, and the usage so far can evaluate as successful. This thesis wants to find out if inexperienced individuals also manage to solve tasks on maps that involved geospatial data. The study also wants to determine if the number of elements in each micro-task has an impact on how well individuals solve the microtasks. Both number of correctly chosen elements and time is measured. The thesis used a survey hosted through an web-application to gather participant data. The data is then used in statistics methods. The methods will then accept or deny the null hypothesis.

2 | Micro-tasking review

Today, geospatial data is more available than ever. Governments are releasing more and more data and OpenStreetMap 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).

2.1 Human computation

Utilizing the human processing power is still important. Humans are necessary even though our computers are becoming more and more complex. Traditional approaches for solving problems is to focus on improving the software, but often a solution that uses humans cleverly by exploiting the human brains cognitive abilities can create much faster and better results than a software. One of the pioneers of crowdsourcing, Luis von Ahn, created a game called "The ESP game". It solves the problem of labelling images with words. Most images don't have a proper caption associated with them. A fast and cheap method of labelling images is by using humans cleverly. Through "The ESP game" the players label 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). Another game that was created by Luis Von Ahn is called "Peekaboom". Here the players would locate objects in images. Such information is very useful in computer vision research for instance (von Ahn, 2008).

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 human and computers to accomplish tasks that neither can do alone.

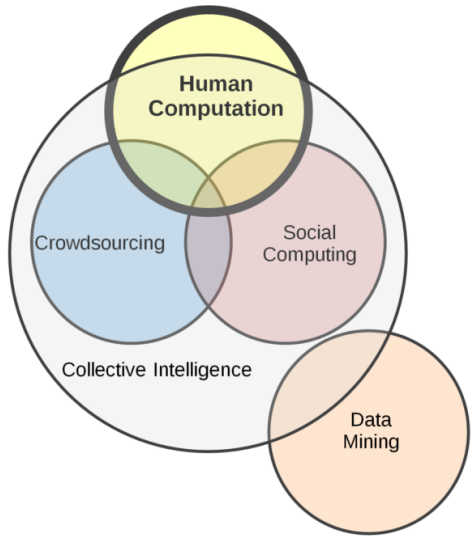

Figure 2.1: Collective intelligence (Quinn and Bederson, 2011)

2.2 Crowdsourcing

The first time the term "crowdsourcing" appeared was in Wires magazine article by Jeff Howe (Howe, 2006). Whereas human computing (section 2.1) 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 an widespread approach to dealing with machine-based computations where we leverage the human intelligence (Gadiraju et al., 2015).

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 use micro-tasking to create geospatial data in areas during crisis to support the help organizations. In machine learning process teams are starting to use micro-tasks too both validate the created data and also create test datasets to the algorithms. *Interpretation and analysis* tasks rely on the individual to use their interpretation skills during task completion. The task can be to choose between two layers, and decide which is best. This is the task-class used in this thesis during the survey. More in section 3.2.

2.3 Micro-tasking

Micro-tasks should not require any special training and a task should be completed within a couple of minutes (Ipeirotis and G., 2010).

With the establishment of micro-task crowdsourcing platforms as Amazon's Mechanical Turk (MTurk; www.mturk.com) 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).

Micro-tasking has already been used to process queries for instance. 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 to a micro-tasking platform. Most cases of micro-tasking usage exploit the large volume capabilities machines have and the cognitive capabilities of humans (Difallah et al., 2016).

The task refers to the activity, production or service the company or organization wants to have done. Gadiraju et al. (2015) findings when analyzing data from MTurk, indicate rapid growth in micro-task crowdsourcing.

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).

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.

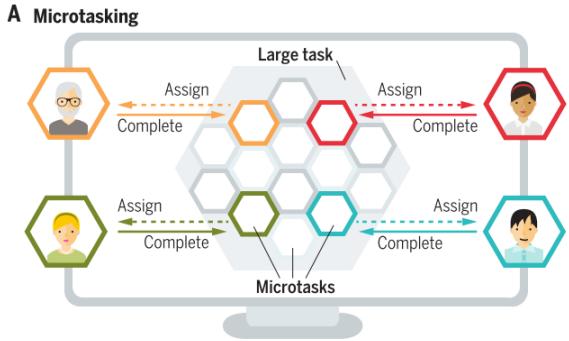

Figure 2.2: Micro-tasking (Michelucci and Dickinson, 2016)

2.3.1 Human Intelligence tasks

Thanks to micro-tasking platforms as Amazon's Mechanical Turk (MTurk), it is possible to build hybrid human-machine system that combines the scalability of computers with the yet unmatched cognitive abilities of the human brain (Difallah et al., 2016).

