Georeferenzierung von nutzergenerierten Daten

Thomas Wider
Matrikelnummer: 04-397-386

GEO 511: Masterarbeit

Geocomputation
Geographisches Institut, Universität Zürich

Betreuung: Prof. Dr. Ross Purves (Fakultätsmitglied)
Dr. Damien Palacio

Zürich, 31. Januar 2013

Kontakt

Autor

Thomas Wider
Scheuchzerstrasse 223
8057 Zürich

twider@geo.uzh.ch

Betreuer

Ross Purves, Damien Palacio
Geocomputation
Geographisches Institut
Universität Zürich-Irchel
Winterthurerstrasse 190
8057 Zürich

ross.purves@geo.uzh.ch | damien.palacio@geo.uzh.ch

Zusammenfassung

Mit der zunehmenden Verbreitung von Internetanwendungen mit geographischem Bezug haben sich die Benutzergruppen von Geographischen Informationssystemen (GIS) in den letzten Jahren auf eine breite Öffentlichkeit ausgedehnt. Die Verwendung und Erstellung von geographischen Informationen ist nicht mehr alleinige Domäne von Spezialisten, sondern auch für Anwender ohne Expertenwissen möglich. Ein Effekt davon sind eine Fülle von nutzergenerierten Daten (UGC), die von vielen Nutzern erstellt und für die Öffentlichkeit verfügbar gemacht werden. Ein typisches Beispiel von UGC ist die gemeinschaftlich erstellte Weltkarte von OpenStreetMap[1].

Die Relevanz der Ergebnisse bei der Suche nach bestimmten Informationen mit räumlichem Kontext ist für Suchmaschinen jedoch verbesserungsfähig, da sie meistens lediglich Schlüsselwörter zur Suche verwenden. Die Ursache ist die Diskrepanz zwischen den räumlichen Konzepten der Nutzer zur Suche und deren Implementierung in Suchmaschinen. Eine mögliche Lösung dieser sogenannten „semantischen Lücke“ ist die Analyse der natürlichen Sprache, welche von Menschen zur Beschreibung von UGC verwendet wird. In Verbindung mit den in den Metadaten teilweise vorhandenen geographischen Koordinanten können lokale Verwendungen von Ortsbezeichnungen extrahiert und in einem umgangssprachlichen Ortsverzeichnis gespeichert werden.

In dieser Arbeit werden dazu georeferenzierte Flickr-Photos[2] aus Grossbritannien und der Schweiz verwendet. Mit den ortsrelevanten Tags lässt sich die Position von anderen Photos mithilfe ihrer textuellen Beschreibung als Zelle eines regelmässigen Gitternetzes vorhersagen. Diese Georeferenzierung dient zum einen zur Evaluation der verwendeten Methode zur Extraktion von ortsrelevanten Bezeichnungnen und zum anderen als mögliche Anwendung zur Anreicherung von UGC mit geographischen Koordinaten. Mit einer Zellgrösse von 1 km lassen sich für rund 40 % aller Photos die korrekte Zelle und für über 65 % eine Zelle innerhalb eines Radius von 5 km um die korrekte Zelle herum vorhersagen.

Weiter zeigt die Arbeit auf, dass Nutzer Photos ohne geographische Koordinaten im Durchschnitt mit mehr Schlüsselwörtern beschreiben und dabei mehr Ortsnamen verwenden. Zudem weist die Arbeit den Einfluss von Nutzerverzerrungen auf die Ergebnisse nach und zeigt den Effekt deren Filterung.

[1] http://www.openstreetmap.org/
[2] http://www.flickr.com/

Summary

In the past few years applications with geographic context on the Internet became increasingly accessible to a broader public and caused a shifting in the use of Geographic Information Systems (GIS). The usage and creation of geograpic information – formerly restricted to specialsts – is now open for a broad range of users without expert knowledge. A wide choice of publicly available user-generated content (UGC) is one of the outcomes of this recent development. A typical example of UGC is the collaborative world map OpenStreetMap[3].

However, the relevance of the results of a query for specific information in a spatial context is limited for search engines using only keyword search methods. A reason for this shortcoming is the discrepancy between the human concepts to perceive and describe spatial knowledge and the corresponding implementation in search engine methods. A possible solution to solve the problem of the so-called „semantic gap“ lies in the analysis of the natural language humans use to describe UGC where geographic coordinates are provided in the metadata. Subsequently, local use of placenames can be extracted and stored to a vernacular placename gazetteer.

For this master thesis georeferenced photos from Flickr[4] serve as a source to extract the most relevant local designations for places in Great Britain and Switzerland. This allows the prediction of the position in a grid of equally sized cells for photos only using their textual decriptions. This georeferencing evaluates the method used to extract placenames on the one hand and shows a possible application for the enrichment of UGC with geographic coordinates on the other hand. With a cell size of 1 km it is possible to predict the correct cell for approximately 40 % or a cell within a radius of 5 km around the correct cell for approximately 65 % of the photos.

Furthermore, the thesis shows that users tend to describe photos without geographic coordinates more detailed and with more toponyms. Moreover, the thesis explores the influence of user bias on the results and shows the effects of filtering.

[3] http://www.openstreetmap.org/
[4] http://www.flickr.com/

Vorwort

Geographische Informationswissenschaft (GIScience) – am Anfang meines Studiums wusste ich nicht einmal, dass es diese Disziplin überhaupt gibt. Die Schnittmenge von Geographie und Informatik weckte allerdings meine Neugierde und bildete die Motivationsgrundlage für mein Studium, welches mit der vorliegenden Masterarbeit seinen Abschluss gefunden hat. Welche Informationen in Daten – im Speziellen in nutzergenerierten Daten (UGC) – stecken, wie diese extrahiert und das darin enthaltene Wissen nutzbar gemacht werden können, war mein Ansporn für die Themenwahl dieser Arbeit.

Zahlreiche Personen haben mich im Laufe der Arbeit in allen möglichen Bereichen beraten und unterstützt. Deshalb danke ich ganz herzlich...

... meinen beiden Betreuern Ross Purves und Damien Palacio für die zahlreichen Ratschläge, Hilfestellungen und die konstruktive Kritik.

... Arzu Çöltekin und Kenan Bektaş für die Unterstützung bei der Prozessierung auf der Workstation.

... Stephanie Tuggener und Ralph Straumann für das Korrekturlesen und die hilfreichen Kommentare.

... Irina Meister, die mich während den letzten Monaten meiner Arbeit liebevoll umsorgt hat.

... meinen Eltern für die stetige Unterstützung meiner Ausbildung während der letzten Jahre.

... meinen Freunden, Kollegen und Bekannten für die moralische Unterstützung und die aufmunternden Worte bei unzähligen Mittagessen und Kaffeepausen.

Danke, merci, grazie, thank you!

Tom, 31. Januar 2013

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungen

API	*Application Programming Interface* – Programmierschnittstelle
CH	*Switzerland* – Schweiz (Confoederatio Helvetica)
GB	*Great Britain* – Grossbritannien (England, Wales und Schottland)
GIR	*Geographical Information Retrieval* – Abruf geographischer Informationen
GIS	*Geographic Information System* – Geographisches Informationssystem
GIScience	*Geographic Information Science* – Geographische Informationswissenschaft
GPS	*Global Positioning System* – globales Navigationssatellitensystem des US-Verteidigungsministeriums
IR	*Information Retrieval* – Abruf von Informationen
KDE	*Kernel Density Estimation* – Kerneldichteschätzung
MBR	*Minimum Bounding Rectangle* – Minimal umgebendes Rechteck
ML	*Machine Learning* – maschinelles Lernen
MLE	*Maximum Likelihood Estimation* – Schätzung der maximalen Wahrscheinlichkeit
NB	*Naive Bayes* – probabilistischer Klassifikator basierend auf Bayes' Theorem
NER	*Named Entity Recognition* – Verfahren zur Identifizierung und Klassifikation von Textelementen in vordefinierte Kategorien
SVM	*Support Vector Machines* – Klassifikator zur Unterteilung von Objekten in Klassen mit Hilfe von statistischem Lernen
TF-IDF	*Term Frequency, Inverse Document Frequency* – Vorkommenshäufigkeit, inverse Dokumenthäufigkeit
UGC	*User-Generated Content* – nutzergenerierte Daten
VGI	*Volunteered Geographic Information* – UGC mit geographischem Bezug
WGS84	*World Geodetic System 1984* – Geographisches Referenzsystem für den weltweiten Gebrauch

1 Einleitung

Dieses Kapitel zeigt die Problemstellung dieser Arbeit und begründet den Stellenwert und die Motivation des Themas.

1.1 Problemstellung und Motivation

Die zunehmende Verbreitung von Internet-Anwendungen mit räumlichem Bezug hat in den letzten Jahren zu einer Verschiebung und Erweiterung der Benutzergruppen von Geographischen Informationssystem (GIS) geführt. Als prominente Beispiele können der Kartendienst von Google[1] oder der virtuelle Globus der NASA[2] angeführt werden. Während die Benutzung von GIS früher vor allem Experten vorbehalten war, ist der Zugang zu geographischen Informationen und Anwendungen im Zuge der fortschreitenden Entwicklung für eine breite Öffentlichkeit ohne Expertenwissen möglich geworden (Sui & Goodchild, 2011).

Mit dem Wachstum des Internets erfolgte auch die Erkenntnis, dass ohne geeignete Methoden zur Suche und zum Empfangen von relevanten Informationen aus der Fülle aller verfügbaren Informationen kaum ein Nutzen gezogen werden kann. Verfügbare Suchmethoden sind meist auf die Suche mit Schlüsselwörtern beschränkt und können die von den Nutzern verwendete Semantik in den Suchanfragen – also die Bedeutung der verwendeten Begriffe – nicht verarbeiten (Egenhofer, 2002), was als semantische Lücke bezeichnet wird (Smeulders et al., 2000). Oft erfolgen die Suchanfragen in einem geographischen Kontext und beinhalten somit räumliche Semantik, die von der Suchmethode nicht interpretiert werden kann (Sanderson & Kohler, 2004).

Ein Beispiel der semantischen Lücke illustriert eine Suchanfrage nach dem „Mitteland“ im Kartendienst der Suchmaschine von Google. Diese liefert als Resultat einen Standort inmitten eines Waldes im Kanton Appenzell Ausserrhoden. Das „Mitteland“ ist zwar eine verbreitete umgangssprachliche Ortsbezeichnung und hat auch eine räumliche Ausdehnung (Purves et al., 2005), welche allerdings durch die fehlende Definition von randscharfen Grenzen, der fehlenden Verwendung als offiziellen Ortsnamen und durch technische Hürden nur in den Köpfen der Menschen existiert, nicht aber in der Suchmaschine.

[1] http://maps.google.ch/
[2] http://worldwind.arc.nasa.gov/

Die Disziplin des *Geographical Information Retrieval* (GIR) beschäftigt sich deshalb unter anderem mit der Verbesserung der Qualität von Suchanfragen mit geographischem Bezug (Jones & Purves, 2008).

Eine mögliche Lösung zur Beseitigung der semantischen Lücke ist der Einbezug von menschlichen Raumkonzepten in GIS, wie sie von der naiven Geographie gefordert wird (Egenhofer & Mark, 1995). Diese Raumkonzepte können durch die Analyse der qualitivativen Beschreibung in Form der natürlichen Sprache untersucht und mit geeigneten Methoden in GIS implementiert werden (Smith & Mark, 2001).

Eine mögliche Quelle zur Analyse der menschlichen Sprache und Erfassung von räumlichen Phänomenen sind UGC, wie beispielsweise die in dieser Arbeit verwendeten Flickr-Photos. Da diese Daten sowohl einen Raumbezug in der Form von geographischen Koordinaten als auch einen semantischen Bezug in der Form einer textuellen Beschreibung haben, sind sie zur Analyse menschlicher Raumvorstellungen im Sinne der naiven Geographie potentiell geeignet.

1.2 Ziele der Arbeit

Die im vorherigen Abschnitt 1.1 aufgezeigte Schwäche von Suchmaschinen bei der Suche und dem Empfangen von geographischen Informationen stellt die Hauptmotivation dieser Masterarbeit dar. Als übergeordnetes Ziel gilt die von Egenhofer & Mark (1995) formulierte Forderung nach dem Einbezug von menschlichen Raumkonzepten in GIS und somit die Überbrückung der semantischen Lücke (Smeulders et al., 2000) zwischen der menschlichen Wahrnehmung von Raum und dessen qualitativer Beschreibung in natürlicher Sprache einerseits und der technischen Implementierung in GIS andererseits.

