

# Localizing the Internet – Linking Web N-Grams to Geographic Space

Master thesis GEO 511

29.09.2015

Jérôme Sautier

Student ID number: 10-735-736

Supervised by

Dr. Curdin Derungs

Prof. Dr. Ross Purves (Faculty Member)

Geocomputation

Department of Geography

University of Zurich



## **Contact**

### **Author**

Jérôme Sautier

Junkerriet 11

FL-9496 Balzers, Liechtenstein

jerome.sautier@outlook.com

### **Supervisor**

Dr. Curdin Derungs | Prof. Dr. Ross Purves

Geographic Information Science (GIS) | Geocomputation

Department of Geography

University of Zurich - Irchel

Winterthurerstrasse 190

CH-8057 Zurich, Switzerland

curdin.derungs@geo.uzh.ch | ross.purves@geo.uzh.ch

*“Züri isch eus!”*

Rugby Union Zurich

## Preface

In the beginning of my studies I was clueless. I had no idea how multifaceted geography was. Especially, the combination of geography and computer science caught my attention and motivated me to focus on geographic information science (GIScience). Step by step this led to discovering new disciplines and culminated to the present master thesis.

It was a rough ride writing this thesis. Nevertheless, diving into a new field and learning new things encouraged me to choose Web n-grams as topic. The simple fact of obtaining the frequency of a word and word arrangements on the Web fascinated me.

Honestly, I never thought I would make it this far. Let alone, write an over 100 page master thesis. I might just be a little less clueless than at the beginning of my studies. Therefore, I would like to thank all the people who made this possible:

- A huge thank you to my parents, who supported my decision to study geography and financed my education.
- Dr. Curdin Derungs for the assistance, the long discussions, patience, advice and feedbacks.
- Prof. Dr. Ross Purves for giving me advice and assisting me in the time Dr. Curdin Derungs was gone.
- Ricky Loynd from Microsoft Research for giving me a user token and thus allowing me to use the Microsoft Web N-Gram Service.
- My rugby teams in Zurich and Liechtenstein, which are like a second family to me. The promotion of the Rugby Union Zurich to League B in the Swiss National Rugby League was one of my highlights. Moreover, I will always cherish the memories with the Liechtenstein Rugby Union. The participation in the Rugby Europe 7s Men Division C in Bosnia and Herzegovina followed by the rugby trip to South Africa was an extraordinary experience. I had a great time with both teams this year. You pushed me physically and mentally to the next level.
- My friends, who put a beer in my hand when I desperately needed one. And always thought I could only become a teacher by studying geography.



## Abstract

N-grams are combinations of  $n$  words and are prominently used to structure and index large natural language data. The recent release of  $n$ -gram collections from Google and Microsoft make  $n$ -grams accessible to a broader audience. These collections return the frequency/probability of arbitrary  $n$ -word combinations (i.e.  $n$ -grams, where  $n=1:5$ ) in books or on the Web. In that sense, they allow easy access to a large amount of data on the Web which cannot be obtained through simple Web searches. Such  $n$ -gram collections are rarely used in geography, even though they could provide considerable information for geographic analysis.

The aim of this master thesis is the exploration and analysis of place names at different granularities in the Microsoft Web  $n$ -gram collection. This is accomplished by using the Microsoft Web N-Gram Service API and methods usually situated in the field of information retrieval (IR), geographic information retrieval (GIR) and data mining.

The occurrence of place names (also known as toponyms) is investigated by initiating queries in form of lists to the Microsoft Web N-Gram Service API. The initiated lists have different levels of granularity: continent, country, capital city and city. For each entry in the list the joint probability is returned, which expresses the likelihood of  $n$ -words occurring in the Microsoft Web  $n$ -Gram collection. The resulting joint probabilities are statistically examined in terms of distribution, minimum value, maximum value, mean, standard deviation, range and number of words constituting a place name. A possible cause for different joint probabilities is inspected with a correlation between place name joint probability and their corresponding number of inhabitants. Additionally, the representation of triplets in the form of <topic><spatial relationship><place name> is investigated with the help of the autocompletes from the Microsoft Web N-Gram Service. It is explored if the first 1000 autocompletes can resolve basic topological relations such as country in continent, country bordering country and city in country. These are further investigated on accuracy and coverage. The linking of geographic features/sports activities and place names is also done with triplets and autocompletes, while the plausibility of selected results is evaluated by simple Web searches on Bing. Finally, the strongest associations of geographic features/sports activities and countries are mapped for Europe.

Results indicated that the place name ambiguity highly increased with finer granularities. Moreover, a positive correlation between place name joint probabilities and population was statically proven at country and capital city level. The topological relations were better resolved in terms of accuracy and coverage for coarser granularities. The results of using Web

n-grams to link geographic features/sport activities to space mostly returned accurate representations and helped to gain new insights into the world. Overall, the findings were promising and Web n-grams could have future implementations in GIR for resolving ambiguity, interpretation of vague place names or help to quantify vague spatial language. Nevertheless, the impact of toponym ambiguity in Web n-grams is and stays a problem and should be addressed in future work.

## Zusammenfassung

N-Gramme sind aufeinanderfolgende Fragmente von n Wörtern und werden häufig verwendet um grosse Textdaten zu strukturieren und indexieren. Die kürzlich Veröffentlichung der N-Gramm Korpora von Google und Microsoft ermöglichen einem breiterem Publikum den Zugang zu N-Grammen. Diese Korpora geben die Häufigkeit/Wahrscheinlichkeit von beliebigen N-Worten (d.h. N-Gramme, wo  $n=1:5$ ) in Büchern oder dem Web zurück. In diesem Sinne erlauben N-Gramm Korpora leichten Zugang zu Unmengen an Daten auf dem Web, die durch einfache Web-Suchen nicht erhalten werden können. Solche N-Gramm Korpora werden nur selten in der Geographie verwendet, obwohl sie bedeutende Beiträge für geographische Analysen anbieten könnten.

Das Ziel dieser Masterarbeit ist die Exploration und Analyse von Ortsnamen in verschiedenen Detailebenen im Microsoft Web N-Gram Korpus. Dies wird erreicht, indem die Programmierschnittstelle vom „Microsoft Web N-Gram Service“ und Methoden im Gebiet des „Information Retrieval“ (IR), „Geographic Information Retrieval“ (GIR) und Data-Mining genutzt werden.

Das Auftreten von Ortsnamen (auch bekannt als Toponyme) wird untersucht, indem Suchanfragen – in der Form von Listen – an die Programmierschnittstelle vom „Microsoft Web N-Gram Service“ initiiert werden. Die beauftragten Listen bestehen aus verschiedenen Detailebenen: Kontinent, Staat, Hauptstadt und Stadt. Für jeden Eintrag in der Liste wird die Wahrscheinlichkeit berechnet. Die resultierenden Wahrscheinlichkeiten werden statistisch ausgewertet in Bezug auf Verteilung, Minimum, Maximum, Mittelwert, Standardabweichung, Spannweite und die Anzahl Worte eines Ortsnamens. Mögliche Gründe für die Entstehung von Wahrscheinlichkeiten werden untersucht mit einer Korrelation. Es wird angenommen, dass eine Relation zwischen Einwohneranzahl und Ortsnamenwahrscheinlichkeit besteht. Zusätzlich, werden Drillinge in der Form <Thema><räumliche Beziehung><Ortsname> in den Autovervollständigungen des „Microsoft Web N-Gram Service“ erforscht. Es werden einfache topologisch Beziehung auf ihre Genauigkeit und Vollständigkeit in den ersten 1000 Autovervollständigungen überprüft. Diese topologischen Beziehung sind unter anderem: Staat in Kontinent, Staat angrenzenden an Staat und Stadt in Staat. Ebenso, werden geographische Objekte/Sport Aktivitäten durch Drillinge und Autovervollständigungen mit Ortsnamen verknüpft. Die Plausibilität der Resultate wird vereinzelt überprüft durch einfache Suchanfragen in Bing. Schlussendlich, werden die stärksten Verknüpfungen zwischen geographischen Objekten/Sport Aktivitäten und europäischen Staaten kartiert.

Die Resultate verdeutlichen, dass die Mehrdeutigkeit in Ortsnamen mit zunehmendem Detail steigt. Ausserdem, wurde eine positive Korrelation zwischen Ortsnamenwahrscheinlichkeit und Population statistisch erwiesen. Die topologischen Beziehungen waren vollständiger und hatten höhere Genauigkeiten in abnehmenden Detailebenen. Die Verknüpfungen von geographischen Objekten/Sport Aktivitäten und Ortsnamen, mithilfe von Web N-Grammen, ergaben hauptsächlich getreue Darstellung der Welt und halfen neue Erkenntnisse über die Welt zu gewinnen. Insgesamt, waren die Web N-Gramm Ergebnisse vielversprechend und könnten in Zukunft mögliche Implementierung in GIR haben. Das Problem von Mehrdeutigkeit in Web N-Grammen bleibt bestehen und sollte in künftigen Arbeiten thematisiert werden.

## Résumé

N-grammes sont des combinaisons de n-mots et sont utilisées fréquemment pour structurer et indexer des grandes données de texte. L'introduction récente de corpus de Google et Microsoft facilitent l'accès des n-grammes à un public plus large. Ces corpus donnent la fréquence/probabilité de n-mots arbitraires (où n peut prendre la valeur de 1 à 5) dans des livres ou sur le Web. Dans ce sens, ils permettent un accès facile à une grande quantité de données sur le Web qui ne peut pas être obtenu par de simples recherches sur le Web. Ainsi corpus de n-gramme sont rarement utiliser pour la géographie, alors même qu'il pourrait fournir des informations considérables pour l'analyse géographique.

L'objectif de cette thèse de master est l'exploration et l'analyse des noms de lieux, à différents niveaux de détail dans la collection de n-gramme de Microsoft. Ceci est accompli en utilisant l'interface de programmation du « Microsoft Web N-Gram Service » et des méthodes habituellement situées dans le domaine de recherche d'information (RI), de récupération d'information géographique (« geographical information retrieval ») et d'exploration de données.

L'occurrence des noms de lieux (connu aussi sous le nom de toponyme) est analysée en initiant des « queries » sur de listes successives dans l'interface de programmation du « Microsoft Web N-Gram Service ». Les listes initiées ont différents niveaux de détail : continent, pays, capitale et ville. Pour chaque élément des listes successives, la probabilité conjointe est calculé, pour exprimer la probabilité de n-mots apparaissant dans la collection de Web n-gramme de Microsoft. Les résultats des probabilités conjointes sont examinés statistiquement en termes de distribution, valeur minimale, valeur maximale, moyenne, écart-type, amplitude et nombre de mots constituant un toponyme. Une cause possible pour des probabilités différentes peut généralement être due à la corrélation entre le nombre d'habitants et la probabilité du toponyme lui-même. En outre, la représentation de triplets de la forme < sujet > < relation spatiale > < toponyme > est analysée examinées à l'aide d'autocomplétions fournies par le « Microsoft Web N-Gram Service ». Une recherche sur les 1000 premières autocomplétions peuvent aider à résoudre des relations topologiques comme pays dans un continent, le pays et ses voisins et la ville dans le pays. Pour chacune, une analyse précise de la réalité est effectuée. Le lien des objets géographiques où activités sportives et les toponymes se fait aussi avec des triplets et des autocomplétions. La plausibilité des résultats est évaluée avec des recherches Web sur Bing. Finalement, les associations les plus forts des objets géographique où activités sportives et les pays son cartographiés pour l'Europe.

Les conclusions montrent que l'ambiguïté du toponyme augmenté fortement en fonction des niveaux de détail recherché. En outre, une corrélation entre la probabilité conjointe des toponymes et la population a été prouvée statistiquement au niveau des pays et des capitales. Les relations topologiques sont mieux résolues en termes de précision que la réalité pour des niveaux de détail moindre. La plupart des résultats de l'utilisation de Web n-gramme pour relier des objets géographiques où activités sportives à l'espace a donné des représentations justes. Également, ils ont aidé à acquérir de nouvelles connaissances du monde. Dans l'ensemble, les résultats étaient prometteurs et Web n-grammes pourraient avoir des implémentations en récupération d'information géographique (« geographical information retrieval ») pour résoudre l'ambiguïté, l'interprétation des toponymes vagues ou aider à quantifier le langage spatial. Pourtant, l'impact de l'ambiguïté toponymique dans le Web n-gramme reste à améliorer et analyses complémentaires devraient être entreprises.

# Contents

1	Introduction.....	1
1.1	Context .....	1
1.2	Motivation .....	2
1.3	Objective and Research Questions .....	2
1.4	Outline.....	4
2	Background.....	5
2.1	Information Retrieval .....	5
2.2	Geographical Information Retrieval.....	8
2.3	Gazetteer.....	11
2.4	Toponym Recognition.....	12
2.4.1	Gazetteer Lookup .....	12
2.4.2	Rule Based.....	13
2.4.3	Machine Learning.....	13
2.5	Toponym Ambiguity .....	14
2.5.1	Geo/Non-Geo Ambiguity .....	14
2.5.2	Geo/Geo Ambiguity .....	15
2.6	Toponym Resolution .....	15
2.6.1	Map Based.....	15
2.6.2	Knowledge Based.....	16
2.6.3	Data Driven .....	17
2.7	Data Mining.....	18
2.8	N-Gram.....	19
3	Data.....	21
3.1	Microsoft Web N-Gram Service .....	21
3.2	Lists .....	22
3.2.1	Continents.....	22
3.2.2	Countries .....	22
3.2.3	Capitals Cities.....	23
3.2.4	Cities.....	23
3.2.5	Geographic Features.....	24
3.2.6	Sports Activities .....	25
3.3	Map.....	25
3.4	Software.....	25

4	Methodology .....	27
4.1	Data Pre-Processing .....	29
4.1.1	Programming .....	29
4.1.2	Place Name Lists .....	29
4.2	Queries.....	30
4.2.1	Place Name Lists .....	30
4.2.2	Triplets.....	30
4.3	Analysis .....	32
4.3.1	Place Name Joint Probability .....	32
4.3.2	Additional Information .....	33
4.3.3	Zipf Distribution.....	33
4.3.4	Correlation and Index .....	34
4.4	Evaluation.....	36
4.4.1	Ground Truth .....	36
4.4.2	Correctly Retrieved Spatial Relations .....	38
4.4.3	Relevant Found Spatial Relations .....	39
4.5	Explorative Approach.....	39
4.6	Map Visualization .....	42
5	Results.....	45
5.1	Distribution and Statistics of Place Name Joint Probabilities .....	45
5.1.1	Continent Name Joint Probability .....	45
5.1.2	Country Name Joint Probability .....	48
5.1.3	Capital City Name Joint Probability.....	52
5.1.4	City Name Joint Probability .....	55
5.2	Correlation of Place Name Joint Probabilities .....	59
5.2.1	Country Name Correlation .....	59
5.2.2	Capital City Name Correlation .....	61
5.2.3	City Name Coefficient of Determination .....	62
5.3	Correctly Retrieved and Relevant Found Spatial Relations .....	63
5.3.1	Country in Continent .....	63
5.3.2	Capital City in Continent.....	64
5.3.3	City in Continent .....	65
5.3.4	Country bordering Country .....	65
5.3.5	Capital City in Country .....	67
5.3.6	City in Country .....	68
5.4	Explorative Approach.....	71

5.4.1	Geographic Features .....	71
5.4.1.1	Countries to follow Geographic Feature .....	71
5.4.1.2	Cities to follow Geographic Feature.....	79
5.4.2	Sports Activities .....	85
5.4.2.1	Countries to follow Sport Activity .....	85
5.4.2.2	Cities to follow Sport Activity .....	89
6	Discussion .....	95
6.1	Characteristics of Place Names occurring in Web N-Grams.....	96
6.2	Describing the World with Web N-Grams .....	99
6.3	Evaluation of using Web N-Grams to link information to place names.....	101
7	Conclusion .....	105
7.1	Achievements .....	105
7.2	Findings .....	107
7.3	Outlook.....	109
	References .....	111
	Appendix .....	119
A	Code .....	119
A.1	nGramProbability .....	119
A.2	nGram.....	119
A.3	nGrams .....	120
A.4	retrievedFound.....	120
B	Lists .....	120
B.1	Countries .....	121
B.2	Capital Cities .....	122
B.3	Cities.....	123
B.4	Geographic Features.....	123
B.5	Sports Activities .....	123
C	Additional Results .....	124
C.1	Country Name Correlation per Continent .....	124
C.2	Capital City Name Correlation per Continent .....	126
C.3	Spatial Autocorrelation.....	128
C.4	Geographic Feature in/near Country .....	128
C.5	Geographic Features near City .....	130
C.6	Sport Activity in Country .....	132
C.7	Sport Activity in City .....	134



## Figures

Fig. 4.1: Overview of data usage and methodological steps .....	28
Fig. 5.1: Continent name joint probability on the Microsoft Web N-Gram Service in percentage .....	46
Fig. 5.2: Continent name distribution compared to an adjusted Zipf distribution .....	47
Fig. 5.3: Box plots of continent name joint probability based on the number of words .....	48
Fig. 5.4: Country name joint probability on the Microsoft Web N-Gram Service in percentage .....	49
Fig. 5.5: Spatial distribution of country name joint probabilities on the Microsoft Web N-Gram Service .....	51
Fig. 5.6: Country name distribution compared to an adjusted Zipf distribution .....	51
Fig. 5.7: Box plots of country name joint probability based on the number of words .....	52
Fig. 5.8: Capital city name joint probability on the Microsoft Web N-Gram Service in percentage .....	53
Fig. 5.9: Capital city name distribution compared to an adjusted Zipf distribution .....	54
Fig. 5.10: Box plots of capital city name joint probability based on the number of words .....	55
Fig. 5.11: City name joint probability on the Web in percentage with a cut off at 0.01% .....	56
Fig. 5.12: City name distribution compared to an adjusted Zipf distribution .....	58
Fig. 5.13: Box plots of city name joint probability based on the number of words .....	58
Fig. 5.14: Rank correlation between country name joint probability and population .....	60
Fig. 5.15: Rank correlation between capital city name joint probability and population .....	61
Fig. 5.16: Influence of population ranges on the coefficient of determination between city name joint probability rank and population rank .....	62
Fig. 5.17: Percentage of the correctly retrieved country names bordering a country name .....	66
Fig. 5.18: Percentage of the relevant found country names bordering a country name .....	67
Fig. 5.19: Percentage of the correctly retrieved country names for the spatial relation country in capital .....	68
Fig. 5.20: Percentage of the correctly retrieved country names for the spatial relation city in country .....	69
Fig. 5.21: Percentage of the relevant found country names for the spatial relation city in country .....	70
Fig. 5.22: Country city name density for the spatial relation city in country .....	70
Fig. 5.23: Conditional probability of the 20 most likely countries to follow “delta in” .....	71
Fig. 5.24: First webpage recommendations on Bing for the query “delta in Romania” .....	72
Fig. 5.25: Conditional probability of the 20 most likely countries to follow “lake in” .....	73
Fig. 5.26: Conditional probability of the 20 most likely countries to follow “sea near” .....	73
Fig. 5.27: Conditional probability of the 20 most likely countries to follow “valley in” .....	74
Fig. 5.28: First webpage recommendations on Bing for the query “valley in France” .....	74
Fig. 5.29: Conditional probability of the 20 most likely countries to follow “forest in” .....	75
Fig. 5.30: Conditional probability of the 20 most likely countries to follow “beach in” .....	75
Fig. 5.31: Conditional probability of the 20 most likely countries to follow “mountain in” .....	76
Fig. 5.32: First webpage recommendations on Bing for the query “mountain in England” .....	76
Fig. 5.33: Conditional probability of the 20 most likely countries to follow “volcano in” .....	77
Fig. 5.34: Conditional probability of the 20 most likely countries to follow “glacier in” .....	78
Fig. 5.35: Spatial distribution of most probable geographic feature to precede European country .....	79
Fig. 5.36: Conditional probability of the 20 most likely cities to follow “river near” .....	80
Fig. 5.37: First webpage recommendations on Bing for the query “river near Hastings” .....	80
Fig. 5.38: Conditional probability of the 20 most likely cities to follow “sea near” .....	81

Fig. 5.39: Conditional probability of the 20 most likely cities to follow “forest near” .....	81
Fig. 5.40: First webpage recommendations on Bing for the query “forest near Izmir” (left) and “forest near Smolensk” (right).....	82
Fig. 5.41: Conditional probability of the 20 most likely cities to follow “hill near” .....	82
Fig. 5.42: Conditional probability of the 20 most likely cities to follow “mountain near” .....	83
Fig. 5.43: Conditional probability of the 20 most likely cities to follow “volcano near” .....	84
Fig. 5.44: First webpage recommendations on Bing for the query “volcano near Cartagena” .....	84
Fig. 5.45: Conditional probability of the 20 most likely countries to follow “climbing in” .....	85
Fig. 5.46: Conditional probability of the 20 most likely countries to follow “marathon in” .....	86
Fig. 5.47: First webpage recommendations on Bing for the query “marathon in Lithuania” .....	86
Fig. 5.48: Conditional probability of the 20 most likely countries to follow “rugby in” .....	87
Fig. 5.49: Conditional probability of the 20 most likely countries to follow “skiing in” .....	87
Fig. 5.50: Conditional probability of the 20 most likely countries to follow “tennis in” .....	88
Fig. 5.51: Spatial distribution of most probable sport activity to precede European country .....	89
Fig. 5.52: Conditional probability of the 20 most likely cities to follow “football in” .....	90
Fig. 5.53: Conditional probability of the 20 most likely cities to follow “parkour in” .....	90
Fig. 5.54: First webpage recommendations on Bing for the query “parkour in Mardin” .....	91
Fig. 5.55: Conditional probability of the 20 most likely cities to follow “sailing in” .....	92
Fig. 5.56: Conditional probability of the 20 most likely cities to follow “skiing in” .....	92
Fig. 5.57: Conditional probability of the 20 most likely cities to follow “skydiving in” .....	93
Fig. 5.58: First webpage recommendations on Bing for the query “skydiving in Manchester” (left) and “skydiving in Empuriabrava” (right).....	93
Fig. C.1: Rank correlation in Africa between country name joint probability and population .....	124
Fig. C.2: Rank correlation in Asia between country name joint probability and population .....	124
Fig. C.3: Rank correlation in Australia between country name joint probability and population .....	125
Fig. C.4: Rank correlation in Europe between country name joint probability and population .....	125
Fig. C.5: Rank correlation in North America between country name joint probability and population.....	125
Fig. C.6: Rank correlation in South America between country name joint probability and population.....	126
Fig. C.7: Rank correlation in Africa between capital city name joint probability and population .....	126
Fig. C.8: Rank correlation in Asia between capital city name joint probability and population .....	126
Fig. C.9: Rank correlation in Australia between capital city name joint probability and population.....	127
Fig. C.10: Rank correlation in Europe between capital city name joint probability and population.....	127
Fig. C.11: Rank correlation in North America between capital city name joint probability and population .....	127
Fig. C.12: Rank correlation in South America between capital city name joint probability and population .....	128
Fig. C.13: Conditional probability of the 20 most likely countries to follow “stream in” .....	128
Fig. C.14: Conditional probability of the 20 most likely countries to follow “river in” .....	129
Fig. C.15: Conditional probability of the 20 most likely countries to follow “ocean near” .....	129
Fig. C.16: Conditional probability of the 20 most likely countries to follow “plain in” .....	129
Fig. C.17: Conditional probability of the 20 most likely countries to follow “desert in” .....	130
Fig. C.18: Conditional probability of the 20 most likely countries to follow “hill in” .....	130

Fig. C.19: Conditional probability of the 20 most likely cities to follow “ocean near” .....	130
Fig. C.20: Conditional probability of the 20 most likely cities to follow “lake near” .....	131
Fig. C.21: Conditional probability of the 20 most likely cities to follow “valley near” .....	131
Fig. C.22: Conditional probability of the 20 most likely cities to follow “beach near” .....	131
Fig. C.23: Conditional probability of the 20 most likely countries to follow “cricket in” .....	132
Fig. C.24: Conditional probability of the 20 most likely countries to follow “cycling in” .....	132
Fig. C.25: Conditional probability of the 20 most likely countries to follow “football in” .....	132
Fig. C.26: Conditional probability of the 20 most likely countries to follow “hiking in” .....	133
Fig. C.27: Conditional probability of the 20 most likely countries to follow “mountaineering in” .....	133
Fig. C.28: Conditional probability of the 20 most likely countries to follow “parkour in” .....	133
Fig. C.29: Conditional probability of the 20 most likely countries to follow “snowboarding in” .....	134
Fig. C.30: Conditional probability of the 20 most likely cities to follow “cricket in” .....	134
Fig. C.31: Conditional probability of the 20 most likely cities to follow “cycling in” .....	134
Fig. C.32: Conditional probability of the 20 most likely cities to follow “darts in” .....	135
Fig. C.33: Conditional probability of the 20 most likely cities to follow “golf in” .....	135
Fig. C.34: Conditional probability of the 20 most likely cities to follow “hiking in” .....	135
Fig. C.35: Conditional probability of the 20 most likely cities to follow “rugby in” .....	136
Fig. C.36: Conditional probability of the 20 most likely cities to follow “snowboarding in” .....	136



## Tables

Table 3.1: Total number of place names per continent .....	24
Table 5.1: Continent name joint probability in percentage .....	46
Table 5.2: Joint probability of different variations for the place name America.....	46
Table 5.3: Statistical parameters of the continent name joint probabilities.....	47
Table 5.4: Top five country name joint probabilities in percentage.....	49
Table 5.5: Bottom five country name joint probabilities in percentage .....	49
Table 5.6: Statistical parameters of the country name joint probabilities .....	50
Table 5.7: Top five capital city name joint probabilities in percentage .....	53
Table 5.8: Bottom five capital city name joint probabilities in percentage.....	53
Table 5.9: Statistical parameters of the capital city name joint probabilities .....	54
Table 5.10: Top five city name joint probabilities in percentage.....	56
Table 5.11: Bottom five city name joint probabilities in percentage .....	57
Table 5.12: Statistical parameters of the city name joint probabilities.....	57
Table 5.13: Spearman correlation coefficient of country joint probability rank and population rank .....	60
Table 5.14: Spearman correlation coefficient of capital city joint probability rank and population rank .....	62
Table 5.15: Triplets investigated on correctly retrieved and relevant found spatial relations .....	63
Table 5.16: Percentages of correctly retrieved and relevant found continent names for the spatial relation country in continent.....	64
Table 5.17: Percentages of correctly retrieved and relevant found spatial relations per continent for the triplet country in continent .....	64
Table 5.18: Percentages of correctly retrieved and relevant found continent names for the spatial relation capital city in continent.....	65
Table 5.19: Percentages of correctly retrieved and relevant found continent names for the spatial relation city in continent.....	65
Table 5.20: Percentages of correctly retrieved and relevant found country names for the spatial relation country bordering country.....	67
Table 5.21: Percentages of correctly retrieved and relevant found country names for the spatial relation capital city in country .....	68
Table 5.22: Percentages of correctly retrieved and relevant found country names for the spatial relation city in country .....	70
Table 6.1: Statistical parameters of all place name joint probabilities.....	97
Table B.1: List of countries.....	121
Table B.2: List of capital cities .....	122
Table B.3: List of geographic features .....	123
Table B.4: List of sports activities .....	123



## Abbreviations

API	<i>Application Programming Interface</i>
ASCII	<i>American Standard Code for Information Interchange</i>
GIR	<i>Geographical Information Retrieval</i>
GIS	<i>Geographic Information System</i>
ID	<i>Identification</i>
idf	<i>inverse term frequency</i>
IR	<i>Information Retrieval</i>
tf	<i>term frequency</i>
tf-idf	<i>term frequency - inverse term frequency</i>
WGS84	<i>World Geodetic System 1984</i>

## Definitions

conditional probability	<i>The likelihood of a word given preceding n-words in the Microsoft Web n-gram collection.</i>
corpus	<i>A collection to be searched (usually a set of documents).</i>
gazetteer	<i>A place name dictionary.</i>
geocoding	<i>Relating identified place names to a single referent.</i>
geoparsing	<i>Identifying place names in unstructured content.</i>
gold standard	<i>Document(s) where all place name occurrences and their definite referent are known.</i>
joint probability	<i>The likelihood of n-words occurring in the Microsoft Web n-gram collection.</i>
n-gram	<i>A contiguous sequence of n items from a given text.</i>
precision	<i>The proportion of returned documents/place names which are relevant.</i>
recall	<i>The proportion of possibly relevant documents/place names which were returned.</i>
referent	<i>A particular entity being referred to (i.e. Zurich is a place name with multiple possible referents).</i>
toponym	<i>A place name.</i>



# 1 Introduction

*This chapter introduces the objective and motivation of the thesis. Further focus will be on the context, research questions, significance of this dissertation and giving an overall outline of the chapters to follow.*

## 1.1 Context

The Web is omnipresent in our everyday life. Be it uploading and sharing photos on Flickr<sup>1</sup>, posting our current lunch on Facebook<sup>2</sup>, tweeting about current events on Twitter<sup>3</sup>, searching directions on Google Maps<sup>4</sup> or searching for a specific topic on Google<sup>5</sup> or Bing<sup>6</sup>. Hence, the Web is an ever-growing source of information and pool of social interaction. In this network place names – also known as toponyms – are indispensable. They are used in conversations, messages, maps, news and online services (Hill 2006). It is expected that the majority of documents and webpages contain geographical information in the form of place names and spatial relations. Some authors estimate that approximately 80% of all information – and thus also the information on the Web can be linked to geographic space (Abdelmoty & El-Geresy 2008; Franklin 1992; Huxold 1991).

The extraction of this geo-referenced information is known as geographical information retrieval (GIR). It focuses on indexing, searching, retrieving and browsing geo-referenced information sources (Larson 1996). In particular, the emphasis lies in accessing unstructured documents on the Web and improving the retrieval of geographically specific information (Jones & Purves 2008).

A different perspective is offered by the recent release of n-gram collections from Google (Cohen 2010; Whitney 2010) and Microsoft in 2010 (Oiaga 2010). These collections return the frequency/probability values of arbitrary n-word combinations (i.e. n-grams, where n=1:5) in books and on the Web. In the case of Google, the probability of n-words found in all the sources printed from 1500 to 2008 is returned (“Google Books Ngram Viewer” 2013). These are based on all the scanned books and magazines in Google Books<sup>7</sup>. In the case of Microsoft, the probability of n-words on the Web is returned. These are based on Web snapshots taken in

---

<sup>1</sup> <https://www.flickr.com/>

<sup>2</sup> <https://www.facebook.com/>

<sup>3</sup> <https://twitter.com/>

<sup>4</sup> <https://maps.google.com/>

<sup>5</sup> <https://www.google.com/>

<sup>6</sup> <https://www.bing.com/>

<sup>7</sup> <https://books.google.com/>

2010 and 2013 (“Microsoft Web N-Gram Services” 2014). Ultimately, this allows observing the distribution and characteristics of place names on the Web.

### 1.2 Motivation

GIR mainly focuses on assigning geographic identifiers to textual words in unstructured documents and solving ambiguities (Horak et al. 2011). The first step is known as geoparsing and the second as geocoding. The use of Web n-gram services is different. It needs an input such as a word or word combinations to search for. Hence, Web n-gram services allow searching for word(s) frequencies on the Web. To the authors’ knowledge the frequency of place names on the Web has possibly never been examined.

The use of Web n-grams in the scientific field of geography is also relatively new. Recent studies in this area have been done by using Web tri-grams to explore spatial relations (Derungs & Purves 2014a). However, further researches making use of Web n-grams in geography are scarce. This is quite surprising, as Web n-grams allow easy access to a large amount of data on the Web. The data cannot be obtained by a simple Web search query on Bing or Google. With the use of Web n-grams all the place name frequencies on the Web can be obtained and analyzed. The interesting results can then be verified by querying for them on Web search engines such as Bing or Google.

Given these facts, the main motivation of this master thesis is the research gap of using Web n-grams in geography and exploring the characteristics of place names on the Web. As a byproduct, this research may have further use for other scientific fields or help in solving place name ambiguities in GIR.

### 1.3 Objective and Research Questions

The primary objective of this work is exploring and analyzing place names in the n-gram collection offered by Microsoft (“Microsoft Web N-Gram Services” 2014). Being that, the thesis follows an explorative approach and investigates the characteristics of place names of different geographical hierarchy levels on the Web. The obtained “big data” are then statistically evaluated. Thus, a descriptive analysis of the data takes place with the aim to derive information and making it accessible for human consumption.

The main emphasis, during the exploration of places names on the Web, will be in answering the following research questions:

---

**RQ1** *What are the characteristics of place names occurring in Web n-grams, in terms of spatial coverage or ambiguity?*

---

The first research question centers on a descriptive analysis of place name Web probabilities. Accordingly, the purpose is to get a first glimpse of the (spatial) distribution of place name Web probabilities at continent, country, capital city and city level. These probabilities are then statistically evaluated by observing the minimum, maximum, mean, standard deviation and range. The distribution of place name probabilities is also inspected on resembling a Zipf distribution and on the number of words constituting a place name. In a second step, the possible causes for high place name probabilities are investigated. This is accomplished by testing correlations. Particularly, if a high place name probability is accompanied by a high number of inhabitants. This is followed by an inspection of simple spatial relations between place names. Generally, all of these steps consider how the ambiguity changes at different levels of granularity.

---

**RQ2** *To what degree can spatial information retrieved from Web n-grams be used to describe the world?*

---

The second research question follows an explorative approach and aims to analyze the usability of Web n-grams to describe the world. This is performed by assigning toponyms to geographic features and sports activities. The overall goal is to obtain new insights regarding geographic features and sports activities related to toponyms, on the basis of Web n-grams. Finally, the plausibility of these results is verified.

---

**RQ3** *Can an application, using Web n-grams to link information to place names, be evaluated?*

---

The third research question tries to clarify if the obtained probabilities and results of spatial relationships between place names can be evaluated. This is done by observing the spatial relationships of place names on the Web and comparing them to ground truth. Further insights of the accuracy of Web n-grams are acquired by checking the plausibility of the results received through assigning geolocations to geographic features and sports activities. The interesting results or the results in question are then searched on Bing.

## 1.4 Outline

The master thesis is organized as follows. In chapter 2, the first part provides the theoretical background in the field of GIR with attention to the methods and concepts utilized. Secondly, the use of n-grams in other fields and the emergence of Web n-grams are introduced. Chapter 3 presents the data and software used for obtaining the results in respect to source and structure. The succeeding methods performed on the data are discussed in chapter 4. Especially, the steps undertaken for the pre-processing, processing, analysis, possible evaluation and visualization of the data are described. The results are then presented in chapter 5 and will be the basis for answering the research questions. Subsequently, these results are critically examined in chapter 6 in regard to the research questions and literature. The interpretation and discussion of the research questions take center stage in this chapter. Finally, the findings of this thesis are summarized in chapter 7 with an outlook of future work and possible implementations of Web n-grams.

## 2 Background

*The research conducted in this thesis is relative new and relates to different scientific fields. Therefore, the relevant areas of research are introduced in this chapter. The emphasis is on the retrieval of geographical information, while deeper insights on the different methods and concepts used in this field are given. Moreover, an overview of different scientific fields using n-grams is presented.*

### 2.1 Information Retrieval<sup>8</sup>

Information retrieval (IR) is an academic field of study situated in interdisciplinary field of information science. It deals with finding material of unstructured nature relevant to an information need within large collections which are usually stored on computers. The material of unstructured nature usually refers to text documents which have not been brought into a format easily processed by computers. Nevertheless, most data contain some sort of structure. Text documents for instance have headings, paragraphs, footnotes and sentences which are built on language and grammar.

The idea of computerized IR and the first computer based searching systems started in the late 1940s (Cleverdon 1991; Liddy 2005; Sanderson & Croft 2012). Eventually, the amount of data stored on computers grew steadily and the need of searching archives of text gained in importance. The term IR finally emerged in the 1950s (Mooers 1950; Singhal 2001). In these times, the focus of document retrieval was based on author, title and keywords. Besides that, the few users of information search were specialists trained in the daily use of complex systems. These specialists mainly engaged in a specialized topic such as library catalogues, legal searches or financial information.

The capabilities of retrieval systems rapidly grew with the increase in processor speed and storage capacity of computers (Sanderson & Croft 2012). Furthermore, the systems shifted from manual library-based approaches of acquiring, indexing and searching information to increasingly automated methods (Sanderson & Croft 2012). The main method used by commercial systems in this timespan was the Boolean search. It basically searched for the specified keywords in the collection and only returned documents containing these keywords. The next big change happened with the introduction of the World Wide Web in the early 1990s. With that, a huge amount of information became available to everyone with a computer and an Internet connection. As a result, the search for information became

---

<sup>8</sup> The statements in this section are generally based on the information covered in the chapter 1, 2 and 6 of Manning et al. (2008).

## 2 | Background

mainstream. The sources of digitized unstructured data grew and so did the global access to these enormous quantities of data (Sanderson & Croft 2012). This made IR a necessity for searching and finding relevant information in the huge data pool called the World Wide Web. Consequently, the algorithms developed for IR were employed for searching the Web (Singhal 2001). These methods consisted of using tokens and normalizing them into terms used for the index. The tokens can be seen as sequence of characters in a document which are then grouped (into words) for further processing. In the second step the actual normalization took place. The normalization assigned actually equivalent tokens, which had small differences, to the same term in the index. These were then used for the method known as ranked search. Contrary to the Boolean search, the ranked search returned documents based on the term frequency. This means that the more often a query term (the search in question) occurred in the document the more likely the document had to do with that query. One of the fundamental concepts used in IR to help weighing these terms in documents is known as *tf-idf* (term frequency and inverse document frequency). The *idf* in this weighing scheme is defined as follows:

$$idf_t = \log \frac{N}{df_t}$$

(Manning et al. 2008; Robertson 2004)

The  $idf_t$  is the inversed document frequency of a term  $t$  which is defined as the logarithm to base  $N$  (total number of documents in collection) divided by the document frequency of the term  $df_t$ . This provides information to how common the term is in the collection as a whole. Thus, the *idf* of rare term is high and the *idf* of a frequent term is likely to be low. The *idf* is then combined with the term frequency *tf* to produce a composite weight for each term in each document. This term frequency provides information to how often the term occurs in the document. Finally, the *tf-idf* weighting scheme assigns a weight to term  $t$  in document  $d$  by applying the formula:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

(Manning et al. 2008)

That way each term  $t$  is assigned a weight in document  $d$ . As a result, the weight is highest when the term occurs many times within a small number of documents and lowers when the term occurs fewer times in a document or many documents. This weight is lowest when the

term occurs in virtually all documents. Given that, the simplest way to score and rank documents with the *tf-idf* is by applying the following formula:

$$Score(q, d) = \sum_{t \in q} tf - idf_{t,d}$$

(Manning et al. 2008)

The formula is known as the overlap score measure. For a query  $q$  each document  $d$  is given a score. This score is based on the sum of *tf-idf* weights (expressed in the formula above) for each queried term found in the document  $d$ . Accordingly, the documents with the highest scores are at the top of the retrieved document list.

In recent times, more focus has been given to exploiting query logs. Ergo, user behavior is analyzed in form of commissioned queries, click patterns and reformulation of queries. This enabled researches to develop more efficient query processing techniques based on understanding the user's intent. (Sanderson & Croft 2012)

The measuring of effectiveness of such retrieval systems is of great importance in the field of IR. Therefore, two metrics are used to quantify the effectiveness of a retrieval method. These two metrics are called precision and recall (formulas are introduced in section 2.2). The precision defines the proportion of documents returned which are relevant, while the recall describes the proportion of possibly relevant documents which are returned. The metrics are generally calculated in a controlled empirical experiment where a set of topics (often expressed as queries) are searched in a corpus.

Overall, the use of Web search engines has increasingly gained popularity in the last years. The Web is searched for different topics every day by hundreds of millions of people. This makes the retrieval of information omnipresent. People need Web searches for private use, education or business. This also includes enterprise, institutional and domain-specific searches on centralized file systems for corporate internal documents, patents or research articles.

In general, any user of Web search engines can approve that IR works and succeeds in finding the documents in question. In some cases, these search queries are refined by using place names to obtain more specific results (Gan et al. 2008). For instance, the search for a hotel is refined by adding a place name to the query. A person going on vacation in Berlin would therefore search for "hotels in Berlin". This query might be followed up by a search for attractions or landmarks near Berlin. Based on the location and cluster of landmarks the person might even search for hotels in the center or east of Berlin. Geography and spatial

relations are, in that sense, an important part of IR and are also subject to research (Abdelmoty & El-Geresy 2008; Fu et al. 2005). However, Web search engines are far from perfect. The use of geographic references in queries mostly come short and are treated as any other arbitrary word in Web search engines. Consequently, a research field dealing with these problems emerged. This domain combines IR and spatial information and is commonly referred to as geographic information retrieval (GIR).

### 2.2 Geographical Information Retrieval

The retrieval of geographic information in documents is known as geographical information retrieval (GIR). It focuses on indexing, searching, retrieving and browsing geo-referenced information sources and designing systems to accomplish these tasks (Larson 1996). A newer and more specific definition of GIR is given by Jones & Purves (2008: 220):

GIR is ... concerned with improving the quality of geographically specific information retrieval with a focus on access to unstructured documents such as those found on the Web.

The significance of geographic information is obvious, as the majority of documents and webpages contain some sort geographical information in the form of place names or spatial relations. Hence, location is a common element of information found on the Web. Some authors even estimate that approximately 80% of all information – and thus also the information on the Web can be linked to geographic space (Abdelmoty & El-Geresy 2008; Franklin 1992; Huxold 1991). Moreover, 12.7%-14.8% of queries submitted to Web search engines contain a place name (Sanderson & Kohler 2004; Zhang et al. 2006). GIR can help in distinguishing this geographic information from arbitrary words.

Nevertheless, several aspects of GIR need improvements. These key issues are identified by Jones & Purves (2008) as:

- detecting geographical references within text documents and in queries
- disambiguating place names
- geometric interpretation of vague place names and vague spatial language
- indexing documents (geographic context and non-spatial thematic content)
- ranking the relevance of documents with respect to geography as well as theme
- developing effective user interfaces
- evaluation of GIR

The first point in the list refers to automatically identifying place names in text. This step is known as geoparsing (Gelernter & Balaji 2013; Horak et al. 2011). The main focuses herein lies on recognizing toponyms, while using specific methods to accomplish this task (further discussed in chapter 2.4). This is normally followed by toponym resolution. It is also referred to as geocoding and relates identified place names to a single referent (Horak et al. 2011; Zubizarreta 2009). Thus, allowing the assignment of a geometry. The goal of toponym resolution is therefore to solve place name ambiguities. The different types of place name ambiguities are further described in chapter 2.5, while the various approaches to solve ambiguity are explained in chapter 2.6. Overall, both of the steps above – toponym recognition and toponym resolution – rely on gazetteers to recognize and disambiguate place names. These gazetteers are nothing more than toponym dictionaries. Further information on gazetteers and their structure is provided in chapter 2.3.

The third aspect of GIR, needing improvement, is the use of vague place names and vague relations. Vague place names are areas such as the Swiss Mittelland or the Highlands of Scotland. These regions have no well-defined boundary and the extent of the place differs from many people's perception (Hart & Dolbear 2007; Jones et al. 2008). Additionally, the spatial language accompanying place name terminology can be as vague as some of the place names (Jones & Purves 2008). Examples of spatial language include terms such as near, close, between and north/east/south/west of. Lastly, it is important to interpret such terms and locate vague places to facilitate the analysis of geographic context in documents (Davies et al. 2009; Jones & Purves 2008).

More challenges of GIR include the indexing of documents, document relevance scores and effective user interfaces. The indexing of documents is mainly concerned with assigning a spatial index that records which documents relate to a particular region. This is usually done by assigning a document footprint based on the geographic reference in it. The challenges lie in the occurrence of multiple geographic references in a document and finding efficient ways to combine text and spatial indexes. The relevance of documents based on a search query, is a further hurdle for GIR. Especially, calculating and assigning scores to documents based on the query term frequency and spatial references poses a challenge. Furthermore, how Web search engines should be designed to facilitate the search for documents is important. Therefore, GIR focuses on efficient ways to design Web search engines based on geographic scope, user needs and feedback. (Jones & Purves 2008)

The evaluation in GIR stands at the end of the processing chain and is similar to IR. This means the effectiveness of the retrieval method is quantified by precision and recall. In IR the

## 2 | Background

focus is on the amount of documents retrieved. However, in GIR the recognition of geographic information and toponym resolution in documents take center stage in the evaluation. This is normally done by comparing the automatic retrieval of toponyms in document(s) compared to a gold standard. The gold standard represents the document(s), where all place name occurrences and their definite referent are known.

The precision in GIR stands for the percentage of correctly retrieved or correctly disambiguated place names in a document or set of documents (Grover et al. 2010; Leidner 2007). This is manifested in the following formula:

$$precision = \frac{\text{relevant toponyms} \cap \text{retrieved toponyms}}{\text{retrieved toponyms}} \quad (\text{Leidner 2007})$$

The precision is a measure of how many selected place names are relevant. In the case of toponym resolution, the ratio would simply be the correctly disambiguated toponyms divided by the total of toponyms the system tried to resolve. Whereas, the recall is the percentage of place names found or correctly resolved place names in regard to the gold standard (Grover et al. 2010; Leidner 2007). The ratio for the recall is expressed as:

$$recall = \frac{\text{relevant toponyms} \cap \text{retrieved toponyms}}{\text{relevant toponyms}} \quad (\text{Leidner 2007})$$

The recall is a measure of how many relevant place names are selected in a document or set of documents. Regarding toponym resolution, the ratio is expressed as the number of correctly resolved place names and the number of all place name instances. On the one hand, a system has a high precision if it correctly recognizes a large amount of toponyms (toponym recognition) or assigns the right referents to the detected toponyms (toponym resolution). This means a low precision is caused by a high number of wrongly recognized toponyms or wrongly resolved toponyms. On the other hand, a system has a high recall if it finds the majority of the place names (toponym recognition) or assigns the right referent to the relevant place names (toponym resolution). This recall would be low if a number of the relevant place names have not been found or are wrongly resolved compared to the gold standard. In general, the goal of a system recognizing and resolving toponyms is to find the right balance between precision and recall. As an increase in precision usually causes a lower recall, while an increase in recall normally results in a lower precision.

### 2.3 Gazetteer

A gazetteer is a geospatial dictionary of geographic names, concisely a toponym dictionary (Hill 2006). They first started off as printed documents and exist now as digital libraries or knowledge information system (Hill 2006; Goodchild & Hill 2008). Gazetteers consist of toponym entries where each entry has at least a name, a location and a type (Hill 2000). Some even have further information of place names in form of population, area and alternative names (Leidner et al. 2003). The location or spatial footprint of a toponym is mostly represented as the latitude and longitude of a point (Hill 2000; Leidner 2007). Meanwhile, the type in a place name entry is assigned to one geographical feature type class such as country, county, city, airport or any other class (Hill 2000; Leidner 2007). As an example, the place name entry of Paris in a Gazetteer is shown below (“GeoNames” 2014). The example shown is simplified for illustration purposes:

<b>ID:</b>	2988507
<b>Name:</b>	Paris
<b>Coordinates:</b>	48.85341, 2.3488
<b>Alternative Names:</b>	Baariis, PAR, Paarys, Pari
<b>Class:</b>	Capital of a political entity
<b>Hierarchy:</b>	Europe > France > Île-de-France > Paris
<b>Population:</b>	2 138 551

Currently, the number and types of gazetteer resources are increasing as commercial gazetteer become supplemented by volunteered sources of geographic place name data (Smart et al. 2010). These tend to vary in scope, coverage and detail of granularity (Buscaldi 2011). The scope indicates whether a gazetteer is limited to region or a country. The coverage is defined by the number of places listed in a gazetteer, while the detail signifies how fine-grained a gazetteer is with respect to the area covered. One of the more common known gazetteers supplemented by volunteered sources is GeoNames<sup>9</sup>. It boasts a total of 10 million geographical names and consists of over 9 million unique features whereof 2.8 million are populated places and 5.5 million are alternate names (“About GeoNames” 2015).

In a nutshell, gazetteers are an integral part of GIR, notably in toponym recognition and toponym resolution (Amitay et al. 2004; Janowicz & Kessler 2008). The explicit use of gazetteers for toponym recognition is further discussed in the upcoming chapter, while the use of gazetteers for disambiguating place names is presented in section 2.6.

---

<sup>9</sup> <http://www.geonames.org/>

## 2.4 Toponym Recognition<sup>10</sup>

Toponym recognition is the process of identifying toponyms in a text, also known as geoparsing. There are multiple ways to recognize toponyms. Specifically, three fundamental approaches are used to recognize place names in text. These approaches are based on gazetteer lookup, rules and machine learning (Leidner & Lieberman 2011). Anyhow, the following chapters are dedicated to describing the characteristics of the different methods used for toponym recognition.

### 2.4.1 Gazetteer Lookup

The gazetteer lookup based approach analyzes the text by looking at each word or character. In this process, each word or character are searched for occurrences of a predefined set of toponyms. These predefined toponyms are taken from a gazetteer. Therefore, extra care has to be taken in form of multi-word toponyms such as New York City. This then may lead to inefficiencies if a naïve lookup based approach is used. Moreover, this approach can fail if the place name entries in the gazetteer are not used in a geographical sense in texts (Clough 2005; McCurley 2001). This refers to place names entries which are equal to names of people, businesses or common language use. The data quality in the gazetteers is also an issue, as they are incomplete and noisy in nature. Hence, gazetteer lookup is unable to identify place names which are not in the gazetteer and cannot distinguish locations used in a geographical context (Clough 2005; McCurley 2001). In addition, the administrative boundaries of places and names may change over time. This requires an integrated and automated workflow to keep the data up to date in gazetteers. Nevertheless, gazetteer lookup is simple, robust, language independent and often effective which is needed for ungrammatical webpages containing limited context (Clough 2005; Mikheev et al. 1999).

There is also the possibility of using an ontology-based approach, which is nothing more than an organized gazetteer. The set of toponyms are organized as a hierarchy and enable to query for more complex relationships such as containment and adjacency (Smart et al. 2010).

Overall, the gazetteer lookup is a useful first step to identify place name candidates. But, it should be combined with further toponym recognition methods to reduce its flaws. This is currently done in the project of SPIRIT (Spatially aware Information Retrieval on the Internet). The gazetteer lookup is used to identify locations in reasonable time, while context rules are applied to filter out place names used in a non-geographical sense (Clough 2005; Purves et al. 2007).

---

<sup>10</sup> The subdivision of chapters and their contents are built on the propositions from Leidner & Lieberman (2011).

### 2.4.2 Rule Based

The rule based approach uses a set of symbolic rules based on the language of the searched documents. In that sense, regular expression or other more complex ways are used to express rules to identify toponyms in text. A possible set of useful rules is shown below:

- (1) city of [A-Z][a-z]
- (2) near [A-Z][a-z]
- (3) [A-Z][a-z]+berg

In the first example every word starting with “city of” and a capital letter is defined as a toponym. The next example defines a rule where every word arrangement starting with the spatial relation “near” and is followed by a capitalized word, is a place name. Lastly, the third rule applies for mountains in German speaking areas. Thus, a capitalized word followed by lower case character(s) and “berg” is recognized as a mountain. This is a perfect example of the influence of language for rule based approaches. In German speaking countries mountains usually contain “berg”, “horn” or “spitz” in their name. However, most nouns in German are capitalized which makes the use of rule based approaches more difficult. In English few words are capitalized. The capitalization is mostly reserved to names such as surname, company name or geographic name.

Rule based approaches can work well. Nonetheless, they typically have many rules which have to be adapted for specific cases and the set of documents in question. The general strong points of the rule based approach lie in hierarchically processing the documents and identifying clear geographic references. This information can then be used as context for further analysis.

### 2.4.3 Machine Learning

The machine learning based approach aims to learn from data. It consists of four key stages. The first stage is the construction of annotated training data. This is normally a text where the place names have already been identified by somebody. Thus, the training data are nothing but a human-tagged gold standard text (Leidner 2007). The next key step lies in choosing features for the learning algorithm. This is done by moving a sliding window over the text, while at each position a set of properties known as features are computed. The features are then controlled on specific properties like word length, capitalization and presence of propositions followed by a proper noun. These features are essential for the learning algorithm. The third key stage is choosing an appropriate algorithm

(e.g. Naïve Bayes or support vector machine) for the classification of place names. Finally, the fourth key stage is to run the algorithm on the training data. Based on the training data, the feature configurations that are most highly correlated with toponyms are extracted. The evaluation of the algorithm is then done by testing it on an independent evaluation data set.

The major plus of machine learning approaches is their capability to adapt to each case. This makes them flexible and usable for the retrieval of information in different domains and languages (Freire et al. 2011).

### **2.5 Toponym Ambiguity**

The toponym recognition is followed by the toponym resolution. Its main objective is to solve toponym ambiguities. These are situations where an identified place name can refer to multiple possible entities. Most of the place names and names found on the Web are ambiguous (Amitay et al. 2004). Therefore, the resolution of these ambiguities is essential to GIR (Jones & Purves 2008). There are two types of ambiguities: geo/non-geo ambiguity and geo/geo ambiguity (Amitay et al. 2004). In this section, their properties are further described and complemented with examples.

