2015•2016
FACULTEIT WETENSCHAPPEN
master in de informatica

Masterproef
"Lying with maps": A Framework to Create Maps with a Subtext

Promotor:
Prof. dr. Johannes SCHOENING

Copromotor:
De heer Nick MICHIELS

Stein Smeets
Scriptie ingediend tot het behalen van de graad van master in de informatica
Masterproef
"Lying with maps": A Framework to Create Maps with a Subtext

Promotor:
Prof. dr. Johannes Schoening

Copromotor:
De heer Nick Michiels

Stein Smeets
Scriptie ingediend tot het behalen van de graad van master in de informatica
# Contents

Abstract .................................................. 3

Dutch Summary .......................................... 3

Acknowledgements ....................................... 19

1 Introduction ........................................... 21
   1.1 History of Maps .................................... 21
   1.2 Lying With Maps ................................... 24
   1.3 Contribution ....................................... 30
   1.4 Structure and Conventions of the Thesis .......... 30

2 Related Work ........................................... 33
   2.1 Maps ................................................ 33
      2.1.1 Map Types .................................... 33
      2.1.2 Map Elements .................................. 35
      2.1.3 Data Classification ............................ 36
      2.1.4 Color Schemes .................................. 38
   2.2 Perceiving and Generating Maps .................... 39
      2.2.1 Perception ..................................... 39
      2.2.2 Generating Maps ............................... 41
   2.3 Machine Learning ................................... 42
      2.3.1 Data Gathering ................................. 42
      2.3.2 Machine Learning Techniques ................. 43
Abstract

Maps are typically used for communication and are believed to present data accurately. Therefore, people have high trust in the message maps communicate and often use them to empower a statement. Because of this trust, maps are very powerful communication tools. Nevertheless, all maps are still an abstraction and generalisation of the real world. In general, maps reside in 1 of 2 categories. Reference maps are maps which are used to communicate locations. They refer to the current position of the reader and indicate where to go from there. Reference maps answer questions like "where am I" or "where should I go". Thematic maps help the reader to understand the distribution of a phenomenon e.g. the distribution of the population of Europe. Within the field of cartography several design principles have been established over the last centuries to create “good” maps. As maps have a very high persuasion factor, this thesis explores the creation of thematic maps to deliberately mislead people. This is achieved by using cartographic principles to their extremes, but without manipulating the underlying data. Within this thesis the focus is put on choropleth maps. Choropleth maps are a subtype of thematic maps on which enumeration units are colored. The cartographic design principles explored within this thesis are data classification and color schemes. To explore these principles several experiments have been conducted to investigate influencing factors. The experiments focus on whether the communication message of a map is perceived differently by people when the explored design principles change. The experiments show that for some cartographic principles, the communication message of a map can be changed.
Dutch Summary

Kaarten zijn een vorm van communicatiemedia die door mensen gezien wordt als vertrouwelijk en accuraat. Hierdoor zijn ze een belangrijk communicatiemiddel. Maar, kaarten blijven een abstracte weergave van de echte wereld. In het algemeen zijn er twee soorten kaarten: referentiekaarten communiceren locatie en bestemming; thematische kaarten geven de verspreiding van een fenomeen weer zoals het aantal inwoners per land. Omdat kaarten zo overtuigend zijn, is er steeds de drang om kaarten te ontwikkelen die mensen misleiden. In deze thesis ligt de focus op choropleth kaarten. De invloed van het veranderen van kleurenschema's en classificatietechnieken wordt onderzocht.

Inleiding

Kaarten bestaan al meer dan 6000 jaar [19]. De oudst gevonden kaart is een muurschildering die een stad toont samen met omliggende natuurelementen. De eerste kaarten werden gebruikt voor het organiseren van de jacht en plaatselijke bijeenkomsten. Over de jaren heen veranderde de functie van de kaart. Om nieuwe gebieden te ontdekken werden steeds nieuwe kaarten ontwikkeld waarop de wereld werd weergegeven zoals ze gekend was. Figuur 1 toont enkele voorbeelden van pogingen om de wereld in kaart te brengen. MacEachren [5] heeft de rollen en functies van kaarten onderzocht en samengevat in een model. Het model laat ons toe om de functie of rol van een kaart te plaatsen in een driedimensionale omgeving. Figuur 2 toont het model. De eerste dimensie beschrijft of het om een publieke kaart of een kaart voor eigen gebruik gaat.
Een tweede dimensie beschrijft of de getoonde data al dan niet gekend zijn door de lezer. De laatste dimensie geeft aan of er veel of weinig interactie met de kaart nodig is om informatie te kunnen verkrijgen. Zo plaats hij een kaart die gebruikt wordt tijdens een presentatie in de driedimensionale ruimte als een publieke kaart met gekende data en weinig interactie. Het model toont aan dat kaarten gebruikt worden voor verschillende doeleinden.

Toch hebben alle voorgaande kaarten een gemeenschappelijk doel, namelijk het communiceren van informatie. Samen met het feit dat ze als vertrouwelijk en accuraat aangenomen worden, maakt dat het altijd verleidelijk geweest is ze te gebruiken om mensen in een bepaalde gedachtegang te sturen.

**Lieten met Kaarten**

Monmonier [7] gebruikt voorbeelden in zijn boek die aangeven dat kaarten kunnen gebruikt worden om mensen te misleiden. Het is vaak niet duidelijk of de maker van de misleidende kaart dit al dan niet met opzet doet. Soms
kan het verkeerd gebruiken van cartografische principes leiden tot misleidende kaarten. Figuur 3 toont een voorbeeld waar de classificatietechniek met opzet veranderd is, om te laten blijken dat de situatie van hoeveelheid huishoudens zonder telefoon minder ernstig is. Het contrast tussen de linkse en rechtse kaart is duidelijk zichtbaar. Op een vergelijkbare manier werden in het verleden ook kaarten gemaakt om toeristen aan te trekken of overheidsinstellingen te overtuigen (Figuur 1.4, 1.5).

Ook vandaag verschijnen nog steeds misleidende kaarten in kranten en artikels. Zo publiceerde de krant 'Het Laatste Nieuws' een artikel over de slimste steden van Vlaanderen (Figuur 4). Terwijl in de tabel duidelijk te zien is dat Hasselt het hoogst scoort, wekt de kaart een ander idee op. Dit is onder andere te wijten aan het feit dat Hasselt niet werd ingekleurd op de kaart, waardoor ze, mede omwille van haar oppervlakte, naar achteren gedrukt wordt. Een alternatief is een proportionele symbolenkaart in plaats van een choropleth kaart. De oppervlakte van Hasselt zelf is dan veel minder van belang. Het gebeurt ook dat leugens opzettelijk toegevoegd worden op kaarten. Een voorbeeld hiervan zijn de zogenaamde 'paper towns'. Dit zijn steden die enkel bestaan op een kaart, maar niet in de echte wereld. Dit soort leugens wordt toegevoegd door bijvoorbeeld Google om het copyright-recht van de kaart te bewaren. Wanneer deze steden opduiken in externe kaarten, weet Google dat hun kaarten gebruikt werden.
Figure 3: Het opzettelijk veranderen van de classificatie techniek om een situatie beter te doen lijken.

Figure 4: Weergave van de slimste steden van Vlaanderen volgens Agoria [35].
Toedracht van de Thesis

Kaarten zijn overal te vinden in de omgeving van de mens. Heel wat onderzoek geeft aan dat mensen kaarten vertrouwen en dat ze gebruikt kunnen worden om de gedachtegang van een persoon te sturen. Deze thesis bouwt verder op deze onderzoeken door een applicatie te ontwikkelen die automatisch een kaart met een boodschap genereert. Hier omtrent werden enkele experimenten uitgevoerd, om het effect van veranderende cartografische principes te onderzoeken.

De applicatie die ontwikkeld werd in deze thesis noemt MapTales. Deze naam is een samensmelting van 'Map’ en 'Tales’. De naam vertegenwoordigt het idee kaarten te genereren die een bepaald verhaal of boodschap willen overbrengen.