Goodchild & Hill (2008) illustrieren ihre Vorstellung einer zukünftigen Suchmaschine folgendermassen:

> „*At some point in the future, it may be possible to submit a query such as 'find an orange grove five miles north of Bakersfield' to a portal and to obtain a probability density function that has been informed by all of the many sources of relevant information distributed over the Web.*“

In dieser Arbeit wird deshalb angestrebt, aus den textuellen Beschreibungen von georeferenzierten Photos der Online-Plattform *Flickr*[3] die relevanten Ortsbezeichnungen zu extrahieren und daraus ein umgangsprachliches Ortsverzeichnis zu erstellen (Rattenbury & Naaman, 2009). Da die Einträge in diesem Ortsverzeichnis neben verwendeten Toponymen auch umgangssprachliche Ortsbezeichnungen und weitere

[3] http://www.flickr.com/

Begriffe mit räumlicher Aussage umfassen, können sie neben offiziellen Ortsverzeichnissen zur Anreicherung von Suchanfragen mit räumlicher Semantik dienen.

Die Anreicherung mit räumlichem Kontext kann auch als Georeferenzierung der Suchanfrage bezeichnet werden. Die Georeferenzierung – die räumliche Verortung von Informationen durch Zuweisung von geographischen Koordinanten (Hill, 2006) – ist in dieser Arbeit sowohl als mögliche Anwendung von umgangssprachlichen Ortsverzeichnissen als auch zur Evaluation der möglichen Qualität einzuordnen. Die Georeferenzierung von Flickr-Photos ist somit nicht das eigentliche Ziel der Problemstellung dieser Arbeit. Allerdings sind Flickr-Photos aufgrund der öffentlichen Verfügbarkeit und der Popularität der Plattform gut zur Entwicklung und zum Testen der verwendeten Methoden geeignet.

Zusammengefasst sind die Ziele dieser Arbeit...

- ...die Erstellung eines umgangssprachlichen Ortsverzeichnisses aus räumlich relevanten Ortsbezeichnungen in georeferenzierten Flickr-Photos,
- ...die Evaluation der räumlich relevanten Ortsbezeichnungen zur Georeferenzierung von Flickr-Photos nur aufgrund deren textuellen Beschreibung,
- ...die Untersuchung von Ähnlichkeit der Beschreibungen zwischen Flickr-Photos mit geographischen Koordinaten und Photos ohne geographische Koordinaten
- ...und die Analyse des Einflusses der Filterung von nutzergenerierten Daten auf die Resultate der Georeferenzierung.

1.3 Aufbau der Arbeit

Der Aufbau dieser Masterarbeit ist folgendermassen gegliedert:

- Kapitel 2 beleuchtet den wissenschaftlichen Hintergrund dieser Arbeit und fasst die wichtigsten Ansätze der bisherigen Forschung zusammen. Aus den aufgezeigten Forschungslücken werden die Fragestellungen der Arbeit formuliert.
- In Kapitel 3 werden die Untersuchungsgebiete und die Datengrundlagen kurz beschrieben.
- Das Vorgehen und die angewendeten Methoden werden in Kapitel 4 behandelt.
- Kapitel 5 zeigt die detaillierten Ergebnisse der Untersuchungen, Kapitel 6 interpretiert und diskutiert diese ausführlich.
- Kapitel 7 rekapituliert die Erkenntnisse der Arbeit, beantwortet die Forschungfragen und gibt einen Ausblick auf mögliche Verbesserungen.

2 Hintergrund

Diese Arbeit fusst auf einer Vielzahl theoretischer Grundlagen aus verschiedenen Bereichen in- und ausserhalb der Geographie. Bevor auf die Ansätze der Georeferenzierung von UGC eingegangen wird, muss begründet werden, wieso es sinnvoll und nützlich sein kann, solche Daten in der GIScience zu nutzen, wo dabei die Schwierigkeiten liegen und welche theoretischen Grundlagen dafür nötig sind. Am Schluss werden aus den aufgezeigten Forschungslücken die Forschungsfragen abgeleitet.

2.1 Naive Geographie

Mit der Einführung des Begriffs und der Konzepte der *naiven Geographie* fordern Egenhofer & Mark (1995) einen neuen Denkansatz bei der Entwicklung von zukünftigen GIS, da bestehende Systeme für neue Benutzergruppen ohne Fachkenntnisse nicht geeignet sind, ihre alltäglichen Tätigkeiten auszuführen. Dies begründen sie mit dem potentiellen Unterschied zwischen der instinktiven räumlichen Denkweise der Menschen und der abstrakten Umsetzung dieser Raumkonzepte in GIS. Sie stellen das durch Erfahrung und Denken gewonnene Wissen der Menschen des geographischen Raumes ins Zentrum der naiven Geographie und empfehlen, dieses Wissen so zu formalisieren, dass es in Computern implementiert werden kann.

Diese Auffassung teilen Smith & Mark (2001), da GIS die wahrgenommene Welt formalisiert und abstrahiert abbilden. Sie sehen es mit der zunehmenden Verbreitung von GIS und ähnlichen Anwendungen im Internet als nötig und legitim an, in deren Weiterentwicklung neben abstrakten, auch Denkweisen wie sie die naive Geographie vorschlägt, einzubeziehen. Auch Kuhn (2001) fordert, dass GIS menschliche Handlungen unterstützen sollte. Seiner Meinung nach wurde bei der bisherigen Entwicklung von GIS zu wenig Rücksicht darauf genommen, welche Aufgaben auf der Anwenderseite damit ausgeführt werden müssen.

Obwohl es nicht trivial ist, die Denkweisen von Laien über den geographischen Raum in einer Ontologie – einer Sammlung der Definitonen von Grundstrukturen und Begriffen in einer Disziplin (Gruber, 1993) – abzubilden, wäre deren Verfügbarkeit hinsichtlich der Vergleichbarkeit von Resultaten mit ontologischer Forschung benachbarter Wissenschaftsbereiche von Vorteil. Die dafür nötige Beschreibung des Raumes bildet häufig Phänomene ab, die eine subjektiv empfundene Ausprägung und Ausdehnung haben und zudem auch noch von der Massstabsebene der Betrachtung abhängen. Diese Phänomene können qualitativ mit Wörtern der natürlichen Sprache

beschrieben werden. GIS behandeln dieselben Phänomene als mess- und quantifizierbare Entitäten. Damit diese für Nicht-Experten in einer verständlichen, qualitativen Form repräsentiert werden können, braucht es geeignete Umwandlungsmethoden, was wiederum das Vorhandensein der erwähnten Ontologie voraussetzt (Smith & Mark, 2001).

Ein Beispiel dafür zeigen Janowicz et al. (2010) mit der Einbindung von Semantik in eine Geodateninfrastruktur als Webservice basierend auf bereits bestehenden OGC-Standards[1]. Da die Beschreibung der Semantik keine inhärente Eigenschaft von räumlichen Daten ist, können diese mittels eines Webservices mit einer extern vorhandenen Ontolgie verbunden und mit semantischen Informationen angereichert werden.

2.2 Georeferenzierung

Mit Georeferenzierung ist im Allgemeinen das Verknüpfen von Informationen mit einem geographischen Standort gemeint. Eine Georeferenzierung kann entweder explizit über die Angabe von geographischen Koordinaten oder implizit durch die Verwendung von Ortsnamen erfolgen. Im Alltag begegnet man meistens Ortsnamen. So enthalten, je nach Stichprobe, geschätzte 50 bis 70 % aller Textdokumente Ortsnamen. Dies ist relevant für die Suche nach Dokumenten mit räumlichem Bezug, welcher oft über einen Standortnamen als herkömmliches Schlüsselwort in der Suchanfrage hergestellt wird (Hill, 2006).

2.2.1 Suche nach geographischen Informationen

Mit der fortschreitenden Entwicklung des Internets in den letzten Jahren, hat sich auch der Zugang zu geographischen Informationen und die Vielfalt an verfügbaren Daten grundlegend geändert. Dabei zeigte sich, dass gängige Suchmethoden, die meist auf der Suche nach Schlüsselwörtern basieren, zum Auffinden und Benutzen von Informationen im Internet oft nicht den Anforderungen der Benutzer entsprachen. Die Suchmaschinen sind unter anderem nicht in der Lage, die räumliche Bedeutung der von den Nutzern verwendeten Begriffe zu interpretieren, was von Smeulders et al. (2000) als „semantische Lücke“ bezeichnet wird. Da laut Sanderson & Kohler (2004) rund 13 % aller Anfragen bei Suchmachschinen eine Ortsbezeichnung beinhalten, braucht es verbesserte Suchmethoden, welche die räumlichen Vorstellungen der Menschen (siehe Abschnitt 2.1) in Suchanfragen miteinbeziehen (Larson, 1996; Egenhofer, 2002; Jones et al., 2008b). Mit dem explizitem Einbezug der Semantik

[1] Open Geospatial Consortium: http://www.opengeospatial.org/

räumlicher Suchbegriffe soll deshalb die Suche nach geographischen Informationen verbessert werden.

O'Hare & Murdock (2012) formulieren ein Beispiel einer Suchanfrage zur Veranschaulichung der Problematik von menschlichen Raumkonzepten: *„dry cleaner on the way to work"*. Eine Suchmaschine sollte bei einer solchen Anfrage im Optimalfall den Arbeitsweg als Pfad verstehen können und für den Unweg zu einer Trockenreinigung mögliche Lösungen anbieten. Die Relevanz der Lösung könnte im obigen Beispiel über einen möglichsten kurzen Umweg vom üblichen Arbeitsweg bewertet werden.

Dazu braucht es Ortsverzeichnisse, die neben offiziellen Ortsnamen auch umgangssprachliche oder alternative Bezeichnungen für Orte enthalten. Die Erstellung und Verschmelzung mehrerer Ortsverzeichnisse aus unterschiedlichen Quellen zu einem Meta-Ortsverzeichnis könnte bei der Beantwortung der obigen Fragestellung helfen und ist zentral für die Erfüllung der Aufgaben von GIR (Kessler et al., 2009; Smart et al., 2010).

Es wurde viel in die Entwicklung von Systemen zur Suche und zum Empfang räumlicher Informationen aus dem Internet investiert, die dort in grosser Menge und relativ unstrukturiert vorhanden sind. Ein Beispiel dafür ist das Projekt *SPIRIT*[2] von Purves et al. (2007), einer Suchmaschine für unstrukturierte Daten im Internet, welche explizit die räumliche Beziehung und den Kontext von Suchanfragen berücksichtigt und damit zu relevanteren Ergebnissen führt.

Diese Entwicklung wird unter dem Begriff *Geographical Information Retrieval (GIR)* zusammengefasst. Grundsätzliches Ziel von GIR ist der verbesserte Zugang zu geographischen Informationsquellen (Larson, 1996). Die Herausforderungen von GIR bestehen in der Identifizierung von Ortsnamen in Textdokumenten und Suchanfragen, der Auflösung von mehrdeutigen Ortsnamen, der geometrischen Interpretation von umgangssprachlichen Ortsbezeichnungen und räumlicher Sprache, der räumlichem Indexierung von Dokumenten, dem Ordnen von Dokumenten nach räumlicher Relevanz sowie in der Entwicklung von Benutzerschnittstellen und Evaluationsmethoden (Jones & Purves, 2008).

Wichtig ist in disem Zusammenhang die klare Unterscheidung der beiden Konzepte *place* und *space*. Der Begriff *place* verknüpft Geographie mit dem von den Menschen wahrgenommenen Raum und kann aufgrund der Subjektivität nicht nur in Form von Koordinaten ausgedrückt werden (siehe Abschnitt 2.3). Dagegen wird mit *space* eher die abstrakte und geometrische Ausprägung des Raums erfasst, wie sie in Computersystemen gespeichert wird (Fisher & Unwin, 2005; Davies et al., 2009).

[2]Spatially Aware Information Retrieval on the Internet

2.2.2 Ortsverzeichnisse und Identifizierung von Toponymen

Mittels sprachlicher Analyse können Toponyme oder andere Wörter mit Raumbezug in Texten oder Suchanfragen identifiziert und lokalisiert werden (Jones & Purves, 2008). Dieser Prozess wird auch als *Geoparsing* bezeichnet (Guillén, 2008).

Dafür braucht es digitale Ortsverzeichnisse *(Gazetteers)*, die einem Ortsnamen geographischen Koordinaten und weiteren beschreibenden Informationen, wie etwa eine Kategorie zuordnen. Ortsverzeichnisse sind das wichtigste Bindeglied bei der Übersetzung von Ortsnamen in geographische Koordinaten und umgekehrt. Dazu liefern sie die nötige Grundgesamtheit an möglichen Toponymen zur Identifikation beim *Geoparsing* (Hill, 2006; Smith & Crane, 2001). Je nach Anwendungszweck kann die geometrische Repräsentation des Standorts im Ortsverzeichnis als Punkt, Linie oder Polygon vorhanden sein. Für GIR kann es zum Beispiel sinnvoll sein, statt Punktkoordinaten ein minimal umgebendes Rechteck (MBR) zu verwenden, welches dann auch Suchanfragen über räumlichen Beziehungen ermöglicht (Goodchild & Hill, 2008).

