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 Methodology and experiment

1.1 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).

1.2 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 is conducted with a small sample of users. Results from the pilot-test can be used to determine the sample size. The sample size tells us how many responses that are needed (Smith, 2013). The formula for determining the sample size requires the standard deviation, how much variance to expect in the response (Smith, 2013). This standard deviation can be calculated from the pilot-test results.

A pilot test was conducted with a total of eight participants, five experienced and three non-experienced participants aged from 22 to 64 years. 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 was measures with the *System Usability Scale*(SUS). This scale gives an subjective measure of usability and was developed by John Brooke. The *System Usability Scale* questionnaire consists of ten statements where the participants rate their agreement in an 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).

The usability is important to measure. If the participants doesn'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. Usability is an important factor to get enough participants to do the whole survey and not quit halfway.

1.2.1 Execution of the pilot test

The pilot test started with a brief information about this study and the survey. They where 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 was asked to give general feedback on the web-application. The SUS score and the feedback was then used to determine the usability of the web-application and to determine which improvements to be done.

1.2.2 Results from the pilot test

The average SUS score was 84.64 out of 100.

All participants thought that the instructions movie was confusing. It was short, the instructions went too fast and it lacked voice descriptions. The movie needed major improvements.

Overall feedback on the tasks was that it was difficult to understand which building was which and also if the building shape layer was selected or not in question one. The lack of labels on the building was done on purpose to get the task as much as possible realistic. The selection of best fitting shape needed improvements, it had to be made clearer that selection was done by clicking on the shape, not by using the layer control as some thought.

Also, some pages had too much information and long sentences. The task progress bar was removed, no one noticed it, only the survey progress bar on the top right is necessary.

The two oldest participants spent almost twice as much time on the test than the younger. Maybe it where 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 where experienced and the other non-experienced, so this is a surprising result. CHECK THE TIME ON EACH TASK FOR THE OLDER PARTICIPANTS.

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.

1.2.2.1 Statistics result

First the pilot-test result need to be normality tested. As mentioned in 2.1.2 this thesis will use the Anderson-Darling test. The python library Scipy has an Anderson-Darling test function (The Scipy community, 2017), this was used to answer the normality test with the Anderson-Darling hypothesis. Testing the total time data form all participants (in total 32 entries) in Anderson-Darling gave a *p-value* of 0.717. The null hypothesis can not be rejected. When doing the analysis on experienced and non-experienced data set separately the null hypothesis was also not rejected. Notice that the *p-value* is lower because of fewer entries.

Anderson-Darling test
Data: All participants - total time, 32 entries
P-value: 0.717
P-value > 0.05
Normal distributed: Yes

Anderson-Darling test
Data: Experienced participants - total time, 19 entries
P-value: 0.692
P-value > 0.05
Normal distributed: Yes

Anderson-Darling test
Data: Non-experienced participants - total time, 12 entries
P-value: 0.679
P-value > 0.05
Normal distributed: Yes

2 Statistics

2.1 Statistics theory

2.1.1 Hypothesis testing

The null- and alternative hypothesis are statements regarding a difference or an 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, 2013).

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

2.1.2 Normal testing

A *goodness-of-fit* test is used to determine whether a sample of n observations can be considered as a sample from a given specified distribution (Walpole et al., 2012). The Anderson-Darling and the Kolmogorov-Smirow tests stand out as *goodness-of-fit* procedures specialized for small samples (Romeu, 2003). The Anderson-Darling test will be used in this study to test if the observations gathered in this study is normally distributed. The hypothesis for the Anderson-Darling test is:

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

The computations in the Anderson-Darling test differs based on that is known about the distribution. In this study both the expected mean and variance is unknown. In all the Anderson-Darling tests in this study a significance level of 0.05 is used, which gives a confidence interval of 95 %. If the calculated *p-value* is less than the significance level (0.05), the null hypothesis is rejected. The larger the *p-value* the closer match is the data to the normal distribution. The Anderson-Darling statistics is used to calculate the *p+value* for the *goodness-of-fit* test.

2.1.3 Tests on a Single Mean

When the variance is unknown we use the Student t-distribution. The random variables $X_1, X_2, X_3, ..., X_n$ represents the random sample from a normal distribution with unknown μ and σ^2.

2.1.4 Two sample 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 and μ_2 and variances σ_1^2 and σ_2^2. The one-sided hypothesis on two means can be written as:

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

Before doing tests on the two means Levene's Test is used to test if the samples are from populations with equal variances. If they are equal a two-sampled t-test may be used.

Appendices

A Tets

Fbox

Some text esfljsf lskj lksdjflsk slk

Some text kduhaszkdh aszkd-jhs zkjdfh skdj skd

dwkjdkwjdh wkjdhw kjdh wkjhd qwkjhd kwd qw .

text

dwkjdkwjdh wkjdhw kjdh wkjhd qwkjhd kwd qw .

Bibliography

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

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

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

Lund Research Ltd (2013). Hypothesis Testing - Significance levels and rejecting or accepting the null hypothesis.

Romeu, D. J. L. (2003). Anderson-Darling: A Goodness of Fit Test for Small Samples Assumptions. *START*, 10(5).

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 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.

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