#### **2.5.1 Geo/Non-Geo Ambiguity**

Geo/non-geo ambiguity refers to the use of terms to name a place as well as a different concept (Ahlers & Boll 2007). Thus, place names can have other non-geographic meanings (Amitay et al. 2004; Freire et al. 2011). This is also known as semantic ambiguity. Examples of such incidents would be Georgia and Turkey. Georgia can either refer to a country or a person, while Turkey can refer to an animal or a country (Freire et al. 2011). There are also instances where some of the most common words in English are place names: “As” in Belgium, “Of” in Turkey and “To” in Myanmar (Amitay et al. 2004). Most of these problems arise due to the fact that the capitalization of words is ignored or all words in a text are converted to lowercase. Naturally, the ambiguities also vary with language. The place name Stein in the canton Aargau, Switzerland may also refer to stone in German. Additionally, other place names may vary in spelling depending on the language. In such cases, gazetteers containing alternative names in place name entries can provide assistance.

Overall, geo/non-geo ambiguity tends to be tricky because many of the geographic names may have a matching common word in one language or another (Zubizarreta et al. 2009). This is mostly addressed by toponym recognition.

### 2.5.2 Geo/Geo Ambiguity

Geo/geo ambiguity arises when different places share the same name or a place has different possible spellings (Ahlers & Boll 2007; Amitay et al. 2004; Freire et al. 2011). This is also referred to as geographic ambiguity. An example of such occurrences would be London, Paris or Springfield. London is the capital city of the United Kingdom, but also a city in Ontario, Canada. The city Paris exists in France and in the USA, while Springfield occurs 63 times in the USA alone (Amitay et al 2004). Generally, almost every major city in Europe has sister city of the same name in America (Amitay et al. 2004; Freire et al. 2011). In average 28% of the location names worldwide refer to more than one place (Smith & Crane 2001). Fortunately, the context of a text provides useful information to narrow down the possible referent location (Zubizarreta et al. 2009). The resolution of such geo/geo ambiguity is subject to toponym resolution.

## 2.6 Toponym Resolution<sup>11</sup>

Toponym resolution is the process of relating identified toponyms to a single referent and explicit location. This is also called geocoding. The toponym resolution focuses on resolving toponym ambiguities (discussed in the previous section 2.5) and is the next step following the toponym recognition. Usually, humans are quite good at dealing with place name ambiguities. Mostly, due to the fact that people can communicate with each other and have cognitive abilities. Computers are rather bad at solving ambiguities. Therefore, computers need specific methods to aid in the process of disambiguation. These may vary according to the nature of documents. The succeeding sections introduce the different approaches for toponym resolution. In particular, these are made up of the map based, knowledge based and data driven approach (Buscaldi 2011; Buscaldi & Rosso 2008).

### 2.6.1 Map Based

The map based approach uses explicit representation of places on a map. It uses geometry of candidate referents for the toponym resolution. Therefore, the method mainly needs the coordinates of places appearing in a text. The geometry or coordinate are then used for topology (e.g. containment) or distance based measures. However, additional information is required to calculate the distances with the help of a unique referent. These can be available in the text in form of locations with unique referents, location of author/contributor/owner or

---

<sup>11</sup>The subdivision of chapters and their contents are built on the suggestions from Buscaldi (2011), Buscaldi & Rosso (2008).

## 2 | Background

content information. With this information the average distance between unambiguous context toponyms and referents is calculated. For instance, the city Canberra and Newcastle occur in a text document. The city Canberra is an unambiguous place name located in Australia, whereas Newcastle is an ambiguous place name. Therefore, the distances between Canberra and the possible referents for Newcastle are calculated. Based on the distances, it is more likely that Newcastle refers to the city in Australia rather than Newcastle in the United Kingdom. This approach is usually very sensitive to changes in context. Consequently, it is necessary to remove places that are very far away on average from the other locations in the text (Smith & Crane 2001). This was done in Smith & Crane (2001), where all candidate referents were mapped and weighted by the number of time they appear in the text. Afterwards, a centroid was calculated and the candidate referents more than two standard deviations away from the centroid were discarded. Then a new centroid was calculated and the remaining place name referents were given scores based on importance, proximity to other toponyms and proximity to centroid. Based on these scores, a number of referents were kept and the final centroid was calculated. Finally, the candidate referent closest to the final centroid was chosen as the correct location.

The inclusion of external knowledge and the position of the source of a text can also greatly improve the map based approach (Buscaldi & Magnini 2010). For example, the occurrence of the place name Paris in a Texan newspaper is more likely to refer to Paris in Texas, United States rather than to Paris in France. The source of information can therefore be very useful for the disambiguation of local text collections. Buscaldi & Magnini (2010) have even shown that 76.2% of place names mentioned in an Italian newspaper are located within 400km of the city where the newspaper is published. Moreover, ambiguous toponyms tend to be found closer together, which means that ambiguous toponyms are spatially autocorrelated (Brunner & Purves 2008). This may cause problems for map based approaches and call for another disambiguation method.

### 2.6.2 Knowledge Based

The knowledge based approach is similar to rule based approach (in section 2.4.2), as it uses simple rules to decide between candidate referents. These rules mostly compare the area, population or population density between the possible locations referring to the place name in the text (Clough et al. 2004; Li et al. 2003). Thus, more populated places are more likely to be mentioned in a text. These heuristics were used by Amitay et al. (2004) and Rauch et al. (2003). An example of disambiguating a place name based on population can be demonstrated

with Hamburg. The city Hamburg can refer to multiply cites: Hamburg in Germany or Hamburg in Pennsylvania, USA. Based on the data offered by the gazetteer GeoNames (“Download” 2014), Hamburg in Germany has a population of 1 739 117 and Hamburg in the USA has a population of 4 289. Therefore, the place name Hamburg is assigned to Hamburg in Germany. Gazetteers are often used to obtain place name information for knowledge based disambiguation (Olligschlaeger & Hauptmann 1999; Rauch et al. 2003). Other knowledge based rules for disambiguating place names include: one sense per discourse, default sense or hierarchical containment (Bensalem & Kholadi 2010). The hierarchical containment is based on the fact that places in the context are usually contained in the same region or geographical area. For instance, a text mentions the city London and the United Kingdom. The city London is part of the United Kingdom and therefore must refer to London, UK rather than London, Ontario in Canada.

A completely different knowledge based approach is used by Overell (2009). The approach takes advantage of Wikipedia by using article templates, categories and links to other articles in Wikipedia. With this information co-occurrence based rules are generated. These imply that specific areas are highly related to specific properties or co-occurrences. An example would be London and snow. This might imply London, Ontario in Canada, since snow is more common in Canada rather than the United Kingdom.

In summary, knowledge based toponym resolution is built on rules derived from the knowledge of locations or investigated texts at hand.

### 2.6.3 Data Driven

The data driven approach is similar to machine learning approach (in section 2.4.3), as it uses standard machine learning methods (Overell & R ger 2008). Commonly, the approach is not used for toponym resolution. This is due to the lack of geographically tagged data and the problem of unseen place names. Hence, the method would require a large accurate corpus of annotated ground truth (Overell & R ger 2008). If such a corpus existed, na ve methods or latent semantic classification could be applied (Grossman & Frieder 2004). Nevertheless, supervised classifiers have been implemented with mixed results. The advantage of these methods is that they can exploit non-geographical content and build probabilistic models (Roberts et al. 2010). These contain spatial relationships between non-geographical entities and places. Thus, allowing referring a place name to a single referent based on a known person or organization residing at a fixed location.

### 2.7 Data Mining

The previous chapters showed how geographical information can be extracted and disambiguated from the Web and documents. However, there is also a scientific field which focuses on the extraction of implicit, previously unknown and potentially useful information from data (Witten & Frank 2005). This field is known as data mining and is part of knowledge discovery (Fayyad et al. 1996; Maimon & Rokach 2010). The main emphasis lies on discovering meaningful new correlation, patterns and trends by mining large databases (Lew & Mauch 2006; Maimon & Rokach 2010; Waller & Fawcett 2013). These tasks are accomplished by using statistical, computational, machine learning, artificial intelligence and data visualization techniques (Lew & Mauch 2006). Usually, data mining and knowledge discovery is an iterative process. It involves multiples steps, including data selection, cleaning, preprocessing, transformation, interpretation and evaluation (Fayyad et al. 1996). Overall, data mining and knowledge discovery is exploratory in nature and uses more inductive than traditional statistical methods (Hirji 2001; Mennis & Guo 2009). It has already been used in the medical, manufacturing, aerospace, chemical, marketing, finance, telecommunication and Internet agent sector (Fayyad et al. 1996; Lew & Mauch 2006).

Data mining has two primary goals: prediction and description (Clifton 2014; Fayyad et al. 1996; Maimon & Rokach 2010). The prediction is often referred to as supervised data mining (Maimon & Rokach 2010). It uses prediction oriented methods for the automated building of a behavior model which obtains new and unseen samples and is able to predict values of one or more variables related to the sample (Clifton 2014; Maimon & Rokach 2010). These prediction oriented methods include regression, neural networks, decision trees, support vector machines. In summary, the emphasis of prediction in data mining is in predicting unknown values or future values (Fayyad 1996).

The descriptive data mining includes unsupervised and visualization aspects of data mining (Maimon & Rokach 2010). To accomplish this task the data is clustered or divided into groups to find specific patterns (Clifton 2014). The methods used in this process are clustering, anomaly detection, association rule learning, principal component analysis or affinity grouping.

Generally, the two methods – predictive and descriptive – have no sharp boundary and can flow into one another (Fayyad 1996). Caution also has to be taken, that the data mining process does not lead to an endless search of what the data really say (Barton & Court 2012).

In this master thesis, the information found in the n-gram collection from Microsoft is mined by using methods commonly associated with GIR. The approach is mainly explorative and

statistically analyzes the data in the collection, while visualizing the results. Further information to n-gram and their use is found in the next section.

## 2.8 N-Gram

An n-gram is a contiguous sequence of n items for a given sequence of text. The items can refer to words or characters. These sequences are typically collected from a text corpus (“n-gram” 2014). The use of n-grams spans a variety of scientific fields. For instance, n-grams and associated frequency/probabilities are popular for structuring and indexing large natural language corpora such as the Web. These mainly concern computational linguistics and probability (Brown et al. 1992; Goodman 2001). Other fields making use of n-grams include: machine translation (Manning et al. 2008; Mariño et al. 2006), query auto-completion (Nandi & Jagadish 2007; Manning et al. 2008), malware detection and internet security (Abou-Assaleh et al. 2004), plagiarism (Stamatatos 2011) and identifying genome sequences (Tomovic et al. 2006). Additionally, n-grams have been of use in digital humanities to study the change of language, grammar or cultural trends over time (Michel et al. 2011).

The introduction of n-gram collections from Google (Cohen 2010; Whitney 2010) and Microsoft (Oiaga 2010) in 2010 makes n-grams available to the general public. These collections return the frequency/probability values of arbitrary n-word combinations (i.e. n grams, where n=1:5) in books and on the Web. Google returns the probability of n-words found in all the sources printed from 1500 to 2008 is returned (“Google Books Ngram Viewer” 2013). These are based on all the scanned books and magazines in Google Books. Additionally, different searches are possible. The searches include: part-of-speech tags, wildcards, inflections, arithmetic compositions and corpus selection (“Google Books Ngram Viewer” 2013). The part-of-speech tags differentiate between the searched word as verb or noun. The wild card searches return the ten most popular words following a word, while the inflections return grammatical variations. Lastly, these searches can be initiated on different language corpora. This means the frequency of a word can be searched in all the scanned books in English, German, French and other languages. All of these data can either be accessed by commissioning single queries on the Google Ngram Viewer<sup>12</sup> webpage or downloading the raw data in form of various alphabetical GNU Zipped Archive files.

In the case of Microsoft, the probability of n-words on the Web is returned. These are based on Web snapshots taken in 2010 and 2013 (“Microsoft Web N-Gram Services” 2014). The n-gram collection can be accessed with an open access API and returns word probabilities and

---

<sup>12</sup> <https://books.google.com/ngrams>

word auto-completes. These are returned if a query or multiple queries are commissioned to the API. Further information to the n-gram collection of Microsoft is offered in chapter 3.1.

Altogether, it seems surprising that n-grams have barely been used for linking the information on the Web to geographic space. To the authors' knowledge no research has been done in the context of n-grams and place names, except for Derungs & Purves (2014a). They investigated vague spatial relations with web tri-grams. This was done by deriving near spatial relationships for a large number of populated places in Great Britain. Their results showed that linking n-grams to space can have considerable potential. However, the limitation of only accessing tri-grams instead of whole text documents brought challenges in understanding possible cases of ambiguity. These potential cases of ambiguity were so high that 90% of the initial data set had to be discarded. Hence, ambiguity in n-gram corpora should not be ignored.

Finally, Web n-grams provide a rich resource to explore geographic concepts (Derungs & Purves 2014a). They could offer new possibilities and insights in place name characteristics on the Web.

### 3 Data

*The chapter gives an in-depth look at the Microsoft Web N-Gram Service and its capabilities. This is followed by the inputs used for the Web N-Gram Service, which consist of place name lists, a list of geographic features and a list of sports activities. As well as supplementary information in respect to the place names is described. Lastly, the software used for data preprocessing, processing, statistical analysis and visualization is presented.*

#### 3.1 Microsoft Web N-Gram Service<sup>13</sup>

The Microsoft Web N-Gram Service is an open access API from Microsoft Research in partnership with Microsoft Bing. It is made available to the public via a cloud-based platform to drive discovery and innovation in Web search, natural language processing, speech and related areas through real-world Web-scale data. This permits easy access to petabytes of data. The data includes content types such as document body, document title, anchor texts, query and all the Web documents indexed by Bing in the EN-US market. The total amount of webpages is at the order of hundreds of billions, whereas the spam and low quality webpages are excluded by using Bing's proprietary algorithms. Then the Web documents are downloaded, parsed and tokenized by Bing, while the text is lowercased and the punctuation is removed. Additionally, the frequency of n-words (with  $n = 1:5$ ) on the Web and autocompletes of n-words are made available. The former is based on web models and n-gram models from Web snapshots taken in 2010 and 2013, while the latter is established with query models and n-grams models from Bing queries in 2010 and 2013. These n-gram models are based on the CALM algorithm from Wang & Li (2009) which dynamically adapts the n-gram models as Web documents are crawled. As a consequence, it is ensured that the model is kept up to date with the Web contents and that duplicated contents do not have an outsized impact in biasing the n-gram statistics.

The frequency of initiated n-words in the corpus of Web snapshots is returned as logarithmic probability. It serves to quantify how likely a word occurs in the corpus. The probability is returned as two different measures: joint probability and conditional probability. The joint probability is the likelihood of n-words occurring on the Web, whereas the conditional probability is the likelihood of a word given preceding n-words. For the autocompletes a list of most likely words following n-words is returned with their corresponding conditional probability.

---

<sup>13</sup> The statements in this section are largely based on Microsoft Research ("Microsoft Web N-gram Services" 2014) and Wang et al. (2010).

## 3.2 Lists

The inputs used for the Microsoft Web N-Gram Service consist of place name lists, a list of geographical features and a list of sports activities. Additionally, the population and area of place names are used for further analysis of the data. Their contents and sources are described in this section. An overview of the total number of place names per continent is presented in table 3.1. It also displays the number of cases where a country and thus its cities can either be associated with Europe or Asia.

### 3.2.1 Continents

The list of continents was self-made and is identical to the seven-continent model. Hence, the list contains a total of seven continents. These are as follows: Africa, Antarctica, Asia, Australia, Europe, North America and South America.

### 3.2.2 Countries

A list of all countries was taken from Wikipedia (“List of sovereign states” 2014). The list contains 206 states and is divided by the United Nations (UN) system (“Member States” 2014) into three categories: 193 member states, two observer states and eleven other states. Moreover, the sovereignty of 190 states is undisputed and the sovereignty of 16 states is disputed. In summary the list (“List of sovereign states” 2014) takes all states into account which:

- a) Consider themselves sovereign (e.g. through a declaration of independence) and satisfy the declarative theory of statehood (possessing a permanent populations, a defined territory, a government and a capacity to enter into relations with other states).
- b) Are recognized as a sovereign state by at least one UN member state.

The estimates of the population per country were downloaded as a Microsoft Excel file from the UN website of the Department of Economic and Social Affairs (“Population Division” 2014). These are based on the 2012 revision and contain the total population of both sexes annually for the years 1950 to 2010. However, some disputed states are not represented or a missing information to the number of inhabitants. The figures of inhabitants per country are presented in thousands, whereas merely the records of the year 2010 are taken into consideration for the further analysis.

A list of countries by total area in square kilometers is taken from Wikipedia (“List of countries and dependencies by area” 2014). This total area refers to all land and water areas within the international boundaries and coastlines of a country. The data is mostly taken from

the United Nations Statistics Division and missing/disputed countries are supplemented through other sources. Nevertheless, the data also contains gaps and not all countries have an assigned total area.

### 3.2.3 Capitals Cities

The list of capitals cities was taken from Wikipedia (“List of national capitals in alphabetical order” 2015) and contains a total of 214 capitals. However, the list was reduced to the capital cities belonging to the countries listed in chapter 3.2.1. In fact some of these contain multiple capitals cities. The reason for this is the inclusion of different cities which operate as the seat of some or all part of the government. Hence, the seat of the administration, executive, government, judicial bodies, legislature and others are distributed to different cities.

The population estimates were also taken from Wikipedia (“List of national capitals by population” 2015), while the numbers of inhabitants per capital are taken from different sources and their estimates are of different years. The population statistics are based on the official capital city area and omit the wider metropolitan districts. Nevertheless, the data is also subject to gaps and some capitals are missing or miss a population estimate.

### 3.2.4 Cities<sup>14</sup>

A list of cities was downloaded as tab-delimited text file from GeoNames (“Download” 2014). This list contains all cities with a population over 1000 inhabitants or seats of administrative division. More precisely, the list of cities also refers to all towns, villages, populated places and other agglomerations of buildings where people live and work. It also has information to place names, such as the place name in ASCII characters, alternate names, latitudes, longitudes, feature class, population, elevation, time zone and modification date. These facts and figures are occasional available for the 143 337 “cities” in the list.

Overall, the data is licensed under a Creative Commons Attribution 3.0 License (“Creative Commons” 2015) and is without warranty or representation of accuracy, timeliness or completeness. Additional information to the wide variety of data sources used by GeoNames can be found under data source tab on the website (“Datasources used by GeoNames in the GeoNames Gazetteer” 2014).

---

<sup>14</sup> The statements in this section are largely based on the information offered by GeoNames (“Download” 2014; “GeoNames” 2014).

**Table 3.1: Total number of place names per continent**

Continent Name	Number of		
	Country Names	Capital Names	City Names
Africa	56	60	4211
Antarctica	0	0	1
Asia	47	48	38008
Australia	16	16	1464
Europe	45	45	61966
North America	23	23	24148
South America	12	14	6245
Cases of Asia & Europe	7	8	7294
<b>Sum of Place Names</b>	<b>206</b>	<b>214</b>	<b>143337</b>

### 3.2.5 Geographic Features

The list of geographic features was made with the help of the empirical study from Battig & Montague (1969) and Smith & Mark (2001). In the case of Battig & Montague (1969), 442 students from the Universities of Maryland and Illinois were tested on their responses in a number of verbal categories. The students had 30 seconds to write down their response for each of these categories, while one of the categories was a natural earth formation. The responses with the highest count in this category were used to make up a list of geographic features. Smith & Mark (2001) conducted a similar study where students had to write down as many examples as they could think of in 30 seconds for each category. These categories were made up of five variations of geographic features: a kind of geographic feature, a kind of geographic object, a geographic concept, something geographic and something that could be displayed or portrayed on a map. Additionally, only the terms which were mentioned by 10% of the subjects in a category were displayed in the results. The results in the category “a kind of geographical feature” and “a geographic concept” aided in choosing a list of geographic features. Finally, a portion of the responses in the studies were used for the list of geographic features. These were mainly focused in the physical domain of geography, while some additional geographic features were added by hand. The final list is made up of 15 geographic features. The appendix B.4 can be viewed for further information regarding the geographic features constituting the final list.

### 3.2.6 Sports Activities

The list of sports activities is made up of different sports activities listed in the Wikipedia article “list of sports” (2015). The list from Wikipedia contained several different sport activities. Therefore, a selection of different sports was conducted. This was done by intuitively choosing popular sports activities or sports activities bound to certain places. The final list contains 42 different one worded sports activities. These can be thoroughly looked at in the appendix B.5.

### 3.3 Map

A map of the world was taken from the University of Zurich data spaces server. The map is provided by the company ESRI<sup>15</sup> and was produced by DeLorme<sup>16</sup>. The world map from 2012 contains the geometry of countries, water bodies, ocean and latitude and longitude grids. The geometry of the countries is supplied as vector representation, while the reference system used is WGS84. In addition, the names of all countries are available in a data table.

### 3.4 Software

In this master thesis, different kind of software has been used. The development platform for accessing the Microsoft Web N-Gram Service API was NetBeans<sup>17</sup> IDE 8.0.1. The implementation was done with the computer programming language Java<sup>18</sup>. Additionally, Notepad++<sup>19</sup> was used to adjust the obtained results before importing them into Microsoft Excel. In Microsoft Excel, the different lists were created and the results were analyzed. Hence, it was used for processing the data and analyzing them. Further statistical investigations such as boxplots and correlations were done with the help of IBM SPSS Statistics 21. The cartographic processing and visualization of the data was done with ArcGIS<sup>20</sup> 10.2.2. Furthermore, ColorBrewer 2.0 (Brewer 2013) assisted in creating appealing color palettes and schemes for ArcGIS.

---

<sup>15</sup> <http://www.esri.com/>

<sup>16</sup> <http://www.delorme.com/>

<sup>17</sup> <https://netbeans.org/>

<sup>18</sup> <https://www.java.com/>

<sup>19</sup> <https://notepad-plus-plus.org/>

<sup>20</sup> Maps throughout this thesis were created using ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license. Copyright © Esri. All rights reserved. For more information about Esri® software, please visit [www.esri.com](http://www.esri.com).



## 4 Methodology

*The following chapter describes the handling of the data and the steps taken to process the data. In particular, it takes a closer look at the data pre-processing, the queries used for the Microsoft Web N-Gram Service, the data used for further analysis of the data, the method for evaluating the data and the visualization of the data.*

The main objective of this master thesis is to analyze the characteristics of place names and spatial relationships in the corpus of the Microsoft Web N-Gram Service. This is accomplished by initiating queries to the Microsoft Web N-Gram Service and examining the returned probabilities. These should assist in addressing the research questions from section 1.3. An overview, of all the steps involved to obtain the necessary results and utilized inputs, is available in figure 4.1. Furthermore, the respective chapters discussing the method or data are also illustrated in the figure.

At the center of the project is the Microsoft Web N-Gram Service (chapter 3.1). To the left side of the figure, are the queries (chapter 4.2) which are commissioned to the Web N-Gram Service. Hence, the left side of the method overview illustrates the inputs (lists in 3.2) and the data pre-processing (section 4.1) which took place. The triplets (section 4.2.2) and triplets for the explorative approach (chapter 4.5) are created with the help of the lists in section 3.2. These are commissioned as doublets to the Microsoft Web N-Gram Service.

To the right side of the figure, are the probabilities which are returned from the Microsoft Web N-Gram Service. These are complemented with additional information (chapter 4.3.2) and newly calculated information. The information to place name population was added from the sources listed in section 3.2. The data is then analyzed in regard to place name probabilities (section 4.3.1), Zipf distribution (section 4.3.3) and correlation (section 4.3.4). Additionally, ratios are calculated in the chapter correlation for further use in the explorative approach. The matches from the 1000 autocompletes of the places names doublets are compared with the assembled ground truth (chapter 4.4.1). This step serves as an evaluation (chapter 4.4) of place name relationships on the Web. For this purpose, two measures are used: correctly retrieved spatial relations (section 4.4.2) and relevant found spatial relations (section 4.4.3). The matches from the 1000 autocompletes of the explorative approach (chapter 4.5) doublets are returned in a separate step. Finally, all the received data is analyzed (in chapter 5) and visualized in form of charts, tables or maps (chapter 4.6).

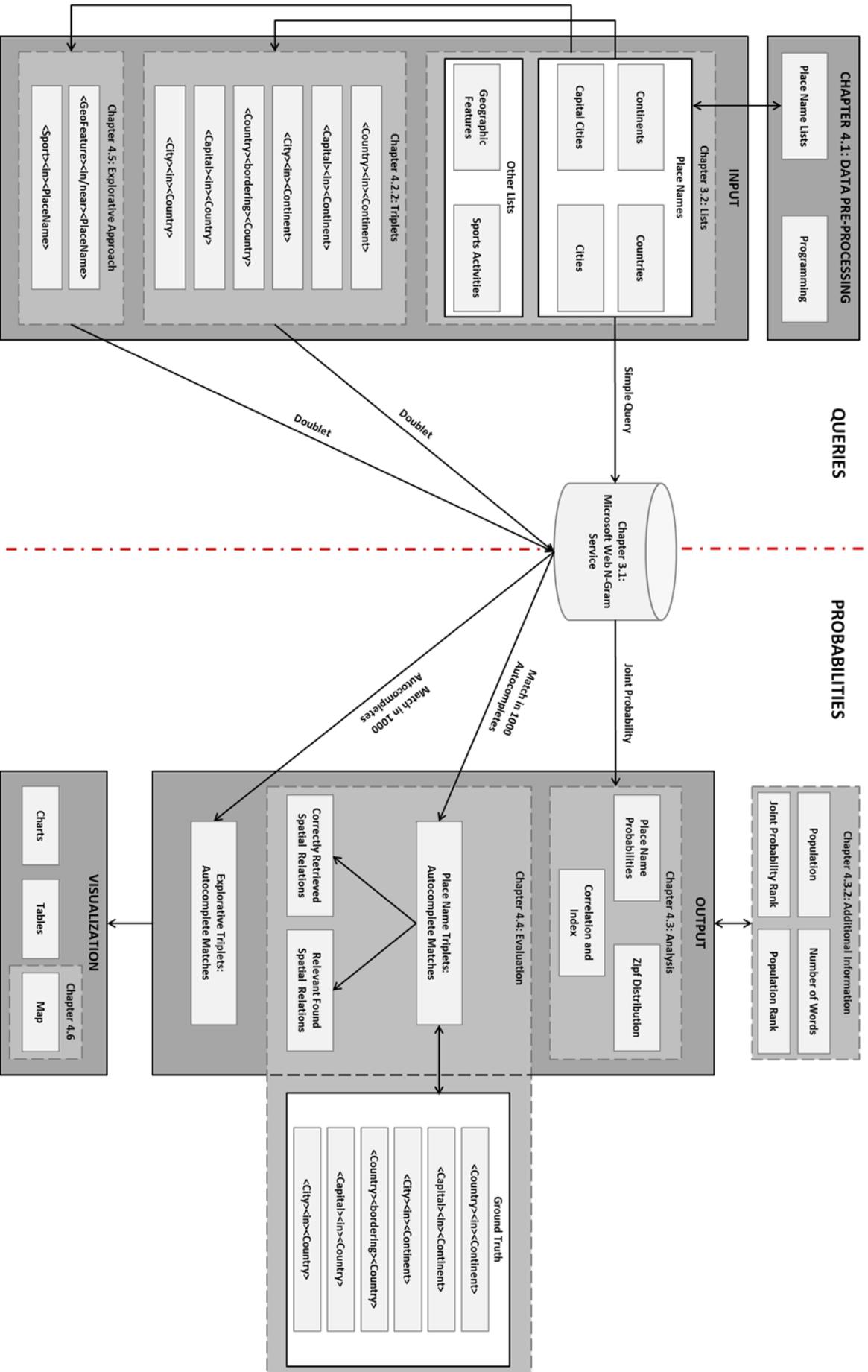


Fig. 4.1: Overview of data usage and methodological steps

## 4.1 Data Pre-Processing

The pre-processing of the data was essentially rewriting the provided code for accessing the Microsoft Web N-Gram Service API and allowing it to realize additional operations. Furthermore, the lists used as inputs had to be cleaned up and adapted to suit the applied code to access the API.

### 4.1.1 Programming

Microsoft Research (“Microsoft Web N-Gram Service Quick Start” 2014) offered a quick start guide for writing an application which uses their Web N-Gram Service. In the example given, the development platform for Web services was Visual Studio and the programming language was C#. However, with Java a more popular and already accustomed programming language was used for this thesis. The development platform used was NetBeans IDE. Therefore, the code provided by Microsoft Research had to be rewritten into Java. This was done with the help of Prof. Dr. Ross Purves and Dr. Curdin Derungs. The resulting code can be found in the appendix A.1.

In addition to this a user-specific token had to be obtained to make use of the Microsoft Web N-Gram Service. For that purpose a user token had to be requested from Microsoft.

With the prior code (appendix A.1) only the joint probability, the conditional probability and the autocompletes of a predefined term could be obtained. Thus, the code was adapted in a second step to allow using a list of terms to be searched for joint probability, conditional probability and autocompletes. In this case, the list of terms was a list of place names. Additionally, the resulting probabilities were directly written to a semicolon separated text file. The mentioned Java code can be looked up in the appendix A.2.

### 4.1.2 Place Name Lists

The place name lists mentioned in section 3.2 were used as inputs for the Microsoft Web N-Gram Service. However, the lists had to be adjusted and shortened. For the list of place names taken from Wikipedia (chapter 3.2.2 and chapter 3.2.3) the diacritics had to be removed. This had to be done, since the Microsoft Web N-Gram Service did not recognize these characters and returned insignificant joint probabilities. Therefore, the place name list was changed to ASCII format to ensure the place name characters could be recognized by the service.

For the list of cities in section 3.2.4, an ASCII format was provided by GeoNames. Hence, the diacritics in place names were already removed. The only pre-processing which had to be

done, was removing additional information of place names which were included in the place name itself. These were mostly started by a slash to refer to an alternative place name or parentheses and squared brackets to specify the corresponding country/district of a city. To remove this unnecessary information the place names were split into strings where the occurrence of slashes, parentheses and squared brackets occurred. Afterwards, the first string was taken and written to a text file. This made sure a list of cities was created without additional information contained in the place name itself. The code and list are available in the appendix A.2 and appendix B.3.

### 4.2 Queries

An input has to be commissioned to the Microsoft Web N-Gram Service to obtain its joint probability, conditional probability and autocompletes. In regard to this, the commissioned inputs are referred to as queries. These consist of two different methods, which are further discussed in the upcoming sections.

#### 4.2.1 Place Name Lists

The lists of place names were used as queries to obtain the joint probability for each place name. Thus, the joint probability for each continent, country, capital city and city was acquired. These queries had a simple construct, as they iterated through a list of place names to assign the joint probability for each place name. The received joint probabilities were simultaneously written to a semicolon separated text file with their corresponding place names. This allowed gaining first insights into the characteristics of place names in the Microsoft Web N-Gram Service.

#### 4.2.2 Triplets

The further analysis of place names on the Web was done by analyzing spatial relationships of place names, geographic features and sports activities to place names. These spatial relationships were represented as triplets. Therefore, the triplets had the following structure (similar to Jones & Purves 2008; Perea-Ortega et al. 2009; Pu et al. 2009):

**<topic><spatial relationship><place name>**

The topic in this scheme is either represented as a place name, a geographic feature or a sport activity. Furthermore, each single topic is represented in a list of equal categories. This is followed by a spatial relationship which is either a preposition (i.e. in or near) or a word

which signifies a spatial relation (i.e. bordering). At the end of the triplet comes a place name, which is obtained from the list of continents, countries, capital cities or cities.

For these triplets a new Java code is used which can be looked up in the appendix A.3. The code is built to compare the doublet <topic><spatial relationship> followed by the first 1000 autocompletes with the list of place names. If a match is found the conditional probability of the doublet <topic><spatial relationship> is assigned to a text file. The procedure is similar for the place names which contain more than one word. These are compared with the first 1000 autocompletes and if the first word of a place name occurs in the autocompletes the next 1000 autocompletes of the doublet and the first word are searched. For each step the conditional probability is written to the text file and separated by the sign “/”. Thus, if a place name would consist of two words, the code would return two conditional probabilities and for three words it would return three conditional probabilities. The place names containing more than three words were ignored, since the Microsoft Web N-Gram Service only allows searching for five words and the triplet contains at minimum two words at the beginning with <topic><spatial relationship>.

Furthermore, the list of place names at the end of the triplet need to contain an ID. This ID is also written to the same text file, to identify which place name has a match with the preceding doublet. The place name IDs are separated by “&”, when multiple place name matches occur. More importantly, the list of countries with an ID was slightly adjusted by changing the country name of “United States” to “the United States”. This was done, because the triplets were very sensitive in the case of the United States. In the English language the United States normally has an article at the beginning. As example the words “Texas in the United States” would occur more on the Web than “Texas in United States”. This is supported by comparing their corresponding probabilities with each other. For the first word gathering the joint probability was 0.0000005420%, while the second word gathering has a joint probability of 0.0000000438%.

At the end, the information is saved in a semicolon separated text file where the first column is the topic, the second column is the matching place name ID and the third column contains the conditional probability.

The lists and received results were adjusted in a second step. The list of cities used as topic was shortened to contain only words as long as three words. Additionally, the triplet of the correct spatial relationships of cities (listed in chapter 4.4.1) containing more than five words were removed from the list of cities. Hence, the correct triplet of a city could contain more than six words e.g. “New York City in the United States”. In that case, the city “New York

City” was removed from the list used as input. For the place name list of countries and capital cities this step was done after obtaining the results of the triplets. This was easier to do afterwards since, the number of entries for countries or capital cities were significantly smaller. Meaning, the countries/capital cities containing more than three words and countries/capital cities whose correct triplet contained more than five words were removed.

### 4.3 Analysis

The analysis of the received semicolon separated text files was done in Microsoft Excel, while further information was added to the received joint probabilities of place names. In this section, the procedure used for the statistics of the joint probabilities, the added information, the distribution of the joint probabilities in relation to the Zipf distribution and the correlation between population and joint probability are discussed.

#### 4.3.1 Place Name Joint Probability

The place name joint probabilities were imported into Microsoft Excel. The logarithmic joint probabilities were then converted to joint probabilities in percentage. The place names consisting of more than five words were also removed before analyzing the data. After that, the total number of place names, minimum joint probability, maximum joint probability, mean joint probability, standard deviation joint probability and the range of the joint probability was calculated. This was done for every list of place names mentioned in section 3.2.

Additionally, the joint probabilities of the place names were displayed in bar charts ranging from the smallest to the biggest joint probability. The bar chart of the city joint probability was slightly adjusted, as the data was ambiguous and was extremely scattered. This led to a distorted bar chart and therefore a cut off at the joint probabilities over 0.01% was performed. This concerned 232 city names which had a higher joint probability as displayed on the bar chart. Furthermore, for each place name list the top five place names joint probabilities and bottom five place name joint probabilities were displayed in a table. Place names occurring more than once were ignored in this table. They were only mentioned by listing the number of occurrence of a place name in brackets. Overall, the table helped to improve the comprehension of the place name distribution in the bar charts.

### 4.3.2 Additional Information

The Excel table containing the place name joint probabilities was used to add further information to each place name. As mentioned in the previous chapters the joint probabilities received from the Microsoft Web N-Gram Service were in a logarithmic scale. Therefore, these values were converted to a linear scale in percentage. The values were calculated in percentage to make the values more readable, as some of the values were considerably small and were smaller than  $10^{-10}$ .

In a second step, the population and area for each place name was added to the Excel table. The area was only added to the list of country joint probabilities, while the population was added to the joint probabilities of countries, capital cities and cities. Out of these area values and population data, the inhabitants per square kilometer were calculated for each country. Further information of the origin of the values can be taken from the chapter data in section 3.2.

In the third step, the number of words constituting a place name was calculated. This was done in Excel by counting the number of characters and subtracting the number of characters excluding word separators. A word separator in this case included all the characters the Microsoft Web N-Gram Service recognized as word divider. The word dividers used to distinguish the number of words are the following (an example of the mentioned character is given in the brackets):

- space ( )
- hyphen (-)
- apostrophe (')

Subsequently, the place names longer than five words were removed in the Excel file with the place name lists and joint probabilities.

### 4.3.3 Zipf Distribution

The distribution of the place name joint probabilities was compared to the Zipf distribution. The Zipf distribution is an empirical observation on certain statistical regularities commonly used in quantitative statistics (Montemurro 2001). It states that the word frequency in English is an inverse power law with the exponent close to one, if the words are aligned according to their ranks (Li 1992; Zipf 1932). These ranks are defined by the frequency of words occurring in a corpus, while the rank one is assigned to the most frequent word. Equally, the other words are assigned a rank which corresponds to its frequency. Given these facts the Zipf distribution can be defined as the following function:

$$p(r) = \frac{C}{r^\alpha}$$

(Auerbach 1913; Zipf 1932)

The variable  $p$  in this function represents the estimated frequency of a word in a corpus, while  $C$  and  $\alpha$  are constants (Montemurro 2001; Newman 2004). Additionally, the variable  $r$  stands for the corresponding rank of a word. The special case of the function, where  $\alpha = 1$  and  $C$  equals the frequency of the most frequent word, is known as Zipf's Law (Li 1992; Zipf 1949).

For the comparison with the place name joint probabilities each place name had a rank assigned to them based on their joint probability. The most probable place name was assigned the rank one, while the proceeding ranks were assigned to lower place name joint probabilities. Secondly, Zipf's Law was used for plotting the Zipf distribution. The variable  $p$  was represented as the estimated place name joint probability, while  $r$  represented the rank of a place name. Thus, the frequency of a place name was replaced by the place name joint probability in the Microsoft Web N-Gram Service corpus. The constant  $C$  was the most probable place name in the list, while  $\alpha$  was kept at the value one. Finally, the distribution of place name joint probabilities by rank was compared with the estimated joint probability calculated with Zipf's Law. This was done by plotting both functions in Microsoft Excel and comparing them with one another.

Additionally, the number of words constituting a place name was analyzed. This was done by exporting the information containing the number of words per place name (calculated in section 4.3.2) for each place name list into SPSS. Furthermore, the joint probability of the corresponding place names was imported into SPSS. In a second step, the place name joint probabilities by word count were plotted as a boxplot. The total number of place names with a certain number of words was also displayed. Finally, this gave some insights into the distribution of place name joint probabilities based on their word count.

#### 4.3.4 Correlation and Index

Possible correlations between place name joint probabilities and their corresponding population were tested. For that purpose the joint probabilities and the number of inhabitants per place name were ranked. However, the entries which contained no information to the total population were removed beforehand. The distribution of the values was shown with the help of a scatterplot in Microsoft Excel, where the joint probability rank was a function of the

population rank. Hence, this equals a rank correlation by Spearman. Additionally, a linear trend line and the coefficient of determination ( $R^2$ ) were displayed to show a possible trend and the strength of the correlation. This coefficient of determination equals the square of the Pearson product moment correlation coefficient. Lastly, the coefficient of determination helps to quantify the proportion of explained variance to the total variance and how much of the statistical dispersion of a variable  $x$  is explained by a linear dependence of variable  $y$ . The scatterplot of joint probabilities and population showed each value as rank. Therefore, the highest value is located at rank one. This means the  $y$ -axis ranges from highest (rank one) value to the lowest value (last rank) and the same goes for the  $x$ -axis, where the highest value is situated at the left and decreases to the right. For the further analysis of a possible correlation the joint probability ranks and population ranks were exported into SPSS. In principle, we assumed that the continuity, the homoscedasticity and a linear relation of the variables were given. Furthermore, all the tested variables are in a metric scale (interval or ratio). A two-tailed bivariate correlation by Spearman was conducted, as the values were already transformed into ranks. The significance level in these tests was set at 1%. Finally, the resulting correlation coefficient and  $p$ -value gave insights if a correlation exists between the variables.

In ArcGIS a spatial autocorrelation was conducted for the global distribution of the countries and their joint probabilities. At first, all the islands belonging to a country were removed to make sure that only the primary land of a country was taken into account. The possible autocorrelation was tested in ArcGIS by using the Spatial Autocorrelation (Morans I) in the ArcToolbox with the contiguity of edges/corners for the conceptualization of spatial relationships. In this case, the countries that share a boundary, node or overlap influence the target country. Hence, all countries which share the specified geometry with another country are weighted as one for the specified country, while all other countries which do not share any geometry with the specified country are weighted as zero. The report of the spatial autocorrelation was returned as an HTML file containing the Moran's Index, the expected index, the variance, the  $z$ -score and the  $p$ -value. These values provide information if a spatial autocorrelation is existent.

The second step looked at the influence of city population on the coefficient of determination. This was done by calculating the coefficient of determination between joint probability rank and population rank for a selected range of city population. Therefore, the joint probability rank and population ranks had to be recalculated for the new range of city names. The ranges included the city names with a population over 1 000, 10 000, 100 000, 1 000 000, 10 000 000.

Finally, the results would give insight how the strength of the correlation between joint probability rank and population rank changes based on the selection of city names by population.

In the third and final step, indices were constructed for the city names. These should indicate possible ambiguous and overrepresented city names on the Web compared to the rank correlation of joint probability and population. The step was only conducted for city names, as the ambiguity in continent, country, and capital city names was manageable. Furthermore, the use of indices for locating ambiguous place names in the categories continent, country and capital city was counterproductive. It led to the exclusions of ambiguous place names, but also excluded popular places frequently presented on the Web. The index used for the city names was constructed with a ratio of joint probability rank over population rank, similar to Derungs & Purves (2014a). Therefore, the function of the ratio was as follows:

$$\frac{\textit{Joint Probability Rank}}{\textit{Population Rank}}$$

The ratio indicated if a place name was overrepresented on the Web compared to its population. A value of 1 indicated a perfect match between joint probability and population rank. Values bigger than 1 indicated an underrepresentation of a city name on the Web compare to its population, while values smaller than 1 indicated an overrepresentation of a city name on the Web compared to its population. This means the lower the index/ratio was, the more likely a city name was overrepresented on the Web. Thus, the city name was probably ambiguous. The ratio was further used for the explorative approach in chapter 4.5.

### 4.4 Evaluation

The autocompletes obtained for place name triplets with the help of Microsoft Web N-Gram Service and the Java code (appendix A.3) were assessed on their accuracy and completeness. For that purpose ground truth was needed. This section, explains how the ground truth was acquired and compared to the received autocompletes of the triplets. Lastly, the Java code used for the comparison of the ground truth with the received autocompletes for the triplets can be looked up in the appendix A.4.

#### 4.4.1 Ground Truth

The ground truth data of each place name triplet list was obtained with the help of Wikipedia ArcGIS and GeoNames. Additionally, some information was supplemented by hand. Each

ground truth list of a triplet in Microsoft Excel contained the place name and the hierarchical higher place name it belonged to as an ID in a column. These were then saved as semicolon separated text files, while the place names IDs were separated by “&”. The place name IDs were identical to the place name ID list mentioned in chapter 4.2.2. As an example Zurich would have Switzerland as a hierarchical higher place name and therefore it would be saved as “Zurich; 179” in the semicolon separated text file.

Wikipedia (“List of national capitals in alphabetical order” 2015; “List of sovereign states and dependent territories by continent” 2015) was used for the ground truth data of the triplets <country><in><continent>, <capital><in><continent> and <capital><in><country>. In some cases the countries were not clearly assigned to a continent. These cases were countries associated with either Europe or Asia. To resolve this predicament, the countries having a double association were assigned the concerning continents Asia and Europe. Furthermore, the continent Australia was set synonymous to Oceania. This means that all the islands countries which had no clear continent association and were referred as Oceania, were assigned to the continent Australia. The ground truth of the triplet <capital><in><continent> was derived from the other triplets <country><in><continent> and <capital><in><country>. The same procedure, mentioned further above, was used for capitals lying in a country with a double association.

The ground truth of the triplet <country><bordering><country> was derived with the help of ArcGIS. The map offered by ESRI and DeLorme (see chapter 3.3) was used to obtain the bordering countries of each country. For that purpose the function polygon neighbors in the ArcToolbox was used. The received database table was exported as text file and then imported into Microsoft Excel for further modifications. These mainly consisted of bringing the dataset into the same format as the other triplets. Hence, a column contained the country names and the proceeding columns contained the bordering country IDs. Each country ID was in a separate column. Furthermore, the ArcGIS country names had to be slightly adjusted to match the list of country names used in this thesis. In a second step, the bordering countries obtained through ArcGIS had to be converted into an ID to fulfill format specifications. Additionally, missing spatial relationships were adjusted by hand. This mainly concerned disputed countries and their neighboring countries.

The ground truth of the triplets <city><in><continent> and <city><in><country> was obtained with the help of GeoNames. The list of cities taken from GeoNames (“Download” 2014) already contained the information which city belonged to which country. Therefore, the only necessary step was assigning the country ID to the country abbreviations

in the GeoNames list. For the triplet <city><in><continent> the continents IDs were derived with the help of the triplet <city><in><country> and <country><in><continent>. Hence, cities have the possibilities to be referred to two continents if their country was associated with two continents.

At the end, the number of words for each triplet was counted. The triplets containing more than five words were removed from the list. This was to assure that the Microsoft Web N-Gram Service could obtain the correct triplets, since it was only possible to query for a maximum of five words.

### 4.4.2 Correctly Retrieved Spatial Relations

The correctly retrieved spatial relations represent the ratio of the correct spatial relations divided by the amount of total retrieved relations. The formula is expressed as follows:

$$\text{correctly retrieved spatial relations} = \frac{\text{correct spatial relations}}{\text{retrieved spatial relations}}$$

The correct spatial relations refer to the correctly received place names from the 1000 autocompletes for a doublet <place name><spatial relationship> compared to the ground truth. The retrieved spatial relations are all the correct and false place names received from the 1000 autocompletes for the doublet. This information is contained in the ground truth. This ratio was calculated for every entry, except for the entries which received no place name ID from the autocompletes. For the triplet <country><bordering><country> a further exception was done. This exception consisted of given islands which had no possible neighbor the ratio one. Therefore, the countries which had no possible neighbor and did not have any autocomplete place name match were 100% correct. The remaining triplet entries were compared to the ground truth entries. Therefore, both lists needed the same length and be ordered identically. This was achieved by ordering the preceding place name alphabetically and comparing the place name IDs with each other. The calculation of the ratio for every entry was done with the Java code in appendix A.4. The results were saved to a semicolon separated text file containing the correctly retrieved spatial relations and the relevant found spatial relations (in section 4.4.3). In the last step, the text file was imported into Microsoft Excel and the ratios were converted into percentages. Additionally, the mean ratio of the correctly retrieved spatial relations was calculated. The mean helped indicating how many of the obtained spatial relations were relevant.

### 4.4.3 Relevant Found Spatial Relations

The relevant found spatial relations represent the ratio of the correct spatial relations divided by the amount of total relevant spatial relations. To clarify, the formula was expressed as follows:

$$\text{relevant found spatial relations} = \frac{\text{correct spatial relations}}{\text{relevant spatial relations}}$$

The correct spatial relations refer to the correctly received place names from the 1000 autocompletes for a doublet <place name><spatial relationship> compared to the ground truth. The relevant spatial relations signify the place names which should finish the doublet. This information is contained in the ground truth. The ratio was calculated for every entry, while the amount of relevant spatial relations represents the number of relations listed in the ground truth. Yet again, the triplet <country><bordering><country> was an exception. The countries which had no neighboring country and had no assigned place name ID from the autocompletes were given the value one. The remaining triplets were calculated with the help of the defined ratio. These steps were done with the same Java code in section 4.4.2 and the resulting ratios were also saved to a semicolon separated text file. The resulting text file contained the first place name of the triplet, correctly retrieved spatial relations ratio and the relevant found spatial relations ratio. These were then imported into Microsoft Excel and the ratios were converted to percentages, while the mean value of correctly retrieved spatial relations and the relevant found spatial relations was calculated. The means were displayed in a table for comparison. All in all, the mean of the relevant found spatial relations indicated how many of the relevant spatial relations were found.

## 4.5 Explorative Approach

The explorative approach focused on assigning locations to geographic features and sports activities. This was done with the help of triplets (see chapter 4.2.2). These triplets had the following structure (Jones & Purves 2008; Perea-Ortega et al. 2009; Pu et al. 2009):

**<topic><spatial relationship><place name>**

The topic was either a geographic feature (list in 3.2.5) or a sport activity (list in 3.2.6), while “in” and “near” were used as spatial relationships. For the place names, only the list of countries and cities were used. Furthermore, these lists were shortened to investigate a more local area. The inspected area was chosen based on the first results in sections 5.1 and 5.3.

#### 4 | Methodology

Europe was one of the continents with numerous countries containing a high joint probability, while the spatial relationships – country bordering country and city in country – contained high percentages of correctly retrieved spatial relationships. Hence, most place names seemed to be well represented on the Web and the retrieved spatial relationships in Europe were mostly correct. Accordingly, the investigation was done with the country and city names in Europe. These lists were further reduced, as the Microsoft Web N-Gram Service only allowed a maximum of five words to be queried. Consequently, only the place names made up of three or less words were taken into the list. Moreover, the country United Kingdom was split into: England, Northern Ireland, Scotland and Wales. This was mainly done, since geographic features and especially sports activities are more likely associated to the countries comprising the United Kingdom. As an example, the football teams participate as England, Northern Ireland, Scotland and Wales in competitions. They do not participate as the United Kingdom. These countries mainly participate as the United Kingdom in other competitions such as the Olympic Games. Additionally, an article was added to country names where it seemed appropriate (e.g. “Czech Republic” was changed to “the Czech Republic”). This was done to improve the results. Finally, the total of 55 European Countries were assigned a new ID. The ID was used to identify which country was found in the 1000 autocompletes following the doublet <topic><spatial relationship>.

The list of European city names was also shortened and possible ambiguous city names were removed. This was done with the help of the introduced index in chapter 4.3.4. Therefore, the joint probability rank and population rank was recalculated for all European cities containing three or less words. Then the index, represented as the ratio of joint probability rank over population rank, was calculated for each city (Derungs & Purves 2014a). Index values lower than one, were overrepresented on the Web and were possible ambiguous city names. The city names, where the population rank was at least twice as big as their corresponding joint probability rank, were discarded. Hence, the city names which were twice as small as their expected value (population rank) were overrepresented and removed. This means that all entries with an index value smaller than 0.5 were removed from the list. In the second step, all duplicate city names were erased. The final list contained 48554 city names, while each city name contained an ID.

The doublets <topic><spatial relationship> were then commissioned to the Microsoft Web N-Gram Service and the autocompletes were investigated on possible matches with the place name lists. The spatial relationship varied for the doublets of geographic features followed by country. For the most part, the spatial relation “in” was used. The spatial relation “near” was

used for geographic features such as sea and ocean. The doublets of geographic features followed by city names only used the spatial relationship “near”. The spatial relation “in” was used for all the sport activities doublets.

The obtained result were then imported into Microsoft Excel and plotted as bar charts. These illustrated the most likely place names (country or city) to follow a geographic feature or sport activity. Furthermore, only the first 20 countries or cities with the highest conditional probability were displayed in the bar chart. The reduction of place names was done to guarantee the readability, as the inclusion of more entries made the corresponding place names in the bar chart unreadable. In addition, the mean of all conditional probabilities per geographic feature or sport activity was displayed in the bar chart. This helped to identify which place names were proportionally more probable to follow a specific topic compared to other place names. A collection of the obtained bar charts are discussed in chapter 5.4, while the most interesting or suspicious results are verified by querying the triplet directly on Bing. Each query is conducted after deleting the browser data in the operated Web browser. This should guarantee that the Web query results are not influenced by previous searches or preferences. Finally, a screenshot is made of the first results returned by the Web search engine. This helps to analyze the possible cause of a high conditional probability.

In the last step, the results of geographic features and sports activities were visualized per country on separate maps. For that purpose, the most likely geographic feature/sport activity to precede a country had to be identified. This would imply that the entry with highest conditional probability per country would be the most likely geographic feature/sport activity to precede the country. However, some geographic features/sports activities and place names might be overrepresented on the Web compared to others. Therefore, the *tf-idf* (introduced in chapter 2.1) was used to normalize the conditional probabilities. The *tf-idf* made sure that frequent words did not influence the outcome of the most likely geographic feature/sport activity to follow a country. The  $idf_t$  was calculated with the following formula:

$$idf_t = \log \frac{N}{df_t}$$

(Manning et al. 2008; Robertson 2004)

In this formula  $N$  was set at one hundred billion, since the total number of webpages in the Microsoft Web N-Gram Service is at the order of hundreds of billions. The document frequency of the term  $df_t$  was the joint probability of the doublet <topic><spatial relationship>. This step was also done with the joint probability of country

names. The resulting  $idf_t$  values were then multiplied with the corresponding conditional probabilities to obtain a value similar to the  $tf-idf$ . As an example, the  $idf_t$  values of Germany and “river in” were multiplied with the conditional probability that “river in” is followed by Germany. This made sure that frequent doublet “river in” was normalized and had less impact compared to the less frequent geographic features. The resulting  $tf-idf$  values helped to identify which geographic feature/sport activity was most likely to precede a country. Finally, these results were visualized in a map.

### 4.6 Map Visualization

The visualization of the data was done with the help of ArcGIS. The source of the provided maps can be looked up in chapter 3.3. At first, the map dataset had to be joined with the data tables to visualize the results per country. For this purpose the country names in the text files with the of the join probabilities, correctly retrieved spatial relations, relevant found spatial relations, geographic features in country and sports activities in country had to be slightly adjusted. This made sure that every country name with a geometry had a match when joining the tables with each other. Thus, an inner join was chosen for the tables and all entries without a match were removed.

The country name joint probabilities were displayed as five classes which were chosen through natural breaks (Jenks) in ArcGIS. The sequential single hue color scheme was adjusted and taken from ColorBrewer 2.0 (Brewer 2013). The color scheme ranged from light red to dark red.