Leidner (2004) zeigt wichtige Punkte bei der Wahl eines geeigneten Ortsverzeichnisses auf. Der Abdeckungsbereich eines Ortsverzeichnisses kann von kleinräumigen, lokalen Katasteranwendungen bis hin zu globalem Massstab reichen. Dies zeigt sich auch in der variierenden Anzahl an enthaltenen räumlichen Referenzen, welche jedoch nie vollständig sind. Die Einträge eines Gazetteers können sich während der Zeit verändern, somit ist auch die Aktualiät ein wichtiges Kriterium. Zudem muss die Granularität, das heisst die räumliche Präzision der Toponyme berücksichtigt werden. Weitere Unterschiede gibt es hinsichtlich der verfügbaren Zusatzinformationen, welche die Toponyme ergänzen. So enthält der in dieser Arbeit verwendete, jährlich aktualisierte *1:50 000 Scale Gazetter* rund 250'000 Toponyme und ist laut Ordnance Survey (2012) für Grossbritannien das detaillierteste erhältliche Produkt. Die geographischen Koordinaten der Toponyme beziehen sich allerdings auf den Mittelpunkt einer Zelle eines 1 km-Gitternetzes, womit eine Nutzung für grössere Massstäbe eingeschränkt ist. Weitere Informationen über die verwendeten Ortsverzeichnisse in dieser Arbeit finden sich im Abschnitt 3.3.

2.2.3 Mehrdeutige Ortsnamen und Disambiguierung

Bei der Identifizierung von Toponymen kommt es oft zu Mehrdeutigkeiten. Viele Ortsnamen sind nicht eindeutig, sondern bezeichnen mehrere Standorte auf der Erdoberfläche. Durch den Prozess der Disambiguierung können mehrdeutigen Toponyme aufgelöst und mit dem zutreffendsten Standort assoziiert werden. Smith & Crane (2001) haben beobachtet, dass 92 % aller identifizierten Toponyme in einer digitalen historischen Bibliothek möglichweise mehrdeutige Standorte haben. Oftmals werden

Toponyme auch in einem nicht-geographischen Kontext verwendet, zum Beispiel als Bestandteil von Personen- oder Organisationsbezeichnungen. Ebenso mehrdeutig sind Toponyme, die sich auf ein anderes Konzept beziehen. Dies kann beispielsweise im Titel einer Sportnachricht „*Schweiz gewinnt sensationell gegen Spanien*“[3] vorkommen, wo sich die Ländernamen auf die jeweiligen Fussball-Nationalmannschaften beziehen. Dies ist auch der Fall, wenn Toponyme in beschreibender Weise verwendet werden, etwa in *der Kaiser von China* (Hu & Ge, 2005; Leveling & Hartrumpf, 2008).

Eine weitere Herausforderung sind umgangssprachliche Ortsnamen, die neben offiziellen Namen in Ortsverzeichnissen häufig zur Benennung von Orten verwendet werden (siehe Abschnitt 2.3).

Hu & Ge (2005) unterscheiden drei verschiedene Vorgehensweisen der Disambiguierung. Korpus-basierte Methoden wenden statistische Analyse und maschinelles Lernen (ML) auf einen Korpus mit Trainingsdaten an, der bereits disambiguierte Toponyme enthält. Daraus lassen sich Klassifikatoren zur Disambiguierung von Toponymen in anderen Dokumenten erstellen. Allerdings ist diese Methode auf zuverlässige Trainingsdaten angewiesen, deren Erstellung gemäss Leidner (2004) sehr aufwendig ist. Regel-basierte Analyse ermöglicht die Disambiguierung durch die Erkennung von linguistischen Mustern wie dem gemeinsamen Auftreten von Toponymen. Diese Form lässt sich mit Hilfe von vorgegebenen Regeln einfach implementieren. Ansätze, welche geographische Heuristiken verwenden, versuchen das Problem der Mehrdeutigkeit mit Eigenschaften wie etwa der Distanz zu einem geographischen Schwerpunkt zu lösen. Diese Ansätze sind auf externe geographische Daten wie zum Beispiel jene in einem Ortsverzeichnis angewiesen.

Den letzten Ansatz verwenden Smith & Crane (2001), wobei sie in einem ersten Schritt mögliche Verwendungen von Toponymen als Namen herausfiltern. In einem zweiten Schritt werden die Standorte aller mehrdeutigen Toponyme aufgrund ihrer räumlichen Verteilung und ihrer geographischen Relevanz bewertet und zugewiesen. Beaufsichtigtes ML wird von Hu & Ge (2005) zur Auflösung von Mehrdeutigkeiten angewendet. Zuerst werden Toponyme mittels *named entity recognition (NER)* aus Zeitungsartikeln extrahiert. NER ist eine Methode zur Identifikation und Klassifikation von Textelementen in vordefinierte Kategorien (übliche sind <Person>, <Organisation>, <Ort>, <Uhrzeit>, <Datum>, <Geldwert> und <Prozentwert>). Danach werden die mehrdeutigen Toponyme mit einem statistischen Sprachmodell aufgelöst.

Buscaldi & Rosso (2008) benutzen die englische Sprachdatenbank *WordNet*[4] zur Berechnung der konzeptuellen Dichte, welche einen Messwert für die Korrelation

[3] http://www.tagesanzeiger.ch/ Zugriff: 10.12.2012
[4] http://wordnet.princeton.edu/

zwischen einem Wort und seinem verwendeten Kontext berechnet. Einen ähnlichen Ansatz zur Disambiguierung verwenden Overell & Rüger (2008), allerdings auf der Grundlage des gemeinsamen Auftretens von Wörtern beziehungsweise Ortsnamen in einer Wikipedia-Stichprobe.

2.3 Vage Orte und umgangssprachliche Ortsnamen

Im vorherigen Abschnitt 2.2.2 wurde erwähnt, dass Ortsverzeichnisse den Standort der Ortsnamen in unterschiedlicher geometrischer Form repräsentieren können. Dies hat grosse Bedeutung für GIR-Anwendungen, da der Informationsgehalt eines Punktes weniger aussagekräftig ist, als ein MBR oder detaillierte Grenzverläufe. Je realistischer und damit weniger abstrakt eine Repräsentation ist, desto geringer ist auch die Unsicherheit der räumlichen Ausdehnung (Hill, 2006). Der Detailgrad ist abhängig von der Anwendung, wobei für die meisten Aufgaben ein MBR ein hinreichender Kompromiss zwischen den Anforderungen für die Analyse von räumlichen Beziehungen und der Performanz ist (Goodchild & Hill, 2008).

Die oben genannten Fälle beziehen sich implizit auf die Repräsentation von Geometrien als randscharfe Phänomene, wie sie in räumlichen Datenbanken oder GIS implementiert sind. Somit können die räumlichen Beziehungen zwischen den Objekten präzis definiert werden. Schwieriger wird es bei Regionen wie dem *Lake District* in Grossbritannien. Es gibt keine eindeutig definierte Grenze, weshalb das Gebiet nur unscharf abgegrenzt werden kann (Hart & Dolbear, 2007). Zudem ist die Verwendung von umgangsprachlichen Bezeichnungen für vage Orte weit verbreitet und erschwert die Identifizierung über einen eindeutigen Ortsnamen (Davies et al., 2009).

Die nicht klar definierte räumliche Ausdehnung und Benennung von vagen Orten beruht auf dem dem Konzept von *place* (vgl. Abschnitt 2.2.1), welches von Davies et al. (2009) als *„uncomfortable challenge“* für die GIScience bezeichnet wird. In der Geographie werden verschiedene Ansätze von *place* verwendet. Diese basieren unter anderem auf der Vorstellung von Orten als Erfahrungsraum, entstanden durch soziale Interaktion oder definiert durch seine Charakteristik, was eine Implementation in Computersystemen zu einer komplexen Aufgabe macht. Allerdings teilen alle Ansätze die Abwesenheit von randscharfen Grenzen, eines präzis definierten Standorts und eines eindeutigen, identifizierenden Ortsnames.

2.3.1 Anforderungen an Modelle von vagen Orten

Vertreter der naiven Geographie fordern auch die Implementation der Raumkonzepte, die in den Köpfen der Menschen existieren, in GIS (Abschnitt 2.1). Im Gegensatz zu

traditionellen Karten aus Papier, ermöglichen digitale Daten die Modellierung vager Orte mit unscharfen Grenzen und umgangssprachlicher Ortsnamen aus verschiedenen Perspektiven.

Die Implementierung von vagen Orten in einem Computersystem kann je nach Anwendung in unterschiedlicher Form und Detailgrad erfolgen. Davies et al. (2009) definiert mögliche Anforderungen an Modelle:

- einfache Auswahl und Darstellung von Ortsnamen und der dazugehörigen unscharfen Grenzen auf einer Karte
- Suchen nach alternativen umgangssprachlichen Bezeichnungen für einen gegebenen Ortsnamen
- Darstellung der unscharfen Grenzen für die Suche nach einem Ortsnamen

Daraus lässt sich schliessen, dass umgangssprachliche Ortsnamen verknüpft mit ihrer räumlichen Ausdehnung analysiert und erhoben werden sollten. Schliesslich sollten sich mit allen Modellen die beiden primären Aufgaben der Suche nach einem Standort für einen gegebenen Ortsnamen und des Auffindens von Ortsnamen für einen gegebenen Standort durchführen lassen (Davies et al., 2009).

2.3.2 Beschreibung von vagen Orten mit Internetdaten

Laut Montello et al. (2003) sind zwar die technischen und theoretischen Grundlagen zur Implementierung von vagen Orten in Computersysteme vorhanden, allerdings sind diese Konzepte bisher nicht umgesetzt worden (Abschnitt 2.1), weil deren Inhalt auf der Beobachtung von menschlichen Wahrnehmungen basiert und demnach empirische Erhebungen gemacht werden müssten.

Die Analyse von vagen Orten und umgangssprachlichen Ortsbezeichnungen ist aber nicht nur über direkte empirische Erhebung und Beobachtung von Menschen möglich. Folgende Beispiele zeigen, wie nicht randscharf definierbare räumliche Geometrien mittels Internetdaten modelliert werden können.

Purves et al. (2005) modellieren nicht randscharf definierte Gebiete, wie zum Beispiel das Schweizer *Mittelland* oder die schottischen *Highlands*. Dazu extrahieren sie Textdokumente aus dem Internet und georeferenzieren die darin enthaltenen Toponyme. Daraus lässt sich mittels Kerneldichteschätzung (KDE) eine Wahrscheinlichkeitsoberfläche schätzen, welche den Grad an Zugehörigkeit zur gesuchten Region durch einen Dichtewert darstellt. Einen ähnlichen Ansatz verfolgen Jones et al. (2008a) und approximieren aus der Wahrscheinlichkeitsoberfläche eine Grenzlinie der gesuchten Region. Sie evaluieren ihre Methode mit randscharfen administrativen Grenzen und

kommen zum Schluss, dass eine gute Übereinstimmung zwischen geschätzten und gegebenen Grenzen erreicht wurde.

Eine andere Internetquelle zur Modellierung von unpräzisen Regionen und umgangssprachlichen Namen sind Sammlungen von georeferenzierten Photos. Hollenstein & Purves (2010) verwenden georeferenzierte Flickr-Photos (siehe Abschnitt 3.2) zur Bestimmung der Ausdehnung von Stadtzentren mittels durch KDE erstellter Wahrscheinlichkeitsoberflächen. Eine andere Quelle von georeferenzierten Photos ist *Geograph*[5], eine Onlineplattform mit dem Ziel, jeden Quadratkilometer von Grossbritannien und Irland mit repräsentativen Photos und den dazugehörigen Metadaten zu beschreiben. Dykes et al. (2008) benutzen diese Daten zur Erstellung von Baumkarten *(treemaps)*, welche die Hierarchie der verwendeten Begriffe in den Titeln der Photos visualisieren.

Die automatische Erstellung der geographischen Ausdehnung für Suchbegriffe zeigen Grothe & Schaab (2009). Sie verwenden dazu die Daten von georeferenzierten Flickr-Photos und die Methoden von KDE beziehungsweise *Support Vector Machines* (SVM), Methoden aus dem Feld des statistischen Lernens. Die Umsetzung mit SVM bringt dabei bessere Resultate als die KDE-Methode.

