GEO 511 MASTER'S THESIS

Demonstrating Spatial Patterns in the Construction & use of the GeoNames Gazetteer Data

Source: Oxford Internet Institute

Author
Yamgouet Monique Ndam
S 12-745-253
Stockenstrasse 3
8802 Kilchberg, Zürich
moniqueyamgouet@geo.uzh.ch

Supervisor/Faculty Representative
Prof. Dr Ross .S. Purves
ross.purves@geo.uzh.ch

Submission Date: January 2015

Acknowledgements

This thesis wouldn’t have been complete without the help of others:

- Prof Ross Purves for his time, effort and patience with me during this thesis,
- Mr. & Mrs. Helbing for making it possible for me to travel to Switzerland for further studies, their continuous love, support and encouragement during this time, they made sure I was comfortable and lacked nothing which could distract me from my studies.
- Mr. & Mrs. Eyoum; my parents, for their unending parental love, the messages, calls and encouraging words all the way from Cameroon kept me till the end of this academic journey
- Prof Fombe Lawrence Fon for his numerous emails and phone calls to remind me of my potentials encouraging me and letting me know I can do this and make him proud.
- Brenda Nabila my loving aunty and who had my back all the time
- My brothers Jerry, Junior, Frederick Lot and George Willibroad for their fervent prayers that kept me safe and sound
- Besong Enoh Belsham Besong for his unending support and encouragement during the course of my studies.
- Yvonne Engemise, Akum Stephanie, Adeline Bofon and Bahauline Finjap for their sisterly love and concern.
- Mbong Johnson, Hildermar Cruz for their time and support.

Abstract

Placenames also known as toponyms are names of places used by people on daily basis to identify or refer to a particular geographic location, these placenames are registered in books called gazetteers, most often than not, one placename could refer to multiple places and in different locations (ambiguity) and as such, care must be taken when referring a placename to a location. Placenames and also other information related to them are collected by administrative bodies or individuals and stored in books or records called gazetteers, it is uncertain if the information contained by the country's gazetteers have the same information as the world GeoNames gazetteer, as such one could search for a placenames in gazetteer, find some and not find others .Gazetteers are dictionaries of placenames; "they illustrate the largest freely available dictionary of geographic placenames per spatial unit, that is a square of one tenth degree of latitude and one tenth degree of longitude covering the globe used in conjunction with a map or atlas it is based on freely available national gazetteers and data sets as well as volunteered geographic information"(Bruner et al., 2008). Most gazetteers are free and open sourced and they give the exact position of a placename on the map. There are so many kinds of gazetteers, each having different contents and they are usually used for Geocoding (associating places to coordinates) , Geotagging , Geoparsing (recognizing a word as a place or location) and Georeferencing information from news and social network streams (Zhang et al., 2014) to compare data as well as to search and store information.

National newspapers from three chosen countries (Cameroon, Switzerland and Sri Lanka) will be read and the placenames will be identified in their text by hand; these placenames will be used to create a Gold Standard list which will later be used to compare with the placenames from GeoNames data and SwissNames (Switzerland names atlas), the news corpora are both from online (Sri Lanka Daily News online) and published sources (Cameroon Tribune and NZZ). Input of placenames data into the GeoNames gazetteer database for Haiti, Sri Lanka and Somalia will also be analysed to see if there exists any bias in data input or construction of this gazetteer. A dot density map will be produced from the coordinates in the GeoNames data to show the spatial distribution of placenames on space, this dot density map will also be used to produce a Kernel Density Estimation in order to describe the placename distribution and clustering (Cheshire et al., 2012) and to see if placename and population have the same hotpots per country, In order to know if there is a relationship between placenames and population, Regression Analysis will be carried out between placename per region and population count per region to see what kind of correlation exists.

The results will be analysed under various aspects such as number of placenames not available in the GeoNames database but available in the newspapers and atlases. Bias in the construction of the GeoNames data will be demonstrated by a cumulative graph to show the lag that exists in the input data of various countries, dot density map to show placename distribution in space, Regression analysis will be used to see what kind of correlation or relationship exists between population and placenames.

Résumé

Noms de lieux aussi connus que les toponymes sont des noms de lieux utilisés par les personnes sur une base quotidienne pour identifier ou se référer à un lieu géographique particulier, ces toponymes sont enregistrés dans les livres appelés nomenclatures, plus souvent qu'autrement, une toponyme peut se référer à plusieurs endroits et à différents endroits (ambiguïté) et en tant que telle, il faut prendre soin pour désigner un nom de lieu à un emplacement. Noms de lieux et d'autres informations qui leur sont liées sont collectées par des instances administratives ou des individus et stockés dans des livres ou des enregistrements appelés répertoires toponymiques, il est incertain que les informations contenues par les nomenclatures du pays ont les mêmes informations que l'GeoNames mondiaux répertoire toponymique, en tant que telle ne pouvait rechercher pendant toponymes dans les nomenclatures, trouver certains et pas trouver d'autres .Les nomenclatures sont des dictionnaires de noms de lieux; « Elles illustrent le plus grand dictionnaire libre disposition des toponymes géographiques par unité spatiale, ce est un carré de dixième degré de latitude et dixième degré de longitude couvrant le monde entier utilisé en conjonction avec une carte ou un atlas elle est basée sur des nomenclatures nationales librement disponibles et des ensembles de données ainsi que l'information géographique volontaire »(Bruner et al., 2008). La plupart des nomenclatures sont gratuits et open source et ils donnent la position exacte d'un toponyme sur la carte. Il y a tellement de sortes de répertoires toponymiques, chacune ayant des teneurs différentes et ils sont généralement utilisés pour le géocodage (associant endroits aux coordonnées), Géolocalisation, Géosémantique (reconnaître un mot comme un lieu ou d'un emplacement) et des informations Georeferencing de nouvelles et réseaux sociaux flux (Zhang et al., 2014) pour comparer les données ainsi que de rechercher et de stocker des informations.

Les journaux nationaux de trois pays choisis (Cameroun, Suisse et Sri Lanka) seront lus et les noms de lieux seront identifiés dans leur texte à la main; ces noms de lieux seront utilisés pour créer une liste Gold Standard qui sera ensuite utilisé pour comparer avec les toponymes de données et SwissNames toponymie (noms Suisse atlas), les nouvelles corpus sont à la fois de ligne (Sri Lanka Daily Nouvelles en ligne) et les sources publiées (Cameroun Tribune et NZZ). Entrée des toponymes données dans la base de données de nomenclature GeoNames pour Haïti, au Sri Lanka et en Somalie sera également analysé pour voir se il existe un biais dans l'entrée ou à la construction de ce répertoire toponymique données. Une carte de densité de points sera produite à partir des coordonnées dans les données de la toponymie de montrer la distribution spatiale des toponymes sur l'espace, cette carte de densité de points sera également utilisé pour produire une estimation des noyaux de densité pour décrire la distribution des noms de lieux et de regroupement (Cheshire et al., 2012) et pour voir si toponyme et la population ont les mêmes potées par pays, afin de savoir se il ya une relation entre les noms de lieux et de la population, l'analyse de régression sera réalisée entre toponyme par région et la population compte par région pour voir quel genre de corrélation existe.

Les résultats seront analysés sous différents aspects tels que le nombre de noms de lieux ne sont pas disponibles dans la base de données des noms géographiques, mais disponibles dans les journaux et des atlas. Préjugé dans la construction des données de la toponymie sera démontré par un graphique cumulatif de montrer le décalage qui existe dans les données d'entrée des différents pays, carte dot densité de montrer la distribution des noms de lieux dans l'espace, l'analyse de régression sera utilisé pour voir quel genre de corrélation ou relation existe entre la population et les toponymes.

Zusammenfassung

Ortsnamen, auch Toponyms genannt sind Namen von Orten, gebraucht, von den Leuten für die tägliche Identifizierung oder um auf einen bestimmten geografischen Standort zu verweisen. Diese Ortsnamen sind in Büchern, genannt Ortsverzeichnisse, meistens als nicht. Ein Ortsname kann auf verschiedene Standorte verweisen (Mehrdeutigkeit) und als solches muss Vorsicht geboten werden, wenn Ortsname zu einem Standort genommen wird. Ortsnamen und andere Informationen in diesem Zusammenhang werden von Behörden oder Privatpersonen gesammelt und in Bücher oder Unterlagen sogenannten Ortsverzeichnissen gespeichert, es ist ungewiss, ob die Ortsverzeichnisse des Landes die gleichen Angaben enthalten, wie die gleichen Informationen wie die Welt GeoNames Ortsverzeichnisse, die als solche eine Suche sein könnte für einen Ortsnamen im Ortsverzeichnis, finden sie einige und andere finden sie nicht. Ortsverzeichnisse sind Wörterbücher für geografische Namen; „Das grösste freie Lexikon über geografische Ortsnamen pro Raumeinheit zu veranschaulichen sie, das ist ein Platz von einem Zehntel Breitengrad und den zehnten Längengrad für den Globus in Verbindung mit einer Karte oder Atlas verwendet sie auf geografische Informationen“ (Brunner et al., 2008). Die meisten Ortsverzeichnisse sind kostenlos und öffentlich und sie geben die genaue Position eines Ortsnamens auf der Karte an. Es gibt so viele Arten von Orsverzeichnissen, mit jeweils unterschiedlichen Inhalten und sie werden in der Regel für die Geokodierung verwendet (Zuordnung Orte zu Koordinaten), Geotagging, Geoparsing (Erkennung eines Wortes als einen Ort oder Standort) und Georeferencing Informationen von den Nachrichten und Sozial Netzwerk-Strom (Zhang et al., 2014), um Daten zu vergleichen und zu suchen und zu speichern.

. Nationalzeitungen aus drei ausgewählten Ländern (Kamerun, Schweiz und Sri Lanka) werden gelesen und die Ortsnamen in ihrem Text von Hand identifiziert werden; diese Ortsnamen werden verwendet, um eine Gold Standard-Liste, die später verwendet mit den Ortsnamen von GeoNames Daten und SwissNames (Schweizer Namen Atlas), sind die Nachrichten Korpora sowohl online (Sri Lanka Daily News Online) und veröffentlichten Quellen verglichen worden (Cameroon Tribune und NZZ). Eingabe von Ortsnamen Daten in die GeoNames Ortsverzeichnisses Datenbank für Haiti, Sir Lanka und Somalia wird auch analysiert werden, um festzustellen, ob es keine Verzerrungen bei der Dateneingaben oder Bau dieses Ortsverzeichnisses. Eine Punktedichtekarte wird auch von den Koordinaten in den GeoNames Daten erzeugt werden, um die räumliche Verteilung von Ortsnamen auf dem Platz zu zeigen, wird diese Punktdichtekarte auch einen Kernel-Dichte-Beurteilung, um die Ortsnamenverteilung und Zusammenlegung beschreiben produzieren (Cheshire et al., 2012), um zu wissen, ob es einen Zusammenhang zwischen Ortsnamen und der Bevölkerung, Regressionsanalyse wird zwischen Ortsname pro Region und Bevölkerungszahl pro Region durchgeführt werden, um zu sehen, welche Art von Korrelation besteht.

Die Ergebnisse werden unter verschiedenen Aspekten wie die Anzahl der Ortsnamen in den GeoNames Datenbank nicht verfügbar, aber in den Zeitungen und Atlanten zur Verfügung analysiert werden. Als Vorspann in den Bau der GeoNames Daten werden in einem kumulativen Diagramm gezeigt werden, um die Verzögerung, die in den Eingangsdaten der verschiedenen Länder, Punktdichte-Karte vorhanden ist, um die Ortsnameverteilung im Raum zu zeigen, wird die Regressionsanalyse verwendet werden, um welche Art von Korrelation oder die Beziehung zwischen Bevölkerung und Ortsnamen zu sehen.

CONTENTS

List of figuress

List of Tables

Abbreviations

GIS Geographic Information Systems

GDP Gross Domestic Product

IR Information Retrieval

GIR Geographic Information Retrieval

IE Information Extraction

NE Named Entity

NER Named Entity Recognition

MUC Message Understanding Conferences

VGIS Voluntary Geographic Information System

NZZ Neue Zürcher Zeitung

KDE Kernel Density Estimation

DG Digital Gazetteer

1. Introduction

1. 1 Context

The theory of plate tectonics describes the movement of the earth's lithosphere which was explained by the concept of continental drift proposed by a German geologist and meteorologist Alfred Wegener , he hypothesized that, "there was an original gigantic super continent 200 million years ago which he named Pangaea meaning "all the land" in Greek, which consisted all of the earth's landmasses",(Frisch et al., 2010) he said "Pangaea started breaking up into smaller super continents called Laurasia and Gondwanaland during the Jurassic period and by the end of the cretaceous period, the continents were spreading into landmasses that look like our modern day continents[1]". From Wegener's theory we can conclude that the world was once one landmass which was separated by geological processes and is a 'global village' and this phrase has emerged with the help of various scientific research and methods which have narrowed the existing gap between the large spaces that exist on the earth's surface, to what has been illustrated on a map as "a pictured representation of the earth and to help organize our experiences, we categorize the phenomena in various ways and assign names to many of them for convenience" (Hastings, 2008): that is why today we categorize places into streets, cities, villages countries and continents for ease of identification. We can also confidentially say that the "naming" of places came way back from Wegener's era.

This name categorization has made the identification and direction of places much easier as each place has a particular placename accredited to it. "A persistent but not altogether static kind of geographic phenomenon is *place*, a recognizable portion of the world shared across minds, time, and space. The distinction of place from space is at the core of geography" (Hastings 2008), that's where georeferencing comes into play, where the names of places are matched with their location in space and it is a very important aspect of geography because when talking about anything geographic we must include its location: this is inevitable. Though placename ambiguity exits, where multiple places bear the same placename, emphasizing on a particular country for example London Canada and London Great Britain can rectify this, this is termed toponym resolution (Amitay et al., 2004).

[1] http://www2.ametsoc.org/amsedu/online/oceaninfo/samplecourse/oceanchap2.pdf

Majority of placenames have ties with renowned people, rivers and other prominent features attached to the places, the names are not formulated and the naming of the places comes automatically, places are named by tourists, high government officials or organizations. "Some places are named after people for example Washington DC is named after George Washington, events and even prominent features for example Cameroon got its name from Portuguese word *Rio dos Cameroes* in 1472 meaning *Shrimp River*[2]" which was given by the Portuguese explorers which became Cameroon in English. "Naming a place is also a spontaneous cognitive and socially situated linguistic act, part of the naive geographic behavior that people share" (Egenhofer and Mark 1995): people name places, naming a place is not a planned process, most places have been named abruptly from features surrounding them, influential people in that area or certain events that had happened there like eruptions, floods and fire.

In order to recognize all the variables of a map they have all been allocated names termed "placenames" in geography. "Proximate phenomena are directly experienced in our sensory systems and selectively assimilated into our mental maps; distant phenomena reported by others and/or remotely sensed also may be captured in those maps" (Hastings 2008).We automatically store the names of those places in our memories; those which are discovered and named around us or named by other people from distant and unknown areas are also recognized as well. Place of the emergency can only be located if the placename is identified and correctly. GIS use gazetteers to trace locations on the map as such placenames play an essential role in emergency services and makes the rescue easier and faster (Burenhult and Levinson 2008).

Following Hill 2000, "a geographic place is defined by: at least one placename by which it is commonly known and communicated; at least one place type situating it in a classification scheme that also provides conceptual/symbolic bases for it; and finally at least one footprint, georeferenced geometry, locating it in the real-world and optionally designating its areal configuration", a footprint may be a point, line, bounding rectangle, or other polygonal shape (Jones et al., 2003) these are also characteristics of what kind of information one can find in a gazetteer; placename, its place type (if it's a park, building city or cliff), the X and Y coordinates giving it a point location on the map. "Gazetteers provide the translation between place names and coordinate footprints" (Hill 2006) for example if you search for Dolder,

[2]http://en.wikipedia.org/wiki/List_of_country-name_etymologies

Switzerland from the GeoNames Database you get this information: Dolder Grand Switzerland, Zurich-Hotel- N 47° 22´ 19´´ E 8° 34`30'' it gives the place type and foot print. Placename and footprint correspond directly to the first two of four spatial primitives noted by (Golledge 1995), which he calls identity and location. "Place type serves to anchor a place categorically within the human experience of geographic places generally, i.e. medium-scale objects locked to the surface of the Earth that exhibit conjoint metrological and topological connectedness" (Mark et al 1999).

"Placenames aka feature name" (Hill 2006, p.91) are used on daily basis for communication irrespective of age, sex, education and even culture, formally when writing someone's house address or informally when talking about a holidays or arranging the location of an event, as such placenames are inevitable in our day to day activities, we use them sometimes unconsciously and even wrongly. "Placenames are used in conversations, correspondence, reporting and documentation" (Hill 2006).

"Dictionary of placenames are called gazetteers" (Hill 2006 p.91). For some georeferencing applications, e.g. Geoparsing (Hill 2006 pg.100), "the simple existence of a place, confirmed by a placename, is of first importance; place type and footprint details are secondary". Of all the information found in a gazetteer, the most important of them is the placename before finding out if it's a park, hotel or school then later finding out its coordinates: placenames are very important. Similarly, in way finding: during trips or explorations, people are quite tolerant of approximate footprints or they could get lost because it seems that we continually conflate them mentally either due to ambiguity or existence of too many and this can lead to wrong locations (Montello 1998).

In former times, news was written on papers using ink and these were called newssheets, but this changed since the 17thcentury. "When we think of newspapers, we think of them as bringing us news: when we think of news we think of what's happening currently around us. This is a totally inadequate description of news as well as newspapers. A newspaper is not only a source of information; it's a store house of information[3]" Newspapers are very important in many ways; its importance in giving valuable information still holds till date , they give us lots of detail information about various happenings at different levels that are not available anywhere.

[3]https://dhanushka1996.wordpress.com/2010/11/14/newspaper-as-a-source-of-information

anywhere. Newspapers are very important in many ways; its importance in giving valuable information still holds till date, they give us lots of detail information about various happenings at different levels that are not available anywhere. Though magazines and books also give lots of information, they are at a small scale, time consuming to produce and might release the newsfeed later than the internet and TVs but the latter cannot have the details and depth of the newspapers and don't have the possibility of reading it a second time . Newspapers are data banks, people learn English Language Grammar just from reading newspapers, people learn new words including names of places around the world and they see job opportunities and other information on the newspapers.

The existence of an incomplete gazetteer can have a chain of effects and many things go wrong, many placenames and very local placenames for example the name of a clearing in the forest or the shoulder of a mountain are under looked and omitted, Some placenames are not available in the administrative gazetteers of countries, maybe because the feature is too small to be names on the map, or if it is known by more than one name (Piotrowski et al., 2010), people will not be able to know about new places they come across while reading or listening the news, some might locate wrong places instead of booking a flight to Sydney, Australia; you could mistakenly book to Sydney, Nova Scotia, rescue services won't be able to help in cases of emergencies as well. Goodchild 2007 points out that, in the case of an unknown or unclear emergency site, precious minutes will be lost trying to determine and locate the unambiguous location of the incident.

1.1.1 Placenames Ambiguation

The way Toponyms are spelled or punctuated differ between documents especially with the existence of variant names of some places (Hill 2006) as such, drawing attention to the fact that newspapers are large data banks for names of places, when we read through newspapers and we come across new placenames, it is but normal to want to find out where it is located and also other information on the places like its population, capital and even its GDP, sometimes we don't find them in the search engines and sometimes we find them ambiguated (multiple locations having the same placename).

Geotagging involves arbitrating two types of ambiguities: geo/non-geo and geo/geo. A geo/non-geo ambiguity as described by Amitay et al., 2004 are toponyms that do not only

describe geographic places but names, when a placename also has a non-geographic meaning, such as a person's name (e.g. London) or a common word (Turkey). Geo/geo ambiguity arises when distinct places have the same name, as in London, England and London, Ontario. This causes confusion and problems such cases because in the case of geo/non-geo ambiguity, the word that appears in the text might refer to someone's name and not a placename while in the case of geo/geo ambiguity the placename might refer to another location.

Toponym resolution is concerned with finding out which concrete instance of a name is meant in a document and assigning it to the right pair of coordinates, and this is the solution to toponym ambiguity; various methods have been discovered by multiple authors on Toponym resolution (Brunner 2008)

1.1.2 Vernacular Geography (Language & Dialect)

Some place names have one or more variant names and these are different versions of the placenames in different languages, and most of the time the inhabitants of that location use the variant names especially those in the dialect version, some of them have never used the official name of their location while referring it to someone.

Sometimes we search the names and we cannot find them on the search engine or we find them but in different versions. Spellings and punctuation of placenames can vary between documents (Hill 2006). For example, Zürich can be written as Zürich. People searching have to deal with the possibility of different spellings, accents and hyphens when searching a placename. Spelling variations are closely connected to the language and dialect problem, but the most important thing is to have knowledge of the official name of the area.

1.1.3 Modifications over time

As language and spelling evolve through time, so do the names of places (Hill, 2000). Old versions of names are not simply forgotten especially by the aged population group, the transition period between the old name and the new name makes both versions useable at that moment. "Geographical renaming is the changing of the name of a geographical feature or area. This can range from the uncontroversial change of a street name to a highly disputed change of the name of a country. Some names are changed locally but the new names are not

recognized by other countries, especially when there is a difference in language, Sometimes a place reverts to its former name[4]".

" One of the most common reasons for a country changing its name is newly acquired independence, when borders are changed, sometimes due to a country splitting or two countries joining together, the names of the relevant areas can change. This, however, is more the creation of a different entity than an act of geographical renaming. A change might see a completely different name being adopted or may only be a slight change in spelling for example Peking was changed to Beijing[4]"

This is a very important and altogether complex issue because some placenames change over time but the old versions of the name are more in use especially for official matters, this also depends on the age group for example my home town was formally called Victoria and then later changed to Limbe, the older aged group more often than not refer to the city as Victoria unlike the youths who stick to the current name; this also causes the city to be referred as both Victoria and Limbe even on official documents. Change of placename over time leads to the automatic creation of variant names.

1.2 Motivation

To raise awareness on the importance and use of Gazetteer data as this facilitates the process of data dissemination from various sources, an incomplete gazetteer makes this process very difficult and frustrating to the searchers. There are also very few studies and knowledge about the Gazetteer data: Brunner (2008), Kunz 2008 and Dirk Ahlers (2013) .It is evident from the GeoNames map that the placenames are not evenly distributed among the regions and this is also due to the fact that crisis mappers concentrate on countries with political and/or social instability, as such "we see a much higher proportion of placenames in Europe, and a lower density in Africa and Asia" (Brunner et al., 2008)[5] this can be seen in Figure 1.1 below.

"Information overload (also known as infobesity or infoxication) refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information" (Yang C et al., 2003). This is no doubt information including people, places and things; as such an inefficient gazetteer data makes it difficult for information retrieval about people, places and events.

[4]http://en.wikipedia.org/wiki/Geographical_renaming
[5]http://geography.oii.ox.ac.uk/?page=mapping-the-geonames-gazetteer

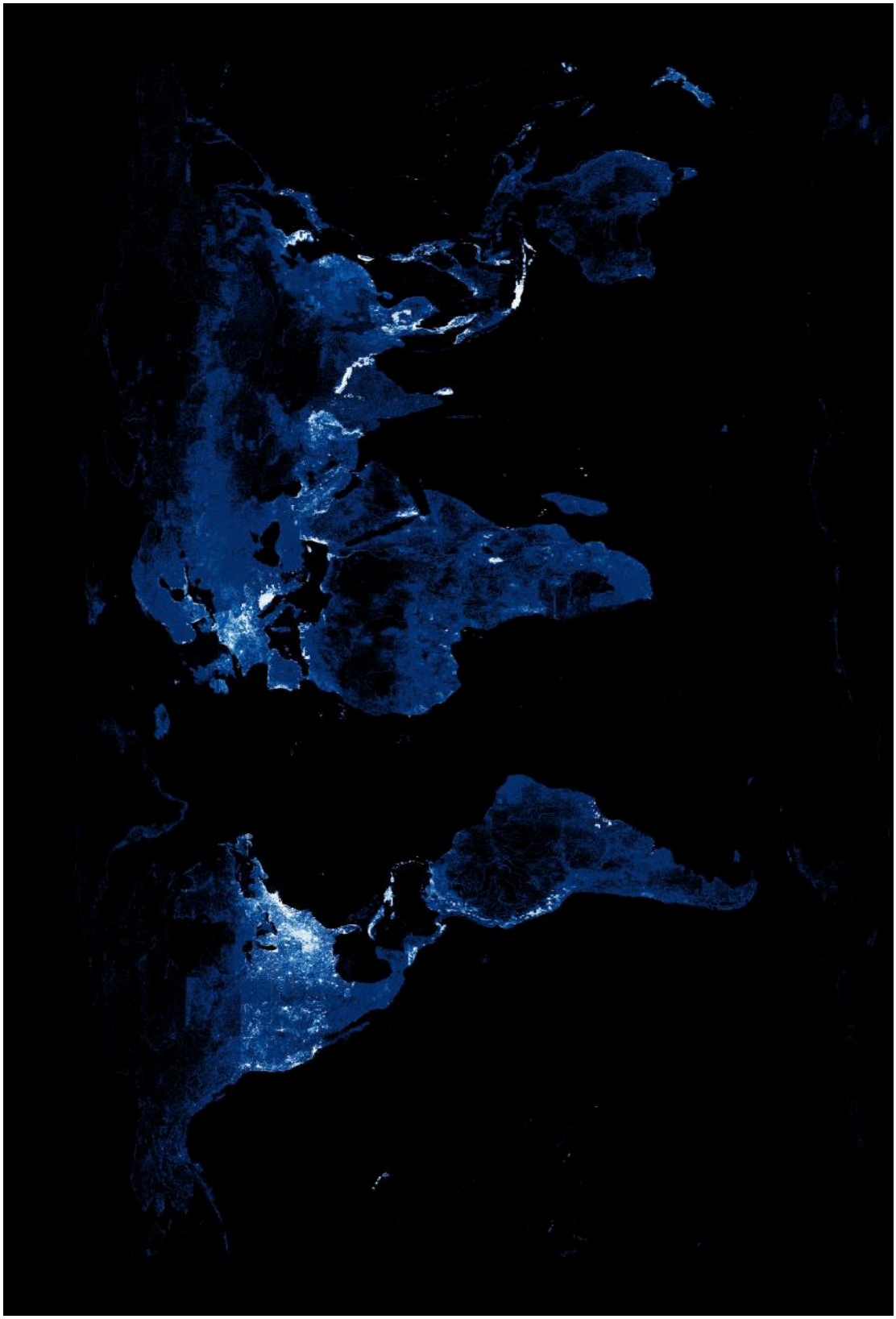


Figure 1.1 Mapping the GeoNames Gazetteer

Source: Oxford Internet Institute

Due to increase in online and published information, which also includes information extraction and retrieval. One problem amongst others is that most place names, and the vast majority of names found on the Web, are ambiguous (Amitay et al, 2004), this ambiguity calls for rectification (toponym resolution) somehow, these placenames can only be verified and rectified somewhere, and one of such places is from the World GeoNames Gazetteer, thus, attention has to be given as to the completeness and reliability of this gazetteer data , the searcher should be able to gain meaningful information after haven visited it.

“More recently, the specific task of determining which place is meant by a particular occurrence of a place name, known as grounding (also referred to as localization), has been gaining attention” (Amitay et al, 2004) following the recent Global climate change trend, we are been faced with numerous natural disasters and as such rescue services need to know exactly where to go to in case of an emergency call and the rest of the world needs to know and have vital information about the said affected area, some placenames have alternate spellings and this confuses the reader or researcher; with an inefficient gazetteer data this process is difficult.

“The past few years, awareness on the naming and description of places have increased and taken on new meaning and importance with the growth of user-generated content on the Web. Sites such as Wikipedia, Wikimapia, and Flicker allow private citizens to create and integrate descriptions of interesting places and to build links between them using standard codes known as Geo-tags”(Goodchild and Hill 2008). People who are not aware of the existence or location about places they read in these sites need to search and know more about them. Sometimes the search engines don’t have what they search for and they have to turn to precise online libraries and books like the gazetteer. As the web continues to grow, having more users and more information, major general-purpose search engines have been faced with serious problems, they are unable to index all the documents on the web, because of the rapid growth in the amount of data stored and the number of documents that are publicly available, this affects the quality of results one gets form the search engine (Barfourush et al., 2002)

1.3 Hypothesis

The goal of this thesis is to find out if there is a spatial pattern in the construction (placename input) and use (availability) of the GeoNames gazetteer data. The following hypothesis will be used as guidelines and the hypothesis stated will be tested.

