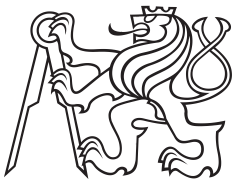
Master's Thesis



Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Computer Graphics and Interaction

User Evaluation of Label Placement Rules

Bc. Lilian Machulda

Supervisor: Ing. Ladislav Čmolík, Ph.D.
Field of study: Open Informatics
Subfield: Human-Computer Interaction
May 2023

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Machulda Lilian** Personal ID number: **483843**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Graphics and Interaction**

Study program: **Open Informatics**

Specialisation: **Human-Computer Interaction**

II. Master’s thesis details

Master’s thesis title in English:

User evaluation of label placement rules

Master’s thesis title in Czech:

Ověření pravidel pro umisťování popisků

Guidelines:

Get familiar with cartographical rules for label placement, especially with rules for label placement of point features. Further, analyze methods for automatic label placement of point features and with online mapping services (e.g., Google Maps) and determine which cartographic rules for label placement are not used by the online map services. Finally, get familiar with possibilities of remote empirical user testing. Based on the analysis, create example maps with labels placed for point features according to the cartographic rules. For the same example maps, place the labels according to the chosen online map service (e.g., Google Maps). Design and conduct remote empirical study which allows to determine if the cartographic rules influence the function of the labels (e.g., the labels can be easily associated with labeled points) or their aesthetics. Evaluate data collected in the study and report the results.

Bibliography / sources:

[1] E. Imhof, Positioning Names on Maps. The American Cartographer, Vol. 2, No. 2, pp. 128-144, 1975.
[2] P. Yoeli, The Logic of Automated Map Lettering. The Cartographic Journal, Vol. 9, No. 2, pp. 99-108, 1972.
[3] M. Kuniavsky, Observing the user experience: a practitioner's guide to user research. Elsevier, 2003.
[4] J. Sauro and J. R. Lewis, Quantifying the user experience: Practical statistics for user research. Morgan Kaufmann, 2016.

Name and workplace of master’s thesis supervisor:

Ing. Ladislav Čmolík, Ph.D. Department of Computer Graphics and Interaction

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **16.02.2023** Deadline for master's thesis submission: **26.05.2023**

Assignment valid until: **22.09.2024**

Ing. Ladislav Čmolík, Ph.D. Supervisor’s signature	Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master’s thesis, the author must state the names of consultants and include a list of references.

.	
Date of assignment receipt	Student’s signature

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Ing. Ladislav Čmolík, Ph.D, for his support and guidance throughout the work on this thesis.

Words cannot express my gratitude to my mom, who unconditionally showed me love and support throughout my studies and especially during hard times.

Finally, I would like to extend my thanks to my friends for their continuous support.

Declaration

I hereby declare that I am the sole author of this thesis and that I have cited all sources per the Methodical guideline No. 1/2009 for adhering to ethical principles when elaborating an academic final thesis.

Prague, May 2023

.....

Abstract

In the last few decades, we have witnessed a massive rise in computer usage [1]. This has also led to the rise of online map services [2].

When we compare the label placement in an online environment with physical maps, we can observe a significant difference in the mindset of such design. For this reason, we have decided to research the influence of label placement methods on the users of online map services.

We will be comparing the label placement methods of Google Maps with two custom-made methods that follow the principles of a label placement ruleset designed by Prof. Eduard Imhof in 1962 (republished in English in 1975) [6].

Our goal is to validate the ruleset in an online environment and compare it with label placement methods of Google Maps - since they seem to have their own principles.

Keywords: HCI, Research, Map, Label, Label Placement

Supervisor: Ing. Ladislav Čmolík, Ph.D.

Abstrakt

V posledních dekádách se výrazně zvýšila používanost počítačů [1]. To mimo jiné vedlo ke vzniku online mapových služeb [2].

Když porovnáme umístění popisků v online prostředí s fyzickými mapami, můžeme si všimnout významného rozdílu, v přístupu k návrhu daného rozmístění. Z tohoto důvodu jsme se rozhodli zkoumat vliv umístění popisků na uživatele mapových služeb.

Porovnávat budeme rozmístění popisků podle Google Maps a dvou vlastních metod, jež dodržují zásady umístění popisků podle Prof. Eduarda Imhofa z roku 1962 (znovu publikované v roce 1975 v angličtině) [6].

Naším cílem je ověřit tyto zásady v online prostředí a porovnat je s metodami Google Maps - jelikož se zdá, že mají vlastní zásady.

Klíčová slova: HCI, Výzkum, Mapa, Popisky, Umístění popisků

Překlad názvu: Ověření pravidel pro umisťování popisků

Contents

Figures

Tables

Introduction

Motivation

In the past few decades, computer usage has been on the rise [1]. During this time, there have been significant advancements in the field of web applications. This has led to the development of numerous web applications, including online map services, that are widely used to this day [2].

The user interface of computer applications often tries to mimic the familiarity of their physical counterparts, a phenomenon also known as skeuomorphism, which helps people create mental models based on their experience and adapt to new technologies [3]. Skeuomorphism was also used in the design of online map services. In this case, the design of the user interface comes from cartography, a field of science that specializes in the graphical representation of geographical data [4].

One of the most difficult problems of cartographic design is labeling - the process of placing labels in association with their corresponding features [5]. Eduard Imhof, a Swiss professor of cartography, published a paper outlining the principles and guidelines for labeling in 1962 [6]. Cartographers try to follow these principles during labeling but they still have to rely on their experience and intuition.

Labeling is also an issue for online map services. Pinhas Yoeli, a German cartographer set the basis for automatic label placement in 1972 with his paper where he mathematically systematized the ruleset by Prof. Imhof [7]. There have been significant advancements in the field of automatic label placement since then but there are still noticeable visual differences from hand-made labeling by a mapmaker.

From our observation, online map services do not follow all of the principles of Prof. Imhof. Since the differences seem to be rather significant and intentional, we have set out to evaluate whether the ruleset might be obsolete in the environment of online map services.

Aims and Objectives

In this thesis, we aim to evaluate Prof. Imhof's general principles of labeling point features in an online environment. For comparison, we will be using the labeling methods of Google Maps, one of the currently most used global online map services [8]. We will be comparing the labeling methods in terms of efficiency and subjective aesthetic preference.

To achieve our aims we have divided the thesis into three main parts. First of all, we will carry out an analysis of the selected labeling methods and their background. The following objective is to design and conduct an experiment based on the analysis. The final objective is to evaluate the data from the experiment and the impact of using Prof. Imhof's ruleset in an online environment.

Part I

Analysis

Chapter 1

Cartography

Cartography is the craft and science of making and studying maps [4]. Cartography has a rich history, possibly reaching times as far back as 25000 BC [9]. The technological capabilities and the medium have changed through time but the underlying principles and techniques have roughly stayed the same [10].

One of the components of cartography is labeling (likewise called lettering, label placement, or simply typography). Labeling is the art of designing and placing labels corresponding to the features depicted on the map [5]. The map features can be generally categorized as *point*[1], *linear*[2], or *area*[3] features [6]. Each one of these categories requires a different labeling approach. We will discuss the labeling principles for all types of features but the focus of this thesis lies in the labeling of point features.

Erwin Josephus Raisz, who is considered to be one of the most notable and influential cartographers of the twentieth century, noted that label placement is one of the most difficult problems of mapmaking. This opinion was shared among his colleagues during the pre-computer era [5]. As such, Eduard Imhof, a Swiss professor of cartography, well known for being the first president of the International Cartographic Association and his influence on relief shading techniques [11], decided to formulate the principles of labeling and the specific practices and rules to follow [6].

1.1 Ruleset by Eduard Imhof

In 1962 Prof. Eduard Imhof published a scientific paper *Die Anordnung der Namen in der Karte* [6]. The paper focuses on the best practices for labeling. It was later translated into English and republished in *The American Cartographer*, a journal currently known as *Cartography and Geographic Information Systems (CaGIS)*, under the name *Positioning Names on Maps*. The paper builds upon Prof. Imhof's experience in mapmaking and describes specific guidelines for the process of labeling. This chapter will summarize Prof. Imhof's work.

[1] For example, mountain peaks or cities on small-scale maps are point features.
[2] For example, rivers and roads are linear features.
[3] For example, countries and bodies of water are area features.

1.1.1 General Principles and Requirements

The paper presents six general principles to consider when labeling a map, though it is noted that there are, of course, exceptions since some rules can be mutually exclusive.

- The first principle talks about label *legibility*. In essence, Prof. Imhof says that labels should be easily read, discriminated, and located. He also mentions that legibility depends on the position of other labels.
- The second principle focuses on *clear graphic association*. That means that the label and the corresponding feature should be effortlessly identified and associated.
- The third principle touches on map *readability*. Prof. Imhof states that overlapping or concealment of other map elements by the label should be avoided.
- The fourth principle says that labels should also help convey *semantic information*, such as the territorial extent or connections.
- The fifth principle notes that *classification and hierarchy* of map elements should be emphasized by the text type choice - namely the type style and size.
- The sixth principle talks about *label distribution*. Perhaps counter-intuitively the goal is not to have evenly dispersed labels throughout the map. Nevertheless, we should also avoid dense clusters of labels.

1.1.2 Point Features

Point Features (likewise called *Position* or *Punctiform Designations*) are defined by Prof. Imhof as either points in the geometric sense (e.g. triangulation points) or objects small enough that the label cannot be placed within them.

General rules

In general, there are eight possible positions for a label - one for each cardinal direction (see Figure 1.1). Prof. Imhof mentions the best five positions in terms of legibility. At the time, due to technological limitations, it was also harder to position labels accurately. As such, he mentions that to improve legibility and execution it is best to avoid placing labels to the left of the point feature.

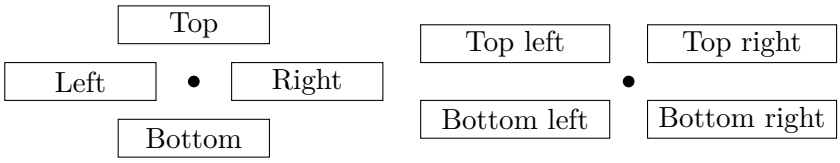


Figure 1.1: Possible label placement positions.

It is worth mentioning that any position is acceptable and with regards to other rules might even be unavoidable. The preferred positions, in reference to Figure 1.1, are ordered from the most preferred to the least preferred.

1. Top right
2. Right
3. Left
4. Top
5. Bottom

It is also very important to keep the label parallel with horizontal grid lines. Depending on the map projection (e.g. conic projection), it may be necessary to curve the label (see Figure 1.2). The label shall never be placed on a horizontal grid line or overlap it. Overlapping vertical grid lines is usually necessary and therefore allowed.

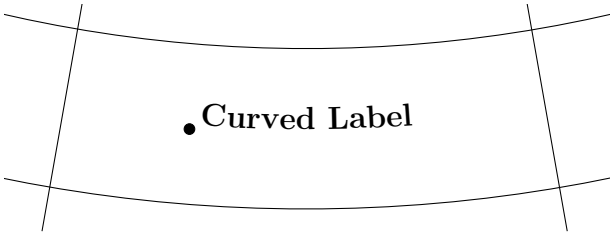


Figure 1.2: Curved label parallel to horizontal grid lines.

Places on Linear Features

To improve clarity, labels should be placed on the same side of a linear feature as the point feature it belongs to. If the point feature lies on both sides of a linear feature (e.g. a river flows through a city), the label should be placed over the linear feature. Alternatively, it may be placed to the right of the linear feature. This rule also applies to territorial borders and other linear map elements.

Places on Ocean Coasts and Lake Shores

This rule is perhaps the trickiest. The problem is that the aesthetic and legibility requirements are mutually exclusive. From an aesthetic standpoint, it would be best to place all labels inland. For legibility and semantics, it is best to differentiate between coastal and non-coastal places. As such, labels for coastal places should be written on a water surface and inland for all other places.

1.1.3 Linear Features

Linear Features (likewise called *Line Features* or *Linear Designations*) are depicted as lines or ribbon-like objects. They most commonly refer to rivers, streets, paths, airlines, and ship courses but on other types of maps may even refer to military lines or migration paths. It should be noted that rivers are considered to be a linear feature only on small-scale maps. On large-scale maps, once the label can fit wholly inside them, they are considered an area feature.

In general, labels should be placed along the linear feature. The label should follow the shape of the feature, though severe curvature should be avoided. When possible, the label should be placed in a location where the feature lies horizontally.

The label should always be placed above the feature and in vertical positions with the first letter towards the bottom of the map. The label should be adequately distanced from the feature and should never lie directly on it. For lengthy features, the label should be repeated at reasonable intervals. The label should also be repeated at junctions, where name changes often occur.

1.1.4 Area Features

Area Features (likewise called *Areal Designations*) are objects on the map within which a label can be placed. Most commonly they refer to territories and bodies of water, but their use is practically limitless. There are some exceptions though. Especially on small-scale maps, it may be better to treat islands and island groups as point features, as that allows us to see more of the land and potentially convey more information about the place (e.g. show roads and cities).

Prof. Imhof mentions that for clear association it is usually best to spread the label across the corresponding area. If the name is too long, it may be better to divide it into multiple lines, though there should be no syllabification. However, hyphenated names may be divided into multiple lines. Every line must begin with a capital letter.

For some areas, to retain legibility, it may be best to tilt the label. In such cases, the label should never be set on a straight line. Instead, the label should always be distinctly curved. Usually, the label is set on a single arc of no more than 60 degrees. For longer names, it may be necessary to use a double curved line.

Chapter 2

Automatic Label Placement

Automatic label placement is the problem of automatically finding adequate positions for labels on maps or other figures. The problem can be further split into internal and external labeling (i.e. whether the labels are placed inside or outside the figure) [12]. We will only focus on the internal labeling of maps, which is the problem cartographers are usually faced with.