Im nächsten Abschnitt wird detailliert beschrieben, wie Daten aus dem Internet für die Analyse von vagen Orten und umgangssprachlichen Ortsbezeichnungen genutzt werden können und welche Probleme dabei beachtet werden müssen. Das Hauptaugenmerk liegt auf der vorher erwähnten Nutzung von georeferenzierten Bildern.

2.4 Extraktion von Informationen aus nutzergenerierten Daten

2.4.1 Potential von nutzergenerierten Daten aus dem Internet

Im Gegensatz zu früher sind die Möglichkeiten von Internetnutzern heute nicht mehr bloss auf das Suchen und Empfangen von Informationen in einem Browser beschränkt. Mit den aufkommenden *Web 2.0*-Anwendungen, *open-source*[6] GIS-Software, erschwinglichen GPS-Geräten zur Ermittlung des aktuellen Standorts auf der Erde durch Satelliten, digitalen Photokameras und fortschrittlichen Mobiltelefonen hat sich die Rolle der Internetnutzer von der Konsumation hin zur vielfältigen Beteiligung und Produktion von Internetinhalten erweitert (Heipke, 2010).

Viele dieser so produzierten UGC haben einen geographischen Bezug, weshalb in diesem Zusammenhang auch der Begriff *Volunteered Geographic Information*

[5] http://www.geograph.org.uk/
[6] http://opensource.org/docs/definition.php

(VGI) gebräuchlich ist (Goodchild, 2007). So tragen zum Beispiel fast eine Million freiwilliger Mitglieder zum Projekt *OpenStreetMap*[7] bei, welches sich die Erstellung und die Verfügbarkeit von detaillierten geographischen Daten als Ziel gesetzt hat und dabei äussert erfolgreich ist (OpenStreetMap Foundation, 2012). Eine andere bereits angesprochene Quelle von VGI sind Plattformen wie Flickr (siehe Abschnitt 3.2) oder *Geograph*, welche georeferenzierte Photos öffentlich zugänglich anbieten.

Diese Entwicklung wird als *NeoGeography* (Goodchild, 2009) bezeichnet und beschreibt die geänderten Verhältnisse zwischen Produzenten und Konsumenten von geographischen Informationen und kartographischen Repräsentationen. Während früher vor allem staatliche Institutionen und spezialisierte private Anbieter geographische Informationen bereitgestellt haben, so sind es heute auch gemeinnützige Bewegungen oder Laien. Die gemeinnützig gesammelten Daten sind oft für alle Internetnutzer zu geringsten Kosten zugänglich und manchmal sogar auch die einzig verfügbare Quelle. Begünstig wurde die Entwicklung durch die bereits erwähnten Technologien, welche die finanziellen Eintrittskosten in die Welt GIS und Kartographie gegen Null tendieren lassen.

Gemäss Goodchild (2007) können UGC als Quelle für die geographische Beschreibung der Erdoberfläche dienen und dabei helfen, die menschliche Wahrnehmung von Orten im Sinne der naiven Geographie zu beschreiben (Abschnitt 2.1 und 2.3.2). Edwardes & Purves (2007) erstellen aus den Beschreibungen von Photos aus der *Geograph*-Sammlung eine konzeptuelle Ontologie der Landschaft und vergleichen diese mit bestehenden Konzepten. Ihre Resultate bestätigen, dass mit UGC das Potential für die Erforschung von menschlicher Wahrnehmung beschrieben durch natürliche Sprache vorhanden ist. Zudem sind mit UGC sehr viel grössere Datenmengen vorhanden, als dies bei bisherigen empirischen Befragungen mit begrenzter Stichprobengrösse der Fall ist. Dies zeigt sich auch bei der Untersuchung der Benennung von geomorphologischen Landschaftselementen von Gschwend & Purves (2012). Sie vergleichen die aus *Geograph*- und Flickr-Photos extrahierten Schlüsselwörter mit geomorphometrischen Klassen. Dabei zeigen sich zum einen die räumliche Variation der verwendeten Sprache und zum anderen auch Unterschiede zwischen den beiden verschiedenen UGC-Quellen.

Auch Purves et al. (2011) heben dieses Potential von nutzergenerierten Daten zur Beschreibung von *place* hervor. Die Identifizierung von räumlichen Bezeichnungen in Textdaten und deren Verknüpfung mit semantischen Grundkonzepten kann durch die Analyse von UGC erreicht werden. Die Autoren verwenden ebenfalls Daten aus den beiden unterschiedlichen Photosammlungen Flickr und Geograph. Damit adressieren sie das Problem der semantischen Lücke zwischen den Bedürfnissen der Benutzer von Suchmaschinen und deren Möglichkeiten (Abschnitt 2.2.1).

[7] http://www.openstreetmap.org/

Dass in Flickr-Metadaten Potential für die Analyse von umgangsprachlichen Bezeichnungen und vagen Orten vorhanden ist, zeigt auch der hohe Anteil von Toponymen in der Beschreibung von bereits georeferenzierten Bildern. Laut Sigurbjörnsson & van Zwol (2008) sind 28% aller verwendeten Schlüsselwörter Ortsnamen, bei Hollenstein & Purves (2010) sind es sogar 35 %. Zudem sind 70 % aller Photos in der Stichprobe mit mindestens einem Ortsnamen als Schlüsselwort gekennzeichnet.

2.4.2 Motivation der Beitragenden von UGC und Tagging-Verhalten

Für die Frage, nach der Motivation von Nutzern an UGC-Projekten teilzunehmen, gibt es mehrere mögliche Antworten. Selbstpromotion oder soziale Präsenz sind mögliche Gründe dafür. Auf UGC-Seiten wie Flickr kann dies durch das Teilen von Photos mit anderen Einzelnutzern oder in Interessengruppen erfolgen. Zudem nutzen viele Anwender die Angebote zum Teilen von Informationen mit Freunden und Bekannten ohne die Berücksichtigung der Tatsache, dass der Zugang auch für die Öffentlichkeit möglich ist. Auch eine Rolle spielen kann die persönliche Befriedigung der Beitragenden von Projekten wie OpenStreetMap oder Geograph, die durch den Beitrag zum Fortschritt des Projekts empfunden wird (Goodchild, 2007; Nov et al., 2008).

Im Fall der Online-Photoplattform Flickr, von welcher die in dieser Arbeit verwendeten Daten stammen, ist auch die Motivation und das Verhalten der Benutzer beim *Tagging* interessant. Unter Tagging wird das Hinzufügen von Schlüsselwörtern beziehungsweise *Tags* zu den Metainformationen eines Photos verstanden. Mit Tags lassen sich einfach Beschreibungen erstellen, die verschiedene Facetten des beschriebenen Objekts wiedergeben (Schmitz, 2006). Die Tags werden meistens von den Benutzern selber hinzugefügt und können aus einem oder mehreren Wörtern zusammengesetzt und frei zugewiesen werden. Das Tagging bringt einen Nutzen bei der Organisation und der Kommunikation von grossen Datenbeständen. So können die Nutzer ihre Photos mit bestimmten Tags organisieren, was die spätere Auffindbarkeit erleichert. Den Betrachtern der Photos werden damit zusätzliche Informationen zum Hintergrund des Photos bereitgestellt. Zudem ermöglicht es den Betrachtern das Auffinden von Photos mit einem bestimmten Tags (Ames & Naaman, 2007; Nov et al., 2008).

Das Teilen von Inhalten und Metainformationen in Form von Tags kommt allerdings nicht bei allen Benutzern in der gleichen Intensität vor. Neue Mitglieder von Flickr tendieren dazu, mehr Photos zu teilen als ältere Mitglieder. Hingegen sind es die älteren Mitglieder, welche mehr Metainformationen mit der Gemeinschaft teilen. Grund dafür könnte sein, dass die bei längerer Mitgliedschaft anfallenden, grösseren Datenbestände mehr Organisation benötigen und diese Nutzer durch die Vernetzung mit anderen Mitgliedern vertrauter im Umgang mit Tagging sind. Allerdings scheint

die soziale Motivation den grössten Einfluss auf das Verhalten beim Teilen von Metainformationen zu haben (Nov et al., 2008, 2010).

Hollenstein & Purves (2010) zeigen auf, dass Ortsnamen die am häufigsten anzutreffenden Schlüsselwörter für georeferenzierte Photos auf Flickr sind (siehe Abschnitt 2.4.1). Somit beschreiben die Nutzer eine Lokalität am häufigsten mit dem Namen des Ortes, zusätzlich zu den bereits vorhanden Koordinaten. Die räumliche Granularität der Ortsnamen kann dabei meistens der Stadtebene zugeordnet werden. Dies bestätigt die Untersuchung von Jones et al. (2008b), welche über 80% aller in Suchanfragen verwendeten Ortsnamen ebenfalls dieser Ebene zuordnen kann.

Da die Tags im Allgemeinen vom Nutzer frei wählbar und zugeordnet werden können, beinhalten sie keine vordefinierte Bedeutung, wie das in Kategorie- oder Ontologiebasierten Anwendungen der Fall ist. In Tagging-Systemen liegen die darin enthaltenen Informationen in unstrukturierter Form vor, worin gemäss Rattenbury et al. (2007) ein Nutzen zur Untersuchung der inhärenten Eigenschaften und zur Extraktion von strukturiertem Wissen liegt. Die Analyse der Tags von georeferenzierten UGC stellt somit eine mögliche Quelle zur Beschreibung der menschlichen Wahrnehmung von Orten dar.

2.4.3 Herausforderungen bei der Verwendung von UGC

Bei der Verwendung von nutzergenerierten Daten gibt es einige Herausforderungen hinsichtlich Repräsentativität, Qualität und Verzerrung der Datengrundlage zu beachten.

Durch eine grosse Stichprobe von nutzergenerierten Daten kann nicht automatisch von einem repräsentativen Querschnitt der Bevölkerung ausgegangen werden. Nicht alle Menschen haben einen gleichwertigen Zugang zum Internet. Ein grosser Teil der nutzergenerierte Daten stammt von Menschen aus Industrieländern, vor allem aus Nordamerika und Europa. Für viele Menschen bestehen zudem Hürden im Hinblick auf die Sprache und das Wissen, welche den Zugang zu UGC-Anwendungen erschweren oder verunmöglichen. Oft sind es technikaffine Leute im Alter zwischen 20 und 30 Jahren, mehr männliche als weibliche, welche zu UGC-Projekten beitragen (Goodchild, 2007; Cox, 2008). Beim Online-Nachschlagewerk *Wikipedia*[8] sind weniger als 20 % der Beiträge von Frauen verfasst, was auf einen markanten Geschlechterunterschied hinweist (Antin et al., 2011). Zudem muss beachtet werden, dass nicht alle Benutzer den gleichen Anteil zu UGC-Sammlungen beitragen. Als Faustregel definiert Nielsen (2006), dass 1 % aller Nutzer für den grössten Teil der Beiträge verantwortlich sind, während sich 9 % nur sporadisch beteiligen. 90 % aller Benutzer tragen nichts zur Produktion der Daten bei und konsumieren dieses lediglich.

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

Die Qualität von nutzergenerierten Daten betrifft verschiedene Bereiche. Wie bei allen Daten mit Raumbezug gibt es auch bei UGC, zu denen verschiedene Nutzer in unterschiedlicher Form Daten beitragen, Erwartungen an die Positionsgenauigkeit. Für eine kleine Stichprobe von Flickr-Photos aus dem urbanen Raum haben Hochmair & Zielstra (2012) einen Medianpositionsfehler von 58.5 m evaluiert. Eine häufige Fehlerursache und Schwäche von georeferenzierten Bildern ist die unklare Angabe des verwendeten Standorts. Bei der Verwendung eines GPS-Sensors wird in den meisten Fällen der Standort der Kamera während der Aufnahme festgehalten. Bei nachträglich georeferenzierten Bildern kann auch der Standort des festgehaltenen Bildmotivs gewählt werden. Für diese Unsicherheit gibt es keine Lösung, weshalb sie bestehen bleiben.

OpenStreetMap-Daten werden im Gegensatz zu staatlichen oder professionellen privaten Anbietern zum grössten Teil von Benutzern ohne Expertenwissen erhoben. Während staatliche Anbieter durch langjährig gereifte Produktionsprozesse und Qualitätskontrollen ein grosses Vertrauen geniessen, muss sich OpenStreetMap dieses Vetrauen erst durch den Beweis von Qualität und *fitness for use* erarbeiten. Heipke (2010) sieht es als Vorteil von nutzergenerierten Daten, dass durch die vielen Nutzer lokales Wissen besser abgebildet werden kann. Zudem ergibt sich aus statistischer Sicht durch die vielen mehrfach kartierten Objekte eine geringere Varianz und Anzahl Fehler. Die grösste Herausforderung bei UGC ist aber die räumlich heterogene Verteilung der Qualität und der Abdeckung, welche bei staatlichen Anbietern in der Regel relativ homogen ist. Vorallem zwischen urbanen und ländlichen Gebieten gibt es grosse Unterschiede in der Gesamtheit der erfassten Objekte. In urbanen Gebieten kann OpenStreetMap durchaus mit professionellen Datensätzen mithalten, in ländlichen Regionen ist hingegen die Abdeckung häufig kleiner (Koukoletsos et al., 2012).