Hypothesis One: The GeoNames gazetteer data is representative and contains all placenames present on the country landscape

Using data from the GeoNames gazetteer, the coordinates of the placenames will be used to create a dot density map, from this dot density map we will see the spatial distribution of placenames in space for the three chosen countries, we want to know if the placenames are evenly distributed spatially, are all the areas covered with placenames or do we see empty spaces on the dot density maps from the GeoNames data. From a glance do we have the same patterns for the three countries? Or do we see more points of placenames for some countries and large empty spaces for some; do we see high density of placenames points in some and no placename points at all in some? From the results I will know if there are more placenames present for some countries in the gazetteer than others.

Hypothesis Two: All existing placenames for the various Countries are registered on the GeoNames Database.

The GeoNames gazetteer, which hypothetically represents placenames from around the world, is expected to be complete, reliable and up to date. Newspapers are data banks and convey large amounts of information from around the world, so I will use placenames present in daily newspapers to verify if they are on the GeoNames database. Taking into consideration country atlases, comparisons will be done to find out if all the placenames on the Switzerland country atlas (SwissNames) present on the GeoNames Gazetteer.

Hypothesis Three: There is a consistent and unbiased input of the placenames into the gazetteer (its construction).

How is the pattern of placenames input into the gazetteer? Is it consistent or are there large gaps of input data time? Is the input tied to aftereffects of certain events? Is there a large existence of placenames input after environmental disasters or political instability? On the normal and peaceful state what’s the placename input and what’s the difference in input after social unrest?

Hypothesis Four: Placenames are related to people: they both have a strong correlation.
Is there any relationship or strong correlation between placenames and population? Places are being named by and after people and are occupied by people; are there more people where there are more placenames? Are areas with dense population having agglomeration of placenames?

1.4 Structure

Chapter One: Introduction
Chapter Two: Current state of Research
Chapter Three: Data
Chapter Four: Methodology
Chapter Five: Results and Interpretation
Chapter Six: Discussion
Chapter Seven: Conclusion

2. Current State of Research

This chapter gives an overview of the areas of research which are relevant to my thesis. The significant terminology is explained and the important literature is briefly summarised. After a brief overview of this chapter in section 2.1, an introduction to Information Retrieval (IR, section 2.2) and Geographic Information Retrieval (GIR, section 2.3) the link is made to Information Extraction (IE, section 2.4) and Name Finding (section 2.5). A detailed explanation of a Gazetteer in section 2.6. IR holds the subfield GIR while Geoparsing is a GIR subtask and corresponds to NER (Name Finding) for toponyms in the field of computational linguistics. NER is a subtask of IE. The focus of this thesis lies on placenames and the completeness of the World GeoNames gazetteer data, which leads us to section 2.7 on Placenames their types, and origins. Newspapers are a data bank for placenames as such we need to examine what they are, their origin , frequency and geographical scope in Section 2.8.

2.1 Overview

Manning et al. (2008, p.1) defines IR as “the finding of material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)”, that is information gotten from computers; people upload and store masses of information in the online and when others are in need of these information, it has to be retrieved from the loads, most of the time they get more than one reply from their query, some having detail of what they search for, others are scanty and sometimes they don’t . Figure 2.1 below describes the link of all the processes used in this thesis.

IR involves web search, text classification and text clustering (Manning et al., 2008) these could be information about people, places and events, this involves a user entering a query into a system or web search engine and the web search engine will in turn give multiple replies depending on the words you type into the search engine.

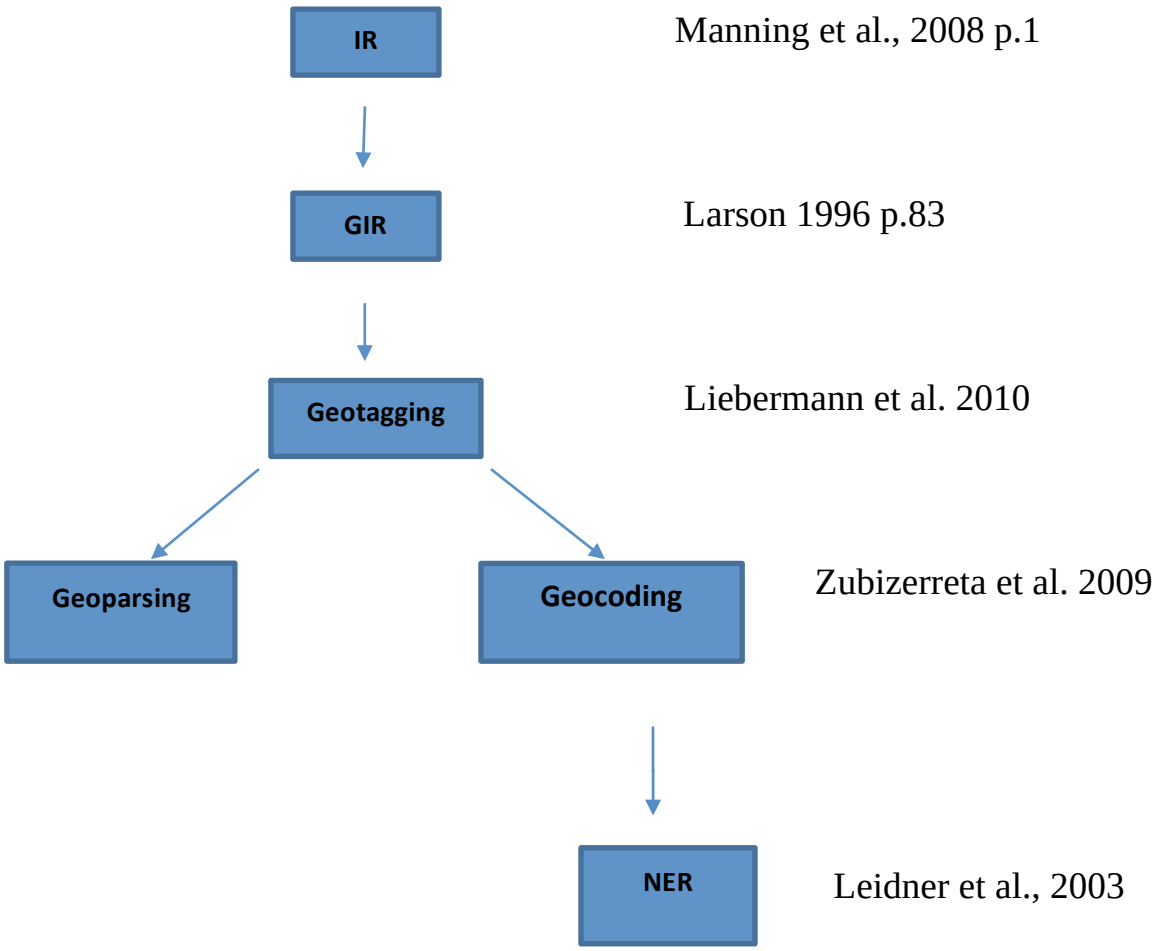


Figure 2.1: Scientific background showing relationship between IR, GIR & Name Finding

According to Kunz (2008), the first definition of GIR was the one given by Larson (1996, p.82) he said: "Geographic Information Retrieval is concerned with providing access to georeferenced information sources. It includes all of the areas that have traditionally formed the core of IR research with an emphasis or addition, of spatially and geographically oriented indexing and retrieval". GIR covers a smaller and more precise scale as opposed to IR, and GIR systems can commonly be broken down into the following stages: Geotagging, text and Geographic indexing, data storage, geographic relevance ranking and browsing results (commonly with a map interface) but this thesis lays emphasis on Geotagging.

According to Lieberman et al., 2010, Geotagging is the identification of geographic terms or words in documents and associating these terms with appropriate geographic locations. Zubizerreta et al., 2009 further divided Geotagging into two important GIR tasks often referred to as Geoparsing and Geocoding. Geoparsing, also known as toponym recognition according Lieberman et al., 2010, is the process of recognizing placenames in a text and linking then to geographic identifiers, one difficulty Geoparsing contends with is ambiguity.

In computational linguistics, Geoparsing is called NER (Leidner, 2007); NER is not only concerned with the identification of locations, but also with finding references to organizations and persons in a document. Geoparsing, or NER for locations, is the focus of this thesis. Since the techniques used to approach this GIR task originate in the field of NER (Brunner, 2008), the topic will be approached from the NER point-of-view. Geocoding on the other hand involves the insertion and enriching of a placename or image with geographic coordinates (Valli & Hannay 2010)

2.2 Information Retrieval

Back in the 90s people preferred to get information from their friends rather than from information retrieval systems, but during the last decade things have changed, this is thanks to the relentless optimization of information retrieval effectiveness which has made it a reliable source of information collection (Manning et al., 2008). Information retrieval can be locally said to be a process of removing or extracting facts from a document or set of files, this could be done by hand or using machine/computer based techniques: this involves going through the whole text, selecting and sorting the necessary information needed.

IR emerged in the 1950s in response to the ever growing amount of digitally accessible data: this data had to be used by several people and they method of extracting this information (Singhal 2001) IR did not begin in the web, it began with scientific publications and library records but soon spread to other forms of content and it was only relevant to several specialized professions such as librarian's and information experts, this changed with the arrival of the WWW in 1990s. A principal driver to this was the creation of WWW by Tim Berners-Lee a British computer scientist and a Belgian scientist Robert Cailliau on March 12, 1989[6].

This revolutionary development brought IR into the spotlight, "efficient access to relevant information became a necessity for lay people as well" (Baeza-Yates and Ribeiro-Neto, 1999). We have two kinds of information retrieval: text based and image based. To improve IR results, search queries are often refined by placenames (Gan et al., 2008); for instance, if you are looking for a restaurant to eat from the web, typing in restaurant is not enough, by

[6]http://www.wikipedia.org

adding the name of the locality like Zurich makes the search much more precise. WWW retrieves information about people, places and things; a user enters a URL into a browser (for example, http://www.google.com). This request is passed to a domain name server (Barfourosh et al., 2002) when we search for information from the web, we receive varying results depending on what command we give in as seen from Tables 2.2 below: the search engines database and manner of search affects the quality of the results, if you are using chrome, Firefox of explorer the results will vary.

Evaluation Parameter	**Description**
web pages Ranking methods	Different parameters used to specify the rank of web pages in returned result list, such as site popularity,
Various display option	If various options are available to rank the returned result, such as by date, by site,
Suggested search	Suggestions for further searching based on the initial search are provided or no. These suggestions can be simple, such as synonyms or alternative search terms, or may be more sophisticated, such as suggestions for searching in different, specialized databases.
Similar searches	If someone locates a web page that is highly relevant to his research issue, It might be interested in finding more pages that are very similar, is it available?
Translated results	Possibility of offering a tool to translate a given result page from one language to another.

Table 2.1: Evaluation parameters of search engines from the perspective of returned results.
Source: Barfourosh et al., 2002

2. 3 Geographic Information Retrieval

"Geographic information retrieval (GIR) or geographical information retrieval is the augmentation of information retrieval with geographic metadata[7]". "Generally, GIR is split into four stages: Information Extraction, Disambiguation, the User Interface and Information Storage" (Overell et al 2006). There is a thin line in between IR and GIR and this is the fact that GIR is specific with locations, there is more beyond just retrieving the information: when you extract what you need from the text or file, you must also identify the particular information: the name of a place, the place type or the coordinates and then the information is filtered and stored.

[7]http://en.wikipedia.org/wiki/Geographic_information_retrieval

According to Kunz (2008), the first definition of GIR was the one given by Larson 1996, p.83: "Geographic Information Retrieval is concerned with providing access to georeferenced information sources. It includes all of the areas that have traditionally formed the core of IR research with an emphasis or addition, of spatially – and geographically -oriented indexing and retrieval".

"Information retrieval generally views documents as a collection of words. In contrast, geographic information retrieval requires a small amount of semantic data to be present (namely a location or geographic feature associated with a document) it is common in GIR to separate the text indexing and analysis from the geographic indexing[8]".

(Purves and Jones 2008, p. 375) defined GIR as "the provision of facilities to retrieve and relevance rank documents or other resources from an unstructured or partially structured collection on the basis of queries specifying both theme and geographic scope" from this definition, the distinction between geographic information retrieval and geographic data retrieval have been brought to light and they emphasized the relatively unstructured nature of the documents.

Purves and Jones also identified several challenges, which GIR is still faced with today:

• "Identification of geographic terms in documents and associating these terms with appropriate geographic locations;

• Ways of indexing large collection saliently for search on both thematic and geographic content.

• Development of search engines and algorithms which can exploit such indexing systems;

• Techniques to combine geographic and thematic relevance in appropriate ways;

• Methods to allow users to formulate queries to such search systems; and

• Design of interfaces and visualization's which allow users to effectively explore and assess returned document sets".

From the above challenges, it is evident that geographers have not attained an ideal GIR system; it still has difficulties in ambiguity, search engines that can index all the available documents in very organized manner such that the documents are easy to retrieve from the search engines.

[8]http://en.wikipedia.org/wiki/Geographic_information_retrieval

The main challenge in Geocoding is dealing with ambiguous toponym that is placenames which can refer to various locations. This problem is called geo-geo ambiguity. Geotagging is composed of two tasks; Geoparsing and Geocoding. (Zubizarreta et al., 2009) .Geoparsing aka toponym recognition is the process of recognizing placenames in a text, it could also be normal words or words which are used as toponyms, in computational linguistics Geoparsing is called NER (Leidner, 2007), it is not only concerned with identification of locations but also finding references to organizations and persons in a document.

Geocoding can be referred to as toponym resolution (Buscaldi and Rosso 2008) which is focused on assigning the correct geographical coordinates to the toponyms recognized during Geoparsing. The main challenge of Geocoding is dealing with ambiguous Toponyms. This will not be treated in depth because this is not the focus of the thesis.

GIR in a nutshell draws us to the fact that when searching information on places, their geographic locations or footprints are very important especially when it comes to ambiguity, the ideal solution to this is getting the precise coordinates of the place you are searching; GIR gives us more precise results.

2.4 Information Extraction (IE)

“IR can be improved with the help of IE” Bear et al., 1998, as such IE can be said to be a sub of IR, they might seem same but are not, IR identifies the text or file having the information while IE does the selection, sorting and filtering of the information according to need . Sarawagi 2008 p.263 defines IE as “... the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources”.

“IE therefore involves the creation of a structured representation (such as a database) of selected information drawn from a text, it also includes restructuring and reducing information from a document” (Grisham 1997).

Grisham 1997 sees information extraction as any method for filtering information from large volumes of texts tagging of particular terms in texts and creating a structured representation. There has been growing interest on developing systems for information extraction. Table 2.5

shows information extracted from a newspaper after said bombing by terrorists in El Salvador

INCEDENT TYPE	bombing
DATE	march 19
LOCATION	El Salvador: San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

Table 2.2 A terrorist report and a template of information extraction
Source: Grisham 1997

Most of the world's news for example is reported in newspapers, radios and TV broadcast .Info extraction has the potential of extracting data with much more precision from such text and this is the focus of one part of this thesis; extracting placenames from newspapers(Grisham 1997)

Figure 2.3 gives us Guidelines or steps on how to extract information from a document but for this thesis only few steps are of importance. I will use just two steps for my research which are:

Name recognition: Identifies various types of proper names and other special forms such as dates and currency amounts, names appear frequently in many types of texts and identifying and classifying them implies further processing: personal names, company names, placenames.

Template Generation: Here we will have a list generated instead of a template, a list of placenames that appeared on the newspapers

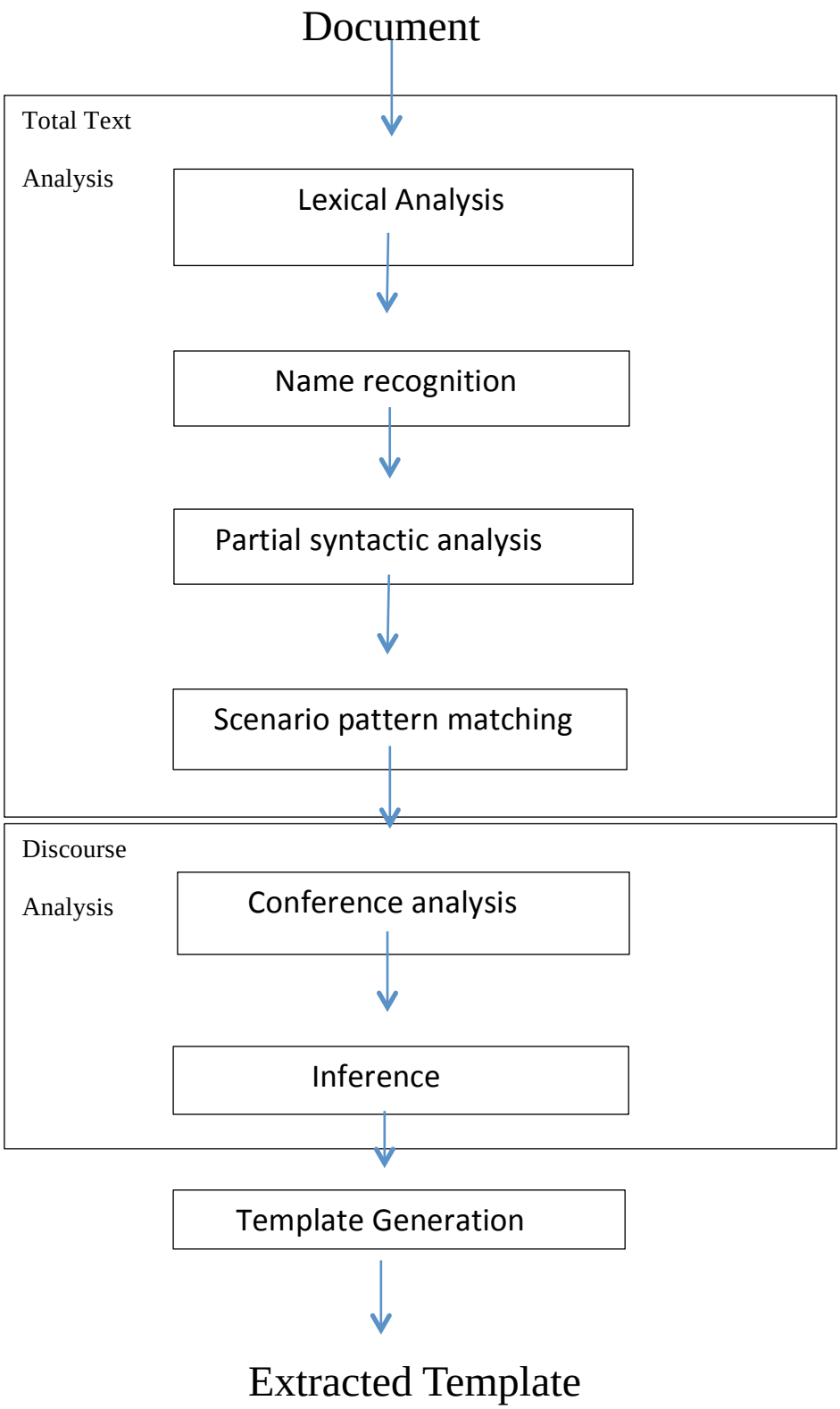


Figure 2.2 Structure of an Information Extraction System
Source: Grisham1997

2.5 Name Finding (NER)

Named Entity (NE)

The MUC-6 coined the term NE and it has since remained an expression referring to automatic language processing (Rossler, 2007). For the NE task, the MUC: Message Understanding Conferences (MUC-6 Appendix, 1995) defined NEs as proper names, acronyms, and perhaps miscellaneous and other unique identifiers, which are categorized as follows:

ORGANIZATION: named corporate, governmental, or other organizational entity (e.g. Novartis, UBS)

PERSON: named person or family (e.g. Ursula Andres, Professor Euler,**)**

LOCATION: name of politically or geographically defined location (cities, Provinces, countries, mountains.)

Rossler, 2007 stated that there is still an incomplete description or definition of NE. Mikheev et al. (1999) and Rossler (2007) settled on a more open definition and suggest that what an NE is depends on the context it is used in. Most of the literature, however, seems to adhere to the original definition stated in the MUC-6 (Nadeau and Sekine, 2007). Likewise, in this thesis the term NE will be used to refer to names of places recognised from newspapers.

Named Entity Recognition (NER)

This is an IE subtask and is described as the finding and classifying of all expressions in a document and could be assigned to one of the following seven categories: organisation (e.g. UN), person (e.g. Monique Ndam), location (e.g. Zurich), time (e.g. four o'clock), date (e.g. 1st October), monetary amounts (e.g. e 5.20) and percent expressions (e.g. 99%) (MUC-6 Appendix, 1995).This can also be seen in Figure 2.3 which illustrates all the word features we can find in a sentence but this thesis is focuses on names of places which begin in capitalised words.

The NE task can, ultimately, be described as recognising NEs, temporal expressions and expressions of quantities in a text. Hence the NE task is also called Named Entity Recognition (NER). In the words of Kozareva 2006, p. 15 “NER consists in detecting the

most silent and informative elements in a Scientific Background text such as names of people, company names, location, monetary currencies, date".

Word Feature	Example text	Intuition
twoDigitNum	90	Two-digit year
fourDigitNum	1990	Four digit year
containsDigitAndAlpha	A8956-67	Product code
containsDigitAndDash	09-96	Date
containsDigitAndSlash	11\09\89	Date
containsDigitAndComma	23,000.00	Monetary amount
containsDigitAndPeriod	1.00	Monetary amount, percentage
otherNum	456789	Other number
allCaps	BBM	Organization
capPeriod	M.	Person name innitial
firstWord	First word of sentence	No useful capitalization information
InitCap	Sally	Capitalised word
Lowercase	can	Uncapitalised word
Other	,	punctuation

Table 2.3 Word features, examples and intuition behind them
Source: Bickel et al., 1997

Since the NE term is not clearly defined, there logically is also a lack of consensus regarding the concept of NER. Some scientists like Kozareva adhere to the original NE task as it was laid out by the MUC-6. Others, like Chieu et al., 2003 reduce NER to four categories: Person, Organization, location and miscellaneous, (Rossler 2007) considers only the MUC-6 entities - that is organisations, persons and locations.

NER is a necessary building block for many IE tasks, such as the mentioned template relationships and scenario template tasks: before answering questions concerning the relationship between NEs or the content of a document, it is necessary to recognise the NEs themselves. Being a prerequisite for IE tasks, NER is consequently also of significance for effective IR (Mikheev et al., 1999). Sekine and Isahara 1999, p. 1 call NER one of the basic techniques in IR and IE.

There are various methods used to perform the task of NER. Two main approaches can be distinguished: the rule-based and the machine learning based approach (Rossler, 2007).

As mentioned above, Geoparsing in computational linguistics is called NER, but since in this thesis I won't be using natural language processing techniques, I will recognize the names from the text manually so I will be using the term Name Finding,

2.5.1 Rule based approach

This approach to has many sub sections but I will lay emphasis on the list based approach

List based approach

Is considered the simplest way of accomplishing NER such as using first names, surnames, companies or locations. A list of locations is usually called a *Gazetteer*, when GIR and NER are concerned with toponym recognition the List-based approach is referred to as the *gazetteer look up approach* (Brunner 2008) it is the oldest method to detect geographic names in text (Jones et al., 2001), the content of the list is also a big challenge in this method: what names are there, how many names are there or should all the variant names be included?. Another problem in this method is the actual detection of the names in the text, small variations such as change in grammar cases might cause the NE to be confusing, list based approach has many drawbacks but however, lists such as gazetteers are still useful tools when applied in combination with other approaches.

2.6 Gazetteer

"A gazetteer is a list of geographic names, together with their geographic locations and other descriptive information[9]". "A geographic name is a proper name for a geographic place or feature, such as Washington, D.C, Gulf of Mexico, Central High School, and Southern France. Geographic places and features include political and administrative areas (e.g., cities, counties, and countries), natural features (e.g., mountain ranges, lakes, and canyons), manmade structures (e.g. Buildings, bridges, and canals), and imprecise areas like Southern California[10]".

"Gazetteer data exist in toponymic authority files like U.S. Board on Geographic Names gazetteers (U.S. Board on Geographic Names, 1998), in published gazetteers like the New York Times Atlas (Mackay, John Bartholomew and Son, & Times Books, 1992), in thesauri

[9]http://www.dlib.org/dlib/january99/hill/01hill.html
[10]http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.39.5867&rep=rep1&type=pdf

of geographic names like the Getty Thesaurus of Geographic Names (Getty Information Institute, 1997), as biogeographic and physiographic regions like U.S. watersheds (Environmental Protection Agency, 1999)" Hill 1998

All named geographic places no matter their size and content are proper features to be represented in a gazetteer (Hill &Zheng 1999). "A digital gazetteer (DG) is a spatial dictionary of named and typed places in some environment, typically the near-surface of the Earth. DGs are proliferating in number and sophistication with the popularity of location-based" (Hastings 2008)

"The essentials of a DG remain: place name, place type, and footprint. The essential utility of a DG is to translate between formal and informal systems of place referencing, i.e. between the ad hoc names and qualitative type classifications assigned to places, on the one hand, and quantitative locations for them, on the other" (Hastings 2008). "The digital gazetteer provides a spatial dictionary of named and typed places in some environment, typically the near surface of the earth" (Hasting 2008)

2.6.1 Origins of gazetteers

"The term gazetteer was originally applied to one who wrote a gazette. It was first used in its modern sense early in the 18th century after the publication (1703) by Lawrence Echard of the Gazetteer's or Newsman's Interpreter, a geographical index. But lists of placenames, with descriptions, had been made as early as the 6thcentury: part of the gazetteer of Stephen of Byzantium, of this time, is exact. The 19th century when geographical knowledge and the need for having geographical facts readily available had both increased greatly, was the great period of development of gazetteer making. Attempts were made to produce complete gazetteers, necessitating several volumes. Famous gazetteers include Johnston's (Scotland, 1850), Blackies (Scotland 1850) Bouillets (France 1857) Ritter's (Germany 1874) Longmans (England 1895) Garollos (Italy 1898) and Lippincotts (United states 1865; now The Columbia Gazetteer of the world, 1998). Later editions of many of these have appeared".(Hill 2006 p.91) Today the model of gazetteer has changed significantly from its early manifestations as printed alphabetical or hierarchical lists of placenames with associated information to those containing more detailed information including place type and footprints.

2.6.2 Uses of Gazetteer

Gazetteers like any other kind of dictionary are used for reference and verification of information which is closely related to placenames. When people come across new placenames on newspapers, adverts or while talking to others, they want to find out to know more about these new places and they can get that only from gazetteers.

Beyond this minimum set of information for each place (name, footprint, and type), a gazetteer can:

- Provide variant names for the same location for example Zurich and Zürich.
- "Trace historical changes in names and in spatial footprints: giving all the names of the place from former times till the current name" (Hill 2006).
- "They play several roles in information management, for info retrieval; they provide translation between informal and formal means of georeferencing e.g. Geoparsing- where placename references in text are looked up in a gazetteer so that geospatial coordinates can be associated with the text" (Hill 2006).
- Contain variant spatial representations for the same feature (a point location, bounding box, detailed boundary, etc. or spatial footprints from different sources) and measurement details.
- Link to (or include) other types of information about the feature (physical dimensions, history, description, Population, etc.)
- Combine information from various sources together for one entry.