Still, there are other important factors to consider. First, we have to consider what map features we will be labeling. For the purpose of this thesis, we will only look into the automatic labeling of point features, a problem also known as *point-feature label placement (PFLP)*. Another essential aspect of automatic labeling is what positions we will consider for the labels. For point features this usually means that we either allow any position or just a handful of fixed positions around the point feature (e.g. one in each of the eight cardinal directions as depicted in Figure 1.1) [13]. The first variant is usually not adopted by online map services, therefore we will only focus on variants of the PFLP problem with fixed positions. The last aspect we will consider is whether we are labeling static or dynamic maps. We will first look into static maps, which was the problem cartographers were historically faced with and then we will extend this with the requirements of labeling dynamic maps which are widely used in online map services today.

The basic PFLP problem was shown to be an NP-hard problem [13]. Through the years there have been many different suggestions for solving this problem efficiently. We will go over some of the initial approaches, which are essentially the foundation for the more sophisticated techniques suggested more recently [14].

Exhaustive Search

Exhaustive search algorithms are quite simple in their nature. We go sequentially over the point features in the map and assign an unobstructed position to the corresponding label [13]. If we get to a label that cannot be positioned (i.e. the label has no unobstructed position or all positions have been tried already) we backtrack to the previous label and try the next unobstructed position. We continue this process until we have successfully labeled all points or until we have exhausted all options. Even though the implementation of exhaustive search algorithms is quite easy, the computational complexity makes them only applicable to small problems [13].

Greedy Algorithms

Greedy algorithms are in a way a relaxed version of the exhaustive search [13]. To reduce computational complexity we compromise on finding the optimal solution. Instead of backtracking we either allow the point and the corresponding label to be completely left out or we simply allow overlaps. To end with a reasonable label placement, heuristics are used to determine the order in which we go over the points [13].

Discrete Gradient Descent

An improvement of the greedy algorithms came in the form of discrete gradient descent [13]. This method repeatedly adjusts the label positions in a way that improves the overall labeling quality. For example, we can label the map using a greedy algorithm where we allow overlapping. After we place all the labels according to the greedy algorithm we consider changing the position of each label to all other possible positions. We calculate a score for each possible change and apply the one that improves the overall score the best. We can then keep repeating this process to improve the overall labeling. Unfortunately, this method has a substantial weakness in the form of local minima. This means that the algorithm might get stuck with a solution that could be further improved upon but the algorithm will not find it because the algorithm would have to consider making changes that do not immediately improve the overall score. But even with this weakness, even the simplest discrete gradient descent algorithms are a dramatic improvement over the previously mentioned alternatives [13].

Further Improvements

Since then more sophisticated techniques have been researched. Notably, Zoraster (1990) reduced the PFLP problem to mathematical programming [15], Verner et al. (1997) proposed a genetic algorithm for the PFLP problem [16], and more recently Mote (2007), Luboschik et al. (2008), and also Pavlovec and Čmolík (2022) proposed algorithms that are very fast and

scalable [17][18][19]. Often, the proposed techniques look at PFLP as an optimization problem with the sole goal of labeling as many features as possible. This might not always be desirable, as this often produces ambiguous labeling. As such, there have been attempts at addressing this weakness. Notably, Rylov and Reimer (2014) proposed a multi-criteria optimization model for PFLP which focuses on the cartographic requirements of labeling [14], such as those proposed by Prof. Imhof (see Section 1.1).

Dynamic Maps

The PFLP problem becomes more complex for dynamic maps. By dynamic maps, we mean maps in an environment that allows panning around and changing the map scale (zooming). There are generally four requirements that should be met when labeling dynamic maps [20].

The first requirement states that labels currently in the view area should not disappear when zooming in or appear when zooming out. This requirement represents the expectation of getting a more detailed view when zooming in and a less detailed one when zooming out. The second requirement expresses that the position (on the map, not the label placement position as discussed earlier) and size of a label should change continuously with the pan and zoom operations. This just means that the label should stay in the same place relative to the map and the size should change proportionally with the zoom level. The label placement position should however stay the same. The third requirement states that except for sliding in or out of the view area, labels should not disappear or appear during panning. The fourth requirement notes that the labeling (i.e. what labels are shown in what positions) should not be dependent on the previous operations. This simply ensures that the map is labeled the same regardless of the starting position and the operations we applied to get to the specific state. All of these requirements could be easily violated by labeling only the features currently visible in the frame [20].

Chapter 3

Online Map Services

The creation of online map services has arguably played a major role in the advancements of the field of cartography today. At the very least it has made cartography more accessible and has had an impact on the expectations people have for maps [21].

3.1 Interactivity

Interactivity plays an important role in the user experience of online map services [22]. The most noteworthy features are in our opinion zoom and pan, allowing the user to move around the map freely. Usually, semantic zoom is used. Semantic zoom is a feature that changes the map scale and controls the level of detail the user sees on the map. Typically, more labels are shown with increasing zoom level, allowing the user to zoom in more to get more context.

Most current online map services also have more advanced features. Presumably, the most widespread are search (allowing the user to search for specific locations and showing them in great detail) and navigation. Some services also feature street view, a feature that allows the user to view the selected area from ground level rather than the standard top-down view.

3.2 Google Maps

Google Maps are currently among the most used global online map services [8]. We have observed that they do not follow the principles of Prof. Imhof. As such, we have decided to conduct further research into their methods of labeling.

3.2.1 Label Placement

To our knowledge, the label placement methods of Google Maps are not publicly accessible, therefore we will attempt to deduce the methods just from our observations.

Point Features

Google Maps distinguish between many different types of point features. These include but are not limited to mountain peaks, public transport stops, cultural venues, restaurants, and other businesses. On small-scale maps, cities and towns are also included.

For the aforementioned businesses and cultural designations, Google Maps use the iconic pin symbol. If there is enough space and the place is notable enough, there is also a text label included. From our observation, there appears to be an order of preferred positions for the text label, presumably to stay consistent. The positions, in reference to Figure 1.1, are ranked below from the most preferred (at the top of the list) to the least preferred (at the bottom of the list).

1. Right
2. Left
3. Top left
4. Bottom left

The symbol used to show the location of cities and towns is a circle rather than a pin. Google Maps seem to use three different styles of circles to denote the size and significance of each city: a small white circle with a black outline for small cities, a bigger white circle with a black outline for larger cities, and a black circle with an offset black outline for the capital city. The order of preferred positions for the text labels is also different.

1. Top
2. Bottom
3. Right
4. Left

If the label cannot be placed in either of these positions, the label and the corresponding symbol will not be shown.

From our observation, Google Maps do not mind when a label overlaps other map elements such as shorelines, rivers, or even region borders. The only exception we observed was with country borders, where Google Maps try to prevent overlapping. That said, there are still cases where some overlapping is allowed as long as the label can be placed in the most preferred position and only a small portion of the label overlaps.

Linear Features

Out of the linear features specified by Prof. Imhof, Google Maps label streets, rivers, and also ship courses. However, Google Maps also label country and province borders on large-scale maps. Other linear features such as railroads and are not labeled. The labeling method stays the same regardless of the map scale.

Labeling of streets, rivers, and ship courses seems to be the same. The label follows the shape of the linear feature and if there is enough space, the label is repeated throughout its length. The label is always placed in the center of the linear feature, even if it causes the label to overlap the edges of the feature.

The method for labeling region borders is different, however. The labels for each respective region are placed parallel to one another, one on each side of the border, and the labels are repeated throughout the length of the border. However, the labels are never curved to the shape of the border. Instead, they are straight but placed parallel to the general direction of the border (see Figure 3.1).

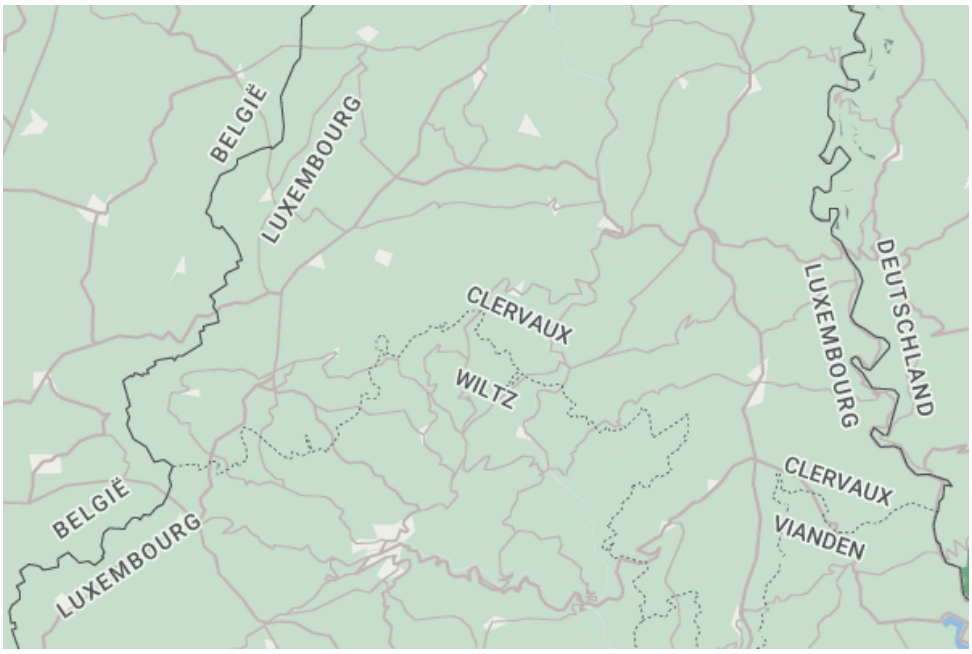


Figure 3.1: Country borders of Belgium, Germany, and Luxembourg and canton borders of Luxembourg. *Map data ©2023 GeoBasis-DE/BKG (©2009), Google.*

For all linear features, the label is always placed so it is the right way up. For strictly vertical labels, it appears that Google Maps prefer to put the first letter toward the top of the map.

Area Features

Google Maps recognize many different types of area features. For example oceans, seas, lakes, parks, and countries. On large-scale maps also cities, towns, and even city districts or neighborhoods. From our observation, the features can be divided into two groups, by their respective label placement rules.

The first group we will consider is bodies of water. This group consists of oceans, seas, gulfs, bays, channels, lakes, and dams. Arguably, we could also include rivers but since it is mostly a linear feature and the labeling rules stay the same regardless of map scale, we will omit it in this section.

There appear to be two ways of labeling bodies of water. The preferred way is to place the label horizontally, which is also preferred by Prof. Imhof since it is easy to read that way. If the label is too long, it is divided into multiple lines. This rule seems to be applied mostly to larger bodies of water. The other way is to make the label follow the shape of the body of water. However, in contrast with the principles of Prof. Imhof, if the label is not horizontal it is not required to be visibly curved. What surprised us is that the method does not change with the map scale. This causes overlapping with land in some cases (see Figure 3.2).

Figure 3.2: Label of Tyrrhenian Sea overlapping land. *Map data ©2023 Google, INEGI.*

The second group includes countries, states, provinces, cities, towns, and city districts or neighborhoods. The labels in this group are always placed horizontally and in the center of the area. The rationale behind this design might be that people usually do not rotate the device displaying the online maps, as you would do with a physical map, but rather use the rotate map feature. As such, Google Maps presumably keep the label level for better readability.

Placing the label horizontally and in the center of the area means that the label may overlap its borders for narrow and small countries, states, or provinces. This may in our opinion cause some ambiguity - as an example, we have selected Equatorial Guinea and Singapore (see Figures 3.3, 3.4).

(a) : A small-scale map showing Equatorial Guinea and bordering countries.

(b) : A large-scale map showing Equatorial Guinea and bordering countries.

Figure 3.3: On a small-scale map, the label of Equatorial Guinea overlaps the country's borders. Intriguingly the label of Gabon, which is a much bigger country, is not displayed at all. That in our opinion causes ambiguity in what land the label of Equatorial Guinea references. *Map data ©2023 Google, INEGI.*

(a) : A small-scale map showing Singapore and bordering countries.

(b) : A large-scale map showing Singapore and bordering countries.

Figure 3.4: On a small-scale map, the label of Singapore overlaps the country's borders and spans over other regions. That in our opinion causes ambiguity in what land the label references. *Map data ©2023 Google.*

Chapter 4

Conclusion

For this thesis, we have decided to focus only on the general principles of labeling point features. This includes the six general principles of labeling and the label position priorities, as discussed in Section 1.1. Specifically, we will focus on label placement for cities, but we think it would be beneficial to do further research into other rules specified by Prof. Imhof. We have come to this decision because each of the rules has some level of variability and would require an experiment on its own. We also felt that applying other rules for point features would add other variables to the experiment and we would not be able to interpret the result for each rule separately.

We have selected to compare three labeling methods. The first method, which we will further call *Legibility*, will follow the principles and rules outlined by Prof. Imhof as closely as possible. The second method, which we will call *Consistency*, will try to place labels on the preferred position, specified by Prof. Imhof, for maximum consistency, even in cases where it will produce clusters of labels. This method is in part inspired by online map services, where label placement consistency seems to be an important factor. The last labeling method that we will compare is the labeling method of *Google Maps*.

We aim to quantify the differences between the selected labeling methods. Therefore, we will conduct quantitative experimental research. The only independent variable in the experiment will be the placement of the labels. As such, we will not allow interactivity (e.g. zoom and pan). Interactivity would add unnecessary complexity to the experiment, as we are only trying to compare the labeling methods, which we can do in a static environment. This will also allow us to maintain equal conditions for all participants.

Part II

Experimental Design

Chapter 5

Hypotheses

In this experiment, we will be comparing multiple labeling methods. Since we have set out to compare both efficiency and subjective aesthetic preference, we will divide our hypotheses into two separate categories.