Die räumliche Verteilung von georeferenzierten Photos auf Flickr sieht ähnlich aus. In Grossbritannien korreliert sie stark mit den urbanen Gebieten, während ländliche, eher dünnbesiedelte Gebiete schwach abgebildet sind. Diese Unterschiede entstehen durch die lokale Bevölkerungsdichte sowie durch die touristische Attraktivität von bestimmten Orten (Grothe & Schaab, 2009; Purves et al., 2011).

Die Beschreibung von UGC mit unstrukturierten Schlüsselwörtern kann im Hinblick auf die Konsistenz nicht immer gewährleistet sein. Es kann sein, dass die Beschreibung der Daten unpräzis, irrelevant oder im schlimmsten Fall gar nicht vorhanden ist. Dieser Nachteil entsteht aufgrund der fehlenden Einschränkung und ist eine Eigenheit von offenen Tagging-Systemen. Diese Schwäche kann aber auch als Stärke gesehen werden, wenn man eine grosse Bandbreite der von den Nutzern verwendeten Sprache einfangen will (Rorissa, 2010; Ahern et al., 2007; Rattenbury et al., 2007). Zudem ermöglichen diese unstrukturierten Daten die Untersuchung

subjektiver Wahrnehmung und Verwendung der Sprache auf lokaler Ebene (Purves et al., 2011). Auf Flickr wurde allerdings auch häufig ein Nutzerverhalten zur Steigerung der Aufmerksamkeit von anderen Nutzern beobachtet, das sich in der Verwendung von vielen, teils auch unpassenden Tags zeigt (Cox, 2008). Moderierte Plattformen wie Geograph, wo solches Verhalten kontrolliert werden kann, haben in Bezug auf dieses Problem einen Vorteil. Dieser Unterschied rührt aus der Tatsache her, dass Flickr primär als Photo-Plattform dient, während Geograph explizit die Beschreibung der Landschaft durch Bilder als Ziel nennt.

Ein wichtiger Aspekt von UGC wurde lange nicht genügend beachtet. Durch die grosse Menge an Daten ging man davon aus, dass einzelne Nutzer in einer Stichprobe nicht in der Lage sind, signifikante Verzerrungen zu induzieren. Neuere Untersuchungen mit Flickr-Daten zeigen, dass dies aber möglich ist, wenn die Beiträge einzelner Nutzer überproportional höher sind als jene der restlichen Nutzer. Dies kann problematisch sein, wenn man annimmt, verschiedene subjektive Perspektiven zu analysieren, in Wirklichkeit aber nur eine einzelne Perspektive betrachtet. Dies führt im schlimmsten Fall zu verzerrten und wenig aussagekräftigen Resultaten (Hollenstein & Purves, 2010; Purves et al., 2011). Purves et al. (2011) bestätigen mit der Untersuchung zweier UGC-Stichproben (Flickr und Geograph) die ungleiche Partizipation in Online-Gemeinschaften gemäss Nielsen (2006). In der Geograph-Stichprobe stammt mit 90 % der grösste Anteil der Daten von bloss 10 % der Nutzer, in der Flickr-Stichprobe 73 %. Es zeigt sich, dass der Beitrag dieser Nutzer für grosse Verzerrungen sorgt und deshalb in einer Analyse berücksichtigt werden muss, beispielsweise durch Filterung.

Geographische Informationen, die aus den räumlichen Denkweisen der Menschen entstanden sind, weisen also Lücken und Fehler auf, sind von unterschiedlichem Detailgrad, haben oft keine scharfen Grenzen und sind möglicherweise durch Einzelnutzer verzerrt. Methoden, die UGC analysieren oder verwenden, müssen deshalb genug robust sein, um mit lücken- und fehlerhaften Information in Datensätzen umgehen zu können (Egenhofer & Mark, 1995). Zudem müssen sie mögliche Verzerrungen durch überdurchschnittlich aktives Nutzerverhalten berücksichtigen.

2.4.4 Methoden zur Georeferenzierung von UGC

Die Verwendung von UGC wurde in den vorherigen Abschnitten ausführlich beschrieben. Der Fokus dieser Arbeit, liegt auf der Georeferenzierung von UGC in Form von Flickr-Photos. Daher werden in diesem Abschnitt einige Ansätze und Methoden der Georeferenzierung genauer besprochen.

Aus den unstrukturierten Beschreibungen von UGC, die meistens in Textform vorliegen, wird strukturierte Information extrahiert. Damit lässt sich ein umgangs-

sprachliches Ortsverzeichnis oder ein statistisches Sprachmodell erstellen. Dieses wird im Kontext dieser Arbeit zur Bestimmung des Standorts von weiteren UGC benutzt. Laut Wing & Baldridge (2011) geht diese Vorgehensweise von der Hypothese aus, dass auch Wörter, die keine Toponyme sind und explizit einen Ort bezeichnen, nützlich zur Bestimmung des Standortes sein können.

Rattenbury & Naaman (2009) zeigen am Beispiel von Flickr-Photos verschiedene Methoden zur Extraktion von räumlicher Information aus georeferenziertem UGC. Eine Methode, die auch in dieser Arbeit verwendet wird ist die *TF-IDF*-Methode (*term frequency, inverse document frequency*, siehe Abschnitt 4.6.1) (Ahern et al., 2007). Diese identifiziert Tags, die als Bezeichnung für eine begrenzte geographische Region verwendet werden, wie zum Beispiel *Hyde Park* oder *Golden Gate Bridge*. Sie basiert auf der räumlichen Verteilung der Tags und der Annahme, dass Tags, die nur in einem begrenzten geographischen Raum und selten ausserhalb davon vorkommen, repräsentativer sind als solche, die räumlich verteilt und weniger geballt sind. Die Methode ist eine leicht modifizierte Variante der ursprünglichen TF-IDF-Methode (z.B. in Salton & Buckley, 1988), die zur Beurteilung der Relevanz von Dokumenten in einer Sammlung für bestimmte Suchbegriffe entwickelt wurde. Damit lassen sich Tags mit räumlicher Relevanz isolieren und extrahieren. Eine andere Möglichkeit ist die Verwendung von Wahrscheinlichkeitsverteilungen. Van Laere et al. (2010) berechnen für jedes Tag die χ^2-Verteilung und wählen jene mit lokalen Maxima als relevante Beschreibung von Standorten. Kessler et al. (2009) demonstrieren den Aufbau eines Bottom-up-Ortsverzeichnisses und identifizieren dazu aus den Metadaten von georeferenzierten Photos einerseits Ortsnamen und andererseits auch die dazugehörige geographische Ausdehnung.

Wie Flickr-Photos aufgrund ihrer Metadaten und durch visuelle Merkmale automatisch einer von zehn Sehenswürdigkeiten einer Stadt zugeordnet werden können, zeigt die Arbeit von Crandall et al. (2009). Dazu wird für jede Sehenswürdigkeit mittels ML ein Klassifikator mit Beispielen von positiven und negativen Vorkommen der Sehenswürdigkeit trainiert. Damit lassen sich anschliessend weitere Photos einem Ort beziehungsweise einer Sehenswürdigkeit zuordnen. Eine Fortsetzung davon findet sich bei Serdyukov et al. (2009), welche ebenfalls Flickr-Photos automatisch einer Lokalität zuweisen. Im Gegensatz zu Crandall et al. (2009) verwenden Serdyukov et al. (2009) allerdings nur die Metadaten (Tags) der Photos. Dazu spannen sie ein die ganze Erdoberfläche umfassendes Gitternetz auf und positionieren georeferenzierte Photos aus einem Trainings-Datensatz in die entsprechende Gitterzelle. Daraus erstellen sie ein Sprachmodell, das für jedes Tag eine Verteilungsfunktion berechnet. Anschliessend kann für ein Photo mit unbekanntem Standort mittels MLE eine Gitterzelle geschätzt werden.

Einen ähnlichen Ansatz wie das vorherige Beispiel verwenden Van Laere et al. (2010). Die automatische Lokalisierung von Photos ist allerdings auf 55 europäische

Städte beschränkt. Die räumliche Aufteilung erfolgt nicht durch ein regelmässiges Gitternetz wie bei Serdyukov et al. (2009), sondern durch ein Clustering-Verfahren[9]. Mit den berechneten Verteilungsfunktionen können anschliessend Photos dem vom Klassifikator als am wahrscheinlichsten geschätzten Cluster zugewiesen werden.

O'Hare & Murdock (2012) zeigen eine Fortsetzung und Erweiterung des Ansatzes von Serdyukov et al. (2009) und verwenden eine grössere Stichprobe für die Erstellung des Sprachmodells. Zudem verändern sie einige Parameter und Methoden des Modells geringfügig und erhalten dadurch signifikant bessere Werte. Die Autoren verweisen darauf, dass die Georeferenzierung von Photos eigentlich nicht das Hauptziel ihrer Forschung ist, sondern ein verbessertes Verständnis von räumlicher Semantik in Suchanfragen im Sinne von GIR (siehe Abschnitt 2.2.1). Allerdings stellen georeferenzierte Flickr-Photos eine gute Möglichkeit zur Validierung der Resulte dar. Das Konzept des Ansatzes lässt sich identisch auch auf andere UGC-Daten anwenden.

Wing & Baldridge (2011) bestimmen den wahrscheinlichsten Standort von Wikipedia-Artikeln und Nachrichten der Microblogging-Plattform *Twitter*[10] ähnlich Serdyukov et al. (2009) oder Van Laere et al. (2010) mit verschiedenen Varianten von ML-basierten Sprachmodellen. Da Wikipedia-Artikel dazu tendieren, viele Toponyme und sonstige Begriffe mit hohem räumlichem Bezug zu verwenden, kann ihr Standort mit einem probabilistischen Sprachmodell gut eingegrenzt werden. Die Lokalisierung der Twitter-Nachrichten entpuppt sich als schwieriger, da die maximale Zeichenzahl einer Nachricht auf 140 Zeichen beschränkt ist und mit 1.3 % (Stand 2012)[11] nur ein Bruchteil aller Nachrichten georeferenziert sind. Dies bestätigen auch Kinsella et al. (2011) und Cheng et al. (2010), die Sprachmodelle als Ansatz zur Georeferenzierung von Twitter-Nachrichten verwenden. Beide Untersuchungen zeigen auf, dass mit Hilfe von Sprachmodellen bessere Resultate erreicht werden, als bei der Georeferenzierung mittels identifizierten Toponymen.

Die Georeferenzierung von UGC dient somit der Evaluation umgangssprachlicher Ortsverzeichnisse aus den textuellen Beschreibungen von georeferenzierten UGC, welche zur Überbrückung der semantischen Lücke (Abschnitt 2.2.1) bei der Suche nach geographischen Informationen potentiell beitragen können.

2.5 Forschungslücken und Fragestellungen

Der Hintergrund der vorliegenden Arbeit wurde in den vorherigen Abschnitten ausführlich besprochen. Die Motivation zur Georeferenzierung von UGC lässt sich

[9] Zusammenfassen von Objekten zu einer Gruppe (Cluster) aufgrund ihrer ähnlichen Eigenschaften (MacKay, 2003)

[10] http://www.twitter.com/

[11] http://twitter.com/rajatgarg79/status/251321658428764160/ Zugriff: 27.09.2012

zusammenfassend mit den folgenden Hauptaspekten begründen.

- Die Konzepte der naiven Geographie fordern eine Implementierung menschlicher Wahrnehmung von Raum in GIS, deren Funktionen zunehmend von Laienanwendern genutzt werden. Zur Beschreibung der subjektiven Wahrnehmung von Raum und der räumlichen Semantik eignet sich die Analyse der natürlichen Sprache, wie sie in UGC vorkommt.
- Bisherige Methoden zur Suche nach Dokumenten berücksichtigen die räumliche Semantik von Suchanfragen nicht explizit. Die Disziplin des GIR beschäftigt sich deshalb mit der Frage, wie räumliche Semantik in Suchmaschinen implementiert werden kann.
- Die Entwicklung von Ortsverzeichnissen, die neben offiziellen Toponymen auch umgangssprachliche Ortsbezeichnungen enthalten oder menschliche Raumkonzepte mit unscharfen Grenzen sowie undeutlicher Ausdehnung unterstützen, ist eine mögliche Lösung zur Überwindung der semantischen Lücke. UGC kann als Grundlage solcher umgangssprachlichen Ortsverzeichnisse dienen. Dazu werden robuste Methoden benötigt, welche grosse Mengen von unstrukturierten Daten verarbeiten und relevante Informationen daraus extrahieren können.