2.3.2 Micro-tasking platforms

2.3.2.1 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. In contains simple tasks such as labeling or segmenting images or tagging content, to more complex tasks such as translating or even edition text (2.3.2.2) (Franklin et al., 2011).

2.3.2.2 Soylent

Is a word processing interface that enables writers to call on Mechanical Turk workers to shorten, proofread, and edit parts of their document on demand. To improve the quality of the work, the Soylen team introduced the Find-Fix-Verify crowd programming pattern. This architecture splits tasks into a series of generating and review stages (Bernstein et al., 2015).

2.3.2.3 Tasking manager

2.3.2.4 Crowdflower

2.3.2.5 CrowdMap

CrowdMap is implemented using CrowdFlower. Is an approach to integrate human and computational intelligence in ontology [1] alignment tasks via microtask crowdsourcing (Sarasua et al., 2012). Ontology is still (2012) one of those problems that we cannot automate completely and having a human in the loop might increase the quality of the results of machine-driven approaches.

2.3.3 Usage

A machine learning company called "developmentSEED" use a micro-tasking solution for cleaning machine learning output data. They have created a GUI web application solution called Skynet Scrubber. 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".

People are good at comparing items, such as how well an image represents a particular concept. We 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 implementations of query operations, a huge cost and time sparing potential can be utilized.

2.3.4 Challenges

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 crowdworker model. Very complex tasks that can't be partitioned are not suitable for solving through micro-tasks.

It is important that operations added to a micro-tasking platform considers the talents and limitations of human workers (Franklin et al., 2011) and this is what this thesis try to examine. What is the limitations of human workers when dealing with 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).

One problem 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

[1] Ontology is a formal naming and definition of types, properties, and interrelationships of the entities that really or fundamentally exists for a particular domain of discourse. Ontologies are created to limit complexity and to organize information. The ontology can then be applied to problem solving.

micro-tasking platforms cannot guarantee on the task completion time (Difallah et al., 2016).

3 | Methodology and experiment

3.1 Experiment

This thesis will try to determine questions regarding micro-tasks around geospatial data. Little research is done on how well inexperienced individuals solve micro-tasks when they involve map interaction and geospatial data.

3.2 Survey

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's 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 informa-

tion 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. Creating cases with position errors

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

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.

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 was added to the movie with voice description. The design on the question one page was also improved by adding color to the text telling the participant which layer was 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. 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 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 maninly help with creating plots to better interpret the data and to more easily visualize the different results.

3.5.3 Preliminary results

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 non-experienced, so this is a surprising result. Figure 3.1 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.

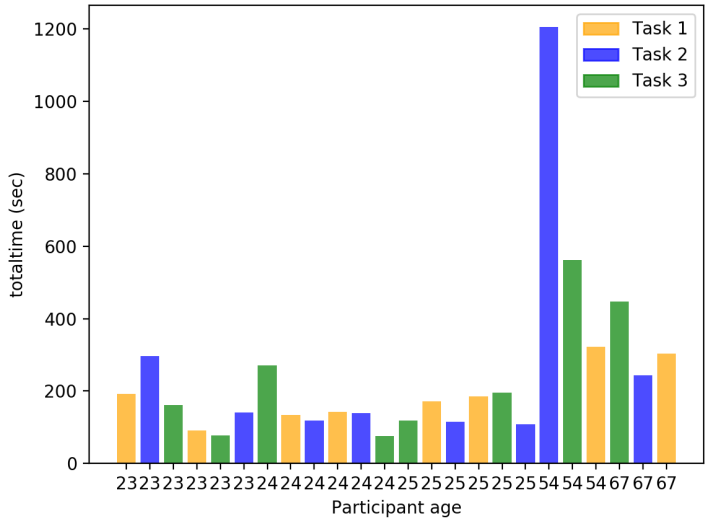

Figure 3.1: Total time - all participants ordered by age

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

In the pilot-test the same building 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 every time the result will not be as useful. Even though the task order varies. Reading of the data in figure 3.1, $\frac{7}{8}$ participants spent less time on the last task, even though the task order varies. Which matches the number of participants who remembered the buildings and meta information from the previous tasks.