2.6.3 Basic & Core Elements of a Gazetteer

The core elements of a gazetteer according to Hill 2006 p.107 remain:

1. Name which could include variant names too.
2. Class/type which could vary according to categories
3. Location: geographic units

There are three other important elements of gazetteers Hill 2006 p.119

1. Relationships between the named places e.g. administrative hierarchies
2. Temporal ranges/aspects of gazetteer data
3. Attribution: documentation of sources of info about a place

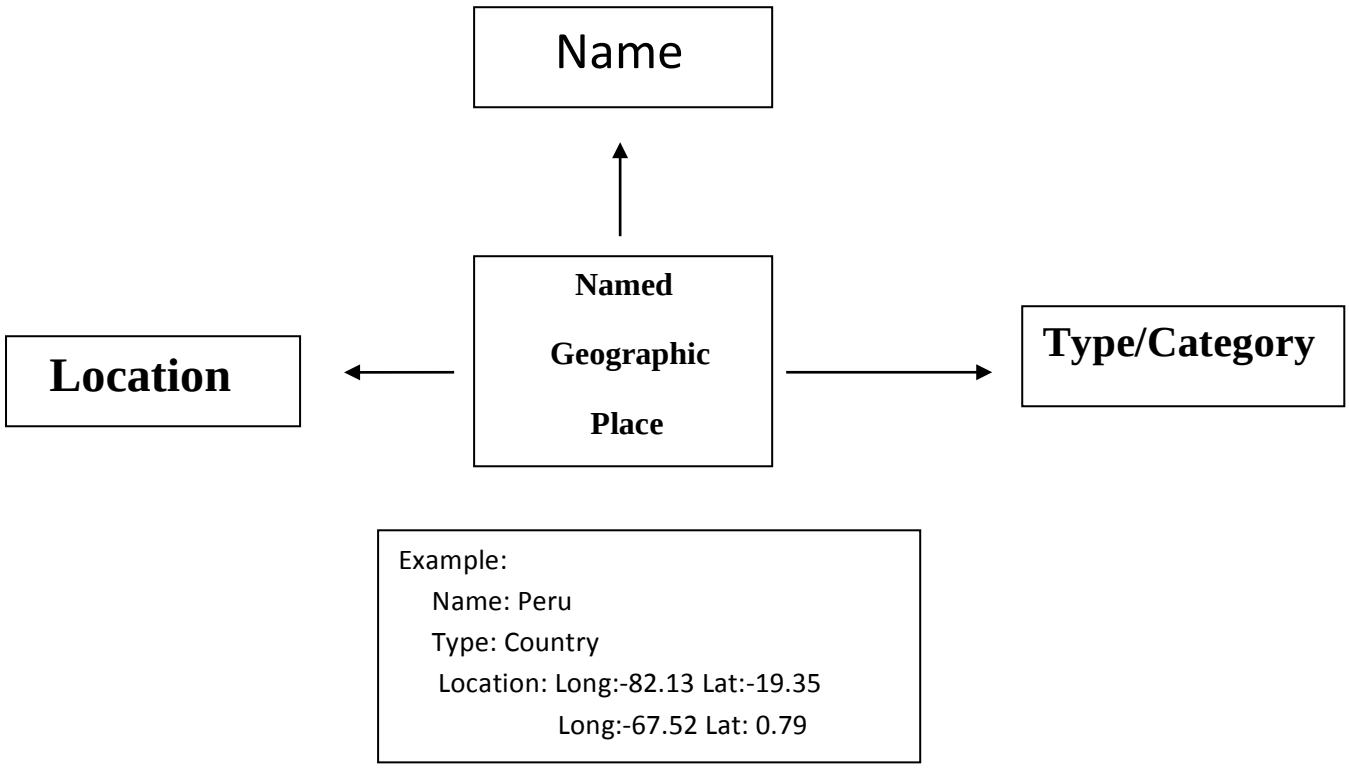


Figure 2.3: Basic components of an entry in a digital gazetteer for a named geographic place
Source: Hill 2006 p.92

Modern digital gazetteers play a central role in searches with a geographic component (Davies et al., 2009) In the context of a GIS, Hill 2006, p. 97 describes gazetteers as translators "between formal and informal means of georeferencing" meaning they supply an unambiguous geographic location (coordinates as an example of a formal means of Georefencing) for every placename (informal means of georeferencing).

Vestavik 2004 has pointed out that ,in the past years, a need for more advanced gazetteers has surfaced, gazetteers should, for example, be capable of dealing and solving the problems attached to vernacular geographic terms (Hollenstein and Purves, 2010), give information about spatial and semantic relationships between places (Vestavik, 2004) and be customizable and support time (Goodchild, 1999).

2.6.4 Sources of gazetteer data

There are numerous sources of data about named places available for building new gazetteers and adding to existing gazetteers.

- "Toponymic authorities such as the U.S Board on Geographic Names (BGN) and the Geographic Names Board of Canada are primary sources of gazetteer data. They are tasked with standardizing placenames of governmental purposes from, from the country/nation level to local government councils. They use set of rules to standardize the form of official names and to exclude some names that would be offensive or otherwise unacceptable". Hill 2006 p.103

- "Other governmental agencies also collect and create gazetteer-type data as a result of community planning zoning, the legal recording of property ownership and community infrastructure building, these are the agencies that were the early adopters of GIS technology". Hill 2006 p.104

- "Creators and publishers of maps, both governmental and private hold vast quantities of gazetteer data in the form of labels foe features in map layers, these maps in turn are the sources traditionally used as a basis for gazetteers. From these map series, coordinate points or grid references for features can be obtained as well as supporting evidence for officially authorized names alternative names and variant spellings. Charts that identify the locations and land marks for air, sea and coastal navigation are ready sources of gazetteer data, as are the gazetteers included as indexes for atlases". Hill 2006 p.104

- "Gazetteer data is also the result of scientific research projects were locations along the path of an expedition, the boundaries of research areas, or specimen collection sites are recorded. Scholarly historical research rediscovers and locates places referenced in the past and genealogical archives, oral histories, and memories contain many placename references". Hill 2006 p.104

- "With a GPS in hand – that is , a global positioning system unit that calculates your position by receiving signals from orbiting satellites ,gazetteer data can be collected by almost anyone. These locations could be the routes of hiking trails or bicycle paths, the boundaries of school athletic fields or the locations of grave sites of interest from family history" Hill 2006 p.105.

- VGIS people who are interested in placenames and their existence voluntarily upload and add information into existing gazetteers so that others can see and use them.

2.6.5 Criteria of gazetteer quality

Hill 2006 p.107 stated that “John Leidner in an extended abstract and presentation titled *Towards a Reference Corpus for Automatic Toponym Resolution Evaluatio*n for a workshop on geographic information retrieval, identified key criteria of gazetteer quality that he used as a criterion for his research project”;

1. Availability: degree to which it is available and not limited by restrictive conditions of us
2. Scope: small communal database, regional/national or worldwide coverage
3. Completeness: degree to which the scope of gazetteer is covered completely
4. Currency: degree to which the gazetteer has incorporated changes
5. Accuracy: number of detectable errors in names, footprints and types
6. Granularity; includes large, well-known features only or features of all sizes and those that are less well known
7. Balance; uniform degree of detail, currency accuracy and granularity across scope of coverage
8. Richness of annotation: amount and detail of descriptive information beyond the basics of name, footprint and type (Hill 2006 adapted from Leidner 2004)

2.7 Placenames

2.7.1 Origins

“In much of the "Old World" (approximately Africa, Asia and Europe) the names of many places cannot easily be interpreted or understood; they do not convey any apparent meaning in the modern language of the area. This is due to a general set of processes through which place names evolve over time, until their obvious meaning is lost. In contrast, in the "New World" (roughly North America, South America, and Australasia), many place names origins are known[11]” The origin of naming of some places very difficult to trace unlike others,

[11]http://en.wikipedia.org/wiki/Place_name_origins23/07

especially those which are not linked to the current state of the place, the place might have gone through a series of developments which have affected its naming over time for example. Places are named either from the rivers in the areas, links to their colonial masters or specific or general words.

2.7.2 Types of Placenames

Place type ranges from parks, hotels, building, rivers, lake, plantation, museum and the list continues, but they can be grouped under human settlements, natural features and others who not in the two categories

2.7.2.1 Human settlements

A settlement is any form of human habitation or dwelling. It could be the name of a building, country, city, village or county e.g. Cameroon, Frankfurt or Maryland.

2.7.2.2 Natural features

Natural features are given names because they occupy space on the earth's surface and they also have coordinates so they are locations. They are given names to distinguish between them; they include features like lakes, mountains, rivers, desert, cliff, coast and the list continues, examples of such placenames include Lake Victoria was named after Queen Victoria by the explorer John Hanning Speke, who was the first European to discover it[12].

Other placenames originate from names of important persons like Washington DC got its name fin honour of George Washington or events.

2.8 Newspapers

"A newspaper (often just called a paper when the context is clear) is a periodical publication containing news, other informative articles (listed below), and usually advertising. A newspaper is usually printed on relatively inexpensive, low-grade paper such as newsprint. The news organizations that publish newspapers are themselves often metonymically called newspapers. Most newspapers now publish online as well as in print. The online versions are called online newspapers or news sites.[13]"

[12]http://en.wikipedia.org/wiki/Place_name_origins 23/07
[13]http://www.quora.com/Newspapers?merged_tid=445589

Figure 2.4 shows an example of how a newspaper looks like, it has different sections of news: political, economic and the social sections. Some local newspapers have sections for national and international news as well. Majority of newspaper are printed on light brown papers and black ink is the standard color though some countries use blue and red.

"All the News That's Fit to Print."

The New York Times.

THE WEATHER

ARMISTICE SIGNED, END OF THE WAR!
BERLIN SEIZED BY REVOLUTIONISTS;
NEW CHANCELLOR BEGS FOR ORDER;
OUSTED KAISER FLEES TO HOLLAND

KAISER FLEES WITH EX-KAISER

Hindenburg Also Believed to be Among Those in His Party.

ALL ARE HEAVILY ARMED

BERLIN TROOPS JOIN REVOLT

Rade Shell Building in Which Officers Vainly Resist.

WAR ENDS AT 6 O'CLOCK THIS MORNING

The State Department in Washington Made the Announcement at 2:45 o'Clock.

ARMISTICE WAS SIGNED IN FRANCE AT MIDNIGHT

Terms include Withdrawal from Alsace-Lorraine, Disarming and Demobilization of Army and Navy, and Occupation of Strategic Naval and Military Points.

GERMAN DYNASTIES BEING WIPED OUT

MORE WARSHIPS JOIN THE REDS

Figure 2.4 Example of a popular published newspaper

Source: http://en.wikipedia.org/wiki/Newspaper

2.8.1 Definition

"A newspaper basically meets four criteria's

Publicity: Its contents are reasonably accessible to the public.

Periodicity: It is published at regular intervals.

Currency: Its information is as up to date as its publication schedule allows. Universality: It covers a range of topics[14]"

From the definition above, a newspaper is a public document, not for private usage, as such it should be made accessible, in Switzerland most newspapers are free to the public like '20 minuten' and 'Blick am Abend' are free while the 'NZZ' is paid but at least 90% of the population read the free ones daily. A newspaper is also a consistent document and has a

[14] http://en.wikipedia.org/wiki/Newspaper

Continuous supply either daily or weekly. It is also very current with the happenings at that moment and it is usually divided into several topics ranging from job openings, sports, political and geographic issues.

2.8.3 Origins of newspapers

"The term newspaper became common in the 17th century. However, in Germany, publications that we would today consider to be newspaper publications were appearing as early as the 16th century. They were discernibly newspapers for the following reasons: they were printed, dated, appeared at regular and frequent publication intervals, and included a variety of news items (unlike single item news mentioned above). The first newspaper however was said to be the *Strasbourg Relation*, in the early 17th century[15]"

The origins of newspapers can be traced way back to the time when they were called newssheets and then later in the 17th century when the printing press was discovered: printed periodicals replaced handwriting and then publishing became popular, this led to the rapid growth of news in printed documents which were called newspapers, this practice is still carried out till data, even the poorest countries in the world publish news as this is very cheap and can easily be circulated.

2.8.4 Categories of Newspapers

The exists different categories of newspapers from different views and perspectives as seen on the table below;

Category	**Example**
Frequency	Daily e.g. NZZ Weekly e.g. NZZ am sonntag
Geographical Scope	Local/regional: The post(Cameroon) National: Cameroon Tribune International: The international Herald Journal
Technology	Print : NZZ Online, Custom: my Yahoo, Twitter

Table 2.4 Overview of newspapers[16]

[15] https://www.nyu.edu/classes/stephens/Collier%27s%20page.html
[16] http://en.wikipedia.org/wiki/Newspaper

3. Data

This chapter throws light to the different data types that will be used in this thesis and their sources, what I will extract from the data to use for my analysis

Newspapers 3.1

Case Studies 3.2

GeoNames Data 3.3

Population data 3.4

3.1 Newspapers sources

3.1.1 NZZ from Zurich, Switzerland

"The ***Neue Zürcher Zeitung*** (**NZZ**, English: "New Journal of Zurich") is a Swiss, German-language daily newspaper, published by the NZZ-Gruppe in Zurich[17]". Figure 3.1 shows how the daily publication of NZZ looks like


Neue Zürcher Zeitung

SCHWEIZER AUSGABE

Fragwürdiger Auftrag des Bundes an Microsoft

USA gehen härter gegen Steuerflucht vor

Immer grössere Verschuldung in der EU

Yuan Sieg der Agrarpolitik

Bericht zur Jura-Frage von Vorsicht geprägt

Figure 3.1 NZZ newspaper

"It has a reputation as a quality newspaper and as the Swiss newspaper of record, the newspaper is known for its detailed reports on international affairs, stock exchange and the intellectual, in-depth style of its articles, it appeared as Zürcher Zeitung and was renamed Neue Zürcher Zeitung in 1821[17]"

[17] http://en.wikipedia.org/wiki/Neue_Z%C3%BCrcher_Zeitung

"In 2002, the newspaper launched a weekend edition, *NZZ am Sonntag* (*NZZ on Sunday*). The weekend edition has its own editorial staff and contains more soft news and lifestyle issues than its weekday counterpart, as do most Swiss weekend newspapers[17]"

3.1.2 Sri Lanka Daily News Online

"It is published in English, on the website www.dailynews.lk, the publisher or parent company is called associated newspapers of Ceylon limited a government owned corporation established in 1918[18]".

Figure 3.2 Sri Lanka daily news online

Source; http://www.dailynews.lk/

3.1.3 Cameroon Tribune

"The ***Cameroon Tribune*** is a major newspaper in Cameroon. It is also available online. It is owned by the government[19]". Originally, it was released in two versions, a French version and an English version but now it has just one which mixes articles in English and in French. It is a state owned newspaper, Figure 3.3 below shows how the Cameroon Tribune newspaper looks like.

[17]http://en.wikipedia.org/wiki/Neue_Z%C3%BCrcher_Zeitung
[18]http://en.wikipedia.org/wiki/Daily_News_%28Sri_Lanka%29
[19]http://en.wikipedia.org/wiki/Cameroon_Tribune

Source: Cameroon Tribune

Figure 3.3 Cameroon tribune

3.2 Case Studies

3.2.1 Cameroon

Short name:	Cameroon
Official name:	Republic of Cameroon
Status:	Independent country since 1960
Location:	Central Africa
Capital:	Yaoundé
Population:	17,795,000 inhabitants
Area:	474,442 km²
Major Languages:	French & English (Official) Cameroonian Pidgin English Fulfulde and Ethnic Languages
Major Religions:	Indigenous beliefs, Roman Catholic, Islam

Fig 3.4 Cameroon Political Map
Source: **http://en.wikipedia.org/wiki/Cameroon**

3.2.2 Switzerland

Status:	Democratic
Location:	Europe
Capital:	Bern
Population:	8,014,000 inhabitants
Area:	41,285km²
Major Languages:	French & English (Official) German,
Major Religions:	French, Italian, Indigenous beliefs, Roman Catholic, Islam

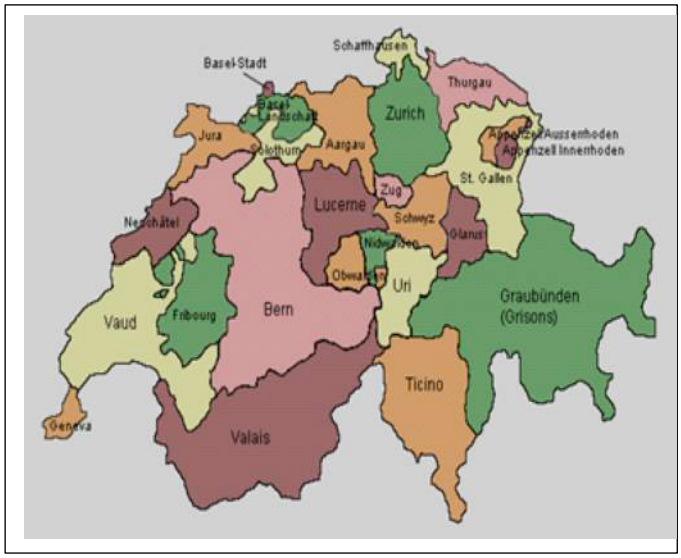


Figure 3.4 Switzerland Political map

Source http://en.wikipedia.org/wiki/Switzerland

3.2.2.1 What is SwissNames?

SwissNames is the name of the official Switzerland national gazetteer, which is based on what is found on its national map. It was compiled and is maintained with up to date modifications from the 'ground' by the Federal Office of Topography; swisstopo. The gazetteer contains approximately 193'000 georeferenced entries. It has different maps scales with varying contents: SwissNames 25, for example, is the set of all toponyms which appear on the national maps scaled 1:25'000. Similarly, there are sets for the other maps issues by swisstopo, which have the scales 1:50'000, 1:100'000, 1:200'000 and 1:500'000. There is also SwissNames Ortschaften (SwissNames settlements), which is the set of names of all municipalities, cities, towns, villages and hamlets that appear on any one of the national maps (Swiss Federal Office of Topography, 2002) and I will use data from it which has a list of all the placenames in Switzerland to compare with the list of names of places in Switzerland from the GeoNames gazetteer. Each entry in the SwissNames consists of a toponym, the coordinates of the object it is referring to and several other attributes such as an ID number and a feature class. (Swiss Federal Office of Topography, 2002).

3.2.2.2 Quality of SwissNames

SwissNames has one of the most comprehensive and detailed sets of Swiss toponyms. The gazetteer covers all of Switzerland homogeneously and is updated yearly. Updates are made according to information received from federal and cantonal authorities and relying on the knowledge gathered by topographers in the field for the overall update of the national maps, which is done every six years and this is very important especially for rescue services, also the non-complex structure of the gazetteer allows for an implementation of SwissNames on various systems (Swiss Federal Office of Topography, 2005).

3.2.3 Sri Lanka

Official name: Sri Lanka

Status: Independent country since 1960

Location: Southern Asia

Capital: Administrative: Sri Jayewardenepura Commercial: Colombo

Population: 20,48 million inhabitants

Area: 65,610 km²

Major Languages Sinhala & Tamil but English is recognised

Major Religions: Buddhism,Hindusm,Islam, Roman Catholicism & Others

Figure 3.5 Sri Lanka Political Map

Source: http://en.wikipedia.org/wiki/Sri_Lanka

3.2 GeoNames Data

All Data were downloaded in April 2013 online from the GeoNames Gazetteer Database as seen below

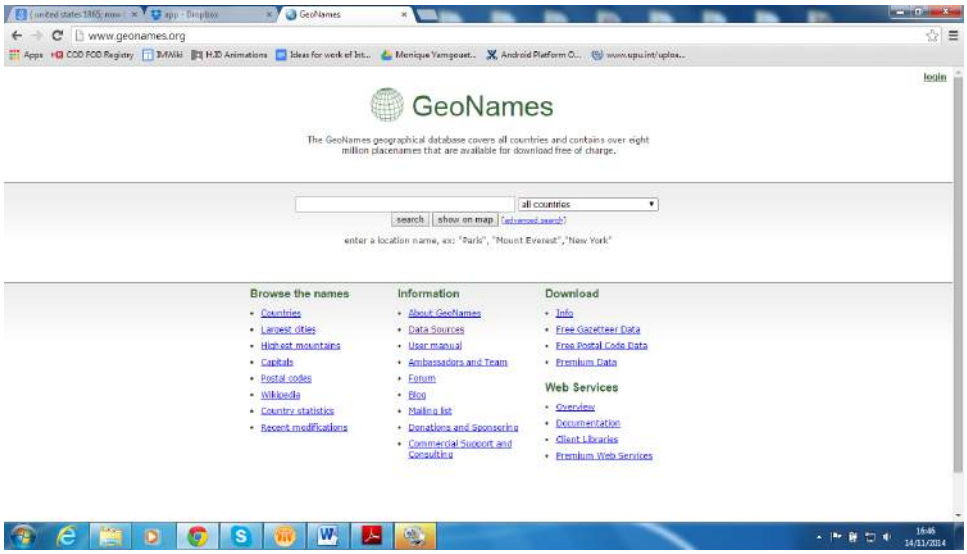


Figure 3.7 Placename search from GeoNames Database

When you click on free gazetteer data you will get something like the Figure below , then you choose the country abbreviation are according to A2 (ISO) standards

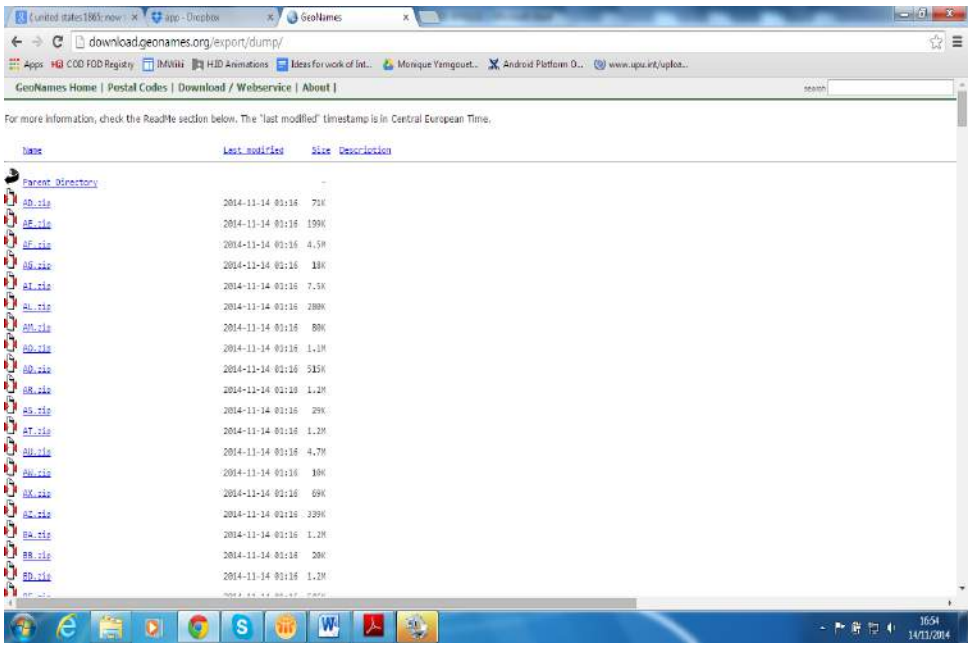


Figure 3.8 Placename download per country showing time and of download plus file size

3.3.1 Data Downloaded

Cameroon: CM.zip 2014-04-10 478k

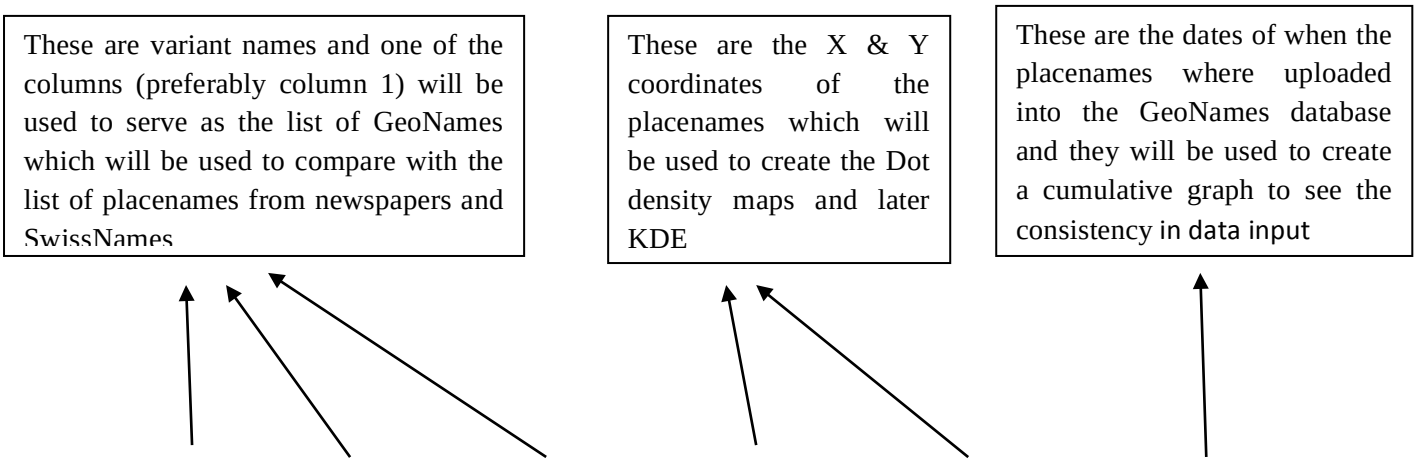
Haiti: HT.zip 2014-04-10 305k

Somalia: SO.zip 2014-4-19 450k

Sri Lanka: LK.zip 2014-04-10 1.0M

Switzerland: CH.zip 2014-04-10 753K

3.3.2 Data sorted from gazetteer download



Name 1	Name 2	Name 3	X	Y	Input Date

Table 3.1 Data extracted from GeoNames data

3.3 Data from Swisstopo

The file containing the list of SwissNames from swisstopo was be gotten from swisstopo and will be opened in ArcGIS

3.4 Population data from GeoHive

3.5.1 Cameroon Population Statistics

FID	Region	Area (sq.km.)	Population 2010 estimate
CM10	Adamawa	63,701	1,015,622
CM11	Centre	68,953	3,525,664
CM04	East	109,002	801,968
CM12	Far-North	34,263	3,480,414
CM05	Littoral	20,248	2,865,795
CM13	North	66,090	2,050,229
CM07	Northwest	17,300	1,804,695
CM07	West	13,892	1,785,285
CM14	South	47,191	692,142
CM09	South-West	25,410	1,384,286
Total		**466,050**	**19,406,100**

Table 3.2 Cameroon Population Statistics
Source: http://www.geohive.com/cntry/cameroon.aspx

3.4.2 Switzerland Population Statistics

FID	Canton	Area (sq.km^2)	Population 2013 estimate
CH01	Zürich	1,728.89	1,425,538
CH02	Bern	5,959.07	1,001,281
CH03	Luzern	1,493.42	390,349
CH04	Uri	1,076.40	35,865
CH05	Schwyz	908.09	151,396
CH06	Obwalden	490.55	36,507
CH07	Nidwalden	276.06	41,888
CH08	Glarus	685.40	39,593
CH09	Zug	238.72	118,118
CH10	Fribourg	1,670.84	297,622
CH11	Solothurn	790.51	261,437
CH12	Basel-Stadt	37.07	189,335
CH13	Basel-Landschaft	517.52	278,656
CH14	Schaffhausen	298.50	78,783
CH15	Appenzell Ausserrhoden	242.94	53,691
CH16	Appenzell Innerrhoden	172.50	15,778
CH17	Sankt Gallen	2,025.45	491,699
CH18	Graubünden	7,105.15	194,959
CH19	Aargau	1,403.79	636,362
CH20	Thurgau	990.87	260,278
CH21	Ticino	2,812.46	346,539
CH22	Vaud	3,212.05	749,373
CH23	Valais	5,224.42	327,011
CH24	Neuchâtel	803.06	176,402
CH25	Genève	282.44	469,433
CH26	Jura	838.81	71,738
Total		**41,284.98**	**8,139,631**

Table 3.3 Switzerland Population Statistics
Source: http://www.geohive.com/cntry/switzerland.aspx

3.4.3 Sri Lanka Population statistics

FID	Region	Area (sq.km.)	Population 2012 estimate
CE36	Western (Basnahira)	3,684	5,821,710
CE35	Uva	8,500	1,259,900
CE34	South (Dakunu)	5,444	2,464,732
CE33	Sabaragamuwa	4,968	1,918,880
CE32	North Western (Wayamba)	7,888	2,370,075
CE38	North (Uturu)	8,884	1,058,762
CE30	North Central (Uturumeda)	10,472	1,259,567
CE37	Eastern(Negenahira)	9,791	7,551,381
CE29	Central (Madhyama)	5,674	2,558,716
Total		**65,610**	**20,263,723**

Table: Table 3.4 Cameroon Population Statistics
Source: http://www.geohive.com/cntry/srilanka.aspx

4. Methodology

In this chapter, placenames from different sources will be compared to that of the GeoNames Gazetteer. Newspapers from three sources (Cameroon tribune, NZZ and Sri Lanka Daily news online) will be collected for a period of 30days and in the next step they will be read and the placenames will be identified. A check will be carried out to see if all the placenames on the local newspapers are on the GeoNames database. Data that was downloaded from the GeoNames gazetteer will be used to create dot density maps to show the distribution of placenames, to see placename input into the gazetteer, to see placename count per region and to know if placename and population have a correlation.