5.1 Efficiency

This category of hypotheses will help us infer differences between the labeling methods in terms of efficiency. To test the differences between the labeling methods, will formulate a global hypothesis comparing all labeling methods at once. If we reject the global hypothesis, we will conduct a post-hoc test where we will compare the labeling methods in pairs. Each hypothesis in this category is in practice two hypotheses, each for a different variant of the assignment task, as described later in Section 6.1.

There are four global hypotheses in this category. The first two will compare the objective differences between labeling methods. The first hypothesis will compare the differences between labeling methods in terms of the error rate and the second in terms of the completion time of the tasks.

Error rate:

H_0^E There is no error rate difference between *Legibility*, *Consistency*, and *Google Maps*.

- Pairwise hypotheses for the post-hoc test:

  $H_0^{E_1}$ There is no error rate difference between *Legibility* and *Google Maps*.

  $H_0^{E_2}$ There is no error rate difference between *Legibility* and *Consistency*.

  $H_0^{E_3}$ There is no error rate difference between *Consistency* and *Google Maps*.

Completion time:

H_0^T There is no difference in completion times between *Legibility*, *Consistency*, and *Google Maps*.

- Pairwise hypotheses for the post-hoc test:

  $H_0^{T_1}$ There is no difference in completion times between *Legibility* and *Google Maps*.

  $H_0^{T_2}$ There is no difference in completion times between *Legibility* and *Consistency*.

  $H_0^{T_3}$ There is no difference in completion times between *Consistency* and *Google Maps*.

Additionally, we will measure subjective scores of correctness and speed of completing the tasks, as reported by the participants. We will use these scores as a secondary indicator of how well the labeling methods compare.

Correctness (subjective):

H_0^C There is no difference in the subjective score of correctness between *Legibility*, *Consistency*, and *Google Maps*.

- Pairwise hypotheses for the post-hoc test:

  $H_0^{C_1}$ There is no difference in the subjective score of correctness between *Legibility* and *Google Maps*.

  $H_0^{C_2}$ There is no difference in the subjective score of correctness between *Legibility* and *Consistency*.

  $H_0^{C_3}$ There is no difference in the subjective score of correctness between *Consistency* and *Google Maps*.

Speed (subjective):

H_0^S There is no difference in the subjective score of speed between *Legibility*, *Consistency*, and *Google Maps*.

- Pairwise hypotheses for the post-hoc test:

  $H_0^{S_1}$ There is no difference in the subjective score of speed between *Legibility* and *Google Maps*.

  $H_0^{S_2}$ There is no difference in the subjective score of speed between *Legibility* and *Consistency*.

  $H_0^{S_3}$ There is no difference in the subjective score of speed between *Consistency* and *Google Maps*.

5.2 Aesthetic Preference

This category has a single family of hypotheses. This family contains hypotheses comparing the subjective aesthetic preferences between pairs of labeling methods.

Aesthetic preference:

$H_0^{A_1}$ There is no aesthetic preference between *Legibility* and *Google Maps*.

$H_0^{A_2}$ There is no aesthetic preference between *Legibility* and *Consistency*.

$H_0^{A_3}$ There is no aesthetic preference between *Consistency* and *Google Maps*.

Chapter 6

Tasks

The tasks in this experiment can be divided into two parts. The first part will serve as the basis for evaluating the effectiveness of each labeling method and the second part will be used for the subjective evaluation of each labeling method.

6.1 Assignment Task

The experiment's main objective is to test the participants' ability to assign the corresponding map features to their labels and vice versa. To compare the efficiency of the labeling methods, we will compare the results in terms of error rate and speed, at which the participants performed the assignment.

The assignment task has two variants, each of which will be evaluated separately. In the first variant, the participant is shown a map with a highlighted map feature and is prompted to select (i.e. click on) the corresponding label as quickly as possible. In the other variant, the participant is shown a map with a highlighted label and is prompted to select (i.e. click on) the corresponding map feature as quickly as possible. An example for each variant of the assignment task is shown in Figure 6.1.

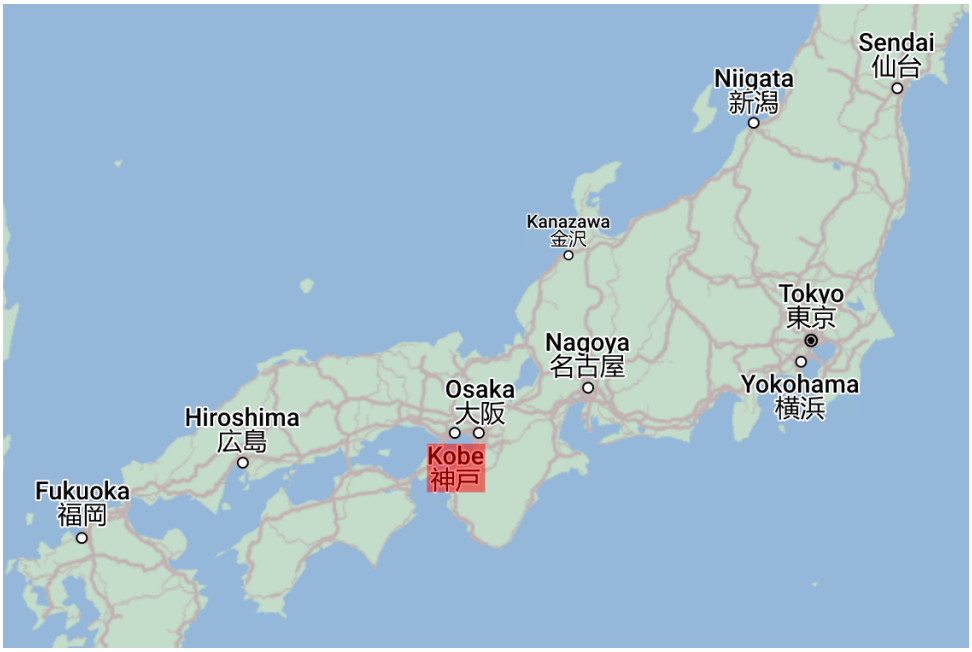
6.2 Subjective Perception

Subjective perception has a significant impact on the rating of usability. People rate aesthetically pleasing interfaces as more usable [23][24]. Perceived aesthetics also have an impact on the completion time of tasks [24].

6.2.1 Rating of Confidence

As an additional indicator of how well the labeling method performs and to gain an insight into the perception of participants on how they performed, there is an intermission after each group of assignment tasks. Each time, the participant is asked to evaluate their performance from the previous tasks.

(a) : First variant of the assignment task. The participant was prompted to select the label corresponding to the highlighted point. *Map data ©2022 GeoBasis-DE/BKG (©2009), Google.*



(b) : Second variant of the assignment task. The participant was prompted to select the point corresponding to the highlighted label. *Map data ©2022 Google, TMap Mobility.*

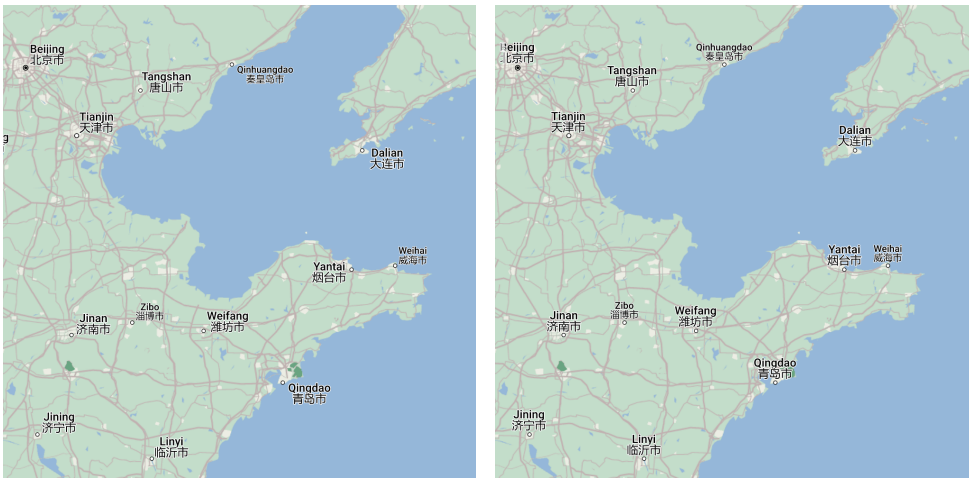
Figure 6.1: An example of the assignment task. Both variants of the task are shown in subfigures (a) and (b), respectively.

The questions are split into two identical parts. One for each variant of the assignment task (i.e. selecting either a point or a label). In each part, the participant is asked to select the option which best represents how they feel about their performance. We use the 5-point agreement Likert scale [25], so they are specifically asked to select how much they agree with the following statements: *I was sure in the assignment*, *I was fast in the assignment*.

6.2.2 Aesthetic Preference

To measure the aesthetic preference of participants we used Two-alternative forced choice (2AFC) tasks. This method was originally used to detect sensory discrimination thresholds [26] but is also used to assess more complex stimuli such as aesthetic preference [27].

In 2AFC tasks, the participant is asked to choose one of two stimuli with regard to the parameter being measured (e.g. which of the two stimuli is more aesthetically pleasing). In our experiment, this means that the participant was shown a pair of maps, each using a different labeling method, and was asked to select the one they preferred aesthetically (see Figure 6.2).



(a) : A map labeled using the *Legibility* labeling method

(b) : A map labeled using the *Google Maps* labeling method

Figure 6.2: An example of the 2AFC task. The participant was shown a pair of maps using a different labeling method and was asked to select the map they liked better. *Map data ©2022 Google, TMap Mobility.*

2AFC tasks have been shown to be significantly more effective in measuring aesthetic preference compared to using rating choice or rating scales [28], which are essentially the only alternatives to the 2AFC method in tasks where the goal is to obtain direct preference measurements [27]. They are also simpler for the participants because they require essentially no memory load [27]. 2AFC tasks have also been shown to have minimal response bias [27] and reduce fatigue and carryover effects when compared to techniques where the participant is asked to assess three or more samples simultaneously [29].

Chapter 7

Methodology

The experiment was designed to have a single independent variable - the label placement. Having a single independent variable will allow us to draw conclusions based directly on the changes in that variable. However, there are other important aspects of the experimental design, which are discussed below.

7.1 Between-subject Design

We use a between-subject design for the assignment tasks and the associated subjective evaluation (see Sections 6.1, 6.2.1). This means that for these tasks, the participants are randomly divided into three disjoint groups, each of which is exposed to a different labeling method. Each group is only exposed to that single labeling method for the duration of these tasks.

A substantial advantage of using between-subject design over within-subject design is the absence of carry-over effects and experimenter demand [30]. Carry-over effects occur when the participant is affected in subsequent experimental treatments by the previous experimental treatments [31]. Experimenter demand is a phenomenon that occurs in experiments where only a single variable is changed between the experiment variants. The participant naturally notices the change in the variable and may consider that a prompt to change their behavior [30].

Carry-over effects are very undesirable for our experiment. Mainly because in subsequent trials the selected answers could be affected by previous trials, distorting the measured error rate of each labeling method and introducing bias. A within-subject design would also make the experiment much longer, which would certainly negatively impact the completion rate [32] and we would therefore collect less data. Evidently, we only have a third of the responses for each labeling method but we can instead test more scenarios and keep the experiment duration the same.

7.2 Maps

To prevent additional variables in the experiment, all three methods will label the same points. The labeling method of *Google Maps* might have a slight advantage compared to the custom-made methods since the map data (i.e. which cities are labeled) is directly taken from Google Maps. This means that only the labels that the labeling algorithm of Google Maps evaluated as readable are shown. The other two methods have to make do with the selected points. This for example means that the *Consistency* labeling method ends up creating clusters of labels in some regions, which is generally discouraged [6]. This was in our opinion the only way to get accurate labeling for the *Google Maps* labeling method and also have the map variants comparable.

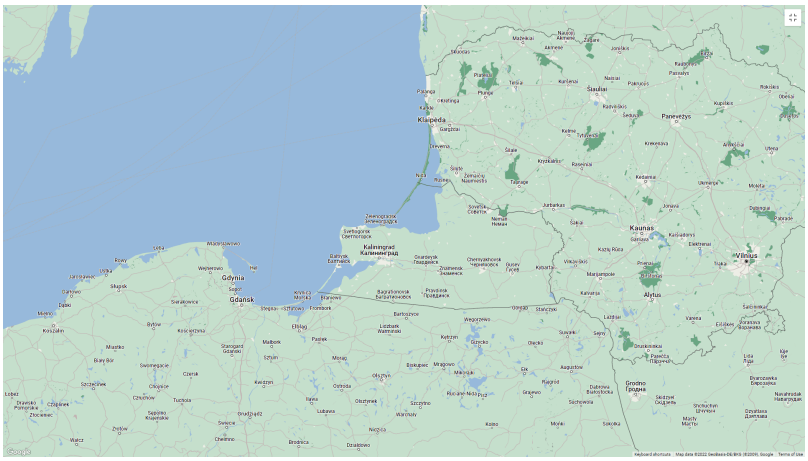
We have selected four regions to test. This allows us to create more test cases to test different labeling principles and also keep the experiment relatively short. Testing multiple diverse regions also minimizes the risk of accidentally selecting a region where one labeling method performs significantly better due to unprecedented circumstances. Keeping the number of regions even also produces fewer combinations needed for counterbalancing order effects.

The first region is eastern China, around the Yellow Sea and near the Korean Peninsula. In this region, Google Maps produce a map with very little interference between labels. Here we hope to test the differences between the preferred labeling positions and also the impact of labels overlapping land borders. The second region is Japan. This region also has labels overlapping land borders and some interference between labels, most notably between Kobe and Osaka. This will let us test the difference between labeling that focuses on consistent label positioning and labeling that tries to maximize legibility at the cost of consistency. The third region is around Kaliningrad, Lithuania, and northern Poland. This region is denser in terms of labels so there is some interference and there are also country borders, which affect the labeling. The fourth region is northern Spain and Portugal. This region is also quite dense and in addition to country borders also deals with province borders.