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.

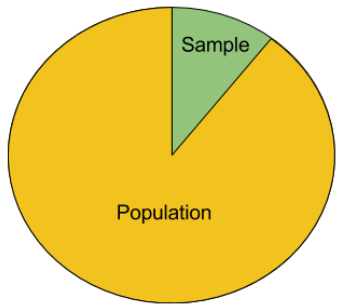

Figure 3.2: Population vs. sample

A sample is a collection of observations and is the subset of a population, illustrated in figure 3.2. 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 is basicly the population.

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 Statistics

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).

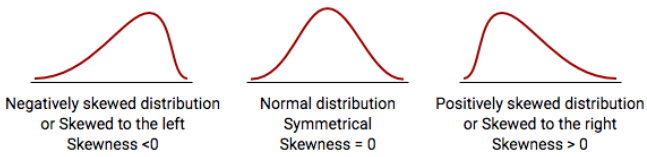

Figure 4.1: Skew (MedCalc Software bvba, 2017)

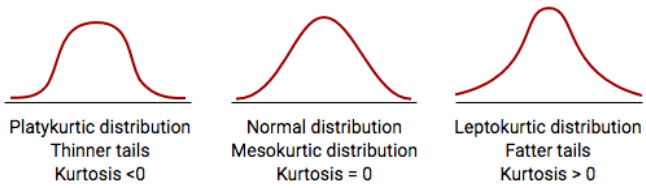

Figure 4.2: Kurtois (MedCalc Software bvba, 2017)

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 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, 2013b).

4.2.2.1 Two sample t-test

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

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

Then the hypothesis refer to a one-tailed two sampled t-test. 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.

Because the one-sided tests can be backed out from the two-sided tests. (With symmetric distributions one-sided p-value is just half of the two-sided pvalue). It goes on to say that scipy always gives the test statistic as signed. This means that given p and t values from a two-tailed test, you would reject the null hypothesis of a greater-than test when p/2 < alpha and t > 0, and of a less-than test when p/2 < alpha and t < 0.

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**:** Mean task time between participants are equal
H_A**:** Experienced participants finish the tasks faster, use less time

H_0: Total number of correct elements between participants are equal
H_A: Experienced participants have a higher number of correct elements

H_0: There are no difference in total number of correct elements between the tasks
H_A: Participants have more correct elements on the one element task

H_0: There are no difference in mean time between the tasks
H_A: Participants finish the one element task faster

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

4.2.2.2 Analysis-of-Variance

Analysis-of-Variance (*ANOVA*) is according to Walpole et al. (2012) a very common procedure used for testing population means. Where a two sample t-test 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 4.2.2.1.

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

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

μ equals the group mean and k represents the number of groups. It is important to check that each group are normally distributed, not only the sample (Lund Research Ltd, 2013b). 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.

In an one-way *ANOVA* test there should be one variable and minimum three independent groups, which is an relevant approach considering the data produced from this thesis survey. There are at least two 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* test 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 testen in the section blabla.

4.2.2.3 Wilcoxon Rank-Sum test

The Wilcoxon Rank-Sum test is an appropriate alternative to the two-sample t-test (see subsection 4.2.2.1) when the normality 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:

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

The alternative hypothesis depends on what the test should determine. If the sample

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

4.2.2.4 Mann-Whitey U test

Skriv om nødvendig, om vi bare bruker tid er det liten sannsynlighet at det et identiske målinger..

4.2.2.5 Kruskal-Wallis test

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

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

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

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

4.3 Survey results

4.3.1 Bar Charts

- All participants ordered by age
- All participants ordered by age, excluded by task 4
- All results in one task, ordered by age

Can use it to explain the data

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."

5.2 Usage potential

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.

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

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.

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

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.

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). 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.

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

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

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.

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/.

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/.

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.

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.