Gerelateerd Onderzoek

Kaarten

Kaarten kunnen onderverdeeld worden in twee grote types: referentie kaarten leggen de focus op locatie en oriëntatie; thematische kaarten tonen de verspreiding van één of meerdere fenomenen. Thematische kaarten kunnen verder onderverdeeld worden in een choropleth kaart, proportionele symbolen kaart en puntdichtheid kaarten. Een voorbeeld van deze subtypes wordt gegeven in Figuur 5. De focus binnen deze thesis ligt op choropleth kaarten. Choropleth kaarten geven een kleur aan een gebied, bijvoorbeeld een land of provincie, op de kaart. Welke kleur gegeven wordt aan een bepaald gebied, is afhankelijk van de overeenkomstige waarde van het gebied. Om de kleur van een gebied te bepalen zijn er twee mogelijkheden binnen choropleth maps: of de data wordt in klassen verdeeld (classed) of dit wordt niet gedaan (unclassed). In het tweede geval krijgt elke waarde binnen een bepaald waardebereik een bepaalde kleur toegewezen. Wanneer de data in klassen worden opgedeeld, krijgt elke klasse een kleur toegewezen. Figuur 6 toont deze twee soorten
choropleth maps. Een gebied op de kaart krijgt de kleur van de klasse waar ze toe behoort. Er zijn verschillende technieken mogelijk om data in klassen te verdelen. De vier meest gebruikte technieken zijn: equal interval, quantiles, standard deviation en natural breaks.

Zoals te zien is in Figuur 7, zijn meerdere kleurenschema’s mogelijk. Deze kleurenschema’s zijn opgedeeld in: qualitative, diverging en sequential schema’s. Om te bepalen welk kleurenschema best gebruikt wordt, wordt naar de data zelf gekeken. Wanneer het gaat over kwalitatieve data (data die verschillende fenomenen weergeven die niet rechtstreeks aan elkaar gelinkt zijn) zijn qualitative kleurenschema’s (Figuur 7) de beste keuze. Wanneer het gaat om kwantitatieve data (data die de verspreiding van een fenomeen beschrijven, zoals aantal inwoners) dan kiest de maker van de kaart best een kleurenschema uit de diverging of sequential kleurenschema’s.
Machine Learning

Onderzoek in machine learning stelt de vraag hoe een computerprogramma zelfstandig kan verbeteren door te leren uit ervaring. Om aan machine learning te doen, zijn in eerste plaats training data nodig. De training data kunnen op verschillende manieren verzameld worden. Ofwel worden deze op voorhand verzameld door een persoon, of de training data worden door het computer systeem verzameld aan de hand van menselijke input die geregistreerd wordt. Een record in de training data bevat steeds een reeks variabelen, waaronder target- en predictorvariabelen. Een machine learning-algoritme zal steeds proberen een target te voorspellen aan de hand van de predictors. Er zijn verschillende technieken die hiervoor gebruikt kunnen worden. Een eerste techniek is *Decision Tree Learning*. Dit algoritme stelt een boomstructuur op
waar de knooppunten overeenkomen met de predictors en de bladeren met de targets. Door de boom te doorlopen kan een voorspelling gemaakt worden. Uit de boom in Figuur 9 kan volgende voorspelling afgeleid worden: Als het regent en de wind is zwak dan speelt men tennis.

Een tweede techniek maakt gebruik van de Bayes theorie. Bayes rule is een algoritme dat aan de hand van kansberekeningen bepaalt welk target voorspeld zal worden. Om deze berekening te kunnen doen, stelt het algoritme een zogenaamde distributietabel op. Figuur 8 toont een distributietabel voor een dataset die fatale auto-ongelukken in België beschrijft.

De tabel bevat voor elke mogelijke predictorcombinatie een distributiepercentage. Wanneer deze tabel berekend is, kan een joint distributionberekening gebruikt worden om een kans te berekenen. Bijvoorbeeld, wat is de kans dat je een fataal auto-ongeluk krijgt in België wanneer je een vrouw van vijftwintig
jaar bent. Formule 1 berekent de kans.

\[ P(a|b) = \frac{P(a \land b)}{P(b)} \]  
\[ = \frac{0.02}{0.27} \]  
\[ = 0.074 \]

Volgens Mitchell [22] is er geen betere manier om een voorspelling te doen. Maar bij grote trainingsets met veel predictors kan deze tabel al snel zeer groot worden. De berekening van een kans wordt dan al snel zeer zwaar voor een computer. Daarnaast kan het voorkomen dat trainingdata onvolledig zijn, omdat er nog geen voorbeeld van een bepaalde combinatie bestaat. In dit geval zal de berekening falen. De Bayes rule lost dit laatste probleem op door denkbeeldige voorbeelden toe te voegen aan de dataset. Dit is minstens één voorbeeld voor elke mogelijke situatie. Maar dit neemt niet weg dat voor grote datasets zware berekeningen dienen gemaakt te worden. De oplossing hiervoor is het Naïve Bayes algoritme. Dit algoritme maakt de aanname dat alle predictors conditioneel onafhankelijk zijn van elkaar. Door deze aanname moet het algoritme enkel het distributiepercentage berekenen van iedere predictor apart, de combinaties moeten niet meer gemaakt worden in de distributeitabell. Met andere woorden, er moet slechts eenmalig berekend worden hoe groot de kans is dat een predictor voorkomt in de dataset wanneer het target positief is. Formule 2 kan dan gebruikt worden om de kans voor een bepaalde set predictorwaarden te berekenen.

\[ P(Y = y_k|X_1, ..., X_n) = \frac{P(Y = y_k) \prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i|Y = y_j)} \]  

Een andere manier om aan machine learning te doen, is gebruik maken van sequential covering of separate and conquer technieken. Deze technieken analyseren de training dataset en leren beschrijvende regels. Regels kunnen van algemeen naar specifiek geleerd worden of andersom. Wanneer een regel geleerd is, worden alle trainingdata, die beschreven worden door de regel uit de training
dataset, verwijderd. Uit de overgebleven data wordt een nieuwe regel geleerd. Op deze manier kunnen uit een training dataset regels opgesteld worden.

MapTales: Een Tool om Kaarten met een Subtext te Genereren.

Binnen deze thesis werd een tool ontwikkeld die automatisch een kaart genereert met een subtext. Figuur 10 toont een overzicht van de componenten van de tool. MapTales bestaat uit twee grote delen. Een user interface (UI) enerzijds en een machine learning component anderzijds (1 2 3). Om de werking van de tool te verduidelijken, wordt volgend scenario gebruikt. Een gebruiker wil een kaart genereren van Europa waarop het aantal inwoners van Duitsland zo laag mogelijk lijkt.

De toolgebruiker gebruikt het menu in Figuur 10 (UI) om het process te starten. De data, het land en de subtext worden ingevuld. Binnen het voorgestelde scenario zijn deze waarden respectievelijk: population dataset, Germany en 'as bad as possible'. Bij een klik op de 'Create map' knop wordt een zogenaamd 'purpose object' aangemaakt. Dit object bevat de gekozen dataset, land en subtext. Naast deze waarden bevat ze nog enkele berekende waarden zoals getoond in Listing 1.
Listing 1: Voorbeeld van een purpose object.

Dit object wordt door middel van een HTTP-request verzonden naar een OpenCPU server. Deze server gebruikt het object om het voorspellen van een geschikte kaartconfiguratie te starten. Omwille van de beperkingen tijdens het verzamelen van training data, bevatten de training data niet alle mogelijke land- en subtextcombinaties. Daarom wordt een filter toegepast op de, door de training gathering interface, verworven training data. Het filteren van de training data gebeurt op basis van het ontvangen purpose object. De training data worden gefilterd op gelijkende waarden, in plaats van specifieke waarden. Op deze manier tracht het algoritme zowel kaartconfiguraties voor landen te voorspellen die in de training data aanwezig zijn, als landen waarvoor geen training data verzameld werden. De data worden onder andere gefilterd op file, rank, map positie en quota-waarden. Na het filterproces worden de overgebleven example data gebruikt om het machine learning
Evaluatie en Toekomstig Werk

De experimenten binnen de thesis worden binnen dit hoofdstuk aangehaald. Verder wordt het machine learning algoritme geëvalueerd en suggesties voor verder onderzoek aangeboden.