Hill (2006) sieht Georeferenzierung – explizit durch Koordinaten oder implizit durch die Angabe eines Ortsnames – als Schlüsselelement in der Beschreibung, der Suche nach Information, Evaluation, Visualisierung und Benutzung von Informationssystemen aller Art.

Da der Zugang und die Suche nach Informationen mit geographischen Kontext für die tägliche Anwendung weder einfach noch intuitiv ist, müssen neue Methoden entwickelt werden, die räumliche Semantik in Suchanfragen interpretieren und nutzen können. Jones & Purves (2008) definieren die Aufgaben von GIR. Dabei sehen sie unter anderem Herausforderungen in der Identifizierung von Ortsnamen und räumlichen Begriffen in der natürlichen Sprache von Suchanfragen, der Mehrdeutigkeit von Ortsnamen, der geometrischen Interpretation von umgangssprachlichen Ortsbezeichnungen sowie der Bewertung von geographischer Relevanz.

Die naive Geographie (Egenhofer & Mark, 1995; Smith & Mark, 2001) sieht eine mögliche Lösung dafür in der Entwicklung von GIS, die subjektive menschliche Auffassungen von Raum berücksichtigen. Durch den enormen Erfolg und die Verbreitung von Web 2.0-Plattformen wie Flickr stehen riesige, bisher ungenutzte neue Datensammlungen zur Verfügung. Mit der Analyse der Metadaten von georeferenzierten nutzergenerierten Daten (UGC) bietet sich die Chance die Konzepte der naiven Geographie mit der natürlichen Sprache in den Metadaten zu verknüpfen und so subjektive, lokale Wahrnehmungen von Raum daraus zu filtern.

Wie räumliche Semantik aus den Tags von georeferenzierten Flickr-Photos extrahiert werden kann, zeigen Rattenbury & Naaman (2009). Sie verweisen auch auf die andere Anwendungsmöglichkeiten wie etwa verbesserter Suche nach Bildern, automatische Ergänzung von räumlichen Metadaten, Vorschläge für passende Tags oder die Erstellung von Daten für Ortsverzeichnisse zur Verbesserung des geographischen Kontexts in Suchanfragen. Purves et al. (2011) zeigen auf, dass trotz grossen Stichproben einzelne Nutzer zu Verzerrungen in den Resultaten führen können.

Ansätze zur Georeferenzierung von UGC findet man bei Crandall et al. (2009) oder Van Laere et al. (2010). Sie beschränken ihre Untersuchungsgebiete auf ausgewählte Städte. Weniger limitiert sind die flächendeckend anwendbaren Methoden mit regelmässigen Gitternetzen von Serdyukov et al. (2009) und O'Hare & Murdock (2012). Bei allen Anwendungen zeigt sich die Überlegenheit von Ansätzen, welche lokale Relevanz von natürlicher Sprache berücksichtigen, gegenüber solchen, die nur anhand von Ortsnamen georeferenzieren.

In dieser Arbeit wird versucht, durch die Kombination von verschiedenen Ansätzen eine Verbesserung der Georeferenzierung von nutzergenerierten Daten zu erreichen. Das Ziel ist eine automatisierbare Extraktion von ortsrelevenaten Beschreibungen aus den Tags georeferenzierter Flickr-Photos. Deshalb lautet die erste Forschungsfrage dieser Arbeit:

Forschungsfrage 1:

> *Mit welcher Genauigkeit können nutzergenerierte Daten ohne explizite Ortsangaben aufgrund ihrer textuellen Beschreibung georeferenziert werden?*

Bisher noch wenig untersucht wurde die Möglichkeit, dass sich aus UGC extrahierte räumliche Semantik auf UGC ohne explizite Angabe von geographischen Koordinaten anwenden lässt. Gschwend & Purves (2012) haben bei der Beschreibungen von geomorphologischen Landschaftselementen Unterschiede zwischen Photos von Flickr und *Geograph* festgestellt. Das Benutzerverhalten beim Tagging wurde zwar untersucht (z.B. Sigurbjörnsson & van Zwol, 2008), allerdings ist bisher nicht hinreichend nachgewiesen, ob Nutzer georeferenzierte Daten ähnlich oder unterschiedlich beschreiben als Daten ohne explizite Georeferenzierung mit geographischen Koordinaten. Die zweite Forschungsfrage widmet sich deshalb diesem Sachverhalt.

Forschungsfrage 2:

> *Wie ähnlich sind sich die textuellen Beschreibungen von Flickr-Photos mit Geotag und solchen ohne Geotag?*

Dass Einzelnutzer auch in grossen Stichproben zu Verzerrungen führen können, haben verschiedene Arbeiten gezeigt (Hollenstein & Purves, 2010; Van Laere et al., 2010; Purves et al., 2011). Die dritte Forschungsfrage gilt deshalb folgender Problematik.

Forschungsfrage 3:

> *Welchen Einfluss hat die Filterung nutzergenerierter Daten auf die Qualität der Georeferenzierung?*

2.6 Begriffsdefinition

Im Kontext dieser Arbeit werden die folgenden Begriffe häufig verwendet:

- *Tags*: Schlüsselwörter zur Beschreibung von Photos
- *Bulk Uploads*: Photos eines Nutzers die mit identischen Schlüsselwörtern (Tags) beschrieben werden
- *Geotag*: explizite Beschreibung des Standortes georeferenzierter Photos mit geographischen Koordinaten
- *Tagscore*: die räumliche Relevanz eines Tags, ausgedrückt als Wert der TF-IDF-Methode von Rattenbury & Naaman (2009)
- *ortsrelevante Tags*: Tags von georeferenzierten Photos, die aufgrund ihrer Tagscore zur Beschreibung eines bestimmten Standorts als relevant bewertet werden
- *Toponym*: Ortsname, der auch in offiziellen Ortsverzeichnissen verwendet wird
- *umgangssprachliche Ortsbezeichnung*: neben Toponymen (z.B. *New York City*) werden auch umgangssprachliche Namen für Orte verwendet (z.B. *Big Apple*)
- *Stoppwörter*: kommen in der natürlichen Sprache häufig vor, haben aber wenig semantische Relevanz und bei sprachlichen Analysen oft entfernt. Beispiele sind *du, sind, hier* oder *zu*.

3 Untersuchungsgebiet und Datengrundlagen

In diesem Kapitel werden die beiden Untersuchungsgebiete gezeigt und ihre Wahl begründet. Es folgt eine detaillierte Beschreibung der Datengrundlage der Online-Photoplattform Flickr, der Ortsverzeichnisse und eine kurze Auflistung der in dieser Arbeit verwendeten Software.

3.1 Untersuchungsgebiet

Als Untersuchungsgebiete für die Fragestellungen dieser Arbeit werden Grossbritannien und die Schweiz ausgewählt (Abbildung 3.1). In Grossbritannien – bestehend aus England, Schottland und Wales – umfasst der Perimeter die Hauptinsel des Vereinigten Königreichs ohne Nordirland. Die beiden Ländern eignen sich als Untersuchungsgebiete aufgrund der Verfügbarkeit von offiziellen Ortsverzeichnissen und der guten Abdeckung mit georeferenzierten Flickr-Photos.

Tabelle 3.1 zeigt die wichtigsten Charakteristika der beiden untersuchten Länder.

Tab. 3.1: Vergleich Grossbritannien – Schweiz (Worldatlas, 2012; CIA, 2012)

	Grossbritannien	Schweiz
Fläche	229'736 km^2	41'277 km^2
Landessprache(n)	Englisch	Deutsch, Französisch, Italienisch, Rätoromanisch
Einwohner (2012)	62.44 Mio.	7.92 Mio.
Urbanisierunggrad (2010)[a]	80 %	74 %
Internetnutzer (2009)	51.4 Mio.	6.1 Mio

[a] Anteil der Gesamtbevölkerung in urbanen Gebieten

Grossbritannien ist flächenmässig fast sechsmal so gross wie die Schweiz, die Bevölkerungszahl ist sogar achtmal grösser. In beiden Ländern nutzen grosse Teile der Gesamtbevölkerung das Internet, weshalb sie sich gut für eine Analyse basierend auf nutzergenerierten Daten eignen. Der Urbanisierungsgrad ist in beiden Ländern sehr hoch, was sich auch in der räumlichen Verteilung der nutzergenerierten Daten widerspiegelt (Abschnitt 5.1).

(a) Grossbritannien

(b) Schweiz

Abb. 3.1: Übersicht der beiden Untersuchungsgebiete mit den grössten Städten
Daten: Natural Earth, Swisstopo

3.2 Flickr

Flickr[1] ist eine Internet-Photoplattform, welche im Februar 2004 von der kanadischen Firma *Ludicorp* lanciert wurde. Ursprünglich als Teil des Onlinespiels *Game Never Ending* geplant, wurde der Nutzen von Flickr – Hochladen, Organisieren, Verwalten und Teilen von Photos im Internet – vor dem Hintergrund von Web 2.0 und erschwinglichen digitalen Photokameras schnell erkannt und weiterentwickelt. Flickr wuchs rapide und wurde im März 2005 von der Internetfirma *Yahoo!* gekauft und weiterbetrieben (Naughton, 2008; Fake, 2005; Cox, 2008).

Im August 2011 durchbrach Flickr die Grenze von 6 Milliarden hochgeladenen Bilder (Kremerskothen, 2011), mittlerweile (Stand November 2012) sind es bereits über 8 Milliarden, wovon laut Cox (2008) etwa 80 % öffentlich zugänglich sind. Der Erfolg von Flickr beruht gemäss Naughton (2008) neben der oben beschriebenen Funktionalität auch auf der Programmierschnittstelle[2] (*Application Programming Interface* – API). Diese erlaubt es ohne grossen Aufwand Funktionalität von Flickr in neue Anwendungen oder Webseiten einzubinden.

Ein zentrales Element von Flickr ist das Versehen von Photos mit Schlüsselwörtern – sogenannten *Tags*. Diese helfen den Nutzern ihre Daten zu organisieren, und der Öffentlichkeit bei der Suche nach Photos mit bestimmten Schlüsselwörtern. Das *Tagging* als Organisationssystem wurde von der ebenfalls zu *Yahoo!* gehörenden Webseite *Del.icio.us*[3] übernommen (Cox, 2008; Nov & Ye, 2010).

Flickr bietet den Nutzern auch die Beschreibung der Photos mit geographischen Koordinaten, auch *Geotags* genannt, an. Dies geschieht entweder per Auswahl der Position auf einer Karte oder mit GPS-Daten, welche in den Metadaten der Photos gespeichert werden können (Ahern et al., 2007). Laut Flickr Code Blog (2009) hatten im Jahr 2009 rund 3.3 % der Photos auf Flickr Geotags, wovon rund zwei Drittel öffentlich einsehbar sind. Auf dieser Datengrundlage basiert die vorliegende Arbeit.

Eine ausführliche Beschreibung des Vorgehens zur Datensammlung folgt in Abschnitt 4.1, statistische Kennwerte und Visualisierungen der Datensätze in Abschnitt 5.1.

[1] http://www.flickr.com/
[2] http://www.flickr.com/services/api/
[3] http://delicious.com/

3.3 Offizielle Ortsverzeichnisse mit Toponymen

Zur Identifizierung von Toponymen und zur Beurteilung der Resultate der Georeferenzierung mit UGC werden in dieser Arbeit zwei offizielle Ortsverzeichnisse aus Grossbritannien und der Schweiz verwendet. Datenlieferanten sind die dafür zuständigen Behörden des jeweiligen Landes, *Ordnance Survey* in Grossbritannien und *Swisstopo* in der Schweiz.

Die beiden Ortsverzeichnisse enthalten neben den Ortsnamen und deren Koordinaten im jeweiligen nationalen Referenzsystem noch weitere Attribute. Tabelle 3.2 fasst die wichtigsten Kennwerte zusammen (Ordnance Survey, 2012; Swisstopo, 2008). Die räumliche Verteilung der Toponyme in Grossbritannien und der Schweiz zeigen die Karten im Anhang (siehe Abschnitt A.4).