1. Overview of Methodology (section 4.1)

2. Data Collection (section 4.2)

3. Name finding (section 4.3)

4. Data Processing (section 4.4)

5. Results Analysis (Section 4.5)

4.1 Overview of methodology

Data Collection

- Collection of daily (online & published) newspapers for 30days (April 2014)

-Download of placenames from GeoNames gazetteer database for Cameroon, Sri Lanka, Switzerland, Haiti & Somalia

-Collection of placenames from country atlas of Switzerland

-Collection of population data for Cameroon, Switzerland, Sri Lanka from GeoHive

Name Finding

Reading Newspapers, recognizing and underlining placenames

Data Processing

-Comparing placenames from Newspapers with that from GeoNames Gazetteer

-Comparing placenames from Swiss country atlas with those of GeoNames Gazetteer

-Analysing data input in GeoNames to show bias in Gazetteer Construction

-Creating three dot density maps for Cameroon, Switzerland & Sri Lanka using coordinates from GeoNames data

-Calculating kernel Density Estimation from dot density maps to describe placename clustering and comparing with population density map

-Comparing placename count and population count per region to get correlation

Result Analysis

-A list of placenames on the local Newspapers not in the list of GeoNames

-A list of placenames in the SwissNames not in the GeoNames

-Cumulative graph illustrating fluctuations in placename input

-Dot density showing placename distribution in space for various countries

- Kernel density maps describing placename clustering and distribution

-Population Density map showing population density on space

-Distribution of placename in space per Country

-Regression analysis showing relationship between placename and population

Figure 4.1 Overview of methodology

4.1.1 Data Collection

-Collection of daily (online & published) newspapers for 30days (April 2014)

-Download of placenames from GeoNames gazetteer Database for Cameroon, Sri Lanka, Switzerland, Haiti & Somalia

-Collection of placenames from country atlas of Switzerland

-Collection of population data for Cameroon, Switzerland, Sri Lanka from GeoHive

Figure 4.2 Schematic Overview of Data Collection

Normally before data analysis you must make sure you have all the data available for your research. I started by collecting and downloading all the data I needed both from online and published sources as seen in Figure 4.2 above.

I went to the GeoNames Database and downloaded data for five countries as seen below, they were in .zip format then I unzipped the file and opened all archives, it opened on an Excel sheet and formatted the respective columns to extract what I need.

Country	**File Name**	**Size**	**Date of Download**
Cameroon	CM.zip	478K	10-04-2014
Haiti	HT.zip	305K	10-04-2014
Sri Lanka	LK.zip	1.0M	10-04-2014
Somalia	SO.zip	450K	10-04-2014
Switzerland	CH.zip	753K	10-04-2014

Table 4.1 Data downloaded from GeoNames Database

From the table above I will need data from Cameroon, Sri Lanka and Switzerland for comparisons with placenames from the three chosen Newspapers, to create dot density maps to see placenames distribution and then KDE to see placename distribution and clustering. For Haiti, Sri Lanka and Somalia I will need just the date of the placename input.

Swisstopo has a list of all placenames in Switzerland and I will add the data onto ArcGIS and extract the list from the attribute table.

GeoHive has Population census data from all countries in the world; I will use the data from it to calculate population density which I will create population density maps well with.

4.1.2 Name Finding

Reading Newspapers, recognizing and underlining placenames

Figure 4.3 Schematic Overview of Name Finding

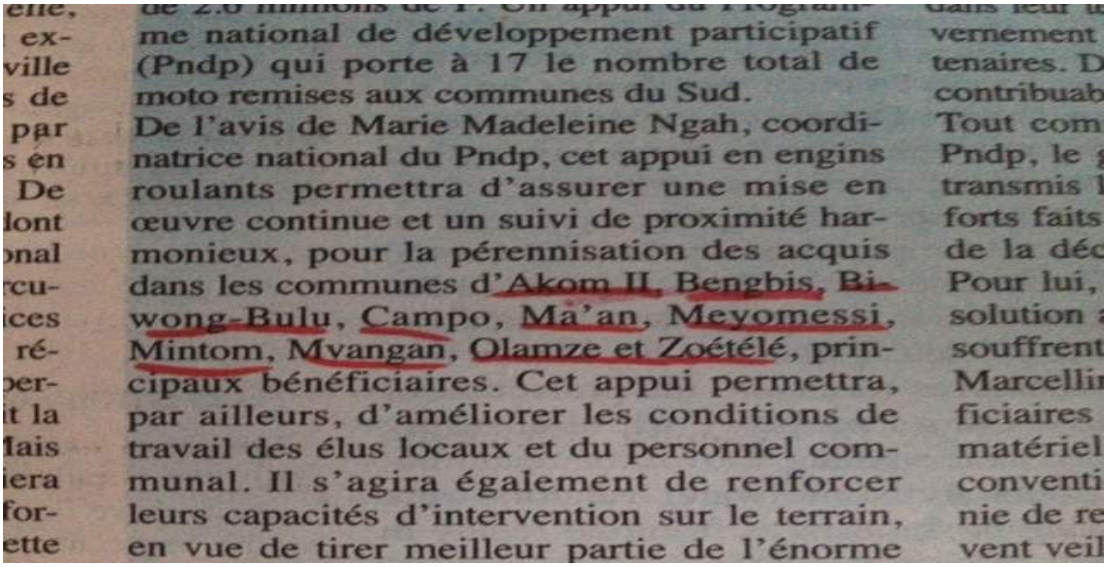

me national de développement participatif
(Pndp) qui porte à 17 le nombre total de
moto remises aux communes du Sud.
De l'avis de Marie Madeleine Ngah, coordi-
natrice national du Pndp, cet appui en engins
roulants permettra d'assurer une mise en
œuvre continue et un suivi de proximité har-
monieux, pour la pérennisation des acquis
dans les communes d'Akom II, Bengbis, Bi-
wong-Bulu, Campo, Mâ'an, Meyomessi,
Mintom, Mvangan, Olamze et Zoétélé, prin-
cipaux bénéficiaires. Cet appui permettra,
par ailleurs, d'améliorer les conditions de
travail des élus locaux et du personnel com-
munal. Il s'agira également de renforcer
leurs capacités d'intervention sur le terrain,
en vue de tirer meilleur partie de l'énorme

Figure 4.4 recognizing placenames and underlining them

Newspapers were collected daily for the month of April. Underlining is used to distinguish and emphasize certain words from others within a text as seen in Figure .4 above; these could be names of places or things, strange/new words, titles of documents or words reproduced as sounds. Underlining is a typographical method which could be done manually or using a computer (equivalent to italicizing in computer). "The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them"(Sebastiani 2002), this involves breaking the words in the sentences into tokens under nouns, pronouns, adjectives, adverbs and conjunctions. This thesis is focused on identifying nouns.

While reading these three newspapers (NZZ; Cameroon Tribune & Sri Lanka Daily news online) and recognizing the placenames, just like in any other text we can come across situations where it will be confusing if the names identified are names of people or names of

places; geo-non geo toponym ambiguity (Amitay et al., 2004) to solve this problem, all the names will be added to the list of verifications (toponym resolution) will be done later. Some placenames have one word tokens like Bern while others have two word tokens like Sankt Gallen so they are not mistaken for various locations. Though the use of hand methods to identify names in text (also known as Geoparsing) is not new, modern machine learning methods have been developed in computational linguistics to do this task fast and easy (Leidner 2007)

"The idea of reducing information from a document to a tabular structure is not new. Its feasibility for sublanguage texts was suggested by Zellig Harris in 1950s" (Grisham 1997) this can also be termed information extraction and it varies according to what information the individual needs from the text, in this thesis we are focusing on extracting placenames from newspapers and filling them in Excel sheets for a period of 30days for the month of April; the newspapers will be read and the recognized, placenames will be underlined with hand.

Grisham 1997 stated that; "If we want to know who has signed contracts over the past year to deliver airplanes or natural gas or which jurisdictions have enacted new restrictions, we must pour over realms of retrieved documents by hand". Grisham tries to explain the kind and quality of information we can retrieve from printed documents, though with the advent of modern technology which has differing methods of saving information on the computer and even in external drives, the importance of keeping documents in printed form still over rules.

4.1.3 Data Processing

-Comparing placenames from newspapers with that from GeoNames Gazetteer

-Comparing placenames from Switzerland country atlas with those of GeoNames Gazetteer

- Analysing data input in GeoNames to show bias in gazetteer construction

-Creating three dot density maps for Cameroon, Switzerland & Sri Lanka using coordinates from GeoNames data

-Calculating kernel Density Estimation from dot density maps to describe placename clustering and comparing with population density map

-Distribution of placename on space for Cameroon, Sri Lanka & Switzerland

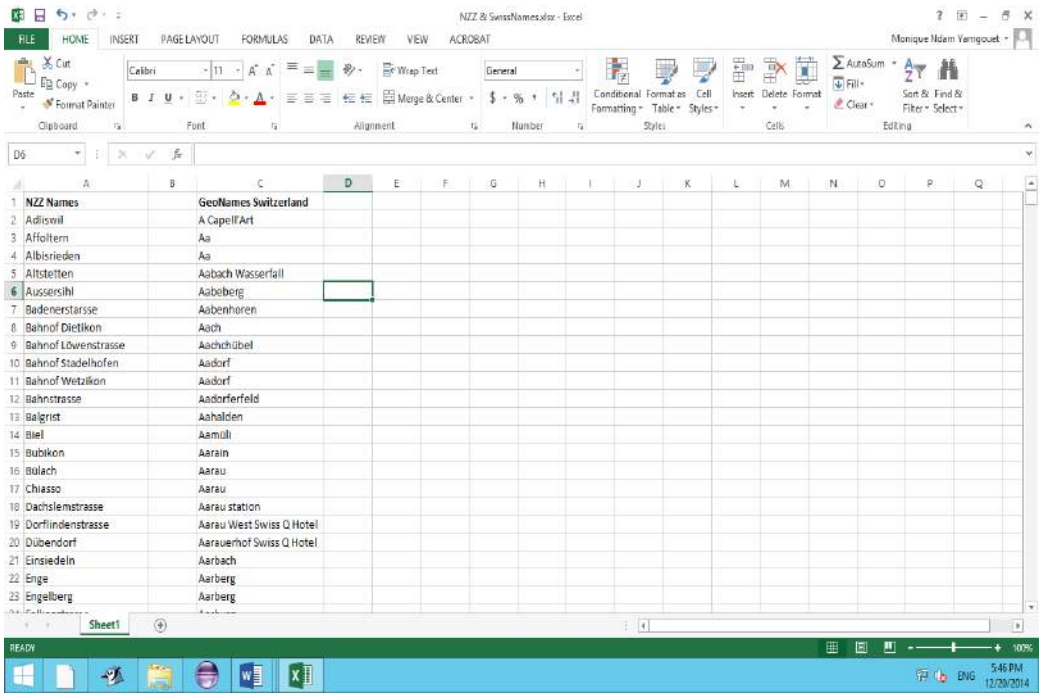
-Comparing placename count and population count per region to get correlation

Figure 4.5 Schematic Overview of Data Processing

I am going to do all the comparisons and map productions in this section for the various countries whose data I already have, the comparisons will be as seen below.

4.1.3.1 Comparing data from Newspapers & GeoNames

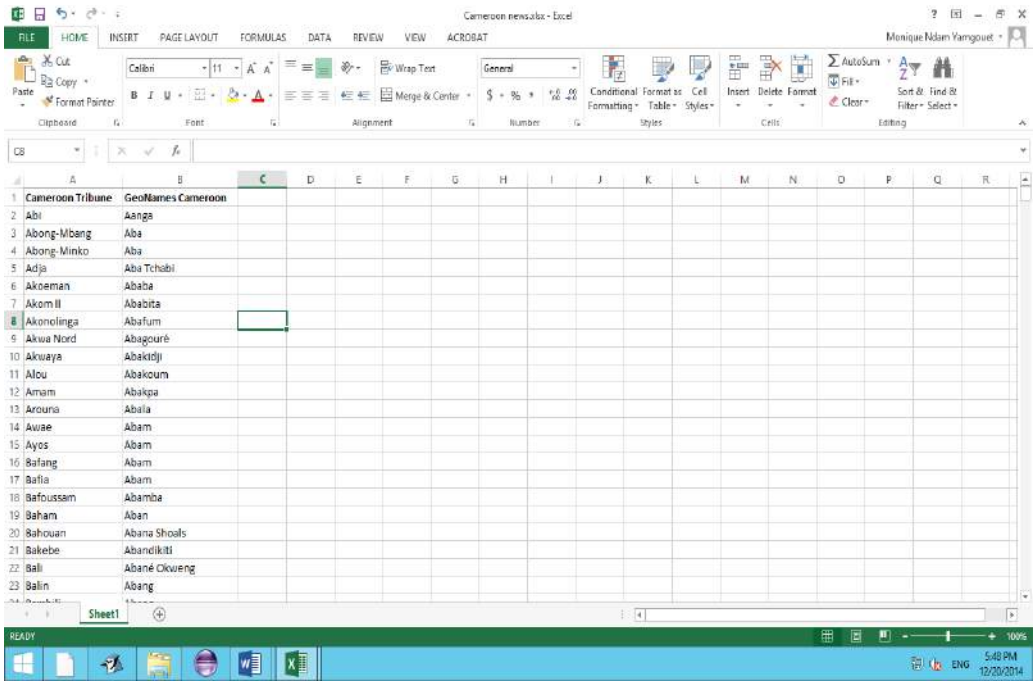
- Switzerland



NZZ Names	GeoNames Switzerland
Adliswil	A Capell'Art
Affoltern	Aa
Albisrieden	Aa
Altstetten	Aabach Wasserfall
Aussersihl	Aabeberg
Badenerstarsse	Aabenhoren
Bahnof Dietikon	Aach
Bahnof Löwenstrasse	Aachchübel
Bahnof Stadelhofen	Aadorf
Bahnof Wetzikon	Aadorf
Bahnstrasse	Aadorferfeld
Balgrist	Aahalden
Biel	Aamüli
Bubikon	Aarain
Bulach	Aarau
Chiasso	Aarau
Dachslernstrasse	Aarau station
Dorflindenstrasse	Aarau West Swiss Q Hotel
Dübendorf	Aarauerhof Swiss Q Hotel
Einsiedeln	Aarbach
Enge	Aarberg
Engelberg	Aarberg

Table 4.2 NZZ & Switzerland GeoNames

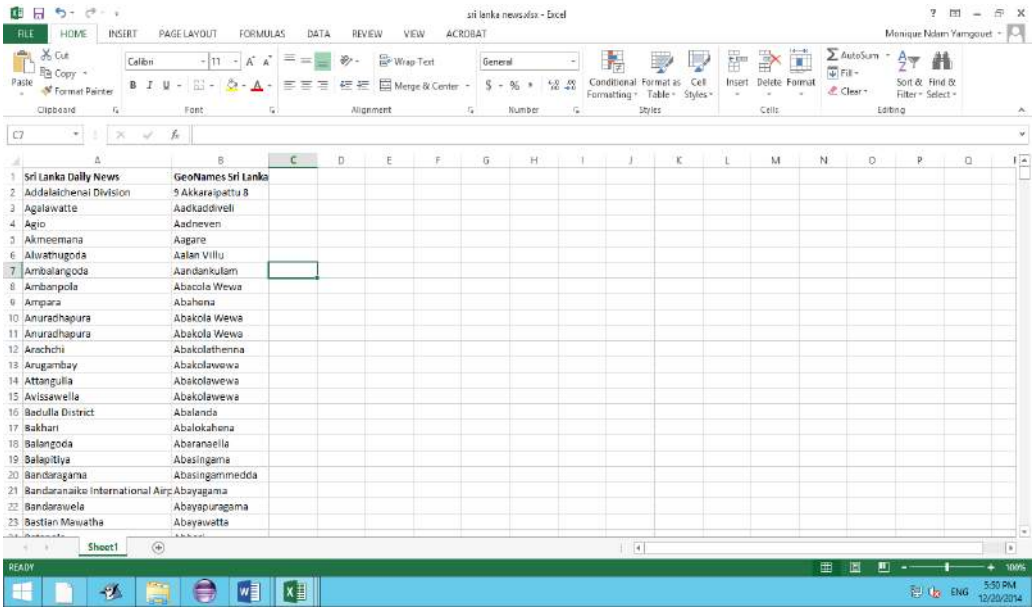
- Cameroon



Cameroon Tribune	GeoNames Cameroon
Abi	Aanga
Abong-Mbang	Aba
Abong-Minko	Aba
Adja	Aba Tchabi
Akoeman	Ababa
Akom II	Ababita
Akonolinga	Abafum
Akwa Nord	Abagourè
Akwaya	Abakidji
Alou	Abakoum
Amam	Abakpa
Arouna	Abala
Awae	Abam
Ayos	Abam
Bafang	Abam
Bafia	Abam
Bafoussam	Abamba
Baham	Aban
Bahouan	Abana Shoals
Bakebe	Abandikiti
Bali	Abané Okweng
Balin	Abang

Table 4.3 Cameroon Tribune &Cameroon GeoNames

- Sri Lanka



Sri Lanka Daily News	GeoNames Sri Lanka
Addalaichenai Division	5 Akkaraipattu 8
Agalawatte	Aadkaddiveli
Agio	Aadneven
Akmeemana	Aagare
Alwathugoda	Aalan Villu
Ambalangoda	Aandankulam
Ambanpola	Abacola Wewa
Ampara	Abahena
Anuradhapura	Abakola Wewa
Anuradhapura	Abakola Wewa
Arachchi	Abakolathenna
Arugambay	Abakolawewa
Attangulla	Abakolawewa
Avissawella	Abekolawewa
Badulla District	Abalanda
Bakhari	Abalokahena
Balangoda	Abaranaella
Balapitiya	Abasingama
Bandaragama	Abasingammedda
Bandaranaike International Airp	Abayagama
Bandarawela	Abayapuragama
Bastian Mawatha	Abayawatta

Table 4.4 Sri Lanka Daily News & Sri Lanka GeoNames

After filling the placenames from the newspapers daily into the excel sheet for a period of 30 days, I emerged with one Gold standard list (List 1) per newspaper, note should be taken that while filling the names into the excel sheet, names that were already been filled into the excel automatically indicated when I started typing the placename and since I am not dealing with placename frequency I will not refill the same name again. After that, I sorted the names alphabetically in descending order then I inserted the names from the GeoNames gazetteer in a separate column (List 2) and also sorted it alphabetically in descending order as can be seen in Tables 4.1, 4.2 and 4.3 respectively

After sorting the Names, I selected and highlighted the two columns, I used conditional formatting, highlighted cell rules then selected format duplicate values, and then I choose the orange color to indicate the duplicate values. The results will highlight names on both lists, but note should be taken that I am interested only on the names that are not highlighted on the first list (News placenames) which indicate the placenames that are not found on the second list (GeoNames list)

4.1.3.2 Comparing SwissNames & Switzerland GeoNames

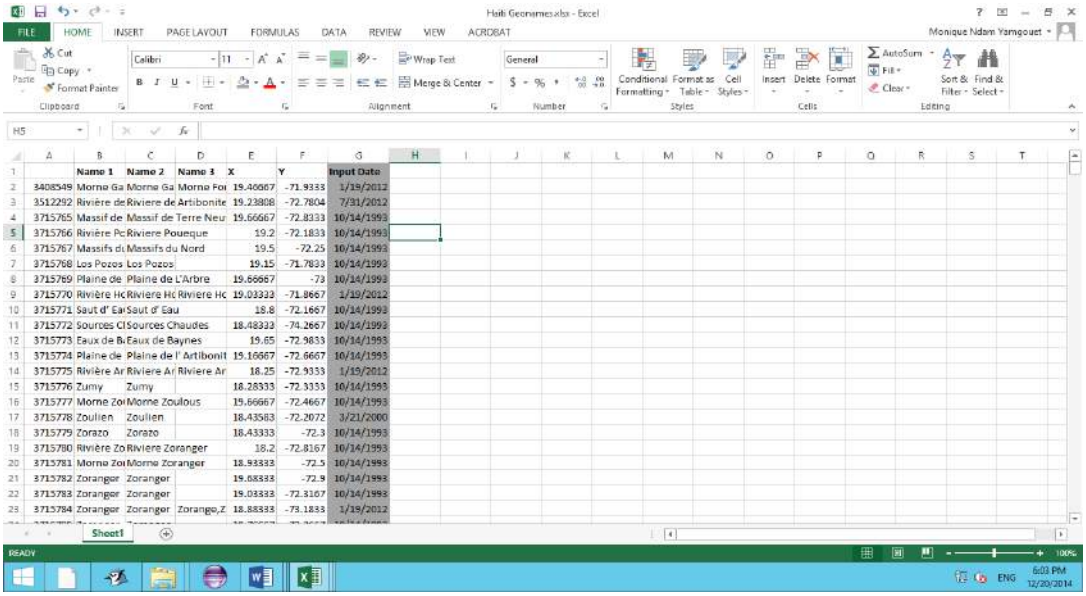
The data file containing the SwissNames from the Swisstopo is inserted into ArcGIS and the list is opened through the attribute table and the names of all the SwissNames is exported from the ArcGIS and opened with an Excel sheet, the names are sorted alphabetically to get the SwissNames (List 1) which is used to compare with the Switzerland GeoNames list (List 2) After sorting the Names I will do the comparison. I won't use excel formatting to do this comparison because the SwissNames list has 156, 755 names while the GeoNames list has just 23,559 names.

4.1.3.3 Data input of Placenames to show bias in construction

The data downloaded from the GeoNames gazetteer data (Table 3.1) will be filtered and sorted from the Table 4.4. , 4.5 and 4.6 illustration according to each year (in descending order from the most recent year to the oldest) and the place names at the beginning of each year will be subtracted from the placename at the end of the year (for example names from 2010 will be subtracted from 2011) to get the total placename input for that year. In the end we will firstly have placename input for all the years then secondly we will have a cumulative add up of the placenames in ascending order to see if as the years went by the placenames input increased or decreased.

4.1.3.3.1 Haiti

Data is from 1993-2014

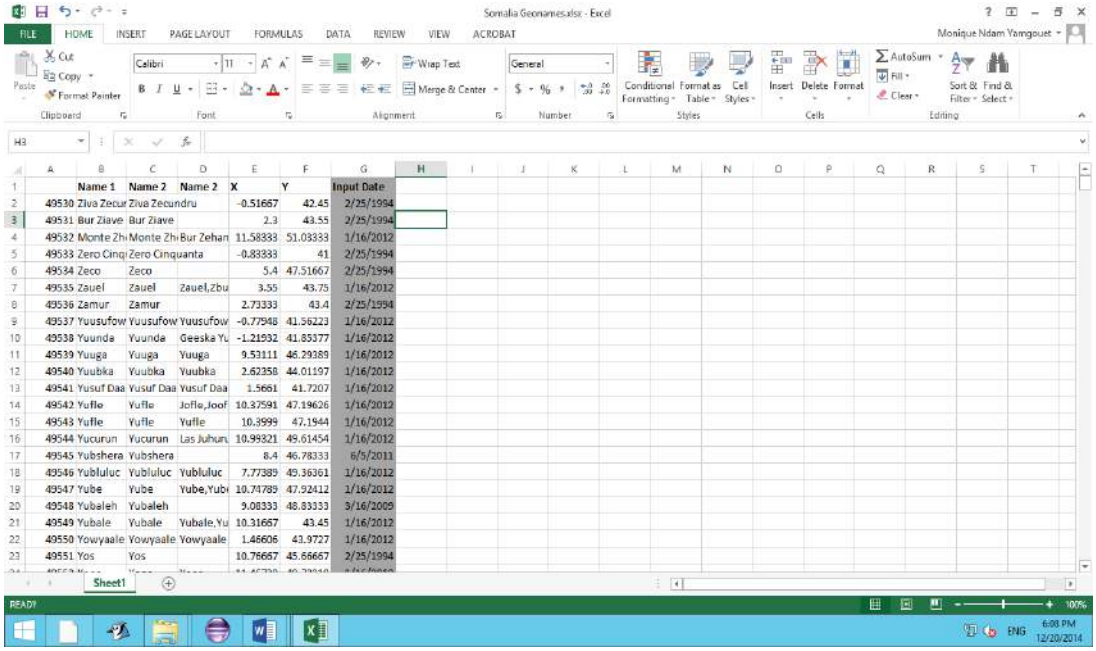


	Name 1	Name 2	Name 3	X	Y	Input Date
2	3408549	Morne Ga	Morne Ga	Morne Foi	19.40667	-71.9333	1/19/2012
3	3512292	Rivière de	Riviere de	Artibonite	19.23808	-72.7804	7/31/2012
4	3715765	Massif de	Massif de	Terre Neu	19.66667	-72.8333	10/14/1993
5	3715766	Rivière Pc	Riviere	Poueque	19.2	-72.1833	10/14/1993
6	3715767	Massifs du	Massifs du	Nord	19.5	-72.25	10/14/1993
7	3715768	Los Pozos	Los Pozos		19.15	-71.7833	10/14/1993
8	3715769	Plaine de	Plaine de	L'Arbre	19.66667	-73	10/14/1993
9	3715770	Rivière Hc	Riviere Hc	Riviere Hc	19.03333	-71.8667	1/19/2012
10	3715771	Saut d' Ea	Saut d' Eau		18.8	-72.1667	10/14/1993
11	3715772	Sources Cl	Sources	Chaudes	18.48333	-74.2667	10/14/1993
12	3715773	Eaux de Ba	Eaux de	Baynes	19.65	-72.9833	10/14/1993
13	3715774	Plaine de	Plaine de l'	Artibonit	19.16667	-72.6667	10/14/1993
14	3715775	Rivière Ar	Riviere Ar	Riviere Ar	18.25	-72.9333	1/19/2012
15	3715776	Zumy	Zumy		18.28333	-72.3333	10/14/1993
16	3715777	Morne Zoi	Morne	Zoulous	19.66667	-72.4667	10/14/1993
17	3715778	Zoulien	Zoulien		18.43583	-72.2072	3/21/2000
18	3715779	Zorazo	Zorazo		18.43333	-72.3	10/14/1993
19	3715780	Rivière Zo	Riviere	Zoranger	18.2	-72.8167	10/14/1993
20	3715781	Morne Zoi	Morne	Zoranger	18.93333	-72.5	10/14/1993
21	3715782	Zoranger	Zoranger		19.08333	-72.9	10/14/1993
22	3715783	Zoranger	Zoranger		19.03333	-72.3167	10/14/1993
23	3715784	Zoranger	Zoranger	Zorange,Z	18.88333	-73.1833	1/19/2012