Five classes with equal intervals were chosen for the correctly retrieved and relevant found spatial relations, while a sequential single hue color scheme was chosen from ColorBrewer 2.0 (Brewer 2013). This color scheme ranged from light green to dark green. Each of the classes contained a range of 20%, since the method for the classes was equal intervals. Hence, this made it easier to identify the countries with a high percentage of correctly retrieved/relevant found spatial relations (dark green) and a low percentage of correctly retrieved/relevant found spatial relations (light green). An exception was made for the correctly retrieved spatial relations: country bordering country. The dataset only contained three distinct values and therefore only three classes were chosen. These had the same green color scheme as the other maps. Additionally, the country city density was displayed on a map for the spatial relation city in country. For this map, five classes were chosen through natural breaks and a sequential single hue color scheme ranging from light red to dark red was applied. The countries Vatican City, Monaco, Nauru, Tuvalu, Malta, Marshall Islands and San

Marino were excluded from the classes. This was done due to the fact that, they distorted the classes with their high city densities.

The maps created in the explorative approach were based on the *tf-idf* values calculated in chapter 4.5. The main focus was on European countries. Therefore, the European countries were extracted from the world map and the remaining countries were grayed out. A slight adjustment was done for the map top sport activity to precede European country. The adjustment was splitting up the geometry of the United Kingdom into four areas: England, Northern Ireland, Scotland and Wales. Afterwards, the highest *tf-idf* value and their corresponding class in the category geographic feature and sport activity were displayed for each European country. The classes in each category were assigned a qualitative color scheme with the help of the colors schemes in ColorBrewer 2.0 (Brewer 2013).

At last, the layout of the maps was finalized in ArcGIS. This consisted of adding a title, legend, scale bar, imprint and sources. The maps were then exported as JPEG image files.



## 5 Results

*In this section, the obtained results are presented and visualized. This is done by following a structured procedure and looking at different levels of granularity. At first the occurrence, distribution and representation of place names in the Microsoft Web N-Gram Service is examined. The second step focuses on correlations between the joint probability which could explain the frequency of place names on the Web. The third step returns a measure for the reliability of spatial relations represented in the Microsoft Web N-Gram Service. Finally, the results of an explorative approach with emphasize on relations between geographic features/sports activities and place names are presented.*

### 5.1 Distribution and Statistics of Place Name Joint Probabilities

The joint probabilities of a place name list and its distribution is investigated in this chapter. This is followed by a table displaying the top five and bottom five joint probabilities of the place name list, while additional statistics of the joint probabilities of the list are shown. Furthermore, the joint probability distribution of the place name list is compared to an adjusted Zipf distribution. Each of these steps is conducted for different levels of granularity: continent, country, capital city and city. Finally, this will give insights into the characteristics of place names on the Web.

#### 5.1.1 Continent Name Joint Probability

The distribution of continent names in the Microsoft Web N-Gram Service is displayed in the bar chart in figure 5.1. It can be observed that Australia is the most probable continent name followed by Africa, Europe and Asia. The joint probability of the continents names North America, South America and Antarctica are considerably smaller than the previously mentioned continent names.

Table 5.1 can be consulted for a closer of continent name joint probabilities. The joint probabilities are shown in percentage up to ten decimals. Additionally, the continent names are ordered from highest to smallest joint probability. The table 5.2 demonstrates different cases of ambiguity for the place name America. It also displays the joint probability in percentage ranging from highest to smallest. It can be observed that the place name United States, the abbreviation USA and America are more frequently represented on the Web compared to North America, U.S.A, United States of America and the United States of America. The latter have considerably smaller joint probabilities and are consequently less frequently represented on the Web.

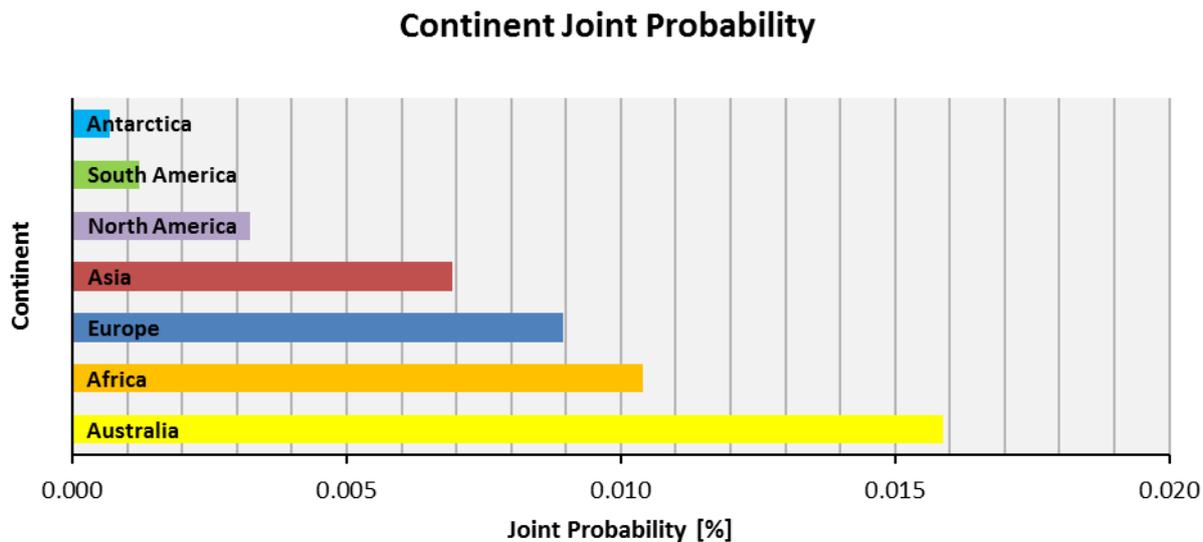


Fig. 5.1: Continent name joint probability on the Microsoft Web N-Gram Service in percentage

Table 5.1: Continent name joint probability in percentage

Continent Joint Probability		
Rank	Continent	Probability [%]
1	Australia	0.0158854675
2	Africa	0.0103992017
3	Europe	0.0089536477
4	Asia	0.0069183097
5	North America	0.0032359366
6	South America	0.0012105981
7	Antarctica	0.0006745280

Table 5.2: Joint probability of different variations for the place name America

Ambiguity: America Joint Probability		
Rank	Country	Probability [%]
1	United States	0.0227509743
2	USA	0.0185780446
3	America	0.0161435856
4	North America	0.0032359366
5	U.S.A.	0.0022080047
6	United States of America	0.0007079458
7	The United States of America	0.0002280342

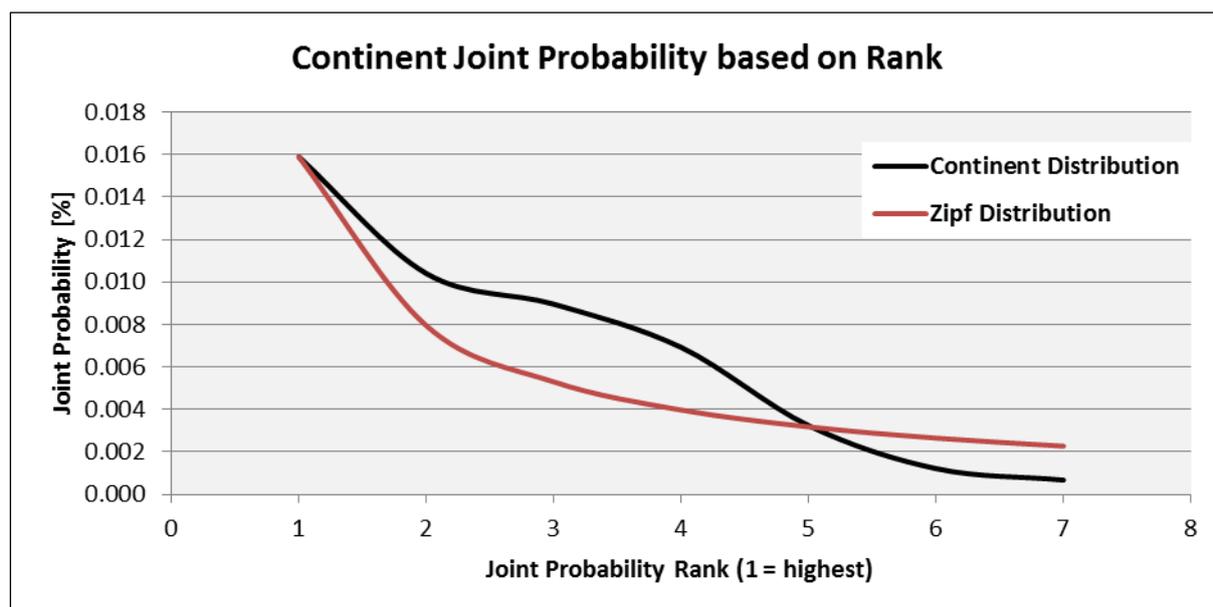
The overall statistics of the seven continent names is displayed in table 5.3. The minimum joint probability in the continent name list lies at 0.00067%, while the maximum joint probability lies at 0.01589%. Therefore, the range lies at 0.00895% and signifies a great

margin between minimum and maximum joint probability. The mean joint probability is at 0.00675%, while the standard deviation is at 0.00509%. This results to fairly scattered joint probabilities, since the standard deviation is almost as big as the mean value. The minimum value and the maximum value also have a large gap between them.

**Table 5.3: Statistical parameters of the continent name joint probabilities**

Statistics: Continent Joint Probability	
n Continents = 7	
Statistics	Probability [%]
Minimum	0.0006745280
Maximum	0.0158854675
Mean	0.0067539556
Std. Deviation	0.0050934509
Range	0.0089536477

The distribution of the continent name joint probability compared to the Zipf distribution can be seen in figure 5.2. Furthermore, the joint probability is plotted in dependency to the rank. This rank is ordered based on the joint probability and starts with the highest joint probability at rank one. The distribution of continent name joint probabilities slightly follows a Zipf distribution. However, the number of continent names is small and there is an offset from rank two to four.



**Fig. 5.2: Continent name distribution compared to an adjusted Zipf distribution**

In figure 5.3 the continent name joint probability distribution per number of words is shown. At the top of the box plots, the total number of a continent names constituted of a certain number of words is displayed. Additionally, the box plot helps to indicate the 25% percentile

(under line of the box), the mean joint probability (line in the box), the 75% percentile (upper line of the box), smallest not extreme joint probability (lower whisker), biggest not extreme joint probability (upper whisker), runaway values (circle: indicating the joint probabilities which are 1.5-3 times bigger than the box) and extremes (star: indicating the joint probabilities which are over 3 times bigger than the box size). Finally, the results show that there are more continent names made up of one word. These also have a higher mean joint probability, but also contain extreme values. One of these extreme values even has a smaller joint probability than all the continent names containing two words. This ratifies the results from the statistics which indicated highly scattered joint probabilities.

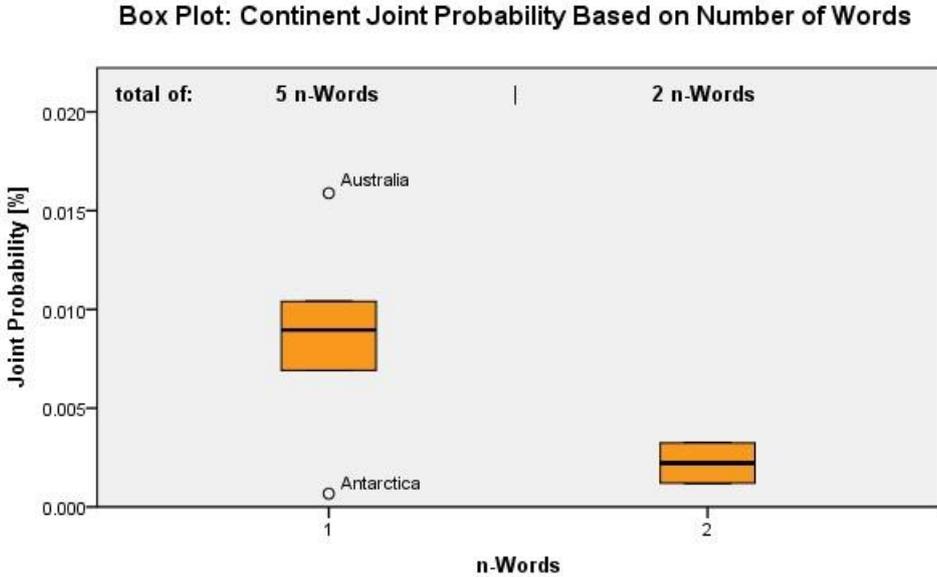


Fig. 5.3: Box plots of continent name joint probability based on the number of words

### 5.1.2 Country Name Joint Probability

The distribution of the country name joint probabilities can be observed in the bar chart in figure 5.4. The country name joint probabilities are further scattered than the continent name probabilities. Moreover, there are few country names with a high joint probability, while the amount of country names with a low joint probability is considerably bigger.

The table 5.4 and 5.5 show the top and bottom five country name joint probabilities. The most frequent country names on the Web are United States, China, Canada, India and Australia. On the other hand the most unlikely country names are Sahrawi Arab Democratic Republic, Transnistria, Nagorno-Karabakh, South Ossetia and Northern Cyprus. These are all countries where the sovereignty is disputed (“List of sovereign states” 2014).

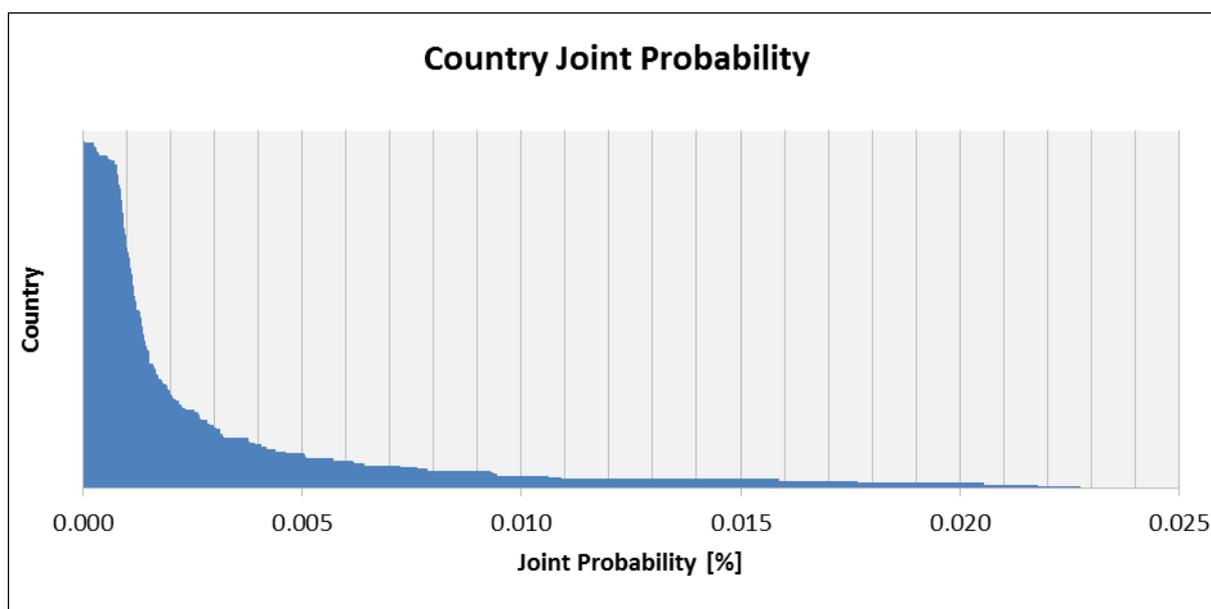


Fig. 5.4: Country name joint probability on the Microsoft Web N-Gram Service in percentage

Table 5.4: Top five country name joint probabilities in percentage

Top Five: Country Joint Probability		
Rank	Country	Probability [%]
1	United States	0.0227509743
2	China	0.0217770977
3	Canada	0.0205589060
4	India	0.0176603782
5	Australia	0.0158854675

Table 5.5: Bottom five country name joint probabilities in percentage

Bottom Five: Country Joint Probability		
Rank	Country	Probability [%]
1	Sahrawi Arab Democratic Republic	0.0000037154
2	Transnistria	0.0000106905
3	Nagorno-Karabakh	0.0000187932
4	South Ossetia	0.0000199526
5	Northern Cyprus	0.0000248313

In table 5.6 the statistics of the 206 country name joint probabilities are demonstrated. The minimum joint probability lies at approximately 0.0000037%, while the maximum joint probability is at 0.02275%. Consequently, the range has a value of about 0.0227447% and is almost as big as the maximum joint probability. This signifies a great margin between minimum and maximum joint probability. Hence, the country name joint probabilities are likely heavily scattered. To put this into perspective the country names and their joint

probabilities can be compared. The country name United States is about 5 600 times more frequently represented on the Web than the country name Sahrawi Arab Democratic Republic. The mean and standard deviation joint probability help to clarify the statement of heavy scatter. The mean joint probability is at 0.00233%, while the standard deviation is at 0.00334%. As a result the joint probabilities are heavily scattered, since the standard deviation exceeds the mean joint probability.

**Table 5.6: Statistical parameters of the country name joint probabilities**

<b>Statistics:</b>	
<b>Country Joint Probability</b>	
<b>n Countries = 206</b>	
<b>Statistics</b>	<b>Probability [%]</b>
Minimum	0.0000037154
Maximum	0.0227509743
Mean	0.0023257535
Std. Deviation	0.0033395510
Range	0.0227472590

The spatial distribution of these country name joint probabilities is revealed in figure 5.5. Generally, the first things to be observed are the low joint probabilities of country names in Africa. Secondly, the country names in North America, Australia and Europe seem to be clustered and have higher joint probabilities. In North America the country names with a high joint probability are United States, Canada and Mexico. The countries in Middle America have lower joint probabilities. In Europe the country names with a high joint probability persist of United Kingdom, France, Germany, Italy, Ireland and Spain. Besides these country names, there are other country names with have a high joint probability. These are more scattered and are as follows: China, India, Japan, Singapore, Brazil, South Africa, Georgia, Turkey, Jordan and Israel.

In figure 5.6 the distribution of the country name joint probability compared to the Zipf distribution is illustrated. The distribution curve of country name joint probabilities is similar to a Zipf distribution. However, the joint probabilities are situated slightly higher than the adjusted Zipf distribution. The country name joint probabilities curve also has more gradually jumps and is not as smooth as the Zipf distribution. Finally, at the end of the country name curve there is sudden drop to a lower joint probability.

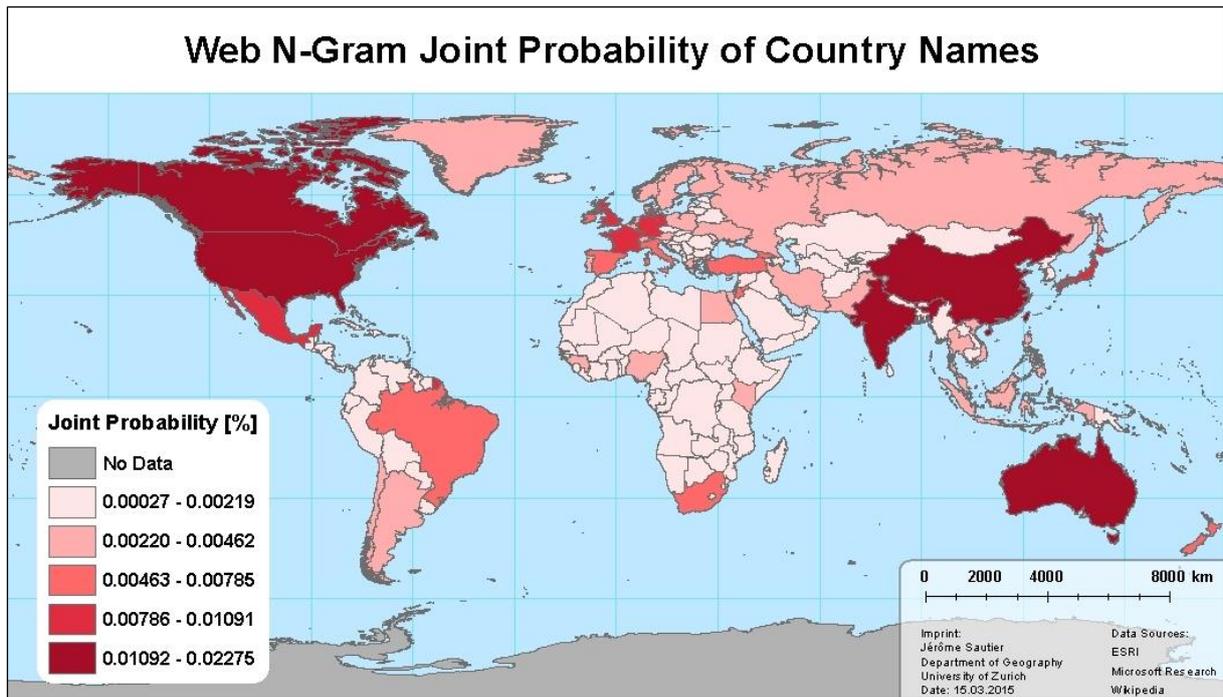


Fig. 5.5: Spatial distribution of country name joint probabilities on the Microsoft Web N-Gram Service

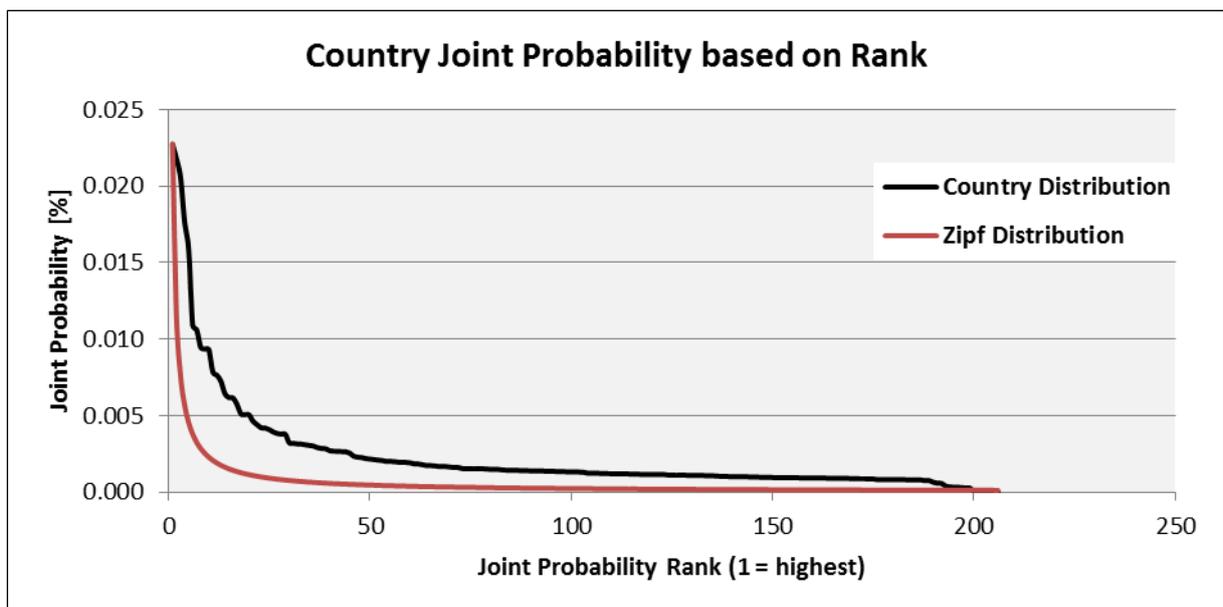


Fig. 5.6: Country name distribution compared to an adjusted Zipf distribution

The figure 5.7 helps to see the distribution of country name joint probabilities based on number of words. There are a total of 165 country names which contain one word, while there are 29 country names which contain two words. The remaining country names consisting of three, four and five words have a total count of six, four and two country names. Moreover, it can be observed that the mean joint probability decreases with the number of words making up a country name. The only exceptions are the country names made up of five words which have slightly higher joint probability than county names made up of four words. Additionally,

the total number of country names decreases in relation to the number of words composing a country name. The number of extreme and outlier joint probabilities also decreases with the number of words. This makes sense, since the number of country names also decrease with the number of words compromising a country name. Hence, there is less variation in country names with a higher word count.

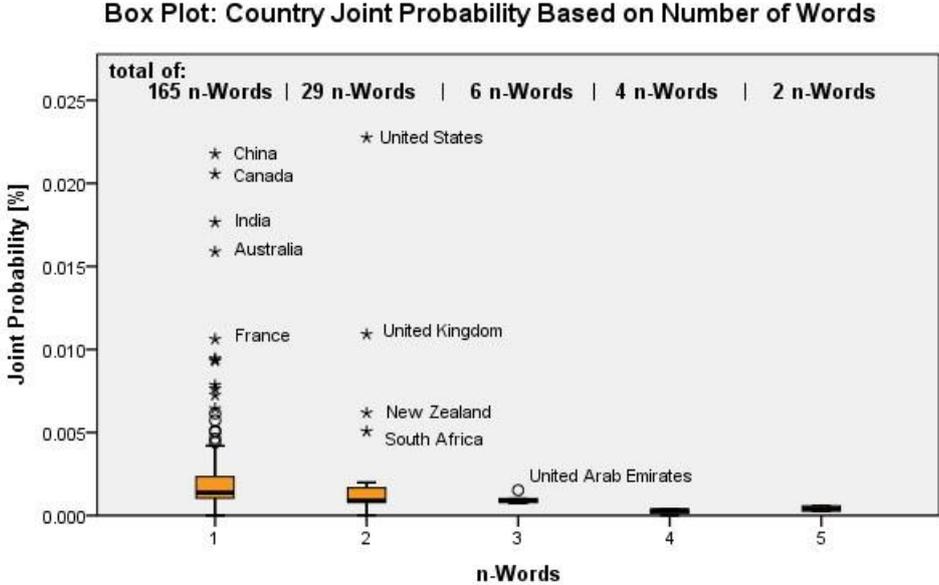


Fig. 5.7: Box plots of country name joint probability based on the number of words

### 5.1.3 Capital City Name Joint Probability

The distribution of the capital city name joint probabilities is displayed in the bar chart in figure 5.8. The capital city names have a very steep distribution with few high joint probabilities and many low high joint probabilities. Thus, the joint probabilities are heavily scattered compared to the continent and country name distributions.

The top and bottom five of the capital city name joint probabilities are shown in the table 5.7 and 5.8. The most recurrent capital city names on the Web are London, Washington, Male, Singapore and Victoria. These city names contains more severe cases of ambiguity compared to the continent and country names. In all these cases, either a geo/geo ambiguity, geo/non-geo ambiguity or both ambiguities are present. Examples of geo/geo ambiguity occurrences are London and Singapore. The city name London is also a city name in Canada and the United States, while the city Singapore is also a country. Examples of geo/non-geo ambiguity cases are Washington, Male and Victoria. The city name Washington can also refer to a former President of the United States, the city name Male also refers to a gender and the city name Victoria may refer to a brand (Victoria’s Secret) or a feminine first name. Anyhow, the

most unlikely capital city names on the Web are Ngerulmud, Sri Jayawardenepura Kotte, Tarawa Atoll, Lobamba and Palikir. These are mostly capital cities of islands states, except for Lobamba the capital city of Swaziland.

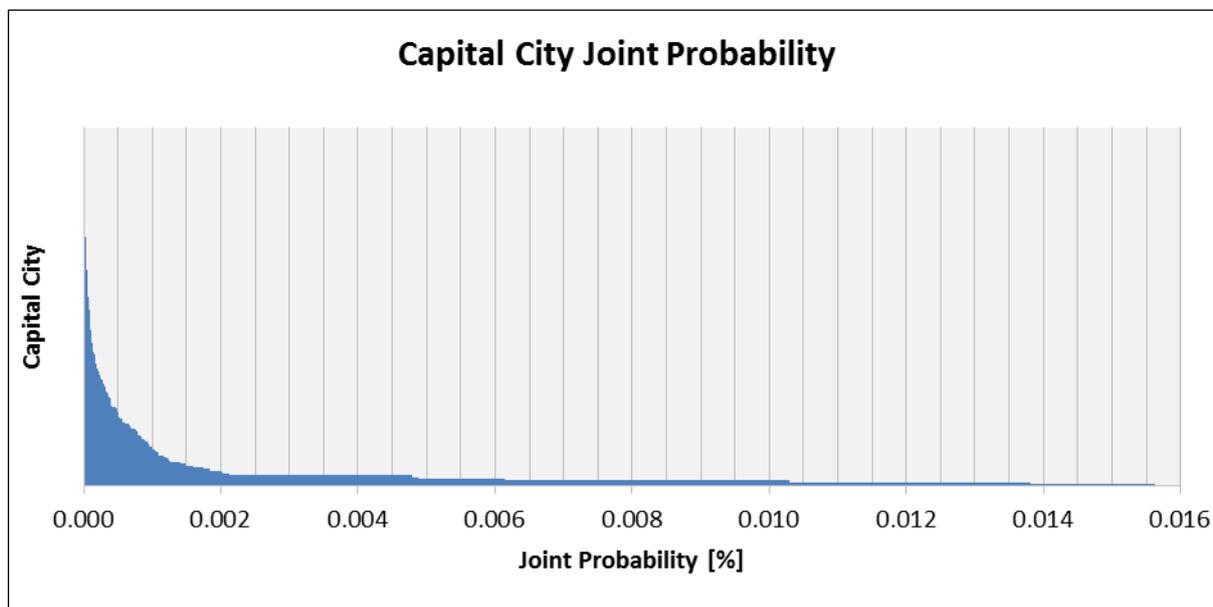


Fig. 5.8: Capital city name joint probability on the Microsoft Web N-Gram Service in percentage

Table 5.7: Top five capital city name joint probabilities in percentage

Top Five: Capital City Joint Probability		
Rank	Capital City	Probability [%]
1	London	0.0156314764
2	Washington	0.0138038426
3	Male	0.0103038612
4	Singapore	0.0061517687
5	Victoria	0.0048752849

Table 5.8: Bottom five capital city name joint probabilities in percentage

Bottom Five: Capital City Joint Probability		
Rank	Capital City	Probability [%]
1	Ngerulmud	0.0000001282
2	Sri Jayawardenepura Kotte	0.0000005070
3	Tarawa Atoll	0.0000007328
4	Lobamba	0.0000009268
5	Palikir	0.0000013964

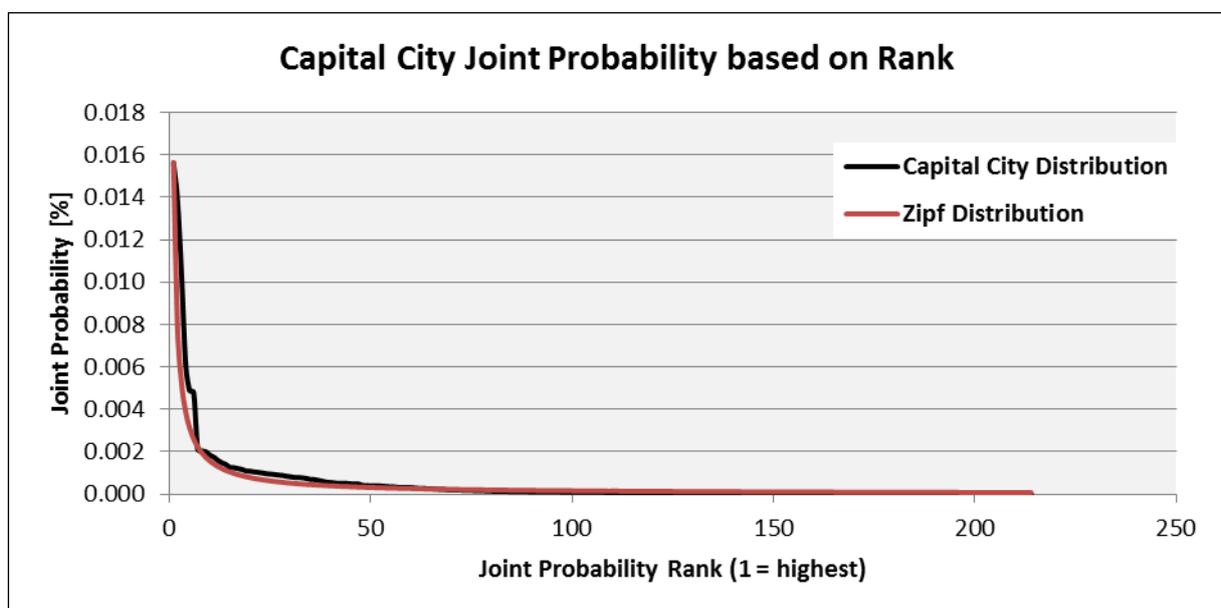
The statistics from the 214 capital city name joint probabilities are displayed in table 5.9. The minimum joint probability lies at approximately 0.00000013%, while the maximum joint probability is at 0.01563%. Consequently, the range has a value of about 0.01563135% and is

basically the same size as the maximum joint probability. This signifies a great margin between minimum and maximum joint probability. Hence, the capital city name joint probabilities are likely heavily scattered. An example can illustrate this statement. The capital city name London is about 130 000 times more frequently represented on the Web than the capital city name Ngerulmud. The mean and standard deviation joint probability help to clarify the statement of heavy scatter. The mean joint probability is 0.00052%, while the standard deviation is at 0.00170%. As a result the capital city joint probabilities are heavily scattered, since the standard deviation exceeds the mean joint probability.

**Table 5.9: Statistical parameters of the capital city name joint probabilities**

<b>Statistics: Capital City Joint Probability</b>	
<b>n Capital Cities = 214</b>	
<b>Statistics</b>	<b>Probability [%]</b>
Minimum	0.0000001282
Maximum	0.0156314764
Mean	0.0005187792
Std. Deviation	0.0017028104
Range	0.0156313482

In figure 5.9 the distribution of the capital city name joint probability is compared to the Zipf distribution. The distribution curve of the capital city name joint probabilities is similar to a Zipf distribution. However, the curve has some abrupt changes in the first 10-30 ranks and is therefore not as smooth as the Zipf distribution. This changes at about rank 50 where the capital city joint probabilities seem to follow the Zipf distribution.



**Fig. 5.9: Capital city name distribution compared to an adjusted Zipf distribution**

The distribution of capital city name joint probabilities based on number of words can be seen in figure 5.10. In total there are 180 capital city names consisting of one word, 27 capital city names made up of two words and seven capital city names containing three words. Furthermore, the mean joint probability seem to have the same size across all word counts making up a capital city. Similar to the country names, the total number of capital city names decreases in relation to the number of words composing a place name. The number of extreme and outlier joint probabilities also decreases with the number of words. Finally, the shorter capital city names have more variation and thus have numerous high joint probabilities. This leads to more outliers.

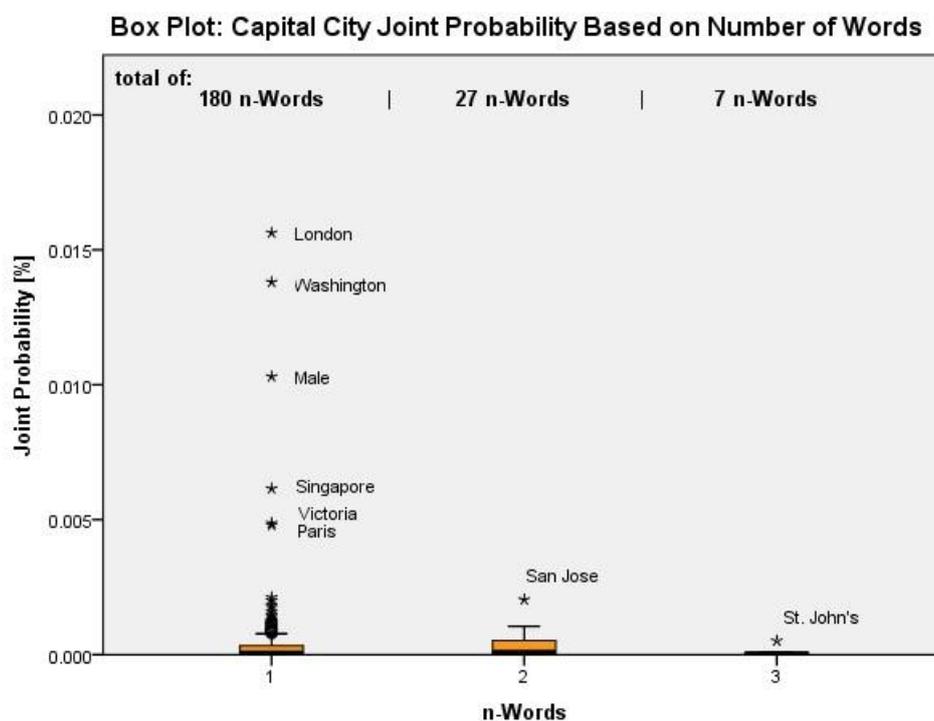
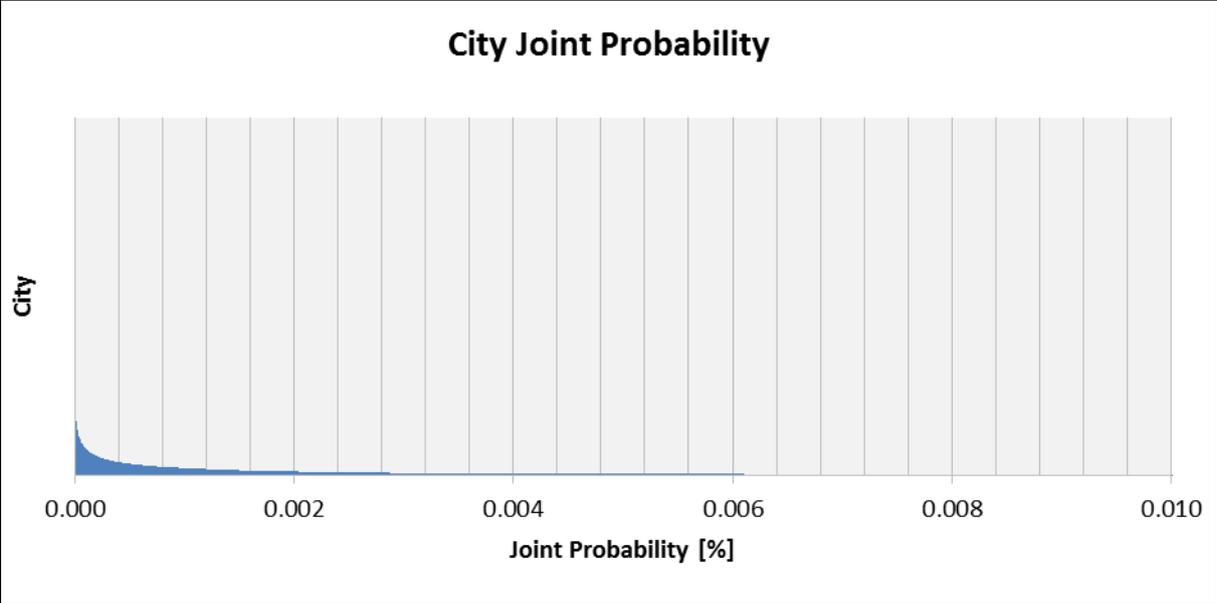


Fig. 5.10: Box plots of capital city name joint probability based on the number of words

#### 5.1.4 City Name Joint Probability

The distribution of city name joint probabilities is very steep and the probabilities are heavily scattered. This trend can be seen in the bar chart from figure 5.11. The bar chart had to be cut off at 0.01% to make the bar chart more visible, since the city name joint probabilities had huge statistical outliers which distorted the chart. Furthermore, it can be observed that there are a large number of city names with a low joint probability and a scare number of city names with a high joint probability. In total there were 232 city names which had a higher joint probability than 0.01% and had their bar chart cut off.



**Fig. 5.11: City name joint probability on the Web in percentage with a cut off at 0.01%**

The top and bottom five city name joint probabilities are illustrated in table 5.10 and 5.11. At first glance it can be detected that the most recurrent city names are common English words such as prepositions, verbs, adverbs and pronouns. Consequently, their joint probability is higher compared to other city names. The number of occurrences is also displayed in the top five city name joint probabilities table. Hence, the city name “As” occurs five times in the list of city names. The bottom five city name joint probabilities on the other hand, contain mostly foreign city names consisting of three or more words. Their joint probability is so small that it cannot be expressed in percentage with ten decimals points. Meaning, there are uncommon and barely occur on the Web.

**Table 5.10: Top five city name joint probabilities in percentage**

Top Five: City Joint Probability		
Rank	City (Number of Occurences)	Probability [%]
1	Of (1)	1.5559656316
2	Is (1)	0.6295061829
3	Are (1)	0.3111716337
4	As (5)	0.3097419299
5	We (1)	0.2142890601

**Table 5.11: Bottom five city name joint probabilities in percentage**

<b>Bottom Five: City Joint Probability</b>		
<b>Rank</b>	<b>City</b>	<b>Probability [%]</b>
1	San Juan Cote Ejido	0.0000000000
2	El Potrero de Sataya	0.0000000000
3	Reforma y Planada	0.0000000000
4	Ejido los Huastecos	0.0000000000
5	Thi Tran Viet Quang	0.0000000000

For the statistics of the 143 252 city name joint probabilities table 5.12 can be consulted. The minimum joint probability is too small to be expressed in percentage with ten decimal points, while the maximum joint probability is at 1.55597%. Hence, the range is about the same size as the maximum joint probability and there is major gap between minimum and maximum value. Moreover, the mean joint probability has value of 0.00014% and the standard deviation is at about 0.00515%. As a result, the large amount of city names and their corresponding joint probabilities are massively scattered.

**Table 5.12: Statistical parameters of the city name joint probabilities**

<b>Statistics: City Joint Probability</b>	
<b>n Cities = 143 252</b>	
<b>Statistics</b>	<b>Probability [%]</b>
Minimum	0.0000000000
Maximum	1.5559656316
Mean	0.0001425559
Std. Deviation	0.0051494061
Range	1.5559656316

However, the increase in place name counts and heavy scatter leads to a closer resemblance of the Zipf distribution. This is shown in the line chart (figure 5.12) comparing the city name probability distribution with the Zipf distribution. Both lines are basically aligned and therefore only the Zipf distribution (red line) is seen as it is on top of the city name distribution. Likewise, the distributions have a very quick drop with an asymptotic behavior converging to zero.

More insights into the characteristics of city names on the Web can be won by looking at the joint probabilities based on number of words. The box plot in figure 5.13 shows the distribution of city name joint probabilities based on number of words. The mean joint probability seem to have the same size across all word counts making up a city. Furthermore, the total number of capital city names decreases in relation to the number of words composing

a place name. The number of extreme and outlier joint probabilities also decreases with the number of words. Yet, the one worded city names have the strongest outliers, while the outliers in the two or more worded city names is moderate.

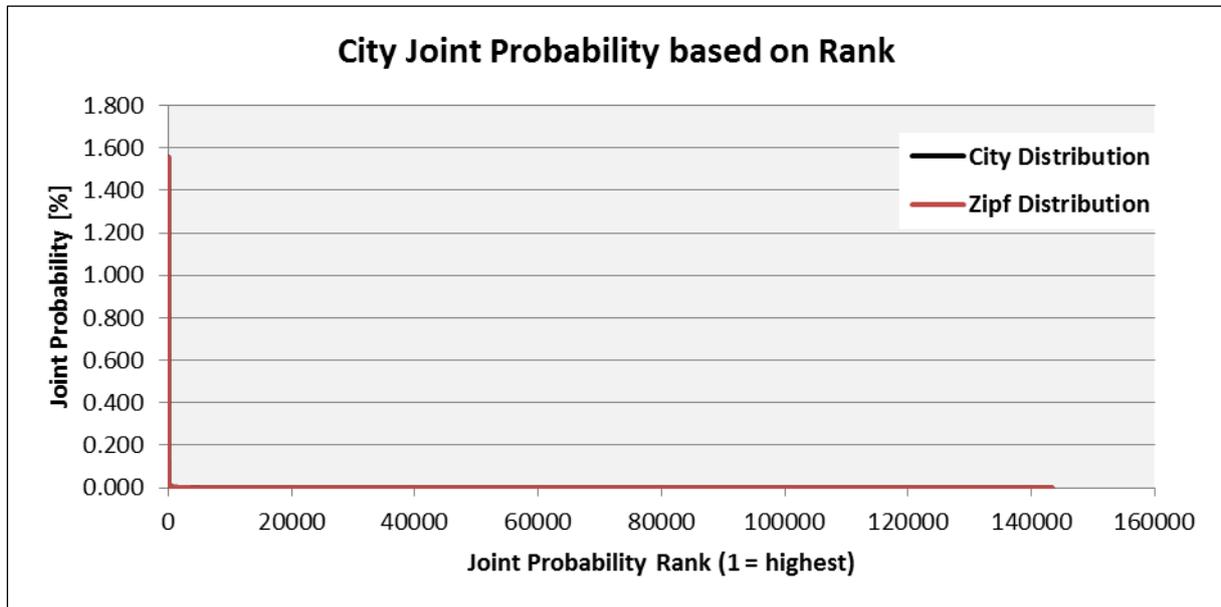


Fig. 5.12: City name distribution compared to an adjusted Zipf distribution

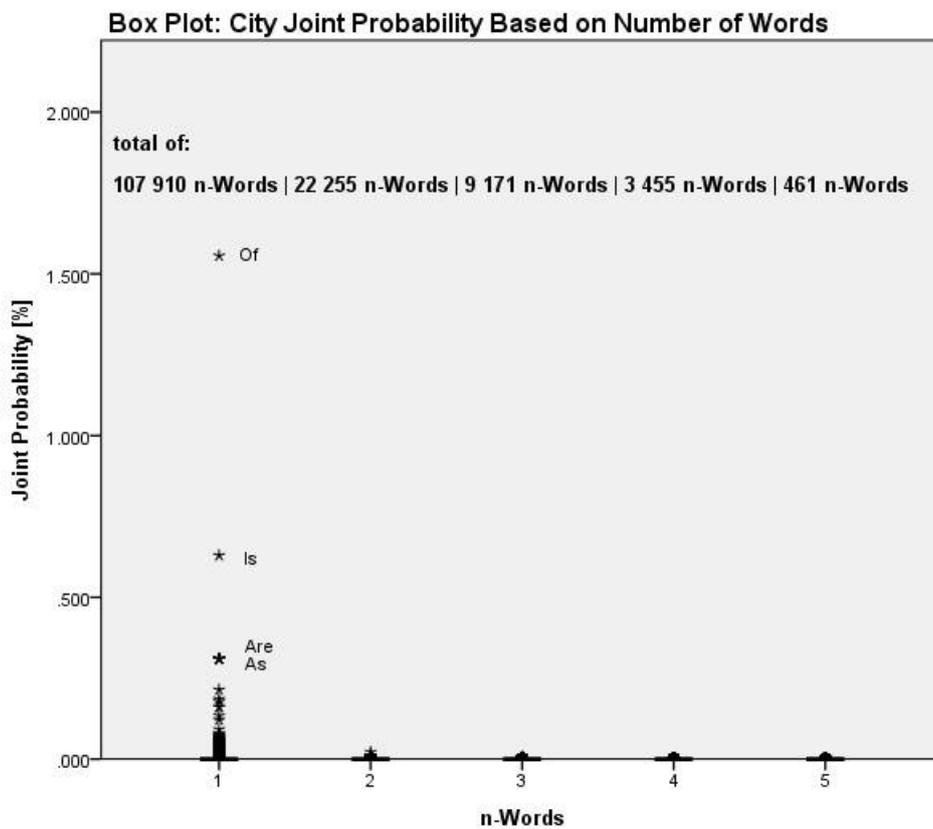


Fig. 5.13: Box plots of city name joint probability based on the number of words

## 5.2 Correlation of Place Name Joint Probabilities

The possible causes for high and low joint probabilities are investigated in this section. In other words, the place name joint probabilities are compared to their corresponding population. This helps to clarify if a high joint probability is created by a high population and vice versa. These steps are done for the country and capital names, as the number of continent names are insufficient and the city names contain several cases of ambiguity. However, the influence of population ranges on the coefficient of determination ( $R^2$ ) between city joint probability rank and city population rank is tested. For additional correlations of place names per continent, see the appendix C.1 and C.2.

### 5.2.1 Country Name Correlation

The data points of country joint probability rank and their corresponding population rank are dispersed in the scatterplot from figure 5.14. Nonetheless, they seem to have a slight upwards trend. This is confirmed by the slope from the linear trend line of the data. Hence, a high joint probability (low rank) likely goes with a high population (low rank). This positive correlation is somewhat visually observable. Additionally, the data points seem to be closer together at the beginning and end of the plot. This would assume that very low joint probability ranks have a very low population rank and vice versa. However, the data points in the middle of the plot seem to be more dispersed. Especially, the population ranks lying around 25 (high population) seem to have a contradictory high joint probability rank (low joint probability). Overall, the data points above the trend line are underrepresented on the Web compared to their population, while the data points under the trend line are overrepresented on the Web compared to their population. The coefficient of determination is moderate with a value of 0.3449. This means the red linear trend line accounts for 34.5% of the total variance between country joint probability rank and country population rank.

The possible correlation between country name joint probability rank and population rank was tested with a two-tailed bivariate correlation by Spearman. The tested null hypothesis is seen below.

**$H_0$**  : No correlation exists between country name joint probability rank and country population rank.

The correlation coefficient has a value of 0.587 and the significance value lies at 0.000 (as seen in table 5.13). The square of the correlation coefficient equals the coefficient of determination. The value of 0.587 signifies a positive correlation between the values of

country joint probability rank and country population rank. Likewise, the p-value (significance value) of 0.000 indicates a correlation, since it is smaller than the significance level of 1%. Therefore, the null hypothesis is rejected. This means the result is statistically very significant. A positive correlation between country name joint probability rank and country population rank is statistically present.

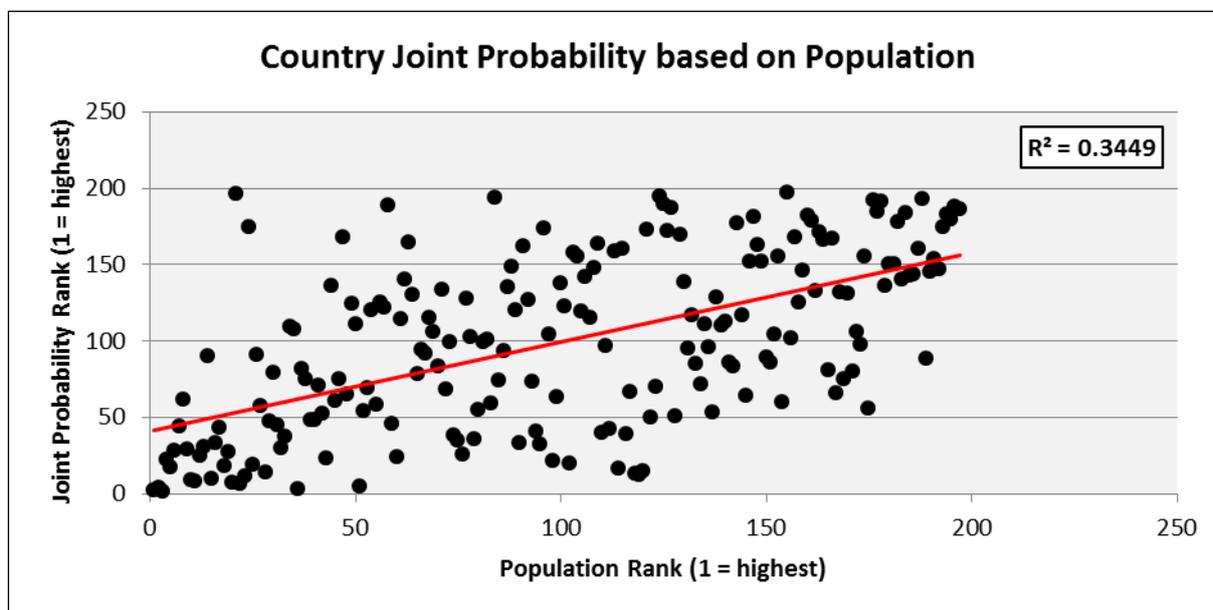


Fig. 5.14: Rank correlation between country name joint probability and population

Table 5.13: Spearman correlation coefficient of country joint probability rank and population rank

**Correlations: Country Name Joint Probability Rank and Population Rank**

		Joint Probability Rank	Population Rank
Spearman's rho	Joint Probability Rank	Correlation Coefficient	1.000
		Sig. (2-tailed)	.000
		N	197
	Population Rank	Correlation Coefficient	.587**
		Sig. (2-tailed)	.000
		N	197

\*\* . Correlation is significant at the 0.01 level (2-tailed).

The country name joint probabilities were also tested on a spatial autocorrelation. A detailed summary of the results and report of the spatial autocorrelation is available in the appendix C.3. The Moran's Index in the test was 0.289, while the z-score was at 4.897. These results indicate a positive spatial autocorrelation between country name joint probabilities. Given the z-score of 4.897, there is less than 1% (significance level) likelihood that the clustered pattern of country joint probabilities could be a result of random chance. Hence, similar country name joint probabilities are spatially clustered. This means country joint probabilities follow Tobler's first law of geography (Tobler 1970). Consequently, near countries have more similar/related joint probabilities than distant countries.

### 5.2.2 Capital City Name Correlation

The data points of capital city joint probability rank and their corresponding population rank are dispersed in the scatterplot from figure 5.15. It also seems the data points are more dispersed with higher population ranks (lower population) and are less dispersed with low population ranks (high population). The linear trend line shows a slight incline. However, this could also be caused by the scattering of the data. The capital cities with a high population (low rank) are overrepresented on the Web (high joint probability). This can be observed with the data points on the lower left lying under the red trend line. The coefficient of determination is low with a value of 0.2018. Meaning, the linear trend line only accounts for 20.2% of the total variance between capital city joint probability rank and capital city population rank.