Differences between the labeling methods can be seen for the whole region of Kaliningrad, Lithuania, and northern Poland in Figure 7.1, and in detail in Figure 7.2. All maps for all labeling methods are available electronically as SVG and JPEG, as a part of Appendix A.

(a) : The whole region around Kaliningrad, Lithuania, and northern Poland labeled in accordance with the *Legibility* labeling method.

(b) : The whole region around Kaliningrad, Lithuania, and northern Poland labeled in accordance with the *Consistency* labeling method.



(c) : The whole region around Kaliningrad, Lithuania, and northern Poland labeled in accordance with the *Google Maps* labeling method.

Figure 7.1: A high-level comparison of the three labeling methods for the Kaliningrad, Lithuania, and northern Poland region. *Map data ©2022 GeoBasis-DE/BKG (©2009), Google.*

(a) : Region around Koszalin (Poland) labeled in accordance with the *Legibility* labeling method.

(b) : Region around Koszalin (Poland) labeled in accordance with the *Consistency* labeling method.

(c) : Region around Koszalin (Poland) labeled in accordance with the *Google Maps* labeling method.

Figure 7.2: A detailed comparison of the three labeling methods for a region around Koszalin (Poland). *Map data ©2022 GeoBasis-DE/BKG (©2009), Google.*

7.3 Counterbalancing Sequential and Order Effects

Sequential and order effects are observable changes in the answers or measurements along the sequence of tasks that are related to the answers from previous tasks or the order of tasks in general [33]. There are two main sequential and order effects that could affect the results of the experiment if not accounted for. The first effect is the so-called *practice effect* (likewise called learning effect). This means that as the participants progress through the experiment, their performance improves as they get more familiar with the task, environment, and even the labeling method [34].

The other effect is the *fatigue effect* (likewise called boredom effect). This means that over the course of the experiment, the participants might start losing focus and in turn their performance might gradually decline or even begin answering randomly [33], which is especially a concern for the 2AFC tasks (see Section 6.2.2).

The fatigue effect might occur due to the repetitive nature of the tasks. However, the fatigue effect should be of less concern due to the nature of the experimental design, though should not be overlooked. The experiment is designed to be relatively short and we provide breaks to the participants between groups of tasks. Either one of the mentioned effects might furthermore affect the subjective scores in our design (see Section 6.2.1).

To compensate for the sequential and order effects in the experiment we use a Balanced Latin Square design. The Balanced Latin Square is a method specifically constructed to counterbalance immediate sequential and other order effects [35]. The method is essentially used to change the order in which each participant completes parts of the experiment. In our experiment, we

use it to change the order of map regions and individual assignment tasks for that region. The tasks for each region are split into two groups, one for each variant of the task and each variant is balanced separately. We also use it for the 2AFC tasks, to change the order of map regions and individual pairs for that region.

For each participant across the number of conditions (e.g. map regions or assignment/2AFC tasks in a sequence), the order of conditions is changed so that the number of conditions preceding a given condition is different and also the condition is immediately preceded by a different condition [35]. An example of the Balanced Latin Square for four conditions can be seen in Table 7.1.

		Condition order			
		1	2	3	4
Participant	1	A	B	D	C
	2	B	C	A	D
	3	C	D	B	A
	4	D	A	C	B

Table 7.1: Balanced Latin Square for four conditions: A, B, C, and D. The fifth participant would be assigned the first row, the sixth participant the second row, etc.

7.4 Experimental Procedure

The experiment was conducted remotely and asynchronously. This means that we were not in direct contact with the participants and they only received text instructions. This approach let us reach more participants and was less time-consuming, than conducting the experiment in a lab environment. The main disadvantage of this approach is that we had less control over the devices the participants use, though we specifically asked the participants to only use a laptop or a desktop computer with a connected mouse.

The experiment was designed to be around 10 minutes long and had a 30-minute time-out. This was in our opinion a good compromise to gather enough data and be short enough to have most participants finish. The experiment starts with a landing page, where the participants are given general instructions and information about the experiment. To begin the testing, participants had to fill in their age and consent to the collection and processing of the data from the experiment.

The testing begins with the assignment tasks (described in Section 6.1). The assignment tasks are split into four groups, each group corresponding to each map region. In each group, the participants first completed a series of four tasks where they had to select a label corresponding to a highlighted point and then a series of four tasks where they had to select a point corresponding to a highlighted label. Before each series of assignment tasks, the participants

were reminded of the instructions and told what would be highlighted and what was their objective to select (i.e. either a point or a label). After each map region, the participants were asked to select how much they agree with statements about their performance, as described in Section 6.2.1. In total, each participant completed 32 assignments ($4\,regions \times (4 + 4\,assignments)$).

After the assignment tasks, the participants moved to aesthetic pairwise comparison. The participants were shown two maps of the same region, each using a different labeling method, and were asked to select the one they liked more. The task is described in more detail in Section 6.2.2. For each region, the participants were shown all three pair combinations of the labeling methods. In total, each participant completed 12 comparisons ($4\,regions \times 3\,pairs$).

Chapter 8

Implementation

We have decided to implement the experiment as an interactive web application. This was a straightforward decision because the experiment was to be conducted remotely and we wanted to make the experiment as easy to complete as possible (i.e. not having to install or download anything). Web applications generally allow good compatibility and are easy to share.

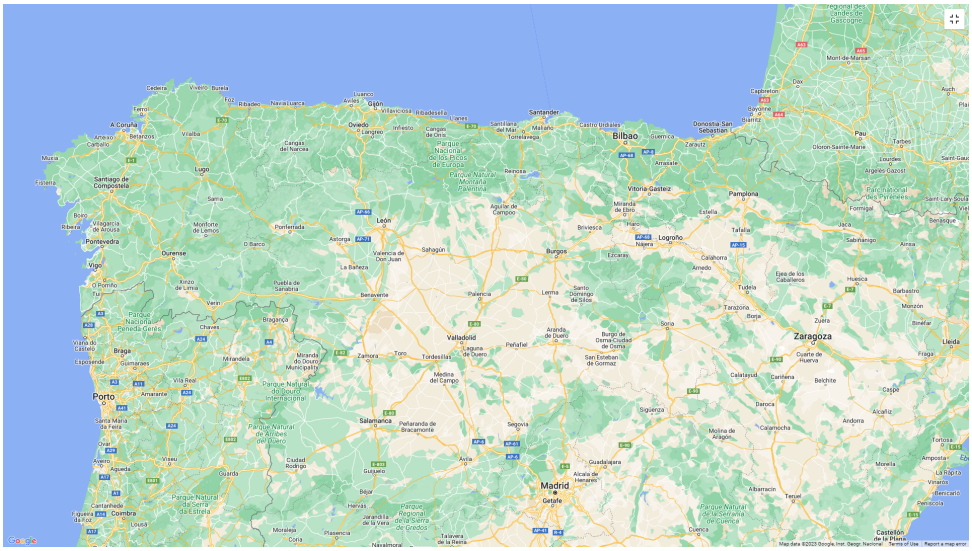
To get consistent and comparable results from the participants, we have only developed the application to be compatible with laptops or desktop computers. We have also urged the participants to only proceed if they have a mouse connected. The experiment is designed and implemented to be displayed exactly the same for all participants. This is important since we need to precisely control the variables of the experiment and even displaying a different typeface could affect the results. Unfortunately, this has caused some compatibility issues and so a small portion of the participants was not able to finish the experiment.

For the data format of the maps, we have decided to use Scalable Vector Graphics (SVG) [36]. SVGs can have multiple layers and each element can have specific attributes, which perfectly satisfies our needs. SVG is a well-supported format that can also be easily edited with graphics editing software and accessed and manipulated via JavaScript [36].

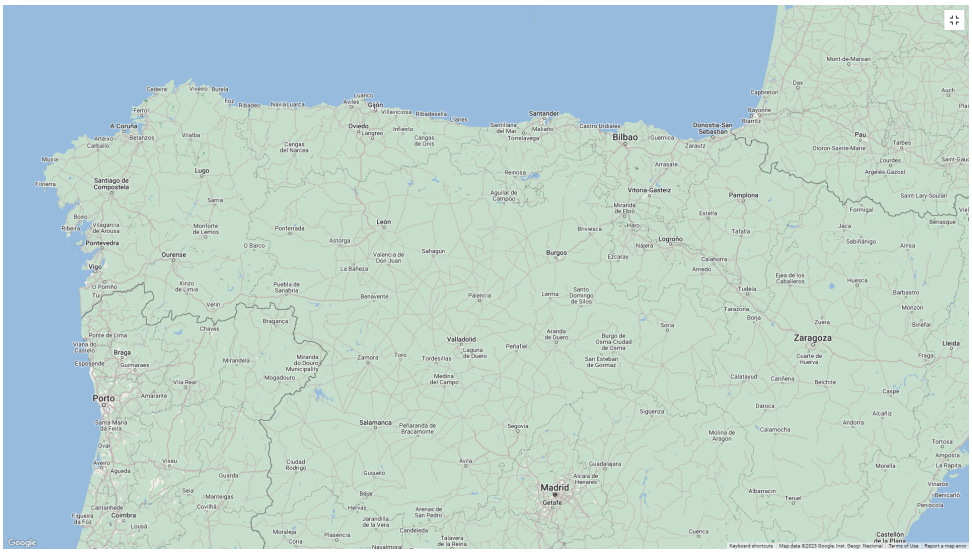
According to Google's official Brand Resource Center [37], we are permitted to take screenshots of Google Maps and use them for the purposes of our experiment without having to obtain a license, as long as the map data is attributed properly. We are also permitted to add custom labels and style the maps with the official *Styling Wizard* available by Google [37].

To make the maps for the experiment, we used the *Styling Wizard* to create a minimalist look with no contouring and only the relevant labels displayed (see Figure 8.1). Our intent was to create maps that look as similar as possible across all regions and to remove potentially distracting features, to have better control over the experiment variables. For each map, we took two screenshots from Google Maps. The first screenshot, with no labels displayed, is used as the base for all three labeling methods. The second screenshot was taken with labels displayed only for the map feature of interest, in our case of cities. We replicated the labels and points from the second screenshot and used them to make all three variants of the specific regions manually.

All maps for all labeling methods are available electronically as SVG and JPEG, as a part of Appendix A. The implementation of the experimental tasks can be seen in Figures 6.1, 6.2, and the source files for the experiment are available electronically as a part of Appendix A.



(a) : Original map style of Google Maps.



(b) : Custom style of Google Maps created for the experiment.

Figure 8.1: A comparison of the original style of Google Maps (a) with the custom one we made for the experiment (b). *Map data ©2023 Google, Instituto Geográfico Nacional.*

For data collection and some statistical processing, we use *Sfinx*. *Sfinx* is a web application for collecting and evaluating data from empiric user experiments [38].

Part III

Evaluation

Chapter 9

Collected Data

The collected data are split into four parts. The parts represent each experimental task and also the data about the participant. We only collected the age of the participants and also noted the labeling method they would be exposed to during the first portion of the experiment (see Table 9.1).

Variable name	Meaning
Age	The age of the participant.
Labeling method	The assigned labeling method for the first portion of the experiment.

Table 9.1: Data collected about the participant.

For the assignment task, we collected a number of variables (see Table 9.2). The main variables are the *completion time* and *error*, but the *labeling method* and *type* are just as important to correctly separate the data before evaluation. The other variables are primarily used for outlier detection.

Variable name	Meaning
Completion time	The completion time of the assignment task in ms.
Expected	The name of the highlighted feature.
Selected	The name of the selected feature.
Error	A binary value indicating if the participant selected a wrong feature.
Labeling method	The labeling method assigned to the participant.
Scenario	The current map.
Type	The variant of the task - either assigning a point to a highlighted label or vice versa.

Table 9.2: Data collected from the assignment task.

The structure of the collected data for subjective evaluation is rather simple (see Table 9.3). We collected the subjective scores regarding confidence in correctness and speed for the previous assignments. The *labeling method* and *type* are used to separate the data before evaluation.

Variable name	Meaning
Confidence correctness	The confidence of the participant in the correctness of the assignment. 5-point Likert scale.
Confidence speed	The confidence of the participant in the speed of the assignment. 5-point Likert scale.
Labeling method	The labeling method assigned to the participant.
Scenario	The current map.
Type	The variant of task - either assigning a point to a highlighted label or vice versa.

Table 9.3: Data collected from the subjective evaluation.

The last part of the experiment was the aesthetic preference evaluation. We used the 2AFC method so the participants were presented with two variants of the same map side by side, each using a different labeling method, and had to choose the one they liked better. This is also represented in the collected variables (see Table 9.4).

Variable name	Meaning
First method	The labeling method used for the map on the left.
Second method	The labeling method used for the map on the right.
Selected method	The preferred labeling method.
Scenario	The current map.

Table 9.4: Data collected from the aesthetic preference 2AFC task.

All collected data, as described in this chapter, are available electronically as a part of Appendix A.

Chapter 10

Data Analysis

The experiment was completed by 108 participants in total. The age of participants ranges from 18 to 73 years (see Figure 10.1). The average age was $\overline{x} = 34.5$. The participants were split into three groups for the assignment task (see Section 6.1) depending on the labeling method they would be exposed to.