Experimenten

Wanneer een choropleth kaart bekeken wordt, kan men elk land plaatsen in een rangorde ten opzichte van elkaar. Het eerste experiment test of deze rangorde aangepast kan worden door de classificatietechniek en het kleurenschema te
veranderen. De resultaten van het experiment werden onderworpen aan een ANOVA-test. Uit het experiment blijkt dat deze rangorde ingerad signific-icant verschilt wanneer de classificatietechniek veranderd wordt. Dit is niet het geval wanneer het kleurenschema verandert. Toch kan uit de resultaten afgeleid worden dat er kleine verschillen bestaan wanneer het kleurenschema verandert. Deze resultaten zijn echter niet consistent voor elk land en elke dataset. Bijvoorbeeld, als de door de kaart lezer waargenomen rang van België in een dataset verhoogt door de data aan de hand van de quantiles-methode te classificeren, dan kan dezelfde techniek in een andere dataset de waargenomen rang van België verlagen.

Omdat er geen specifiek effect gerealiseerd kan worden bij het toepassen van een bepaalde classificatietechniek, maakt MapTales gebruik van machine learn-ing.

**Evaluatie van het Machine Learning Algoritme**

Het ontwikkelde algoritme is getest in 2 fase. De eerste fase test de accu-raatheid van het algoritme wanneer een voorspelling gedaan wordt voor situaties die opgenomen zijn in de training data. De tweede fase test de accu-raatheid wanneer de voorspelde situatie niet aanwezig is in de training data. De resultaten kunnen teruggevonden worden in de grafieken in Figuur 11a en Figuur 11b. Uit de evaluatie is gebleken dat het voorspellingsalgoritme met een gemiddelde accuraatheid van 7% een configuratie-ID kan voorspellen. Wanneer enkel een kleurenschema of classificatietechniek voorspeld wordt, bereiken we respectievelijk een gemiddelde accuraatheid 22% en 33%. Wanneer het voorspellingsalgoritme bij het voorspellen van een ID vergeleken wordt met een algoritme dat random een ID zou toewijzen (1 van 25), doet het ontwikkelde voorspellingsalgoritme 2% beter. Wanneer het kleurenschema of de classificati-etechniek apart voorspeld wordt, is dit 2% en 8%. Dit wil zeggen dat de tool er inderdaad in slaagt om voor een beperkte groep mensen een kaartconfiguratie te voorspellen die voldoet aan een subtext.
(a) Accuraatheid van de voorspelling bij training data evaluatie. (b) Accuraatheid van de voorspelling bij niet in training data evaluatie.

Figure 11: Resultaten van de evaluatie van het voorspelling algoritme.

Toekomstig Werk

Binnen deze thesis werd het effect van het veranderen van kleurenschema en classificatietechniek onderzocht. Er werd met een klein succes een poging gedaan om de gecomuniceerde informatie van een kaart aan te passen zonder de effectieve data aan te passen. Uit de resultaten van de experimenten en evaluaties kan afgeleid worden dat het moeilijk is om een bepaalde kaart-configuratie te voorspellen die voor iedereen dezelfde uitkomst geeft. Het onderzoek kan uitgebreid worden door meer doelpublickericht onderzoek te doen. Verder bestaat de mogelijkheid om de effecten van andere cartografische principes te onderzoeken. Alsook kan gebruik gemaakt worden van bijvoorbeeld eye-tracking software, om de manier waarop een mens kaarten leest te onderzoeken en op deze manier de mens te misleiden.
Acknowledgements

First, I would like to thank Prof. Schöning. I thank him for the opportunity to work on this thesis. I appreciate his advice and expertise during the realisation process of this thesis.

Second, I am grateful to Nick Michiels for dedicating time to review this thesis. His remarks and advices added to the output quality of this thesis.

Third, I would like to thank the people that participated in the experiments within this thesis. I thank them for their time, honest cooperation and justification during the experiments.

Last but not least, I would like to thank and show my deepest appreciation to my friends and family. During difficult and stressful times, there was a listening ear and supporting shoulder. Thank you.
Chapter 1

Introduction

To start the exploration of cartographic principles and their effect on the perceived communication message of a map, this chapter describes a short history of maps. The chapter introduces misleading maps by showing some early and recent examples. This introduction describes the contribution of this thesis to research that explores misleading maps. Last, the structure of this thesis is explained.

1.1 History of Maps

Maps have existed for approximately 8000 years [19]. The oldest preserved map was created around 6200 BC in Catal Hyük Anatolia. It was a wall painting that showed the position of streets and houses of a town in Turkey. Besides streets and houses, the wall painting contained surrounding geographic features such as a volcano. Back then, the main function of a map was organizing hunts and local gatherings [38]. As cartography evolved over the centuries, maps became useful for more than hunting. They became an important tool for navigation and exploration of unknown areas. As a result, several attempts were made over the years to create a map of the world. Figure 1.1 shows some of these attempts. Maps like these were used by travellers to navigate. This lead to the discovery of new places e.g. America was discovered in the 15th
Figure 1.1: Maps of the world as known at the time. a) Eratosthenes - 250BC b) Ptolemy - 2nd Century c) Catalan world map - 15th Century d) Mercator - 16th Century.

century by Columbus [1]. Until the 16th century, only reference maps were made [3]. It was in this century that scientist started creating thematic maps showing demographic data e.g. tides, elevation and wind. This new type of map lead to new map roles. A combination of reference maps and thematic maps are used in wars. They help to organise attacks by informing the soldiers about strategic travel roads and the surroundings of these roads. Thematic maps are also used in spacial ordering e.g. to map building and farming areas.

MacEachren [5] summarises the roles and functions of maps in his model shown in Figure 1.2. He developed a three-dimensional space to define the use of maps. The first dimension reflects whether the map is created for own private use or for public display. Second, whether the created map is used for illustrating known or unknown phenomena. The last dimension characterises a map on the amount of human interaction needed to get information
Figure 1.2: MacEachren’s model of map functions. Every map can be placed in this three-dimensional space describing the data, audience and interaction of the map [5].

from the map. This is called human-map interaction. An example of human-map interaction is the use of Google maps which often requires zooming into a map to get detailed information. According to MacEachren all maps can be categorised into this three-dimensional space. As suggested in Figure 1.2, most maps that are used to present a phenomenon can be placed in the three-dimensional space on the position of 4. This means these maps are public, showing a phenomenon that is known on low human-interaction maps. Maps that are used to explore, analyse or synthesise a phenomenon are placed elsewhere in the three-dimensional space. The discussed examples of map roles and functions have 1 thing in common. All of them serve a communication purpose. Therefore, it has always been tempting to use them to weigh people in a direction of thought. Monmonier [7] took a deeper look in how this is done. The next section describes some typical examples.
Figure 1.3: Class breaks can be manipulated to yield choropleth maps supporting politically divergent interpretations [7].

1.2 Lying With Maps

Monmonier uses illustrative examples in his book to explain how maps can be used to lie and mislead. As the background of the map creator is usually unknown, it is hard to tell whether they made misleading maps on purpose or not. Due to incorrect use of cartographic principles, bad decisions sometimes result in misleading maps. Figure 1.3 shows an example where the cartographer deliberately changed the class breaks to alter the communication message of the map. Although the same data are displayed on both maps, the situation of the housing units lacking a telephone seems worse on the left side map. Advertisement is an other proficiency where communication messages of maps are changed to make phenomena more attractive to people. Figure 1.4 illustrates how a map made by engineers gets transformed into a map for advertisement. The right side map shows compact, well connected, short rail roads whereas the actual rail roads are a lot longer. For a tourist that wants to travel, the right side map looks a lot more attractive.
In the past, maps have not only been used to seduce the town board or for advertisement. Maps have also been used for political propaganda on a large scale. One of the many propaganda maps produced during World War II is shown in Figure 1.5. This map was ordered by the former German government during World War II. By showing that the land coverage of Germany is a lot smaller than the land coverage of Great Britain and its allies, the map communicates to its reader that not Germany, but Great Britain is the aggressor. This way, the German regiment targeted to gain sympathy and help from other countries. Even today, misleading maps are published in newspapers and articles. Some recently published examples are described here.