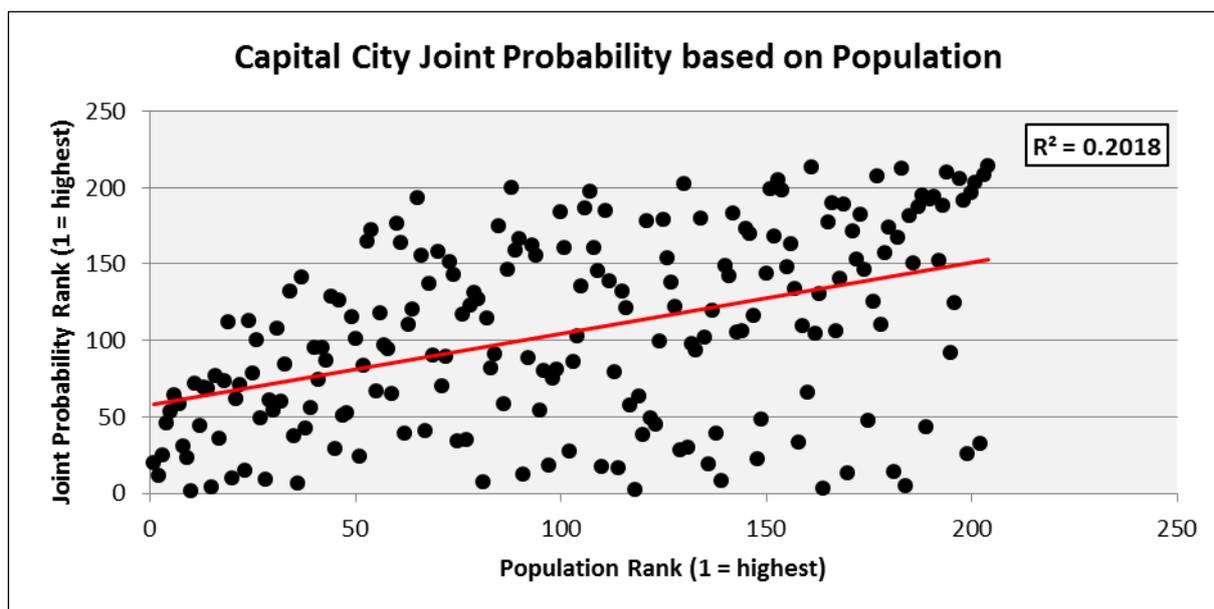


Fig. 5.15: Rank correlation between capital city name joint probability and population

The possible correlation between capital city name joint probability rank and population rank was tested with a two-tailed bivariate correlation by Spearman. The tested null hypothesis is seen below.

$H_0$  : No correlation exists between capital city name joint probability rank and capital city population rank.

The correlation coefficient has a value of 0.447 and the significance value lies at 0.000 (as seen in table 5.14). The correlation coefficient signifies a positive moderate correlation between the values of capital city joint probability rank and capital city population rank. The null hypothesis is rejected, since the p-value is smaller than the significance level of 1%. The

result is therefore statistically very significant (less than 1% likelihood that the correlation is a result of random chance) and a correlation between capital city name joint probability rank and capital city population rank is present.

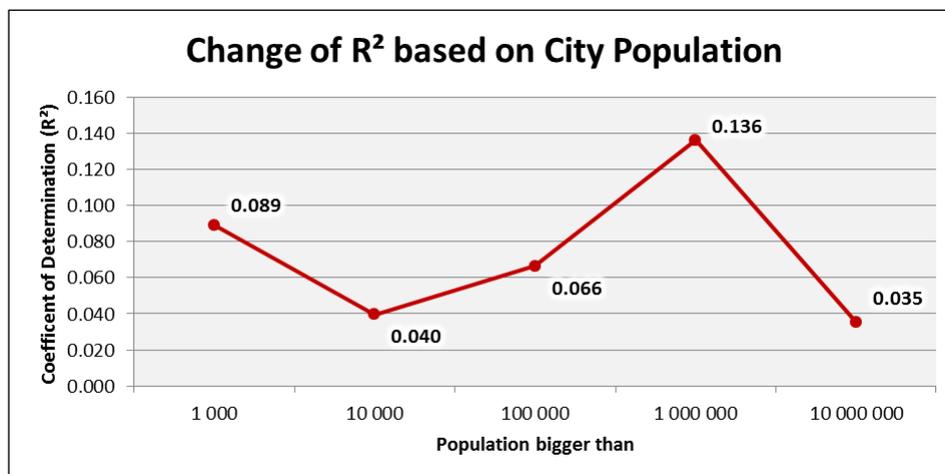
**Table 5.14: Spearman correlation coefficient of capital city joint probability rank and population rank**

Correlations: Capital City Name Joint Probability and Population Rank				
			Joint Probability Rank	Population Rank
Spearman's rho	Joint Probability Rank	Correlation Coefficient	1.000	.447**
		Sig. (2-tailed)	.	.000
		N	204	204
	Population Rank	Correlation Coefficient	.447**	1.000
		Sig. (2-tailed)	.000	.
		N	204	204

\*\* . Correlation is significant at the 0.01 level (2-tailed).

### 5.2.3 City Name Coefficient of Determination

The change of the coefficient of determination ( $R^2$ ) based on city population is shown in figure 5.16. It can be observed that the  $R^2$  between city joint probability rank and city population rank first gets smaller when cities with a higher population are selected. Hereafter, the exclusion of city names with lower populations does not bring an increase in  $R^2$ . This means that the data is still heavily scattered and no correlation can be observed between joint probability and population of cities. However,  $R^2$  slightly increases if only cities with a population higher than one million are taken. The selection of cities with a population over ten million reduces  $R^2$  again, since only 16 cities fulfill this criteria. Generally, the coefficient of determination stays small over all the population ranges. Accordingly, the exclusion of less populated cities does not improve the correlation of joint probability and population. Thus, ambiguity is likely present in all city population ranges.



**Fig. 5.16: Influence of population ranges on the coefficient of determination between city name joint probability rank and population rank**

### 5.3 Correctly Retrieved and Relevant Found Spatial Relations

In this part, the reliability of the Microsoft Web N-Gram Service is tested with the help of basic topological relations in the form of triplets. This is done for certain triplets: country in continent, capital city in continent, city in continent, country bordering country, capital city in country and city in country (an overview is given in table 5.15). Furthermore, the percentages of the retrieved spatial relations and relevant spatial relations are presented. These percentages serve as a measure of reliability for the Microsoft Web N-Gram Service.

**Table 5.15: Triplets investigated on correctly retrieved and relevant found spatial relations**

Investigated Place Name Triplets: <a><spatial relation><b>		
Place Name <b>	Continent	Country
Place Name <a>		
Country	in	bordering
Capital City	in	in
City	in	in

#### 5.3.1 Country in Continent

The spatial relation country in continent has a percentage of 49.19 correctly retrieved spatial relations, while the relevant found spatial relations are at 89.39% (as seen in table 5.16). The mediocre percentage of correctly retrieved spatial relations means that the autocompletes found for a country name followed by “in” has many mismatched continents. However, the percentage of relevant found spatial relations is considerably bigger than the percentage of correctly retrieved relations. Therefore, for the most doublets <country name><in> the correct continent name is found, but also many incorrect continent names are found. The percentages for correctly retrieved and relevant found spatial relations per continent are displayed in table 5.17. It can be observed that the percentage of correctly retrieved spatial relations per continent have a wide range of different values. The percentage of 28.65 correctly retrieved spatial relations for North America signifies that the countries in North America mostly also referred to other continents. Likewise, the continents Asia and South America have low percentages with their respecting values being at 35.45 and 32.72. The highest percentage of correctly retrieved spatial relations is found in Africa with a value of nearly 70%. This means that approximately 70% of the country names in Africa were only matched to the continent

name Africa in the autocompletes. Inversely, for the majority of the countries the relevant continents are found in the autocompletes. All countries in Africa, South America and Asia/Europe (countries which can refer to Asia or Europe) are assigned to the relevant continent name. The relevant continent name of the countries in North America and Australia was somewhat found in the autocompletes. The low percentages for countries in Australia are likely caused by the division of continents. That is to say, the islands states in the Pacific Ocean belonged to the continent Australia in the ground truth, rather than Oceania.

**Table 5.16: Percentages of correctly retrieved and relevant found continent names for the spatial relation country in continent**

Country in Continent		
Correctly Retrieved Spatial Relations	[%]	49.19
Relevant Found Spatial Relations	[%]	89.39

**Table 5.17: Percentages of correctly retrieved and relevant found spatial relations per continent for the triplet country in continent**

Country in Continent		
Continent	Spatial Relations	
	Correctly Retrieved [%]	Relevant Found [%]
Africa	69.62	100.00
Asia	35.45	87.23
Australia	52.86	56.25
Europe	48.71	97.78
North America	28.65	63.16
South America	32.72	100.00
Asia or Europe	50.95	100.00

### 5.3.2 Capital City in Continent

The percentages of correctly retrieved and relevant found spatial relations are presented in table 5.18 for the spatial relation capital city in continent. The correctly retrieved spatial relations lie at 73.22%, while the relevant spatial relations are at 32.28%. In this case the percentage of the correctly retrieved spatial relations is higher than the percentage of the relevant found spatial relations. In other words, several continent names are not found for the doublet <capital city name><in>. This is manifested in the percentage of relevant found spatial relations. The percentage states that only 32.38% of the relevant spatial relations are found, while several are missing or wrong. Nevertheless, the continent names found for a doublet were to 73.72% correct.

**Table 5.18: Percentages of correctly retrieved and relevant found continent names for the spatial relation capital city in continent**

Capital City in Continent		
Correctly Retrieved Spatial Relations	[%]	73.72
Relevant Found Spatial Relations	[%]	32.38

### 5.3.3 City in Continent

In table 5.19 the spatial relation city in continent and its corresponding percentages of correctly retrieved and relevant spatial relations is displayed. The correctly retrieved spatial relations have a value of 22.27%, while the relevant found spatial relations lie at 1.44%. Accordingly, a tiny extent of the continent names are found for the doublet <city name><in>. Moreover, the majority of the retrieved spatial relations for the doublet are incorrect or surplus continents are retrieved with the relevant continents.

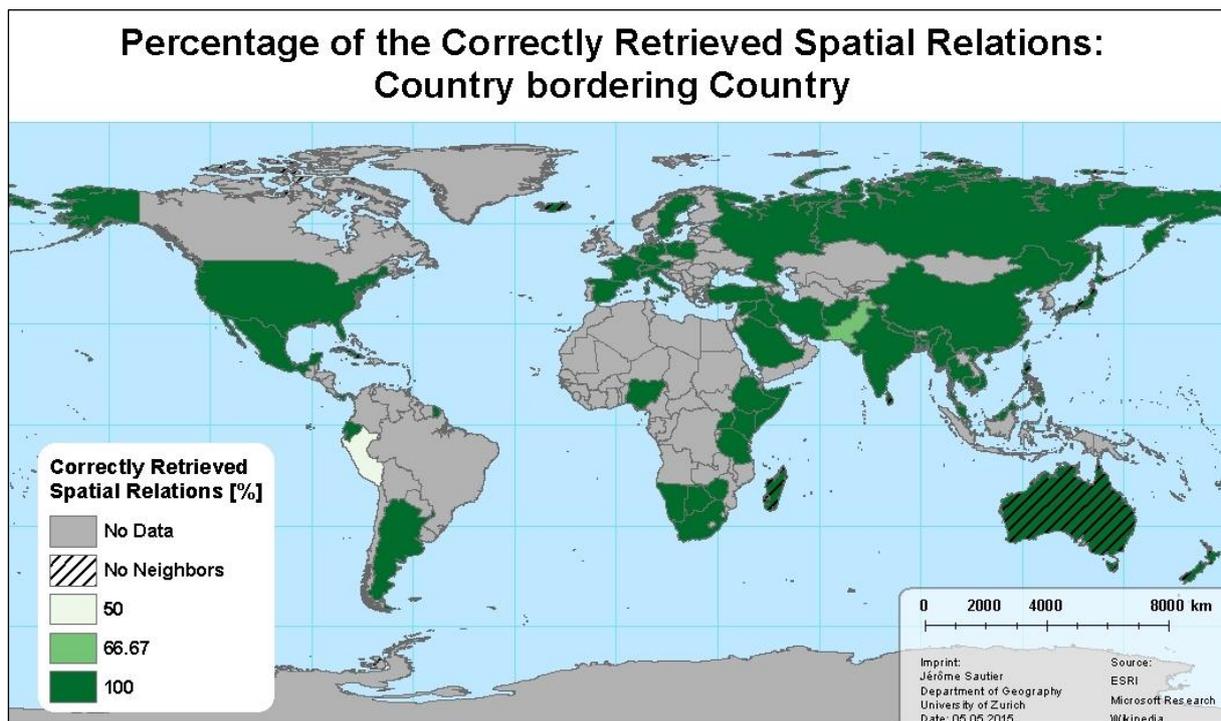
**Table 5.19: Percentages of correctly retrieved and relevant found continent names for the spatial relation city in continent**

City in Continent		
Correctly Retrieved Spatial Relations	[%]	22.27
Relevant Found Spatial Relations	[%]	1.44

### 5.3.4 Country bordering Country

The spatial relation country bordering country helps to investigate the neighboring properties of country names on the Web. The percentages of correctly retrieved spatial relations and the relevant found spatial relations are seen in table 5.20. The percentage of the correctly retrieved spatial relations is almost perfect, while the percentage of the relevant found spatial relations is rather low. Ultimately, this means that 98.96% of the retrieved spatial relations obtained the correct bordering country from the autocompletes and in total 28.19% of all relevant spatial relations were found. The spatial distribution of the correctly retrieved and relevant found spatial relations is illustrated in figure 5.17 and 5.18. There are only three classes for the percentage of correctly retrieved spatial relations. Thus, most of the neighboring countries retrieved for a country were correct. The countries with moderate percentages of correctly retrieved spatial relations are Peru and Pakistan. The striped countries contained no neighboring countries and therefore if no neighboring country was retrieved it

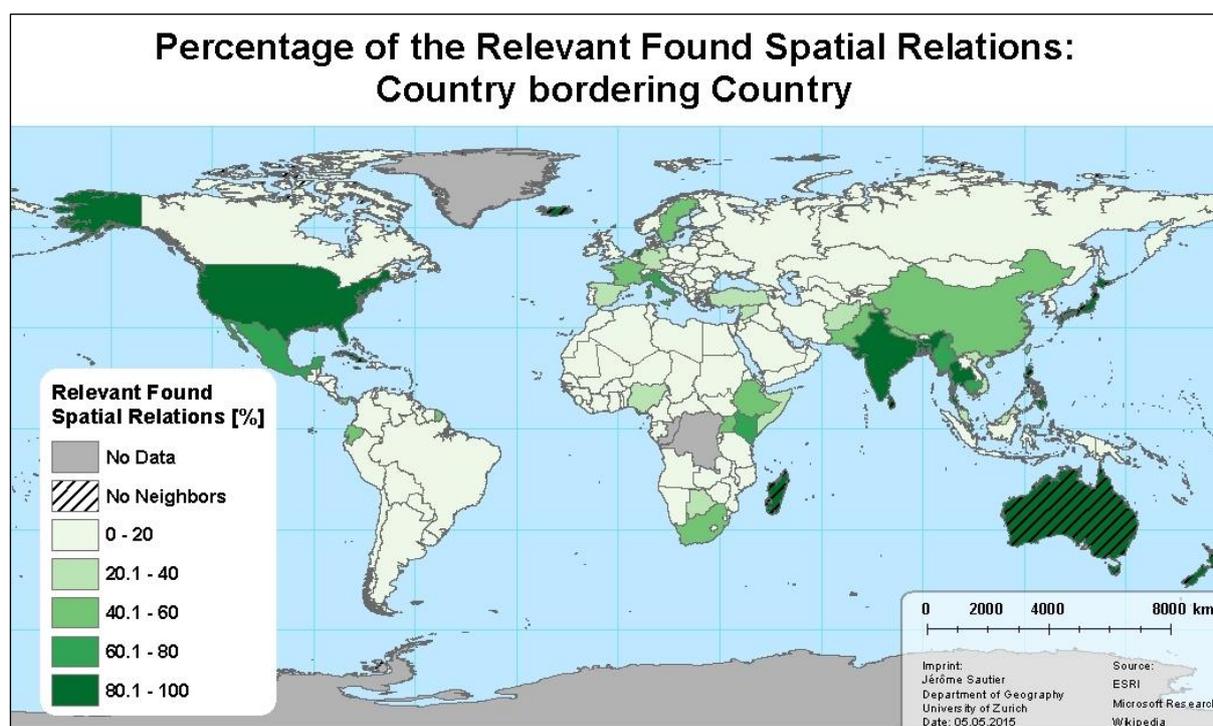
was classified as correct. For the rest of the countries, not a single country was retrieved or the triplets were too long to be retrieved. It can be observed that there are clusters with high percentages of relevant found relations in Asia and Europe. Additionally, the United States, Mexico, Australia, New Zealand, parts of Central East Africa, Madagascar and South Africa have a high percentage of relevant found spatial relations. For these country names the number of relevant found bordering country names is high. Though, some of these countries have no neighboring countries. These are striped in figure 5.18. On the other hand, the countries with a low percentage of relevant found spatial relations are focused in South America, big parts of Africa, Eastern Europe and the Middle East. For these country names the retrieved neighboring country names from the autocompletes were wrong, incomplete or blank (no match found). The cases of countries having no bordering country are striped. These mostly have a high percentage of relevant found spatial relations, since finding no neighboring country name is correct. Finally, the countries with no data were no countries per se or contained too long triplets and therefore could not be searched in the Microsoft Web N-Gram Service.



**Fig. 5.17: Percentage of the correctly retrieved country names bordering a country name**

**Table 5.20: Percentages of correctly retrieved and relevant found country names for the spatial relation country bordering country**

Country bordering Country		
Correctly Retrieved Spatial Relations	[%]	98.96
Relevant Found Spatial Relations	[%]	28.19



**Fig. 5.18: Percentage of the relevant found country names bordering a country name**

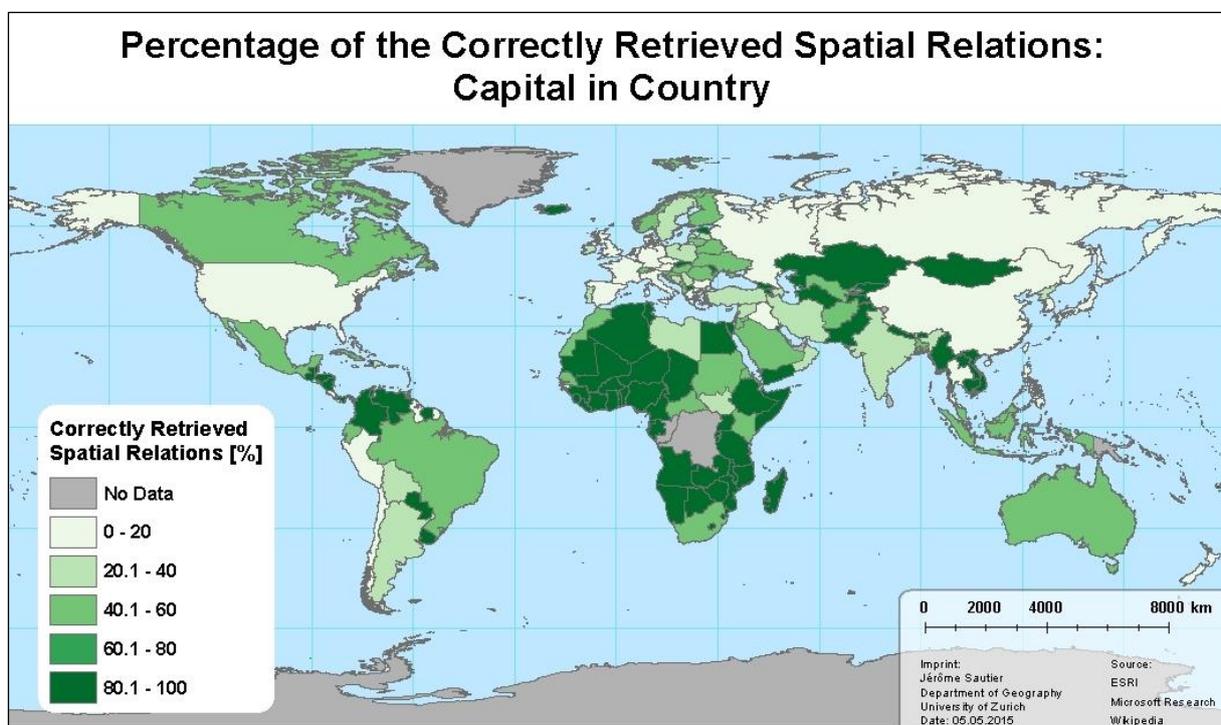
### 5.3.5 Capital City in Country

The percentages of correctly retrieved and relevant found spatial relations are presented in table 5.21 for the spatial relation capital city in country. For the total correctly retrieved spatial relations a value of 64.27% was received, while the relevant found spatial relations have a value of 94.58%. Therefore, the percentage of the correctly retrieved spatial relations is smaller than the percentage of the relevant found spatial relations. This indicates that autocompletes contain most of the relevant country names for the doublet <capital city><in>. However, numerous false country names are retrieved which decreases the percentage of correctly retrieved spatial relations. The spatial distribution of correctly retrieved spatial relations is illustrated in figure 5.19. The map shows low correctly retrieved spatial relations in the United States, the west coast of South America, Europe, Russia and parts of East and South Asia. Accordingly, multiple incorrect countries are retrieved for the capitals in those countries. This could be a sign of possible ambiguity. In contrary, the rest of South America,

Africa, Eastern Europe and parts of Central Asia have high percentages of correctly retrieved spatial relations for capital in country. Especially, the whole continent of Africa stands out with high correctly retrieved percentages. This means that for most capitals in Africa only the corresponding country name was retrieved.

**Table 5.21: Percentages of correctly retrieved and relevant found country names for the spatial relation capital city in country**

Capital City in Country		
Correctly Retrieved Spatial Relations	[%]	64.27
Relevant Found Spatial Relations	[%]	94.58

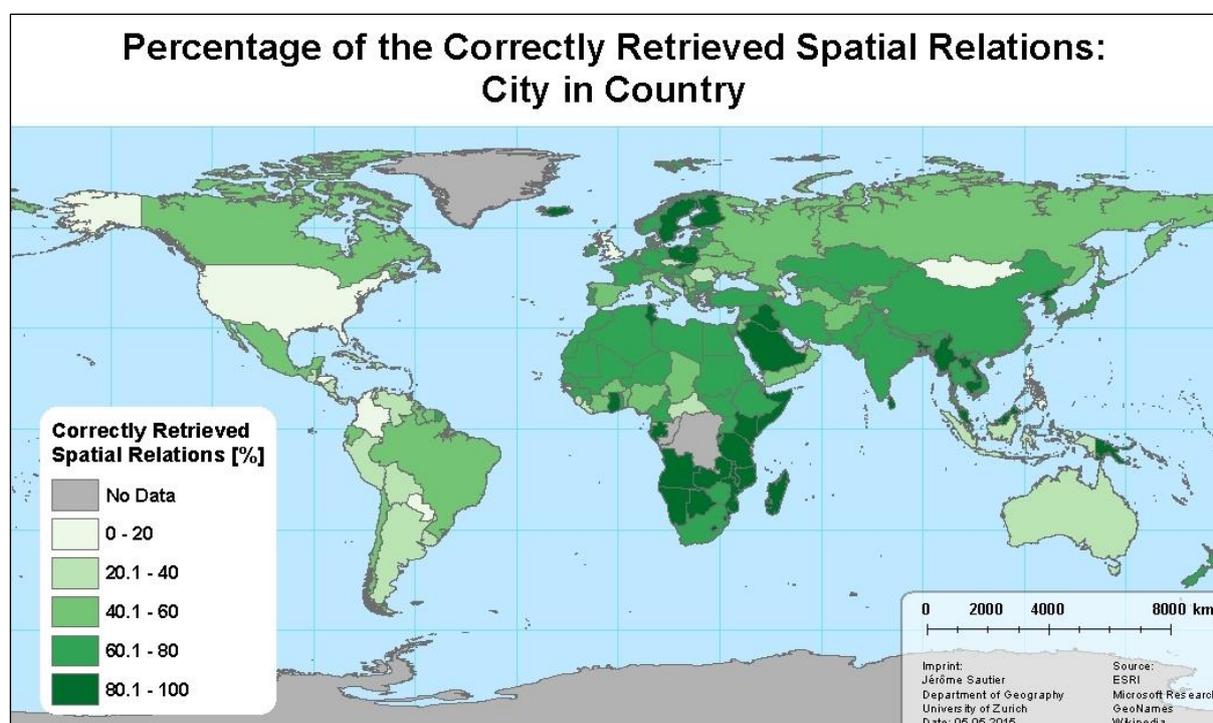


**Fig. 5.19. Percentage of the correctly retrieved country names for the spatial relation country in capital**

### 5.3.6 City in Country

In table 5.22 the percentages of correctly retrieved spatial relations and the relevant found spatial relations are displayed for the spatial relation city in country. The percentage of the correctly retrieved spatial relations is mediocre, while the percentage of the relevant found spatial relations is small. Eventually, this means that 41.33% of the retrieved spatial relations obtained the correct country from the autocompletes and in total 7.85% of all relevant spatial relations were found. The spatial distribution of the correctly retrieved and relevant found spatial relations is illustrated in figure 5.20 and 5.21. At first glance, it can be perceived that the cities in the United States, Colombia, Paraguay, the United Kingdom, Mongolia and

Philippines are associated with wrong country names. Consequently, their corresponding percentage of correctly retrieved spatial relations is the lowest. This may indicate that the city names in these places are subject to ambiguity. The same thing applies for parts of South America, Indonesia and Australia which also have low percentages. Seemingly, the native English speaking and Spanish speaking countries have lower percentages of correctly retrieved spatial relations. Alternatively, moderately high percentages of correctly retrieved spatial relations are observed in Europe, Africa, Middle East, South and East Asia. In these places the cities are mostly associated to the correct country name. On the other hand, the relevant found spatial relations percentages are much lower over all the countries. This means that most city names are not associated with their relevant country in the autocompletes. In total 0-20% of these countries are linked to their corresponding city names. The only cluster which can be observed in the map is a higher rate of relevant found countries for the city names in Africa and the Middle East. Especially, in the south and west of Africa a cluster of high percentages of the relevant found spatial relations are witnessed. A possible cause for this low values might be the number of cities per country. The figure 22<sup>21</sup> displays the city name density for each country. The map clearly signifies a high city name density in Europe. Therefore, the high number of city names may contain less popular or unknown city names. These are likely not associated with their corresponding country in the autocompletes.

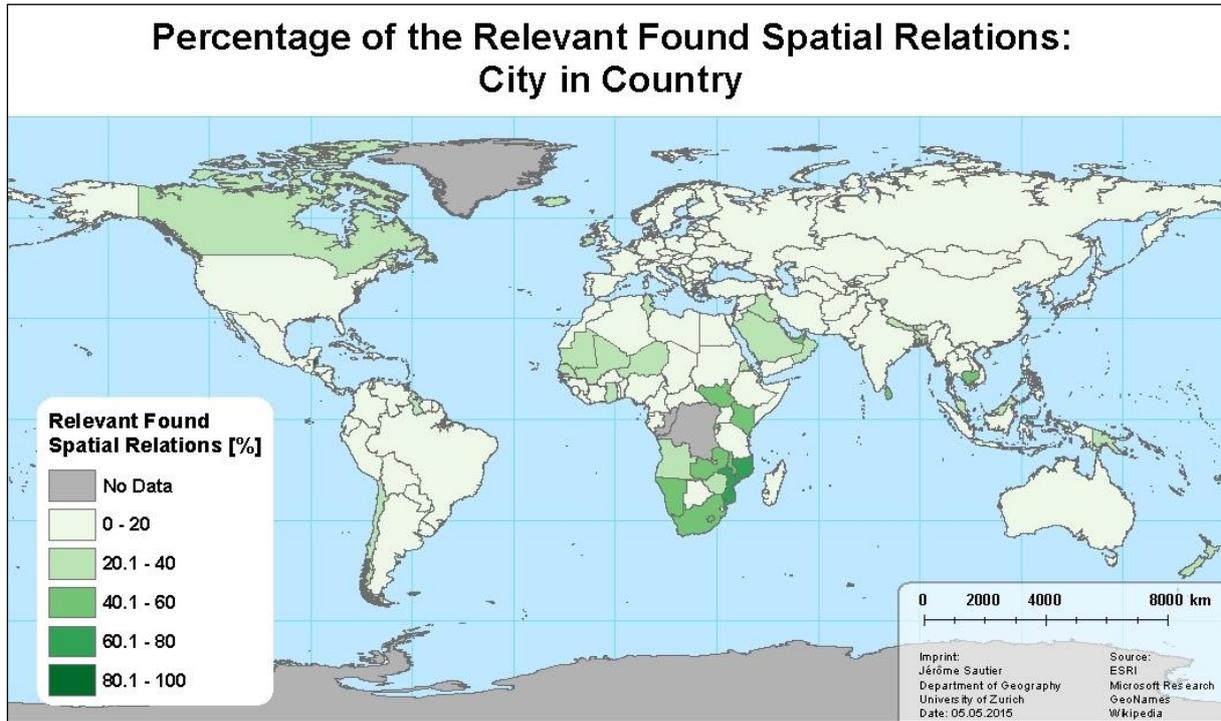


**Fig. 5.20: Percentage of the correctly retrieved country names for the spatial relation city in country**

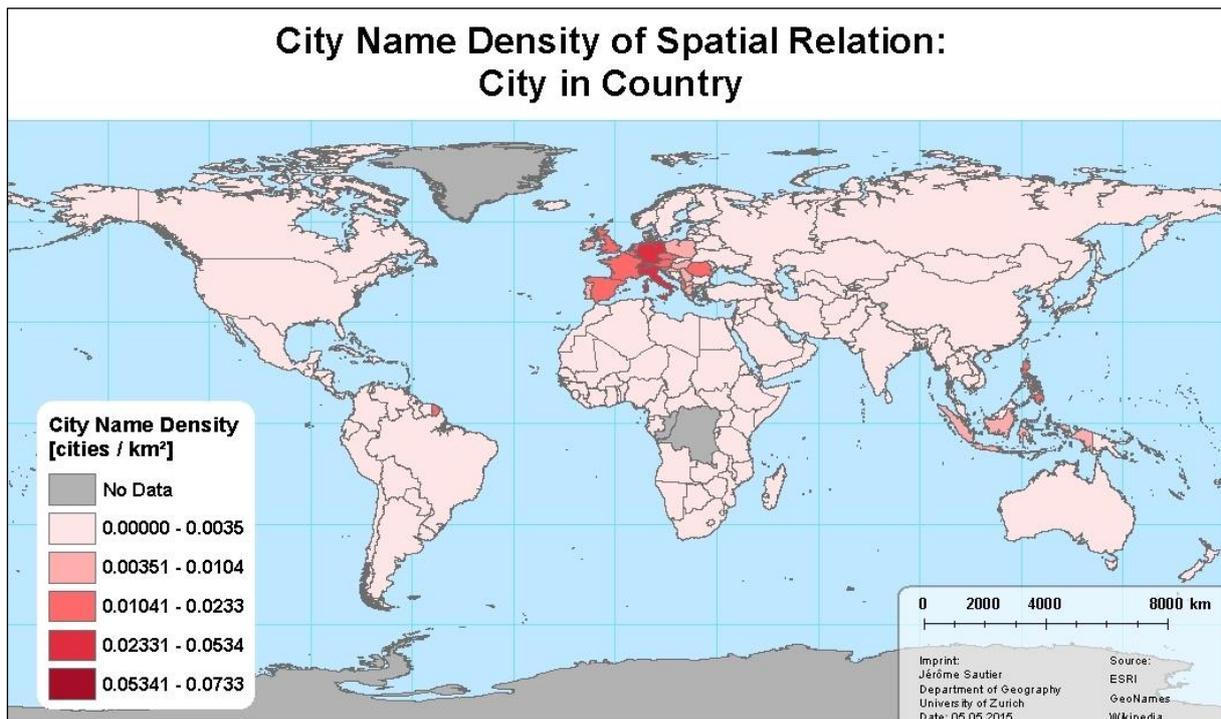
<sup>21</sup> The city name density of Vatican City, Monaco, Nauru, Tuvalu, Malta, Marshall Islands and San Marino were excluded from the map, as they distorted the classes with their high city name densities.

**Table 5.22: Percentages of correctly retrieved and relevant found country names for the spatial relation city in country**

City in Country		
Correctly Retrieved Spatial Relations	[%]	41.33
Relevant Found Spatial Relations	[%]	7.85



**Fig. 5.21: Percentage of the relevant found country names for the spatial relation city in country**



**Fig. 5.22: Country city name density for the spatial relation city in country**

## 5.4 Explorative Approach

The results of assigning locations to geographic features and sports activities are explored in this chapter. At first, the conditional probabilities of geographic features followed by European countries or cities are looked at. Additionally, the spatial distribution of countries and geographic features is investigated. These steps are repeated for the sports activities. The plausibility of the results is questioned in each step and verified by querying for suspicious results on a Web search engine. Finally, possible connections are identified by comparing the results between geographic features and sport activities.

### 5.4.1 Geographic Features

The place names to follow a geographic feature are explored in this section. This is done for a proportion of the results on country and city level. The appendix C.4 and C.5 can be examined for further examples.

#### 5.4.1.1 Countries to follow Geographic Feature

The countries most likely to follow “delta in” and their corresponding conditional probabilities are illustrated in figure 5.23. In total eleven countries were retrieved for the geographic feature delta. It can be observed that delta is most likely to be followed by Romania. In fact, Romania is at least five times more likely to follow “delta in” than all the other European countries. Hence, this must mean that a number of deltas or a famous delta is located in Romania.

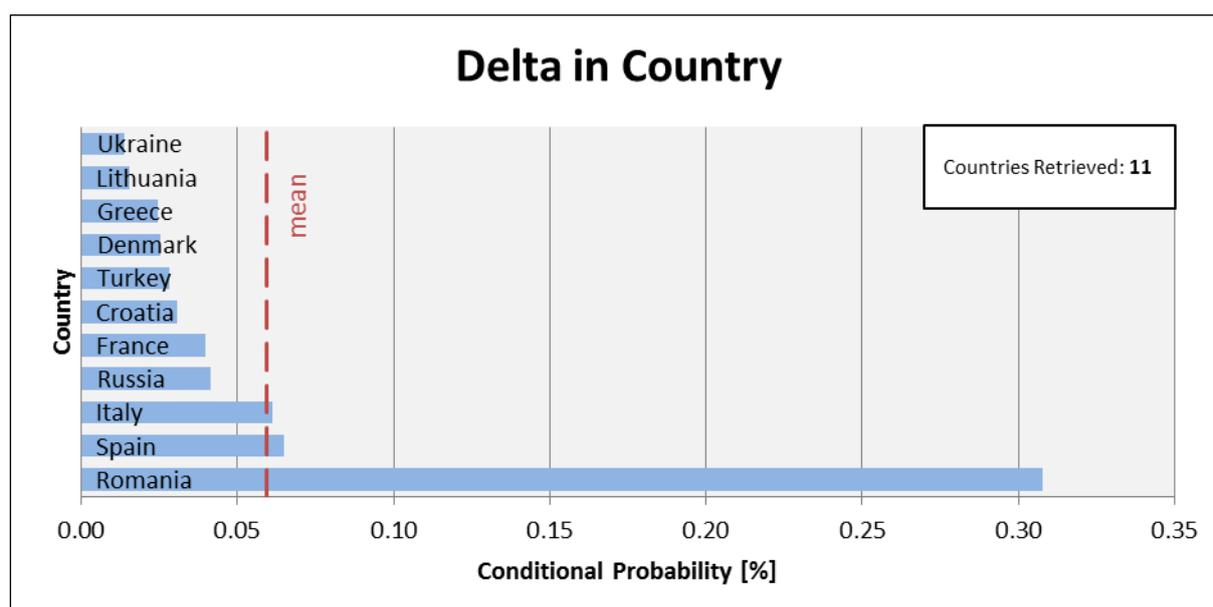


Fig. 5.23: Conditional probability of the 20 most likely countries to follow “delta in”

A Web search on Bing (see figure 5.24) clarifies a possible causes for the high conditional probability of Romania. The delta in Romania refers to the Danube Delta situated between Romania and Ukraine. It is the second largest river delta in Europe and the Romanian part of the Danube Delta is part of the UNESCO list of World Heritage Sites (“Danube Delta” 2007). This is most likely the cause for the high conditional probability of “delta in” followed by Romania.

7,290,000 RESULTS

**Danube Delta - Wikipedia**, the free encyclopedia  
[en.wikipedia.org/wiki/Danube\\_Delta](https://en.wikipedia.org/wiki/Danube_Delta) ▾  
 The **Danube Delta** is the second largest river **delta** in Europe, after Volga **Delta**, and is the best preserved on the continent. The greater part of the **Danube Delta** lies ...  
 Geography and geology · Inhabitants · History · Environment and issues

**Danube Delta - Official Travel and Tourism Information**  
[romaniatourism.com/delta.html](http://romaniatourism.com/delta.html) ▾  
 The **Danube Delta (Delta Dunarii)** - ... **Romania** A VISIT TO THE **DANUBE DELTA** usually begins in the **Romanian** town of Tulcea, a two-hour drive from Constanta.  
 Banat & Crisana · The Carpathian Mountains · Geography · Romania · Architecture

**Delta Air Lines - Official Site**  
[www.delta.com](http://www.delta.com) ▾  
 Official website of **Delta Airlines** including trip bookings, check-in, flight status, and travel information.  
 Book a Flight · Check In · Flight Status · SkyMiles · Flight Schedules

**Delta Dunării - Wikipedia**  
[https://ro.wikipedia.org/wiki/Delta\\_Dunării](https://ro.wikipedia.org/wiki/Delta_Dunării) ▾  
**Delta** Dunării este plasată, din punct de vedere geologic, într-o regiune mobilă a scoarței terestre numită Platforma Deltei Dunării (regiunea predobrogeană).

**Turismul in Romania - Delta Dunarii - Poze, Informatii**  
[www.romanianmonasteries.org/ro/romania/delta-dunarii](http://www.romanianmonasteries.org/ro/romania/delta-dunarii) ▾  
 Bratele Dunarii. Se poate spune ca **Delta** incepe sa se contureze de la Patlageanca unde se bifurca in doua brate, Bratul Chilia la nord si Bratul Tulcea la sud.

**Danube Delta, Romania - Lonely Planet**  
[www.lonelyplanet.com/romania/danube-delta](http://www.lonelyplanet.com/romania/danube-delta) ▾  
 The **Danube** port of Tulcea (pronounced tool-cha) is the largest city in the **delta** and the main entry point for accessing the region. It's got good bus and minibus ...  
 Highlights · Places · Transport · Activities · Sights

**Booking.com: 142 hotels in Danube Delta, Romania. ...**  
[www.booking.com/region/ro/danube-delta.html](http://www.booking.com/region/ro/danube-delta.html) ▾  
 Find hotels in **Danube Delta, Romania**. Book online, pay at the hotel. Good rates and no reservation costs. Read hotel reviews from real guests.

**Delta Dunarii, Fotografii, Turism Ecologic - Rezervatie ...**  
[www.delta-dunarii.ro](http://www.delta-dunarii.ro) ▾  
**Delta Dunarii** - Peisaj mirific, unic in Europa, paradisul pescarilor, cazare - turism, plimbări cu barci de agrement pe diferite trasee turistice

**DANUBE DELTA, ROMANIA - YouTube**  
[www.youtube.com/watch?v=6X1P320KVUk](http://www.youtube.com/watch?v=6X1P320KVUk) ▾  
 By Dimitra Stasinopoulou · 11 min · 2.1K views · Added 10/3/2013  
 Video embedded · PHOTOGRAPHS: Dimitra Stasinopoulou  
[www.dimitrastasinopoulou.com](http://www.dimitrastasinopoulou.com) MUSIC: Iosif Ivanovici "Valurile Dunării", Tudor ...

**Danube Delta (Tulcea, Romania): Hours, Address, Top ...**  
[www.tripadvisor.co.uk/...Reviews-Danube\\_Delta\\_Southeast\\_Romania.html](http://www.tripadvisor.co.uk/...Reviews-Danube_Delta_Southeast_Romania.html) ▾  
 ★★★★★ Rating: 4.5/5 · 177 reviews · 263 photos · Address:, Tulcea, Romania  
**Danube Delta**, Tulcea: See 177 reviews, articles, and 263 photos of **Danube Delta**, ranked No.1 on **TripAdvisor** among 9 attractions in Tulcea.

Fig. 5.24: First webpage recommendations on Bing for the query “delta in Romania”

The distribution of the 20 most likely countries to follow “lake in” is illustrated in figure 5.25. In total 27 countries were retrieved for this topic. The European country best known for lakes or mostly associated with lakes is Finland. This makes sense, since Finland is covered with multiple lakes. Additional countries with a conditional probability bigger than the mean conditional probability include: Italy, France, England, Georgia, Turkey, Scotland, Russia, Switzerland, Wales and Sweden. However, caution is advised for ambiguous country names. The lakes in Georgia are most likely referring to the lakes in the state of Georgia, United States.

The countries most likely to follow the geographic feature sea are seen in figure 5.26. In total 17 countries were retrieved for the topic. As expected, the countries which are close to the sea have high conditional probabilities. The most frequent countries following “sea near” seem to be situated in or at the Mediterranean Sea. Malta for example has the highest conditional probability. This is probably caused due to the fact that it is completely surrounded by the Mediterranean Sea.

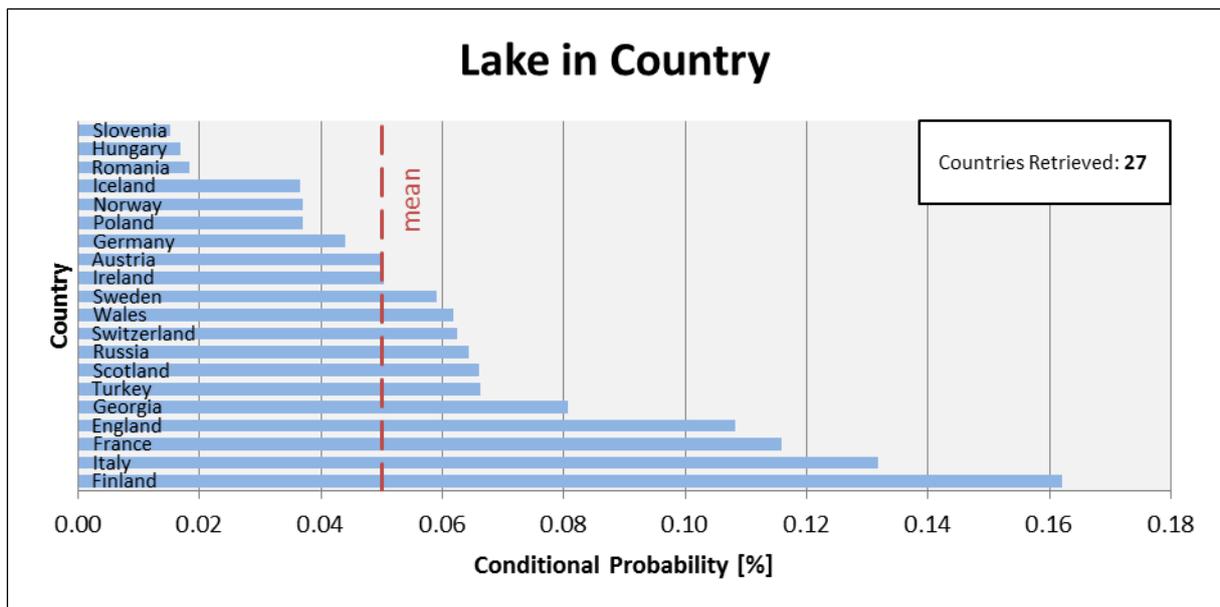


Fig. 5.25: Conditional probability of the 20 most likely countries to follow "lake in"

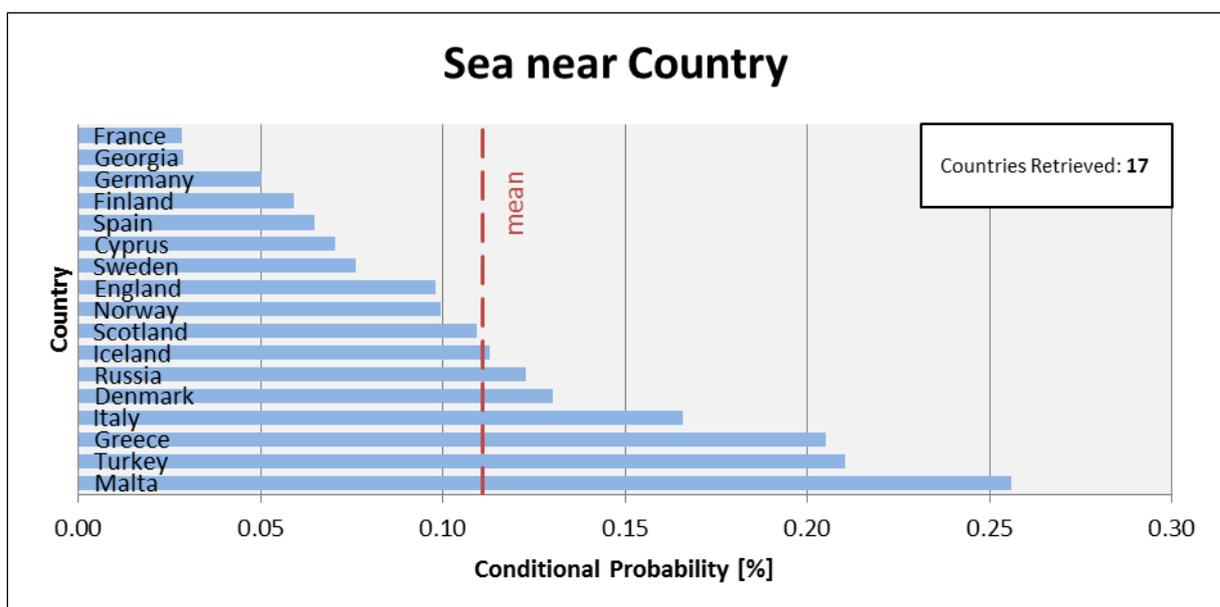


Fig. 5.26: Conditional probability of the 20 most likely countries to follow "sea near"

The countries most likely to follow "valley in" and their corresponding conditional probabilities are illustrated in figure 5.27. In total 23 countries were retrieved for the geographic feature valley. It can be observed that valley is most likely to be followed by France. The conditional probability of France is even twice as big as all other countries following "valley in". Consequently, France must be known for its valleys or must have a well-known valley. A Web query was commissioned on Bing to investigate the possible cause for the high conditional probability, regarding valley in France. The proposed webpages for the query "valley in France" are displayed in figure 5.28. The most frequent word appearing in this search is the Loire Valley and wine valleys. The Loire Valley is known for its

abundance of vineyards, fruit orchards, artichoke and asparagus fields (“Loire Valley” 2007). Additionally, the central part of the Loire River Valley is also part of the UNESCO list of World Heritage Sites (“Loire Valley” 2007). This most likely causes valley to be heavily associated with France.

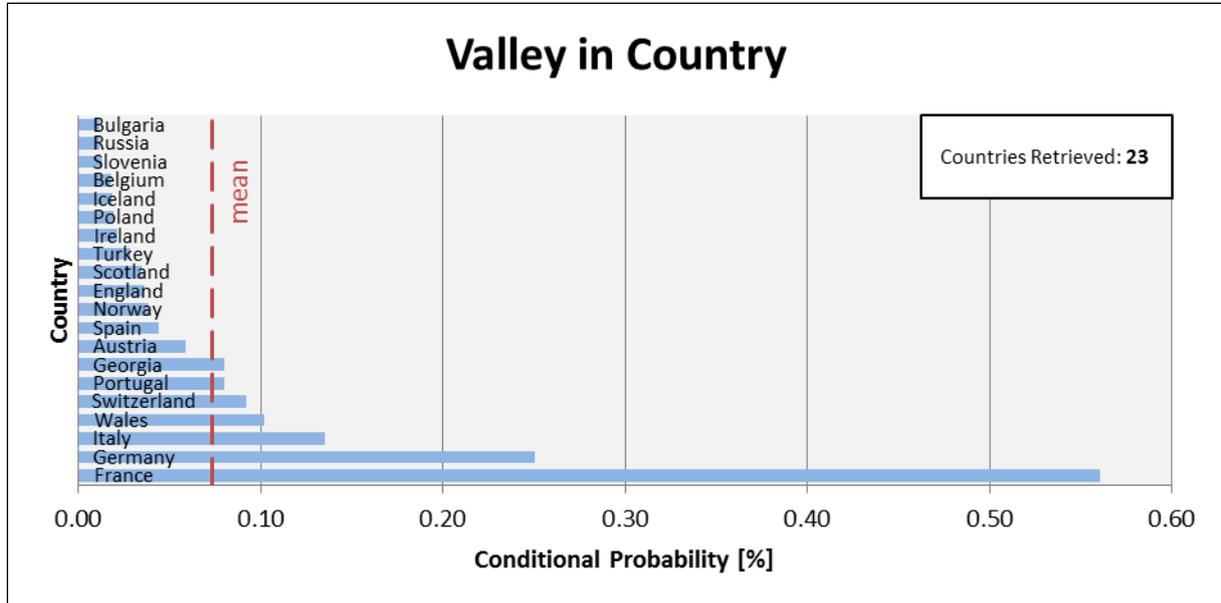


Fig. 5.27: Conditional probability of the 20 most likely countries to follow “valley in”

14,100,000 RESULTS

**Loire Valley - Wikipedia, the free encyclopedia**

[en.wikipedia.org/wiki/Loire\\_Valley](http://en.wikipedia.org/wiki/Loire_Valley)  
 The Loire Valley, spanning 280 kilometres (170 mi), is located in the middle stretch of the Loire River in central France, primarily within the administrative region ...  
 Geography and climate · Wine · Culture · Architecture  
 Loire · Château d'Amboise · Château De Villandry

**Category:Valleys of France - Wikipedia, the free ...**

[en.wikipedia.org/wiki/Category:Valleys\\_of\\_France](http://en.wikipedia.org/wiki/Category:Valleys_of_France)  
 Pages in category “Valleys of France” The following 18 pages are in this category, out of 18 total. This list may not reflect recent changes ...

**Loire Valley Tourism: holidays in Loire, tourist information**

[www.loirevalleytourism.com](http://www.loirevalleytourism.com)  
 Representing the Loire Valley, the Comité Régional du Tourisme Centre (CRT) offers all the practical information the tourists need to organise their France holiday ...

**Images of valley in france**



See more images of valley in france

**Loire Valley Tourism: Best of Loire Valley, France ...**

[www.tripadvisor.co.uk/Tourism-g187196-Loire\\_Valley\\_Centre...](http://www.tripadvisor.co.uk/Tourism-g187196-Loire_Valley_Centre...)  
 Loire Valley Tourism: TripAdvisor has 157,203 reviews of Loire Valley Hotels, Attractions, and Restaurants making it your best Loire Valley resource.

**Explore Each French Wine Valley - Basic Wine Knowledge**

[www.basic-wine-knowledge.com/french-wine-valley.html](http://www.basic-wine-knowledge.com/french-wine-valley.html)  
 Learn About Each French Wine Valley. No visit to France would be complete without a visit to a French wine valley. There are several to choose from, so regardless of ...

**Loire Valley Chateau Map - About.com Travel**

[goeurope.about.com/od/loirevalley/ss/loire-chateau-map.htm](http://goeurope.about.com/od/loirevalley/ss/loire-chateau-map.htm)  
 The Loire Valley of France is notable for its wines and the castles or chateaux that are found in profusion along the Loire and Cher rivers.

**Loire Valley Tours - France.com**

[tours.france.com/loire-valley-tours](http://tours.france.com/loire-valley-tours)  
 Best tours to the Loire Valley from Paris, from cultural tours to wine tours. Visit the Loire river valleys, Loire castles, chateaux of Chambord, Chenoncau, Villandry ...

**Loire Valley - France Travel Guide**

[www.francetravelguide.com/loire-valley](http://www.francetravelguide.com/loire-valley)  
 Background The Loire River gives this valley its lifeline. It is the longest river in France (630 miles) and the last wild river in all of Europe. Once the

**The Loire Valley Travel Guide | Fodor's Travel**

[www.fodors.com/world/europe/france/the-loire-valley](http://www.fodors.com/world/europe/france/the-loire-valley)  
 Expert picks for your The Loire Valley vacation, including hotels, restaurants, entertainment, shopping, top attractions, and more.

**The Lovely, Lively Lot Valley | France Today**

[www.francetoday.com/.../2012/11/26/the\\_lovely\\_lively\\_lot\\_valley.html](http://www.francetoday.com/.../2012/11/26/the_lovely_lively_lot_valley.html)  
 The longest river in southwest France, the Lot winds for almost 300 miles, affording a new surprise at every bend.

Fig. 5.28: First webpage recommendations on Bing for the query “valley in France”

The distribution of the 20 most likely countries to follow “forest in” is illustrated in figure 5.29. In total 34 countries were retrieved for this topic. The European country mostly associated with forests is Germany. This makes sense, since the Black Forest in Germany is a

well-known forest in Europe. Additionally, the countries England and France are also highly associated with forest compared to the other countries.