Table 4.5 Haiti Data Input

4.1.3.3.2 Somalia

Data is from 1994-2014

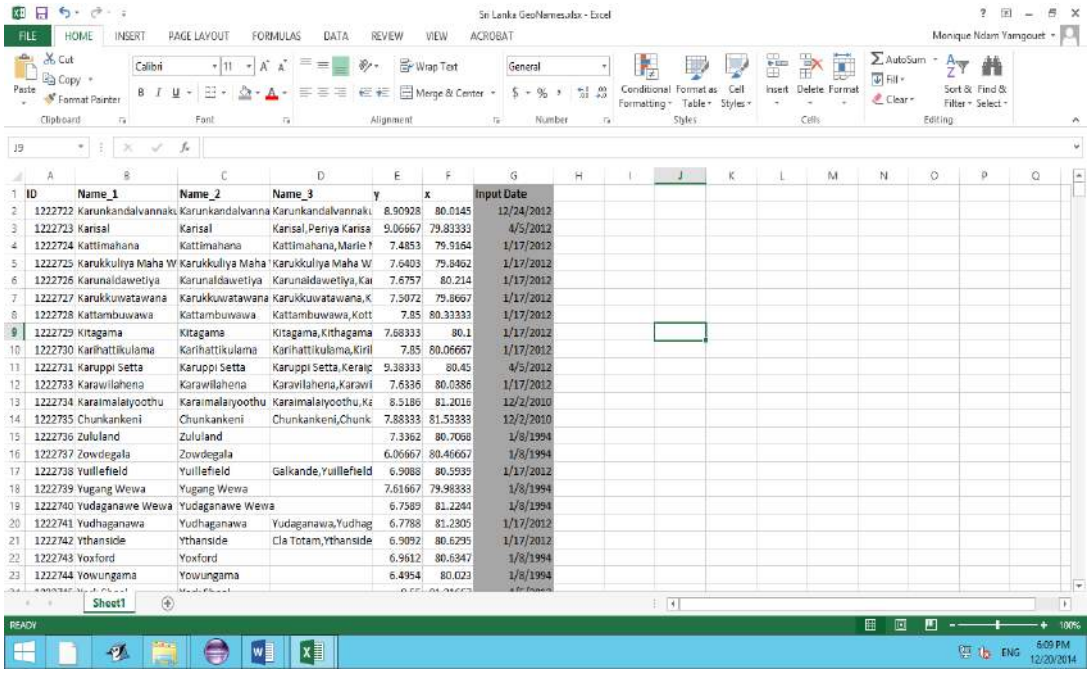


	Name 1	Name 2	Name 2	X	Y	Input Date
2	49530 Ziva Zecur	Ziva Zecundru		-0.51667	42.45	2/25/1994
3	49531 Bur Ziave	Bur Ziave		2.3	43.55	2/25/1994
4	49532 Monte Zhi	Monte Zhi	Bur Zehan	11.58333	51.03333	1/16/2012
5	49533 Zero Cinqi	Zero Cinquanta		-0.83333	41	2/25/1994
6	49534 Zeco	Zeco		5.4	47.51667	2/25/1994
7	49535 Zauel	Zauel	Zauel,Zbu	3.55	43.75	1/16/2012
8	49536 Zamur	Zamur		2.73333	43.4	2/25/1994
9	49537 Yuusufow	Yuusufow	Yuusufow	-0.77548	41.56223	1/16/2012
10	49538 Yuunda	Yuunda	Geeska Yu	-1.21932	41.85377	1/16/2012
11	49539 Yuuga	Yuuga	Yuuga	9.53111	46.29389	1/16/2012
12	49540 Yuubka	Yuubka	Yuubka	2.62358	44.01197	1/16/2012
13	49541 Yusuf Daa	Yusuf Daa	Yusuf Daa	1.5661	41.7207	1/16/2012
14	49542 Yufle	Yufle	Jofle,Joof	10.37591	47.19626	1/16/2012
15	49543 Yuffe	Yuffe	Yuffe	10.3999	47.1944	1/16/2012
16	49544 Yucurun	Yucurun	Las Juhun	10.99321	49.61454	1/16/2012
17	49545 Yubshera	Yubshera		8.4	46.78333	6/5/2011
18	49546 Yubluluc	Yubluluc	Yubluluc	7.77389	49.36361	1/16/2012
19	49547 Yube	Yube	Yube,Yub	10.74789	47.92412	1/16/2012
20	49548 Yubaleh	Yubaleh		9.08333	48.83333	3/16/2009
21	49549 Yubale	Yubale	Yubale,Yu	10.31667	43.45	1/16/2012
22	49550 Yowyaale	Yowyaale	Yowyaale	1.46606	43.9727	1/16/2012
23	49551 Yos	Yos		10.76667	45.66667	2/25/1994

Table 4.6 Somalia Data Input

4.1.3.3.3 Sri Lanka

Data is from 1994-2014



ID	Name_1	Name_2	Name_3	y	x	Input Date
1222722	Karunkandalvannaku	Karunkandalvanna	Karunkandalvannaku	8.90928	80.0145	12/24/2012
1222723	Karisal	Karisal	Karisal,Periya Karisa	9.06667	79.83333	4/5/2012
1222724	Kattimahana	Kattimahana	Kattimahana,Marle I	7.4853	79.9164	1/17/2012
1222725	Karukkuliya Maha W	Karukkuliya Maha	Karukkuliya Maha W	7.6403	79.8462	1/17/2012
1222726	Karunaldawetiya	Karunaldawetiya	Karunaldawetiya,Kar	7.6757	80.214	1/17/2012
1222727	Karukkuwatawana	Karukkuwatawana	Karukkuwatawana,K	7.5072	79.8667	1/17/2012
1222728	Kattambuwawa	Kattambuwawa	Kattambuwawa,Kott	7.85	80.33333	1/17/2012
1222729	Kitagama	Kitagama	Kitagama,Kithagama	7.68333	80.1	1/17/2012
1222730	Karihattikulama	Karihattikulama	Karihattikulama,Kiril	7.85	80.06667	1/17/2012
1222731	Karuppi Setta	Karuppi Setta	Karuppi Setta,Keraic	9.38333	80.45	4/5/2012
1222733	Karawillahena	Karawillahena	Karavilahena,Karawi	7.6336	80.0386	1/17/2012
1222734	Karaimalaiyoothu	Karaimalaiyoothu	Karaimalaiyoothu,Ka	8.5186	81.2016	12/2/2010
1222735	Chunkankeni	Chunkankeni	Chunkankeni,Chunk	7.88333	81.53333	12/2/2010
1222736	Zululand	Zululand		7.3362	80.7068	1/8/1994
1222737	Zowdegala	Zowdegala		6.06667	80.46667	1/8/1994
1222738	Yuillefield	Yuillefield	Galkande,Yuillefield	6.9088	80.5939	1/17/2012
1222739	Yugang Wewa	Yugang Wewa		7.61667	79.98333	1/8/1994
1222740	Yudaganawe Wewa	Yudaganawe Wewa		6.7589	81.2244	1/8/1994
1222741	Yudhaganawa	Yudhaganawa	Yudaganawa,Yudhag	6.7788	81.2305	1/17/2012
1222742	Ythanside	Ythanside	Cla Totam,Ythanside	6.9092	80.6295	1/17/2012
1222743	Yoxford	Yoxford		6.9612	80.6347	1/8/1994
1222744	Yowungama	Yowungama		6.4954	80.023	1/8/1994

Table 4.7 Sri Lanka Data Input

4.1.3.4 Spatial Pattern of Placenames from GeoNames

”Dot mapping is a cartographic presentation method to visualize quantitative absolute values. It is a suitable method for representing spatial distribution patterns” (Kimerling 2009) I want to illustrate the spatial distribution of placenames from the GeoNames data to see the different patterns per the countries chosen. Using ArcGIS, shape files for Cameroon, Sri Lanka and Switzerland are added to the software. For each country I have to create a dot density map from the coordinates in the GeoNames data. Separate excel sheets have been made just with the coordinates (X and Y) of the placenames called Cameroon points.xslx, Sri Lanka points.xslx and Switzerland points.xlsx. After highlighting only the shape file for each respective shape file in Data view mode, I clicked on the Geodatabase symbol on the right side in the Arc Catalogue to import table (single), a pop up window will appear and in the input section choose the required excel file for the respective country shape file, rename the output of the new table and click ok.

The new table will have the name you inserted in the output section of the pop up window. In the Geodatabase section search for the new table and right click on it, select the option create feature class from XY table, another popup window will come up choose the section in the excel file you want to select and use and also select which coordinate system you wish to use and then click ok. The resulting table will be points and then drag and drop this table on the shape file it will emerge into a dot density map. (ArcGIS Desktop Help)

The shape files are in WGS_1984 coordinate system and I have to project each country data into its country’s projection so I used the project tool and converted the map projections to their respective country projections. (ArcGIS Desktop Help)

4.1.3.5 Creating KDE from Dot density maps of placenames to compare with population density map

To create a KDE from a dot density map we need the layer of the dot density map to be highlighted, then we go to the spatial analyst tool and choose kernel density, a pop up window will emerge and in the input field choose the said layer and select the output cell size and click ok, from the KDE layer right click and go to properties, categories and edit as desired. (ArcGIS Desktop Help). The bandwidth and cell size choice are very important in

measuring KDE.

The data from GeoHive will be used to calculate population density with the formula

Population density $\frac{\text{Population}}{\text{Area (Km}^2\text{)}} = \text{Density/Km}^2$ (1)

After haven had the population density for each region or canton, I will create an excel file for each country with the various sections as seen in table 4.7 below, I will join the excel file to the shape file to get the population density map for Cameroon, Sri Lanka and Switzerland (ArcGIS Desktop Help)

Polygon	Region	FIPS	Population	Area	Population Density

Table 4.8 Format for excel file in creating population density map

After haven done so, I will compare the two maps to see if the placename clustering from the KDE is in the like locations as high population densities.

4.1.3.6 Correlation between Population Count and placename count of GeoNames per region or canton

I will use population Count and placename count per region or canton to know if where there are more people we have more placenames and vice versa. To get placename count per region from the dot maps in ArcGIS, right click on the said placename dot density map layer and click join, choose join from another layer based on spatial location, choose save as shape file and fall inside option and then join the data and you will have a new layer containing the join output, open the attribute table and highlight the fields you need (Region or name) and click summarize region and choose the option to add field on the existing data or table , go back and open the new output table to get the list of summary of placename count per each region or canton. (ArcGIS Desktop Help)

When I get this list of placename count per region, I will like to get placename and population count correlation, to know if there is a relationship between population and placename I have to use regression analysis.

A linear Regression analysis will be made simply using excel where placename count will be the independent variable (x) and population count will be the dependent variable (y), I want to know if a change in x will affect y.

4.1.3.7 Distribution of placename in space

Haven gotten placename count per region or canton and total placename per country, I will like to know the correlation of placenames in space for each of the three countries using the formula below

$$\frac{\text{Number of Toponyms}}{\text{Area}} = \text{Toponym/km}^{-2} \qquad (2)$$

4.1.4 Results Analysis

A list of placenames on the local Newspapers not in the list of GeoNames	Dot density showing placename distribution in space for various countries
A list of placenames in the SwissNames names distribution	Kernel density maps describing placename clustering and distribution
Cumulative graph illustrating fluctuations in placename input	Population Density map showing population density on space
	Regression analysis showing relationship between placename and population
	Distribution of placename in space per Country

Figure 4.6 Schematic overview of result analysis

The results will be analyzed under various aspects; Using the results from the placenames list comparisons and those from the mappings. On the side of the name list comparisons we will have a count of names present on the local newspaper which are not present in the gazetteer

data; we will also have a count of placenames which are on the SwissNames that are not in the GeoNames gazetteer data.

I will have three cumulative graphs showing the sequence of placename input into the GeoNames database for Haiti, Somalia and Sri Lanka.

With the maps, I will have three dot density maps for Sri Lanka, Cameroon and Switzerland showing placename distribution on the landscape and I will use the dot density maps to create kernel density maps to show the hotspots for the GeoNames data and then the population density map will show population distribution in space which will be used to compare with placename count to see if there is any correlation between placename count per region and population count per region and this results will be illustrated in a linear graph.

Using placename count per country and total surface area per country I will calculate placename distribution in space per country.

5. RESULTS & INTERPRETATION

In this chapter I am going to show all the results of my research and interpret them, they will be a mixture of lists (from placename comparisons), graphs (data input into the GeoNames gazetteer and Correlation between placename and population) and maps (dot density maps to show placename distribution in space, KDE from the dot density maps and population density maps) this chapter is detail and reproduction of what section 4.1.4 stated.

5.1 Results from List comparisons

Using Excel to conditionally format, differentiate and Highlight values which are on the news list and GeoNames list, those which are not highlighted on the news list are not on the GeoNames list

5.1.1 NZZ and Switzerland GeoNames

Switzerland GeoNames	NZZ News (April)	Number of names not in GeoNames
23,599	107	75

Table 5.1 Comparisons between NZZ and Switzerland GeoNames

From the above results we see that out of 107 names found on the NZZ for a period of 30days, only 32 were found in the GeoNames gazetteer and 75 were not found.

Summarily only 29.9 % (32/107 x 100) of the names found on the NZZ were found in the GeoNames database and we can therefore confidently say that the GeoNames data is incomplete.

5.1.2 Cameroon Tribune and Cameroon GeoNames

Cameroon GeoNames	Cameroon Tribune News (April)	Number of names not in GeoNames
21,339	226	134

Table 5.2 Comparisons between Cameroon Tribune and Cameroon GeoNames

From the above results we see that out of 226 names found on the Cameroon Tribune for a period of 30days, only 92 were found in the GeoNames gazetteer and 134 were not found

Summarily we only 40.7% (92/226 x 100) of the names found on the Cameroon Tribune were found in the GeoNames database and we can therefore confidentially say that the GeoNames data is incomplete.

5.1.3 Sri Lanka Daily News Online and Sri Lanka GeoNames

Sri Lanka GeoNames	Sri Lanka Daily News Online (April)	Number of names not in GeoNames
47,146	247	138

Table 5.3 Comparisons between Sri Lanka Daily News Online and Sri Lanka GeoNames

From the above results we see that out of 247 names found on the Sri Lanka Daily News for a period of 30days, only 109 were found in the GeoNames gazetteer and 138 were not found

Summarily we only 44.1% (109/247 x 100) of the names found on the Cameroon Tribune were found in the GeoNames database and we can therefore confidentially say that the GeoNames data is incomplete.

From the analysis of the three newspapers, we saw from the results the number of placenames not found in the GeoNames database but care should be taken that some names are found on the gazetteer but they are in different form for example in Cameroon Tribune we had Mbam-et-Djerem which was marked not found in the gazetteer list because it was there as Mbam Djerem National Park.

5.1.3 SwissNames and Switzerland GeoNames

Switzerland GeoNames	SwissNames
23,599	156,755

Table 5.4 Comparison between SwissNames and Switzerland GeoNames

As seen from the table above, the data from GeoNames gazetteer has just about 15% (23,600/156,755 x 100) of the number of names on the SwissNames as such there was no need to do a comparison, this is also very evident that the GeoNames gazetteer is incomplete and unreliable; one can search for a placename and not find it there.

5.2 Results from placenames data input in GeoNames

5.2.1 Haiti

Date	Number of placename input	Cumulative total
1993	------	11466
1994	1024	12490
1995	54	12544
2000	414	12958
2003	3	12961
2006	100	13061
2007	8	13069
2009	3	13072
2010	42	13114
2011	12	13126
2012	2123	15249
2013	326	15575
2014	47	15622

Table 5.5 Placename input Haiti

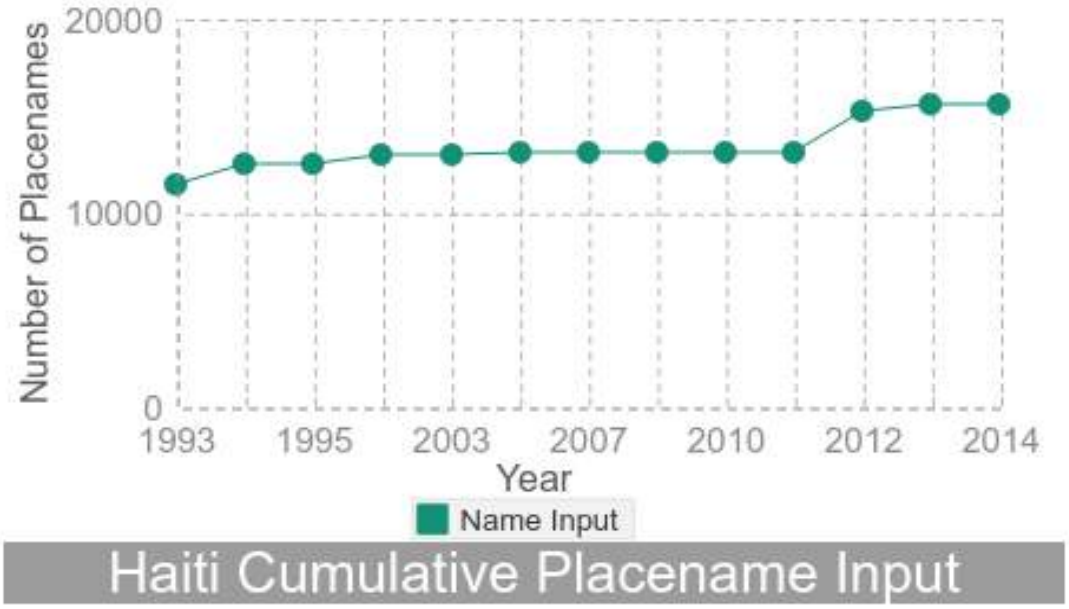


Figure 5.1 Haiti Cumulative placename input

From the above table, we see the number of placename input on the second column and the cumulative total on the third column. This data set ranges from 1993-2014(April) when the data was downloaded from the database. The table illustrates that during this time frame, the lowest placename input was in 2003 and 2009 having a total of 3 placename input, there was a steady increase until 2011 while the highest placename input was in 2012 having a total of 2,123 placename inputs and I can say that this was an after effect of the Haiti earthquake in 2010 where during this time people whose lives were endangered were twitting and posting information on social media[20] and we could say that new and unknown places in Haiti came into light during and after the earthquake and after the recovery mostly the inhabitants and rescue teams and services uploaded more data into the GeoNames database led to the enormous input and after 2012 the input was steady again.

As such we can say that there is a bias in placename input or construction of the GeoNames gazetteer, as in times of no emergencies there is very little input of data and after emergency cases we have very high data input.

[20] http://voices.nationalgeographic.com/2012/07/02/crisis-mapping-haiti/

5.2.2 Sri Lanka

From Table 5.5 below, we see the number of placename input on the second column and the cumulative total on the third column. The data set ranges from 1994-2014 (April) when the data was downloaded from the database. The table illustrates that during this time frame, the lowest placename input was in 2003 where they had no placename input into the database and the highest placename input was in 2012 with 17, 388 names added.

Sri Lanka is one of those countries that have experience long-term civil wars which have affected the country politically and economically, and from Figure 5.2 we see a drastic increase in placename input into the data and this is because of the end of the civil war in 2009. ”When, after 26 years of military campaign, the Sri Lankan military defeated the Tamil Tigers in May 2009 bringing the civil war to an end[21]”.

The total number of placenames in the gazetteer for Sri Lanka sum up to 47,146 which is close to what is on Table 5.6 :47,125 which gives a difference of 21 placenames and this are placenames with no input dates.

Date	Number of placename input	Cumulative total
1994	--------	15890
1996	95	15985
1998	748	16733
1999	183	16919
2000	38	16954
2003	0	16954
2007	58	17012
2008	0	17012
2010	12191	29203
2011	31	29234
2012	17388	46622
2013	411	47033
2014	92	47125

Table 5.6 Sri Lanka placename cumulative input

[21] http://en.wikipedia.org/wiki/Sri_Lankan_Civil_War

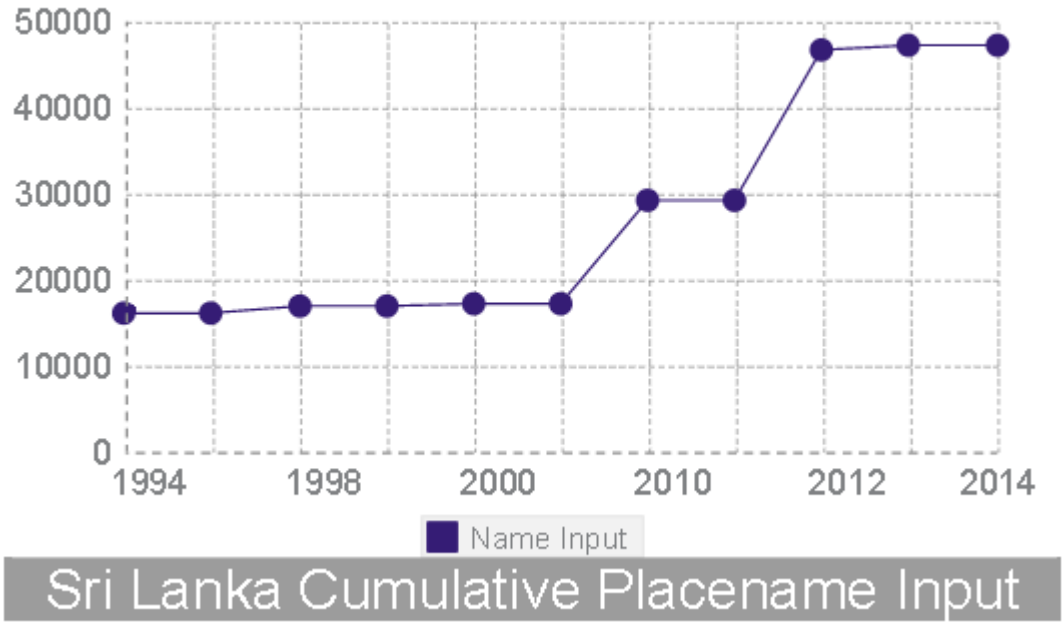


Figure 5.2 Sri Lanka Cumulative placename input

5.2.3 Somalia

Date	Number of placename input	Cumulative total
1994	-----	1944
1996	0	1944
1997	2	1946
2001	66	2012
2006	21	2033
2007	45	2078
2008	121	2199
2009	559	2758
2010	6	2764
2011	1199	3963
2012	13258	17221
2013	56	17277
2014	141	17418

Table 5.7 Somalia cumulative placename input

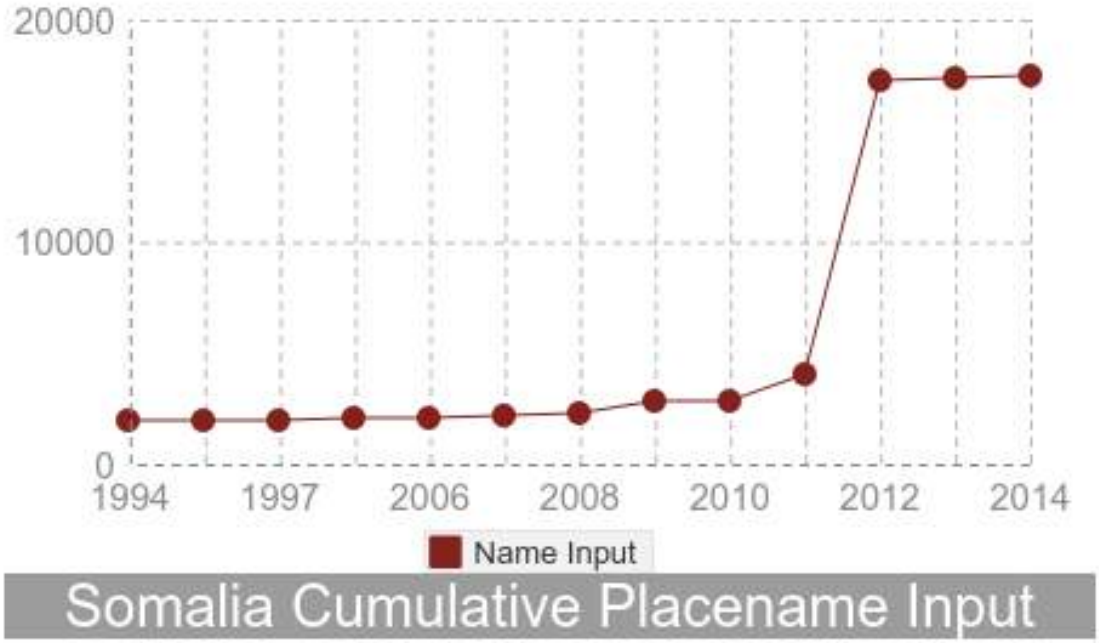


Figure 5.3 Somalia cumulative placename input

From Table 5.6 above and graph, we see the number of placename input on the second column and the cumulative total on the third column. The data set ranges from 1994-2014 (April) when the data was downloaded from the database. The table illustrates that during this time frame, the lowest placename input was in 1996 where they had no placename input into the database and the highest placename input was in 2012 with 13, 258 names added.

Just like Sri Lanka, Somalia also has a long history of civil wars even till date. We see a drastic increase in placename input in 2012 due to the commencement of the joint military operation between the Somali military and multinational forces in 2011[22], 2012 also came with a lot of political reformations in Somalia and this law and order can account for the high placename input.

[22]http://en.wikipedia.org/wiki/Somali_Civil_War

5.3 Results from spatial distribution of placenames in GeoNames

The maps in Figure 5.4, 5.5, 5.6 have the following details to their creation

Details	**Cameroon**	**Sri Lanka**	**Switzerland**
Geometry	Point	Point	Point
Projected Coordinate System	Douala_1948_AEF_West	SLD99_Sri_Lanka_Grid_1999	CH1903_LV03
Projection	Transverse Mercator	Transverse Mercator	Hotline_Oblique_ Mercator_Azimuth_Center
Geographic Coordinate	GCS_Douala	GCS_SLD99	GCS_CH1903
Dot size	4pt	4pt	3pt
Number of points	21,338	47,146	23,559

Table 5.8 Summary of dot density map information

From the maps below, ne dot represents one placename as such they are one-to-one dot maps and we see dots spread across the landscapes: at least all regions on the maps have placenames.

5.3.1 Switzerland

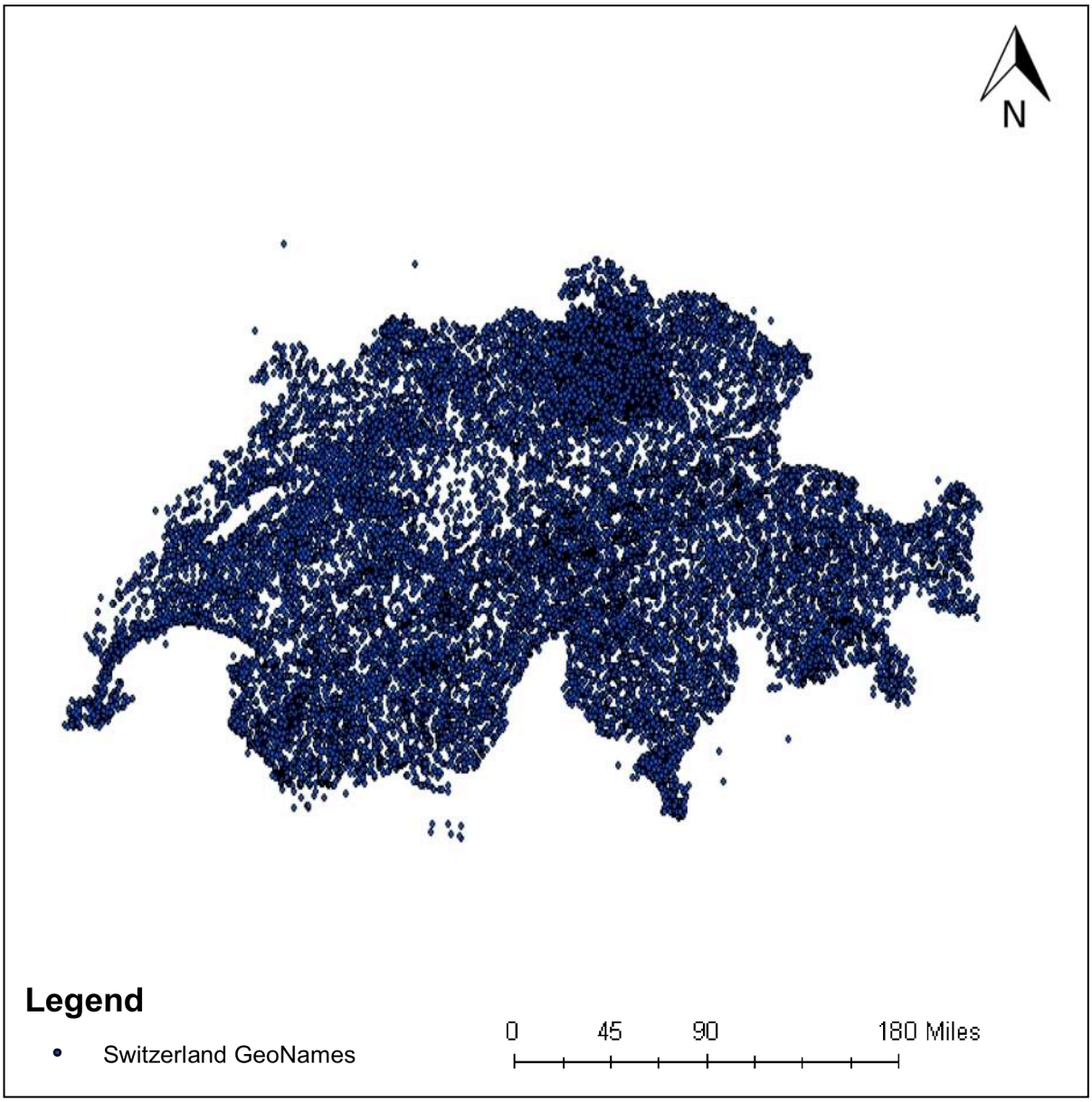


Figure 5.4 Placename distributions from GeoNames Data for Switzerland

From the map in Figure 5.4 above, we see dense clusters of placenames around Zürich Uri and Graubünden and empty spaces around Lucerne. Also from Uri to St Gallen we patches of clusters in between the empty spaces. The placenames points are neither unevenly distributed or show any pattern. In table 5.7 we see that Switzerland has a total surface area of 41,285km^2 and a total of 23, 599 placename points, but as seen from the map above, there are some placename points out of the Switzerland defined boundary: 301 points in total, further

proof of this will be given in section 5.8 were we see placename count per region on the map from the attribute table.