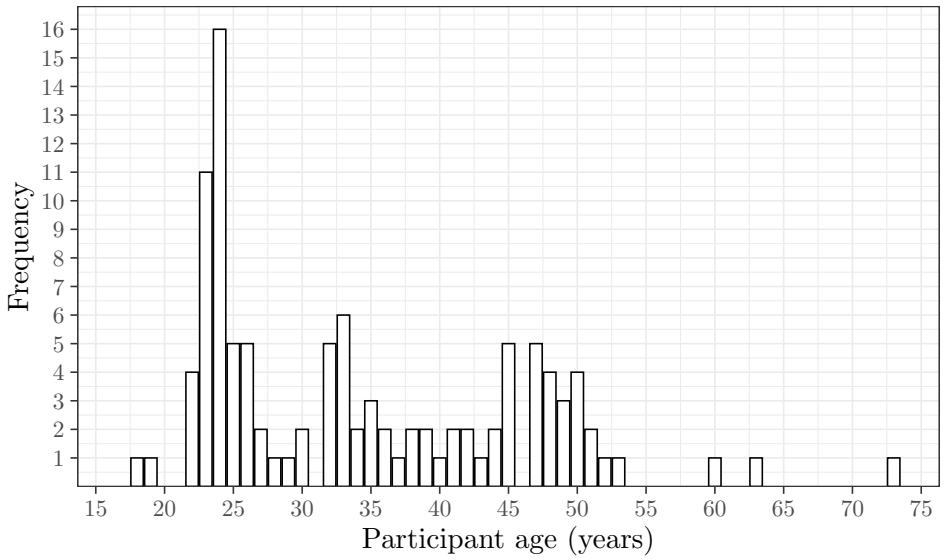


Figure 10.1: Histogram of ages of all participants.

The first group was exposed to the *Legibility* labeling method. This group had 36 participants of ages ranging from 18 to 53 years ($\overline{x} = 35$, $\sigma = 10.75$). The second group, exposed to the *Consistency* labeling method had 36 participants of ages ranging from 19 to 73 years ($\overline{x} = 35.1$, $\sigma = 13.31$). The third group, exposed to the *Google Maps* labeling method had 36 participants of ages ranging from 22 to 60 years ($\overline{x} = 33.2$, $\sigma = 9.94$).

Because the experiment was conducted remotely we looked for participants who might not have taken the experiment seriously. We first looked at the total number of incorrect answers during the assignment task (see Figure 10.2). This has left us with three potential outliers. After analyzing their answers further we concluded that neither of them took the experiment seriously and removed them from the data set.

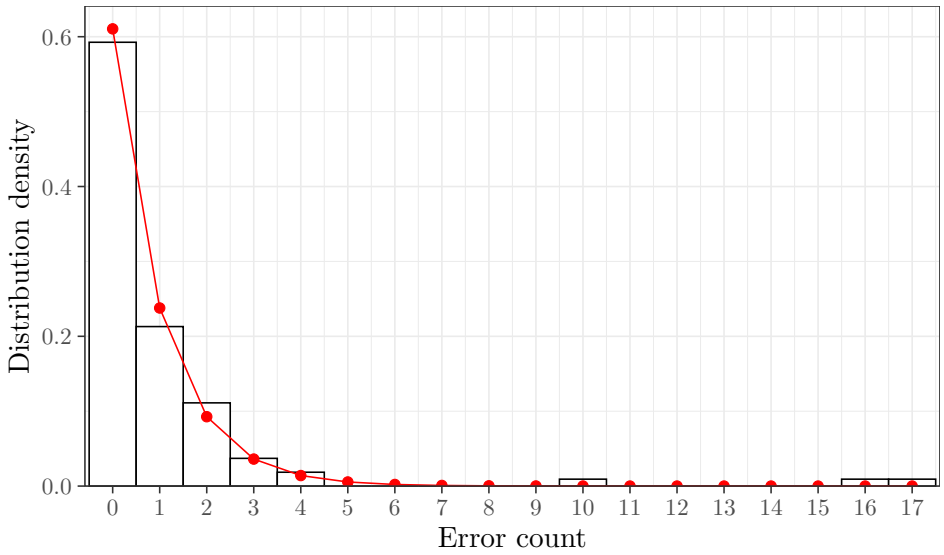


Figure 10.2: Geometric distribution fitted to the total number of errors per participant. We detected three potential outliers with 10, 16, and 17 errors, respectively.

We also looked at the completion times and the completion time per participant seemed to be adequate. We then looked at individual completion times and noticed some outliers on either side of the spectrum and decided to especially look into the low values. We discovered a couple of answers where the completion time was exceptionally short (under 500ms) and the selected answers seemed random. Since the answers were by different participants each time we concluded that they were most likely misclicks and removed them. The answers where the participant might have been distracted and took exceptionally long to answer should not negatively impact the error rate and as such will be kept in the general data set.

10.1 Error Rate

The error rate is an indicator of how many mistakes the participants made during the assignment task. We have calculated a 95% confidence interval for each labeling method, see Figure 10.3. Confidence intervals give us an estimate of the range in which the true value lies (with a given level of confidence) [39] but they do not necessarily tell us if there is a statistically significant difference between two means [40], especially for multiple pairwise comparisons where we

have to control for Type I error of the whole hypothesis family and therefore also change the confidence level for each confidence interval. Consequently, the confidence intervals should only be used as a visual aid (e.g. to tell us the direction of the effect in case of a significant result [39]).

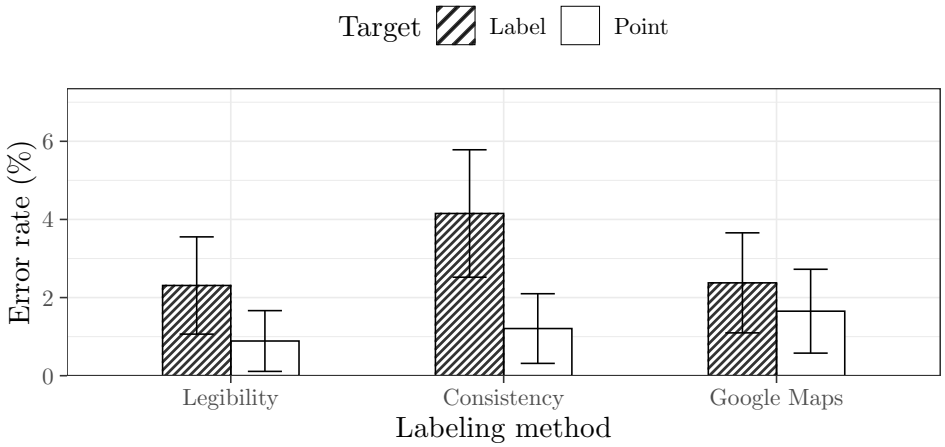


Figure 10.3: 95% confidence intervals of the error rates, split by labeling method.

To test if there is any significant difference between the error rates of our labeling methods, we will use Pearson's chi-squared test of independence for a 3×2 contingency table. Each row in the contingency table refers to a labeling method and the columns to the counts of correct and incorrect answers, respectively. We chose Pearson's chi-squared test because it was shown to give the most accurate results for the test of independence [41].

If the chi-squared test result for our 3×2 table is significant, we will conduct a post-hoc test, to test for differences between each pair of labeling methods. For the post-hoc test, we will use multiple pairwise (2×2 contingency table) Pearson's chi-squared tests with p-value adjustments, to control the family-wise Type I error.

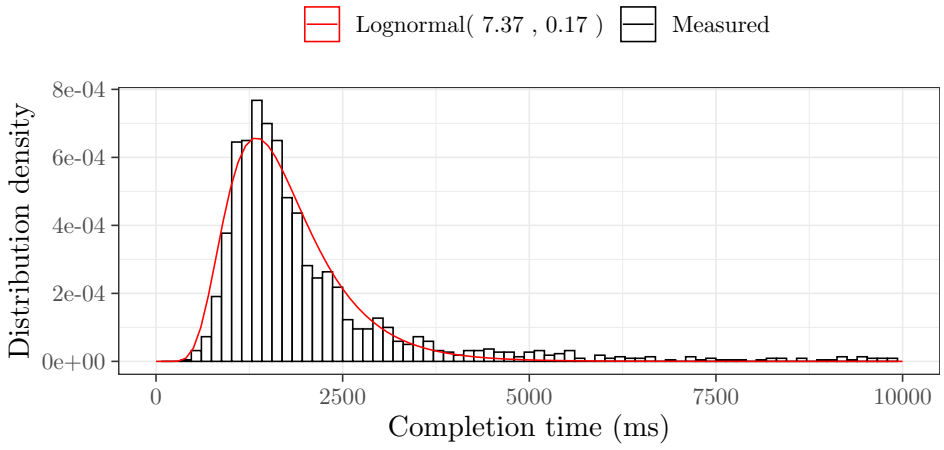
The results in Table 10.1 allow us to conclude that we do not have enough evidence to reject either of the null hypotheses at a significance level $\alpha = 0.05$. This implies that for either variant of the assignment task, we do not have evidence to claim that there is a significant difference in error rates between any of the labeling methods. Therefore, we will not conduct a post-hoc test.

Target	Hypothesis	P-value
Label	H_0^E	0.0828
Point	H_0^E	0.4097

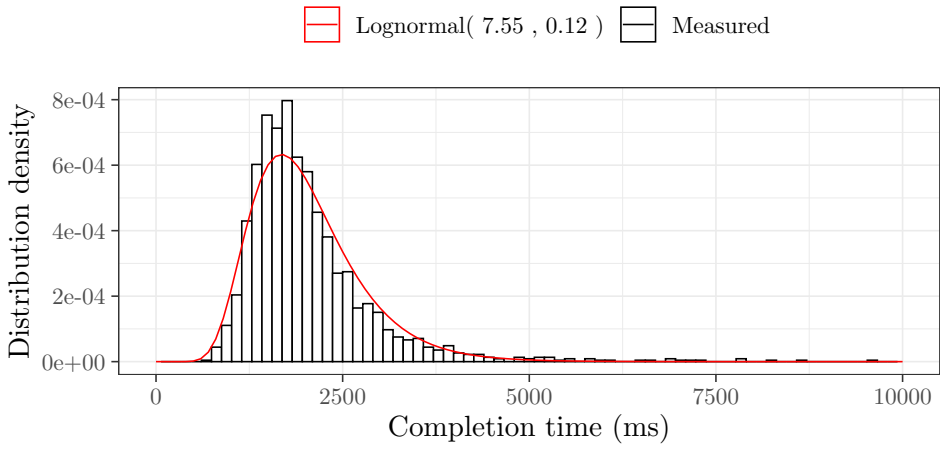
Table 10.1: Calculated p-values for Pearson's chi-squared test, testing for difference in error rates between all labeling methods.

10.2 Completion Time

The completion time is an indicator of how long one assignment took. Since the completion times commonly follow a lognormal distribution [42], as also shown for our data in Figure 10.4, we can log-transform the completion times. Applying log-transformation to this type of data is generally recommended to reduce non-normality and heteroscedasticity [43]. For data such as ours, the log-transformation also increases statistical power and reduces Type I error of ANOVA and two-sample t-tests [44][45].



(a) : Completion times for the first variant of the assignment task (i.e. the participant was asked to select a label corresponding to a highlighted point).



(b) : Completion times for the second variant of the assignment task (i.e. the participant was asked to select a point corresponding to a highlighted label).

Figure 10.4: Log-normal distributions fitted to the measured completion times from the assignment tasks.

The log-transformation also has another benefit for us. The log-transformed data can be further used to compare the geometric means of completion times [42]. The geometric mean of completion times is a very useful measure because it was shown to be the best indicator of the average completion time [43]. To compute the geometric mean of the completion times, we simply compute the mean of the log-transformed data and re-transform the result to the original scale [42][43]. By using ANOVA and two-sample t-tests on the log-transformed data, we are essentially comparing the geometric means of the non-transformed data. Moreover, we have also used the log-transformed data to calculate confidence intervals of the geometric means, see Figure 10.5.

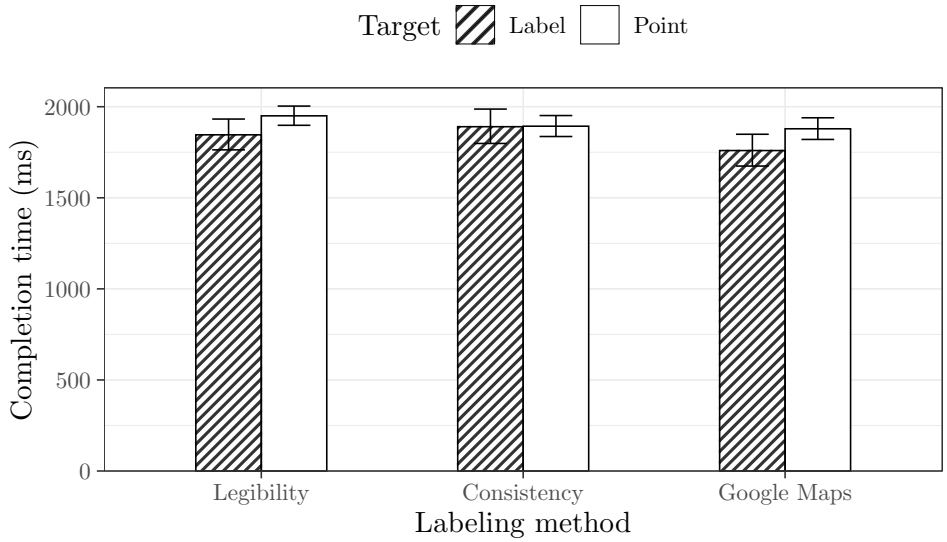
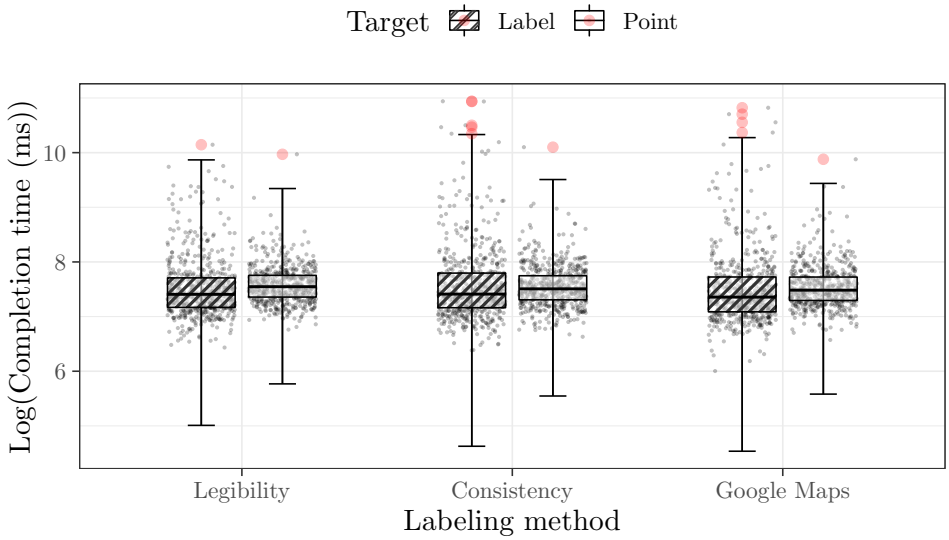
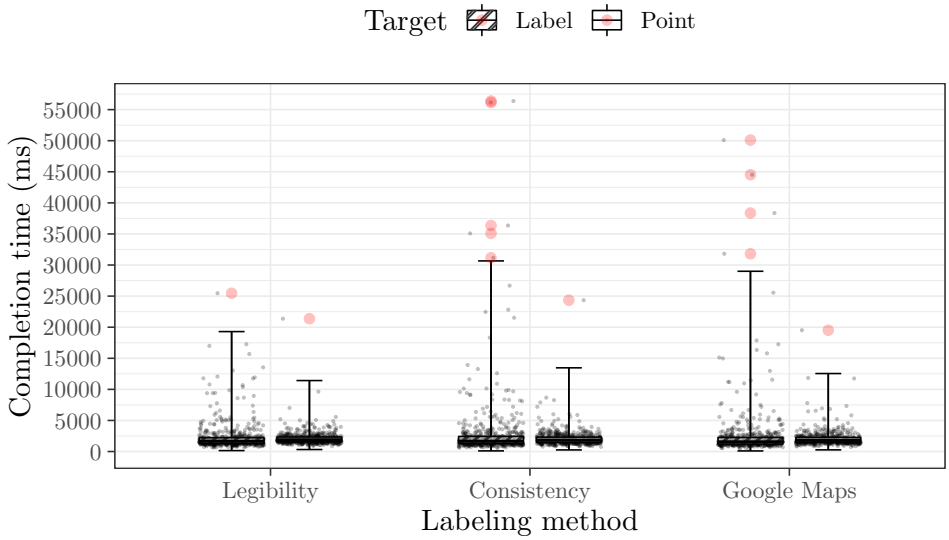