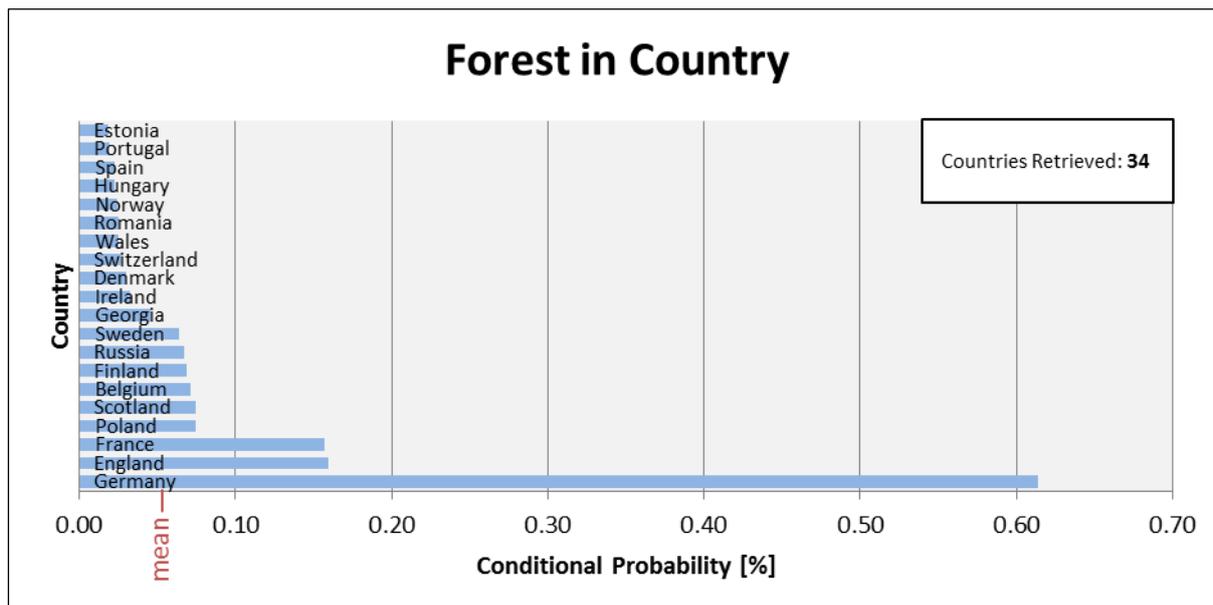


Fig. 5.29: Conditional probability of the 20 most likely countries to follow “forest in”

The countries most likely to be associated with beach are displayed in figure 5.30. In total 28 countries were retrieved for the geographic feature beach. The European countries mostly associated with beach are: Spain, Greece, France, Turkey, Italy, Portugal and Wales. Spain is proportionally more likely to be linked to beach compared to the other countries.

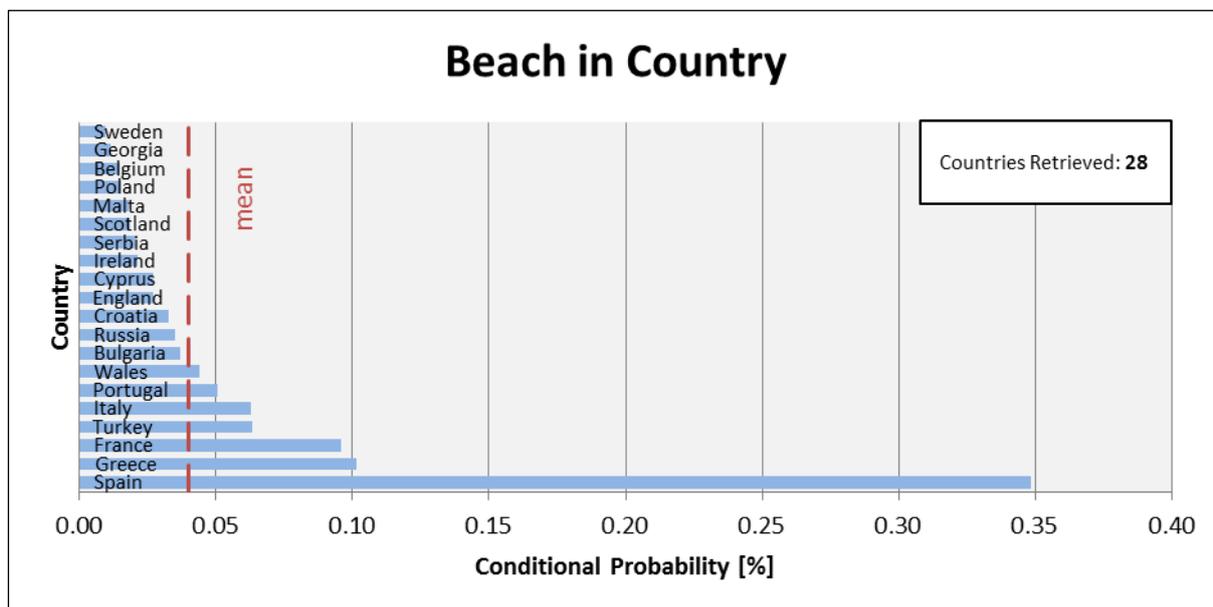


Fig. 5.30: Conditional probability of the 20 most likely countries to follow “beach in”

The distribution of the 20 most likely countries to follow “mountain in” is illustrated in figure 5.31. In total 37 European countries were retrieved for the geographic feature mountain. The countries mostly likely to be related with mountain are: England, Georgia, Wales, France,

Spain, Greece, Italy, Switzerland, Scotland, Germany, Ireland and Norway. These countries have a higher conditional probability compared to the mean conditional probability. Yet again, the probability of Georgia is likely influenced by the state Georgia. The suspicious high conditional probability for mountain in England was further investigated by a Web query on Bing, since England is not really known for high mountains. A screenshot of those results is available in figure 5.32. The Web search showed that a lot of low mountains and hills exist in England. Moreover, the mountains in the United Kingdom seem to be very popular for hikers and mountain-bikers. This possibly influences the results and links mountain with England.

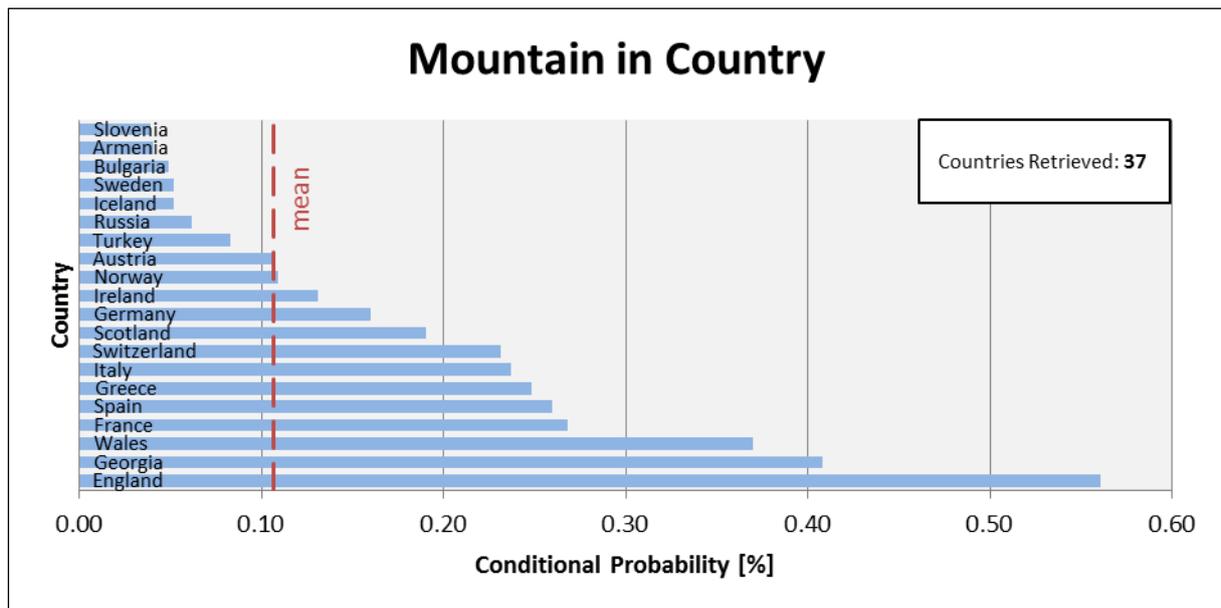


Fig. 5.31: Conditional probability of the 20 most likely countries to follow “mountain in”

15,400,000 RESULTS

List of **mountains and hills of the United Kingdom** ...  
[en.wikipedia.org/.../List\\_of\\_mountains\\_and\\_hills\\_of\\_the\\_United\\_Kingdom](https://en.wikipedia.org/List_of_mountains_and_hills_of_the_United_Kingdom) ▼  
 This is a links page to the **hills and mountains** to be found in the **United Kingdom** (England, Northern Ireland, Scotland, and Wales), and includes lists of the **highest** ...  
 Highest **mountains** in ... · Highest **mountains** in ... · Highest **mountains** in ...

**Mountains and hills of England** - Wikipedia, the free ...  
[en.wikipedia.org/wiki/Mountains\\_and\\_hills\\_of\\_England](https://en.wikipedia.org/wiki/Mountains_and_hills_of_England) ▼  
 The **mountains and hills of England** comprise very different kinds of terrain, from a **mountain** range which reaches over 3,000 foot (910 m) high, to several smaller ...  
 Northern **England** · Midlands · South East **England** · South West **England**

The **Highest Mountains in England** - Walking Englishman  
[www.walkingenglishman.com/mountains.html](https://www.walkingenglishman.com/mountains.html) ▼  
 Although **England** is the poor relation on the British Mainland in **mountain** scale there are still over 200 which reach over 2,000 feet (610 metres) in height.

Images of **mountain in england**  
[bing.com/images](https://bing.com/images)  
  
 See more images of **mountain in england**

Map of **UK mountains** - World Map  
[www.mapsofworld.com/united-kingdom/thematic-maps/uk-mountains-map.html](https://www.mapsofworld.com/united-kingdom/thematic-maps/uk-mountains-map.html) ▼  
 UK **Mountains** Map - Find out the all **mountains** are in the countries of **United Kingdom** (UK).

**Highest mountains in Britain | UK hills & mountain** ...  
[www.themountainguide.co.uk/highest](https://www.themountainguide.co.uk/highest) ▼  
 List of peaks in the British Isles with Prominence greater than 600 meters. Shows **mountains** in descending order of relative height

Walk The 179 **Mountains in England** - go4awalk.com  
[www.go4awalk.com/uk-mountains-and-hills/mountains-in-england.php](https://www.go4awalk.com/uk-mountains-and-hills/mountains-in-england.php) ▼  
 Peak Bagging The **Mountains in England** is a way of discovering and experiencing new **hills and mountains** that you might not otherwise have thought of walking on ...

MountainSummits.com - Hiking New **England's Mountains**  
[www.mountainsummits.com](https://www.mountainsummits.com) ▼  
 A database of **mountains** in New **England** providing all the information you need to plan your hike to specific **mountains** including such information as trail descriptions ...

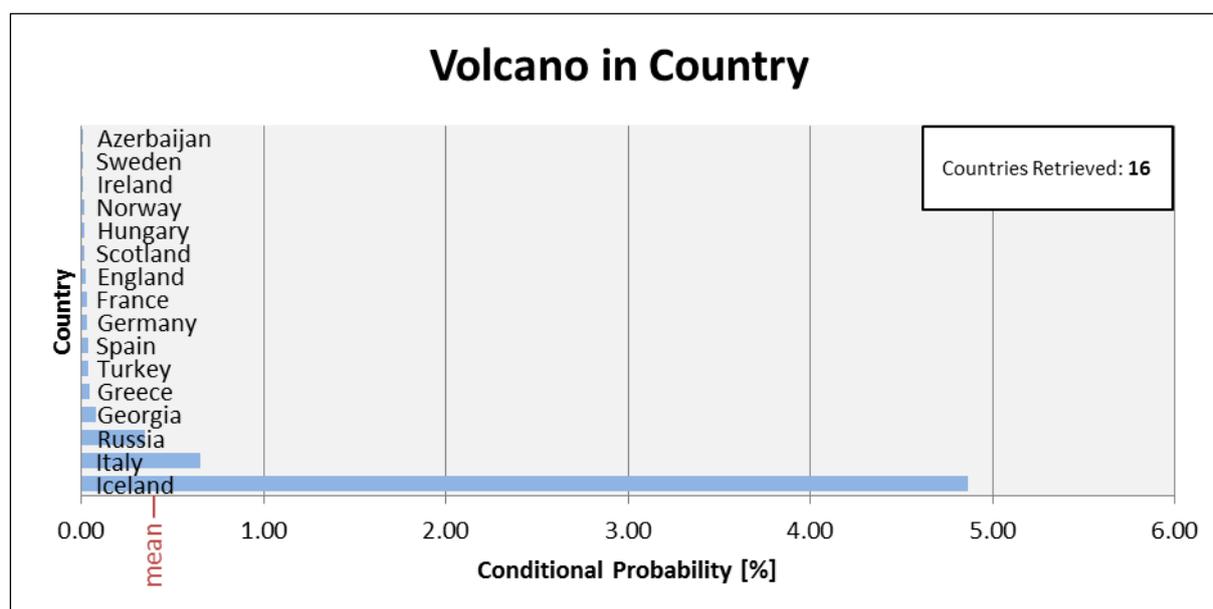
Best to **UK Mountains Views | Welsh Mountains | Visit** ...  
[www.visitwales.com/explore/mountains/10-best-views](https://www.visitwales.com/explore/mountains/10-best-views) ▼  
**Mountains** define the landscape in Wales. Here are ten of the best places to stop and admire them. Best scenic **mountain** views to enjoy, paint or explore.

Top 10 **mountain** biking routes in Britain - **Telegraph.co.uk**  
[www.telegraph.co.uk/.../Top-10-mountain-biking-routes-in-Britain.html](https://www.telegraph.co.uk/.../Top-10-mountain-biking-routes-in-Britain.html)  
 9/26/2008 - From the **hills** of Wales to the challenging terrain of Yorkshire, Nicholas Roe picks out the finest **mountain**-biking routes in Britain.

**Mountain Hikes UK** - Home  
[mountainhikesuk.weebly.com](https://mountainhikesuk.weebly.com) ▼  
 Home of **Mountain Hikes UK**, everything you need to get hiking in the **UK** from skills to news, you can find it all here.

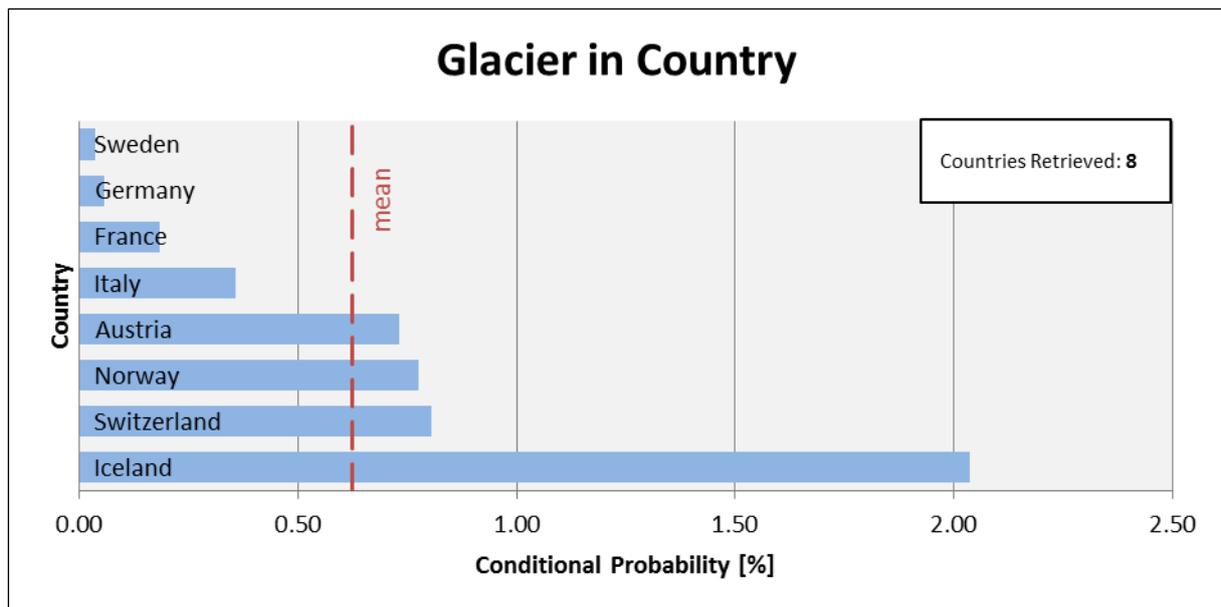
Fig. 5.32: First webpage recommendations on Bing for the query “mountain in England”

The countries most likely to follow the geographic feature volcano are: Iceland, Italy and Russia. The remaining conditional probabilities of the other countries can be observed in figure 5.33. In total 16 countries followed the geographic feature volcano in the 1000 autocompletes. Overall, the results seem plausible since Iceland and Italy are well known for their volcanoes. Italy has a long history with volcanoes, while Iceland lies on the Mid-Atlantic Ridge. Additionally, the Eyjafjallajökull volcano in Iceland was prominent in the news when it erupted in 2010 and led to air travel disruptions. This may be the cause why Iceland is so prominently associated with volcanoes compared to other countries, because the corpus of the Microsoft Web N-Gram Service is based on Web snapshots taken in 2010 and 2013. Thus, a variety of websites in the Microsoft Web N-gram Service might contain news to the eruption of the Eyjafjallajökull.



**Fig. 5.33: Conditional probability of the 20 most likely countries to follow “volcano in”**

For the countries following glacier no unexpected result were obtained. However, the expected countries were successfully retrieved in the 1000 autocompletes. The countries most likely to follow glacier are: Iceland, Switzerland, Norway and Austria. These are probably also the most prominent countries in Europe associated with glaciers. All of these countries also contain some sort of prominent glacier. Iceland got a variety of prominent glaciers such as the mentioned Eyjafjallajökull which covers the top of a volcano. Switzerland got the Aletsch Glacier and others, while Norway and Austria also have a wide variety of glaciers.

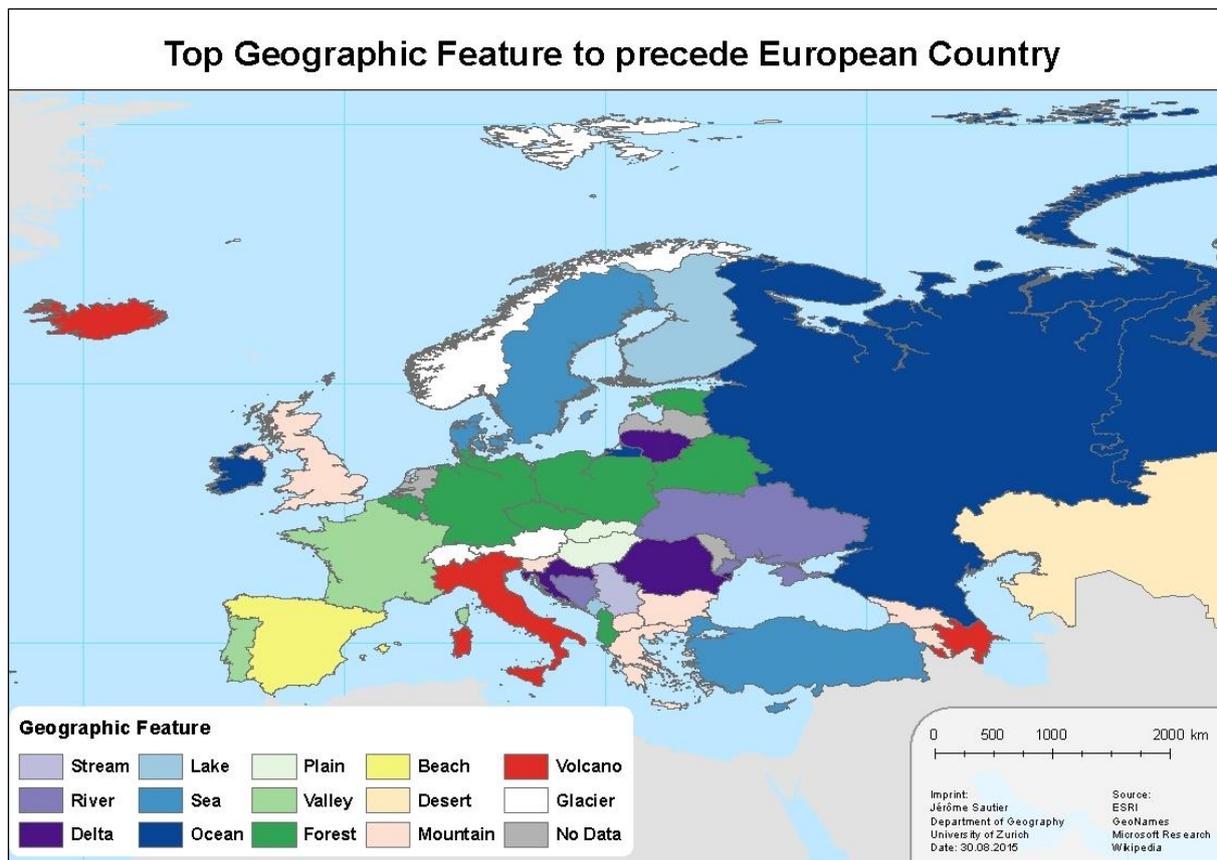


**Fig. 5.34: Conditional probability of the 20 most likely countries to follow "glacier in"**

The spatial distribution of all these results is manifested in figure 5.35. It illustrates the most likely geographic feature to precede a European country. This could be a possible indicator for prominent physical geographic features or attractions in a country. It can be perceived that Lithuania Ukraine, Romania, Serbia, Bosnia-Herzegovina and Croatia are heavily linked to streams, rivers or deltas. In the case of Serbia, this is caused by ambiguity. The reason being is that a lot of television, movie or sports streams are linked to Serbia. The association of Romania and the Ukraine are likely caused by the Danube Delta and its rivers.

The country Finland and Montenegro are most likely preceded by lake. The countries Denmark, Norway and Turkey are heavily linked to the sea. This is primarily caused by the fact that their actually situated at the sea. These corresponding seas also contain the word sea in them (e.g. the Mediterranean Sea/Black Sea near Turkey and Baltic Sea/North Sea for Denmark and Finland). On the other hand, Russia and Ireland are more likely to be linked to the ocean. This seems to be caused by the fact that Ireland is situated in the Atlantic Ocean and the coast of Russia lies at the Pacific and Arctic Ocean.

A cluster of countries linked to forests is seen in the middle of Europe which spans from Belgium to Belarus. Additionally, plains are the geographic feature most likely related to Hungary and Slovakia. For France and Portugal valley is the most likely geographic feature to precede them. Beaches are the most likely geographic feature in Spain, while the desert is prominent in Kazakhstan. Mountainous areas or famous mountains are to be expected in the United Kingdom and in the area of Greece. Volcanic regions, on the other hand, are expected in Iceland, Italy and Azerbaijan. Last of all, glaciers are the most well-known geographic feature in Switzerland, Austria and Norway.



**Fig. 5.35: Spatial distribution of most probable geographic feature to precede European country**

#### 5.4.1.2 Cities to follow Geographic Feature

The cities most likely to follow “river near” and their corresponding conditional probabilities are illustrated in figure 5.36. In total 44 cities were retrieved for the geographic feature river. It can be observed that river is most likely to be followed by the cities: Hastings, Koblenz, Florence, Rome, Paris and London. Most of these cities are popular European cities which have a river flowing through them. However, a quick Web search shows that the high conditional probability for the city Hastings is caused by geo/geo ambiguity. The screenshot of the Web search strengthen this (illustrate in figure 5.37). Most of the cities proposed by Bing refer to cities in the state of Michigan or Minnesota. Moreover, the rivers they refer to are the Mississippi River and the Thornapple River which are all located in the United States. The Mississippi River is one of the most well-known rivers in the United States and therefore heavily influences the results. Interestingly, the European city Hastings is also found on the search results from Bing. Even the Wikipedia page for the city Hastings in England is listed. However, no river is mentioned or associated with the European city Hastings.

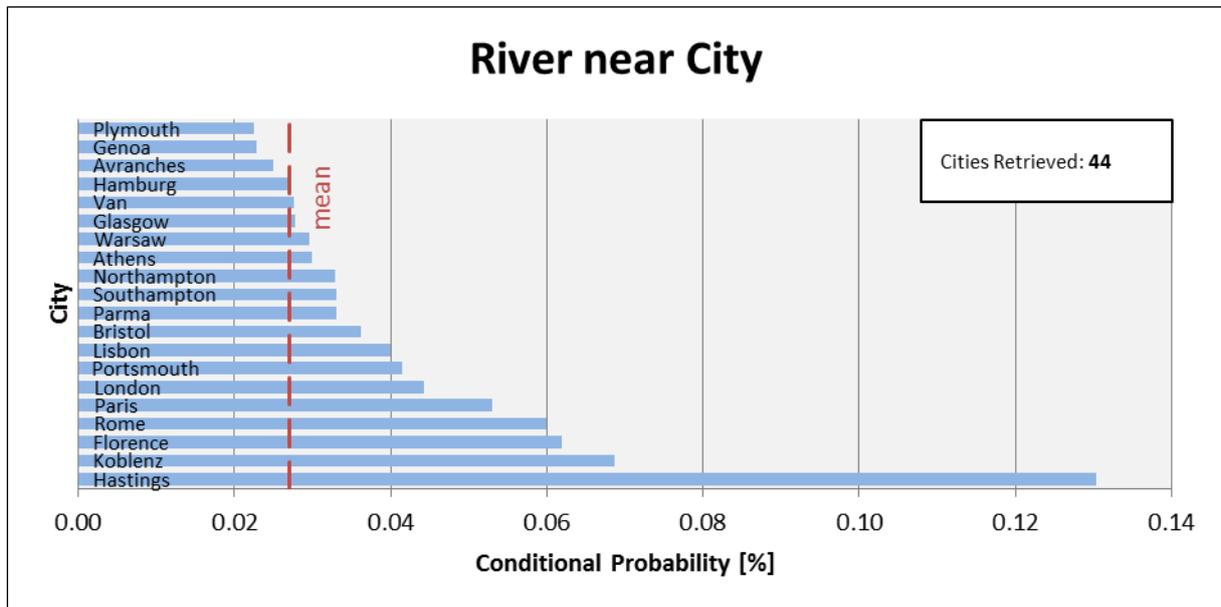


Fig. 5.36: Conditional probability of the 20 most likely cities to follow “river near”

526,000 RESULTS

**U Rent Em Canoe Livery | Canoe, Tube, Kayak Rentals in ...**  
[urentemcanoe.com/index.html](http://urentemcanoe.com/index.html)  
 Located on the scenic **Thornapple River** in downtown **Hastings**, U Rent Em Canoe Livery is West Michigan's oldest, largest and friendliest canoe, kayak and tube Memorial - Rates - Frequent Floater - QR Codes - Angler Update

**River in Hastings, Minnesota with Reviews & Ratings - ...**  
[www.yellowpages.com/hastings-mn/river](http://www.yellowpages.com/hastings-mn/river)  
 Find 553 listings related to **River in Hastings** on YP.com. See reviews, photos, directions, phone numbers and more for **River** locations in **Hastings, MN**.

**Mississippi River at Hastings - Lock and Dam 2 ...**  
[water.weather.gov/ahps2/hydrograph.php?wfo=mpx&gage=hstm5](http://water.weather.gov/ahps2/hydrograph.php?wfo=mpx&gage=hstm5)  
 NOTE: Forecasts for the **Mississippi River at Hastings - Lock and Dam 2** Tailwater are issued routinely during the navigation season, and as needed at other times of ...

**The Best Hastings Hotels - TripAdvisor**  
[www.tripadvisor.com/Hotels-g43146-Hastings\\_Minnesota-Hotels.html](http://www.tripadvisor.com/Hotels-g43146-Hastings_Minnesota-Hotels.html)  
 Book the Best **Hastings Hotels** on TripAdvisor: Find 58 traveler reviews, 24 candid photos, and prices for **hotels in Hastings**, Minnesota, United States.

**Hastings Scenic Mississippi & Vermillion River Bike Trails**  
[hastingsmn.org/recreation.html](http://hastingsmn.org/recreation.html)  
**Hastings** lies at the junction of the Mississippi, St. Croix, and Vermillion **rivers**, which serve as a stunning backdrop for hiking, biking ...

**Hastings - Wikipedia**, the free encyclopedia  
[en.wikipedia.org/wiki/Hastings](http://en.wikipedia.org/wiki/Hastings)  
 Coordinates. **Hastings** / ˈ h eɪ s t ɪ ŋ z / is a town and borough in the county of East Sussex, within the historic county of Sussex, on the south coast of England.  
 History - Government - Geography and climate - Demography - Economy

**Hastings Trails and Maps | TrailLink**  
[www.traillink.com/city/hastings-mi-trails.aspx](http://www.traillink.com/city/hastings-mi-trails.aspx)  
 Find **Hastings, Michigan** walking, running and bike trails with detailed information, reviews, photos and trail maps on TrailLink.

**Hastings on Hudson Restaurants, Westchester County ...**  
<https://www.zomato.com/westchester-county/hastings-on-hudson...>  
 Restaurants in **Hastings on Hudson**; **Hastings on Hudson**, Westchester County Restaurants - Menus, Reviews, Photos for Restaurants, Pubs, Lounges, and Bars in **Hastings** ...

**U Rent Em Canoe Livery | Canoe, Tube, Kayak Rentals in ...**  
[urentemcanoe.com/general.html](http://urentemcanoe.com/general.html)  
 Located on the scenic **Thornapple River** in downtown **Hastings Michigan**, U RENT EM CANOE LIVERY is Barry County's oldest and largest canoe rental and has been a ...

**Whispering Waters Campground in Hastings, Michigan**  
[whisperingwatersonline.com](http://whisperingwatersonline.com)  
 Whispering Waters **Campground in Hastings, Michigan**, Barry County. A member of ARVC Michigan (Association of RV, Parks and Campgrounds)

Fig. 5.37: First webpage recommendations on Bing for the query “river near Hastings”

The cities most probable to be associated with sea are displayed in figure 5.38. In total 81 cities were retrieved for the geographic feature sea. The cities with the highest conditional probability consist of: Great Yarmouth, Perpignan, Hastings, Stockholm, Brighton, Hayling Island, Baku and Istanbul. All of these cities are located in Europe and are also located near the sea. This time, the relation “sea near Hastings” is really referring to the town in the county of East Sussex, on the south coast of England.

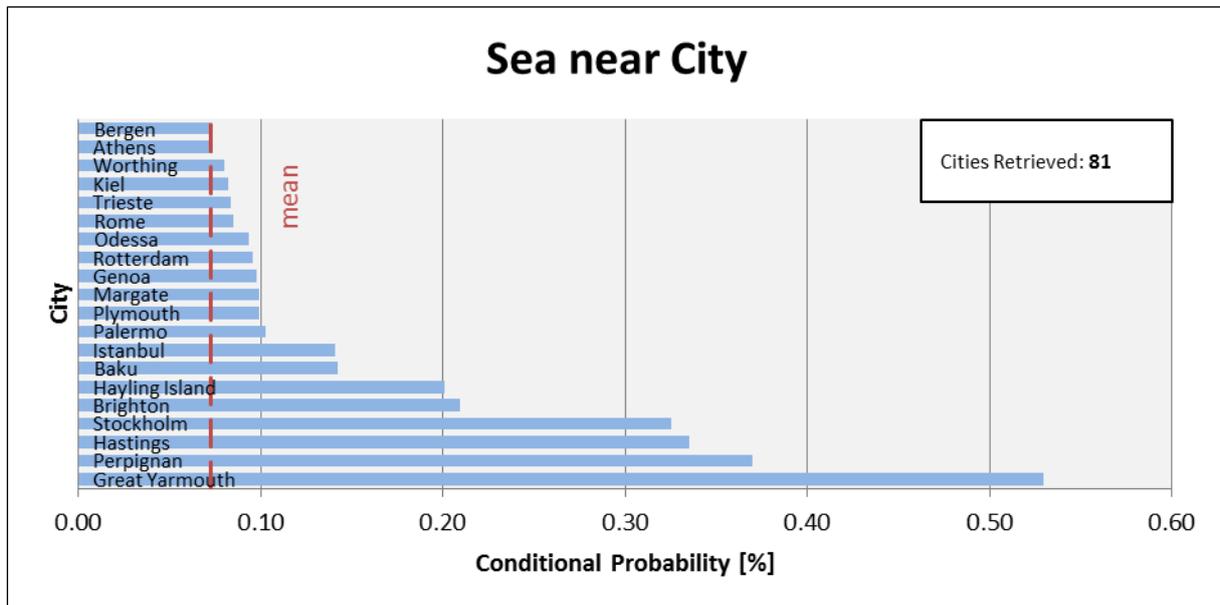


Fig. 5.38: Conditional probability of the 20 most likely cities to follow “sea near”

The top 20 cities preceding forest are available in figure 5.39. The cities Izmir and Smolensk stand out the most and are considerably more likely to follow “forest near” than the other cities. A screenshot of the Web search recommendations for these two cities is shown in figure 5.40. The Web search for “forest near Izmir” gave no conclusive results. Only that the term likely refers to a touristic area in Izmir containing the term forest. Therefore, a variety of hotels and restaurants near the area are proposed. However, Izmir has nothing to do with an actual forest. On the other hand, the Web search for “forest near Smolensk” produced some interesting results. The Katyn forest near Smolensk was used by the Soviet secret police for a mass execution of Polish nationals in 1940 (“Katyn massacre” 2015). This event is known as the Katyn massacre and is a possible cause for the strong relation of forest and Smolensk.

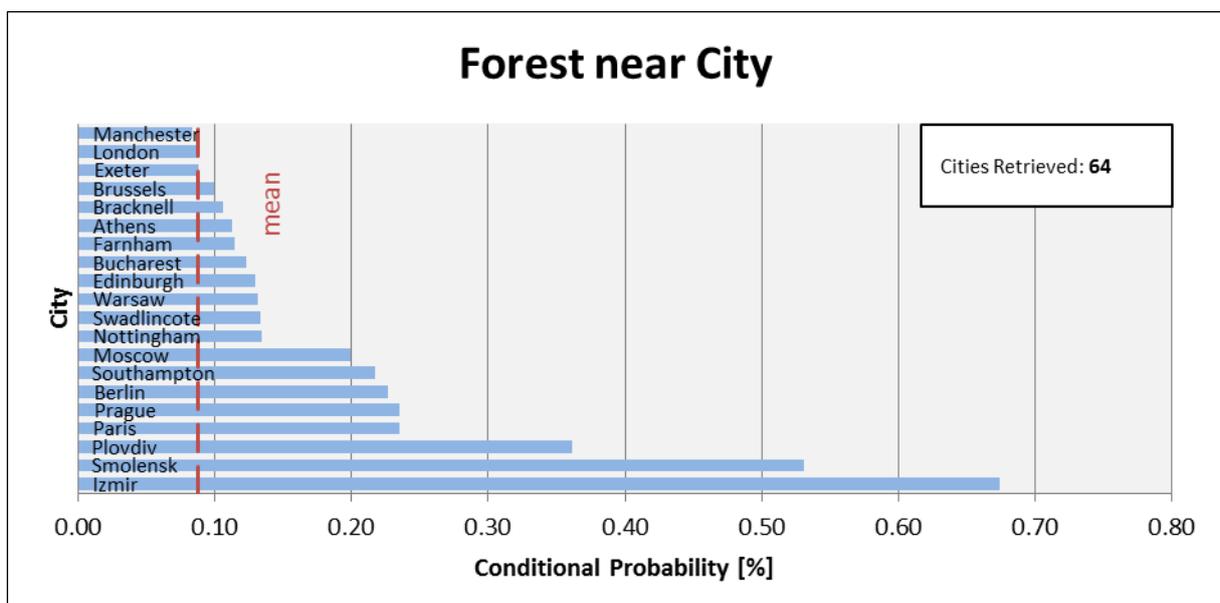


Fig. 5.39: Conditional probability of the 20 most likely cities to follow “forest near”

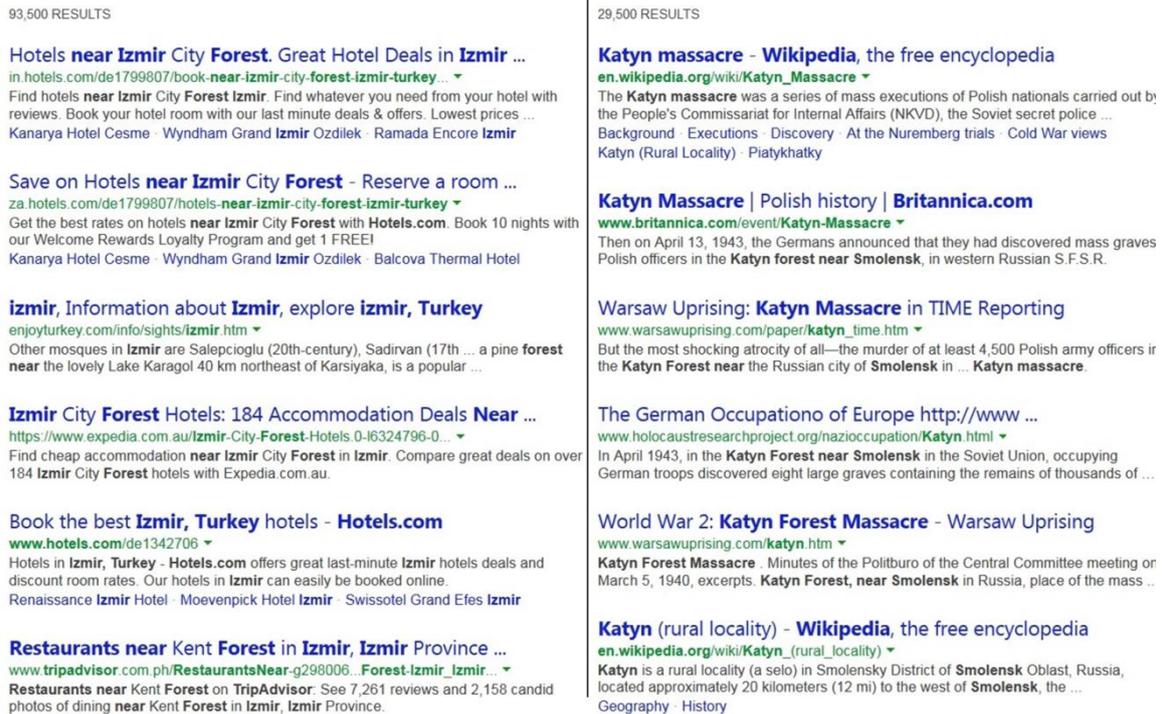


Fig. 5.40: First webpage recommendations on Bing for the query “forest near Izmir” (left) and “forest near Smolensk” (right)

The most likely cities referring to hill also provide some interesting results. The conditional probability of cites to follow “hill near” is visible in figure 5.41. Additionally, a grand total of 101 cities were retrieved for the doublet <hill><near>. The majority of these cities referring to hill are situated in the United Kingdom. This is possibly caused due to the fact that a variety of place names, streets and stations in the United Kingdom contain hill in their name. Notting Hill, for example is a district in the west of London. Furthermore, hill may also refer to a family name. Thus, ambiguity has a huge influence on the outcome of these results.

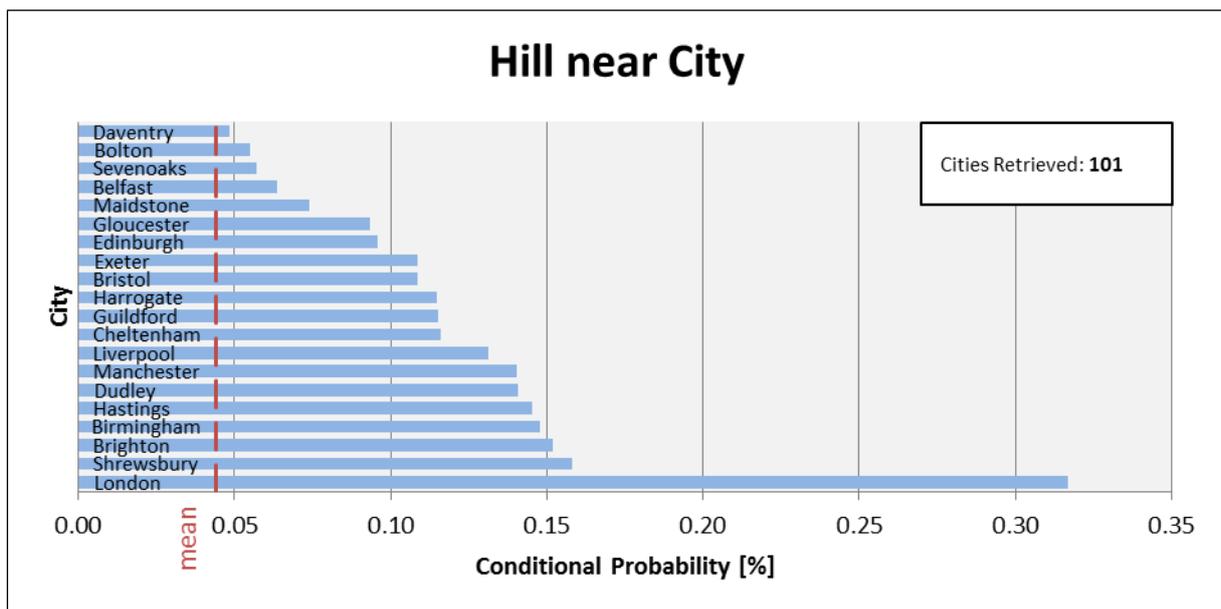


Fig. 5.41: Conditional probability of the 20 most likely cities to follow “hill near”

The most cities referring to mountain also have a real mountain near them. The overall distribution of these cities and their corresponding conditional probabilities is seen in figure 5.42. In total 24 cities were retrieved which followed the doublet <mountain><near>. The only suspicious entry is the high conditional probability of Prague, since it does not have any mountains near its surroundings. A quick Web search helps to comprehend why Prague has such a strong relation with mountain. The search showed that an historic battle took place near Prague known as the Battle of White Mountain (“Battle of White Mountain” 2008). This event involved no actual mountain. It was the name of a plateau used in the battle. Consequently, the historic event may cause the conditional probability of Prague to stick out compared to other cities.

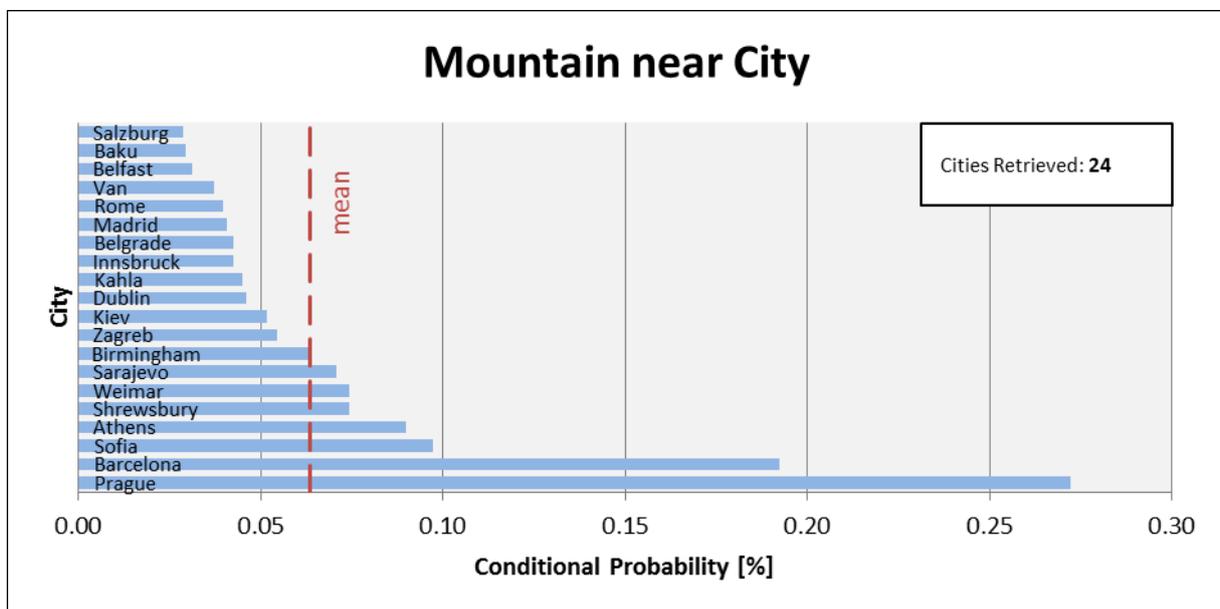


Fig. 5.42: Conditional probability of the 20 most likely cities to follow “mountain near”

The results for volcano followed by cities are illustrated in figure 5.43. In total only four cities were retrieved which mostly seem plausible. A number of volcanoes are situated in Iceland and are also near Reykjavik, while an active volcano with the name Mount Etna is situated between the cities Messina and Catania. The only city name being dubious is Cartagena. A Web search on “volcano in Cartagena” gives clarity as to why a high conditional probability is introduced. The results are seen in figure 5.44. The city Cartagena seems to be heavily linked with volcano, as it is famous for the volcanic mud baths or mud volcano. These are touristic attractions and increase the association of Cartagena and volcanoes. However, the city Cartagena is not referring to the city in Europe. It is rather referring to the city in Colombia and is another case of geo/geo ambiguity.

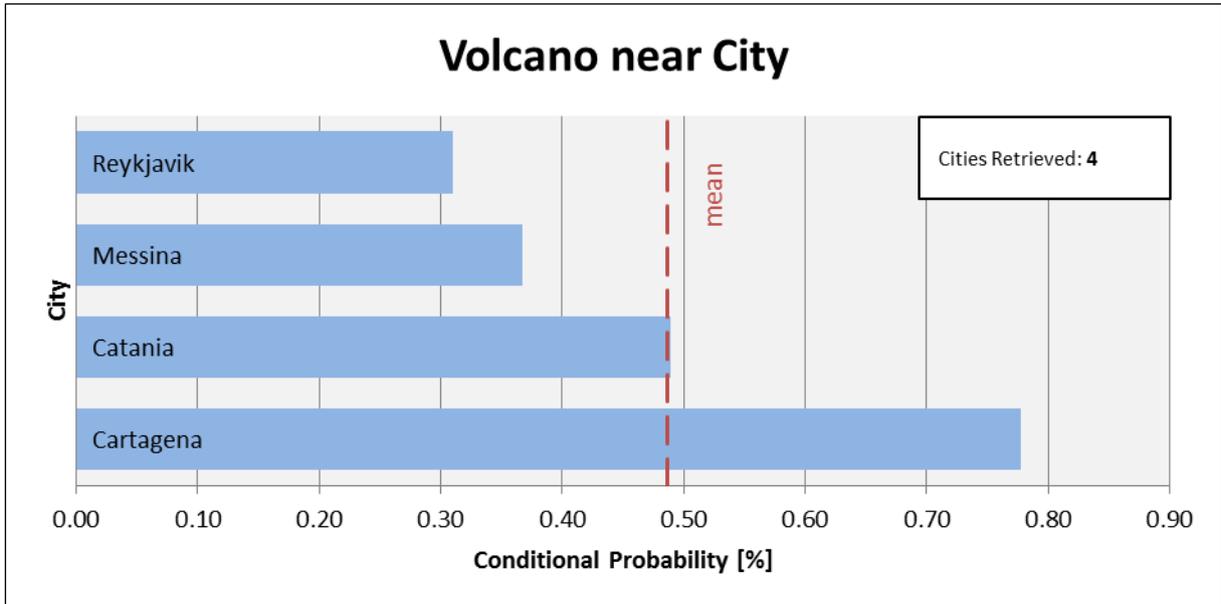


Fig. 5.43: Conditional probability of the 20 most likely cities to follow “volcano near”

39,400 RESULTS

**El Totumo volcano near Cartagena, Colombia - YouTube**

[www.youtube.com/watch?v=NBdGCr7L3I](http://www.youtube.com/watch?v=NBdGCr7L3I)  
 By jknudson · 3 min · 2.1M views · Added 6/5/2007  
 Video embedded · Mud Bath in the volcano...They say the volcano is a couple hundred meters deep but the mud is so thick that you just float on top, even if you stand up ...

**Volcan de Lodo El Totumo (Mud Volcano) (Cartagena ...**

[www.tripadvisor.com/Attraction\\_Review-g297476-d1223572-Reviews...](http://www.tripadvisor.com/Attraction_Review-g297476-d1223572-Reviews...)  
 ★★★★★ Rating: 4/5 · 679 reviews · 269 photos · Address: , Cartagena, Colombia  
 8/2/2015 - Book your tickets online for Volcan de Lodo El Totumo (Mud Volcano), Cartagena. See 679 reviews, articles, and 269 photos of Volcan de Lodo El Totumo (Mud ...

**VOLCANO & MANGROVE TOURS - Cartagena Connections**

[www.cartagenaconnections.com/volcano-mangrove-tours.html](http://www.cartagenaconnections.com/volcano-mangrove-tours.html)  
 PRICES/SCHEDULE A) The WORKS (Volcano, Lunch and Swim at gorgeous secluded beach, Mangrove Tour) 90,000 pesos (approx USD \$45) 8.30am - 4.30pm

**El Totumo volcano near Cartagena, Colombia - YouTube**

[www.youtube.com/watch?v=3O9MQkmYZoo](http://www.youtube.com/watch?v=3O9MQkmYZoo)  
 By Volcano Videos · 3 min · 99 views · Added 2/19/2015  
 Video embedded · Please Subscribe to my Channel :) - - 2014 - 2015 volcano, volcanoes, volcano 2015, volcano eruption, volcano full movie, volcano eruption ...

**Volcanic Mud Baths of El Totumo - TravelMuse**

[www.travelmuse.com/articles/off-beat/totumo-volcano-mud-bath](http://www.travelmuse.com/articles/off-beat/totumo-volcano-mud-bath)  
 Volcanic Mud Baths of El Totumo. Our editor gets covered in mud during a cultural spa experience to the El Totumo volcano near Cartagena, Colombia.

**A Tour to El Totumo Mud Volcano near Cartagena**

[www.uncovercolombia.com/en/.../item/...mud-volcano-near-cartagena.html](http://www.uncovercolombia.com/en/.../item/...mud-volcano-near-cartagena.html)  
 If you are looking for things to do in Cartagena during your holiday in Colombia, a tour to El Totumo Volcano could be one good option for a day trip from Cartagena

**Volcan de Lodo El Totumo (Mud Volcano) (Cartagena ...**

[www.tripadvisor.ca/Attraction\\_Review-g297476-d1223572-Reviews...](http://www.tripadvisor.ca/Attraction_Review-g297476-d1223572-Reviews...)  
 ★★★★★ Rating: 4/5 · 692 reviews · 270 photos · Address: , Cartagena, Colombia  
 Volcan de Lodo El Totumo (Mud Volcano), Cartagena. JOIN; LOG IN · USD  
 Cartagena. ... Near Volcan de Lodo El Totumo (Mud Volcano) Top-rated ...

**Totumo Volcano and Mud Baths Day Trip from Cartagena ...**

[www.lonelyplanet.com/.../totumo-volcano-mud-baths-day-trip-cartagena](http://www.lonelyplanet.com/.../totumo-volcano-mud-baths-day-trip-cartagena)  
 Book your adventure - Escape Cartagena for the day and discover the healing powers of the mud baths at Totumo Volcano! Climb up and inside this small volcano and dive ...

**The Mud Volcano near Cartagena - In Search of Wonder**

[www.insearchofwonder.com/wonder-found-el-totumo-the-mud-volcano](http://www.insearchofwonder.com/wonder-found-el-totumo-the-mud-volcano)  
 Couples travel blog guide to visiting the mud volcano near Cartagena, Colombia. One of the most unique experiences of our lives.

**Near Cartagena, Colombia, the Mud Volcano, a ...**

[www.cartagenainfo.net/noticias/MudVolcano/index.html](http://www.cartagenainfo.net/noticias/MudVolcano/index.html)  
 The next few hours at the mud volcano are weird. Very weird. Next to an otherwise flat(ish) meadowy marshland this 15 m mud cone rises from the shoreline of Ciénaga ...