The first example was published in a Belgian newspaper. Figure 1.6 shows a map that contains information about the "smartest cities" in Belgium. The results are calculated by combining scores on several categories. The covered categories are digital organisations, air quality and energy use. The numeric table shows that Hasselt scores highest of 20 Belgian cities. However, when looked at the map itself it seems like Leuven, Gent and Kortrijk are the smartest cities. This impression is created by the fact that Hasselt has no color assigned on the map. Besides this fact, people could argue whether the chosen map type is best for displaying the data. As Hasselt represents a
small area on the map and therefore is visually pushed back by bigger cities, proportional symbols may have been a better choice instead of some sort of choropleth map. A second example was published in another Belgian newspaper. Christmas and Sinterklaas are two holidays in Belgium on which people give presents to each other. Figure 1.7 shows the amount of money spent per child on these days. The red hat and present represent the average amount of money for Flanders. The yellow hat and present represent the province of Limburg. The map shows that the difference in spending money between Flanders and Limburg on Sinterklaas (symbolised with hats) is three to four times as much in comparison with this difference on Christmas (symbolised with presents). But when looked at the numbers, the difference is only twice as big. Figure 1.8 shows a similar case, in this example the red hats which represent the percentages of presents between 51 to 75 euro and above 100 euro both show the same value, namely 6%. But the drawn hats symbolising
Figure 1.6: Map of the smartest cities according to Agoria [35]. By not assigning a color to Hasselt, the map gives the wrong impression that Hasselt is not the smartest city.

The two latter examples show maps that seem to use proportional symbols, but they are not proportional to the values of the attribute the symbol represents. This results into misleading maps.

Figure 1.7: Mapping of the amount of money spend on Sinterklaas and Christmas for Flanders (red) and Limburg (yellow) [37].
Figure 1.8: Mapping of money spend during ‘Sinterklaas’ for Flanders and Limburg [36].

The third example shows a misleading map due to bad color scheme use. Figure 1.9 shows the life expectancy in the bottom income quartile of the United States. By assigning the darkest colors to the lowest class intervals, it looks like the life expectancy is high in countries where it is low. This use of color schemes is counter intuitive and could be confusing for the map reader.

Figure 1.9: Life expectancy in the bottom income quartile of the United States [26]. The used inverted color scheme could be confusing to the readers of this map.
The last example shows that some maps are deliberately altered e.g. to remain the copyright. Paper towns are imaginary towns that are added to a map. Google, for example, adds a lot of Paper towns to their maps. In fact, they add so many paper towns that someone was able to create a paper town road trip along the east coast of the United States as shown in Figure 1.10. If the same towns appear on maps that are not owned by Google, they know their maps were used. This last examples differs from the others in the fact that here, the underlying data are changed. Within this thesis, the underlying data will not be changed in order to create misleading maps.

![Google adds a lot of paper towns to their maps. This image shows a road trip along Google maps' paper towns](image)

Figure 1.10: Google adds a lot of paper towns to their maps. This image shows a road trip along Google maps’ paper towns [29].
1.3 Contribution

People have been surrounded by maps for a long time. Maps tell them where they are, what the weather will be like and how many people voted for their favorite president candidate. As maps have been around for such a long time, people place trust in maps and believe what they show. But what if the map creator does not like the truth? Would he or she have a choice in how to display the data? Should people blindly believe what a map suggests?

In the past research has been done to prove that misusing cartographic principles can influence the perceived communication message of the map [7]. This thesis contributes to this research by developing software which automatically creates a choropleth map with a subtext without changing the original underlying data. Within this thesis, only changing cartographic principles is allowed. For example, it is allowed to change the classification technique, projection or color scheme. It is only permitted to change these in a cartographic justified manner e.g. no new classification techniques are invented. In order to develop the tool, after conducting several experiments and analysing the results, machine learning is used to determine which map configuration is best to communicate a given subtext.

1.4 Structure and Conventions of the Thesis

The thesis is organised as follows. In Chapter 2, the reader is introduced to basic cartographic principles, map elements and machine learning. Chapter 2 is followed by the description of the developed tool in Chapter 3. Chapter 4 outlines the results from the experiments conducted in this thesis to investigate the effects of changing cartographic principles. In the last Chapter, Chapter 5, the thesis draws to a conclusion and looks into possible further research.

In order to correctly understand this thesis description, following words are clarified. The perception of a map is the communication message that is un-
derstood by the map reader. A *purpose* is the message that the map creator wants to transport to the map readers. Other words for purpose are *map purpose*, *subtext* or *communication purpose*.

The developed tool within this thesis is called MapTales. This is a concatenation of the word 'map', as it creates maps, and the word 'tales'. Tales represents the fact that the tool user wants to communicate a story or subtext when creating the choropleth map.
Chapter 2

Related Work

This chapter introduces the two underlying core concepts of this thesis, maps and machine learning. In Section 2.1 map types, map elements, data classification and color schemes are described. The second core concept, machine learning, is put forward in Section 2.3. In this section data gathering and three machine learning techniques are clarified. Next to these core concepts, related work around perception and map generation is described.

2.1 Maps

2.1.1 Map Types

There are mainly two types of maps. Reference maps display the boundaries, names and unique identifiers of geographic areas, as well as roads, railroads, coastlines, rivers and lakes. They focus on the location and names of features and their geographic relation to each other [17, 13] e.g. a road map of Hasselt shown in Figure 2.1. Thematic maps focus on the spatial variability of a specific distribution or theme e.g. population density or average annual income [17]. Thematic maps can be branched into several subtypes [28] i.e. choropleth maps, proportional symbol maps and dot density maps. Figure 2.1 shows the European population using these thematic map subtypes.
Proportional symbol maps are easy to read and compare data values but the symbols may overlap which decreases readability [3]. Second, dot density maps use points to represent values. Each point may represent 1 within the displayed phenomenon or a predetermined value e.g. 1 dot = 2500 people. The research within this thesis focuses on choropleth maps. Choropleth maps are thematic maps in which areas are distinctly colored or shaded to represent the value of a particular phenomenon in each area [13]. An other name for a defined area on a choropleth map is enumeration unit. Most Choropleth maps attach the displayed data to an enumeration unit e.g. provinces or countries [8]. Important for creating sensible choropleth maps is that the used data values are expressed in the same format e.g. percentages, rates or phenomenon per square mile.
2.1.2 Map Elements

A map contains several elements as shown in Figure 2.2. All maps have a map body. It is the portion of the map that displays 1 or more data layers. In Figure 2.2 the map body shows the world habitat areas which is clarified by the map title. Grid lines on maps define the coordinate system. They are numbered to provide a unique reference to features displayed on the map. Other common elements on a map are scale, projection information and a north arrow. Some maps contain an inset map. This is a small map displayed within a larger map. An inset map might show a detailed part of the map at a larger scale or the extent of the existing map drawn at a smaller scale within the context of a larger area [13]. As it shows a map in a larger or smaller scale, it needs to have its own scale representation. Besides the described elements, the author and data source are sometimes present on a map. The legend serves as the decoder for the symbology used in the map body, it contains colors or symbols and translates them to their meaning. Map symbols are used on reference maps as well as thematic maps. They vary in size, color, texture, orientation and shape e.g. reference maps can use railroad symbols and thematic maps can use symbols in different sizes to show the distribution of a phenomenon. Multiple map icons can be used to display varying elements on 1 map e.g. a water drop icon indicates water whereas a cow icon can indicate farming areas. As this thesis focuses on choropleth maps, the legend attached to the map provides the translation of a color to the corresponding classification interval. Chapter 2.1.3 describes how these classification intervals are implemented in a map design.
2.1.3 Data Classification

Data classification is the process of analysing the distribution of values in a dataset and grouping them into intervals [3]. Whereas this thesis focuses on choropleth maps, there are mainly two types: classed and unclassed choropleth maps [8]. Unclassed choropleth maps assign a color to each enumeration unit based on its relative position in a color range. As humans can only differentiate a limited amount of colors, for choropleth maps it is important not to exceed 10 different color values [4]. Ignoring this will make it impossible for most humans to see the difference. Classed choropleth maps divide the data values into classes. According to the research of Evens [32] class identification is without errors when five or less classes are used. Therefore, within this thesis data are classified into five classes. Figure 2.3 shows an example of the life expectancy in Africa on a classed and unclassed choropleth map. There are numerous techniques for classifying data. Four of the most common classification techniques are: equal interval, quantiles, standard deviation and natural breaks. Each technique uses a different calculation and has its own
characteristics. *Equal intervals* are best used when the raw data are evenly distributed [14]. The classes are calculated by subtracting the lowest from the highest data value and divide the result by the desired number of classes. Each class now represents an equal data range. Second, *natural breaks* are used to classify similar data values into the same class. Natural breaks in the data are identified by finding points that minimise the value differences within the class intervals, and maximises the differences in between the class intervals [14]. The *quantiles* classification technique puts an equal amount of data in each class. This is done by dividing the number of values by the chosen number of classes. All data values are ordered and assigned to a class interval. Since the class assignment of quantiles is based on ranked data, quantiles are most useful classifying ordinal data [3]. Last, *standard deviation* is a classification technique that points out how much data varies from the average. Class breaks are created with equal value ranges that are a proportion of the standard deviation. This is usually done at intervals of one, one half, one third, or one fourth [3]. The effect of using the different classification techniques on a dataset covering oil use over the world [12], is shown in Figure 2.4. The effect of each classification technique depends on the distribution of the data values within the dataset.
2.1.4 Color Schemes