Tab. 3.2: Übersicht der offiziellen Ortsverzeichnisse

	Grossbritannien	Schweiz
Produkt	1:50 000 Scale Gazetteer	SwissNames25
Datenlieferant	Ordonance Survey	Swisstopo
Anzahl Toponyme	$>$ 250'000	$>$ 150'000
Toponyme pro km^2	1.1	3.8
Anteil mehrdeutige Toponyme	11.8 %	14.4 %
Stand	2012	2008
Massstab	1:50'000	1:25'000
Präzision	1km	keine Angaben
Geometrische Repräsentation	Punkt	Punkt

3.4 Verwendete Software

Die Sammlung, Analyse und statistische Auswertung der Daten wurde in der Programmiersprache Python[4] implementiert. Die einzelnen Arbeitsschritte wurden mit Hilfe der in Tabelle 3.3 aufgeführten Zusatzpakete und eigens programmierten Skripten ausgeführt. Für die kartographische Aufbereitung und Visualisierung der Ergebnisse wurde die Software ArcGIS 10.0 der Firma Esri[5] verwendet.

Tab. 3.3: Verwendete Python-Pakete

Paketname	Anwendungsbereich
flickrapi 1.4.2	Schnittstelle zur Flickr-API zur Sammlung der Flickr-Metadaten
Numpy 1.6.1	Lineare Algebra und Statistik
pyproj 1.9.0	Transformation von geographischen Koordinaten
pyshp 1.1.4	Importieren von geographischen Daten (Shapefiles)
NLTK 2.0	Filterung von Stoppwörtern
Matplotlib 1.1.0	Erstellung von Diagrammen

[4] http://python.org/
[5] http://www.esri.com/

4 Methodik

Im Folgenden werden alle Teilschritte und die verwendeten Methoden der Arbeit, welche zur Beantwortung der in Abschnitt 2.5 formulierten Forschungsfragen nötig sind, ausführlich beschrieben.

Das Hauptziel dieser Arbeit ist die Georeferenzierung nutzergenerierter Daten – oder konkreter Flickr-Photos – mithilfe eines umgangssprachlichen Ortsverzeichnisses sowohl in Grossbritannien als auch in der Schweiz. Dieses wird durch die Extraktion von ortsrelevanten Tags aus georeferenzierten Flickr-Photos erstellt (vgl. Abschnitt 1.2).

Abbildung 4.1 zeigt einen Überblick des methodischen Vorgehens und gibt für jeden Teilschritt den entsprechenden Abschnitt in diesem Kapitel an.

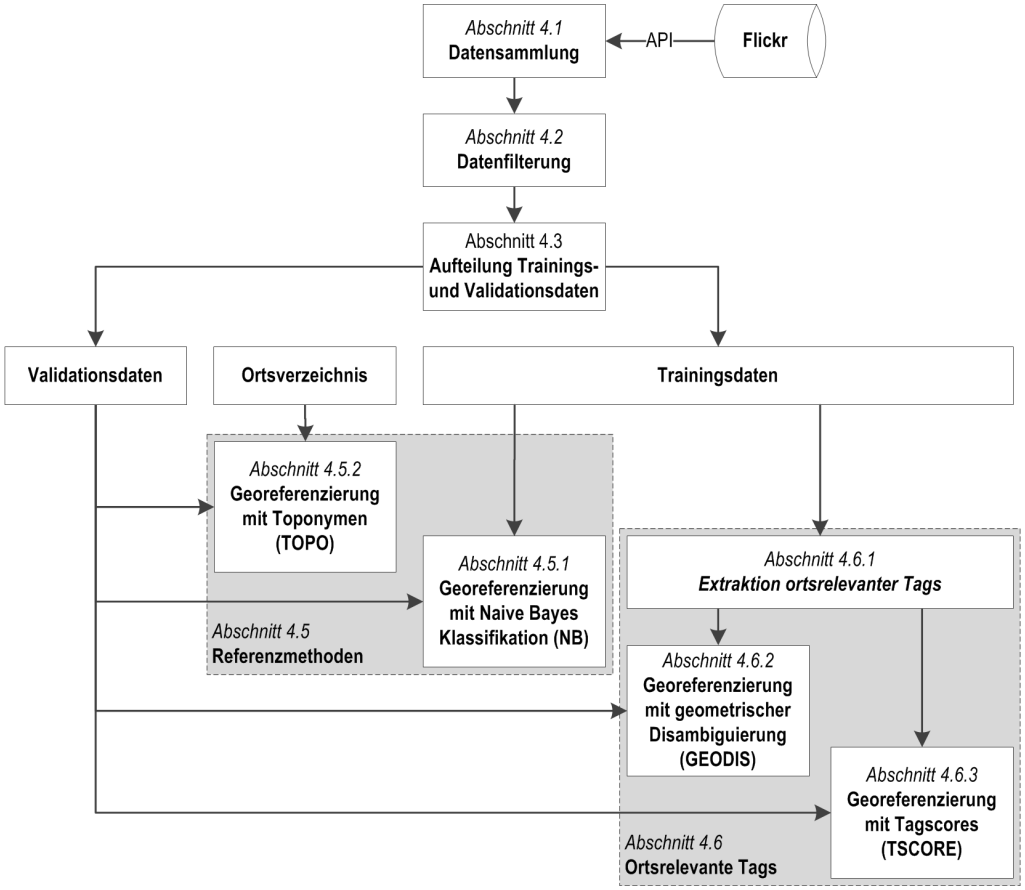


Abb. 4.1: Überblick der methodischen Teilschritte

Nach der Datensammlung (Abschnitt 4.1) und der Datenfilterung (Abschnitt 4.2) werden die georeferenzierten Flickr-Photos in einen Trainings- und einen Validationdatensatz aufgeteilt (Abschnitt 4.3). Die Trainingsdaten dienen einerseits zur Extraktion ortsrelevanter Tags (Abschnitt 4.6.1) und andererseits zur Erstellung eines ML-Klassifikators. Die Photos in den Validationsdaten werden anschliessend mit den verschiedenen Methoden georeferenziert.

Zum Vergleich der Ergebnisse wird die Georeferenzierung mit zwei Referenzmethoden durchgeführt (Abschnitt 4.5). Die TOPO-Methode verwendet die offiziellen Ortsverzeichnisse (siehe Abschnitt 3.3) zur Identifizierung von Toponymen in den Tags der Bilder. Die NB-Methode schätzt die Position der Photos mit einem probabilistischen ML-Klassifikator (Abschnitt 4.5.1).

Zur Extraktion ortsrelevanter Tags aus den Photos der Trainingsdaten wird die TF-IDF-Methode von Rattenbury & Naaman (2009) verwendet (Abschnitt 4.6.1). Die ortsrelevanten Tags werden in einem umgangssprachlichen Ortsverzeichnis gespeichert und dienen den beiden Methoden GEODIS und TSCORE zur Georeferenzierung der Photos im Validationsdatensatz. GEODIS verwendet eine einfache geometrische Disambiguierung, TSCORE nutzt normierte Tagscores der TF-IDF-Methode zur Bestimmung des Standortes.

Die Wahl der Methoden korrespondiert mit den Zielen dieser Arbeit (vgl. Abschnitt 1.2) und lässt sich folgendermassen begründen:

- In den Forderungen der naiven Geographie und den Unzulänglichkeiten bei der Suche nach räumlich bezogenen Informationen (vgl. Abschnitt 2.2.1) gründet das Ziel der Erstellung umgangssprachlicher Ortsverzeichnise. Dafür sind georeferenzierte Flickr-Photos eine mögliche Quelle von UGC (siehe Abschnitt 2.4.1).
- Die Informationen in den Tags der Flickr-Photos liegen in unstrukturierter Form vor, ohne a priori räumliche Semantik zu enthalten. TF-IDF ist eine geeignete Methode um ortsrelevante Tags und somit räumliche Semantik für ein umgangssprachliches Ortsverzeichnis zu extrahieren (vgl. Abschnitt 2.4.4).
- Die verwendete Variante der TF-IDF-Methode berücksichtigt neben der räumlichen Verteilung eines Tags auch die Verteilung unter Nutzern und somit – im Gegensatz zur „naiven“ ML-Klassifikation – auch die umgangssprachliche Repräsentativität.
- Durch die Resultate der Georeferenzierung kann der Nutzen des umgangssprachlichen Ortsverzeichnisses evaluiert und mit den Resultaten der Referenzmethoden verglichen werden. Ähnliche Untersuchungen haben das Potential und die Überlegenheit von umgangssprachlichen Ansätzen gegenüber solchen, die nur Toponyme verwenden, aufgezeigt (vgl. Abschnitt 2.4.4).

- Mit der vorgängigen Filterung und Minimierung von Nutzerverzerrungen werden aussagekräftigere Resultate erwartet (siehe Abschnitt 2.4.3).
- Die verwendete Methodik mit einem regelmässigen Gitternetz kann räumlich flächendeckend und in unterschiedlichen Auflösungen verwendet werden. Zudem ist der verwendete Ansatz verhältnismässig simpel in der Implementierung, die Berechnung effizient und nachvollziehbar.

Das Vorgehen zur Untersuchung des Einflusses der Datenfilterung auf die Ergebnisse der Georeferenzierung und ein Vergleich der textuellen Beschreibungen von Flickr-Photos mit Geotag und ohne Geotag zeigen die beiden letzten Abschnitte 4.7 beziehungsweise 4.8 in diesen Kapitel.

4.1 Datensammlung

Über die API von Flickr lassen sich sämtliche öffentlich zugänglichen Photos durchsuchen und die gewünschten Metadaten abrufen. Die Suche nach bestimmten Photos kann entweder thematisch mit verschiedenen Attributen wie Tags oder für Photos mit Geotag auch räumlich durch die Definition einesMBR eingeschränkt werden. Für die beiden Eckpunkte des MBR (unten links, oben rechts) sind die geographischen Koordinaten in Dezimalgrad im Referenzsystem WGS84 zu definieren.

Der Zugriff auf die API erfolgte mit dem API-Wrapper *flickrapi*[1] für Python. Da je nach verwendeten Suchparametern eine potentiell grosse Anzahl an Photos im Ergebnis möglich ist, begrenzt die API die maximale Anzahl Photos pro Suchanfrage automatisch auf die zuerst gefundenen 4'000. Ausserdem ist die Nutzung der API auf 3'600 Anfragen pro Stunde limitiert (eine Anfrage entspricht ungefähr einem Photo).

Durch die Verwendung zweier zusätzlicher Suchparameter, die ein sich automatisch an die Datenmenge anpassendes Zeitfenster für den Upload-Zeitpunkt der Photos definieren, konnten Suchanfragen mit zu vielen Photos im Ergebnis verhindert werden[2]. Das Überschreiten der erlaubten API-Nutzung wurde mit variablen Pausen zwischen den einzelnen Suchanfragen gelöst. Das Python-Skript konnte somit über längere Zeiträume unbeaufsichtigt die nötigen Metadaten empfangen und in ein Textfile speichern.

Für die vorliegende Arbeit wurden je ein Datensatz für Grossbritannien (GB) und die Schweiz (CH) mit den Metadaten aller öffentlich zugänglichen, georeferenzierten

[1] http://stuvel.eu/flickrapi/
[2] Die Idee wurde aus Quellcode von R. Purves übernommen

Photos[3] innerhalb des jeweiligen umgebenden Rechtecks extrahiert. Für den Vergleich von georeferenzierten Photos und solchen ohne Georeferenzierung wurde zusätzlich ein Datensatz für Grossbritannien extrahiert. Die Einschränkung der Suche nach Photos innerhalb von Grossbritannien erfolgt dabei mit häufig vorkommenden Schlüsselwörtern. Tabelle 4.1 zeigt eine Übersicht der extrahierten Datensätze.

Tab. 4.1: Extrahierte Flickr-Datensätze

Datensatz	Suchmethode	Parameter	Zeitraum	Anzahl
GB mit Geotag	räumlich mit MBR	-8.830, 49.684[1] 1.949, 61.041[2]	1.1.2004 – 31.5.2012	6'663'048
CH mit Geotag	räumlich mit MBR	5.891, 45.773[1] 10.557, 47.853[2]	1.1.2004 – 24.11.2012	876'182
GB ohne Geotag	thematisch mit Tags	england, unitedkingdom, uk, scotland, britain, greatbritain	1.1.2011 – 31.12.2011	2'627'983

[1] MBR-Ecke unten links (Längen-, Breitengrad WGS84)
[2] MBR-Ecke oben rechts (Längen-, Breitengrad WGS84)

Tabelle 4.2 zeigt, welche Attribute aus den Metadaten der Photos empfangen wurden. Alle Photos lassen sich über das Attribut `photoId` eindeutig identifizieren und mit `photoOwner` einem Benutzer zuordnen[4]. Die geographischen Koordinaten (`photoLatitude`, `photoLongitude`) sind in Dezimalgrad im Koordinatensystem WGS84 vorhanden. Die Präzision der Georeferenzierung (`photoAccuracy`) wird mit einer Zahl im Bereich von 1 bis 16 angegeben, die der Massstabsebene beziehungsweise der Zoomstufe der Karte während dem manuellen Georeferenzieren entspricht (1: Welt, 3: Land, 6: Region, 11: Stadt, 16: Strasse) (Flickr, 2012). Wenn in den Metadaten eines hochgeladenen Photos bereits Koordinaten vorhanden sind, wird immer die höchstmögliche Präzision zugeordnet. Es muss beachtet werden, dass aus einer hohen Präzision (Detailierungsgrad der Koordinaten) nicht automatisch auf eine hohe Lagegenauigkeit (Abweichung vom realen Standort) eines Geotags geschlossen werden kann (vgl. auch Abschnitt 2.4.3). Die Tags werden in ihrer Rohform wie vom Benutzer eingegeben mit Akzenten, Sonder- und Leerzeichen gespeichert.