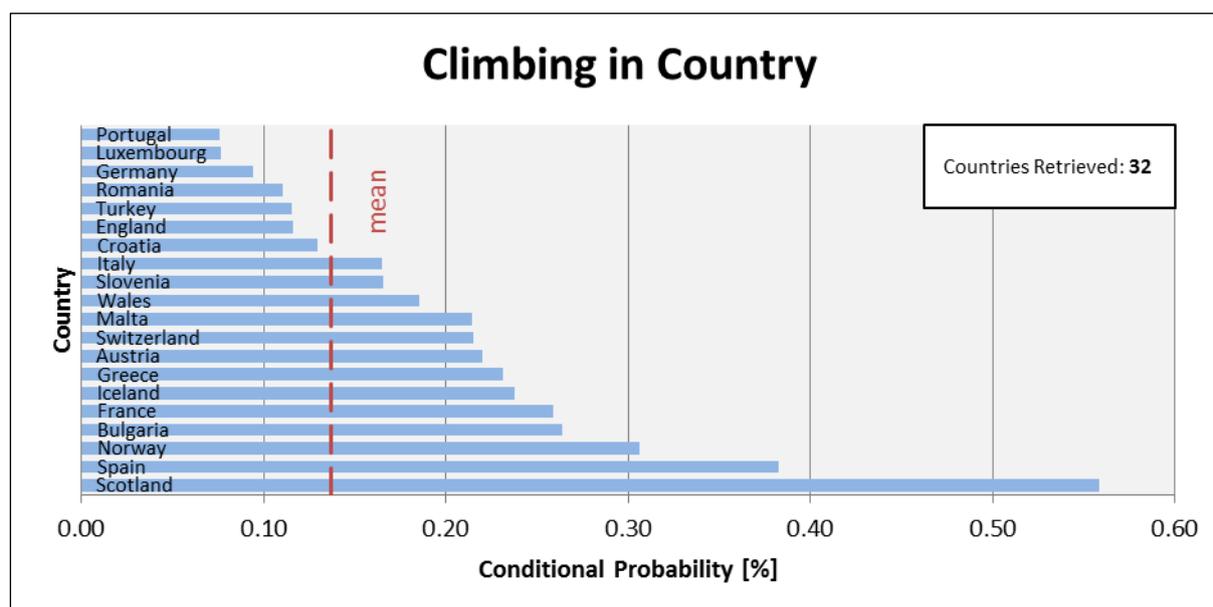
Fig. 5.44: First webpage recommendations on Bing for the query “volcano near Cartagena”

## 5.4.2 Sports Activities

The place names to follow a sport activity are explored in this section. This is done for a proportion of the results on country and city level. The appendix C.6 and C.7 can be examined for further examples.

### 5.4.2.1 Countries to follow Sport Activity

The countries most likely to follow “climbing in” and their corresponding conditional probabilities are displayed in figure 5.45. In total 32 countries were retrieved for climbing in the 1000 autocompletes. The countries heavily related with climbing are: Scotland, Spain, Norway, Bulgaria and France. These countries also seem to be heavily associated with mountains in section 5.4.1.1. This seems only logical, since mountains or indoor halls have to be available for climbing.



**Fig. 5.45: Conditional probability of the 20 most likely countries to follow “climbing in”**

The distribution of the countries conditional probability for marathon is illustrated in figure 5.46. The countries which are mostly associated with marathons are: Lithuania, Greece, Russia, Germany and Spain. In this distribution Lithuania appears to have a considerably high conditional probability. This high probability is inspected by querying for the respective triplet on the Web search engine Bing. The proposed webpages can be seen in figure 5.47. The Web results show that there is quite a famous marathon in Lithuania known as the Danske Bank Vilnius Marathon. This marathon was introduced in 2001 and has gained popularity in recent years. Consequently, the conditional probability is influenced and higher compared to the other countries.

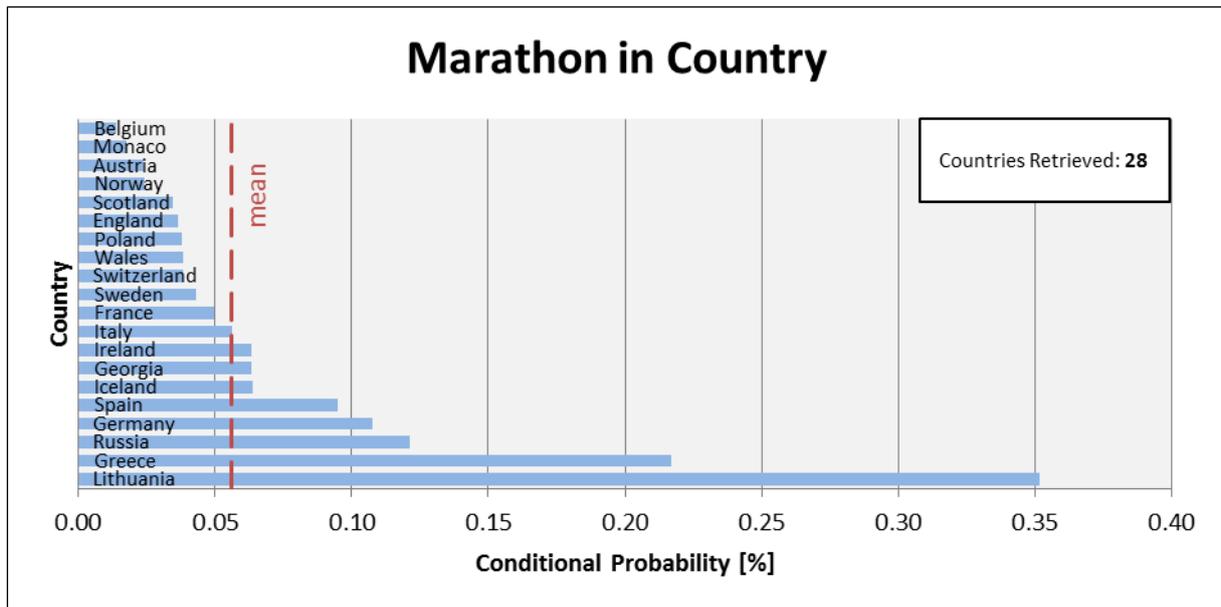


Fig. 5.46: Conditional probability of the 20 most likely countries to follow “marathon in”

1,260,000 RESULTS

**Vilnius Marathon - Official Site**  
[www.vilniusmaratonas.lt/en](http://www.vilniusmaratonas.lt/en)  
 Danske Bank Vilnius maratonas ... Photos from „Danske Bank Vilnius maratonas” 2015 ... We invite you to participate in the biggest running event in Lithuania ...  
 Gallery · Hotels · Europa City · Lithuania Post Relay · Vilnius Marathon

**2015 - 2016 Lithuania Marathon Calendar**  
[marathons.ahotu.com/calendar/marathon/lithuania](http://marathons.ahotu.com/calendar/marathon/lithuania)  
 Browse our Marathon Calendar for races organized in Lithuania among 3 races.

**Vilnius Maratonas - Vilnius, Lithuania, Sep 13 2015**  
[marathons.ahotu.com/event/vilnius-maratonas](http://marathons.ahotu.com/event/vilnius-maratonas)  
 Detailed information on Vilnius Maratonas, provided by ahotu Marathons with news, interviews, photos, videos, and reviews.

**News about Marathon In Lithuania**  
[bing.com/news](http://bing.com/news)

All around town: US Soldiers take part in **Marathon** in ...  
 United States Army · 4 days ago  
 All around town: US Soldiers take part in **marathon** in Lithuania. A colorful mix of clothing worn by the participants in the 12th annual Vilnius Marathon...

**International Vilnius Marathon 2015 - Race Details ...**  
[www.marathonrunnersdiary.com/.../europe-marathons/vilnius-marathon.php](http://www.marathonrunnersdiary.com/.../europe-marathons/vilnius-marathon.php)  
 The first Vilnius Marathon was run in 2001. Find out more and register for the International Vilnius Marathon.

**Category:Marathons in Lithuania - Wikipedia**, the free ...  
[en.wikipedia.org/wiki/Category:Marathons\\_in\\_Lithuania](http://en.wikipedia.org/wiki/Category:Marathons_in_Lithuania)  
 Pages in category “Marathons in Lithuania” This category contains only the following page. This list may not reflect recent changes ...

All around town: US Soldiers take part in **Marathon** in ...  
[www.army.mil/article/155364/All\\_around\\_town\\_US\\_Soldiers\\_take\\_part...](http://www.army.mil/article/155364/All_around_town_US_Soldiers_take_part...)  
 9/13/2015 · All around town: US Soldiers take part in **marathon** in Lithuania. Before the 12th annual Vilnius Marathon began, volunteers from local schools participated ...

**Lithuania Race Calendar 2015/2016 - Run Infinity**  
[runinfinity.com/calendar/lithuania](http://runinfinity.com/calendar/lithuania)  
 Lithuania Race Calendar 2015/2016. ... Held in the city of Vilnius, the capital of Lithuania, the Vilnius Marathon is Organized by Municipality of Vilnius city, ...

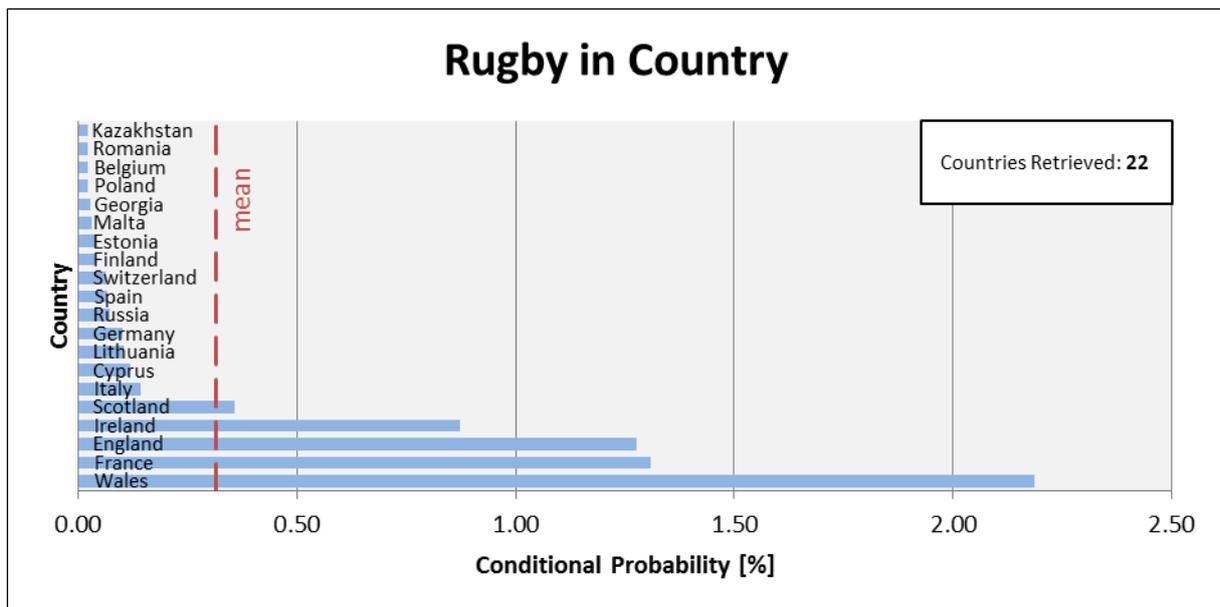
**International Vilnius Marathon - Race Details**  
[www.marathonguide.com/races/racedetails.cfm?MIDD=2563130915](http://www.marathonguide.com/races/racedetails.cfm?MIDD=2563130915)  
 International Vilnius Marathon Information by MarathonGuide.com - the complete marathon resource and community. Complete directory of marathons, marathon ...

**News: All around town: US Soldiers take part in Marathon ...**  
<https://www.dvidshub.net/news/175885/all-around-town-us-soldiers...>  
 9/13/2015 · VILNIUS, Lithuania – The 12th annual Vilnius Marathon was run today that wound it's way through 13 miles of downtown and Old Vilnius center where U.S ...

**Vilnius Marathon 2014/2015 - Date, Registration, ...**  
[runinfinity.com/race/vilnius-marathon](http://runinfinity.com/race/vilnius-marathon)  
 Vilnius Marathon is held in the city of Vilnius, the capital of Lithuania. The marathon starts and finishes at Cathedral Square.

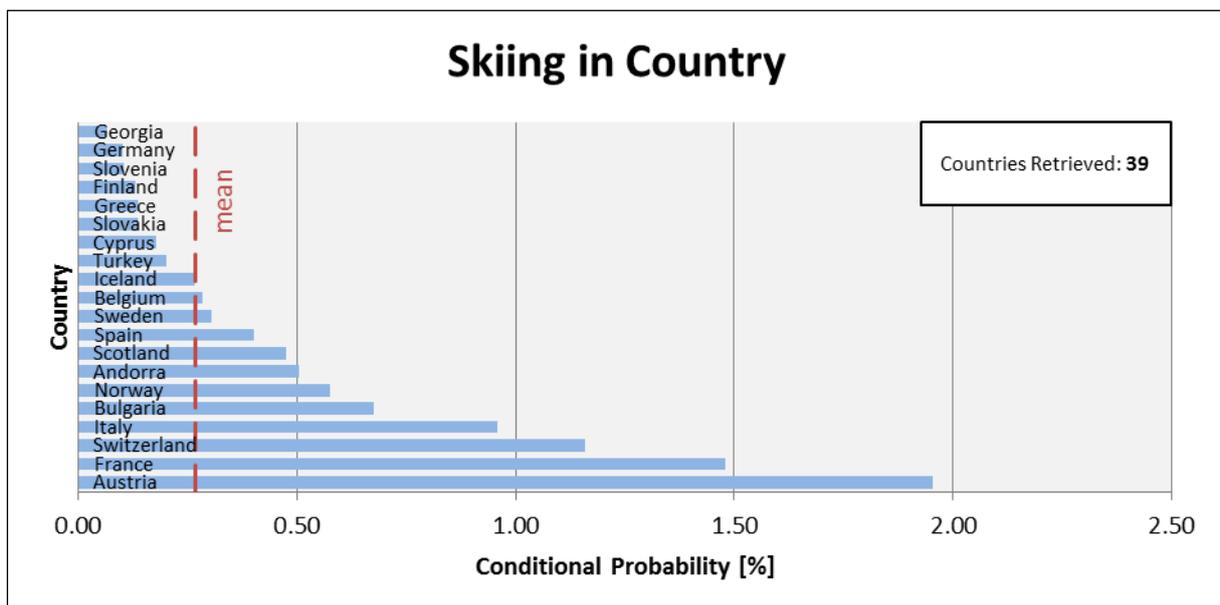
Fig. 5.47: First webpage recommendations on Bing for the query “marathon in Lithuania”

The countries most likely to be associated with rugby are: Wales, France, England, Ireland, and Scotland. This can be observed in figure 5.48. The results make sense, because the first six countries participate in an annual international rugby union championship in Europe. This tournament is known as Six Nations Championship (“RBS 6 Nations” 2015). As the name suggests, the championship is held between the six biggest rugby union nations in Europe: England, France, Ireland, Italy, Scotland and Wales. This coincides with the conditional probabilities received from the Microsoft Web N-Gram Service, because the top six countries related to rugby are the identical with the teams participating in the Six Nations.



**Fig. 5.48: Conditional probability of the 20 most likely countries to follow “rugby in”**

The countries most likely related to skiing are: Austria, France, Switzerland and Italy (seen in figure 5.49). These countries are also heavily linked to mountains or even glaciers. This makes total sense, since the countries are part of the Alps. Furthermore, snow is necessary to be able to ski which is usually available in high and cold regions. Overall, the most popular ski areas and resorts are situated in these countries. Therefore, the relation between skiing and European countries is well represented in the Microsoft Web N-Gram Service.



**Fig. 5.49: Conditional probability of the 20 most likely countries to follow “skiing in”**

The countries most likely to be known for tennis are the following: England, Spain, France, Scotland, Serbia and Germany. These results are likely caused by tennis players or tennis tournaments associated with a country. The high conditional probability of England is likely caused by Wimbledon, the oldest tennis tournament in the world. The high probability of France is likely also caused by a tournament, namely the French Open or Roland Garros. The other high probabilities are most likely caused by current famous players: Rafael Nadal (Spain), Andy Murray (Scotland) and Novak Djokovic (Serbia). Switzerland has a strangely low probability, even though they have a famous tennis player in Roger Federer.

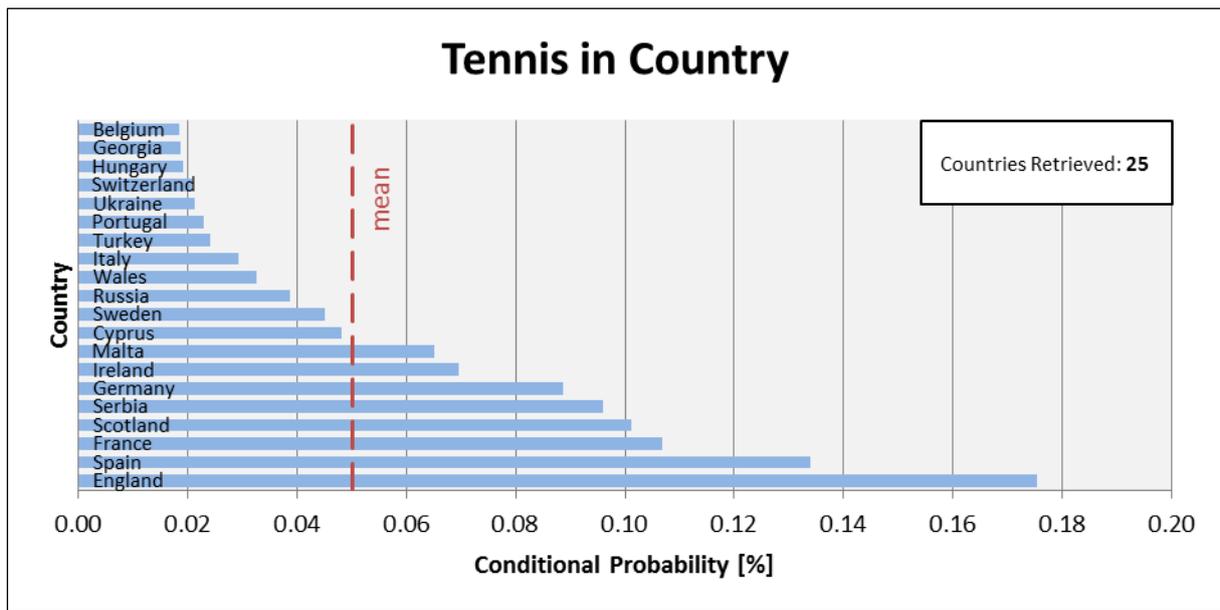


Fig. 5.50: Conditional probability of the 20 most likely countries to follow “tennis in”

The spatial distribution of all sport activities per country is illustrated in figure 5.51. For each country the most likely sport activity to precede it is demonstrated. Each sport activity is normalized for a country to ensure that common sports on the Web do not influence the spatial distribution of top sport activities. Overall, the spatial distribution of most likely sports activities to precede a country may be an indicator for the most popular sport in a country or the most suited sport activity for a country. This may also involve well-known athletes or teams of a country which excel in specific sport.

At first glance, the clusters in the Alps and East Europe are detected. The Alps seem to be a very popular place for snowboarding and consequently the countries Austria, France, Italy and Switzerland are most likely associated with snowboarding. The other cluster involves handball which seems to be the most popular sport in East Europe and Denmark. Additionally, the countries near the sea seem to be popular for sports bound to water. For example, Greece and Croatia seem to be hotspots for sailing. Cyprus is also most likely to be

followed by snorkeling which requires the sea. Finally, a small cluster for golf is observable in Spain and Portugal.

The other sport activities are more heavily scattered and are mostly situated in one or two countries. For instance, the United Kingdom and Ireland are heavily divided. Ireland is most likely referred to the sport hurling which is a very local sport of Irish and Gaelic origin. Football is the sport most likely related to Northern Ireland. However, it seems that in Northern Ireland term soccer is more commonly used the instead of football. Rugby is probably the most popular sport in Wales, while cricket is preferred in England. Lastly, Scotland is most likely associated with mountaineering.



**Fig. 5.51: Spatial distribution of most probable sport activity to precede European country**

#### 5.4.2.2 Cities to follow Sport Activity

The cities most likely related to football are: London, Barcelona, Manchester and Sheffield. The remaining top 20 cities linked to football can be seen in figure 5.52. In total 143 cities were linked with football. The most likely cities associated with football contain prominent football clubs. Several football clubs are located in London: Arsenal, Chelsea, Tottenham Hotspur, West Ham and many more. Manchester and Barcelona are also well known for their popular and successful football clubs. Sheffield also contains a strong relation with football,

because it is one of the oldest football clubs in history. Nevertheless, one of the relations seems odd. The city Javea is to the authors' knowledge not a famous for football. Moreover, the clarification through a Web search failed. The only thing which can be said is that the city Javea has a football club. However, the cause for such a strong bond between football and Javea on the Web is unexplained.

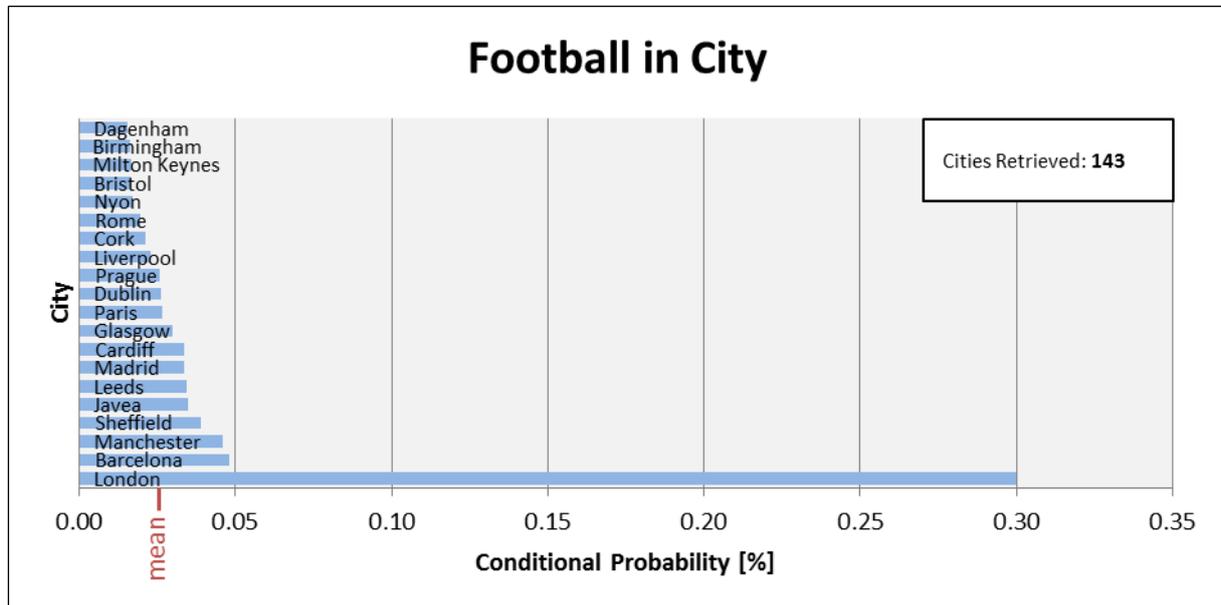


Fig. 5.52: Conditional probability of the 20 most likely cities to follow “football in”

The distribution of the cities conditional probability for parkour is displayed in figure 5.53. Under the most frequent cities associated with parkour are several populated urban cities. The cities Liverpool, Istanbul, Paris and Mardin are remarkably linked to parkour compared to the other cities.

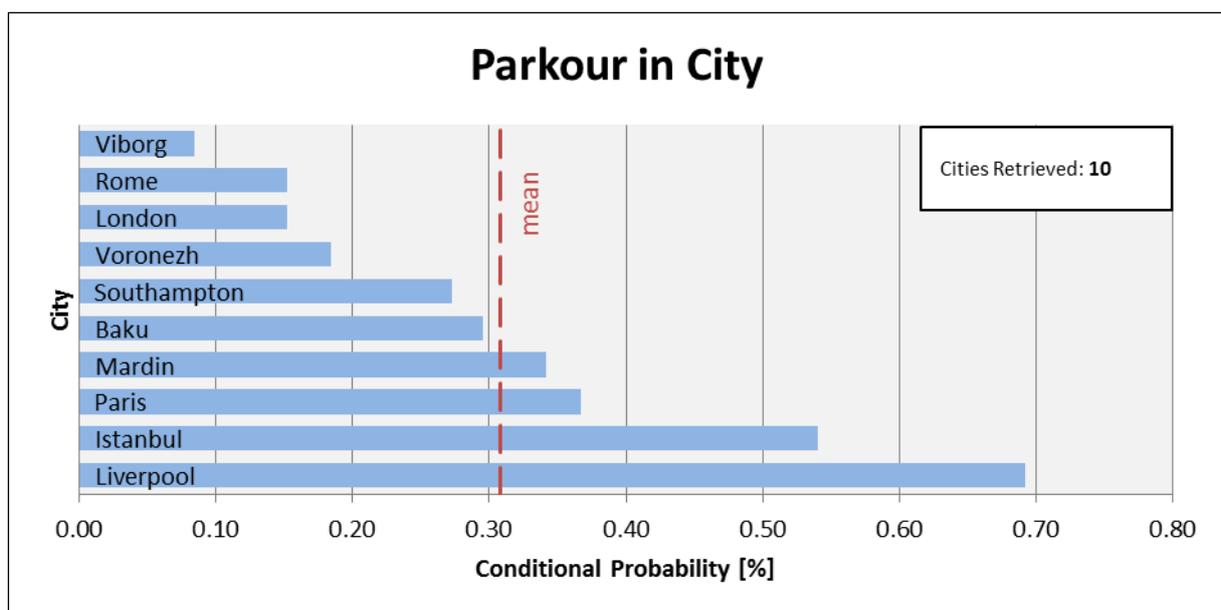


Fig. 5.53: Conditional probability of the 20 most likely cities to follow “parkour in”

The high conditional probability of Mardin is possibly caused by one of the most well-known free runners Ryan Doyle who promoted the cities Parkour possibilities. The Web search results can be observed in figure 5.54. The prominence of Mardin and parkour is also strengthened by the fact that Red Bull helped in making a video of Ryan Doyle doing a parkour in Mardin. Hereafter, the city became known to a broader audience and the popularity of Mardin rose as a parkour area.

8,830 RESULTS

**Ryan Doyle parkour in Mardin - YouTube**  
[www.youtube.com/watch?v=IUTXXMdQnio](http://www.youtube.com/watch?v=IUTXXMdQnio)  
 By Red Bull · 3 min · 810K views · Added 8/11/2011  
 Video embedded · Climb to new heights and visit <http://win.gs/1aXUVRM> for more parkour! Follow one of the world's most creative and innovative free runners and parkour ...

**Ryan Doyle parkour in Mardin Kurdistan - YouTube**  
[www.youtube.com/watch?v=5aP1v8wLgKo](http://www.youtube.com/watch?v=5aP1v8wLgKo)  
 By sami ebd0 · 3 min · 33 views · Added 2/19/2015  
 Video embedded · training by sami ebd0

**Parkour in Mardin – Fubiz Media**  
[www.fubiz.net/2011/08/25/parkour-in-mardin](http://www.fubiz.net/2011/08/25/parkour-in-mardin)  
 Une excellente initiative de la part de Red Bull avec la mise en scène de la discipline du parkour par l'un des plus doués, Ryan Doyle. Une vidéo qui s

Videos of **parkour in mardin**  
[bing.com/videos](http://bing.com/videos)

 Ryan Doyle parkour in YouTube	 Ryan Doyle parkour in Dailymotion	 Ryan Doyle: Parkour in YouTube	 Ryan Doyle parkour in FLIXYA
--	--	---	---

See more videos of **parkour in mardin**

**Parkour in Mardin - Rojaks Site**  
[www.rojaksite.com/parkour-in-mardin](http://www.rojaksite.com/parkour-in-mardin)  
 Follow one of the world's most creative and innovative free runners and parkour artists, Ryan Doyle, as he takes his skills to the next level in Mardin.

**Parkour In Mardin - The Awesomer**  
[theawesomer.com/parkour-in-mardin/120101](http://theawesomer.com/parkour-in-mardin/120101)  
 ★★★★★ Rating: 1/1 · 51 ratings  
 Freerunner Ryan Doyle takes us on an enchanting and enthralling parkour run through the exotic locale of Mardin, located in Turkey. Have you ever seen pigeons do ...

**Ryan Doyle - Parkour in Mardin/Turkey ...** Translate this page  
[www.whudat.de/ryan-doyle-parkour-ind-mardinturkey-clip](http://www.whudat.de/ryan-doyle-parkour-ind-mardinturkey-clip)  
 Einer der größten Parkour-Runner unserer Zeit und nimmt seinen Job ziemlich ernst: Ryan Doyle. In diesem Clip läuft er durch die historische Kulisse von Mardin.

**Parkour In Mardin Turkey - Ryan Doyle - Unfinished Man**  
[www.unfinishedman.com/parkour-in-mardin-turkey-ryan-doyle](http://www.unfinishedman.com/parkour-in-mardin-turkey-ryan-doyle)  
 Parkour athlete and Prince Charles look-alike Ryan Doyle is the star of this video and shows off his Parkour moves in Mardin, Turkey.

**TasteLikePizza - Parkour in Mardin**  
[tastelikepizza.com/item/2011/08/parkour-in-mardin](http://tastelikepizza.com/item/2011/08/parkour-in-mardin)  
 Parkour in Mardin - We surf the internet for fun video's and images so you don't have to

**Red Bull Films: Parkour in Mardin, Turkey | Digital ...**  
[www.digitalbuzzblog.com/redbull-films-parkour-in-mardin-turkey](http://www.digitalbuzzblog.com/redbull-films-parkour-in-mardin-turkey)  
 ★★★★★ Rating: 5/5 · 1 review  
 Here is a little eye-candy for Monday, with the latest Parkour short film from Red Bull! Now I must admit, I really don't understand Parkour, actually I just don ...

**Ryan Doyle Parkour and Free Running**  
[www.redbull.com/en/athletes/1331578990766/ryan-doyle](http://www.redbull.com/en/athletes/1331578990766/ryan-doyle)  
 Watch what happens when parkour meets freeski. Freeski Freerunner Ryan Doyle puts the GB Freeski team through their paces during a parkour masterclass.

Fig. 5.54: First webpage recommendations on Bing for the query “parkour in Mardin”

The distribution of the 20 most likely cities to follow “sailing in” is illustrated in figure 5.55. In total 41 countries were retrieved for this topic. Most of these cities are situated directly at the sea which makes the results more plausible. The cities Weymouth and Dun Laoghaire are well known sailing spots filled with yacht clubs (“ISAF Sailing Worldcup” 2015; “VOLVO Dun Laoghaire Regatta” 2015). Both of those locations also have well-known boat races. Weymouth and Portland is one of the venues for the ISAF Sailing World Cup, while in Dun Laoghaire the VOLVO Dun Laoghaire Regatta takes place (“ISAF Sailing Worldcup” 2015; “VOLVO Dun Laoghaire Regatta” 2015).

The results for skiing followed by cities are displayed in figure 5.56. In total 41 cities were retrieved. The majority of those cities are also located in countries which are associated with skiing (see figure 5.49). Additionally, a high number of Austrian cities have a conditional probability above average. This coincides with Austria having the highest conditional probability in countries following skiing. Sochi also has a high conditional probability. It is

likely caused due to the fact that the 2014 Winter Olympics took place in Sochi. The city with the most likely associated with skiing is Zakopane. The probabilities are mostly due to it being a popular destination for skiing and having hosted several Nordic skiing and ski jumping competitions (“Zakopane” 2015).

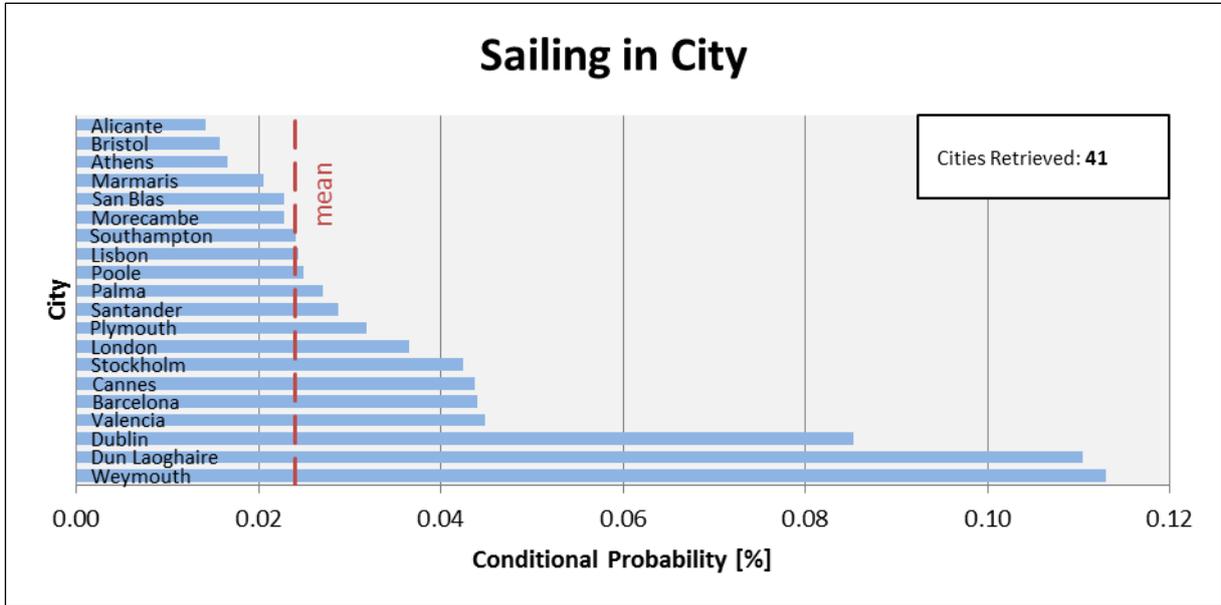


Fig. 5.55: Conditional probability of the 20 most likely cities to follow “sailing in”

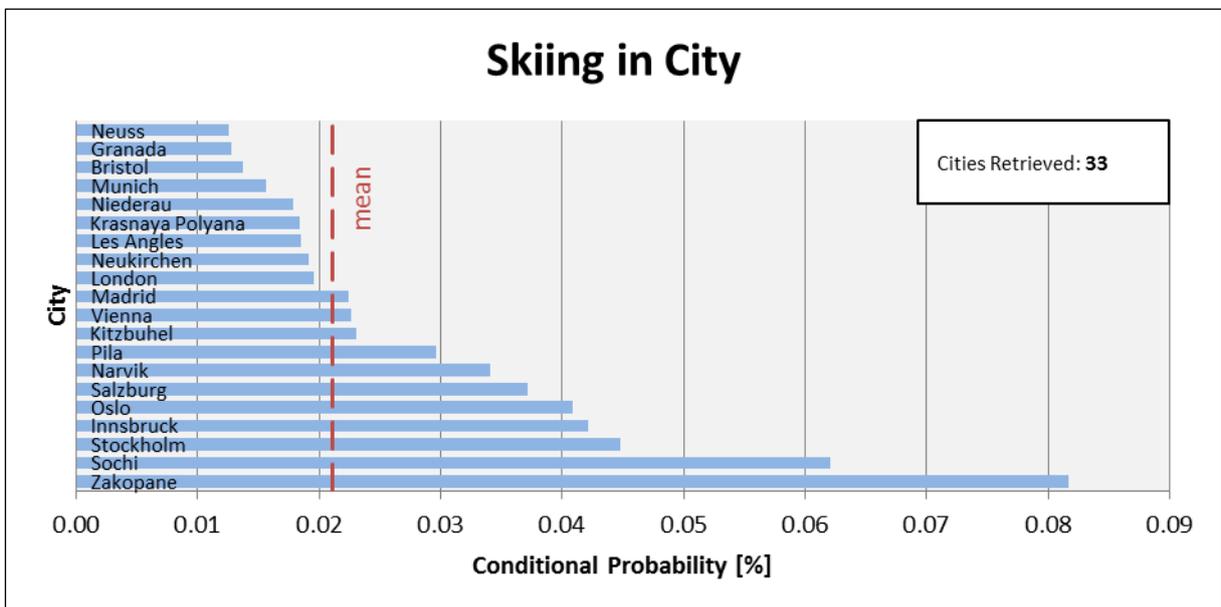


Fig. 5.56: Conditional probability of the 20 most likely cities to follow “skiing in”

The cities most likely to be associated with skydiving are shown in figure 5.57. In total only 7 cities have been linked with skydiving. These consist of: Manchester, Empuriabrava, Helsinki, Prague, Birmingham, Scalea and Dublin. The possible causes for the two highest conditional probabilities are investigated through a Web search on Bing. The webpage recommendations are manifested in the Web screenshot (seen in figure 5.58). These

demonstrate that near Manchester an indoor skydiving facility is available. Furthermore, tandem skydives and outdoor skydiving is offered. Nevertheless, the world's "most popular" skydiving zone is in Empuriabrava, Spain. This is claimed by the recommended webpages which also offer skydiving courses in Empuriabrava. Finally, these circumstances are the likely causes for the strong relations between skydiving and Manchester/Empuriabrava.

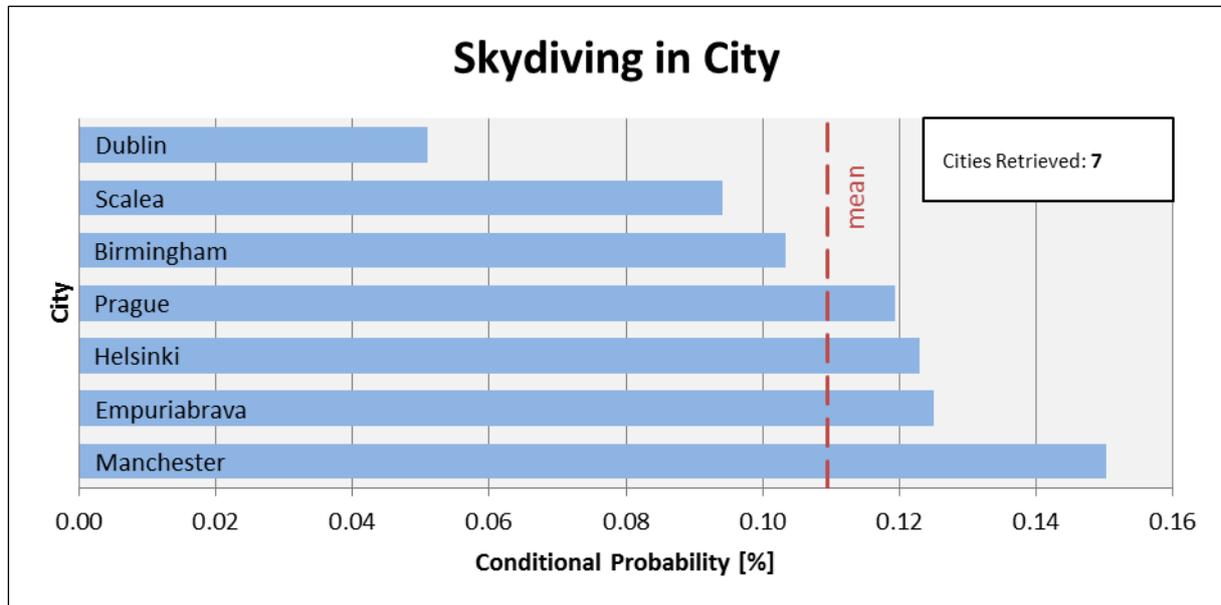


Fig. 5.57: Conditional probability of the 20 most likely cities to follow "skydiving in"

347,000 RESULTS

**Airkix: Indoor Skydiving** at Basingstoke, Milton Keynes ...

[www.airkix.com](http://www.airkix.com)

Located in Basingstoke, **Manchester** & Milton Keynes, Airkix brings in exciting opportunity for **indoor skydiving**, ideal for individuals as well as groups and corporate ...

[Manchester](#) · [Milton Keynes](#) · [Basingstoke](#) · [Contact Airkix](#) · [Vouchers](#) · [Pix](#)

**Airkix Indoor Skydiving in Manchester**

[www.airkix.com/plan/manchester.aspx](http://www.airkix.com/plan/manchester.aspx)

Get ready to try out Airkix's **indoor skydiving** centre in **Manchester**. Perfect for thrill seekers, professional **skydivers** and a great day out with family and friends!

[Cancellation Policy](#) · [Schools](#) · [Basingstoke](#) · [Milton Keynes](#) · [Plan a Visit](#)

**skydiving** near **Manchester** Call 01948 841111

[manchesterskydiving.co.uk](http://manchesterskydiving.co.uk)

Our **Skydiving** Centre is just over an hour from **Manchester**. Located less than 30 miles south of Chester. **Skydive** experiences and lessons close to you. Call us on 01948 ...

**Skydive Manchester. Tandem Skydives / Skydiving** near ...

[www.tandem-sky-dive.co.uk/manchester-tandem-skydive.html](http://www.tandem-sky-dive.co.uk/manchester-tandem-skydive.html)

**Skydive Manchester**. Try **Skydiving**! If you live in or near **Manchester** then Tandem **Skydive** UK is so easy to get to. Book your **skydiving** experience today.

**Manchester Skydiving** - The **Manchester Skydiving** ...

[www.skydivingmanchester.com](http://www.skydivingmanchester.com)

**Manchester Skydiving** delivers an adrenaline rush unlike any other experience in New Hampshire! Try **Skydiving in Manchester**... The Ultimate Adventure!

101,000 RESULTS

**Skydive Empuriabrava - Welcome to The Land of the Sky**

[www.skydiveempuriabrava.com/en](http://www.skydiveempuriabrava.com/en)

Intro. WELCOME TO THE LAND OF THE SKY YOU ARE IN THE LAND OF THE SKY. **Skydive Empuriabrava**, the world's most popular **skydiving** zone, is set in unique, ...

**Skydive Empuriabrava - Wikipedia**, the free encyclopedia

[en.wikipedia.org/wiki/Skydive\\_Empuriabrava](https://en.wikipedia.org/wiki/Skydive_Empuriabrava)

**Skydive Empuriabrava** is the brand that has been commercially operating **Empuriabrava Aerodrome** (on the **Empuriabrava** residential estate of the town of ...

[History](#) · [Services](#) · [Competitions](#) · [Awards](#) · [Exhibitions and ...](#) · [Book of honour](#)

**Skydive Empuriabrava - YouTube**

[www.youtube.com/watch?v=rELd9dZwJ3I](https://www.youtube.com/watch?v=rELd9dZwJ3I)

By Bruno Sees · 3 min · 1.9K views · Added 3/10/2015

Video embedded · My first Tandem jumping in **Skydive Empuriabrava**. One of the best experience of my life without any doubt. Special Thanks to Jose Luis! Snapchat ...

**Skydiving in Empuriabrava - YouTube**

[www.youtube.com/watch?v=FGNSLdodjPk](https://www.youtube.com/watch?v=FGNSLdodjPk)

By Kalyanashis Chakraborty · 6 min · 113 views · Added 7/26/2015

Video embedded · Want to watch this again later? Sign in to add this video to a playlist. My first **skydiving** experience

**Beginners' skydiving courses - Skydive Empuriabrava ...**

[www.skydiveempuriabrava.com/en/beginners-skydiving-courses](http://www.skydiveempuriabrava.com/en/beginners-skydiving-courses)

Do you know what it feels like to **skydive**? If you are here, you may have already experienced the thrill of **skydiving** and freefall in a tandem jump.

Fig. 5.58: First webpage recommendations on Bing for the query "skydiving in Manchester" (left) and "skydiving in Empuriabrava" (right)



## 6 Discussion

*The chapter discusses the obtained results described in the previous section. These results are critically examined with regard to the literature and research questions. The influence of the data and applied methods on the results is also scrutinized. Overall, the main emphasis is on the significance of the acquired results and answering the research questions.*

The primary objective of this master thesis was the exploration of place names on the Web. This was done by investigating the frequency of place names and spatial relations with the Microsoft Web N-Gram Service. As a quick recap, the dissertation set out to answer the following questions:

---

**RQ1** *What are the characteristics of place names occurring in Web n-grams, in terms of spatial coverage or ambiguity?*

---

---

**RQ2** *To what degree can spatial information retrieved from Web n-grams be used to describe the world?*

---

---

**RQ3** *Can an application, using Web n-grams to link information to place names, be evaluated?*

---

All of these research questions are individually addressed in the upcoming sections. In case of the first research question, the emphasis was on the (spatial) distribution of place names at different granularities on the Web. Furthermore, the effect which granularity had on ambiguity is argued. A possible reason for high Web frequencies was inspected with a correlation between place name Web frequency and place name population. The second research question followed an explorative approach and aimed to analyze the usability of Web n-grams to describe the world. This was performed by assigning toponyms to geographic features and sports activities, while checking the plausibility of the results. Finally, the third research question tried to clarify if the received results could be evaluated by comparing spatial relationships with ground truth or querying for them on the Web search engine Bing.

## 6.1 Characteristics of Place Names occurring in Web N-Grams

The spatial coverage of the country joint probabilities revealed high joint probabilities in Europe, North America, China, India and Australia. The joint probabilities in Africa were generally low. Overall, the countries with a high Web frequency also have great political power or a high number of inhabitants. Nevertheless, the differences in joint probability could be of ambiguous, demographic, economic, geographical, historical, political, social or structural nature. The underlying reasons for the Web frequency is a question that remains to be answered.

In total, all place name frequencies/probabilities vary strongly on the Web. This is especially seen by comparing the statistical values at the different granularities (displayed in table 6.1). The place name joint probabilities tend to get more dispersed with increasingly fine granularity. The minimum and mean place name joint probability get smaller with hierarchical lower place name levels. This goes along with an increase in number of place names. The maximum and standard deviation joint probability of each place name category vary, while the range tends to get bigger in categories of finer granularity. The biggest difference is seen in the minimum, maximum and range joint probability of cities. The minimum joint probability in cities is practically non-existent, whereas the maximum of city joint probabilities is extremely bigger compared to the other maximum place name probabilities. Consequently, certain place names are heavily more frequent compared to other place names on the Web. These place names are overrepresented on the Web, which is a potential indicator of ambiguity. Finally, the following could be deduced from these results: **The finer the level of granularity of the place names is the more ambiguity gets introduced.** This fact has also been shown in the research of Brunner & Purves (2012), Derungs et al. (2012) and Hill (2006). In their research, the ambiguity is more pronounced at fine spatial granularities and for natural geographic features types. Leidner & Liebermann (2011) state that the geo/non-geo ambiguity is directly affected by the level of granularity of the considered toponyms. Hence, it is easier to recognize toponyms on country level rather than recognizing them on city level. This is partly due to the comparative smaller number of country place names which provide fewer opportunities for ambiguity (Leidner & Liebermann 2011).

The distribution of place name joint probabilities by rank also changed with the level of granularity. The place name joint probabilities by rank tend to follow a Zipf distribution more closely in finer place name granularities. Thus, following can be deduced: **The finer the granularity of place names is the more does the distribution of place name joint**

**probabilities by rank resemble a Zipf distribution.** This would imply that the finer the granularity of place names is the more likely they follow the same inverse power law of English word frequencies (Li 1992; Zipf 1932). Therefore, place names in finer granularities are treated or tend to act like ordinary English words.

The number of words constituting a place name also had an influence on the place name joint probability. One worded place names tended to have a higher mean joint probability than place names made up of two or more words. Furthermore, the place names containing fewer words had a higher discrepancy between joint probabilities. This is also mainly due to the number of one worded place names being greatly higher than more worded place names. However, it could be said that **the toponym Web frequency and ambiguity decreases with word count.** The decrease in Web frequency and ambiguity makes sense, since one worded place names could also be part of multi-worded place names. Additionally, one worded place names are more common compared to multi-worded place names. The ambiguity presumably also decreases, because the more words make up a place name the unlikelier it is to have multiple referents or a different meaning in the context.

**Table 6.1: Statistical parameters of all place name joint probabilities**

Statistics: Place Name Joint Probability					
Statistics	Probability [%]				
n Place Names	Continent 7	Country 206	Capital City 214	City 143 252	All 143 679
Minimum	0.0006745280	0.0000037154	0.0000001282	0.0000000000	<b>0.0000000000</b>
Maximum	0.0158854675	0.0227509743	0.0156314764	1.5559656316	<b>1.5559656316</b>
Mean	0.0067539556	0.0023257535	0.0005187792	0.0001425559	<b>0.0024352611</b>
Std. Deviation	0.0050934509	0.0033395510	0.0017028104	0.0051494061	<b>0.0051447361</b>
Range	0.0089536477	0.0227472590	0.0156313482	1.5559656316	<b>1.5559656316</b>

The population was tested as a possible factor influencing joint probabilities in chapter 5.2. The results showed a statistically significant positive correlation between country/capital city name joint probability and population. This indicates that population has a definite influence on place name joint probabilities at country and capital city level. Therefore, the following fact can be drawn: **On country and capital city level high place name joint probabilities are accompanied by high number of inhabitants, while low joint probabilities are accompanied by low number of inhabitants.** This means populated places are more frequently represented on the Web than less populated places. However, exceptions exist and on city level the joint probabilities are heavily influenced by ambiguity which makes it hard

to derive any connection between population and Web frequency. These findings correspond with a popular heuristic used for toponym resolution in GIR. The knowledge based approach mostly uses population data to disambiguate place names (Clough et al. 2004; Li et al. 2003). The applied rules in those cases generally imply that if multiple place referents exist the one with the highest population is the correct referent (Amitay et al. 2004; Rauch et al. 2003). Hence, it is assumed that a place with a high population is more likely to be mentioned in a document or the Web than a place with a lower population (Amitay et al. 2004; Rauch et al. 2003). This corresponds with the findings on country and capital city level in this thesis.

In addition, a positive spatial autocorrelation was found between country name joint probabilities. This indicates the spatial distribution of places also influences the joint probability. Moreover, **spatially adjacent country names have similar joint probabilities and are spatially clustered**. Consequently, the country name joint probabilities follow Tobler's first law of geography (Tobler 1970). This means spatially near country names and their joint probabilities are more related than distant countries and their corresponding joint probability.

Overall, the obtained results and ambiguity of different place name granularity levels depend strictly on the resources used (Buscaldi 2011). The continents could have been divided into less or more continent names, whereas the countries could have only contained member states of the United Nations ("Member States" 2014). The capital cities could have been reduced to only one per country and the list of cities could have been taken from another resource than GeoNames. Additionally, the definition of a city in GeoNames was very vague. The GeoNames list of cities with over 1000 inhabitants contained administrative divisions, towns, villages and populated places. Moreover, the granularity of the gazetteer had a big influence on the results. Most gazetteers contain list of popular or relevant places. Usually, the gazetteers which are finer grained yield noise and may contain unpopular places due to short term media spotlight caused by important events (Leidner 2007; Shaw 2003). The dataset from GeoNames comes from other public and open gazetteers and can be edited by multiple users similar to Wikipedia (Ahlers 2013). These sources vary in quality, scope, resolution and age which are a further source of concern. Ultimately, this influences the quality and accuracy of the data and may introduce noise. A simple example of the entries in the list of cities from GeoNames illustrates the problem of data quality. The cities Cologne, Germany and Milan, Italy were written in their native language in the ASCII format. Accordingly, the places were

listed as Koeln and Milano. Other place name entries could also be wrongly spelled or not even exist, which of course has an influence on the obtained results.

Another problem would be the Microsoft Web N-Gram Service. The collection of documents in the database contains only webpages indexed by Bing in the EN-US market. This reduced the search to only English terms, since other languages are definitely underrepresented in the corpus compared to English. Therefore, the use of native city names already has an influence on the joint probability of that certain place name. Milan for example is approximately twice as frequent as the native name Milano in the Microsoft Web N-Gram Service. Additionally, the Microsoft Web N-Gram Service did not distinguish between capitalized words as the documents in the collection were transformed to lowercased text. The inclusion of capitalized words would have greatly simplified the search for places on the Web. A further limitation was the word limit of n-grams. Only the probability for word arrangements containing up to five words could be searched. Consequently, the biggest drawback of Web n-grams was the lack of context. This makes it near impossible to resolve ambiguity, since the Web n-grams lack additional context data. The same drawback of working with Web n-grams was mentioned by Derungs & Purves (2014a). Hereafter, ambiguous place names such as Australia, Georgia and York are overrepresented on the Web. These are influenced by a multitude of factors. Australia can either be a country or continent, while Georgia can be a country, state or a female forename. On the other hand, the city York is heavily influenced by the city New York. The city New York is a very popular city and contains the word York which heavily increases the Web frequency of York. All these factors should be considered when looking at the produced data. **Consequently, the results only represent the Web frequency of the place name and not the actual place in question.**

## 6.2 Describing the World with Web N-Grams

Some interesting results were returned by linking European locations to geographic features and sports activities with Web n-grams. The strength of the relationship between place names and geographic features/sports activities was expressed as conditional probability. This conditional probability was returned if a place name was present in the first 1000 autocompletes following the geographic feature/sport activity and preposition. Hence, place names with a stronger bond to a topic also had a higher conditional probability. The majority of the received results were plausible. However, a variety of investigated geographic features and sports activities contained unknown associations to place names. In this case, the unknown associations refer to associations which were unknown to the author. These results

were checked on their plausibility with a query on the Web search engine Bing. Surprisingly, the majority of spatial information retrieved from Web n-grams returned accurate representations of Europe. Furthermore, new insights about places were gained. The simplicity of the method for obtaining these results also made it easy to search for a huge variety of topics in relation to toponyms. As a result, **Web n-grams can certainly be used to describe Europe.**

Nonetheless, geographic features and sports activities are only two examples of linking toponyms to certain topics. There could be a variety of other examples such as using user generated content and large corpora to link geographic features to even finer toponym granularities (Derungs & Purves 2014b) or using geographic features to describe landscapes/geomorphometry (Derungs & Purves 2014b; Gschwend & Purves 2012). There is multitude of examples but there is a limit to the total number of possibilities.

Nevertheless, some problems persist when Web n-grams are used for linking spatial information to topics. A few of the received results contain unexplained associations. Thus, the origin of a strong relationship between geographic feature/sport activity and place name cannot always be explained through queries on a Web search engine. Missing relations of an expected association between toponym and geographic feature/sport activity also stay unexplained.