5.3.2Cameroon

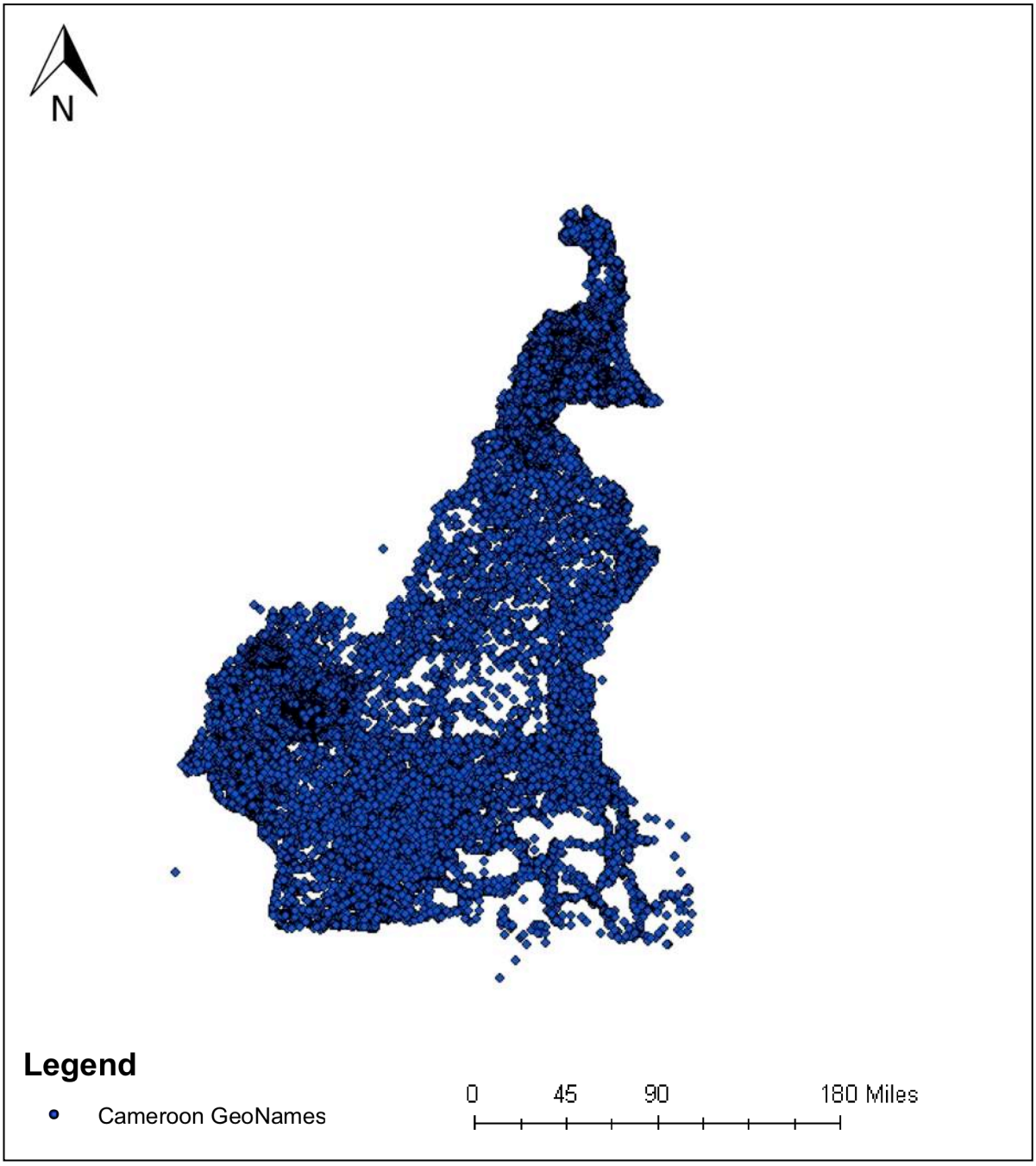


Figure 5.5 Placename distributions from GeoNames Data for Cameroon

From the map in Figure 5.5 above, we see large empty spaces in between the Center region and the southeast regions while there are dense clusters of points in the Far North region and also in between South West, North West and West regions refer to chapter 3.2.1 for names of

various regions. The placename points are neither unevenly distributed or show any pattern. I n table 5.7, we see that Cameroon has a total surface area of 475,440km^2 and a total of 21, 338 placename points, but as seen from the map above, there are some placename points out of the Cameroon defined boundary: 448 points in total, further proof of this will be given in section 5.8 were we see placename count per region on the map from the attribute table.

5.3.3 Sri Lanka

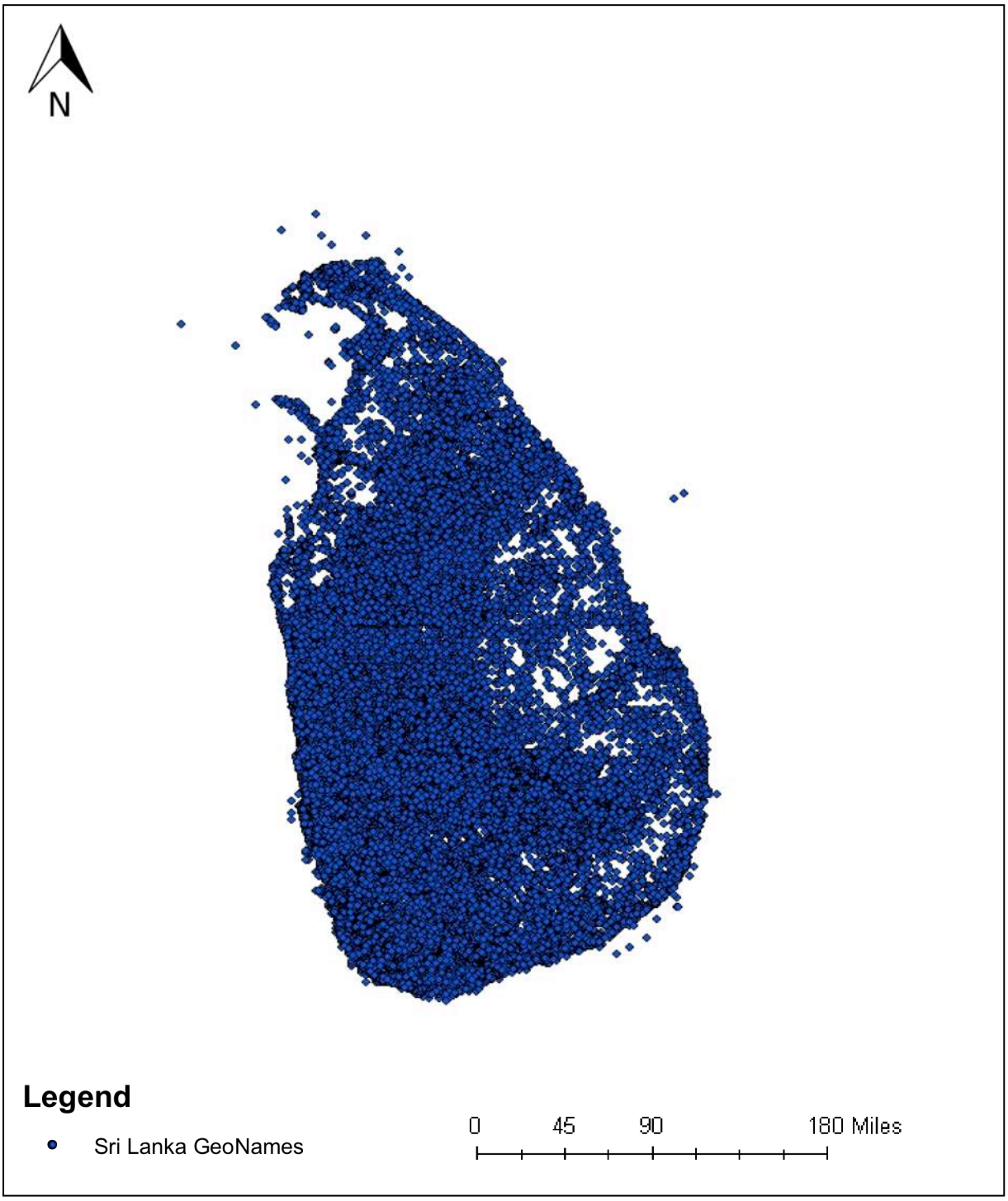


Figure 5.6 Placename distributions from GeoNames Data for Sri Lanka

From the map above, we see a clear cut line between the clustered and sparse regions, we see large spaces with no placenames in the Eastern and Northern regions while we see dense clusters around the Central, Western, North West and Sabaragamuwa regions, refer to Chapter 3.2.3 for names of various regions. The placename points are neither unevenly distributed or show any pattern. I n table 5.7, we see that Sri Lanka has a total surface area of 65,610km2 and a total of 47,146 placename points, but as seen from the map below, there are some placename points out of the Sri Lanka defined boundary: 1708 points in total, further proof of this will be given in section 5.8 were we see placename count per region on the map from the attribute table.

5.4 Kernel Density Estimation

In this section, a brief description of the source of the KDE will be given and the comparison between KDE and population density will be explained in section 5.6.

From the maps below in Figures 5.7, 5.8 and 5.9, the darkest colours represent the areas with the highest surface values and as the colours diminish with increasing distance from the darkest points, so do the value decrease (ArcGIS Desktop Help)

The maps have been made from the points in the dot density maps and the aim is the see where the placename points are clustered: the placename density which can be said as total placename per region or canton divided by the surface area of the canton: gives the amount per unit area

According to Jones et al., 2008. KDE can be calculated using the formula below, which explains that the KDE is the output of the number of points in a certain defined polygon divided by the size of the polygon.

$$pq = \frac{\Sigma pi^{\epsilon C}(q,r)}{\pi r^2} \qquad (3)$$

"Where: *rq* is the density at some location *q; C (q,r*) is a circle centred on q with a radius r; and *pi* are values at points contained within the circle *C(q,r)"* .(Jones et al.,2008)

" Kernel density estimation (KDE) methods add a weight to the values at points *pi* according to their distance from q which smooths the influence of points with distance so that points nearer to q have a greater influence on the density value" (O'Sullivan and Unwin 2003).

From the KDE maps below. The colours depend on one underlying factor: the number of points. The more points in an area the darker the colours and as stated above, in this thesis the KDE is the count of placenames in a region divided by the surface area of that region. For the KDE, if I multiply the KDE/area by the size of the output cell we will get the number of placenames in that cell but that is not the focus of this thesis.

5.4.1 Switzerland

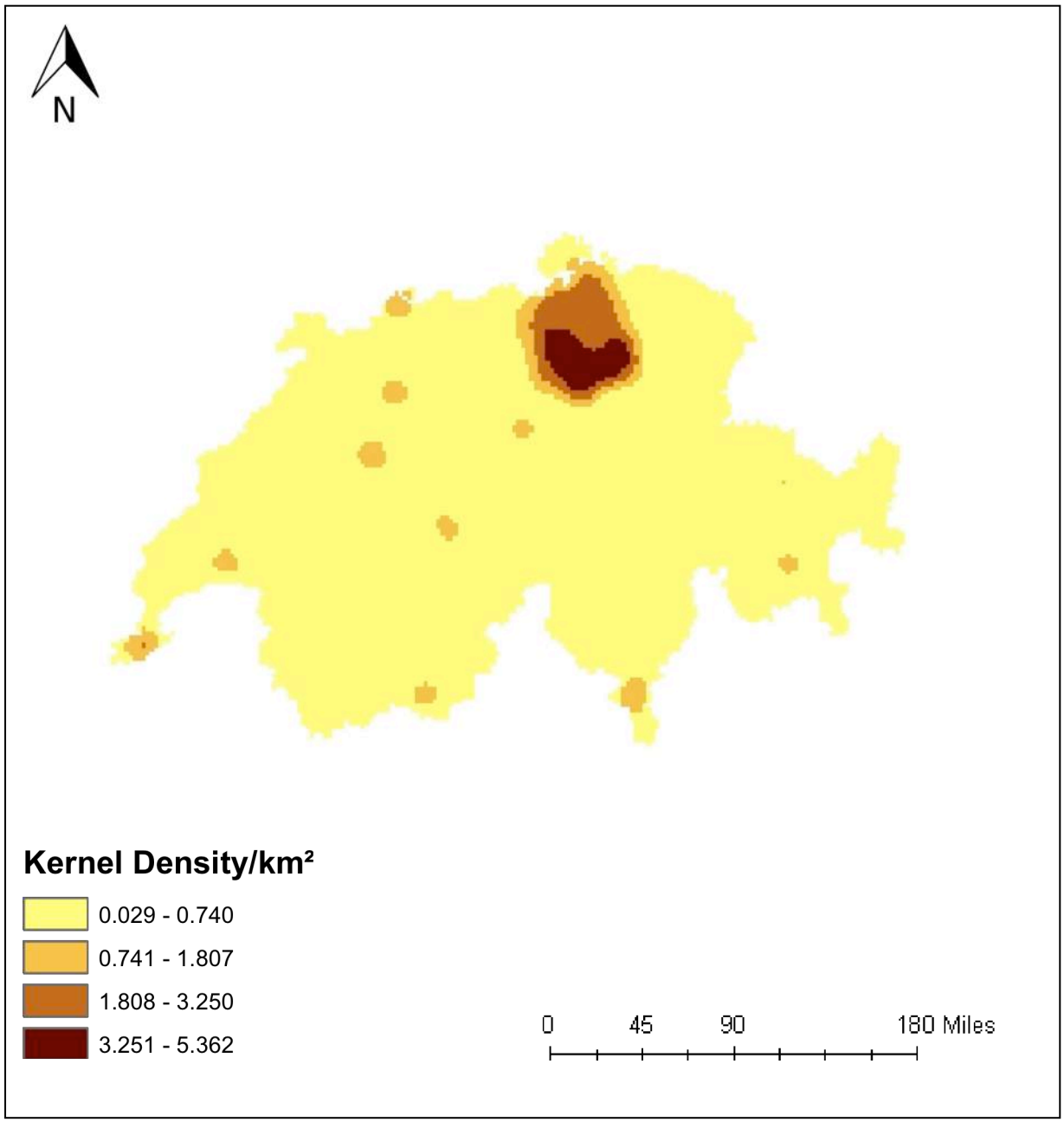


Figure 5.7 Kernel Density Switzerland

Query	Measurement
Data Type	File Geodatabase raster data set
Cell size (x,y)	1880km,1880km
Min	0.0299
Max	5.3624
Std Dev	0.5823
Mean	0.544

Table 5.9 Summary of KDE for Switzerland

5.4.2 Cameroon

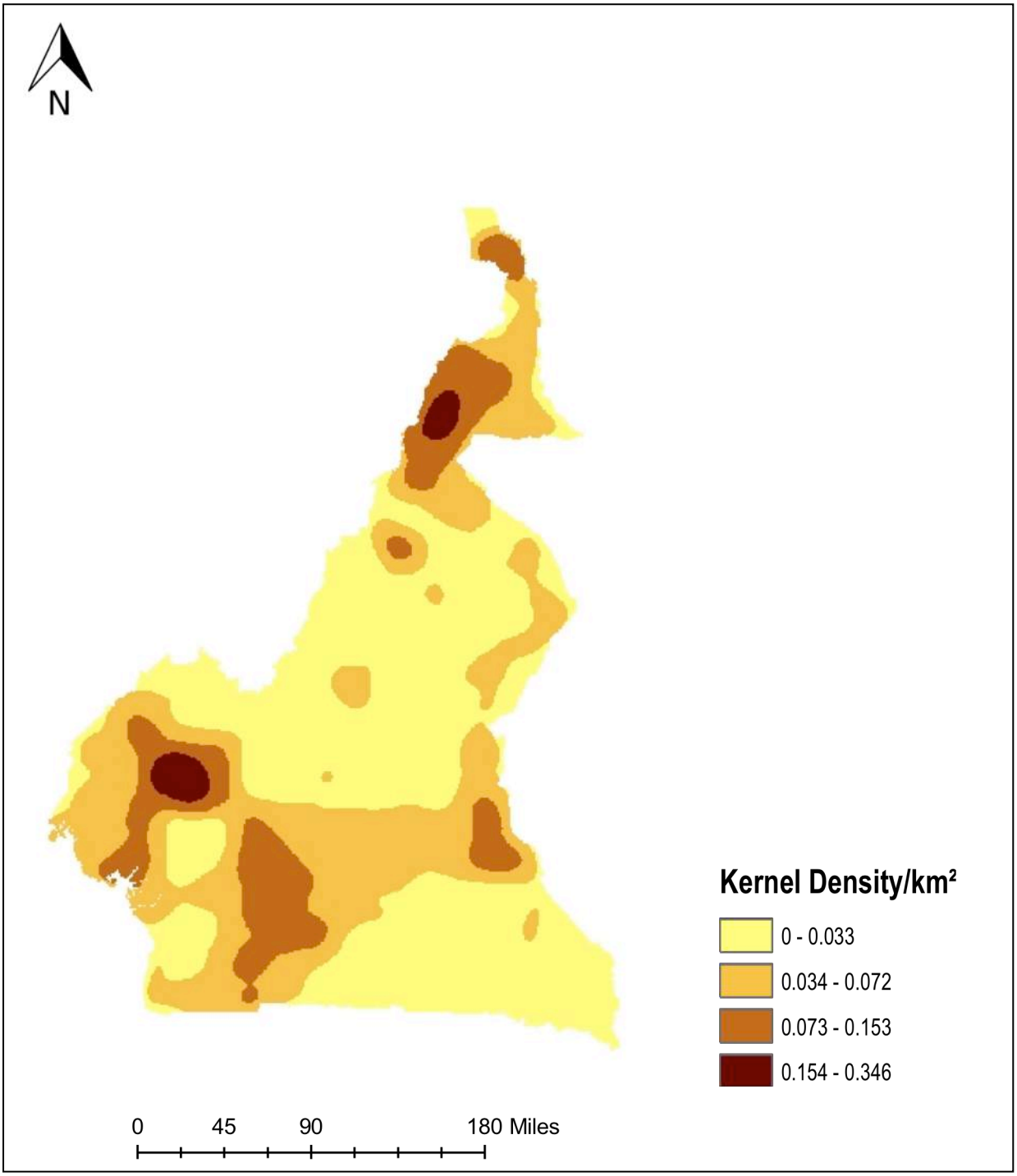


Figure 5.8 Kernel density Cameroon

Query	Measurement
Data Type	File Geodatabase raster data set
Cell size(x,y)	1880km,1880km
Min	0
Max	0.3465
Std Dev	0.0362
Mean	0.423

Table 5.10 Summary of KDE for Cameroon

5.4.3 Sri Lanka

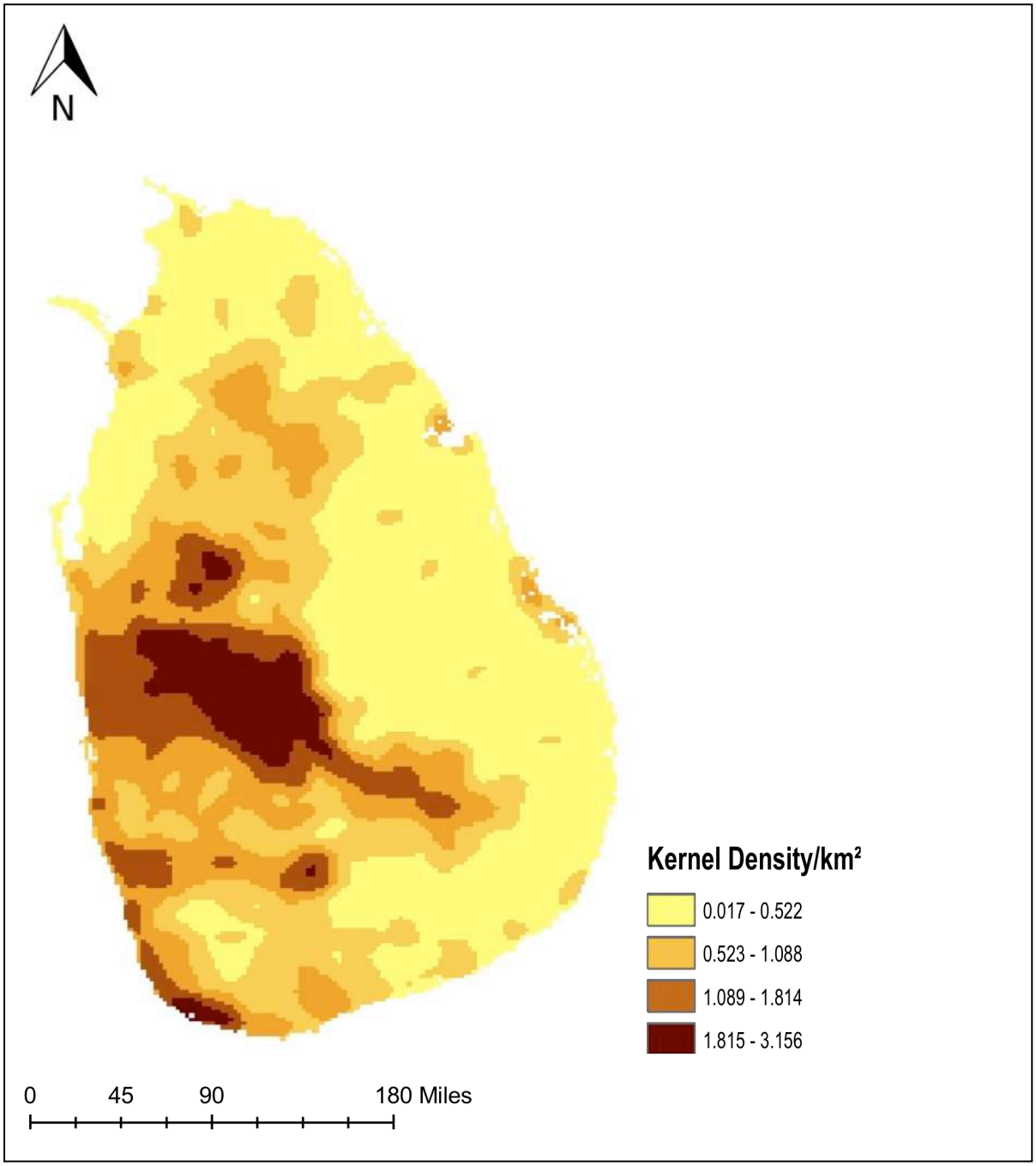


Figure 5.9 kernel density Sri Lanka

Query	Measurement
Data Type	File Geodatabase raster data set
Cell size(x,y)	1880km,1880km
Min	0.0177
Max	3.1566
Std Dev	0.5617
Mean	0.6803

Table 5.11 Summary of KDE for Sri Lanka

5.5 Population density

Population density a measurement of the number of people in an area. It is an average number and is calculated by dividing the number of people by area. Population density is usually shown as the number of people per square kilometer: the number of people relative to the space occupied by them. The maps below are choropleth (shading) maps and illustrates population density. The darker the colour the greater the population density: certain factors affect population density, which range from human to physical.

5.5.1 Switzerland

From my results below, it is evident that canton Basel-Stadt has highest population density in Switzerland for 2013 census data with up to 5107 people per square meter, it is the smallest canton with only 37.0km², followed by canton Geneva with 1662/km², Zurich is the most populated canton with 1,425,538 followed by Bern with 1,001,281 people but they don't have the highest population densities and this is because the surface area of the region affects the population density results. We see areas with second high population density in Zurich, Zug, Aargau and Basel-Landschaft.

We can also see that the map above has the same pattern with the Switzerland topographic map: the highly populated regions are just about 700m maximum and these are the areas on the north side of the map while the south has the lowest population density and it has areas with 200m landscape.

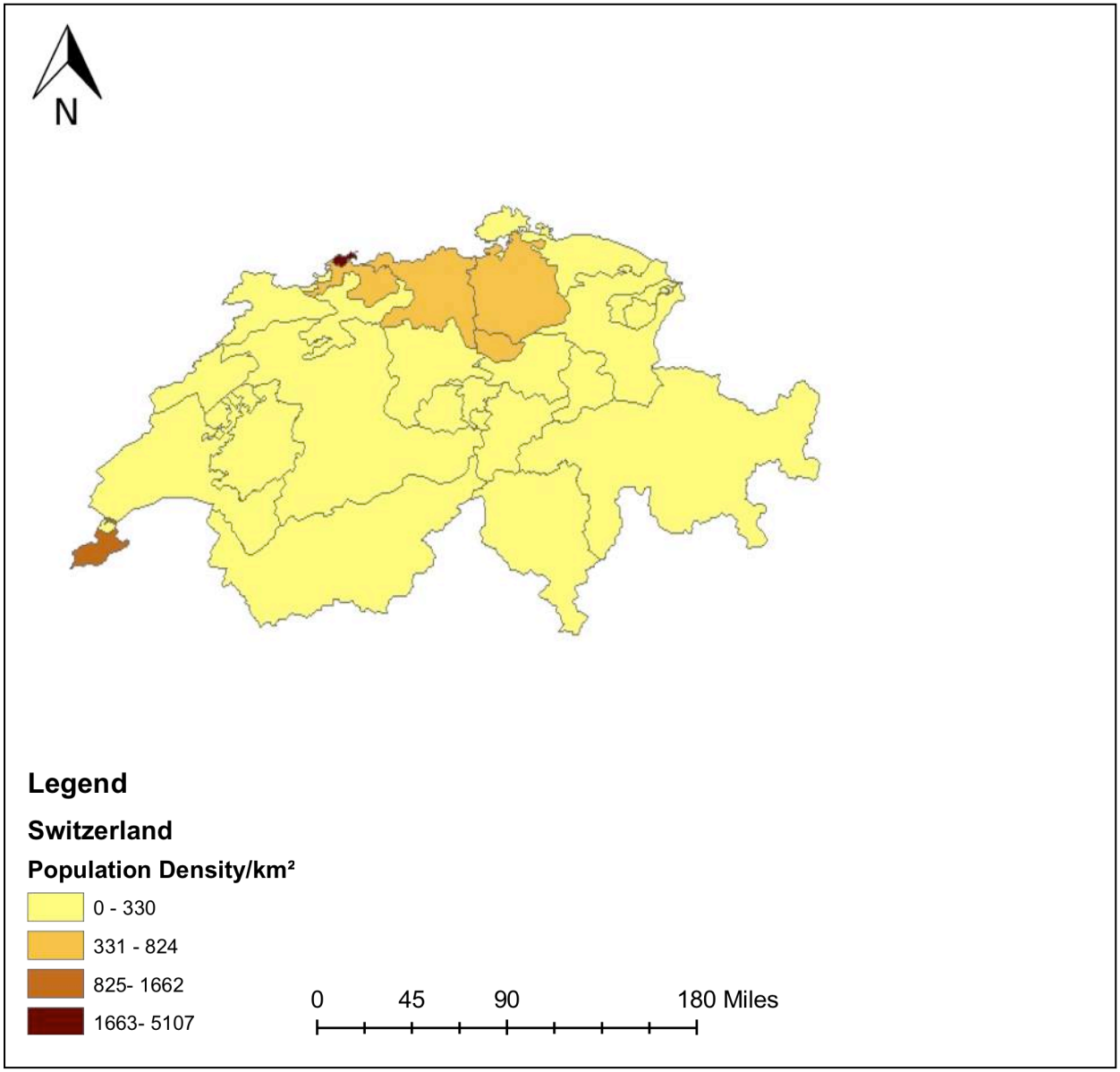


Figure 5.10 Population Density of Switzerland 2013.