Since this thesis studies how to create misleading choropleth maps, color is used to visualise the different class intervals. Choropleth maps can have a great diversity of colors. However, not all color schemes are suitable for all sets of raw data. A key task in deciding which color scheme to use, is analysing the raw data. In general there are two types of raw data. Qualitative data covers different phenomena e.g. different types of vegetation. Quantitative data covers quantities of a phenomenon e.g. population or energy use.

Those two data types have their own suitable collection of color schemes. For qualitative data, a qualitative color scheme is the proper choice. Qualitative color schemes consist of unrelated colors that differ in hue to represent distinctions in the dataset [10, 16].

![Color Scheme Examples](image)

Figure 2.5: Different color scheme examples [11].
Diverging and sequential color schemes as shown in Figure 2.5 are convenient for displaying quantitative data on a choropleth map. Sequential schemes order data from high to low, displaying the highest values as a dark shade and the lowest values as a light shade [16]. Diverging color schemes are used to point out the differences between the lower and higher values in contrast with the average of a corresponding quantitative dataset [10]. Not respecting the color scheme usage as described, decreases the map readability and results into maps that are hard to interpret by the reader.

2.2 Perceiving and Generating Maps

2.2.1 Perception

The Cambridge dictionary defines perception as a belief or opinion [27]. In [34], Kennedy summarises 39 studies about perception. Research shows that a perception is often the same for a lot of people. When participants are asked to describe the graphs in Figure 2.6 which both show the same underlying data. The bar chart is described as showing discrete values and the line chart is described as showing a trend. This shows that when a graph is made to show trends, lines should be used to transmit the correct communication message [40]. Other research explored whether people could best indicate proportion on bar charts or pie charts. This research showed that when the number of components in the charts increased, the perceived proportion on pie charts were more accurate.

Hollands and Spence [33] investigated the performance of pie-, bar- and line charts while displaying the evolution of a phenomenon. They found that pie charts failed to show change. Bar- and line charts performed similarly. They found that it is caused by the fact that people draw imaginary lines between the bars. This results in a graph as shown in Figure 2.7.
Figure 2.6: Charts showing the same data using bars and lines [40]. Although they show the same data, they are perceived differently.

Figure 2.7: An illustration of imaginary lines drawn by the human brain while reading a bar chart [34].
These examples show that the perception of people can be changed by using different ways to display the same data. This thesis explores choropleth maps in order to determine whether a belief or opinion can be changed by changing cartographic principles while creating choropleth maps.

2.2.2 Generating Maps

A lot of tools are available to create or edit maps. Some of these tools like QGIS are stand alone programs, others are online applications like Color Brewer [11], ScribleMaps [24] and ZeeMaps [25]. Next to these existing tools, developers are provided with libraries to create tools themselves. Libraries like MapBox [15] and Leaflet [9] provide the components developers need to create maps. Regardless which tool or library is used, all of them need data to be able to produce maps. On the one hand, data is needed to draw geographic elements like borders or lakes. On the other hand, data that describes the phenomenon that is displayed on the map is needed. This data can be gathered by professional instances or volunteers. Data about the world and everything in it are widely available and occur in tons of different formats. Therefore, the developers that wish to use data gathered by others, sometimes need to process the data and re-factor them into a suitable format for the tool or library they want to use. Recently Freeman created a great example of a map generating tool. To do this he used data which is provided by the United States Census Bureau. He developed a Twitter bot [39] that automatically generates county-level maps based on American Community Survey data. This bot generates one choropleth map per hour. Figure 2.8 shows an example of a map generated by the Twitter bot.
Figure 2.8: Map produced by the Freeman’s Twitter bot. The maps shows the percentage of Black or African American people per county.

2.3 Machine Learning

The field of machine learning is concerned with the question of how to develop computer programs that automatically improve with experience [6]. In order to make machine learning possible, training data has to be gathered. After the machine learning is completed, an accuracy test is performed to evaluate the used algorithm. To solve a single problem, a number of machine learning techniques are possible. This Section describes data gathering and three of the most used machine learning techniques.

2.3.1 Data Gathering

Training data are data that the machine uses to learn about a topic. Training data are a number of sample entries that have been gathered. A sample
contains a number of attributes (predictors) that are used to calculate a prediction along with target attributes which are the values that the machine attempts to predict. The data are gathered from previous recordings or conducted questionnaires. It is an important part of machine learning, if the data are biased, the result will be as well. Furthermore, the training data are often used to check the accuracy of the machine learning algorithm. This is done by using two thirds of the training data to train the algorithm and one third as an accuracy reference [22]. To implement a successful machine learning algorithm, the training data must be collected carefully. They have to be as representative, bias free and complete as possible. As for the fact that this is almost never the case, a machine learning algorithm takes this into account and sometimes makes assumptions to co-op with the fact that the training data are not perfect. A common phenomenon in machine learning is overfitting. A learned rule or hypothesis over-fits the training data if some other hypothesis that fits the training data less well, actually has a higher accuracy rating over the entire distribution of instances [6].

2.3.2 Machine Learning Techniques

Decision Tree Learning is the first discussed machine learning algorithm. Within decision tree learning the algorithm constructs a tree that represents all learned paths for a given dataset. The tree starts with a root node and branches out showing all learned paths which end in leaf nodes. Nodes contain one of the attributes in the dataset that is used as a predictor. The leaves always contain a value of the target variable. As every predictor is a possible root node, the predictor which provides most information gain is chosen. This information gain is calculated using entropy. Entropy characterises the homogeneity of an arbitrary collection of examples [6]. This entropy is used to calculate the entropy gain for a predictor, thus calculate which predictor reduces the homogeneity of the dataset the most. The process of calculating the entropy gain is repeated to calculate every next node in the decision tree. Figure 2.9 shows
Figure 2.9: Tom Mitchel’s example decision tree [30].

an example of a decision tree based on example data [30] containing records that show under which weather conditions tennis was played. The predictors in this tree are outlook, humidity and wind. The target for this tree is deciding whether or not to play tennis and its leaves contain the answer to that question. Besides over-fitting, decision tree algorithms have to take missing and continuous attributes e.g. temperature, into account. To avoid over-fitting, a tree is post-pruned after over-fitting is allowed. Two techniques are used to prune a decision tree. First, reduced error pruning where every node in the tree is evaluated for pruning. When it evaluates positive, the sub-tree is removed by making the node a leaf node. A node is evaluated positive if the pruned tree does not perform worse than the original tree. This means, if all positive training examples are still evaluated positive after pruning the tree, the sub-tree is removed. In Figure 2.9 the tree is pruned at the outlook predictor when the outlook is 'Overcast'. A second way to prune a tree is rule post-pruning. This prune technique converts the over-fitting tree into rules. Each branch results into one rule. A rule extracted from the decision tree in Figure 2.9 is: IF the outlook is sunny AND the humidity is normal THEN tennis is played. Each rule is minified by removing all predictors that do not worsen the prediction accuracy. The resulting set is ordered by their estimated accuracy. The machine learning algorithm will consider these rules in order to predict a target with given predictors.
The issue of continues values is resolved by creating intervals in which the continues values are categorised. By doing this, the continues values are converted into discrete intervals. As for missing attributes, one possibility is to take the most occurring value of the attribute in the dataset. A second possibility is to assign probabilities to each value of the attribute and create a tree branch for every possible path that could be followed from the node with the missing attribute [6]. Eventually the weight of the branches are calculated and the best one is chosen. This way the missing attribute is replaced by, not the most occurring value, but the value which produces the most probable branch.