More importantly, all of this is heavily influenced by the methodology. Only the first 1000 autocompletes were returned for a doublet <geographic feature><spatial relationship> or <sport activity><spatial relationship>. This means the conditional probability of other place names occurring after the 1000 autocompletes was not retrieved. Fortunately, entries occurring after the 1000 autocomplete had considerably smaller conditional probabilities and were therefore less likely to follow an arrangement of words. Consequently, only the strongest relationships between geographic features/sport activities and places are considered. This is apparent when comparing the total matches between European countries/cities and the doublets <geographic feature><spatial relationship> or <sport activity><spatial relationship>. In total 45 of the 55 countries followed at least one geographic feature doublet, while 52 out of the 55 countries followed at least one sport activity doublet. The cities had a much lower match rate. In total 331 of the 48 554 cities followed at least one geographic feature doublet, while 338 out of the 48 554 cities followed at least one sport activity doublet. Therefore, it can be concluded that toponym with coarser granularity are more likely to be represented in the autocompletes of certain topics. This was mainly due to the high number of cities compared to the small number of countries and the restriction of 1000 autocompletes. The

frequency of a country/city also influences the result, since infrequent place names on the Web are assumed to have less chance to be associated with a topic. Accordingly, it is easier to obtain the spatial coverage of topics in countries rather than cities.

The problem of maximum word restriction, ambiguity and lack of context still persisted with triplets. The use of triplets may reduce the ambiguity but they do not get completely rid of it. Cases of ambiguity were still observable at country level and even more at city level which introduced more opportunities for ambiguity.

All of these results and flaws are comparable to the research done with Google Autocomplete (“The World Through the Eyes of a Search Algorithm” 2014). In this research the properties Google Autocomplete associates with countries was investigated. The result gave insights in physical, economic and socio-demographic properties of countries. These reinforced existing patterns but also returned somewhat puzzling results. Not knowing the full extent of the data source and the lack of understanding the mechanism of query suggestions made it hard to resolve the puzzling results. Moreover, the cases of semantic mismatches and ambiguity had to be removed from the data. But overall, the potential of retrieving geographical information was encouraging. Generally, this corresponds with the findings in this thesis. The potential of Web n-grams is apparent, as it gives insights into certain geographic concepts on the Web. The lack of context makes it hard to comprehend certain results, while ambiguity is present and somewhat distorts the results.

In summary, **Web n-grams were used to describe Europe on the Web and tended to work better for coarser level of granularities such as countries. Due to this, the same is assumed for the whole world. Thus, Web n-grams can be used to describe the world.** However, ambiguity is still present and has to be dealt with properly.

### **6.3 Evaluation of using Web N-Grams to link information to place names**

The evaluation of using Web N-grams to link information to place names was attempted with a method usually used in GIR. Normally in GIR, the automatic retrieval and disambiguation of toponyms in document(s) is compared to a gold standard. Then the precision and recall are calculated to quantify the effectiveness of the applied retrieval method (Grover et al. 2010; Leidner 2007). However, no place names were recognized or disambiguated in a set of documents. Rather, doublets in form of <place name><spatial relation> were commissioned to the Microsoft Web N-Gram Service and the 1000 autocompletes of the doublets were returned. The actual evaluation was attempted by inspecting these 1000 autocompletes with the correct triplets in the form of <place name><spatial relation><place name>. Yet, no gold

standard was available to calculate the precision or recall. Besides, the Microsoft Web N-Gram Service only returned the autocompletes which made it hard to identify if the place names in the returned 1000 autocompletes were really referring to the actual place. Therefore, new metrics were introduced with correctly retrieved spatial relations (see chapter 4.4.2) and relevant found spatial relations (see chapter 4.4.3). These metrics were similar to precision and recall, except they were adapted to the 1000 autocompletes. The gold standard or rather the ground truth (see chapter 4.4.1) had to be created with the help of the available data.

Generally, the introduced ratios gave useful insights into the representation of spatial relationships on the Web. In other words, **the accuracy and completeness of the 1000 autocompletes for certain spatial relations was evaluated**. The data was still subject to ambiguity and did not necessarily represent the actual spatial relations between two places. It rather represented the relation between two place names. In addition, the method was still restricted by the constraints of the Microsoft Web N-Gram Service. This reduced the number of investigated place names, because some triplets were too long to be properly processed by the Microsoft Web N-Gram Service.

The use of the spatial relationships was also critical, since it had a direct influence on the results. The spatial relationship “in” can have multiple meanings. It can be used to indicate time, location, shape, color, size, opinion and other things. Thus, “in” is not only used to indicate a location but also other things. This introduces more cases of ambiguity to the triplets. On the other hand, the spatial relationship “bordering” is very straightforward. This is seen in the received results for the spatial relation country bordering country in section 5.3.4. The majority of the retrieved spatial relations were correct, while the percentage of relevant found spatial relations is low. Consequently, the use of spatial relations has an influence on the ratio of correctly retrieved spatial relations and relevant found spatial relations. This would mean with more specified spatial relations higher percentages of correctly retrieved spatial relations can be obtained, whereas the percentages of relevant found spatial relations is reduced. In view of that, the preposition “in” could have been expressed with a more specific word arrangement such as “contained by” to obtain higher percentages of correctly retrieved spatial relations.

A further problem was the resources used to create the ground truth, as the ambiguity of a toponym is influenced by the resources used to represent the world (Buscaldi 2011). For example, there are two cities named Cambridge in the world according to WordNet, 38 according to Yahoo! GeoPlanet and 40 according to GeoNames (Buscaldi 2011). Moreover, the city names in GeoNames are unevenly distributed which had a direct influence on the

calculated ratios (“Mapping the GeoNames Gazetteer” 2014). This uneven distribution of place names is seen in the map (figure 5.22) from chapter 5.3.6. The resulting map matches with the findings in “Mapping the GeoNames Gazetteer” (2014). As a result, the percentage of relevant found spatial relations was directly influenced in continents/countries with high city name densities. These likely contained an abundant number of unknown cities which lowered the percentage of relevant found spatial relations in the autocompletes. Furthermore, the divisions of continents also had an influence on the spatial relations concerning countries/capitals/cities in continents. For instance, the islands states in the Pacific Ocean and their cities were all contained by the continent Australia in the ground truth, rather than Oceania.

The third problem was the applied methodology. The ground truth only contained one correct answer for every spatial relation. This means every city could only refer to one correct referent. Therefore, the ratio of correctly retrieved relations was automatically lowered for cities with multiple duplicates. For instance, the city Paris had two entries in the ground truth. One of these entries was Paris in France and the other was Paris in the United States. The 1000 autocompletes for these two returned France and the United States. For the first entry France is the correct referent, however United States is also returned. This led to a lower percentage of correctly retrieved spatial relations, since multiple country referents were retrieved. The same applies for the entry referring to Paris in the United States. Fortunately, this made the correctly retrieved spatial relations ratio a very good indicator of possible geo/geo ambiguity in place names. Uncommon spatial relations such as capital city in continent also tended to have lower percentages of relevant spatial relations. Lastly, the limit of autocompletes also had an influence on the calculated ratios. A higher limit of autocompletes returned higher ratios of relevant found spatial relations, whereas it lowered the ratios of correctly retrieved spatial relations.

Ultimately, the methodology, the conceptualization of place names and place name relationships greatly influence the evaluation of place name relations in the autocompletes.

Another alternative to evaluate the use of Web n-grams to link information to space was offered by verifying the plausibility of the results on a Web search engine. The occurrence of most relations could be comprehended through simple Web searches but not really evaluated. This was mostly due to the fact that no ground truth was available for certain topics in/near a country/city. Overall, this was a tedious and inefficient way to verify results and is not recommended for an evaluation.

In summary, **an application making use of Web n-grams can be evaluated to a certain degree.** It gives insights how spatial relations of place names are represented in the 1000 autocompletes of the Microsoft Web N-Gram Service in terms of accuracy and coverage. However, the evaluation does not look at actual spatial relationships between places. It looks at the frequency of possible place name relationships on the Web. The word limit of the Microsoft Web N-Gram Service also restricts the search to a limited number of places. Moreover, ground truth is mostly unavailable and has to be generated. This can be tedious task and might not be the best solution to evaluate applications making use of Web n-grams.

## 7 Conclusion

*The last chapter reiterates the achievements and findings of the dissertation. The limitations and possibilities of Web n-grams are further discussed in the findings. This is concluded by an outlook of possible implementations for Web n-grams and future work.*

### 7.1 Achievements

The primary objective of this master thesis was the exploration of place names and spatial relationships in the Microsoft Web N-Gram Service. This gave insights into the representation of place names and spatial relationships on the Web. The task was accomplished by querying for different granularities of place names and more complex relationships in the Microsoft Web N-Gram Service. The resulting achievements are listed below:

- Web n-grams were used in the field of geography to gain new insights into geographic concepts.
- Several Java codes were constructed which access the Microsoft Web N-Gram Service API and return the joint probability, conditional probability and autocompletes of a list of place names, topics or complex relationships.
- The joint probability (frequency) of 143 679 place names in the hundreds of billions of webpages from the Microsoft Web N-Gram Service was investigated and statistically interpreted. This was done for different levels of granularity: continent, country, capital city and city. The number of investigated place names for each level was as follows:
  - Seven continent names
  - 206 country names
  - 214 capital city names
  - 143 252 city names
- The joint probabilities of each granularity level were analyzed in terms of:
  - top five
  - bottom five
  - minimum
  - maximum
  - mean

## 7 | Conclusion

- standard deviation
  - range
- The joint probabilities of each granularity level were investigated on following a Zipf distribution and number of words constituting a place name.
- The spatial distribution of country name joint probabilities was visualized on a map.
- A positive correlation between place name joint probability and population was statistically proven at country and capital city level.
- A positive spatial autocorrelation between country name joint probabilities was statistically proven.
- New metrics were introduced to evaluate the accuracy and coverage of the autocompletes in the Microsoft Web N-Gram Service.
- The accuracy and coverage for the 1000 autocompletes of the following triplets was analyzed:
  - Country in continent
  - Capital city in continent
  - City in continent
  - Country bordering country
  - Capital city in country
  - City in country
- An application was developed which makes use of Web n-grams to link information and place names. The application makes it easy to obtain a variety of results in a short time and offered a way to describe the world through Web n-grams. Furthermore, the obtained results helped to gain new insights into the real world and the world on the Web.

## 7.2 Findings

The investigation of place names and complex relationships in the Microsoft Web N-Gram Service returned a variety of results. The resulting discoveries are summarized below.

The granularity of place names had a huge influence on ambiguity and the distribution of joint probabilities by rank. It was observed that place names at finer granularities tend to be more heavily scattered and contain more possibilities for ambiguity. Additionally, the distribution of place name joint probabilities by rank at fine granularities strongly resembles a Zipf distribution. Hence, place names at fine granularities tend to act like ordinary English words. The amount of words constituting a place name also had a direct influence on the joint probability and ambiguity. It was witnessed that place names made up of fewer words had a higher joint probability and were more prone to ambiguity. A further factor influencing the joint probability was the population and spatial distribution of places. Country names and capital city names with high joint probabilities are usually accompanied by high number of inhabitants and vice versa. Moreover, spatially adjacent country names have similar joint probabilities and are spatially clustered. Although, the underlying reason for certain place name Web frequencies could be of ambiguous, demographic, economic, geographical, historical, political, social or structural nature.

Web n-grams can be used to link information to geographic space in form of triplets. Thus, Web n-grams are capable of describing the world. This generally works better for coarser granularities. However, a proper evaluation of such an application is difficult due to lack of context. Ground truth is usually unavailable and has to be created by hand. The accuracy and completeness of spatial relations in the 1000 autocompletes can be assessed with the introduced metrics: correctly retrieved spatial relations and relevant found spatial relations. The metric correctly retrieved spatial relations was found to be a good indicator for ambiguity, especially geo/geo ambiguity. An alternative “evaluation” is offered by verifying unknown or suspicious strong relationship between topics and place names on a Web search engine. This works best for small number of samples and only helps to gain insights into why certain topics and place names are heavily linked.

Overall, there were two main drawbacks of using Web n-grams to link information to place names: the input data and the Microsoft Web N-Gram Service itself. The data inputs can be changed, while the constraints of the Microsoft Web N-Gram Service persist.

The data sources used to create the list of place names, geographic features, sport activities and ground truth had substantial influence on the results. For instance, the GeoNames gazetteer was used for the list of city names. It contained a huge number of place names

entries. However, the city names were sometimes written in their native language, incomplete, wrong, were spatially biased or had an inconsistent classification. This made it hard to understand which entries were actual cities since the list included administrative divisions, towns, villages and populated places. Hence, a lot of noise was introduced. Consequently, more is not always better. The list of city names could and maybe should have been reduced to a certain area. This would have been easily accomplished by using a local gazetteer instead of a global gazetteer. The creation of the ground truth was based on these place name lists. It was not always clear which country was contained by which continent. This was especially the case for countries between Europe and Asia. Additionally, the islands states in the Pacific Ocean were contained by the continent Australia instead of Oceania. Generally, the conceptualization of place names and their spatial relationships is difficult and complex. Decision had to be made for certain place names and spatial relationships which influence the obtained results. These decisions are neither completely correct nor wrong. It all depends on the classification and view of the individual.

The terminology used for spatial relationships also influenced the results. The use of more explicit terms increased the percentage of correctly retrieved spatial relations in the autocomplete, while reducing the percentage of relevant found spatial relations.

The search restriction of up to five words in the Microsoft Web N-Gram Service was the biggest unchangeable drawback. It only returned the probabilities and autocompletes of short place names or triplets. This word restriction greatly limited the search for place names or specific relationships. While the lack of context made it near impossible to resolve ambiguity in the received probabilities or autocompletes. Consequently, all of the results in this master thesis only represent the Web frequency of place names and not of actual places. The same implies for the results dealing with spatial relationships.

In conclusion, the results clearly demonstrated the potential of Web n-grams and their ability to describe geographic concepts. They provide a rich resource to explore spatial relationships between topics/place names and place names. Moreover, the developed approach allows the use of different spatial scales and can easily be extended by a variety of spatial relations. Nevertheless, the advantage of being able to simply query for Web n-grams (instead of having access to whole documents) also brings challenges in understanding potential ambiguity and local context. These challenges should be addressed in future work.

### 7.3 Outlook

The main deficits of Web n-grams are the lack of context and the resulting difficulty of understanding potential ambiguity. Therefore, future work with Web n-grams should focus on offering possibilities to deal with, recognize, remove, or solve ambiguity. The recognition and solving of ambiguities in Web n-grams is rather impossible. Therefore, the main emphasis should be on removing and dealing with ambiguity. A possible approach for removing ambiguity would be removing ambiguous place names from the investigation. This was done in this dissertation by building a ratio between joint probability rank and population rank. However, finding a proper threshold for the ratio was difficult and would need further investigation. Another approach of dealing with ambiguity would consist of querying for known ambiguities in the Microsoft Web N-Gram Service and subtracting their joint probability from the actual term in question. For instance, the joint probability of “London Canada” would be subtracted from the joint probability “London” to obtain the actual joint probability of London, United Kingdom. This would remove possible geo/geo ambiguity and bring the joint probabilities of places closer to their actual joint probability. Furthermore, it could easily be implemented by using all place names of a global gazetteer such as GeoNames. Cases of geo/non-geo ambiguity would be harder to implement, since attaining all possible cases are hard to come by.

There are numerous imaginable implementations for Web n-grams. This dissertation already showed how Web n-grams can be used to link geographic features and sports activities to toponyms. The distribution of geographic features in a place or better a country could be verified by comparing the results with tags on Flickr. The dominant land use class in countries could be searched with Web n-grams and compared to the country mean of the CORINE (coordination of information on the environment) land cover. Finally, Web n-grams could help in solving some of the challenges in GIR. They could be used as disambiguation method, where the likeliest place referent is assigned to the recognized place name in the text. Moreover, they could help in geometric interpretation of vague place names (e.g. Mittelland) or help in quantifying vague spatial language such as “near”.

These are only some possible implementations of Web n-grams. They could have several more implementations in geography in the future. Nevertheless, their flaws have to be acknowledged and properly dealt with.



## References

- Abdelmoty, A.I. & El-Geresy, B.A. (2008). Modeling Spatial Relations between Places to Support Geographic Information Retrieval. In: I. Lovrek, R.J. Howlett & L.C. Jain (eds.), Knowledge-Based Intelligent Information and Engineering Systems. 12th International Conference, KES 2008, Zagreb, Croatia, September 3-5, Proceedings, Part II. *Lecture Notes in Computer Science*, Berlin, Germany: Springer, Vol. **5178**: 689-694.
- Abou-Assaleh, T., Cercone, N., Keselj, V. & Sweiden, R. (2004). N-gram-based Detection of New Malicious Code. In: *Proceedings of the 28th Annual International Computer Software and Applications Conference, COMPSAC '04*. IEEE: Vol. **2**, 41-42.
- About GeoNames. (2015). *GeoNames*. Retrieved August 17, 2015, from <http://www.geonames.org/about.html>
- Ahlers, D. (2013). Assessment of the accuracy of GeoNames gazetteer data. In: *Proceedings of the 7th Workshop on Geographic Information Retrieval, GIR '13*. Orlando, FL: ACM, 74-81.
- Ahlers, D. & Boll, S. (2007). Location-based Web Search. In: A. Scharl & K. Tochtermann (eds.), *The Geospatial Web. How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*. Advanced Information and Knowledge Processing. London, United Kingdom: Springer, 55-66.
- Amitay, E., Har'El, N., Sivan, R. & Soffer, A. (2004). Web-a-Where: Geotagging Web Content. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '04*. New York, NY: ACM, 273-280.
- Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, **59**: 74-76.
- Barton, D. & Court, D. (2012). Making Advanced Analytics Work for You. *Harvard Business Review*, **90**: 78-83.
- Battig, W. F. & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, **80**(3): 1-46.
- Battle of White Mountain. (2008, December 11). *Wikipedia, The Free Encyclopedia*. Retrieved September 5, 2015, from [http://en.wikipedia.org/wiki/Battle\\_of\\_White\\_Mountain](http://en.wikipedia.org/wiki/Battle_of_White_Mountain)
- Belkin, N.J. & Croft, W.B. (1992). Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM – Special issue on information filtering*, **35**(12): 29-38.
- Bensalem, I.; & Kholladi, M.K. (2010). Toponym Disambiguation by Arborescent Relationships. *Journal of Computer Science*, **6**(6): 653-659.
- Brewer, C.A. (2013). ColorBrewer 2.0. Color advice for cartography. *Colorbrewer2.org*. Retrieved May 5, 2015, from <http://colorbrewer2.org/>
- Brown, P.F., deSouza, P.V., Mercer, R.L., Della Pietra, V.J. & Lai, J.C. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, **18**(4): 467-479.
- Brunner, T.J. & Purves, R.S. (2008). Spatial Autocorrelation and Toponym Ambiguity. In: *Proceedings of the 2nd international workshop on Geographic information retrieval, GIR '08*. New York, NY: ACM, 25-26.
- Buscaldi, D. (2011). Approaches to Disambiguating Toponyms. *SIGSPATIAL Special*, **3**(2): 16-20.
- Buscaldi, D. & Magnini, B. (2010). Grounding toponyms in an italian local news corpus. In: *Proceedings of Workshop on Geographical Information Retrieval, GIR'10*. New York, NY: ACM.

## | References

- Buscaldi, D. & Rosso, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, **22**(3): 301-313.
- Cleverdon, C.W. (1991). The significance of the Cranfield tests on index languages. In: *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91. New York, NY: ACM, 3-12.
- Clifton, C. (2014, November 28). Data Mining. Knowledge discovery in databases. *ENCYLOPAEDIA BRITANNICA*. Retrieved August 29, 2015, from <http://www.britannica.com/technology/data-mining>
- Clough, P. (2005). Extracting metadata for spatially-aware information retrieval on the Internet. In: *Proceedings of the ACM Workshop on Geographic Information Retrieval, GIR '05*. New York, NY: ACM, 25-30.
- Clough, P., Sanderson, M. & Joho, H. (2004). *Extraction of Semantic Annotations from Textual Web Pages*. Technical Report, University of Sheffield, United Kingdom.
- Cohen, P. (2010, December 16). In 500 Billion Words, New Window on Culture. *New York Times*. Retrieved December 8, 2014, from <http://www.nytimes.com/2010/12/17/books/17words.html?pagewanted=all>
- Creative Commons. Attribution 3.0 Unported. (2015). *Creative Commons*. Retrieved February 25, 2015, from <https://creativecommons.org/licenses/by/3.0/legalcode>
- Danube Delta. (2007, November 30). *Wikipedia, The Free Encyclopedia*. Retrieved August 30, 2015, from [https://en.wikipedia.org/wiki/Danube\\_Delta](https://en.wikipedia.org/wiki/Danube_Delta)
- Datasources used by GeoNames in the GeoNames Gazetteer. (2014). *GeoNames*. Retrieved November 17, 2014, from <http://www.geonames.org/data-sources.html>
- Davies, C., Holt, I., Green, J., Harding, J. & Diamond, L. (2009). User Needs and Implications for Modelling Vague Named Places. *Spatial Cognition & Computation*, **9**(3): 174–194.
- Derungs, C., Palacio, D. & Purves, R.S. (2012). Resolving fine granularity toponyms: Evaluation of a disambiguation approach. In: *GIScience 2012, 7th International Conference on Geographic Information Science, September 18-21, 2012*. Columbus, OH.
- Derungs, C. & Purves, R.S. (2014a). Where's near? Using web tri-grams to explore spatial relations. In: *GIScience 2014: 8th International Conference on Geographic Information Science*, September 23-26, 2014 Vienna, Austria: 158-162.
- Derungs, C. & Purves, R.S. (2014b). From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*, **28**(6): 1272-1293.
- Download. Free Gazetteer Data. (2014). *GeoNames*. Retrieved November 17, 2014, from <http://download.geonames.org/export/dump/>
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, **17**(3): 37-54.
- Franklin, C. (1992). An Introduction to Geographic Information Systems: Linking Maps to Databases. *Database*, **15**(2): 12-21.
- Freire, N., Borbinha, J., Calado, P. & Martins, B. (2011). A Metadata Geoparsing System for Place Name Recognition and Resolution in Metadata Records. In: *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, JCDL '11*. New York, NY: ACM, 339-348.

- Fu, G., Jones, C.B. & Abdelmoty, A.I. (2005). Ontology-Based Spatial Query Expansion in Information Retrieval. In: R. Meersman & Z. Tari (eds.), *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE. OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005*, Agia Napa, Cyprus, October 31 - November 4, 2005, Proceedings Part II. *Lecture Notes in Computer Science*, Berlin, Germany: Springer, Vol. **3761**: 1466-1482.
- Gan, Q., J. Attenberg, A. Markowetz, & Suel, T. (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: ACM, 49-56.
- Gelernter, J. & Balaji, S. (2013). An algorithm for local geoparsing of microtext. *Geoinformatica*, **17**(4): 635-667.
- GeoNames. (2014). *GeoNames*. Retrieved November 16, 2014, from <http://www.geonames.org/>
- Goodchild, M.F. and Hill, L.L. (2008). Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, **22**(10): 1039-1044.
- Goodman, J.T. (2001). A bit of progress in language modeling. *Computer Speech and Language*, **15**(4): 403-434.
- Google Books Ngram Viewer. What does the Ngram Viewer do? (2013). *Google Books Ngram Viewer*. Retrieved December 8, 2014, from <https://books.google.com/ngrams/info>
- Grossman, D.A. & Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics*. Springer.
- Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S. & Ball, J. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A*, **368**: 3875-3889.
- Gschwend, C. & Purves, R.S. (2012). Exploring Geomorphometry through User Generated Content: Comparing an Unsupervised Geomorphometric Classification with Terms Attached to Georeferenced Images in Great Britain. *Transactions in GIS*, **16**(4): 499-522.
- Hart, G. & Dolbear, C. (2007). What's So Special about Spatial? In: A. Scharl & K. Tochtermann (eds.), *The Geospatial Web. How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*. Advanced Information and Knowledge Processing. London, United Kingdom: Springer. 39-44.
- 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. 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18-20, 2000 Proceedings. Lecture Notes in Computer Science*, Berlin, Germany: Springer, Vol. **1923**: 280-290.
- Hill, L. L. (2006). *Georeferencing: The Geographic Associations of Information*. Digital Libraries and Electronic Publishing. Cambridge, MA: MIT Press.
- Hirji, K.K. (2001). Exploring Data Mining Implementation. How large volumes of organizational data can be exploited for sustained competitive advantage. *Communications of the ACM*, **44**(7): 87-93.
- Horak, J., Belaj, P., Ivan, I. Nemeč, P., Ardielli, J. & Ruzicka J. (2011). Geoparsing of Czech RSS News and Evaluation of Its Spatial Distribution. In: R. Katarzyniak, T.-F. Chiu, C.-F. Hong & N.T. Nguyen (eds.), *Semantic Methods for Knowledge Management and Communication. Studies in Computational Intelligence*. Berlin, Germany: Springer, Vol. **381**: 353-367.
- Huxhold, W. E. (1991). *An Introduction to Urban Geographic Information Systems*. New York, NY: Oxford University Press.
- ISAF Sailing Worldcup. An Official Website of the International Sailing Federation. (2015). *Sailing.org*. Retrieved September 7, 2015, from <http://www.sailing.org/worldcup/home.php>

## | References

- Janowicz, K. and Kessler, C. 2008. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, **22**(10), 1129-1157.
- Jones, C.B. & Purves, R.S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, **22**(3): 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.
- Katyn massacre. (2015, September 4). *Wikipedia, The Free Encyclopedia*. Retrieved September 5, 2015, from [https://en.wikipedia.org/wiki/Katyn\\_massacre](https://en.wikipedia.org/wiki/Katyn_massacre)
- Larson, R.R. 1996. Geographic information retrieval and spatial browsing. In L.C. Smith and M. Gluck (eds.), *Geographic information systems and libraries: patrons, maps, and spatial information*. (papers presented at the 1995 Clinic on Library Applications of Data Processing, April 10-12, 1995): 81-124.
- Leidner, J.L. (2007). *Toponym Resolution in Text: Annotation, Evaluation, and application of Spatial Grounding of Place Names*. (Ph. D. thesis, University of Edinburgh).
- Leidner, J.L., Sinclair, G. & Webber, B. (2003). Grounding spatial named entities for information extraction and question answering. In: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references, HLT-NAACL-GEOREF '03*. Morristown, NJ: Association for Computational Linguistics, 31-38.
- Leidner, J.L. & Lieberman, M.D. (2011). Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, **3**(2): 5-11.
- Lew, A. & Mauch, H. (2006). Introduction to Data Mining Principles. In: S. Sumathi & S.N. Sivanandam (eds.), *Introduction to Data Mining and its Applications. Studies in Computational Intelligence*, **29**. Berlin, Germany: Springer, 1-20.
- Li, W. (1992). Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. *Information Theory, IEEE transactions on*, **38**(6): 1842-1845.
- Li, H., Srihari, R.K., Niu, C. & Li, W. (2003). InfoXtract location normalization: A hybrid approach to geographic references in information extraction. In: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references, HLT-NAACL-GEOREF '03*. Morristown, NJ: Association for Computational Linguistics, 39-44.
- Liddy, E.D. (2005). Automatic document retrieval. In: *Encyclopedia of Language and Linguistics*. Elsevier.
- List of countries and dependencies by area. (2014, December 1). *Wikipedia, The Free Encyclopedia*. Retrieved December 3, 2014, from [http://en.wikipedia.org/wiki/List\\_of\\_countries\\_and\\_dependencies\\_by\\_area](http://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_area)
- List of national capitals by population. (2015, February 24). *Wikipedia, The Free Encyclopedia*. Retrieved March 12, 2015, from [http://en.wikipedia.org/wiki/List\\_of\\_national\\_capitals\\_by\\_population](http://en.wikipedia.org/wiki/List_of_national_capitals_by_population)
- List of national capitals in alphabetical order. (2015, February 9). *Wikipedia, The Free Encyclopedia*. Retrieved February 10, 2015, from [http://en.wikipedia.org/wiki/List\\_of\\_national\\_capitals\\_in\\_alphabetical\\_order](http://en.wikipedia.org/wiki/List_of_national_capitals_in_alphabetical_order)
- List of sovereign states. (2014, November 28). *Wikipedia, The Free Encyclopedia*. Retrieved December 1, 2014, from [http://en.wikipedia.org/wiki/List\\_of\\_sovereign\\_states](http://en.wikipedia.org/wiki/List_of_sovereign_states)
- List of sovereign states and dependent territories by continent. (2015, February 20). *Wikipedia, The Free Encyclopedia*. Retrieved March 13, 2015, from [http://en.wikipedia.org/wiki/List\\_of\\_sovereign\\_states\\_and\\_dependent\\_territories\\_by\\_continent](http://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_by_continent)

- List of sports. (2015, June 25). *Wikipedia, The Free Encyclopedia* Retrieved June 25, 2015, from [http://en.wikipedia.org/wiki/List\\_of\\_sports](http://en.wikipedia.org/wiki/List_of_sports)
- Loire Valley. (2007, December 3). *Wikipedia, The Free Encyclopedia*. Retrieved August 30, 2015, from [https://en.wikipedia.org/wiki/Loire\\_Valley](https://en.wikipedia.org/wiki/Loire_Valley)
- Maimon, O. & Rokach, L. (2010). Introduction to Knowledge Discovery and Data Mining. In: O. Maimon & L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*. Springer: 1-15.
- Manning, C.D., Raghavan, P. & Schütze, C. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Mapping the GeoNames Gazetteer. (2014). *Oxford Internet Institute. Information Geographies at the Oxford Internet Institute*. Retrieved December 3, 2014, from <http://geography.oi.ox.ac.uk/?page=mapping-the-geonames-gazetteer>
- Mariño, J.B., Banchs, R.E., Crego, J.M., de Gispert, A., Lambert, P., Fonollosa, J.A.R. & Costa-Jussà, M.R. (2006). N-gram-based Machine Translation. *Computational Linguistics*, **32**(4): 527-549.
- McCurley, S.K. (2001). Geospatial mapping and navigation of the web. In: *Proceedings of the 10th International Conference on World Wide Web, WWW '01*. New York, NY: ACM, 221-229.
- Member States. Member States of the United Nations. (2014). United Nations. Retrieved December 1, 2014, from <http://www.un.org/en/members/index.shtml>
- Mennis, J. & Guo, D. (2009). Spatial data mining and geographic knowledge discovery – An introduction. *Computers, Environment and Urban Systems*, **33**: 403-408.
- Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A. & Aiden, E.L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, **331**(6014): 176-182.
- Microsoft Web N-Gram Services. (2014). Microsoft Research. Retrieved November 30, 2014, from <http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx>
- Microsoft Web N-Gram Service Quick Start. (2014). *Microsoft Research*. Retrieved October 25, 2014, from <http://weblm.research.microsoft.com/info/QuickStart.htm>
- Mikheev, A., Moens, M. & GROVER, C. (1999). Named entity recognition without gazetteers. In: *Proceedings of the 9th conference on European chapter of the Association for Computational Linguistics, EACL '99*. Stroudsburg, PA: Association for Computational Linguistics, 1-8.
- Montemurro, M.A. (2001). Beyond the Zipf-Mandelbrot law in qualitative linguistics. *Physica A: Statistical Mechanics and its Applications*, **300** (3-4): 567-578.
- Mooers, C.E. (1950). Coding, information retrieval, and the rapid selector. *American Documentation*, **1**(4): 225–229.
- Nandi, A. & Jagadish, H.V. (2007). Effective Phrase Prediction. In: *Proceedings of the Conference on Very Large Databases, VLDB '07*. Vienna, Austria: ACM, 219–230.
- Newman, M.E.J. (2004). Power laws, Pareto distribution and Zipf's Law. *Contemporary Physics*, **46**(5): 323-351.
- n-gram. (2014, December 3). *Wikipedia, The Free Encyclopedia*. Retrieved December 4, 2014, from <https://en.wikipedia.org/wiki/N-gram>
- Oiaga, M. (2010, April 29). Microsoft Web N-gram Services Public Beta Opened to Academia. *SOFTPEDIA*. Retrieved December 8, 2014, from <http://archive.news.softpedia.com/news/Microsoft-Web-N-gram-Services-Public-Beta-Opened-to-Academia-140964.shtml>

## | References

- Olligschlaeger, A.M. & Hauptmann, A.G. (1999). Multimodal information systems and GIS: The Informedia Digital Video Library. In: *Proceedings of the 1999 ESRI User Conference*. San Diego, CA.
- Overell, S. (2009). *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. (Ph. D. thesis, Imperial College London).
- Overell, S. & Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, **22**(3): 265-287.
- Perea-Ortega, J.M., Ureña-López, L.A., García-Vega, M. & García-Cumbreras, M.A. (2009). Using Query Reformulation and Keywords in the Geographic Information Retrieval Task. In: C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G.J.F. Jones, M. Kurimo, T. Mandl, A. Peñas & V. Petras (eds.), *Evaluating Systems for Multilingual and Multimodal Information Access*. 9th Workshop of the Cross Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. *Lecture Notes in Computer Science*, Berlin, Germany: Springer, Vol. **5706**: 855-862.
- Population Division. Population Estimates and Projections Section. (2014). *United Nations, Department of Economic and Social Affairs*. Retrieved December 10, 2014, from <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>
- Pu, Q., He, D. & Li, Q. (2009). Query Expansion for Effective Geographic Information Retrieval. In: C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G.J.F. Jones, M. Kurimo, T. Mandl, A. Peñas & V. Petras (eds.), *Evaluating Systems for Multilingual and Multimodal Information Access*. 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. *Lecture Notes in Computer Science*, Berlin, Germany: Springer, Vol. **5706**: 843-850.
- Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S. & Yang, B. (2007). The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, **21**(7): 717-745.
- Rauch, E., Bukatin, M. & Barker, K. (2003). A confidence-based framework for disambiguating geographic terms. In: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references, HLT-NAACL-GEOREF '03*. Morristown, NJ: Association for Computational Linguistics, 50-54.
- RBS 6 Nations. (2015). *RBS 6 Nations*. Retrieved September 7, 2015, from <http://www.rbs6nations.com/en/home.php>
- Roberts, K., Bejan, C.A. & Harabagiu, S.M. (2010). Toponym disambiguation using events. In: *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS '10*.
- Robertson, S. (2004). Understanding Inverse Document Frequency: On theoretical arguments for IDF. *Journal of Documentation*, **60**(5): 503-520.
- Sanderson, M. & Croft, W.B. (2012). The History of Information Retrieval. *Proceedings of the IEEE*, **100** (Special Centennial Issue): 1444-1451.
- Sanderson M, & Kohler, J. (2004). Analyzing geographic queries. In: *Proceedings of Workshop on Geographic Information Retrieval, SIGIR '04*. New York, NY: ACM.
- Shaw, A. (2003). New media approaches to mapping humanitarian response. In: *Proceedings of the 2003 ESRI User conference*. San Diego, CA.
- Singhal, A. (2001). Modern information retrieval: a brief overview. *Bulletin of the IEEE computer society technical committee on data engineering*, **24**(4): 35-43.
- Smart, P.D., Jones, C.B. & Twaroch, F.A. (2010). Multi-Source Toponym Data Integration and Mediation for a Meta-Gazetteer Service. In: S.I: Fabrikant, T. Reichenbacher, M. van Kreveld & C. Schlieder (eds.), *Geographic Information Science*. 6th International Conference, GIScience 2010, Zurich, Switzerland, September 14-17, 2010. *Proceedings. Lecture Notes in Computer Science*, Berlin, Germany: Springer, Vol. **6292**: 234-248.

- Smith, D.A., Crane, G. (2001). Disambiguating geographic names in a historical digital library. In: P. Constantopoulos & I.T. Solvberg (eds.), *Research and Advanced Technology for Digital Libraries*. 5th European Conference, ECDL 2001 Darmstadt, Germany, September 4-9, 2001 Proceedings. *Lecture Notes in Computer Science*, Berlin, Germany: Springer, Vol. **2163**: 127-136.
- Smith, B. & Mark, D.M. (2001). Geographical categories: an ontological investigation. *International Journal of Geographical Information Science*, **15**(7): 591–612.
- Stamatatos, E. (2011). Plagiarism detection Using Stopword n-grams. *Journal of the American Society for Information Science and Technology*, **62**(12): 2512-2527.
- The World Through the Eyes of a Search Algorithm. (2014). *Oxford Internet Institute. Information Geographies at the Oxford Internet Institute*. Retrieved October 25, 2014, from <http://geography.oii.ox.ac.uk/?page=the-world-through-the-eyes-of-a-search-algorithm>
- Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46**: 234-240.
- Tomovic, A., Janicic, P. & Keselj, V. (2006). N-Gram-based classification and unsupervised hierarchical clustering of genome sequences. *Computer Methods and Programs in Biomedicine*, **81**(2): 137-153.
- VOLVO Dun Laoghaire Regatta. 9th-12th July 2015. (2015, September 7). *Dlregatta*. Retrieved September 7, 2015, from <http://www.dlregatta.org/>
- Waller, M.A. & Fawcett, S.E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, **34**(2): 77-84.
- Wang, K. & Li, X. 2009. Efficacy of a constantly adaptive language modeling technique for web scale application. In: *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 19-24 April, ICASSP '09*. Taipei, Taiwan: IEEE, 4733-4736.
- Wang, K., Thrasher, C., Viegas, E., Li, X. & Hsu, B.-J.P. (2010). An overview of Microsoft Web N-gram corpus and applications. In: *Proceedings of the NAACL HLT 2010 Demonstration Session*: 45-48.
- Whitney, L. (2010, December 17). Google's Ngram Viewer: A time machine for wordplay. *CNET*. Retrieved December 8, 2014, from <http://www.cnet.com/news/googles-ngram-viewer-a-time-machine-for-wordplay/>
- Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann.
- Zakopane. (2015, July 27). *Wikipedia, The Free Encyclopedia*. Retrieved September 8, 2015, from <https://en.wikipedia.org/wiki/Zakopane>
- Zhang, V.W., Rey, B., Stipp, E. & Jones, R., (2006). Geomodification in query rewriting. In: *Proceedings of the 2006 Workshop on Geographic Information Retrieval*. Seattle, WA: 23-27.
- Zipf, G.K. (1932). *Selective Studies and the Principle of Relative Frequency in Language*. Cambridge, MA: MIT Press.
- Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort*. Reading, MA: Addison-Wesley.
- Zubizarreta, A., de la Fuente, P., Cantera, J.M., Arias, M., Cabrero, J., Garcia, G., Llamas, C. & Vegas, J. (2009). Extracting geographic context from the Web: Georeferencing in MyMoSe. In: M. Boughanem, C. Berrut, J. Mothe, & C. Soule-Dupuy (eds.), *Advances in Information Retrieval*. 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings. *Lecture Notes in Computer Science*, Berlin, Germany: Springer, Vol. **5478**: 554-561.



## Appendix

*The employed Java code and lists are included in this segment. Moreover, a fraction of the excess results are displayed. The majority of the Java source code and input lists are long documents. Thus, the code and lists have been saved on a CD which is part of every hard copy of this master thesis. The location of the data on the CD is generally specified in the corresponding section.*

### A Code

The codes were written in the computer programming language Java in the development platform NetBeans IDE 8.0.1. Multiple codes were used to access the Microsoft Web N-Gram Service API and processing the data. These codes are available in the folder “code” on the CD accompanying this thesis. The folder contains another folder for each package which is divided in the folders: **build**, **dist**, **nbproject** and **src**. In the “dist” folder, the corresponding JavaDoc of the package is found. The Java class files are found in the folder “build”, while the source files (used code) can be found in the folder “src”. The codes making use of a user token from Microsoft Research have been modified, since the user token should not be used by multiple users. Hence, the user token has been extracted and replaced by “insert user token here”. Another user token for the Microsoft Web N-Gram Service can be requested by sending a mail to [webngram@microsoft.com](mailto:webngram@microsoft.com).

#### A.1 nGramProbability

This is a Java code accessing the Microsoft Web N-Gram Service API. It allows the user to type an input in a dialog box which is then queried in the Microsoft Web N-Gram Service. The joint probability, conditional probability and first 100 autocompletes of the input are returned.

#### A.2 nGram

The package nGram consists of four Java classes: **list**, **Probability**, **readFile** and **writeFile**. The main class Probability accesses the Microsoft Web N-Gram Service and returns the probabilities. Additionally, a code snippet focused on cleaning the list of city names. The class list contains the path of all the lists used as inputs. The readFile class reads the place name text files, while the class writeFile writes the probabilities retrieved from the Microsoft Web N-Gram Service to a text file.

### **A.3 nGrams**

The package nGram consists of three Java classes: **nGramProbs**, **ToponymTree** and **ToponymTreeTest**. The class nGramProbs accesses the Microsoft Web N-Gram Service and returns the probabilities and first 1000 autocompletes. The class ToponymTree defines the toponym tree structure for n-grams and creates hash maps. These contain a place name and place name ID. The main class ToponymTreeTest test the spatial relations of <a><spatial relation><b> between a topic <a> and a place name <b>. The path of the list of topics and place names is also defined. Generally, the code verifies if a place name <b> occurs in the 1000 autocompletes of place name <a>. It then returns the topic the matching place name IDs and conditional probability in a semicolon separated text file.

### **A.4 retrievedFound**

The package retrievedFound consists of four Java classes: **readFile**, **retrievedFound**, **spatialRelation\_list** and **writeFile**. The main class retrievedFound compares the received place names ID from the package nGrams in section A.3 with the constructed ground truth. Then the correctly retrieved and relevant found spatial relations are calculated for each entry and returned. The class list contains the paths of all the inputs and corresponding ground truth data files. The readFile class reads these input and ground truth text files, while the class writeFile writes the correctly retrieved and relevant found spatial relations to a text file.

## **B Lists**

In this appendix, the lists used during the course of the master thesis are shown. The lists can also be found as CSV files on the accompanied CD in the folder “lists”.

**B.1 Countries****Table B.1: List of countries**

Abkhazia	Egypt	Malawi	Sao Tome and Principe
Afghanistan	El Salvador	Malaysia	Saudi Arabia
Albania	Equatorial Guinea	Maldives	Senegal
Algeria	Eritrea	Mali	Serbia
Andorra	Estonia	Malta	Seychelles
Angola	Ethiopia	Marshall Islands	Sierra Leone
Antigua and Barbuda	Fiji	Mauritania	Singapore
Argentina	Finland	Mauritius	Slovakia
Armenia	France	Mexico	Slovenia
Australia	Gabon	Micronesia	Solomon Islands
Austria	Gambia	Moldova	Somalia
Azerbaijan	Georgia	Monaco	Somaliland
Bahamas	Germany	Mongolia	South Africa
Bahrain	Ghana	Montenegro	South Korea
Bangladesh	Greece	Morocco	South Ossetia
Barbados	Grenada	Mozambique	South Sudan
Belarus	Guatemala	Myanmar	Spain
Belgium	Guinea	Nagorno-Karabakh	Sri Lanka
Belize	Guinea-Bissau	Namibia	Sudan
Benin	Guyana	Nauru	Suriname
Bhutan	Haiti	Nepal	Swaziland
Bolivia	Honduras	Netherlands	Sweden
Bosnia and Herzegovina	Hungary	New Zealand	Switzerland
Botswana	Iceland	Nicaragua	Syria
Brazil	India	Niger	Taiwan
Brunei	Indonesia	Nigeria	Tajikistan
Bulgaria	Iran	Niue	Tanzania
Burkina Faso	Iraq	North Korea	Thailand
Burundi	Ireland	Northern Cyprus	Togo
Cambodia	Israel	Norway	Tonga
Cameroon	Italy	Oman	Transnistria
Canada	Ivory Coast	Pakistan	Trinidad and Tobago
Cape Verde	Jamaica	Palau	Tunisia
Central African Republic	Japan	Palestine	Turkey
Chad	Jordan	Panama	Turkmenistan
Chile	Kazakhstan	Papua New Guinea	Tuvalu
China	Kenya	Paraguay	Uganda
Colombia	Kiribati	Peru	Ukraine
Comoros	Kosovo	Philippines	United Arab Emirates
Cook Islands	Kuwait	Poland	United Kingdom
Costa Rica	Kyrgyzstan	Portugal	United States
Croatia	Laos	Qatar	Uruguay
Cuba	Latvia	Republic of the Congo	Uzbekistan
Cyprus	Lebanon	Romania	Vanuatu
Czech Republic	Lesotho	Russia	Vatican City
Democratic Republic of the Congo	Liberia	Rwanda	Venezuela
Denmark	Libya	Sahrawi Arab Democratic Republic	Vietnam
Djibouti	Liechtenstein	Saint Kitts and Nevis	Yemen
Dominica	Lithuania	Saint Lucia	Zambia
Dominican Republic	Luxembourg	Saint Vincent and the Grenadines	Zimbabwe
East Timor	Macedonia	Samoa	
Ecuador	Madagascar	San Marino	

**B.2 Capital Cities****Table B.2: List of capital cities**

Abu Dhabi	Dakar	Manama	San Jose
Abuja	Damascus	Manila	San Marino
Accra	Dhaka	Maputo	San Salvador
Addis Ababa	Dili	Maseru	Sana'a
Algiers	Djibouti	Mbabane	Santiago
Alofi	Dodoma	Mexico City	Santo Domingo
Amman	Doha	Minsk	Sao Tome
Amsterdam	Dublin	Mogadishu	Sarajevo
Andorra la Vella	Dushanbe	Monaco	Seoul
Ankara	El Aaiun	Monrovia	Singapore
Antananarivo	Freetown	Montevideo	Skopje
Apia	Funafuti	Moroni	Sofia
Ashgabat	Gaborone	Moscow	Sri Jayawardenepura Kotte
Asmara	Georgetown	Muscat	St. George's
Astana	Guatemala City	Nairobi	St. John's
Asuncion	Hanoi	Nassau	Stepanakert
Athens	Harare	Naypyidaw	Stockholm
Avarua	Hargeisa	N'Djamena	Sucre
Baghdad	Havana	New Delhi	Sukhumi
Baku	Helsinki	Ngerulmud	Suva
Bamako	Honiara	Niamey	Taipei
Bandar Seri Begawan	Islamabad	Nicosia	Tallinn
Bangkok	Jakarta	Nicosia	Tarawa Atoll
Bangui	Jerusalem	Nouakchott	Tashkent
Banjul	Juba	Nuku'alofa	Tbilisi
Basseterre	Kabul	Oslo	Tegucigalpa
Beijing	Kampala	Ottawa	Tehran
Beirut	Kathmandu	Ouagadougou	Thimphu
Belgrade	Khartoum	Palikir	Tirana
Belmopan	Kiev	Panama City	Tiraspol
Berlin	Kigali	Paramaribo	Tokyo
Bern	Kingston	Paris	Tripoli
Bishkek	Kingstown	Phnom Penh	Tskhinvali
Bissau	Kinshasa	Podgorica	Tunis
Bloemfontein	Kuala Lumpur	Port Louis	Ulaanbaatar
Bogota	Kutaisi	Port Moresby	Vaduz
Brasilia	Kuwait City	Port of Spain	Valletta
Bratislava	La Paz	Port Vila	Valparaiso
Brazzaville	Libreville	Port-au-Prince	Vatican City
Bridgetown	Lilongwe	Porto-Novo	Victoria
Brussels	Lima	Prague	Vienna
Bucharest	Lisbon	Praia	Vientiane
Budapest	Ljubljana	Pretoria	Vilnius
Buenos Aires	Lobamba	Pristina	Warsaw
Bujumbura	Lome	Putrajaya	Washington
Cairo	London	Pyongyang	Wellington
Canberra	Luanda	Quito	Windhoek
Cape Town	Lusaka	Rabat	Yamoussoukro
Caracas	Luxembourg	Ramallah	Yaounde
Castries	Madrid	Reykjavik	Yaren
Chisinau	Majuro	Riga	Yerevan
Conakry	Malabo	Riyadh	Zagreb
Copenhagen	Male	Rome	
Cotonou	Managua	Roseau	

### B.3 Cities

The list of cities contained more than 140 000 entries and was too long to be displayed in this appendix. However the list of cities with a population over 1000 inhabitants can be found on the CD. There are two lists: **cities1000** and **cities1000\_geonames\_clean**. The first list contains the original data from GeoNames. This list contains city names, alternative names, population and further information. The second list is a cleaned up version of the GeoNames list. It only contains the city names in ASCII format, while long city names containing parenthesis or any other special characters have been shortened.

### B.4 Geographic Features

**Table B.3: List of geographic features**

stream	ocean	valley
river	hill	forest
delta	mountain	desert
lake	volcano	beach
sea	plain	glacier

### B.5 Sports Activities

**Table B.4: List of sports activities**

archery	football	sailing
badminton	golf	skateboarding
baseball	handball	skiing
basketball	hiking	skydiving
bowling	hurling	snooker
boxing	judo	snorkelling
climbing	karate	snowboarding
cricket	lacrosse	soccer
curling	marathon	squash
cycling	mountaineering	surfing
darts	parkour	swimming
fencing	polo	tennis
fishing	rafting	triathlon
floorball	rugby	volleyball

## C Additional Results

In this section a portion of the produced results are displayed which did not make it into the main section. This is only a selection of the enormous number of results. Please contact the author of this master thesis for further insights into the remaining outputs.