5.5.2 Cameroon

From the map below, we see the highest population densities in the West and Littoral regions with up to 141 people per square meter followed by the Northwest region and the Far North, the regions with the least population density are the South, East and Adamawa regions. The Center region is the most populated region in Cameroon with up to 3,525,664 people but it doesn't have the highest population density, the Far North region is the second most populated region and it falls in the second rank of high population density zone.

The center region has a large surface area and the highest population in the country but it has just up 54/km² people, the littoral and west regions are the smallest regions and this accounts

for the high population density, the Littoral region has the economic capital of the country: Douala and it is the third most populated region with about 2.8 million people.

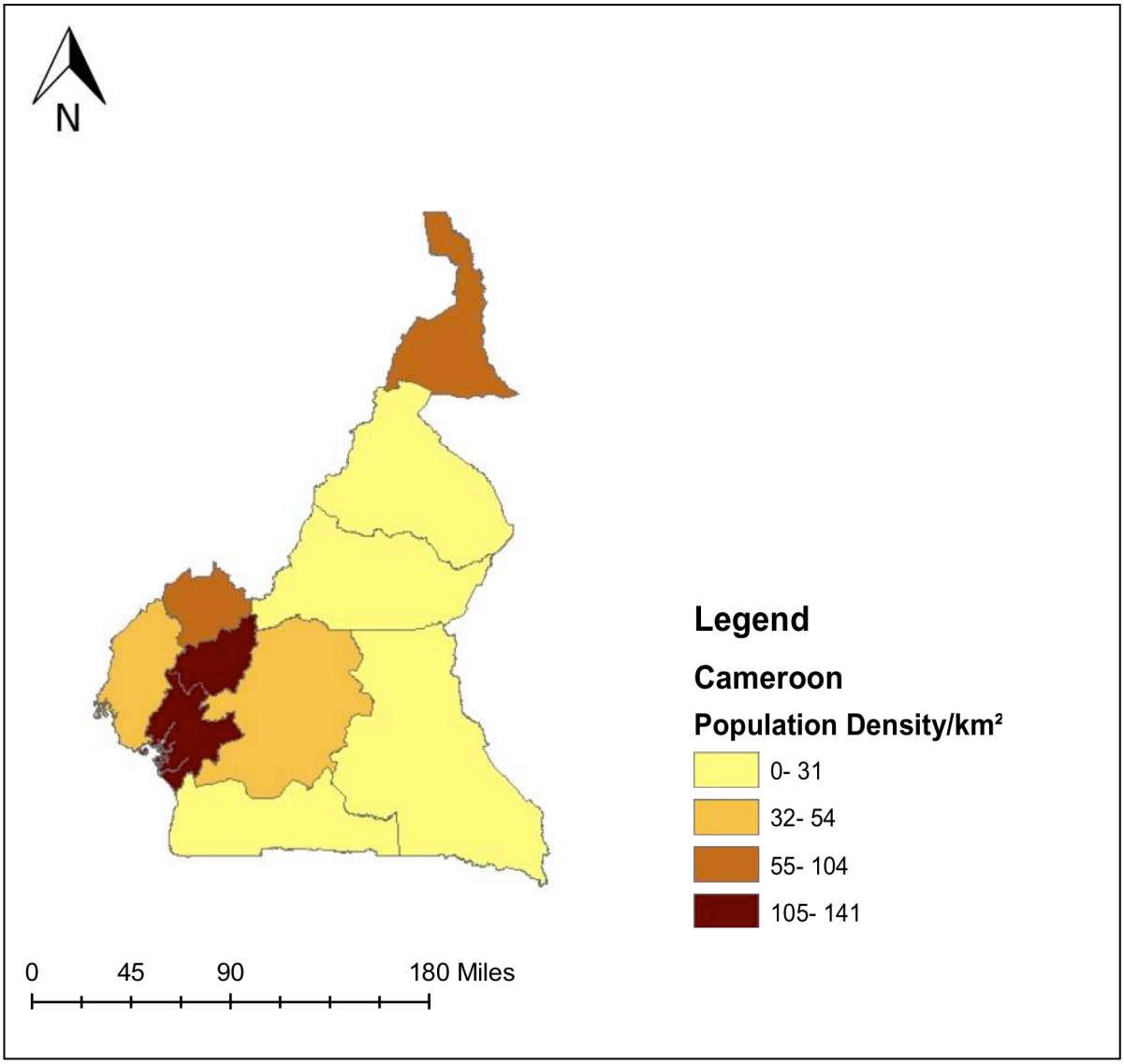


Figure 5.11 Population densities for Cameroon 2010

5.5.3 Sri Lanka

From the map below, the Western region which contains both the country's administrative (Sri Jayewardenepura Kotte) capital and commercial capital (Colombo) has the highest population with a total of 5.8 million people and a population density of up to 1580 people per square meters, we see a cluster of high population density of up to 452 people per square meters around the capital province in the Northwestern region, Sabaragamuwa, South and Central regions and we see that the least populated provinces are the North central(the largest province with 10,472km²) and Northern provinces.

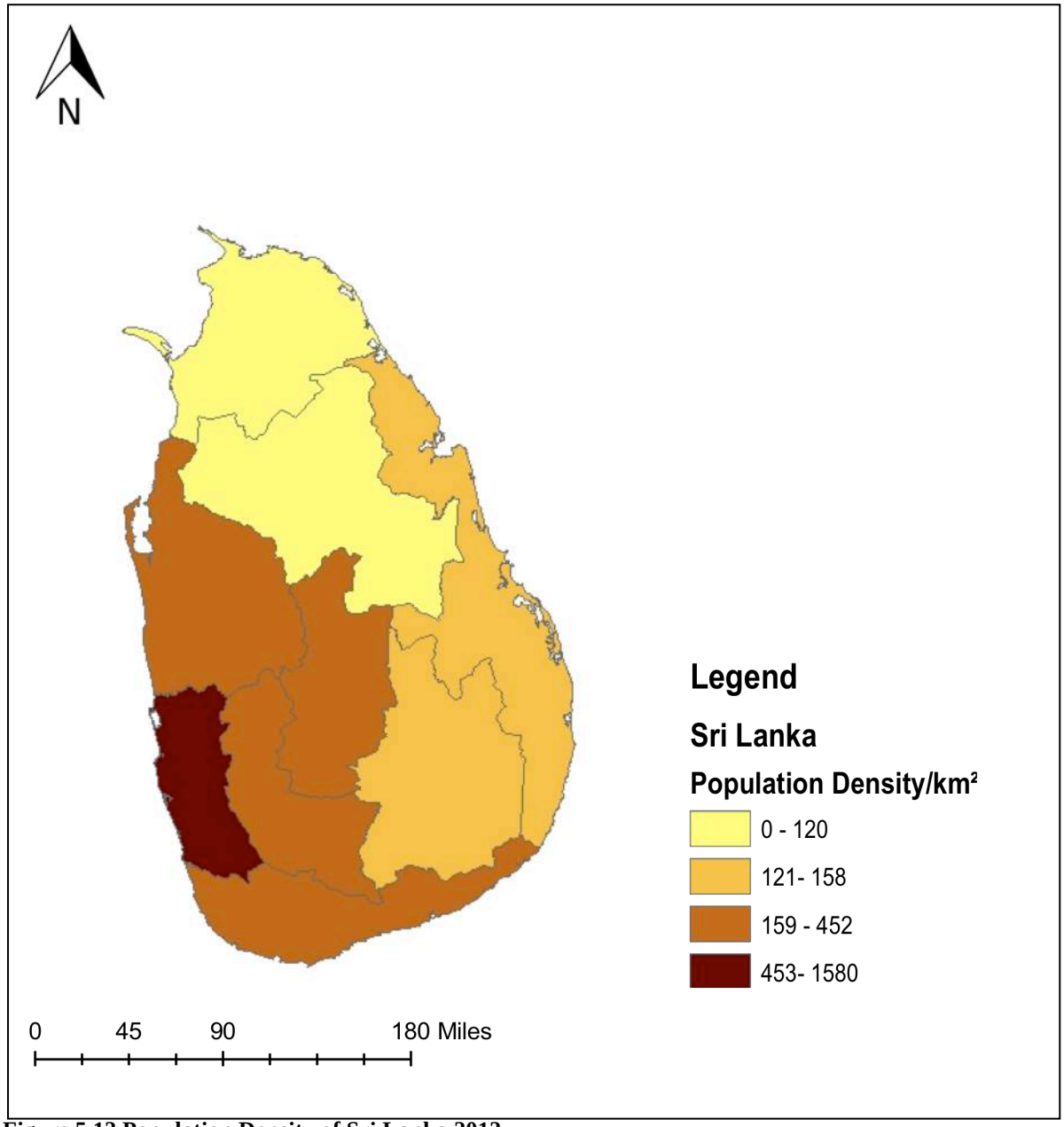


Figure 5.12 Population Density of Sri Lanka 2012

5.6 Population density vs. KDE in GeoNames

The maps of the KDE are results of the placename count divided by the size of the cantons, we see very low figures for the KDE because of the cantons are large and they have few placenames and the out put is a low placename count per region. For population density it is a an output of the number of people divided by the surface area. What I want to compare is to see if the plecename density and the population density have hotspots on the same locations:. Section 5.8 has details of this comparison.

5.6.1 Switzerland population density and Switzerland KDE

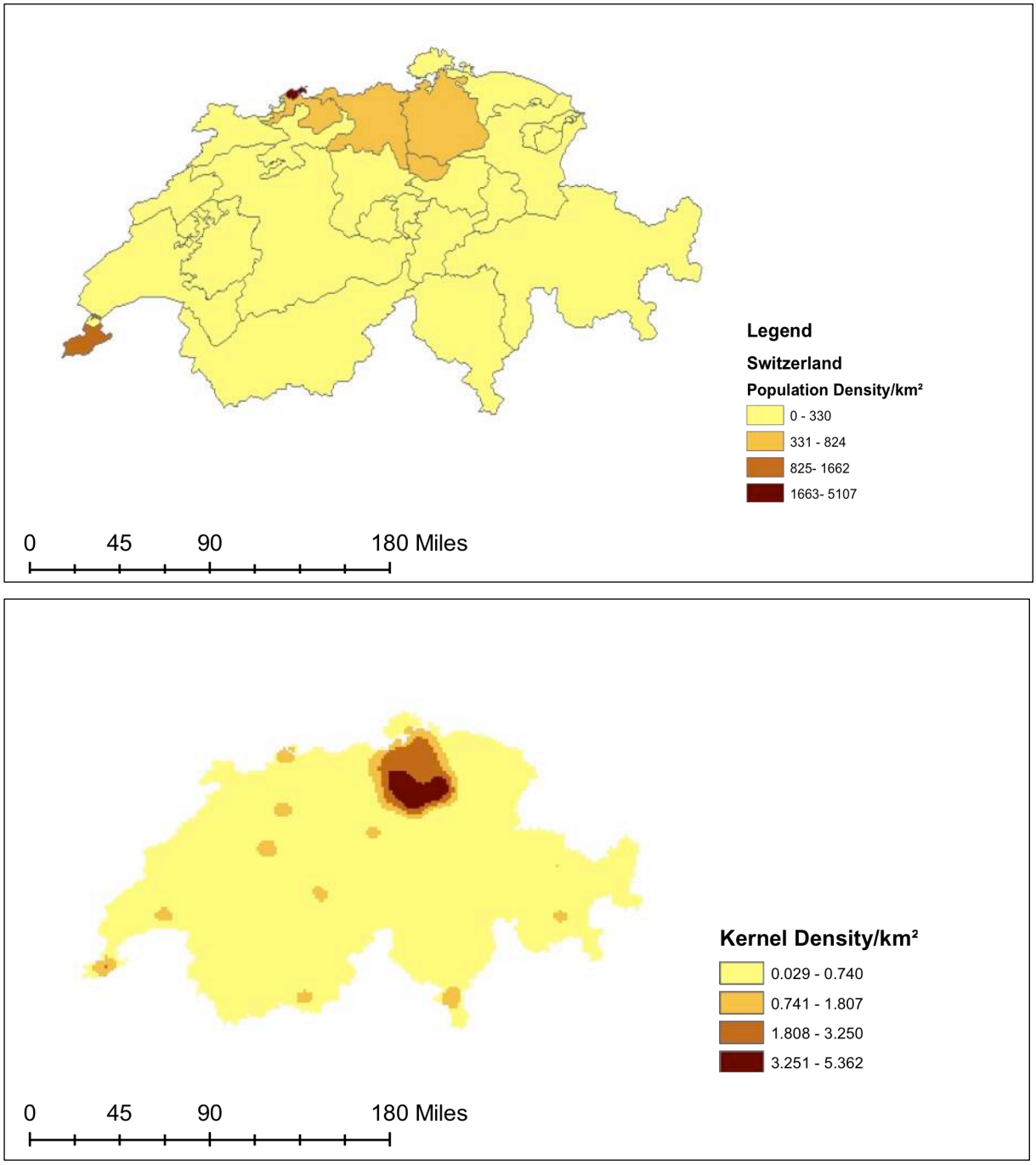


Figure 5.13 Comparison between placename and population density for Switzerland

We see population density and placenames clusters on the two maps: the darker colored sections have more points and the light colors indicate fewer points. Generally, we see a very little link between the placename density and the population density; we don't see any regions on the two maps having the same hotspots, for the population density map we could link its density to social and economic activities which attracts people to this area, on the other hand the placename points shows that there highest clusters are around Zurich which is

the most populated city in Switzerland. For the lighter shades we see a relation on the south side of both maps.

5.6.2 Cameroon population density and Cameroon KDE

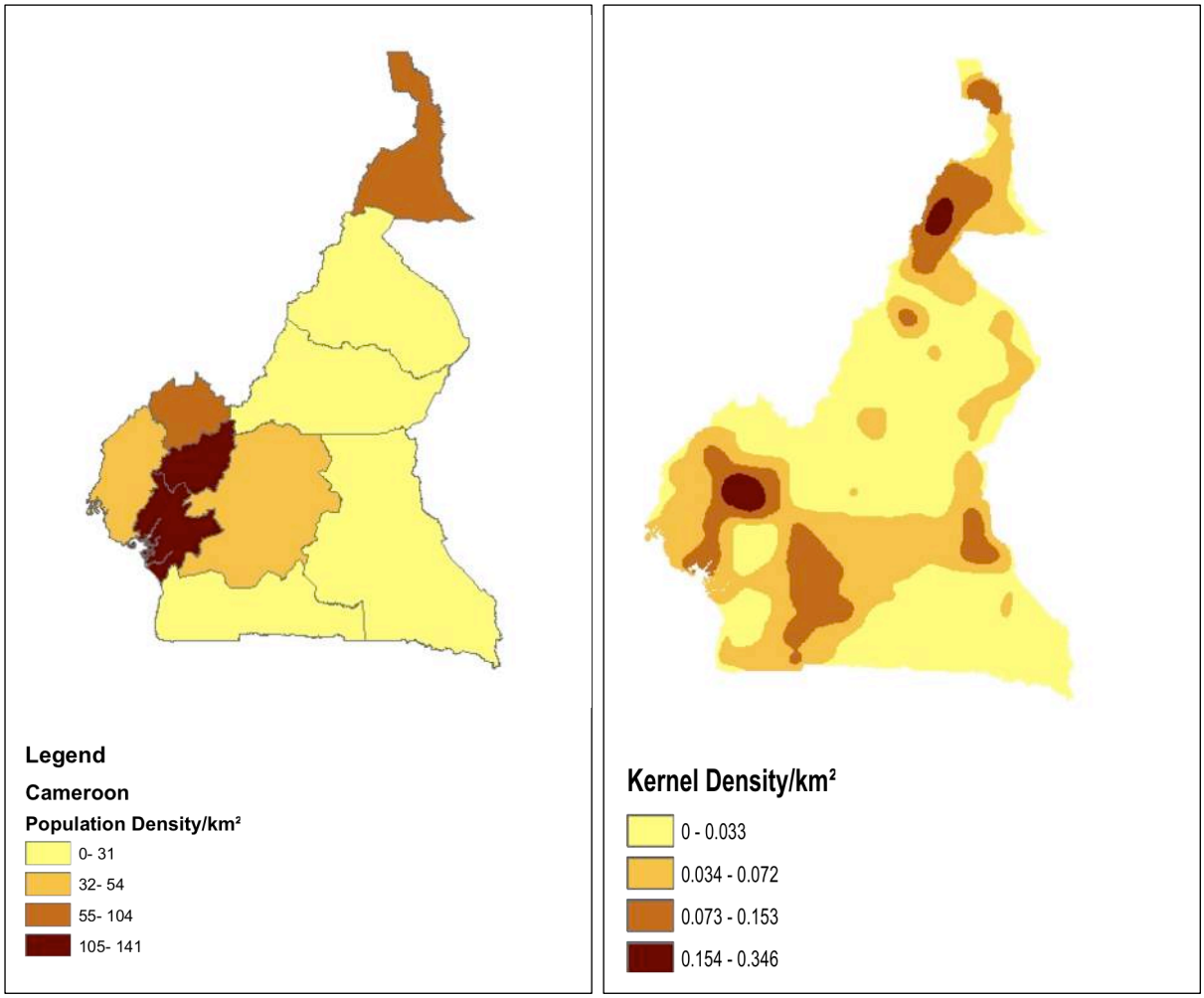


Figure 5.14 Comparison between placename and population density for Switzerland

The map on the left illustrates population density for Cameroon, while that on the right illustrates KDE from the placename points. The darker spots indicate more points while the lighter points indicate fewer points. Looking at these maps closely, we see a like hot spots on the two maps; there is a hotspot for placenames the West and Littoral region which is also the most populated region in the country meanwhile the Southeast and the Adamawa plateau which are the least populated regions also has the least number of placenames, therefore we can say that there is a relationship between placename and population hotspots for Cameroon: where there are more people we have more placenames and where there are less people we have less placenames.

5.6.3 Sri Lanka population density and Sri Lanka KDE

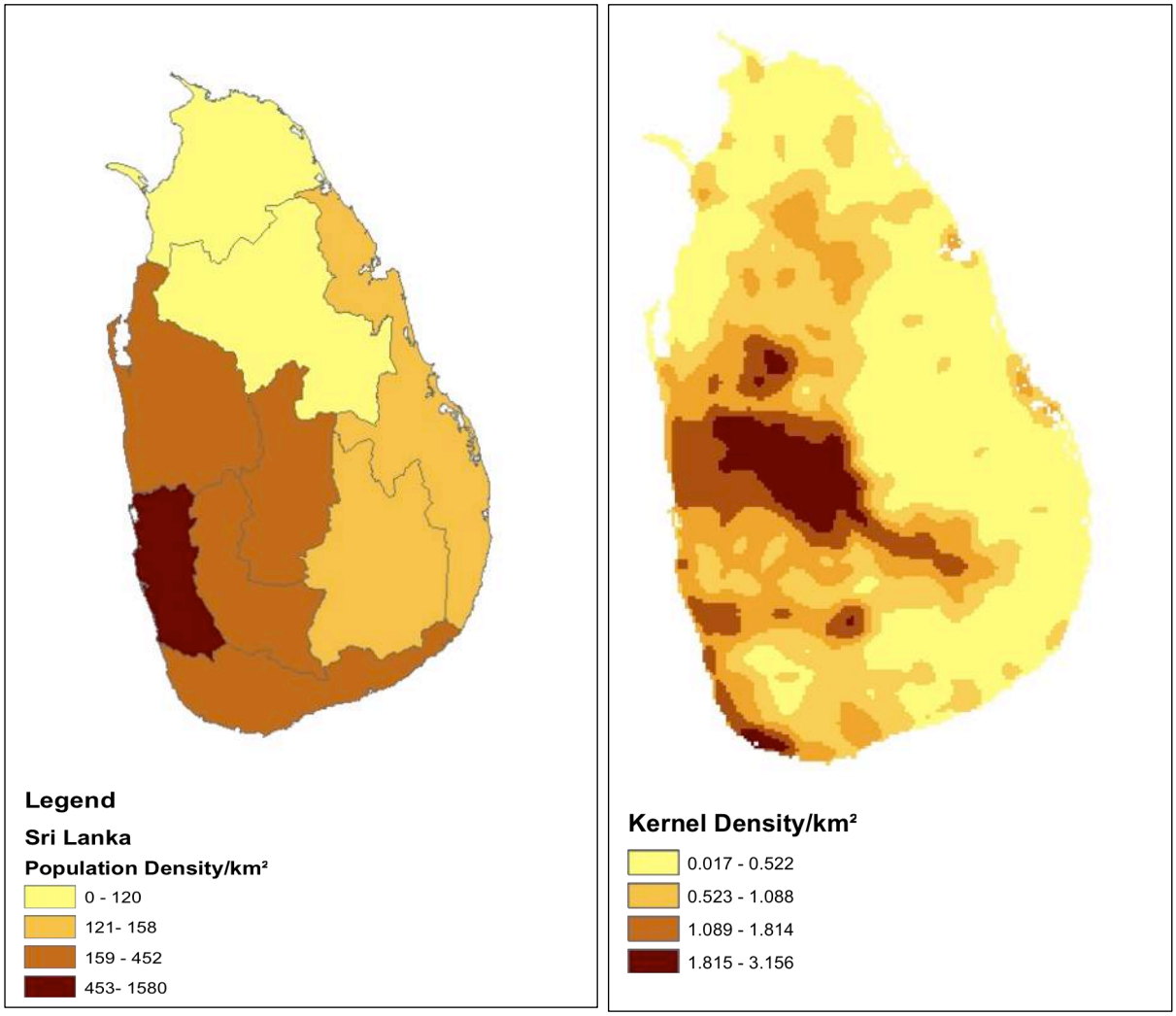


Figure 5.15 Comparison between population and placename for Sri Lanka

The map on the left illustrates population density for Sri Lanka, while that on the right illustrates KDE from the placename points. The darker color indicates more points while the lighter color indicates fewer points. Looking at these maps closely we do not see a relationship between the most populated province the Western province which is the country's capital and the placename map, we instead see five placename hotspots in different regions: one in the Southern region around Galle, another in Sabaragamuwa, the largest hotspot cuts across the Central region the Northwestern and the North central regions, the North of the North Central region has two hotspots as well. There is a relationship in the Northern region for placename and population as it has the least population density and the least placename clusters.

5.7 Distribution of placenames in Space

In addition, I also want to know the distribution of placenames in space for each country: how many placenames can be found per kilometer and which country has more placenames in space. I will use the formula below to do this analysis.

$$\frac{\textit{Density of TOP}}{\textit{Area}} = km\text{-}^2 \qquad (4)$$

Country	Number of Placenames	Surface area	**Distribution in Space**
Cameroon	21,338	475,440km²	0.044km-²
Sri Lanka	47,146	65,610km²	0.718km-²
Switzerland	23,599	41,285km²	0.571km-²

Table 5.12 Placename correlation in space

From the above analysis, we can say that placenames are unevenly distributed on space per country, the patterns of distribution are different, some countries have more placenames per km unlike others, and the surface area of the country could be an influence, how can you analyze this? The results show the total number of placenames found on each kilometer where these countries are found we have more placenames on space in Sri Lanka with the highest number of placenames of the three countries and the least number of placenames per kilometer from Cameroon with the largest surface area of the three countries.

5.8 Correlation between placenames and population

The correlation coefficient *r* measures the strength and direction of a linear relationship between two variables, I have used this method to measure if there is a relationship between population and placename and the results will be interpreted using the scale below:

	Positive	Negative
No Correlation	0 to 0.1	0 to -0.1
Weak Correlation	0.1 to 0.3	-0.1 to -0.3
Medium correlation	0.3 to 0.6	-0.3 to -0.6
Strong correlation	0.6 to 1	-0.6 to -1

Table 5.13 Commonly used scale to interpret correlation coefficient

5.8.1 Cameroon

Region	**Population count**	**Placename count**
Adamawa	1,015,622	1,543
center	3,525,664	3,215
East	801,968	3,056
Far North	3,480,414	3,062
Littoral	2,865,795	1,034
North	2,050,229	2,937
North West	1,804,695	394
South	692,142	1,532
South West	1,384,286	2,028
West	1,785,285	2,089
Total	**19,406,100**	**20,890**
Placenames out of map		21,338-20,890= **448**

Table 5.14 Placename count and population count per regions in Cameroon

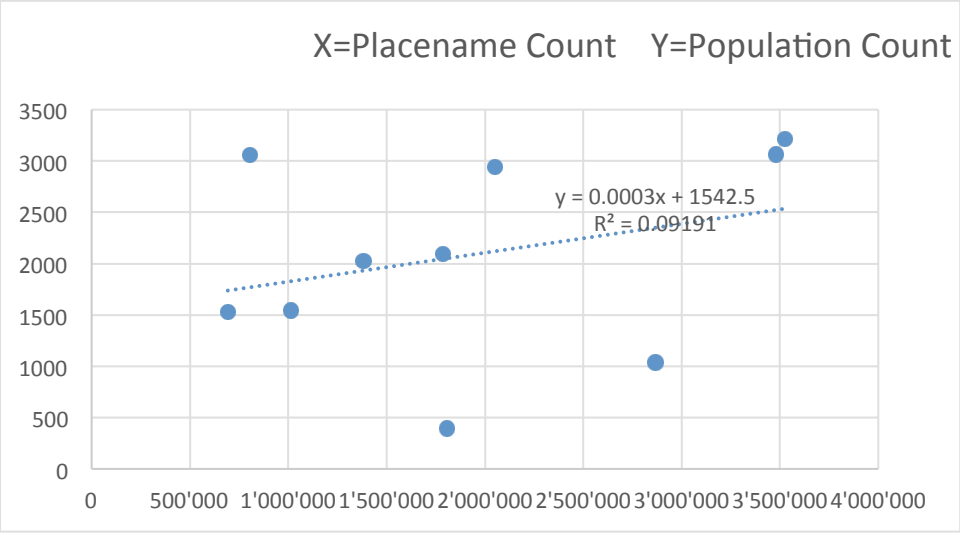


Figure 5.16 Correlation between placename and population count for Cameroon

As seen from the graph above, the correlation between placename count and population count for Cameroon is R^2=0.091 and according to the scale on Table 5.13 there is no correlation.