The second machine learning algorithm is the Bayes rule. The Bayesian learning algorithm, which calculates explicit probabilities for hypotheses, is among the most practical approaches to learning problems [6]. This algorithm is based on the Bayes theorem. The Bayes theorem returns the most probable hypothesis given training data and initial knowledge.

Probabilistic learning algorithms in general, calculate the most probable target by creating a distribution table. This table holds the distribution percentages of all possible predictor combinations in the example dataset where the target is evaluated positive. If the target is not a binary value, a distribution table is calculated for all possible values. Figure 2.10 shows the distribution table for a dataset that covers fatal car accidents in Belgium [18]. The probability of a target value can be calculated by using joint distributions [6]. Joint distributions calculate \( P(a|b) \) or the probability that the target is positive for a given \( b \). For example, what is the probability that a fatal accident occurs as a Belgian female (a) given she is 25 years old (b) using the dataset in Figure 2.10. The probability of this situation can be calculated using Formula 2.1.

\[
P(a|b) = \frac{P(a \land b)}{P(b)}
\]

\[
= \frac{0.02}{0.27} = 0.074
\]

45
In his lecture [22] Mitchell says there is no better way to make predictions than using distribution tables and follows up with a big 'but'. There are some problems with this technique namely continues values and the computational power needed in case of large datasets. But above all, it is not always easy to learn the probability distribution. That is, for a dataset with a 100 different boolean only variables there would be $2^{100}$ rows in the distribution table. In this case there would not be enough people in the world to represent every row in the table. To handle this possible lack of example data, prior knowledge can be used e.g. the probability of a coin flip resulting in heads is 50%. This is exactly what the Bayes rule does. It does this by introducing imaginary data examples that reflect the prior intuition. The more certain a prior intuition is, the more imaginary examples are added to the example data. This also ensures that there is at least one example for every possible outcome of the learned problem.

$$P(X|Y) = \frac{P(Y|X)P(Y)}{P(X)} \quad (2.2)$$

To calculate $P(Y|X)$ the Bayes rule (Formula 2.2) calculates $P(X|Y), P(Y)$ and $P(X)$. This means that for an example set which contains $n$ boolean attributes, the algorithm needs to calculate $P(X_1, \ldots, X_n|Y = 1), P(X_1, \ldots, X_n|Y = \ldots,\ldots$.
and \( P(Y). \) This means to calculate \( P(X|Y) \) \( 2^{2^n} \) + 1 calculations have to be made. To calculate the maximum probability \( P(X) \) does not have to be calculated because it is independent of \( Y. \) Doing these calculation results in even more calculation than the probabilistic learning algorithms which use a distribution table.

The *Naïve Bayes rule* solves this problem [23]. Naïve Bayes assumes:

\[
P(X_1, ..., X_n|Y) = \prod_i P(X_i|Y) \quad \text{i.e., that } X_i \text{ and } X_j \text{ are conditionally independent given } Y, \text{ for all } i \neq j.
\]

This Formula is often written in the form of \( P(X|Y, Z) = P(X|Z) \) which means the probability of \( X \) being true given \( Y \) and \( Z \) is equal to the probability that \( X \) is true given \( Z. \) Given this assumption, the chain rule (Formula 2.3) [31] can be used to derive Formula 2.4.

\[
P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)
= P(X_1|Y)P(X_2|Y) \quad (2.3)
\]

\[
P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)
= P(X_1|Y)P(X_2|Y)
= \prod_i (P(X_i|Y)) \quad (2.4)
\]

Using the Naïve Bayes algorithm which makes the assumption that all parameters are conditionally independent, new data examples are classified using Formula 2.5.

\[
P(Y = y_k|X_1, ..., X_n) = \frac{P(Y = y_k) \prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i|Y = y_j)} \quad (2.5)
\]

With Naïve Bayes calculating the probability of a combination of predictors is reduced to calculating the product of the individual probabilities of every predictor. The probability of every predictor in a dataset is only calculated once. This results in a faster and less power consuming algorithm.
An other way of machine learning is to learn rules directly from the example dataset using *sequential covering* or *separate and conquer* techniques. The algorithms used within these techniques all follow a similar process. Figure 2.11 shows the basic rule learning process. The algorithm learns one rule at the time. Once a rule is learned, the set of examples is reduced to the examples which are not covered by the current set of rules. The remaining examples are used to learn the next rule. This process is repeated until a criteria is reached or until all examples are covered by the rule-set.

Learning one rule at the time is called the **LearnOneRule** strategy. Rules are learned in a general to specific way or visa versa [2]. In the first case the rules are specified until it does not cover any negative examples. In the second case the rules are generalised until they start covering negative examples. In both cases a rule tries to cover as many example data records as possible. When generalising or specifying a rule to learn, entropy or other guiding techniques can be used to determine which part of the rule has to be added or removed.

The algorithm used for machine learning depends on the subject. Some computing problems are more suitable for the described machine learning techniques than others.
Figure 2.11: Rule learning process [2].
Chapter 3

MapTales: A Tool To Create Maps With A Subtext

This chapter provides an overview of the components of the tool developed within this thesis. After the overview, the main components are discussed in detail.

3.1 MapTales Tool Overview

Figure 3.1 shows an overview of the developed components. The two main parts of MapTales are first the tool user interface itself and second the training components. MapTales provides a user with an interface to create maps with a subtext. The training components are used to predict the most likely map configuration suited for a specified subtext. The training components can be divided into three parts. First, an OpenCPU server [20] provides a communication channel between javascript and R. The prediction algorithm that is used in MapTales is developed in R. Second, the training data are gathered using the user interface. This user interface sends the example data to a training data gathering server. The training components are marked ①, ② and ③ respectively in Figure 3.1. The arrows in this figure indicate data flow between the developed components.
3.1.1 MapTales User Interface

The user interface of MapTales, as shown in Figure 3.1 (UI), contains a menu in which three values can be defined: a dataset, target country and subtext. The combination of these three values form a communication message that the user wants to transport with the map. Each dataset contains European countries, ISO country codes and values of a specific phenomenon e.g. population. The subtext-value indicates whether the specified country has to look as good or as bad as possible. When the map purpose is set and the 'create map' button is pressed, the process to create the map is started.

To explain the process, consider the following scenario. The user wants to create a map of Europe showing the population number of Germany as low as possible. To do this the user specifies the map purpose in the MapTales user interface as shown in Figure 3.1 (UI). This means selecting the population
dataset, set Germany as the target country and set the subtext to 'as bad as possible'. When the 'Create map' button is clicked, the chosen dataset, country and subtext are combined with calculated variables in an object which is referred to as *purpose object* within this thesis. This object is similar to an example object as described in Section 3.1.2. As the configuration id, classification technique and color scheme are to be predicted, they are not included in the purpose object. These are the only attributes that differ between the two object types. Listing 3.1 shows the purpose object calculated in case of the proposed scenario.

```json
{
    file: "POP-europe.csv",
    country: "Germany",
    rank: 1,
    toTop: 0,
    toBottom: 46,
    goal: lowest,
    toMiddle: 22,
    pixelCount: 17622,
    pixelQuota: 0.35,
    numNeighbours: 9,
    neighbourQuota: 0.24,
    mapPosition: 5,
    positionQuota: 0.23,
    bigCountryQuota: 0.61
}
```

Listing 3.1: Illustration of the purpose object.

The created purpose object is sent via HTTP to the OpenCPU API that is included in the OpenCPU [20] server. The OpenCPU server uses the *purpose object* to predict four map configuration IDs. The IDs are ordered from most to least probable and returned to the MapTales user interface. For
example, when the purpose object in Listing 3.1 is sent, following configuration IDs are returned: 0, 8, 4, 5. These IDs are a unique identification for a combination of classification technique and color scheme. These combinations are the same combinations that were used during the training phase. The message flow of the HTTP-requests and responses between components is visualised in Figure 3.3.