Figure 10.5: 95% confidence intervals of the completion times, split by labeling method. The confidence intervals were calculated from the log-transformed data and transformed back to the original scale. As such, the means in the confidence intervals refer to the geometrical mean of the non-transformed data [42][43].

Before calculating confidence intervals and conducting further analysis, we removed the most extreme outliers where the participants were most likely distracted. To detect potential outliers we used boxplots (see Figure 10.6). The log-transformed distributions are still positively skewed and show positive excess kurtosis (see Figure 10.7 and Table 10.2).



(a) : Boxplots of the log-transformed completion times.



(b) : Re-transformed boxplots from (a).

Figure 10.6: Boxplots of the log-transformed completion times used to detect potential outliers. Whiskers are based on the $4 \times IQR$ value. We used a conservative coefficient $k = 4$ instead of the usual $k = 1.5$ to only detect outliers outside of the naturally occurring positive skew. The boxplots in (b) should only be used as an interpretation of the boxplots calculated from the log-transformed completion times (a), to show the outliers in the context of the measured values.

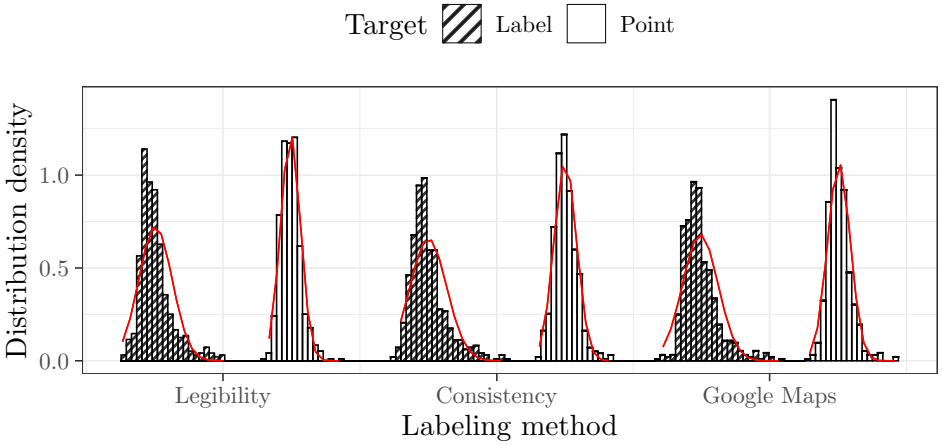


Figure 10.7: Log-transformed distributions for the completion times and fitted normal distributions. This shows that the log-transformed distributions have positive excess kurtosis (leptokurtic distributions) with a positive skew.

Target	Method	μ	σ	Skewness (g_1)	Excess kurtosis (g_2)
Label	Legibility	7.52	0.55	1.45	2.71
	Consistency	7.54	0.61	1.38	2.53
	Google Maps	7.47	0.59	1.42	3.07
Point	Legibility	7.58	0.33	0.72	1.38
	Consistency	7.55	0.37	0.72	1.31
	Google Maps	7.54	0.37	1.09	2.83

Table 10.2: Parameters of the log-transformed distributions. Skewness is calculated as $m_3/m_2^{3/2}$ and excess kurtosis as $(m_4/m_2^2) - 3$, where m_2, m_3 and m_4 are the second, third and fourth sample central moments. Positive excess kurtosis indicates leptokurtic distribution while negative excess kurtosis indicates platykurtic distribution. Excess kurtosis close to zero indicates normal distribution [46].

However, it was shown that ANOVA is quite robust to non-normality, even for distributions with high skewness and excess kurtosis [47][48]. Due to the Central Limit Theorem, we can relax the normality assumption of the t-test [49]. It was also shown that in large samples with heavily skewed data, two-sample t-tests should be even favored over other methods (e.g. Mann-Whitney U test) [50].

To evaluate if there is any difference between the geometric means of each labeling method, we will first use Welch's ANOVA. We decided to use Welch's ANOVA because it shows lower Type I error for non-normal distributions compared to the standard (Fisher's) ANOVA and has similar power [48].

If the ANOVA result is significant (i.e. we reject the hypothesis that all geometric means are the same), we will conduct a post-hoc test to compare the geometric mean differences between each pair of labeling methods.

For the post-hoc set, we settled on multiple Welch's t-tests with p-value adjustments, to control the family-wise Type I error. The Welch's t-test controls Type I error better for groups with unequal variances when compared to other t-tests [51] and as discussed earlier is particularly robust for non-normal distributions.

At a significance level $\alpha = 0.05$, the results in Table 10.3 show that we do not have sufficient evidence to reject either of the null hypotheses. As such, we do not have sufficient evidence to claim that there is a significant difference in completion times between the labeling methods for either variant of the assignment task. Therefore, the post-hoc test is not needed.

Target	Hypothesis	P-value
Label	H_0^T	0.1235
Point	H_0^T	0.164

Table 10.3: Calculated p-values for Welch's ANOVA, testing for difference in geometric means of completion times between all labeling methods.

10.3 Subjective Score Evaluations

To measure the subjective scores, we have used a 5-point agreement Likert scale [25], as described in Section 6.2.1. A rating of 1 (worst) corresponds to option *I disagree* and a rating of 5 (best) to option *I agree*.

There is some debate about what arithmetic operations are appropriate to apply to data from an ordinal scale [52], such as ours. It was originally suggested that certain operations, such as calculating the mean are not appropriate and the data should only be counted [53]. However, the methods used are robust with respect to non-normality and have been shown to provide useful information about the data [52][54], such as detecting significant consistent differences between the groups using ANOVA and two-sample t-tests or computing confidence intervals [52]. We just have to be careful about interpreting the data. We cannot draw any conclusions about the ratio or the interval between group scores [52][54].

To test for differences in the subjective scores of the labeling methods, we will first use Welch's ANOVA. If the results of Welch's ANOVA are significant (i.e. we reject that all labeling methods have the same subjective score), we will conduct a post-hoc test. The post-hoc test will let us infer what labeling methods differ from each other. Specifically, we will test all pairs of labeling methods. For the post-hoc test, we will use multiple Welch's t-tests with p-value adjustments. The advantages of Welch's variants have been discussed in the previous section.

To adjust the p-values we will use the Holm step-down procedures [55] based on the Šidák inequality [56][57]. The Holm step-down procedure controls the family-wise Type I error equally as well as (more conservative) one-step procedures (e.g. Bonferroni correction) but has more power [57].

10.3.1 Correctness Scores

The correctness score corresponds to the responses we received to the statement *I was sure in the assignment*, after completing a group of assignment tasks. As a general indicator of the scores we have calculated 95% confidence intervals for the subjective scores of correctness, see Figure 10.8.

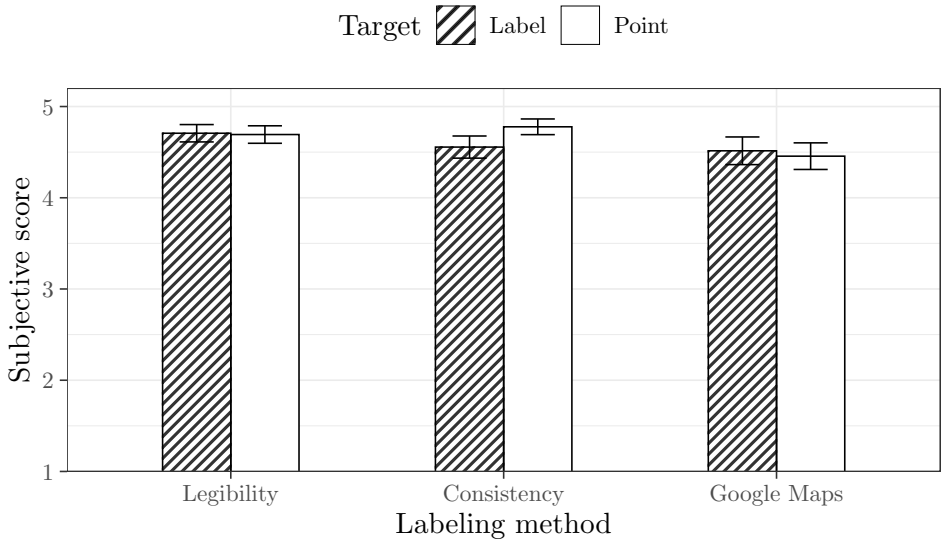


Figure 10.8: 95% confidence intervals of the subjective scores of the correctness of assignment (higher score is better).

At a significance level $\alpha = 0.05$, we can conclude from the results in Table 10.4 that we have sufficient evidence to reject both global hypotheses. This means that there is evidence to support the claim that the subjective scores between the labeling methods differ in both variants of the assignment task. To determine which labeling methods differ from each other, we will conduct a post-hoc test, as described earlier.

Target	Hypothesis	P-value
Label	H_0^C	**0.04558***
Point	H_0^C	**0.00103***

Table 10.4: Calculated p-values for Welch's ANOVA, testing for difference in subjective scores of correctness between all labeling methods. *Null hypothesis rejected at $p < 0.05$.

From the results of the post-hoc test in Table 10.5 we see that for the first variant of the task, we did not find any significant difference for any of the pairwise comparisons. This probably means that the post-hoc test does not have enough power to reject the null hypotheses or that the global test made a Type I error (false positive) [58], possibly due to some violation of Welch's ANOVA assumptions. As such, we will not reject any hypotheses and omit further interpretations.

For the second variant of the task, we can conclude that we have enough evidence to reject two null hypotheses at a significance level $\alpha = 0.05$. There is sufficient evidence to claim that both the *Legibility* and *Consistency* labeling methods are subjectively rated better than the *Google Maps* labeling method.

Target	Hypothesis	Methods	Adjusted p-value
Label	$H_0^{C_1}$	Legibility x Google Maps	0.10374
	$H_0^{C_2}$	Legibility x Consistency	0.10374
	$H_0^{C_3}$	Consistency x Google Maps	0.6778
Point	$H_0^{C_3}$	Consistency x Google Maps	**0.000648***
	$H_0^{C_1}$	Legibility x Google Maps	**0.015508***
	$H_0^{C_2}$	Legibility x Consistency	0.1933

Table 10.5: Calculated p-values for a post-hoc family of hypotheses, testing for difference in subjective score of correctness between all pairs of labeling methods. Welch's t-test was used and p-values were adjusted using the Holm-Šidák step-down procedure[57]. *Null hypothesis rejected at $p < 0.05$.

10.3.2 Speed Scores

The speed score corresponds to the responses we received to the statement *I was fast in the assignment*, after completing a group of assignment tasks. As a general indicator of the scores we have calculated 95% confidence intervals for the subjective scores of speed, see Figure 10.9.

The results in Table 10.6 indicate that we have enough evidence to reject both global hypotheses, at a significance level $\alpha = 0.05$. This means that we have evidence to suggest that there is a difference in the subjective scores between the labeling methods. To determine which labeling methods differ from each other, we will conduct a post-hoc test, as described earlier.

Target	Hypothesis	P-value
Label	H_0^S	**0.00432***
Point	H_0^S	**0.01241***

Table 10.6: Calculated p-values for Welch's ANOVA, testing for difference in subjective scores of the speed between all labeling methods. *Null hypothesis rejected at $p < 0.05$.