5.8.2 Switzerland

Canton	**Population Count**	**Placename Count**
Aargau	636,362	733
Appenzell Innerrhoden	53,691	76
Appenzell Ausserrhoden	15,778	72
Basel-Landschaft	278,656	238
Basel-Stadt	189,335	1140
Bern	1,001,281	2611
Freiburg	297,622	689
Geneva	469,433	312
Glarus	39,593	278
Graubünden	194,959	2844
Jura	71,738	264
Luzern	390,349	491
Neuchâtel	176,402	241
Nidwalden	41,888	139
Obwalden	36,507	223
Sankt Gallen	491,699	945
Schaffhausen	78,783	132
Schwyz	151,396	381
Solothurn	261,437	470
Thurgau	260,278	452
Ticino	346,539	1383
Uri	35,865	2578
Vaud	749,373	1254
Valais	327,011	2339
Zug	118,118	104
Zurich	1,425,538	5909
Total	**8,139,631**	**23,298**
Placenames out of map		23,599-23298=**301**

Table 5.15 Placename count and population count per regions in Switzerland

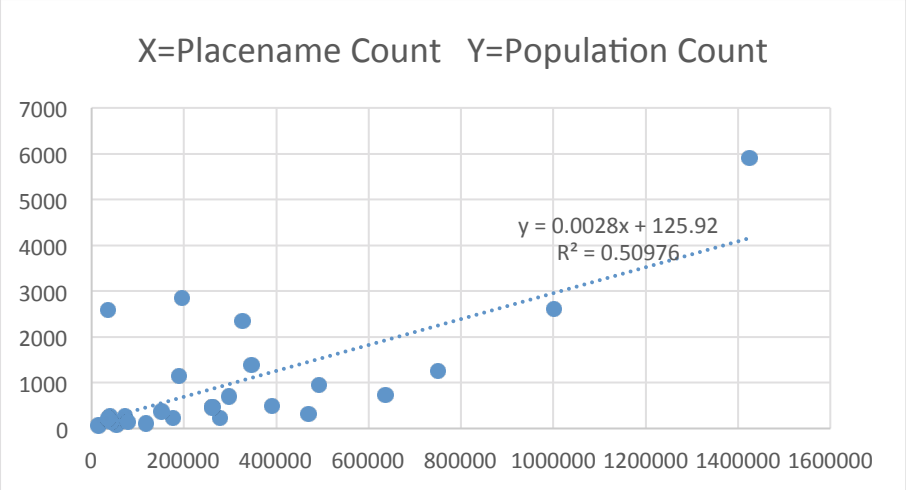


Figure 5.17 Correlation between placename and population count for Switzerland

As seen from the graph above, the correlation between placename count and population count for Switzerland is R^2=0.5098 and according to the scale on Table 5.13 there is a medium positive correlation

5.8.3 Sri Lanka

Province	**Population Count**	**Placename Count**
Basnahira(Western)	5,821,710	4016
Dakunu(Southern)	2,464,732	4638
Madhyama(Central)	2,558,716	6221
Negenahira(Eastern)	1,551,381	2752
Sabaragamuwa	1,918,880	4783
Uturu(Northern)	1,058,762	2976
Uturumeda(North Central)	1,259,567	5208
Uva	1,259,900	4209
Wayamba(Northwestern)	2,370,075	10635
Total	**20,263,723**	**45,438**
Placenames out of map		47,146-45,438=**1708**

Table 5.16 Placename count and population count per regions in Sri Lanka

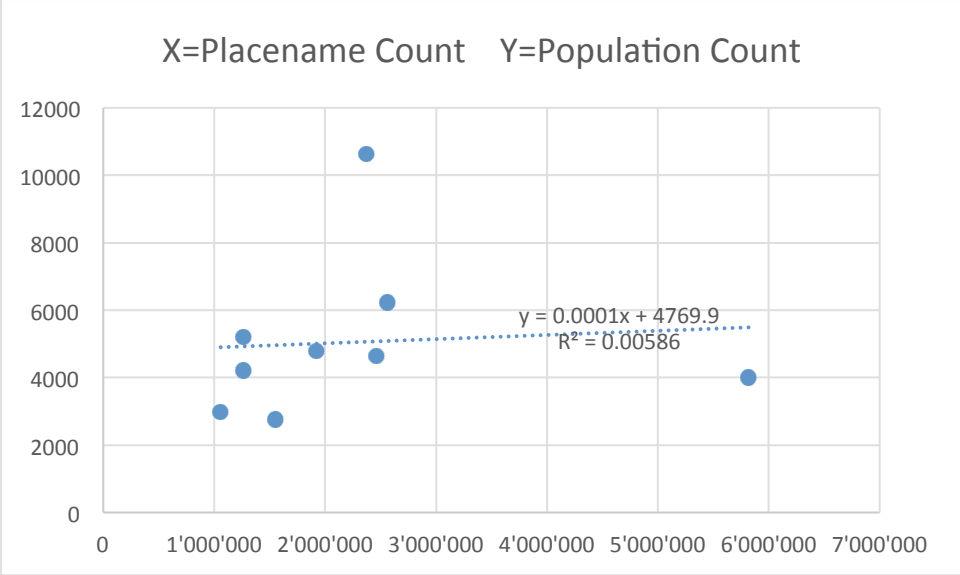


Figure 5.18 Correlation between placename and population count for Sri Lanka

As seen from the graph above, the correlation between placename count and population count for Sri Lanka is R^2=0.005 and according to the scale on Table 5.13 there is no correlation

6. Discussion

This chapter discusses the results described in the previous part of the thesis. First, Placenames are collected from newspapers daily for the month of April. The GeoNames gazetteer data for the various case studies (Cameroon, Sri Lanka and Switzerland) were downloaded, comparisons were made using both excel and maps. In the second section, the research questions are answered and the possible implications of the results are discussed.

Findings 6.1

Discussion of Hypotheses 6.1.1

Problems encountered 6.

6.1 Findings

After all the data collection, results and analysis, the hypotheses stated will be discussed and stated if they are being rejected or accepted.

6.1.1 Discussion of Hypothesis

Hypothesis One: The GeoNames gazetteer data is representative and contains all placenames present on the country landscape

From the GeoNames gazetteer database, the coordinates of the placenames were used to create a dot density maps as seen on Figure 5.4, 5.5 and 5.6. From these dot density maps we saw the spatial distribution of placenames in space for the three countries: Cameroon, Switzerland and Sri Lanka. I wanted to know if the placenames are evenly distributed spatially throughout the landscape, if all the areas are covered with placenames or do we have empty spaces on the dot density maps from the GeoNames data.

Figure 5.4 Shows a dot density map of Switzerland GeoNames points, we see empty spaces on the map indicating that there are no placenames registered for those places in the GeoNames database, there is a cluster of placename points in Zurich and for the rest of the country we see empty spaces with no placenames all over the map, the map illustrates a total of 23,600 placename points with 301 points falling out of the Switzerland boundary as seen in Table 5.14. From Table 5.12, we see a total of 0.571km-^{2} placenames on space unlike with

156,755 placename points in SwissNames we will have a total of 3.79km-^{2} (156,755/41,285km^{2}) placenames on space which is more representative. We also see from Table 5.14 that there are more placenames in some cantons than others and this placename count does not depend on the size of the canton.

Figure 5.5 shows a dot density map of Cameroon GeoNames points; we see empty spaces of no placename points on the maps indicating that there are no placenames registered in the database for these regions, there is a cluster of placename points on the Center region as well as Far North and East region. The map illustrates a total of 21,338 placename points with 448 points falling out of the Cameroon boundary as seen in Table 5.13. From table 5.12 we see a total of 0.044km-^{2} placenames on space for Cameroon. From table 5.13 we see that there are more placenames in some regions than others.

From Figure 5.6 we see a dot density map of Sri Lanka GeoNames points; we see clusters of placenames Northwestern, Central and North central provinces and large empty spaces in the Northern and Eastern provinces. The map illustrates a total of 47,166 placename points with 1708 falling out of the Sri Lankan boundary as seen on Table 5.15. From table 5.12 we see a total of 0.718km^{2} placenames on space for Sri Lanka. From Table 5.15, we see that there are more placenames in some regions than others.

Summarily, to answer the hypothesis of whether the GeoNames gazetteer data is spatially even and if the placenames points spread out across the map, from my analyses and results, I can say that the GeoNames gazetteer data is not spatially representative: the maps have shown that though the points are distributed across the landscape, there are large areas with empty spaces and placenames which have not been recorded into the database and there are a lot of points which are recorded under the countries but fall out of the country borders: it is confusing if the points represent the countries or the neighboring countries. Hypothesis one is therefore rejected which means the GeoNames gazetteer is not representative

Hypothesis Two: All existing placenames for the various Countries are registered on the GeoNames Database.

From the analyses and results of my first research question, I came to the conclusion that there are empty spaces on the dot density maps which is an evidence that not all areas on the country had placenames, my second research question was to find out if the empty spaces

represent no placenames or unnamed places or if they represent placenames that are not found in the gazetteer but exist in the said countries.

The GeoNames gazetteer which hypothetically represents placenames from around the world is expected to be complete, reliable and up to date. Newspapers are data banks and convey large amounts of information from around the world and I chose local newspapers from the various countries because most of the time, the placenames on the local newspapers are from the regions in which the newspaper is produced (Brunner 2008), so I used placenames present in daily newspapers to verify if they are on the GeoNames database. Taking into consideration country atlases, I also used the names found on SwissNames to verify if they were on the Switzerland GeoNames data.

Table 5.1 shows results from the comparison of NZZ and Switzerland GeoNames and from the table we see that out of 107 placenames collected from the NZZ newspaper for a period of 30 days, 75 were not found on the gazetteer and only 32 names were found which accounts for only 29.9% of the total names on the newspaper.

Table 5.2 shows results from the comparison of Cameroon Tribune and Cameroon GeoNames and from the table we see that out of 226 placenames collected from the Cameroon Tribune newspaper for a period of 30 days, 135 were not found on the gazetteer and only 91 names were found which accounts for only 40.7% of the total names on the newspaper.

Table 5.3 shows results from the comparison of Sri Lanka Daily news online and Sri Lanka GeoNames and from the table we see that out of 247 placenames collected from the Sri Lanka newspaper for a period of 30 days, 138 were not found on the gazetteer and only 109 names were found which accounts for only 44.1 % of the total names on the newspaper.

From Table 5.4, we see that while SwissNames has up to 156, 755 placenames GeoNames data has only 23, 599 names for Switzerland which accounts for only 15% of SwissNames data.

From all the analyses above, it is evident that the empty spaces of Figure 5.4(Switzerland), Figure 5.5(Cameroon) and Figure 5.6(Sri Lanka) indicate unregistered placenames and we can further say that there are thousands of placenames names existing in various countries which are not yet registered in the GeoNames gazetteer database. Hypothesis two is therefore rejected which means not all placenames existing in countries are on the GeoNames database.

Hypothesis Three: *There is a consistent and unbiased input of the placenames into the gazetteer (its construction).*

I wanted to know the pattern of placenames input into the gazetteer? Is consistent or if there are large gaps of input data time? If the input is tied to aftereffects of certain events? If there is a large existence of placenames in areas that have been faced with environmental disasters or political instability while peaceful countries have little or no records of placenames?

In order to answer this research question I decided to do an analysis on the input of placename data into the gazetteer. Table 3.1 represents the types of information extracted from the GeoNames data and the input data is one of them. So I sorted the data for each of the countries as they appeared in Table 4.5(Haiti 1993-2014) Table 4.6(Somalia 1994-2014) and Table 4.7 (Sri Lanka 1994-2014), after sorting and calculations I emerged with three cumulative graphs.

Figure 5.1 shows placename input for Haiti and we see that there was a very high input of placenames in 2012 of up to 15, 249 placenames unlike in the other years and we can link this to the after effect of the Haiti earthquake in 2010 which led to the emergence of new placenames data from tweets and other sources.

Figure 5.2 shows placename input for Sri Lanka which experienced a very high input of placenames in 2012 of up to 17,388 placenames unlike in the previous years and we can link this to the end of the civil wars which came to an end in 2012 after 26 years of fighting.

Figure 5.2 shows placename input for Somalia which experienced a very high input of placenames in 2012 of up to 41, 285 placenames unlike the previous years and this is due to the political reforms that came into the country in 2012, Sri Lanka still faces wars till date.

Summarily, I can say from my results that there is a bias in the construction of the GeoNames gazetteer data, the input of placenames is not consistent, we have yearly inputs of placenames no doubt but from the analyses above it can be concluded that the input is higher after certain periods and this is because after these wars or natural disasters, victims, rescue teams and the world at large have more information concerning these countries. It is very complex because,

if we have information on these placenames before the wars or natural disasters, it will be easier for the rescue teams and the victims to save lives but the analyses shows that the placenames come into light only after the disasters. Hypothesis three is therefore rejected which means there is a bias in the input or construction of the GeoNames gazetteer

Hypothesis Four: Placenames are related to people: they both have a strong correlation.

I wanted to find out if there is any relationship or correlation between placenames and population since places are being named by and after people and are occupied by people; if there more people where there are more placenames and if areas with dense population have clusters of placenames as well.

A KDE was made from the dot density maps to get the clustering and hotspots of the placename points as seen if Figure 5.7 for Switzerland, Figure 5.8 for Cameroon and Figure 5.9 for Sri Lanka. Population density maps were also created for the various countries as seen in Figure 5.10 for Switzerland, Figure 5.11 for Cameroon and Figure 5.12 for Sri Lanka.

Comparisons were made between the population density maps and KDE: Figure 5.13 shows a comparison between Switzerland population density and KDE and we saw that there was little or no relationship between the two maps, no regions had the same hotspots though we saw that the most populated canton which is Zurich had the highest placename points but in terms of population density and placename density they were not same.

Figure 5.15 shows comparison between Cameroon population density and KDE and we saw a link between the two maps, we saw that on the two maps we had same shades of hotspots on the West and Littoral regions and Far North regions as well which are the dense clustered regions, also we saw lighter shades of clusters on both maps for South and Adamawa regions which are the least populated and least named regions.

Figure 5.13 shows a comparison between Sri Lanka population density and KDE and from the two maps we see that there is no relationship between the most populated province which is the western province and the placename clusters, we see a relationship between the next highly dense populated regions with the placename map which are the Northwestern regions , Central, Sabaragamuwa and the Southern region , we also see that the least populated region

which is the Northern region has the least placenames too, so we can say that there is a relationship between placename hotspot and population density for Sri Lanka and Cameroon unlike very less relationship for Switzerland.

After comparing the maps I decided to do a more detail comparison using population count and placename count per region as seen on Table 5.13 for Cameroon, Table 5.14 for Switzerland and Table 5.15 for Sri Lanka and then the correlation graphs and for Cameroon and Sri Lanka show no correlation while Switzerland shows a medium correlation. Hypothesis four is therefore rejected which means there is no correlation between placename and population.

6.2 Problems

It was difficult for me to collect news from Sri Lanka daily news online because most of the time the news for a certain day was mixed with that from the previous day and this was a bit confusing, so I had to read each of the stories and memorise them and made sure it wasn't the same story and also this was strenuous because the placenames kept repeating with the repeated storyline and frequency of placename was not my focus but placename count**.**

Some placenames on the newspapers were in English (Cameroon Tribune and Sri Lanka daily news) and some in French (Cameroon Tribune), in German (NZZ) and in Tamil (Sri Lanka daily news) and this made it difficult to identify the placenames from the news script because I could not easily identify what word was a placename or another word in the dialect.

The population density data from GeoHive had census results for different years for each country and it was difficult for me to choose which year to use because I needed the population count for the same year for each country, but what I finally decided to do was to use the most recent population census count which was 2013 for Switzerland, 2012 for Sri Lanka and 2010 for Cameroon.

While comparing the placenames form the newspapers and those form the gazetteer it was difficult for me to actually determine which placenames were not on the gazetteer because of accents, hyphens and some places had more than one variant name which was in different languages and this made it a bit confusing especially when comparing the placenames from the newspapers with those from the gazetteer.

It was difficult to sum the number of placename input per year using excel because some placenames did not have any date of input as such, the input data was calculated per input data and consideration was not given if it summed to the total number of placenames for that country because it was obviously less.

After haven downloaded data from the GeoNames database it was difficult to detect which row was the X coordinate and which was the Y coordinate so I did a guess and joint the coordinates to the country shape file and the map was in the opposite direction then I knew I had put the coordinates opposite so I had to change then and redo the joint to get the real dot density map.

After haven made the dot density maps using WGS_1984, I used the dot maps to create KDE and the KDE seemed wrong for all the countries they were not visually appealing despite trying different cell size outputs and bandwidths, so I realized that the KDE was wrong because I used the WGS_1984 projected coordinate system for the various countries and when I changed the dot maps to the various country projected coordinate systems the KDE emerged right and was correct.

To create the population density maps, I had to join the data on the excel sheet to the said country shape file, after which I went to properties to the symbology section and chose the number of classes I wanted the data to divided into and also the color ramps and in doing so the population density maps for Cameroon and Sri Lanka emerged with some polygons missing and when I went to the attribute table I realized that there were two problems: the first problem was that I had not numbered the FID for the various regions and provinces correctly and the second problem was that the regions in Cameroon were labeled in French and those in my data were labeled in English this was same with Sri Lanka population density data were the names on the shape file were in Tamil and mine were in English so I had to verify the Tamil version of each of the Names on my data and match with those on the country shape file attribute tables to get the correct population density maps with all the polygons present.

In order to get the placename count per region, I had to do a spatial join were each polygon will have a summary of the exact placename points that fall inside and in the end I had an

attribute table with the number of placenames for each region but for all the three countries I had an extra column with no name which also had placename count and I wondered from which polygons the placename counts were from and I had to redo the summary of the attribute table several times but then I realized that these placename count didn’t fall in any of the polygons or regions they were the placenames that featured out of the country maps as seen on Figure 5.4, 5.5 and 5.6

7. Conclusion

7.1 Achievements

The following points sum up to what is the output of my research:

- List (count) and percent of placenames not in the GeoNames gazetteer for Cameroon, Switzerland and Sri Lanka.
- Count and percentage of placenames found in SwissNames not found in the GeoNames gazetteer.
- Placename distribution per km-² for Cameroon, Sri Lanka and Switzerland
- Graphs illustrating cumulative inputs of placenames into the GeoNames gazetteer for Sri Lanka, Somalia and Haiti
- One-to-one dot maps illustrating placename distribution on the landscapes of Cameroon, Sri Lanka and Switzerland.
- Kernel Density Estimation maps for placename distribution for Switzerland, Sri Lanka and Cameroon
- Population density choropleth maps illustrating population density for Switzerland (2013), Cameroon(2010), Sri Lanka(2012)
- Placename count per region (Cameroon), per canton (Switzerland) and province (Sri Lanka) as of April 2014 when the data was downloaded.
- Correlation results from regression analyses between placename count and population count per region for Cameroon, Sri Lanka and Switzerland

7.2 Insights

Placename input is based on certain situations that pull the attention of crises mappers and other people who upload information to the Gazetteer, what about the countries that have already been evaluated and recognized as disaster prone areas and still yet have no data about it: it would be easier to carry out field work in these areas to collect information on their place type and footprints so that in case of emergencies it would make rescue services and evacuation easier.

We could use other methods to determine existing places like using streetlights data or using street light tracker, where one can zoom into the map and identify places by streetlights existence, but what of villages in the developing areas that don’t have electricity supply till date

There is no coordination of the gazetteer data input; data that is uploaded is not verified: the names of the places and their coordinates. Points appearing out of the out of the country maps are probably due to uncoordinated input of placename data into the gazetteer

7.3 Future work

Java or any other programming could be used to get a detail results from the comparison between placenames from the newspapers and those from the GeoNames data or any other source and this would give more details from the comparisons because you can program it to compare the list based on your own criteria.

Point pattern analysis could be done using the dot density maps to get the exact pattern of the points on the maps for the chosen countries; if they rare random, clustered or uniform.

The KDE maps could be produced in different bandwidths and resolutions to see different shades of clusters, they could also be used to calculate number of placename per output cell.

Geographically weighted Regression and kernel regression to see which relationship exists at different points on the maps

It is hoped that work and knowledge gained form this work will help to improve the input (construction) and usage of the GeoNames gazetteer data.

BIBLIOGRAPHY

Baeza-Yates, R. A. and B. Ribeiro-Neto (1999). Modern Information Retrieval. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Barfourosh, A. A., Nezhad, H. R., Anderson, M. L., & Perlis, D. (2002). Information retrieval on the World Wide Web and active logic: A survey and problem definition.

Bear, J., D. Israel, J. Petit and D. Martin (1998). Using information extraction to improve document retrieval. In Proceedings of the Sixth Text Retrieval Conference (TREC-6, pp.367 - 377.

Bikel, D., R. Schwartz, and R. Weischedel (1999). An algorithm that learns what's in a name. Machine Learning 34 (1{3), 211-231.

Bikel, D., S. Miller, R. Schwartz, and R. Weischedel (1997). Nymble: a high-performance learning name finder. In Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP), Washington, D.C., pp. 194-201.

Borthwick, A. E. (1999). A maximum entropy approach to named entity recognition. Ph.D. thesis, New York University, New York, NY, USA. AAI9945252.

Brants, T. and Google Inc. (2004). Natural language processing in information retrieval. In Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands, pp.1-13.

Brunner, T. (2008). Geographic Information Retrieval: Identifikation der geographischen Lage von Zeitungsartikeln. Master's thesis, Universitat Zurich.

Burenhult, N. and S. C. Levinson (2008). Language and landscape: a cross-linguistic perspective. Language Sciences 30 (2/3), 135-150.

Buscaldi, D. and P. Rosso (2008). Map-based vs. knowledge-based toponym disambiguation.

In Proceeding of the 2nd international workshop on Geographic information retrieval,

GIR '08, New York, NY, USA, pp. 19-22. ACM.

Cheshire, J. A., & Longley, P. A. (2012). Identifying spatial concentrations of surnames. *International Journal of Geographical Information Science*, *26*(2), 309-325.

Chieu, H. L. and H. T. Ng (2003). Named entity recognition with a maximum entropy approach. *In Proceedings of the seventh conference on Natural language learning at NAACL*

2003 - Volume 4, CONLL '03, Stroudsburg, PA, USA, pp. 160{163. Association for Computational Linguistics.

Davies, C., I. Holt, J. Green, and L. Diamond (2009). User needs and implications for modelling vague named places. Spatial Cognition & Computation: An Interdisciplinary Journal 9 (3), 174-194.

E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In Proc. SIGIR, p 273–280. ACM, 2004.

F. Bilhaut, T. Charnois, P. Enjalbert, and Y. Mathet.Geographic reference analysis for geographic document querying. In *Workshop on the Analysis of Geographic References*, Edmonton, Alberta, Canada, May 2003. NAACL-HLT.

Frisch, W., Meschede, M., & Blakey, R. C. (2010). *Plate Tectonics: Continental Drift and Mountain Building*. Springer.

Gan, Q., J. Attenberg, A. Markowetz, and T. Suel (2008). Analysis of geographic queries in a search engine log. In Proceedings of the first international workshop on Location and the web, LOCWEB '08, New York, NY, USA, pp. 49{56. ACM.

Gelling, M., & Cole, A. (2000). *The landscape of place-names*. Shaun Tyas.

Gelling, Margaret (1993) Placenames in the landscape

Getty Information Institute. (1997). Thesaurus of Geographic Names. http://www.ahip.getty.edu/tgn_browser/(Date accessed 03.05.2014)

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. Geo-Journal 69 (4), 211-221.

Goodchild, M. F., & Hill, L. L. (2008). Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, *22*(10), 1039-1044.

Goodchild, M. F. (1999). The future of the gazetteer. Presented at the digital gazetteer information exchange workshop, Oct 13-14. http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/Perspectives/Goodchild.htm. Transcribed and edited from audiotape. (Date accessed: 04.10.2014).

Grishman, R. (1997). Information extraction: Techniques and challenges. In *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology* (pp. 10-27). Springer Berlin Heidelberg.

Harada, Y., & Sadahiro, Y. (2005). A quantitative model of place names as a georeferencing system. In *Proceedings of GeoComputation*.

Hill, L. L., Frew, J., & Zheng, Q. (1999). Geographic names. *D-Lib Magazine*, *5*(1), 17.

Hill, L. L. (2000). Core elements of digital gazetteers: Placenames, categories, and footprints. In *J. Borbinha & T. Baker (Eds.), Research and Advanced Technology for Digital Libraries: Proceedings of the 4th European Conference, ECDL 2000*, pp. 280-290.Springer.

Hill, L. L. (2006). Georeferencing: the geographic associations of information. Digital libraries and electronic publishing. Cambridge, MA: MIT Press.

Hill, L. L., & Zheng, Q. (1999). Indirect geospatial referencing through place names in the digital library: Alexandria Digital Library experience with developing and implementing gazetteers. Proceedings of the American Society for Information Science Annual Meeting, Washington, D.C.,Oct. 31- Nov. 4, 1999, pp. 57-69.

Hollenstein, L. and R. S. Purves (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. Journal of Spatial Information Science 1, 21-48.

Hollenstein, L. (2008). Capturing vernacular geography from georeferenced tags. Master's thesis, University of Zurich, Institute of Geography.

J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th VLDB Conference*, Cairo, Egypt, 2000.

J. T. Hastings (2008) Automated conflation of digital gazetteer data, International Journal of Geographical Science, 22:10, 1109-1127.

Jones, C. B., H. Alani, and D. Tudhope (2001). Geographical information retrieval with ontologies of place. In Spatial Information Theory, LNCS 2205, pp. 322{335. Springer.

Jones, C. B., Abdelmoty, A. I., & Fu, G. (2003). Maintaining ontologies for geographical information retrieval on the web. In *On the Move to Meaningful Internet Systems 2003: Coop IS, DOA, and ODBASE* (pp. 934-951). Springer Berlin Heidelberg.

Jones, C. B. and R. S. Purves (2008). Geographical information retrieval. International Journal of Geographical Information Science 22, 219-228.

Jones, C. B., Purves, R. S., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, *22*(10), 1045-1065.

Kimerling,A.J (2009) Cartography and Geographic Information Science, 36(2) 165-182.

Kessler, C., Janowicz, K., & Bishr, M. (2009, November). An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems* (pp. 91-100). AC

Kozareva, Z. (2006). Bootstrapping named entity recognition with automatically generated gazetteer lists. In Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, EACL '06, Stroudsburg, PA, USA, pp. 15-21. Association for Computational Linguistics.

Kunz, R. (2008). Evaluation of spatial relevance in geographic information retrieval. Master's thesis, University of Zurich.

Larson, R. R. (1996). Geographic information retrieval and spatial browsing.

Li, L. T., Pedronette, D. C. G., Almeida, J., Penatti, O. A., Calumby, R. T., & da S Torres, R. (2012, November). Multimedia multimodal geocoding. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems* (pp. 474-477). ACM.

Leidner, J. L. (2007). Toponym Resolution in Text. Ph. D. thesis, University of Edinburgh.

Leidner, J. L., G. Sinclair, and B. Webber (2003). Grounding spatial named entities for information extraction and question answering. In Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1, HLT-NAACL-GEOREF '03, Stroudsburg, PA, USA, pp. 31-38. Association for Computational Linguistics

Lieberman, M., H. Samet, and J. Sankaranarayanan (2010). Geotagging with local lexicons to build indexes for textually-specied spatial data. In 2010 IEEE 26th International Conference on Data Engineering (ICDE), pp. 201-212.

Liu, X., S. Zhang, F. Wei, and M. Zhou (2011). Recognizing named entities in tweets. To appear in ACL 2011.

Manning, C. D., P. Raghavan, and H. Schutze (2008). Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press.

Mikheev, A., M. Moens, and C. Grover (1999). Named entity recognition without gazetteers. In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, EACL '99, Stroudsburg, PA, USA, pp. 1-8. Association for Computational Linguistics

Montello, D. R., M. F. Goodchild, J. Gottsegen, and P. Fohl (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. Spatial Cognition & Computation 3 (2-3), 185-204.

MUC-6 Appendix (1995). Appendix C: Named entity task definition (v2.1). In MUC6 '95: Proceedings of the 6th conference on message understanding, Stroudsburg, PA, USA, pp. 317{332. Association for Computational Linguistics.

Overell, S. E. and S. Ruger (2006). Indentifying and grounding descriptions of places. In SIGIR Workshop on Geographic Information Retrieval, pp. 14{16.

Piotrowski, M. (2010, February). Leveraging back-of-the-book indices to enable spatial browsing of a historical document collection. In *Proceedings of the 6th Workshop on Geographic Information Retrieval* (p. 17). ACM.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Sarawagi, S. (2008). Information extraction. Found. Trends databases 1, 261-377.

Singhal, A. (2001). Modern information retrieval: a brief overview. Bulletin of the IEEE computer society technical committee on data engineering 24, 2001.

Stephens, Mitchell, NYU.edu, "History of Newspapers", Collier's Encyclopedia.

Stevenson, M. and R. Gaizauskas (2000). Using corpus-derived name lists for named entity recognition. In Proceedings of the sixth conference on Applied natural language processing, ANLC '00, Stroudsburg, PA, USA, pp. 290{295. Association for Computational Linguistics.

Swiss Federal Office of Topography (2009). Geographic names http://www.swisstopo.admin.ch/internet/swisstopo/en/home/topics/toponymie.html.(Date accessed: 20.11.2014).

U.S. Federal Geographic Data Committee. (1998). *Content Standard for Digital Geospatial Metadata*. Available: http://fgdc.er.usgs.gov/metadata/contstan.html.

Vestavik, Ø. (2004). Geographic information retrieval: an overview. *Department of Computer and Information Science, Norwegian University of Technology and Science*. Valli,

C., & Hannay, P. (2010). Geotagging Where Cyberspace Comes to Your Place. In *Security and Management* (pp. 627-632).

Yang, C.C.; Chen, Hsinchun; Honga, Kay (2003). "Visualization of large category map for Internet browsing". *Decision Support Systems*35 (1): 89–102

Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, *2014*(9), 37-70.

Zubizarreta, A., P. de la Fuente, J. Cantera, M. Arias, J. Cabrero, G. Garcia, C. Llamas,and J. Vegas (2009). Extracting geographic context from the Web: Georeferencing in MyMoSe. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy (Eds.), Advances in Information Retrieval, Volume 5478 of Lecture Notes in Computer Science,pp. 554-561. Springer Berlin / Heidelberg.

Personal Declaration:

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis

Yamgouet Monique Ndam