The most probable ID, thus 0, is used to proceed. The other three IDs are stored for later use. Using this ID, MapTales executes an ID look-up in a pre-defined object containing all unique identified configurations. By doing this, the tool determines the predicted classification technique and color scheme. MapTales creates classed choropleth maps, the dataset needs to be classified into classification intervals according to the predicted classification technique. To do this, it sends a second HTTP-request to component \( \text{(1)} \) as shown in Figure 3.1. This request contains the name of the dataset and the predicted classification technique. Component \( \text{(1)} \) uses these values to calculate the classification intervals. When calculated, these intervals are returned to MapTales. Within MapTales, color is used to visualise the values of each enumeration unit. Therefore MapTales combines the received intervals and predicted color scheme. At this point the combination of classification interval and color scheme as shown in Figure 3.2 is set and can be used. To draw the choropleth map, MapTales uses Leaflet [9]. Leaflet requires MapTales to provide it with the borders of each enumeration unit in order for it to draw a map. In this case the enumeration units are the European countries. The borders along with a number of attributes e.g. name and ISO code of the enumeration units are
defined in a geoJSON object. Listing 3.2 shows a snippet from the geoJSON
object. The illustrated part, called feature, describes Germany.
Before the passing of this geoJSON object to Leaflet, the geoJSON object is
combined with the dataset. By comparing the ISO values in the dataset and
geoJSON object, each feature is assigned the correct phenomenon value. When
the geoJSON object is passed to Leaflet, these values can be used by Leaflet
while it iterates through all features to draw them. The phenomenon value
of each feature is evaluated against the predicted classification intervals and
their assigned color. By doing this, the color of the feature can be retrieved.
To complete the predicted map, Leaflet generates a legend that translates the
used color scheme to the classification intervals. The map resulting from the
prediction made for the proposed scenario is shown in Figure 3.1 (4).
As can be seen in Figure 3.1 (UI) the menu contains a slider. This slider can
be used to iterate over the alternative maps. The slider has four different po-
sitions, these positions correspond to the four previously predicted IDs. When
the slider is moved, the current map is removed and the former explained pro-
cess starting from the ID look-up is repeated to draw the new map. This slider
was introduced as a solution to the fact that the created map is the result of
a prediction based on perception and therefore, the possibility exists that the
drawn map seems unrealistic. For example, more than 80% of the countries
on the map are assigned the same color due to the effects of using different
classification techniques. When this happens, the tool user may want to use
an alternative map to support his or her purpose.
Listing 3.2: Feature of the geoJSON object which describes Germany.
Figure 3.3: Illustration of the message flow during the creation of a map with a subtext.

3.1.2 MapTales Training and Predicting Component

MapTales collects training data through the user interface that is shown in Figure 3.1. Twenty five people contributed to gathering the training data that is used to train the tool described within this thesis. To gather training data, four maps are repeatedly presented to the user. As MapTales attempts to predict the classification technique and color scheme, these cartographic principles differ in the presented maps. The maps are created using Leaflet in a similar way as described in Section 3.1.1.

Every time four maps are presented, the user is asked to select the map which best suits a given purpose e.g. choose the map that shows the highest population number for Germany.

When one out of four maps is chosen, the map configuration of this map and the given purpose is used to create an example object. Listing 3.3 shows an
illustration of an example object. Next to the purpose and map configuration, several other attributes are calculated and added to the object. These attributes will be used by the prediction algorithm as described later in this section. Besides the rank of the target country within the dataset, the relation with neighbouring- and similar sized countries are calculated and added to the example object. A relation e.g. between France and its neighbours, is calculated by dividing the average data value of the neighbouring countries by the data value of France. This results in a value that reflects a relation. Within this thesis, these values are named quota values.

```json
{
    file: "CO2-europe.csv",
    country: "France",
    configId: 4,
    colorscheme: 1,
    rank: 18,
    toTop: 17,
    toBottom: 29,
    goal: highest,
    toMiddle: 5,
    pixelCount: 22437,
    pixelQuota: 1.36,
    numNeighbours: 7,
    neighbourQuota: 1.66,
    mapPosition: 5,
    positionQuota: 1.50,
    bigCountryQuota: 1.16,
    classification: "Quantiles"
}
```

Listing 3.3: Illustration of an example object.
To determine the map position of a target country, Figure 3.4 is used as a reference. Once the example object is completely calculated. As shown in Figure 3.1, this object is sent to the data gathering server using an HTTP-request. The server collects all received example objects and bundles them into one file (i.e. comma separated value file). The collection of example objects in the file is used to train the prediction algorithm of MapTales.

![Map divided into areas](image)

Figure 3.4: This figures show how the map is divided into areas.

The prediction algorithm starts to work when MapTales sends a prediction request to the OpenCPU server as shown in Figure 3.1. For explanatory purposes, imagine that the OpenCPU server receives a prediction request containing the purpose object as shown in Listing 3.1. The subtext within this purpose object is to create a choropleth map that makes the German population look as low as possible. First, the prediction algorithm takes the training data file and applies a filter. This filtering is based on the purpose object included in the prediction request. By filtering the training data on similar rather than exact values, the prediction algorithm predicts not only countries
The filter process starts by removing all example records that belong to a different dataset. Next, the training data are filtered on rank, amount of pixels covered by a country and position on the map. The rank, pixel count and quota values do not have to be exactly the same, but within a 10% margin. If they are not within the margin, they are removed from the training data. The resulting sub-file is filtered by removing all example records whose quota values are not within a 10% margin of the quota values specified in the request object. For example, when filtering the dataset on a neighbour quota of 1. All entries with a corresponding quota between 0.9 and 1.1 will be kept. All others will be excluded from the training dataset. It may occur that no example record attribute values are within the specified margin. When this happens, the records that were to be removed by the filter, remain in the training dataset. When the filter process is completed, the representative example entries remain. Only these entries are used by the prediction algorithm. The table in Figure 3.5 shows the results of a test that indicated that when using a 10% range, the highest accuracy is obtained. Figure 3.6 shows the filter process using the purpose object in Listing 3.1 on a limited dataset. The bottom table in Figure 3.6 contains the remaining data entries on which the prediction algorithm is applied. The training data remaining after the filter process, is used to train a Naïve Bayes algorithm. The algorithm is trained to return a list of IDs ordered from most to least probable. The four most probable IDs are returned to the MapTales user interface for further use.

<table>
<thead>
<tr>
<th>Range</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>35%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>6%</td>
<td>7%</td>
<td>6%</td>
<td>5%</td>
<td>5%</td>
<td>5%</td>
<td>5%</td>
</tr>
</tbody>
</table>

Figure 3.5: Experiment to determine the filter margin. Results show the highest accuracy for a range of 10%.
Figure 3.6: An example of the training data filtering process.
Chapter 4

Evaluation and Future Work

To create maps with a subtext, within this thesis several experiments were conducted to determine if the communication message of the map changes while changing cartographic principles. The results are discussed in this chapter. Besides the experiments, the prediction algorithm used by the tool described in Chapter 3.1.2 is evaluated.

4.1 Pre-study

A study was conducted in order to investigate the effects on the perceived communication message when changing the classification technique or color scheme. The experiment consisted of two stages. In the first stage the effect of changing classification techniques was explored. The second stage studied the effect of changing the color scheme. To do this, a pre-study tool was developed and used in both stages. The tool consists of a simple user interface as shown in Figure 4.1. To question the effects, in both stages of the experiment, 36 maps in total were created using QGIS [21]. These maps are displayed one by one along with a task. Each task asks the user to rank a given country on a scale. To determine the effect of the tested cartographic principle, this rank is compared to the actual rank of a country within the dataset.
4.2 Experiment Results

The first experiment (pre-study) was introduced to determine whether changing cartographic principles has an effect on the perceived rank of a country on a map. As this thesis concentrates on choropleth maps, classification and color schemes are investigated. Therefore the first experiment consists of two stages. The first stage investigates the effect of changing the classification technique. The second stage explores the effect of changing the color scheme of a choropleth map. Both stages use the same setup and methods to study the effects.

Both experiment stages questioned an average of 20 participants. In both experiments, four countries were used to question the cartographic principles. By ordering the countries on their value in a descending order, every country is ranked within a dataset. One of the countries is highly ranked, one is a low
ranked country and two are in the middle range ranks. By questioning these four countries, the experiment gives closure on whether the effect of changing a cartographic principle is the same for all countries in the dataset. In the first stage, four classification techniques were questioned. This results in 16 different maps in the first stage. The second stage investigated five different color schemes. This means that 20 maps were shown to each participant. The participants were asked to identify the rank of four countries.