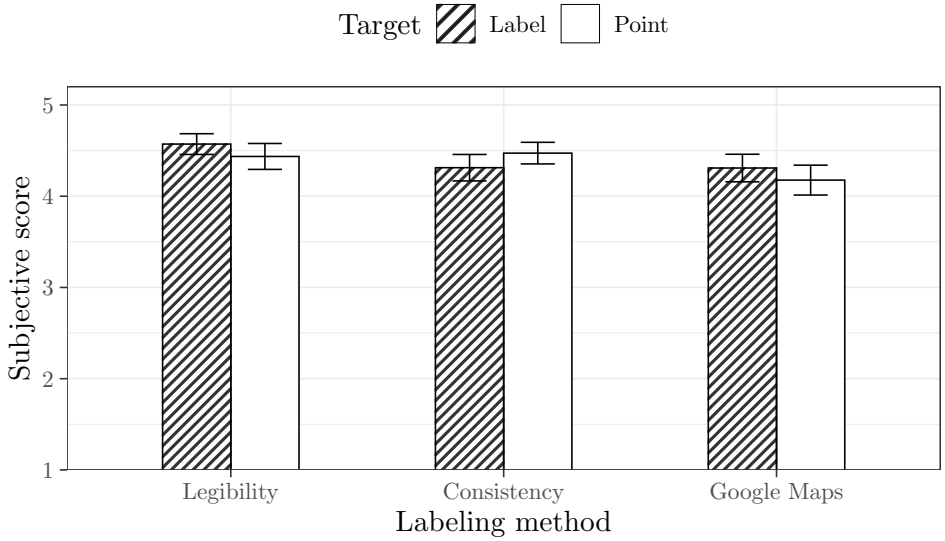


Figure 10.9: 95% confidence intervals of the subjective scores of the speed of assignment (higher score is better).

The results of the post-hoc test (see Table 10.7) show that at a significance level $\alpha = 0.05$ we have strong evidence to reject several null hypotheses. There is sufficient evidence to claim that for the first variant of the assignment task, the *Legibility* labeling method is subjectively rated better than both *Google Maps* and *Consistency* labeling methods. For the second variant both *Legibility* and *Consistency* are rated significantly better than the *Google Maps* labeling method.

Target	Hypothesis	Methods	Adjusted p-value
Label	$H_0^{S_1}$	Legibility x Google Maps	**0.019293***
	$H_0^{S_2}$	Legibility x Consistency	**0.019293***
	$H_0^{S_3}$	Consistency x Google Maps	0.9723
Point	$H_0^{S_3}$	Consistency x Google Maps	**0.012393***
	$H_0^{S_1}$	Legibility x Google Maps	**0.03744***
	$H_0^{S_2}$	Legibility x Consistency	0.6966

Table 10.7: Calculated p-values for a post-hoc family of hypotheses, testing for difference in subjective score of speed between all pairs of labeling methods. Welch's t-test was used and p-values were adjusted using the Holm-Šidák step-down procedure[57]. *Null hypothesis rejected at $p < 0.05$.

10.4 Aesthetic Preference

To measure the aesthetic preference, we used the 2AFC method, as described in Sections 6.2.2, 7.4. We did not find any potential outliers (e.g. participants who always selected the variant on the same side of the screen) and therefore used all collected answers for the evaluation.

10.4.1 Mere-exposure effect

The results of the aesthetic preference testing might have been affected by using a between-subject design for the assignment task. This is known as the mere-exposure effect [59]. The mere-exposure effect means that exposing a participant to a stimulus more frequently will make them like it more [59]. In our case, the participants were always only exposed to one labeling method.

Another case of the mere-exposure effect could be caused by participants using Google Maps regularly. We will disregard that since that would happen even in an isolated experiment and is therefore unavoidable under our conditions.

To test for the mere-exposure effect caused by the between-subject design, we will test to see if there is a significant difference between participants preferring the labeling method they were exposed to previously and the other labeling methods. To test this we fill formulate a family of hypotheses as follows:

$H_0^{M_L}$ There is no difference in preference of *Legibility* compared to other labeling methods among participants who were exposed to *Legibility* in the first part of the experiment.

$H_0^{M_C}$ There is no difference in preference of *Consistency* compared to other labeling methods among participants who were exposed to *Consistency* in the first part of the experiment.

$H_0^{M_G}$ There is no difference in preference of *Google Maps* compared to other labeling methods among participants who were exposed to *Google Maps* in the first part of the experiment.

We will test each hypothesis using the one-tailed exact binomial test ($H_0 : \pi = 0.5$, $H_A : \pi > 0.5$). For each hypothesis, we only take the responses of participants exposed to the labeling method of interest and test the frequency of them preferring that labeling method over the others.

The results in Table 10.8 show that at a significance level $\alpha = 0.05$, we have sufficient evidence to reject each of the three hypotheses. We can conclude that for all three groups, the participants significantly preferred the labeling method they were exposed to previously. Since we have rejected all three null hypotheses, we have enough evidence to claim that the responses were affected by the mere-exposure effect and are therefore biased. As such, we decided to rerun this part of the experiment with different participants.

Hypothesis	P-value
$H_0^{M_C}$	**0.000039***
$H_0^{M_G}$	**0.000755***
$H_0^{M_L}$	**0.00395***

Table 10.8: Calculated p-values for a family of hypotheses testing for the mere-exposure effect - comparing if participants previously exposed to a labeling method also prefer that labeling method aesthetically. *Null hypothesis rejected at $p < 0.05$.

10.4.2 Evaluation

To evaluate the aesthetic preference between the three labeling methods we will use multiple two-tailed exact binomial tests with p-value adjustments ($H_0 : \pi = 0.5$, $H_A : \pi \neq 0.5$). In each binomial test, we will test a pair of labeling methods, as defined in Section 5.2. This means that we will compare the frequencies one labeling method was picked over the other during the 2AFC task.

Because the results of the first experiment were biased due to the mere-exposure effect and did not show any significant difference in aesthetic preference, we decided to conduct a second, isolated experiment, with just the aesthetic preference evaluation. The second experiment was completed by 25 participants, with ages ranging from 22 to 34 years ($\overline{x} = 25.2$, $\sigma = 3.1$), none of whom were deemed to be an outlier. From the results of the second experiment in Table 10.9, we can conclude that at a significance level $\alpha = 0.05$ we do not have enough evidence to reject either of the null hypothesis. This means that we did not find any significant difference in aesthetic preference between any of the labeling methods.

Hypothesis	Methods	Adjusted p-value
$H_0^{A_1}$	Legibility x Google Maps	0.1332
$H_0^{A_3}$	Consistency x Google Maps	0.2713
$H_0^{A_2}$	Legibility x Consistency	0.4841

Table 10.9: Calculated p-values for two-sided binomial tests comparing the frequencies one labeling method was picked over the other during an aesthetic pairwise comparison in the second experiment.

Chapter 11

Results and Discussion

The first experiment was completed by 108 participants. We did not find substantial evidence to support the claim that either of the labeling methods is better in terms of efficiency (i.e. task completion time or error rate). We want to reiterate that the point features for all map regions were selected by the labeling algorithm of Google Maps, as discussed in Section 7.2. The results indicate that in maps with relatively low label density, as is standard for Google Maps, violating some of the principles of Prof. Imhof does not negatively impact usability. However, we have noticed that in areas with a higher label density such as the one in Figure 7.2, labeling methods that prefer label placement consistency often end up creating clusters of labels, which can transfer into higher error rate, as is indicated in Figure 10.3. Even though the error rate difference was not significant for our experiment, we think that studying maps with higher label density might reveal a greater difference between the labeling methods.

Surprisingly, we found statistically significant differences between the subjective scores of confidence. For the variant of the assignment task, where a label is highlighted and the participant is prompted to select the corresponding point, participants exposed to either one of the labeling methods based on the principles of Prof. Imhof (*Legibility* and *Consistency*) reported significantly higher scores of confidence in both the correctness and speed of the assignment, than those exposed to the labeling method of *Google Maps* ($p < 0.05$).

For the other variant of the assignment task, where the participants were prompted to select a label, participants exposed to the *Legibility* labeling method reported significantly higher scores of confidence in the correctness of the assignment than those exposed to either one of the *Consistency* or *Google Maps* labeling methods ($p < 0.05$).

We also have strong evidence to claim that the mere-exposure effect affects aesthetic preference in online maps ($p < 0.05$). As such, we have conducted a second experiment, with only aesthetic preference evaluation. The second experiment was completed by 25 participants. However, we did not find evidence to suggest that there is a significant difference in aesthetic preference between the three labeling methods.

We should also consider that the labeling principles of Prof. Imhof were evidently designed for languages using left-to-right scripts, which is also apparent from the label position preferences. This might negatively impact the usability of maps in languages that use right-to-left scripts. Google Maps should however not be impacted by the script direction, since the preferred label position is centered.

As it stands, we think that adopting the tested general principles of labeling point features by Prof. Imhof is probably not beneficial for global online map services with low map density.

Summary

This thesis aimed to evaluate Prof. Imhof's general principles of labeling point features in an online environment and compare them with the labeling methods of Google Maps, which have their own different principles. We fulfilled this aim to its full extent.

The first partial goal was to carry out an analysis of the labeling methods and the underlying principles of labeling. The analysis has helped us identify the main differences between the labeling methods and set the basis for the experimental design. We mainly focused on the rules of labeling point features, where we found great differences between the labeling methods. We decided to only compare the general placement rules, to limit the number of variables in the experiment.

The second goal was to design an empirical study based on the analysis. We decided to compare the labeling methods in terms of efficiency and aesthetic preference. Therefore, the study is split into two parts. To prevent carry-over effects, the first part of the experiment used a between-subject design. Since our goal was to quantify the differences between the labeling methods and we had to split the participants into disjoint groups, due to the between-subject design, we needed a great number of participants for the experiment. For these reasons, we designed a remote experiment, which allowed us to reach more participants than if we conducted the experiment in a lab. The experiment was completed by 133 participants in total.

The third and last partial goal was to evaluate the data from the experiment. We first conducted an experiment that included both parts of the study, which was completed by 108 participants. We however noticed that the results of the second part were affected by the mere-exposure effect caused by the between-subject design of the first part. As such, we conducted a second experiment that only included the second part of the study (i.e. aesthetic preference evaluation). The second experiment was completed by 25 participants.

All three labeling methods performed equally well in terms of efficiency (i.e. task completion time and error rate) and also aesthetic preference. However, participants exposed to the labeling methods based on the principles of Prof. Imhof subjectively reported significantly higher scores of confidence during the assignment ($p < 0.05$).

As it stands, we did not find justifications to adopt the tested general principles of labeling point features by Prof. Imhof in online map services with relatively low label density. However, we believe that further research into maps with higher label density, other rules for point features, and also for other map features would be beneficial to further validate and improve the labeling methods used in online map services, to further improve the usability of said services.

Bibliography

[1] RYAN, Camille. *American Community Survey, Reports Computer and Internet Use in the United States: 2016* [online]. U.S. Census Bureau, August 8th, 2018. [Accessed 2023-1-1]. Retrieved from: `https://www.census.gov/content/dam/Census/library/publications/2018/acs/ACS-39.pdf`

[2] KRÓL, Karol. Evolution of online mapping: from Web 1.0 to Web 6.0. In: *Geomatics, Landmanagement and Landscape*, No. 1, 2020, pp. 33-51. Retrieved from: doi:10.15576/GLL/2020.1.33

[3] TAYLOR, James and Nicoló DELL'UNTO. Skeuomorphism in Digital Archeological Practice: A Barrier to Progress, or a Vital Cog in the Wheels of Change? In: *Open Archaeology*, Vol. 7, No. 1, 2021, pp. 482-498. Retrieved from: doi:10.1515/opar-2020-0145

[4] PLETCHER, Kenneth. Cartography. In: *Britannica* [online]. [Accessed 2023-1-1]. Retrieved from: `https://www.britannica.com/science/cartography`

[5] COOK, Karen Severud. Labeling of Maps. In: *MONMONIER, Mark. The History of Cartography, Volume 6: Cartography in the Twentieth Century*. Chicago: University of Chicago Press, 2015, pp. 738-743. ISBN 978-0226534695

[6] IMHOF, Eduard. Positioning Names on Maps. In: *The American Cartographer*, 1975, Vol. 2, No. 2, pp. 128-144. Retrieved from: doi:10.1559/152304075784313304

[7] YOELI, Pinhas. The Logic of Automated Map Lettering. In: *The Cartographic Journal*, 1972, Vol. 9, No. 2, pp. 99-108. Retrieved from: doi:10.1179/000870472787352505

[8] CECI, L. Leading mapping apps in the United States in 2022, by downloads. In: *Statistia* [online]. Feb 7, 2023. [Accessed 2023-4-28]. Retrieved from: `https://www.statista.com/statistics/865413/most-popular-us-mapping-apps-ranked-by-audience/`

[9] WOLODTSCHENKO, Alexander and Thomas FORNER. Prehistoric and Early Historic Maps in Europe: Conception of Cd-Atlas. In: *e-Perimetron*, 2007, Vol. 2, No. 2, pp. 114-116. ISSN 1790-3769.

[10] HARLEY, J. B., David WOODWARD, Matthew H. EDNEY, and Jude LEIMER. *The History of Cartography*. Chicago: University of Chicago Press, 1987-2020.