### C.1 Country Name Correlation per Continent

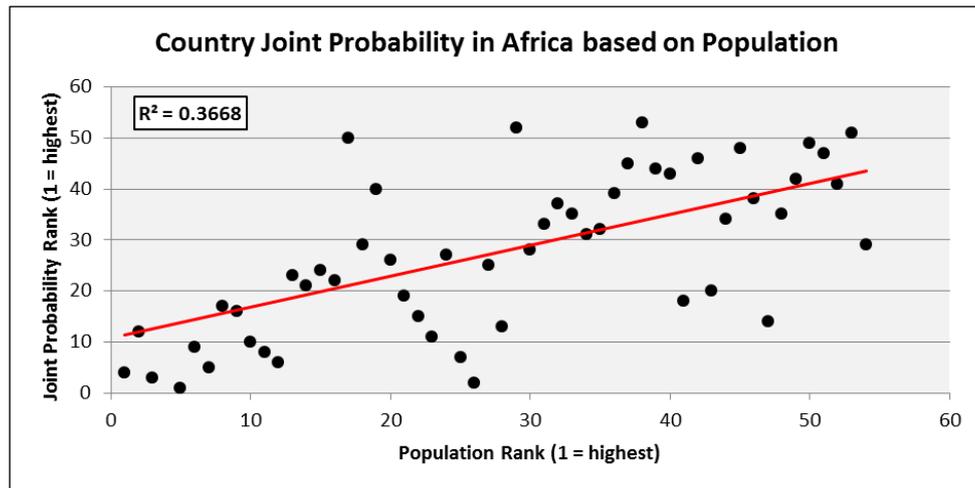


Fig. C.1: Rank correlation in Africa between country name joint probability and population

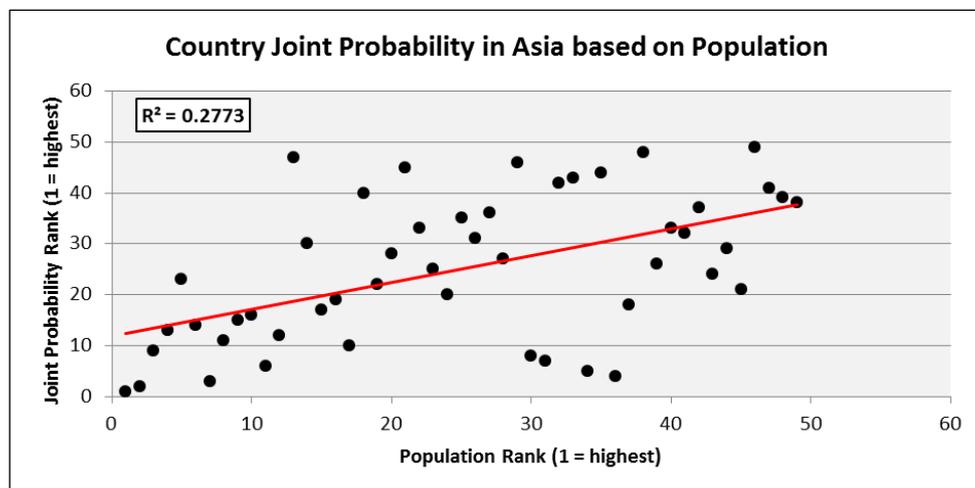


Fig. C.2: Rank correlation in Asia between country name joint probability and population

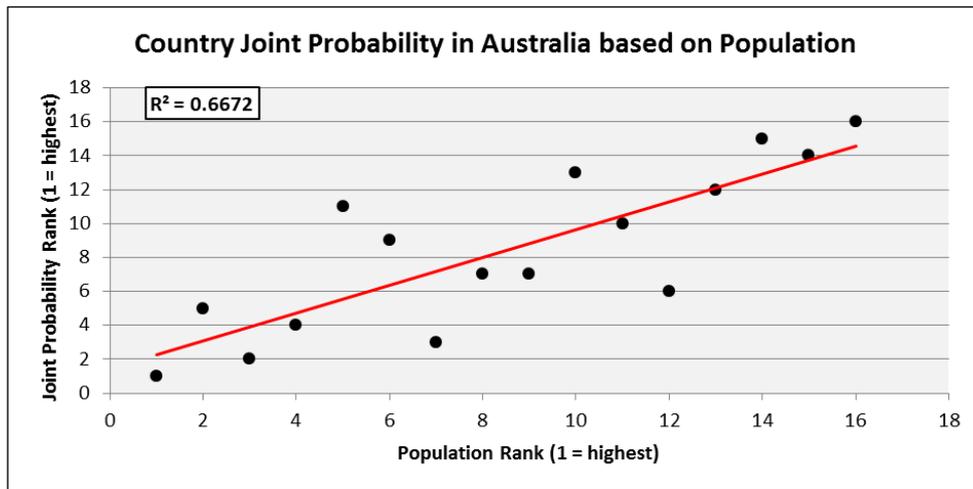


Fig. C.3: Rank correlation in Australia between country name joint probability and population

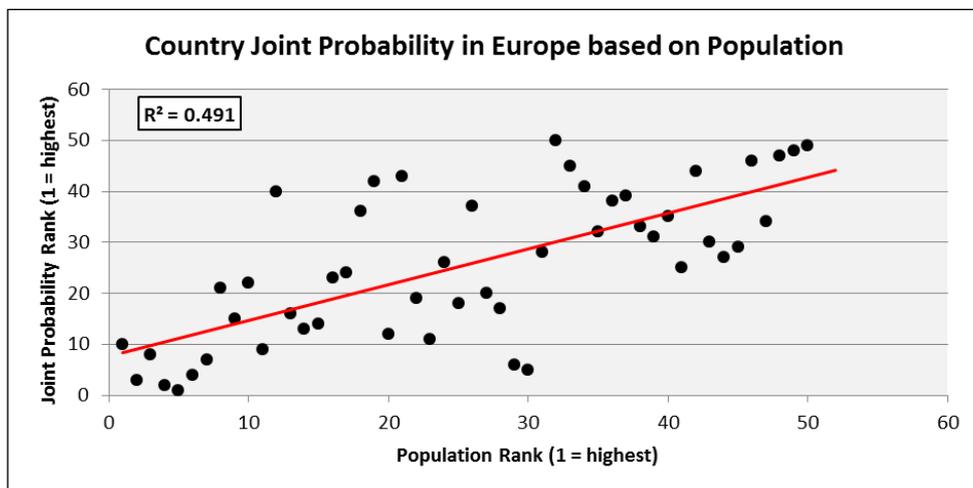


Fig. C.4: Rank correlation in Europe between country name joint probability and population

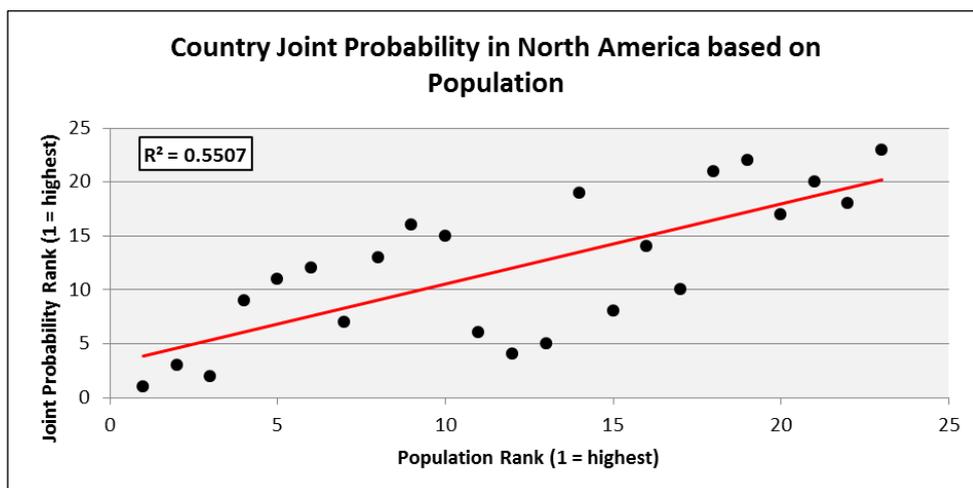


Fig. C.5: Rank correlation in North America between country name joint probability and population

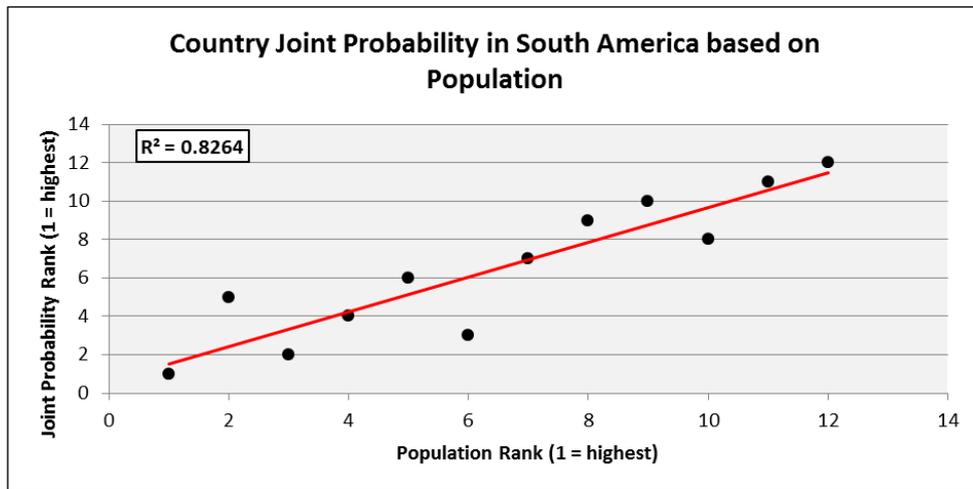


Fig. C.6: Rank correlation in South America between country name joint probability and population

## C.2 Capital City Name Correlation per Continent

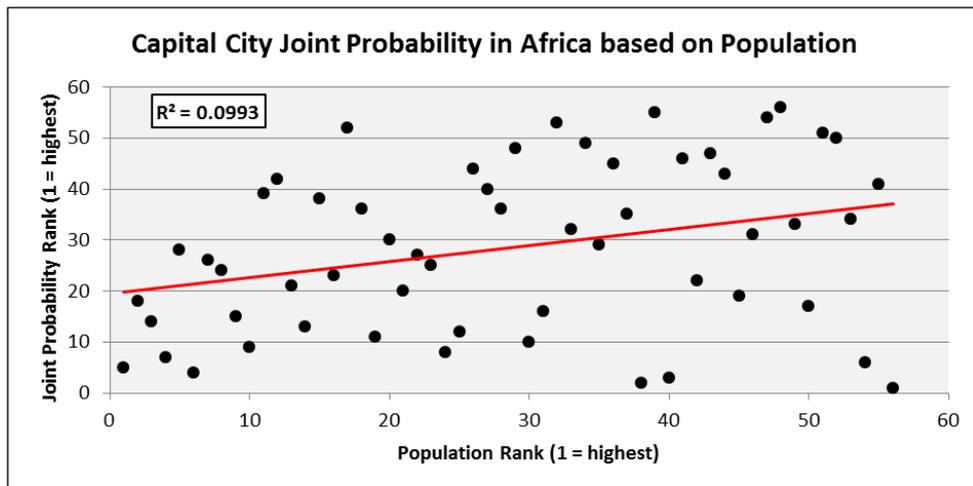


Fig. C.7: Rank correlation in Africa between capital city name joint probability and population

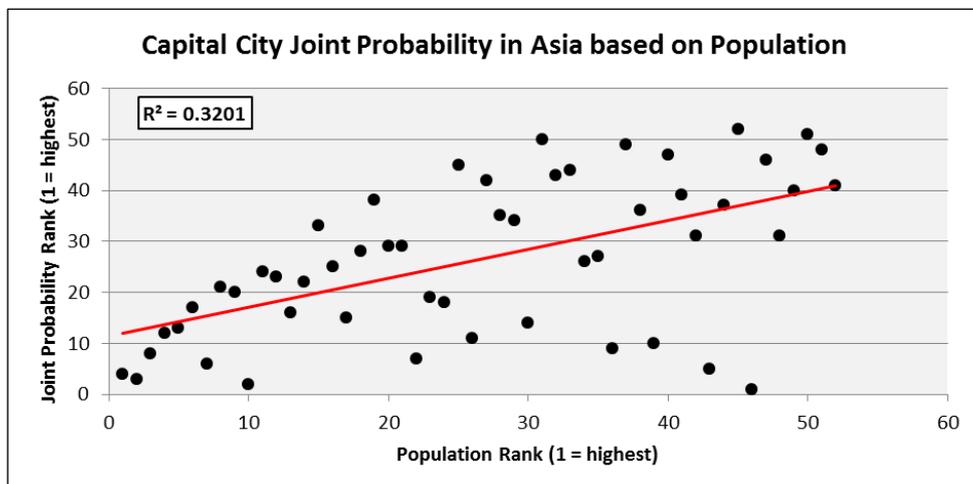


Fig. C.8: Rank correlation in Asia between capital city name joint probability and population

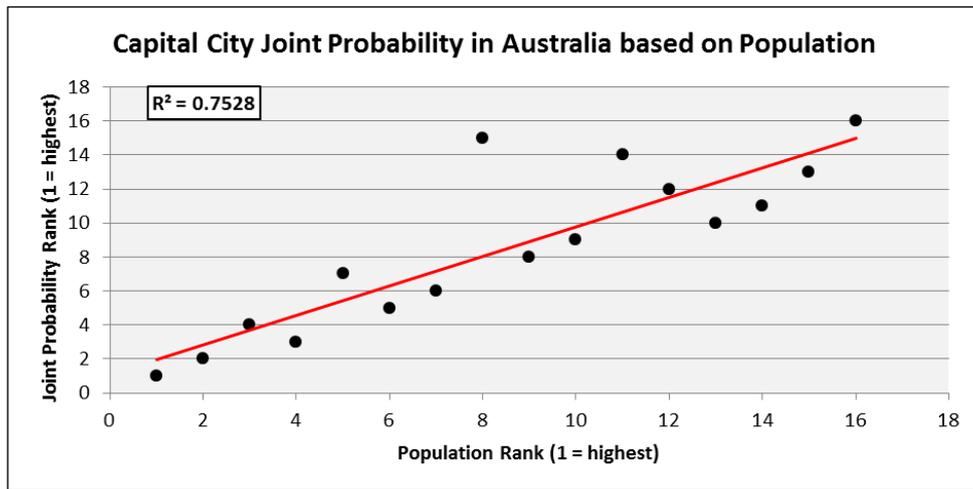


Fig. C.9: Rank correlation in Australia between capital city name joint probability and population

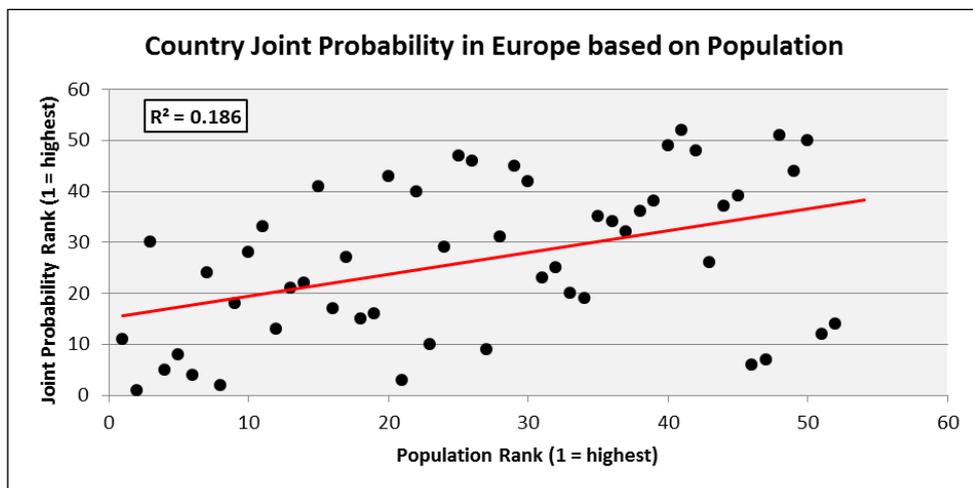


Fig. C.10: Rank correlation in Europe between capital city name joint probability and population

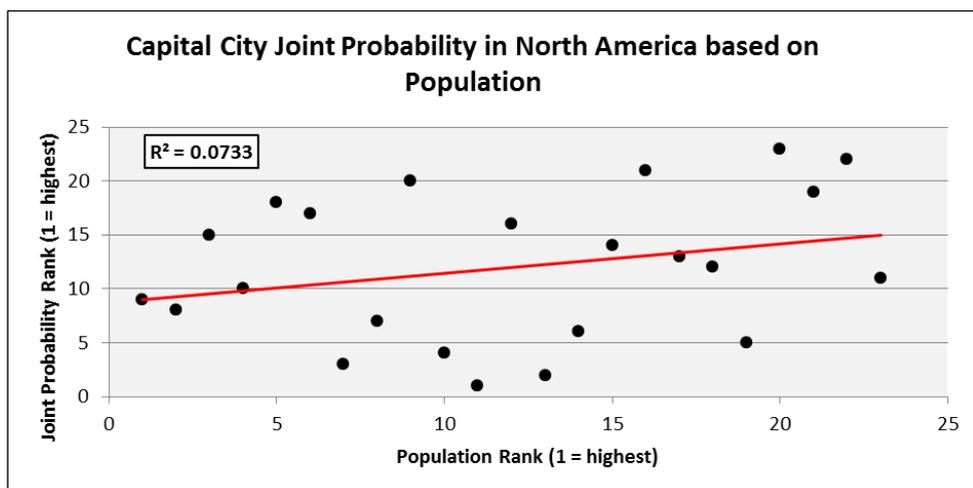


Fig. C.11: Rank correlation in North America between capital city name joint probability and population

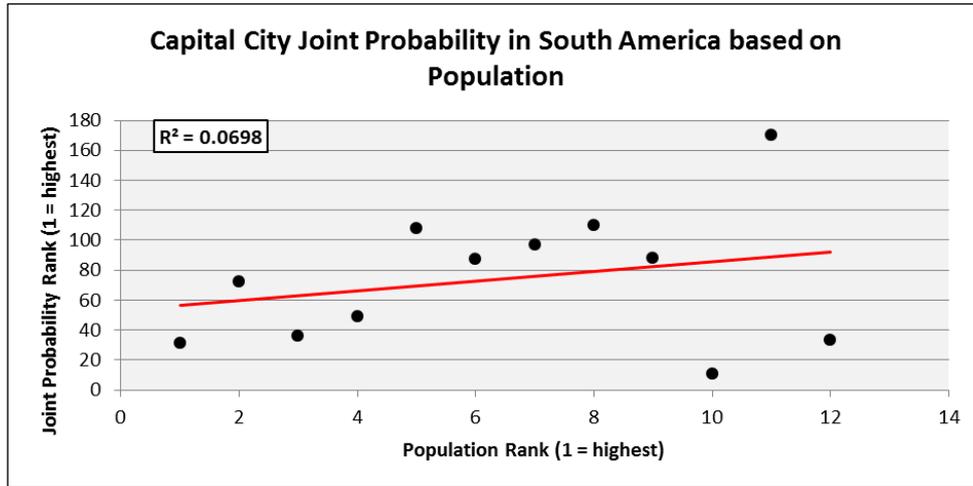


Fig. C.12: Rank correlation in South America between capital city name joint probability and population

### C.3 Spatial Autocorrelation

A detailed summary of the results from the spatial autocorrelation in ArcGIS is available on the accompanied CD. These results are in the folder SpatialAutocorrelation. The folder contains the report to the spatial autocorrelation as HTML file with details to the Moran’s Index, expected index, variance, z-score and p-value.

### C.4 Geographic Feature in/near Country

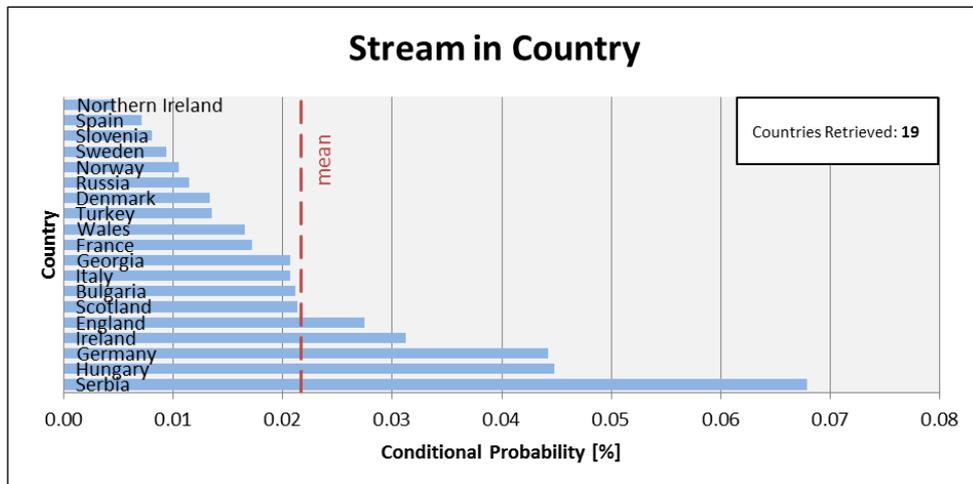


Fig. C.13: Conditional probability of the 20 most likely countries to follow “stream in”

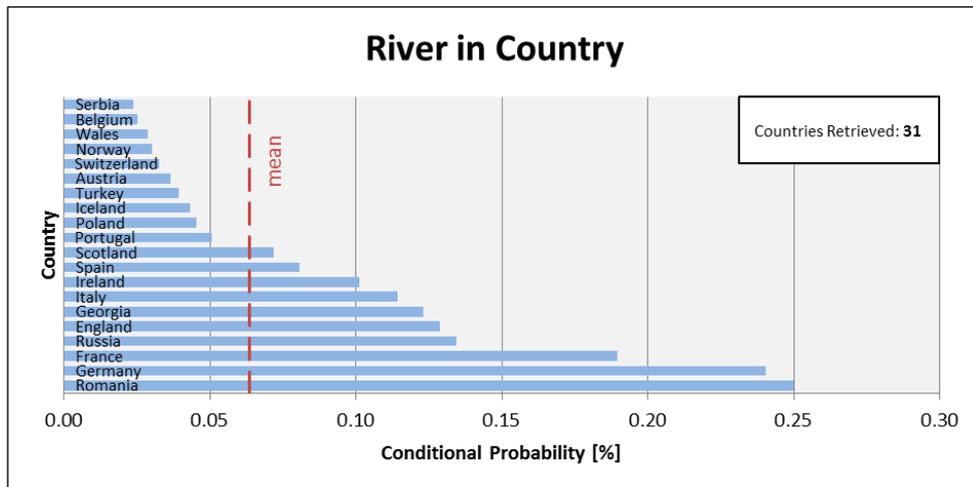


Fig. C.14: Conditional probability of the 20 most likely countries to follow “river in”

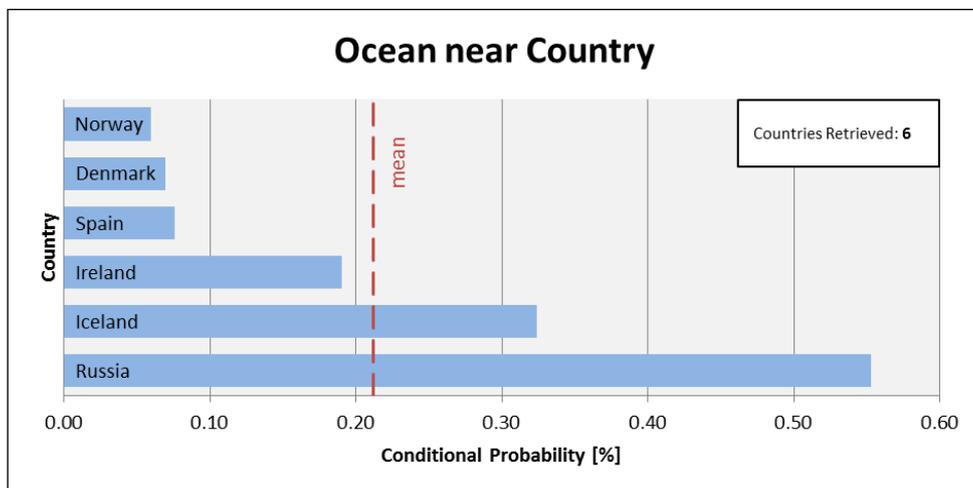


Fig. C.15: Conditional probability of the 20 most likely countries to follow “ocean near”

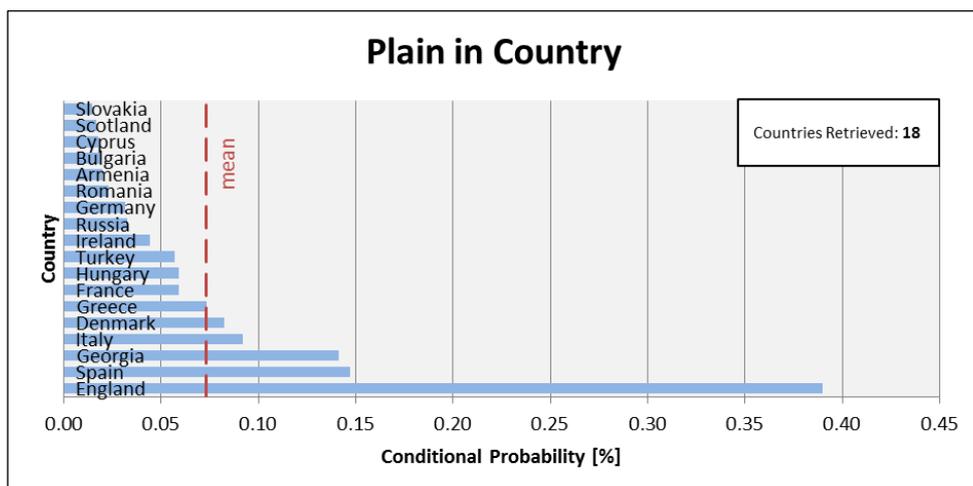


Fig. C.16: Conditional probability of the 20 most likely countries to follow “plain in”

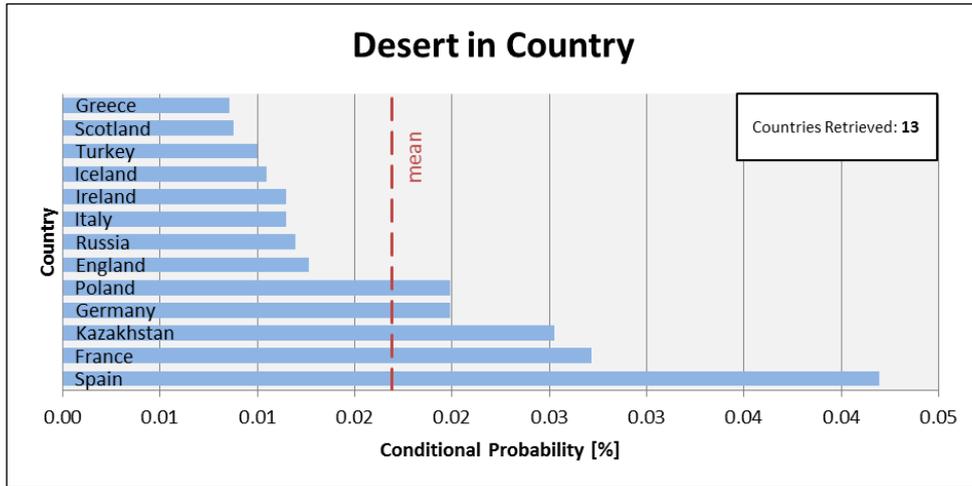


Fig. C.17: Conditional probability of the 20 most likely countries to follow “desert in”

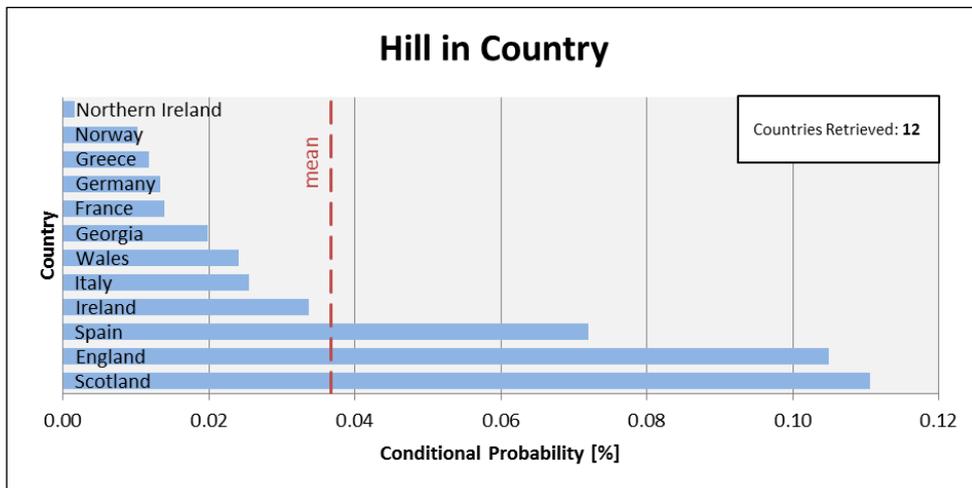


Fig. C.18: Conditional probability of the 20 most likely countries to follow “hill in”

### C.5 Geographic Features near City

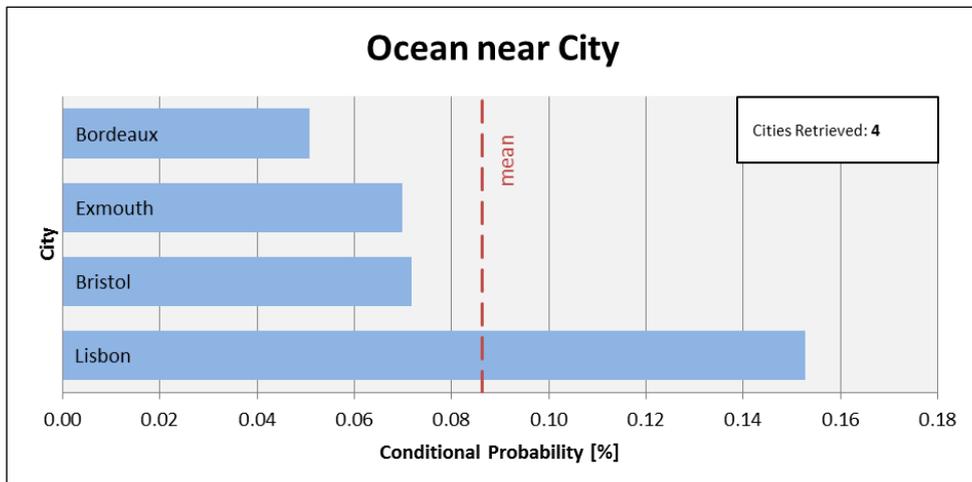


Fig. C.19: Conditional probability of the 20 most likely cities to follow “ocean near”

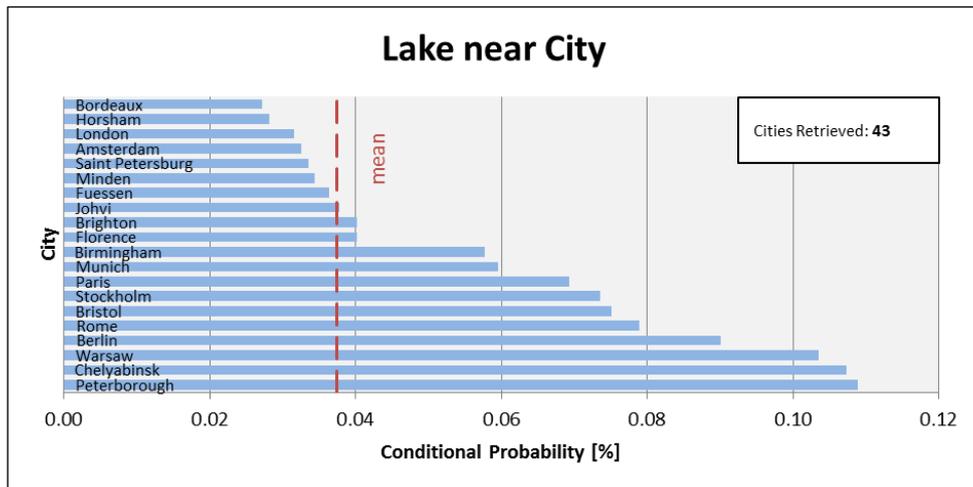


Fig. C.20: Conditional probability of the 20 most likely cities to follow “lake near”

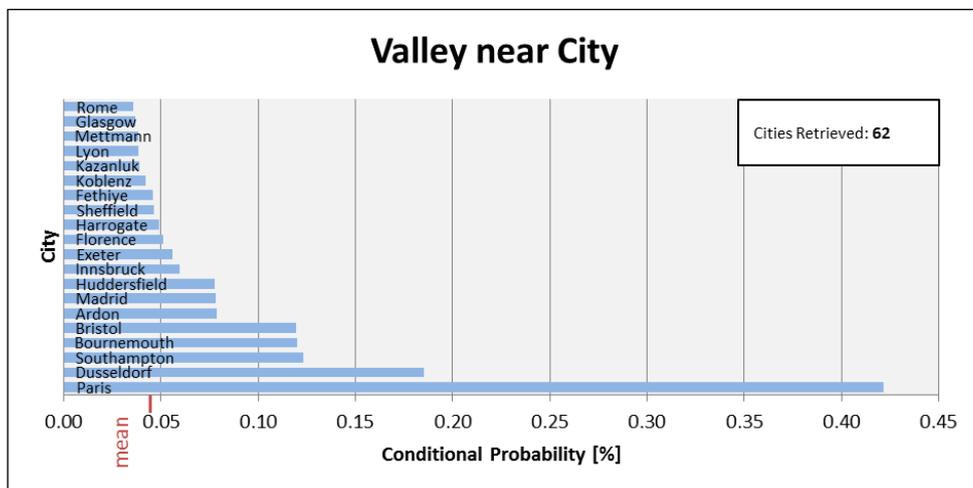


Fig. C.21: Conditional probability of the 20 most likely cities to follow “valley near”

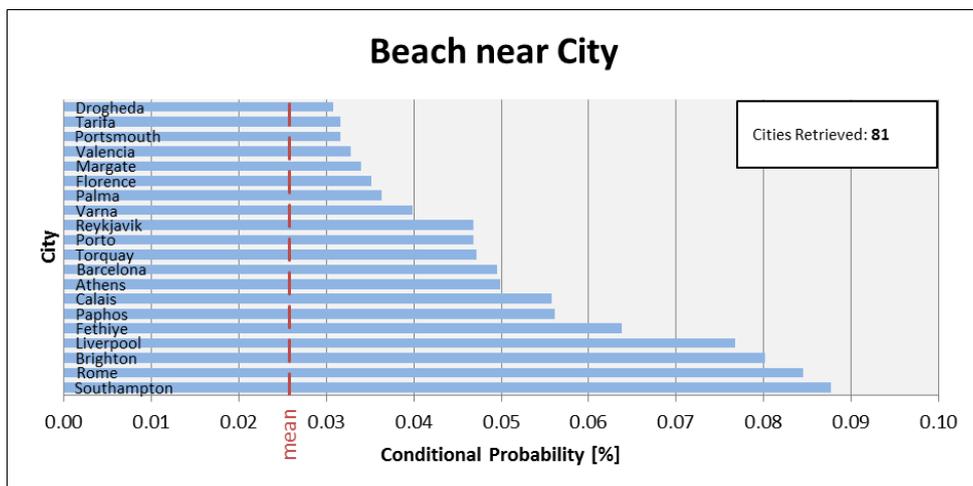


Fig. C.22: Conditional probability of the 20 most likely cities to follow “beach near”

### C.6 Sport Activity in Country

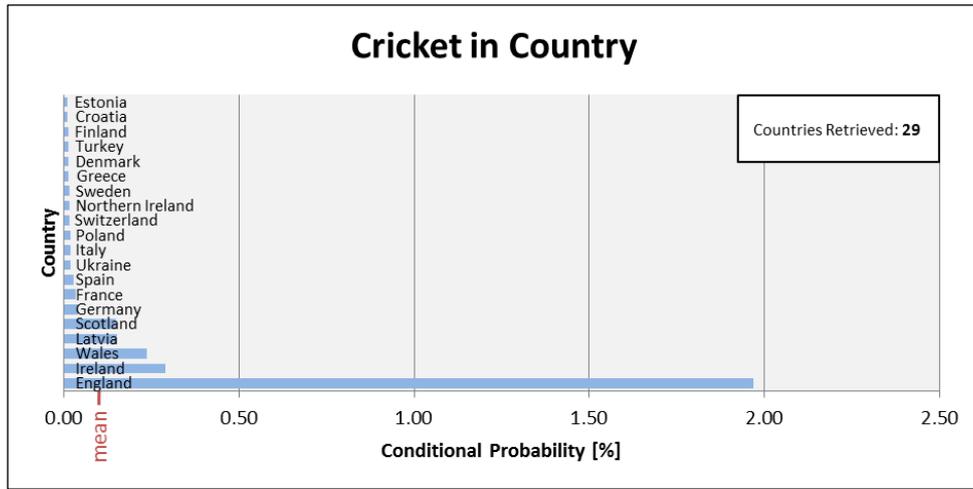


Fig. C.23: Conditional probability of the 20 most likely countries to follow “cricket in”

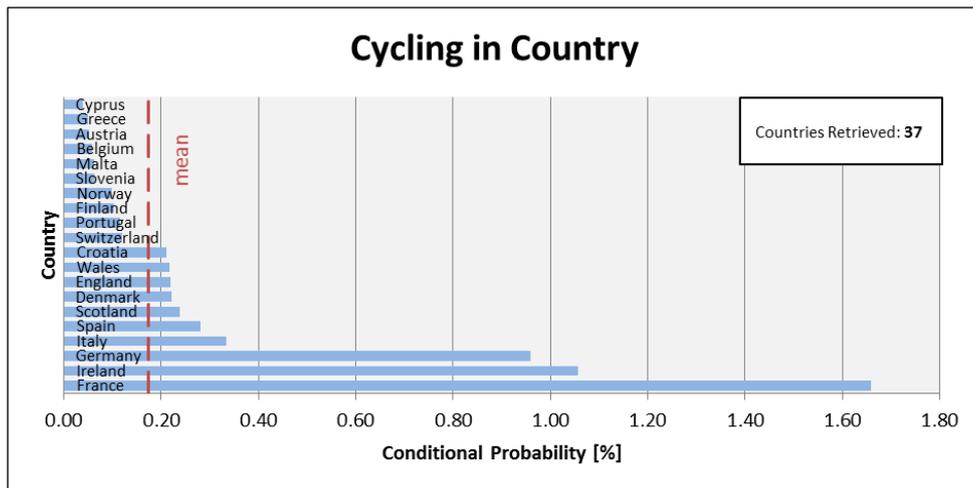


Fig. C.24: Conditional probability of the 20 most likely countries to follow “cycling in”

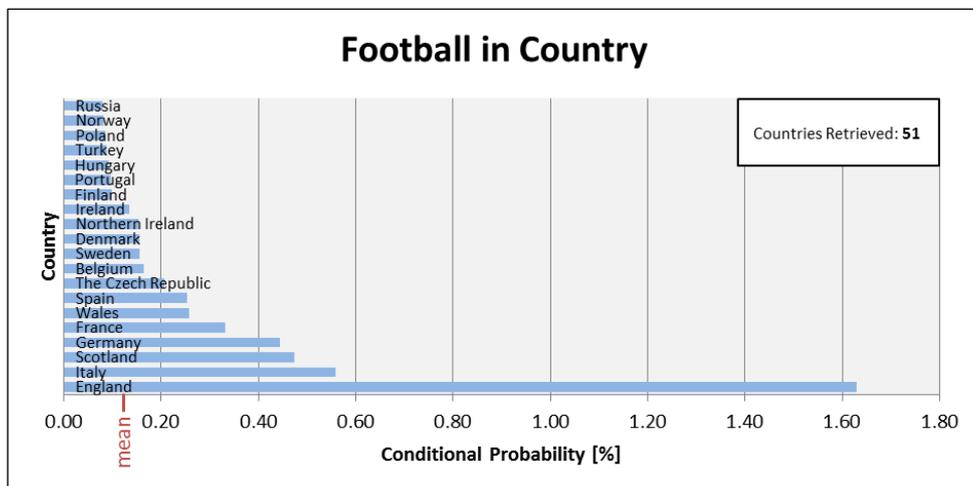


Fig. C.25: Conditional probability of the 20 most likely countries to follow “football in”

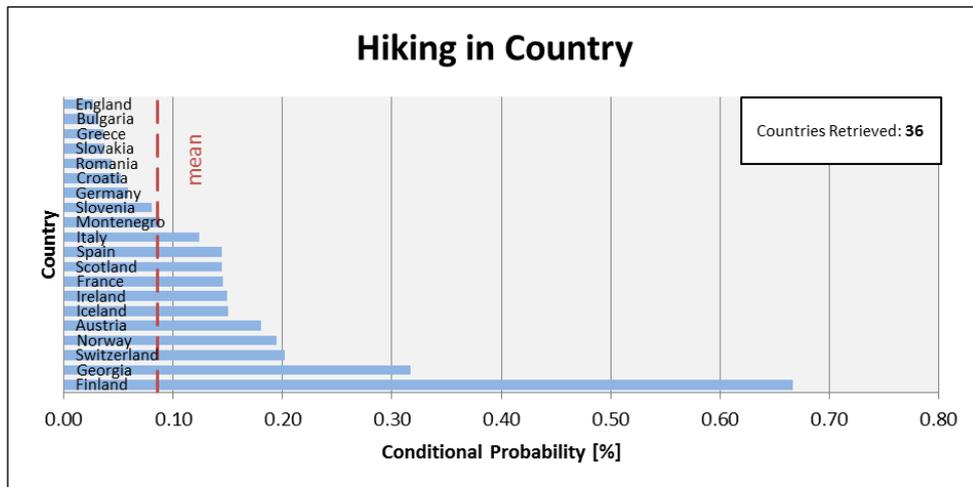


Fig. C.26: Conditional probability of the 20 most likely countries to follow “hiking in”

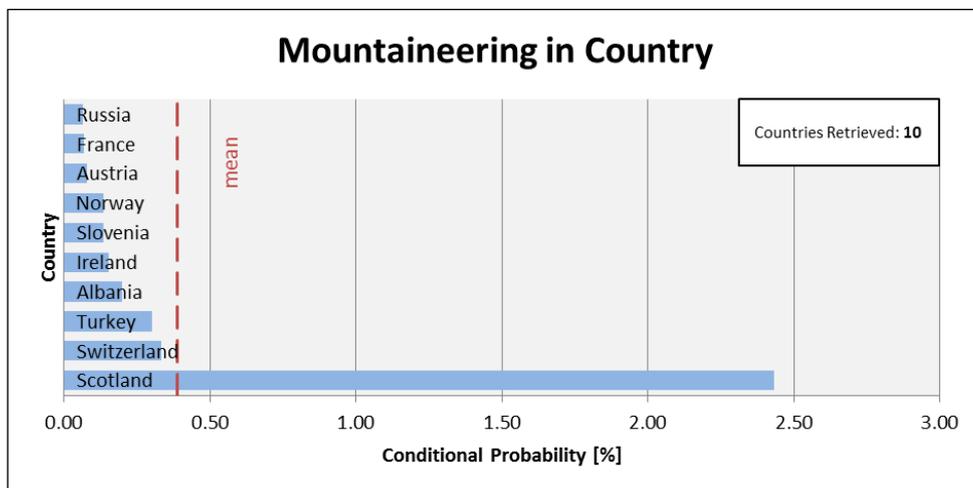


Fig. C.27: Conditional probability of the 20 most likely countries to follow “mountaineering in”

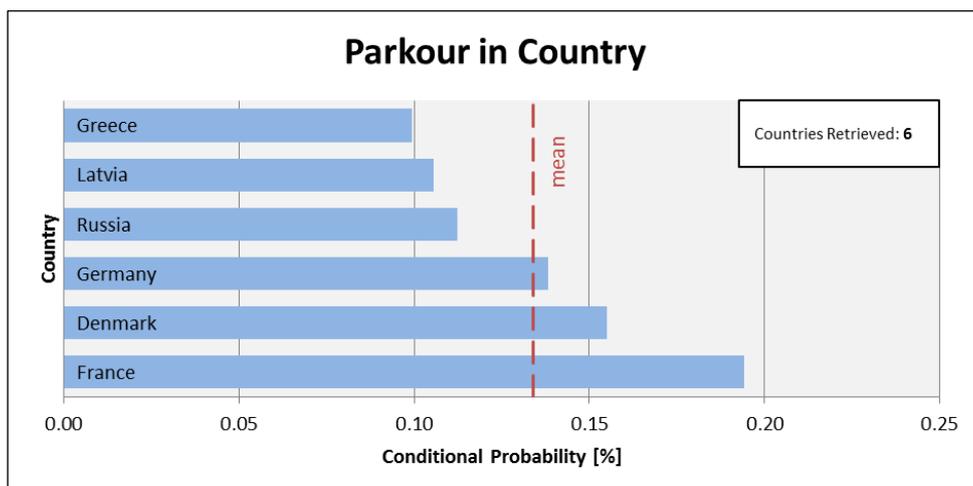


Fig. C.28: Conditional probability of the 20 most likely countries to follow “parkour in”

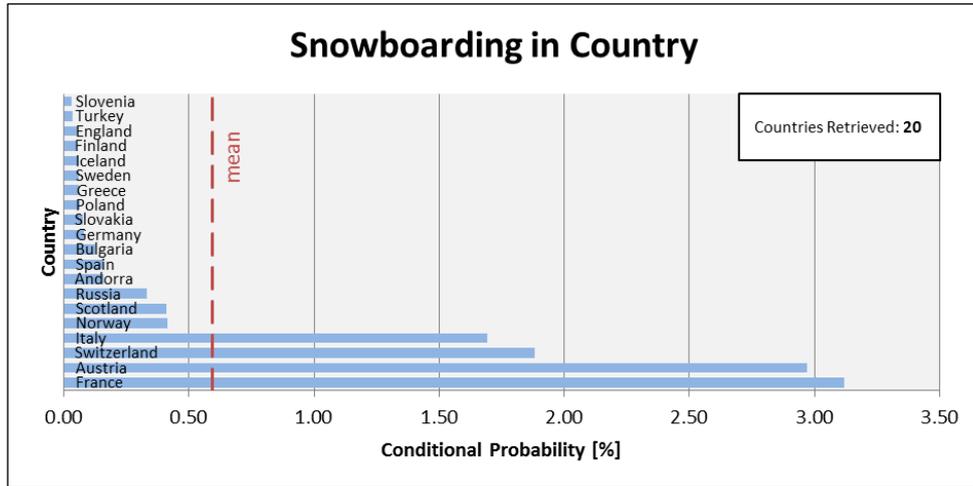


Fig. C.29: Conditional probability of the 20 most likely countries to follow “snowboarding in”

### C.7 Sport Activity in City

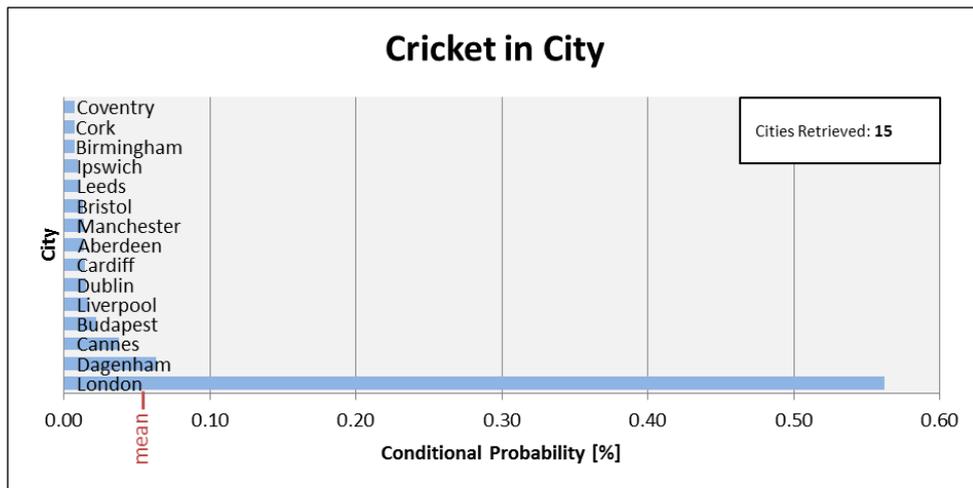


Fig. C.30: Conditional probability of the 20 most likely cities to follow “cricket in”

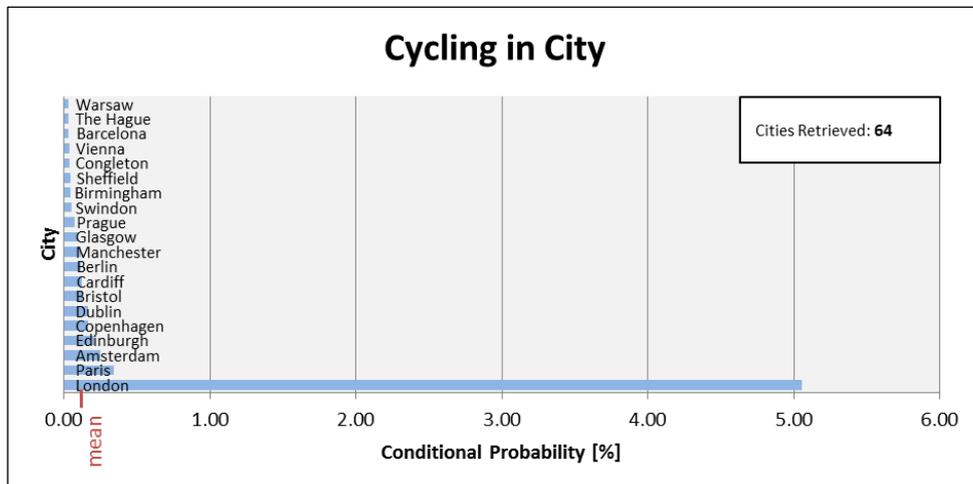


Fig. C.31: Conditional probability of the 20 most likely cities to follow “cycling in”

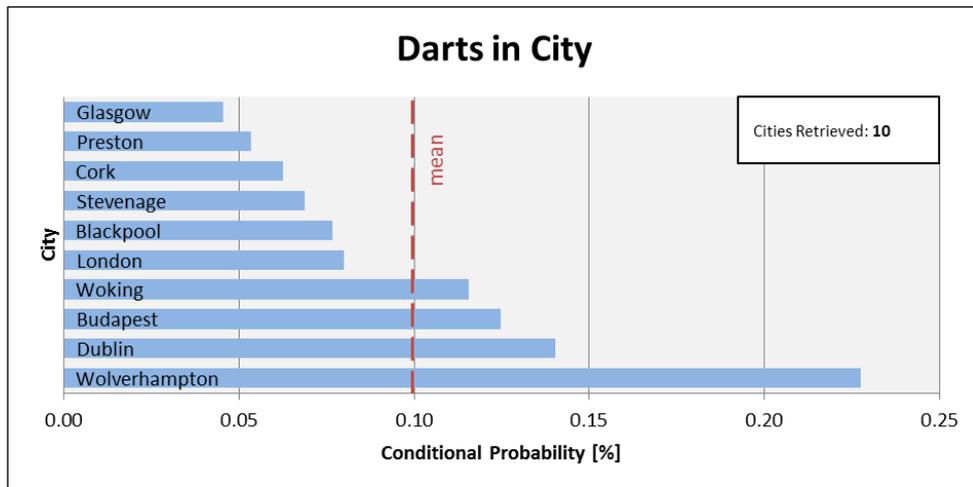


Fig. C.32: Conditional probability of the 20 most likely cities to follow “darts in”

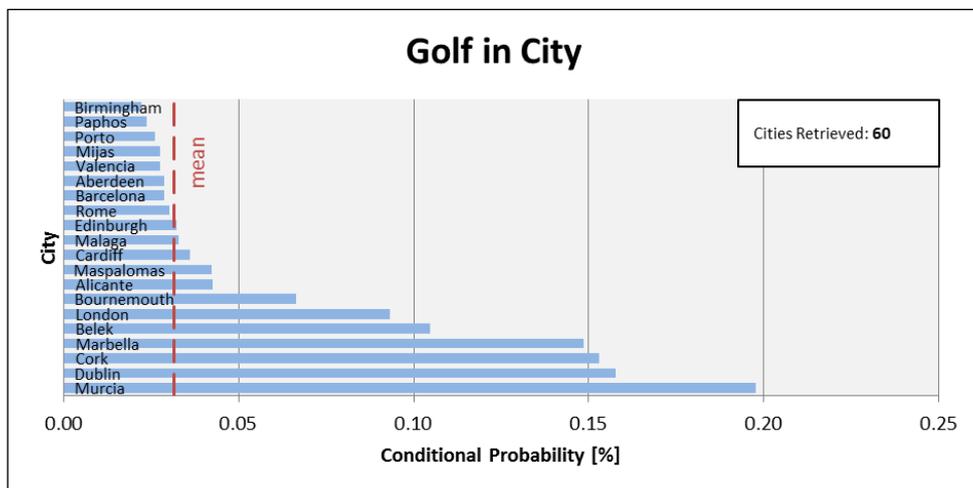


Fig. C.33: Conditional probability of the 20 most likely cities to follow “golf in”

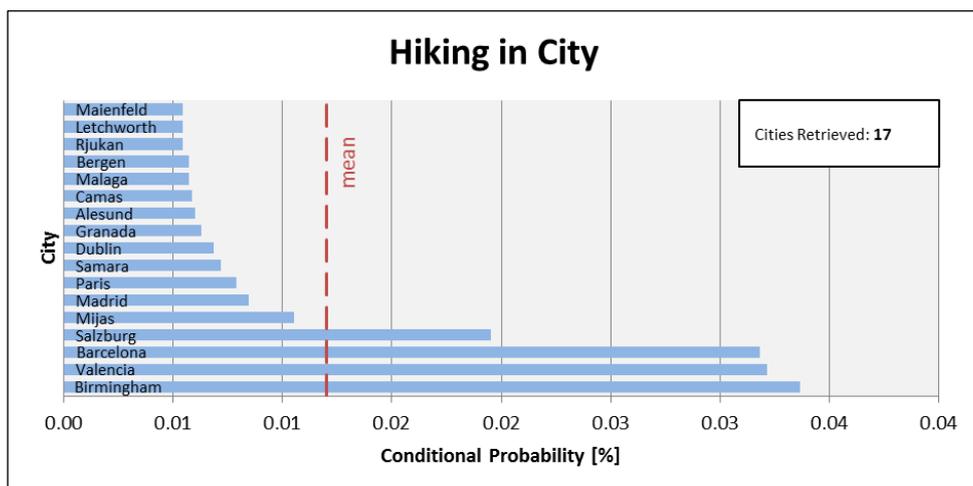


Fig. C.34: Conditional probability of the 20 most likely cities to follow “hiking in”

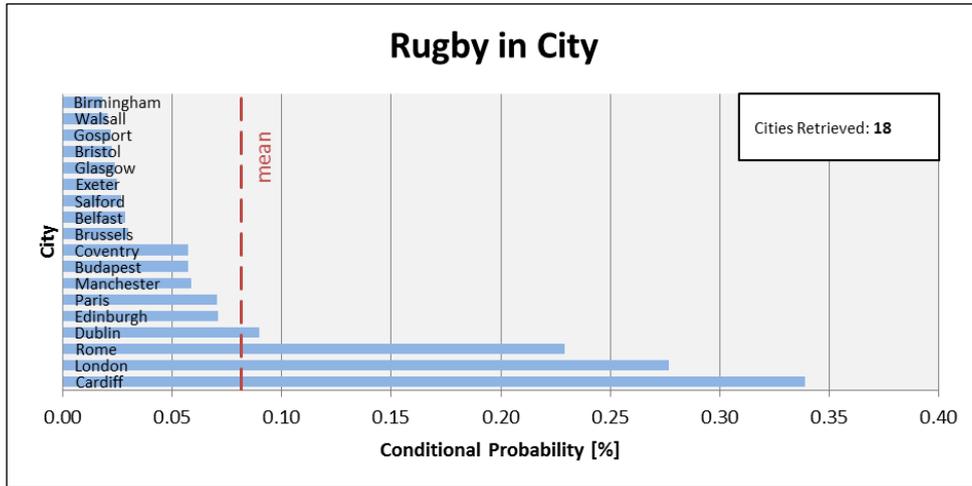


Fig. C.35: Conditional probability of the 20 most likely cities to follow “rugby in”

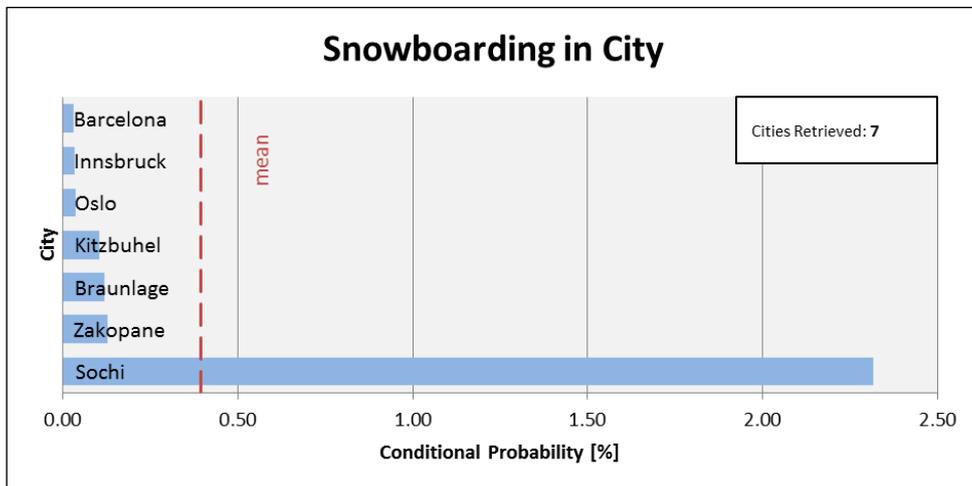


Fig. C.36: Conditional probability of the 20 most likely cities to follow “snowboarding in”

## **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.

Place and date :

---

Signature:

---

Jérôme Sautier