The rank, identified by a participant, is compared to the actual rank of the country within a dataset. The difference between those values is registered to determine the effect. To evaluate both experiments an ANOVA-test is used. The ANOVA-test on the first stage shows that for three out of four questioned countries, there is a significant difference in perception between classification techniques. Nevertheless, the effects are not the same for all three countries. Appendix A.4 shows these effects in detail. The graphs show that the effects of changing the classification techniques are not the same for high ranked countries and middle range countries. The effects between two countries that are ranked in the middle range of a dataset are neither. For example, The Netherlands and the United Kingdom are ranked 29 and 31 out of 45 respectively in the dataset. The change of classification techniques does not have the same effect on the perceived rank of The Netherlands and The United Kingdom. Using the quantiles classification technique, The Netherlands are ranked higher by the participants whereas The United Kingdom is ranked lower.

These results show that the communication message of a map can be influenced by using different classification techniques, but this effect is different for each country. Due to the nature of classification techniques, this effect is also different for each dataset. The histograms in Appendix A.1 show the variance in results for every country. The green dots indicate the actual rank, whereas the blue dots indicate the average perceived rank. The histograms confirm that the effect on perception is different for each country. For countries that are in the mid-range of the ranks e.g. The Netherlands and United Kingdom,
the perceived rank is spread. This indicates that the perceived rank of middle range countries is different for each individual and the effect is therefore even harder to learn.

Part two of the experiment explores the effect of changing the used color scheme. The ANOVA-test for this stage showed no significant difference in perceived rank between the color schemes. Nevertheless, the histograms in Appendix A.2 show that small differences in perception occur. And thus may add to change the perceived communication message of choropleth maps for some targeted readers.

Neither part of the experiment shows a closing conclusion whether how to use a classification technique or color scheme to change the map perception in a certain direction. Especially the effects of color schemes are unique for each individual. To be able to target the largest part of the future map readers, machine learning is used within MapTales to determine which classification technique and color scheme to use in order to acquire the intended communication message.

4.3 Evaluation of the Prediction Algorithm.

The training data that is used by the prediction algorithm is gathered by the use of the training data gathering interface as shown in Figure 3.1. Besides training the prediction algorithm, the training data is used to evaluate the algorithm. Two thirds of the training data is used to train the machine learning algorithm. The remaining training examples are used as an evaluation set.

The training dataset contains almost 800 example entries. This means that, for the evaluation, around 550 examples are used to train the machine learning algorithm whereas the remaining 250 examples are used to evaluate. Figure 4.2a shows the accuracy of the algorithm over the amount of used training examples. The prediction algorithm is able to predict the configuration ID

66
with an accuracy of 8%. This means that 8% of the predicted configuration IDs was correct during the evaluation. Not using the prediction algorithm and just randomly guess 1 configuration ID out of 25 would intuitively result in a 5% accuracy. This means, by using the prediction algorithm a gain of 3% is accomplished. While predicting the color scheme and classification technique, an accuracy of 23% and 36% is reached. As there are five possible color schemes and four possible classification techniques, this is a gain over random guessing of 3% and 9% respectively. These results are in line with the findings discussed in Chapter 4.2. The best prediction results are obtained when predicting the cartographic principles that show significant differences in the perceived communication message. It is reasonable that the cartographic principle that showed no significant difference is harder to predict.

(a) Accuracy of the prediction algorithm in training data evaluation.

(b) Accuracy of the prediction algorithm not in training data evaluation.

Figure 4.2: Prediction algorithm evaluation results.

A second evaluation of the prediction algorithm is done to determine the accuracy when predicting map configuration IDs for subtexts that are not in the training data. For example, predict which map shows the highest population for France. To do the second evaluation, instead of using only two thirds of the example data to train the algorithm. All example data is used as training data. For the evaluation, new example data is gathered containing
examples that are not in the training data. This way, the accuracy of new subtexts can be measured. The result of the second evaluation is shown in Figure 4.2b. While predicting a map configuration ID for a subtext that is not in the training dataset, the prediction algorithm reaches an average accuracy of 7%. The accuracy of a classification and color scheme prediction are respectively 31% and 22%. This means that the algorithm still gains 2%, 6% and 2% respectively over guessing, even when subtexts for countries that are not in the training data are handled.
Chapter 5

Conclusion

As maps have always been a powerful tool in communication, they have been used in countless situations and applications. Sometimes maps are used to pull a map reader in a certain direction. During the creation of these maps, cartographic principles are used to change the appearance of a certain phenomenon. Within this thesis, we explored how the use of classification techniques and color schemes can influence the perception of a choropleth map. This was done by conducting several experiments in which participants were asked to rank a country or select a map that is most suitable for a given subtext. These experiments showed that changing the classification technique influences the communication message of a choropleth map. When the color scheme is changed, no significant effects were registered. But the data showed that small differences in perception did occur. Therefore, machine learning was used to create a tool, called MapTales, that predicts a map configuration for a specific communication message. The evaluation of the prediction algorithm used by the tool showed that the algorithm is able to predict map configurations with an average gain of 2.5% over guessing accuracy. The gain of 2.5% is connected to the fact that the perceived communication message is subjective and often different for each individual. Nevertheless, the algorithm is able to predict configurations with a better than guessing accuracy. By doing this research, the first steps are taken in exploring what cartographic principles can
be used to change the communication message of a choropleth map. To fully understand which other cartographic principles can be used, more research is needed. Map projection, legend and distortion are cartographic variables that have not been studied within this thesis. To investigate these cartographic variables, similar experiments to the experiments within this thesis can be done. An other way of registering data about how people look at maps is the use of eye-tracking software. The results in this thesis show that it is hard to predict a misleading choropleth map for all map readers. Therefore, interesting research can be done in creating these maps for a specific group of readers. One could for example compare experiment results between a group that grew up with digital maps and a group that did not. Maybe one could go even further and explore whether the map perception is different for people with certain interests or news paper preference.
Bibliography

Books and Journals


**Electronic Resources**


[26] K. Styles. *This big data project shows poorest in the US now have same live expectancy as the Sudan.* http://thenextweb.com/us/2016/04/12/big-data-project-shows-best-

[27] Cambridge University. *Cambridge Dictionaries Online.*

[28] PennState University. *Thematic Maps.*

https://www.google.com/maps/d/embed?mid=z9TiRzXYwE-A.kouYXvd1t9rU&msa=0&ie=UTF8&t=m&ll=35.924645%2C-74.399414&spn=15.98893%2C26.323242&z=5&output=embed&w=600&h=450.


[31] WikiPedia. *Chain rule (probability).*
Articles


Appendix A

Appendix

A.1 Classification Experiment Summary

A.2 Colorisation Experiment Summary

A.3 All Population Set Map Configurations.

A.4 Effects on the Perceived Rank while Changing the Classification Technique
Figure A.1: Summary of the results of the classification perception experiment.
Figure A.2: Summary of the results of the colorscheme perception experiment.

<table>
<thead>
<tr>
<th></th>
<th>Divergent</th>
<th>Equal Blue</th>
<th>Equal Red</th>
<th>Logic Red</th>
<th>Logic Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greece</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
</tr>
<tr>
<td>Italy</td>
<td><img src="image6" alt="Graph" /></td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
<td><img src="image9" alt="Graph" /></td>
<td><img src="image10" alt="Graph" /></td>
</tr>
<tr>
<td>The Netherlands</td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
<td><img src="image13" alt="Graph" /></td>
<td><img src="image14" alt="Graph" /></td>
<td><img src="image15" alt="Graph" /></td>
</tr>
<tr>
<td>United Kingdom</td>
<td><img src="image16" alt="Graph" /></td>
<td><img src="image17" alt="Graph" /></td>
<td><img src="image18" alt="Graph" /></td>
<td><img src="image19" alt="Graph" /></td>
<td><img src="image20" alt="Graph" /></td>
</tr>
</tbody>
</table>
Figure A.3: All possible configurations for the population data set.
Figure A.4: Graphs showing the effects on perceived rank of a country while changing classification techniques.
Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling: "Lying with maps": A Framework to Create Maps with a Subtext

Richting: master in de informatica-multimedia
Jaar: 2016

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Smeets, Stein

Datum: 15/06/2016