[11] ORMELING, F. J. Eduard Imhof: talentvol en geleerd. In: *Kartografisch Tijdschrift*, 1986, No. 3, pp. 15-18. [Accessed 2023-1-17]. Retrieved from: `https://icaci.org/eduard-imhof-1895-1986/` (Original work: `https://dehollandsecirkel.courant.nu/issue/KGT/1986-07-01/edition/null/page/17`)

[12] NIEDERMANN, Benjamin. *Automatic Label Placement in Maps and Figures: Models, Algorithms and Experiments*. Karlsruhe, 2017. Dissertation Thesis. KIT, Fakultät für Informatik, Institut für Theoretische Informatik. Retrieved from: doi:10.5445/IR/1000068424

[13] CHRISTENSEN, Jon, Joe MARKS, and Stuart SHIEBER. An empirical study of algorithms for point-feature label placement. In: *ACM Transactions on Graphics*, Vol. 14, No. 3, 1995, pp. 203-232. Retrieved from: doi:10.1145/212332.212334

[14] RYLOV, Maxim A. and Andreas W. REIMER. A Comprehensive Multicriteria Model for High Cartographic Quality Point-Feature Label Placement. In: *Cartographica: The International Journal for Geographic Information and Geovisualization*, Vol. 49, No. 1, 2014, pp. 52-68. Retrieved from: doi:10.3138/carto.49.1.2137

[15] ZORASTER, Steven. The Solution of Large 0–1 Integer Programming Problems Encountered in Automated Cartography. In: *Operations Research*, Vol. 38, No. 5, 1990, pp. 752-759. Retrieved from: doi:10.1287/opre.38.5.752

[16] VERNER, Oleg V., Roger L. WAINWRIGHT, and Dale A. SCHOENFELD. Placing Text Labels on Maps and Diagrams using Genetic Algorithms with Masking. In: *INFORMS Journal on Computing*, Vol. 9, No. 3, 1997, pp. 266-275. Retrieved from: doi:10.1287/ijoc.9.3.266

[17] MOTE, Kevin. Fast Point-Feature Label Placement for Dynamic Visualizations. In: *Information Visualization*, Vol. 6, No. 4, 2007, pp. 249-260. Retrieved from: doi:10.1057/palgrave.ivs.9500163

[18] LUBOSCHIK, Martin, Heidrun SCHUMANN, and Hilko CORDS. Particle-based labeling: Fast point-feature labeling without obscuring other visual features. In: *IEEE Transactions on Visualization and Computer Graphics*, Vol. 14, No. 6, 2008, pp. 1237-1244. Retrieved from: doi:10.1109/TVCG.2008.152

[19] PAVLOVEC, Václav and Ladislav ČMOLÍK. Rapid Labels: Point-Feature Labeling on GPU. In: *IEEE Transactions on Visualization and Computer Graphics*, Vol. 28, No. 1, 2022, pp. 604-613. Retrieved from: doi:10.1109/TVCG.2021.3114854

[20] BEEN, Ken, Eli DAICHES, and Chee YAP. Dynamic map labeling. In: *IEEE Transactions on Visualization and Computer Graphics*, Vol. 12, No. 5, 2006, pp. 773-780. Retrieved from: doi:10.1109/TVCG.2006.136

[21] KENT, Alexander J. A Profession Less Ordinary? Reflections on the Life, Death and Resurrection of Cartography. In: *The bulletin of the Society of University Cartographers. Society of University Cartographers*, 2014, Vol. 48, pp. 7-16. [Accessed 2023-1-10]. Retrieved from: `https://www.researchgate.net/publication/282123268_A_Profession_Less_Ordinary_Reflections_on_the_Life_Death_and_Resurrection_of_Cartography`

[22] CYBULSKI, Paweł and Tymoteusz HORBIŃSKI. User Experience in Using Graphical User Interfaces of Web Maps. In: *ISPRS Int. J. Geo-Inf.*, Vol. 9, No. 7, 2020, Article 412. Retrieved from: doi:10.3390/ijgi9070412

[23] KUROSU, Masaaki and Kaori KASHIMURA. Apparent usability vs. inherent usability experimental analysis on the determinants of the apparent usability. In: *Mosaic of Creativity: CHI '95 Conference Proceedings, Conference on Human Factors in Computing Systems*, 1995, pp. 292-293. Retrieved from: doi:10.1145/223355.223680

[24] SONDEREGGER, Andreas and Jürgen SAUER. The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. In: *Applied Ergonomics* , 2010, Vol. 41, No. 3, pp. 403-410. Retrieved from: doi:10.1016/j.apergo.2009.09.002.

[25] LIKERT, Rensis. A Technique for the Measurement of Attitudes. In: *Archives of Psychology*, Vol. 22, No. 140, 1932, pp. 1-55.

[26] FECHNER, Gustav Theodor. *Elements of Psychophysics, Vol. 1.* Holt, Rinehart and Winston: New York, 1966. (Original work published in 1860)

[27] PALMER, Stephen E., Karen B. SCHLOSS, and Jonathan SAM-MARTINO. Visual Aesthetics and Human Preference. In: *Annual Review of Psychology*, Vol. 64, 2013, pp. 77-107. Retrieved from: doi:10.1146/annurev-psych-120710-100504

[28] SO, Chaehan. Measuring Aesthetic Preferences of Neural Style Transfer: More Precision With the Two-Alternative-Forced-Choice Task. In: *International Journal of Human-Computer Interaction*, Vol. 39, No. 4, 2023, pp. 755-775. Retrieved from: doi:10.1080/10447318.2022.2049081

[29] YANG, Qian and May L. NG. Paired Comparison/Directional Difference Test/2-Alternative Forced Choice (2-AFC) Test, Simple Difference Test/Same-Different Test. In: *Discrimination Testing in Sensory Science*, 2017, pp. 109-134. Retrieved from: doi:10.1016/B978-0-08-101009-9.00005-8

[30] CHARNESS, Gary, Uri GNEEZY, and Michael A. KUHN. Experimental methods: Between-subject and within-subject design. In: *Journal of Economic Behavior & Organization*, Vol. 81, No. 1, 2012, pp. 1-8. Retrieved from: doi:10.1016/j.jebo.2011.08.009

[31] American Psychological Association. Carryover effect. In: *APA dictionary of psychology* [online]. [Accessed 2023-4-29]. Retrieved from: `https://dictionary.apa.org/carryover-effects`

[32] KOST, Rhonda G. and Joel CORREA DA ROSA. Impact of survey length and compensation on validity, reliability, and sample characteristics for Ultrashort-, Short-, and Long-Research Participant Perception Surveys. In: *Journal of Clinical and Translational Science*, Vol. 2, No. 1, 2018, pp. 31-37. Retrieved from: doi:10.1017/cts.2018.18

[33] DAY, Brett, Ian J. BATEMAN, Richard T. CARSON, Diane DUPONT, Jordan J. LOUVIERE, Sanae MORIMOTO, Riccardo SCARPA, Paul WANG. Ordering effects and choice set awareness in repeat-response stated preference studies. In: *Journal of Environmental Economics and Management*, Vol. 63, No. 1, 2012, pp. 73-91. Retrived from: doi:10.1016/j.jeem.2011.09.001

[34] MacKENZIE, Scott I. Designing HCI Experiments: Order effects, counterbalancing, and latin squares. *Human-Computer Interaction: An Empirical Research Perspective*. Morgan Kaufmann, 2013. ISBN 978-0-12-405865-1. Retrieved from: doi:10.1016/C2012-0-02819-0

[35] BRADLEY, James V. Complete Counterbalancing of Immediate Sequential Effects in a Latin Square Design. In: *Journal of the American Statistical Association*, Vol. 53, No. 282, 1958, pp. 525-528. Retrieved from: doi:10.1080/01621459.1958.10501456

[36] Scalable Vector Graphics (SVG) 2. *W3C Candidate Recommendation 04 October 2018* [online]. W3C, 2018. [Accessed 2023-4-29]. Retrieved from: `https://www.w3.org/TR/SVG/`

[37] Products and Services - Geo Guidelines. In: *Brand Resource Center* [online]. Google LLC. [Accessed 2023-2-14]. Retrieved from: `https://about.google/brand-resource-center/products-and-services/geo-guidelines`

[38] LEBEDEVA, Antonina. *Web application collecting and evaluating data from user experiments*. Prague, 2018. Master's thesis. CTU in Prague, Faculty of Electrical Engineering, Department of Computer Science.

[39] DU PREL, Jean-Baptist, Gerhard HOMMEL, Bernd RÖHRIG, and Maria BLETTNER. Confidence Interval or P-Value? In: *Deutsches Ärzteblatt International*, Vol. 106, No. 19, 2009, pp. 335-339. Retrieved from: doi:10.3238/arztebl.2009.0335

[40] JULIOUS, Steven A. Using confidence intervals around individual means to assess statistical significance between two means. In: *Pharmaceutical Statistics*, Vol. 3, No. 3, 2004, pp. 217-222. Retrieved from: doi:10.1002/pst.126

[41] GARCÍA-PÉREZ, Miguel A. and Vicente A. NÚÑEZ-ANTÓN. Accuracy of Power-Divergence Statistics for Testing Independence and Homogeneity in Two-Way Contingency Tables. In: *Communications in Statistics - Simulation and Computation*, Vol. 38, No. 03, 2009, pp. 503-512. Retrieved from: doi:10.1080/03610910802538351

[42] RUMMEL, Bernard. Probability Plotting: A Tool for Analyzing Task Completion Times. In: *Journal of Usability Studies*, Vol. 9, No. 4, 2014, pp. 152-172.

[43] SAURO, Jeff and James R. LEWIS. Average Task Times in Usability Tests: What to Report? In: *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 2347-2350. Retrieved from: doi:10.1145/1753326.1753679

[44] FAYERS, Peter. Alphas, betas and skewy distributions: two ways of getting the wrong answer. In: *Advances in Health Sciences Education*, Vol. 16, No. 3, 2011, pp. 291–296. Retrieved from: doi:10.1007/s10459-011-9283-6

[45] MYERS, Jerome L. *Fundamentals of Experimental Design*. Boston: Allyn & Bacon, 1972. ISBN 0205033350.

[46] DECARLO, Lawrence T. On the meaning and use of kurtosis. In: *Psychological Methods*, Vol. 2, No. 3, 1997, pp. 292-307. Retrieved from: doi:10.1037/1082-989X.2.3.292

[47] SCHMIDER, Emanuel, Matthias ZIEGLER, Erik DANAY, Luzi BEYER, and Markus BÜHNER. Is it really robust? Reinvestigating the robustness of ANOVA against the normal distribution assumption. In: *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, Vol. 6, No. 4., 2010, pp. 147-151. Retrieved from: doi:10.1027/1614-2241/a000016.

[48] DELACRE, Marie, Christophe LEYS, Youri L. MORA, and Daniël LAKENS. Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's F-test instead of the Classical F-test in One-Way ANOVA. In: *International Review of Social Psychology*, Vol. 32, No. 1, 2019, Article 13. Retrieved from: doi:10.5334/irsp.198

[49] LUMLEY, Thomas, Paula DIEHR, Scott EMERSON, and Lu CHEN. The importance of the normality assumption in large public health data sets. In: *Annual Review of Public Health*, Vol. 23, 2002, pp. 151-169. Retrieved from: doi:10.1146/annurev.publhealth.23.100901.140546

[50] FAGERLAND, Morten W. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? In: *BMC Med Res Methodol*, Vol. 12, No. 78, 2012. Retrieved from: doi:10.1186/1471-2288-12-78

[51] DELACRE, Marie, Daniël LAKENS, and Christophe LEYS. Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. In: *International Review of Social Psychology*, Vol. 30, No. 1, 2017, pp. 92-101. Retrieved from: doi:10.5334/irsp.82

[52] SAURO, Jeff and James R. LEWIS. Six Enduring Controversies in Measurement and Statistics: Is it Okay to Average Data from Multipoint Scales? *Quantifying the User Experience*. Morgan Kaufmann, 2012, pp. 242-246. ISBN 978-0-12-384968-7. Retrieved from: doi:10.1016/C2010-0-65192-3

[53] STEVENS, S. S. On the theory of scales of measurement. In: *Science*, Vol. 103, No. 2684, pp. 677-680. Retrieved from: doi:10.1126/science.103.2684.677

[54] NORMAN, Geoff. Likert scales, levels of measurement and the "laws" of statistics. In: *Advances in Health Sciences Education*, Vol. 15, No. 5, 2010, pp. 625-632. Retrieved from: doi:10.1007/s10459-010-9222-y

[55] HOLM, Sture. A Simple Sequentially Rejective Multiple Test Procedure. In: *Scandinavian Journal of Statistics*, Vol. 6, No. 2, 1979, pp. 65-70. Retrieved from: `https://www.jstor.org/stable/4615733`

[56] ŠIDÁK, Zbyněk. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. In: *Journal of the American Statistical Association*, Vol. 62, No. 318, 1967, pp. 626-633. Retrieved from: doi:10.1080/01621459.1967.10482935

[57] LUDBROOK, John. MULTIPLE COMPARISON PROCEDURES UPDATED. In: *Clinical and Experimental Pharmacology and Physiology*, Vol. 25, No. 12, 1998, pp. 1032-1037. Retrieved from: doi:10.1111/j.1440-1681.1998.tb02179.x

[58] CHEN, Tian, Manfei XU, Justin TU, Hongyue WANG, and Xiaohui NIU. Relationship between Omnibus and Post-hoc Tests: An Investigation of performance of the F test in ANOVA. In: *Shanghai Archives of Psychiatry*, Vol. 30, No. 1, 2018, pp. 60-64. Retrieve from: doi:10.11919/j.issn.1002-0829.218014

[59] ZAJONC, R. B. Attitudinal effects of mere exposure. In: *Journal of Personality and Social Psychology*, Vol. 9, No. 2, Pt. 2, 1968, pp. 1-27. Retrieved from: doi:10.1037/h0025848

Appendices

Appendix A

Electronically Submitted Files

The file structure of the electronically submitted files is described below:

- `maps` ..... Full-scale maps of all regions
  - `china`
    - `legibility.svg`
    - `legibility.jpg`
    - `consistency.svg`
    - `consistency.jpg`
    - `google.svg`
    - `google.jpg`
  - `japan`
    - `...`
  - `kaliningrad`
    - `...`
  - `spain`
    - `...`
- `source_files` ..... Source files of the experiments
  - `experiment1`
    - `...`
  - `experiment2`
    - `...`
- `experiment_data` ..... Collected data from both experiments
  - `2AFC.txt`
  - `2AFC_2.txt`
  - `assignment.txt`
  - `participants.txt`
  - `subjective.txt`