[3] Die Photos selber wurden nicht extrahiert, da die Methoden nur deren Metadaten verwenden. Der Begriff *Photo* wird in dieser Arbeit deshalb als Synonym für die Metadaten des Photos verwendet.

[4] Aus Sicht des Datenschutzes ist somit eindeutige Identifizierung der Nutzer möglich. Allerdings lässt sich nur auf Photos zugreifen, die von den Besitzern als öffentlich authorisiert wurden. Zudem werden in dieser Arbeit nur Benutzernummern, jedoch niemals „Klarnamen“ verwendet, die zudem auch Pseudonyme sein können.

Tab. 4.2: Attribute der Photo-Metadaten

Attribut	Beschreibung
`photoId`	eindeutige Photonummer
`photoOwner`	eindeutige Benutzernummer
`photoTitle`	Titel des Photos
`photoDatePosted`	Zeitpunkt der Veröffentlichung
`photoDateTaken`	Zeitpunkt der Aufnahme
`geoTag`	Geotag
`photoLatitude`	Geographische Länge
`photoLongitude`	Geographische Breite
`photoAccuracy`	Präzision der Georeferenzierung
`photoUrl`	Link zum Photo
`photoTagCount`	Anzahl Tags
`photoRawTags`	Tags (in Rohform)

4.2 Datenfilterung

4.2.1 Grundlegende Filterschritte

Bei der Suche über die API von Flickr kann es zu Duplikaten im extrahierten Datensatz kommen, etwa durch Verbindungsunterbrüche. Diese werden nachträglich aus dem Datensätz gefiltert. Da Photos ohne Tags für die Analyse von räumlicher Semantik keinen Nutzen haben, werden sie entfernt. Ebenfalls ausgeschlossen werden Photos mit fehlerhaften Zeitangaben, wie z.B. *00.00.2000*.

Da die Analyse von ortsrelevanten Tags räumlich möglichst fein aufgelöst erfolgen soll, werden nur Bilder mit der höchsten Präzision (`photoAccuracy = 16`) berücksichtigt (siehe Abschnitt 4.1) (Purves et al., 2011). In beiden Datensätzen ist dies bei über 99 % aller Bilder der Fall. Zudem werden Bilder mit mehr als 5 km Abstand von den Landesgrenzen von Grossbritannien und der Schweiz entfernt.

Im nächsten Schritt werden mögliche Stoppwörter in den jeweiligen Landessprachen (GB: Englisch, CH: Deutsch, Englisch, Französisch, Italienisch) aus den Tags entfernt (Manning et al., 2009; Bird et al., 2009). Zusätzlich entfernt wurde eine manuelle Liste von Stoppwörtern (Tabelle A.1). Die Liste wurde aus den 500 häufigsten Tags erstellt und enthält zum einen photospezifische Begriffe wie zum Beispiel *canon, hdr* oder *black and white*, zum anderen Wörter, die beim Upload mit populären Photo-Anwendungen für Smartphones *(Instagram, Hipstamatic)*[5] automatisch hinzugefügt werden. Ebenfalls entfernt werden *Machine-Tags*, die alle dem Schema `Namensraum:Eigenschaft=Wert` folgen (z.B. *geo:lat=8.538*). Sie werden

[5] http://instagram.com/ | http://hipstamatic.com/

ebenfalls meist automatisch zugewiesen werden und dienen zur systematischen Suche nach bestimmten Photos in bestimmten Namensräumen[6].

Danach werden die Tags für die weitere Analyse normalisiert. Umlaute und Akzente werden in ihre Grundform (*Àird Tòranais* wird zu *airdtoranais*) und alle Zeichen in Kleinschreibung umgewandelt sowie Interpunktion und allfällige Leerzeichen entfernt Manning et al. (2009). Die Normalisierung hat den Vorteil, dass zusammengesetzte Tags wie beispielsweise *Hyde Park* zu *hydepark* verbunden werden und so die semantische Eigenschaften gewahrt bleiben.

4.2.2 Entfernung von Nutzerverzerrungen

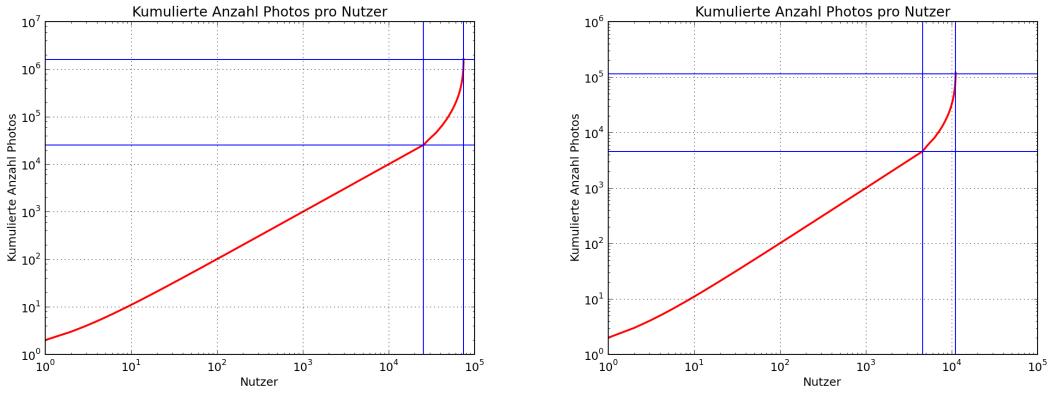
In Abschnitt 2.4.3 wurde bereits der mögliche überproportionale Einfluss von Einzelnutzern in UGC-Stichproben erwähnt. Die dadurch resultierenden Verzerrungen werden durch geeignete Filterung minimiert.

Zuerst werden sogenannte *Bulk Uploads* identifiziert. Diese treten auf, wenn mehrere Bilder eines Nutzers durch Stapelverarbeitung mit identischen Tags hochgeladen werden. Oftmals werden auch die gleichen Geotags für alle Bilder verwendet, was zu einer lokalen Überrepräsentation der vom Nutzer verwendeten Tags führt. Solche Bulk Uploads werden aus dem Datensatz entfernt (Grothe & Schaab, 2009; Van Laere et al., 2010).

Danach werden alle Photos von Nutzern entfernt, die mit überdurchschnittlich vielen oder nur einem einzigen Photo in der Stichprobe vorkommen (Ames & Naaman, 2007; Purves et al., 2011). Ihr Beitrag kann die Repräsentativität der Stichprobe bei der Analyse von räumlich relevanten Tags potentiell verzerren. Der obere Schwellenwert für die maximale Anzahl Photos pro Nutzer, welche zum Ausschluss aus der Stichprobe führen, wurde für die Datensätze von Grossbritannien und der Schweiz manuell anhand der Verteilung in Abbildung 4.2 festgelegt. In Grossbritannien werden Nutzer mit mehr als 10′000, in der Schweiz mit mehr als 1′000 Photos aus dem Datensatz ausgeschlossen.

Tags, die nur von einem einzigen Nutzer oder insgesamt weniger als zehnmal zur Beschreibung von Photos eingesetzt wurden, gelten als nicht repräsentativ zur Ortsbeschreibung und werden deshalb entfernt (Rattenbury & Naaman, 2009; Van Laere et al., 2010).

[6] http://www.flickr.com/help/tags/



(a) Grossbritannien (b) Schweiz

Abb. 4.2: Kumulierte Anzahl Photos pro Nutzer (rot) und die Grenzen (blau) für Benutzer mit einer Anzahl Photos unter-/oberhalb der Schwellenwerte

4.3 Aufteilung in Trainings- und Validationsdaten

Anschliessend an die Filterung werden die verbleibenden Photos in Trainings- und Validationsdaten aufgeteilt. Die Trainingsdaten werden einerseits zur Erstellung eines umgangssprachlichen Ortsverzeichnisses (Abschnitt 4.6.1) und andererseits zum Training eines ML-Klassifikators (Abschnitt 4.5.1) benötigt. Für die Photos in den Validationsdaten wird nur mithilfe ihrer Tags mit allen vier Georeferenzierungsmethoden (NB, TOPO, GEODIS und TSCORE) der mögliche Standort geschätzt. Wie die Photos in den Trainingsdaten enhalten auch jene in den Validationsdaten Geotags. Mit diesen können die Fehlerdistanzen zwischen den geschätzen und den in den Geotags definierten Standorten berechnet werden (vgl. Abschnitt 4.4).

In dieser Arbeit werden zwei verschiedene Zufallsverfahren zur Aufteilung der Photos in Trainings- und Validationsdaten verwendet: (siehe Abbildung 4.3)

- *Zufällig pro Photo (RP)*: für jedes Photo wird zufällig über seine Zugehörigkeit zur Trainings- oder Validationsstichprobe entschieden
- *Zufällig pro Nutzer (RU)*: alle Photos eines Nutzers werden zufällig gesamthaft der Trainings- oder Validationsstichprobe zugeordnet

Mit dem Zufallsverfahren *RP* kann es bei der späteren Georeferenzierung eines Photos in den Validationsdaten theoretisch vorkommen, dass dessen Position hauptsächlich durch ortsrelevanten Tags von Photos desselben Nutzers im Trainingsdatensatz korrekt vorhergesagt werden kann. Obwohl durch die Filterung mindestens zwei verschiedene Nutzer ein Tag verwenden müssen, ist die Möglichkeit von solchen

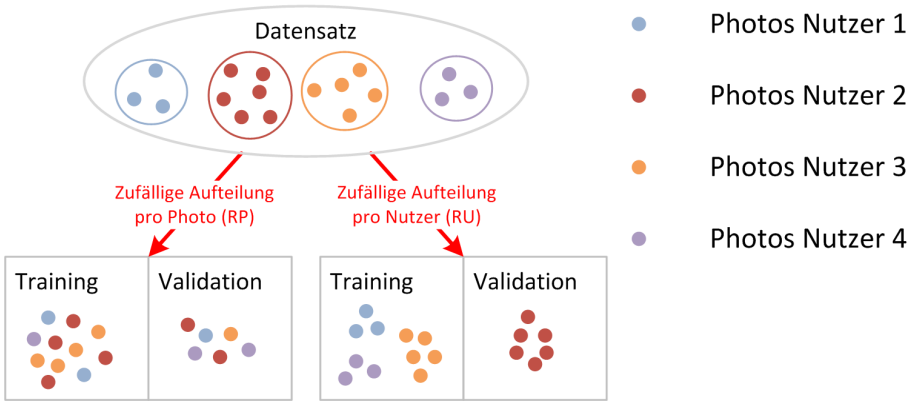


Abb. 4.3: Varianten zur Aufteilung der Photos in Trainings- und Validationsdaten

Verzerrungen potentiell vorhanden. Deren Einfluss lässt sich erkennen, wenn die Resultate der Georeferenzierung von mit *RP* aufgeteilten Daten mit jenen, die durch *RU* aufgeteilt wurden, verglichen werden.

Das Verhältnis zwischen der Anzahl Photos in den Trainingsdaten und jenen in den Validationsdaten beträgt 75:25 (*T75V25*). Zur Überprüfung der Robustheit der Methoden mit weniger Trainingsdaten kommen auch das Verhältnis von 25:75 (*T25V75*) zur Anwendung.

Falls es in den Resultaten im nächsten Kapitel nicht explizit angegeben ist, wird immer das Zufallsverfahren *RP* mit Splitverhältnis *T75V25* verwendet.

4.4 Berechnung der Genauigkeit der Georeferenzierung

Für die Photos in den Validationsdaten wird in den beiden nächsten Abschnitten sowohl mit den Referenzmethoden NB und TOPO (Abschnitt 4.5) als auch mit den Methoden GEODIS und TSCORE mithilfe der ortsrelevanten Tags (Abschnitt 4.6) der Standort bestimmt.

Die beiden Untersuchungsgebiete GB und CH werden dazu in ihrem jeweiligen Bezugssystem (GB: *OSGB1936*, CH: *CH1903*) mit einem regelmässigen orthogonalen Gitternetz in diskrete Zellen eingeteilt (Abbildung 4.4).

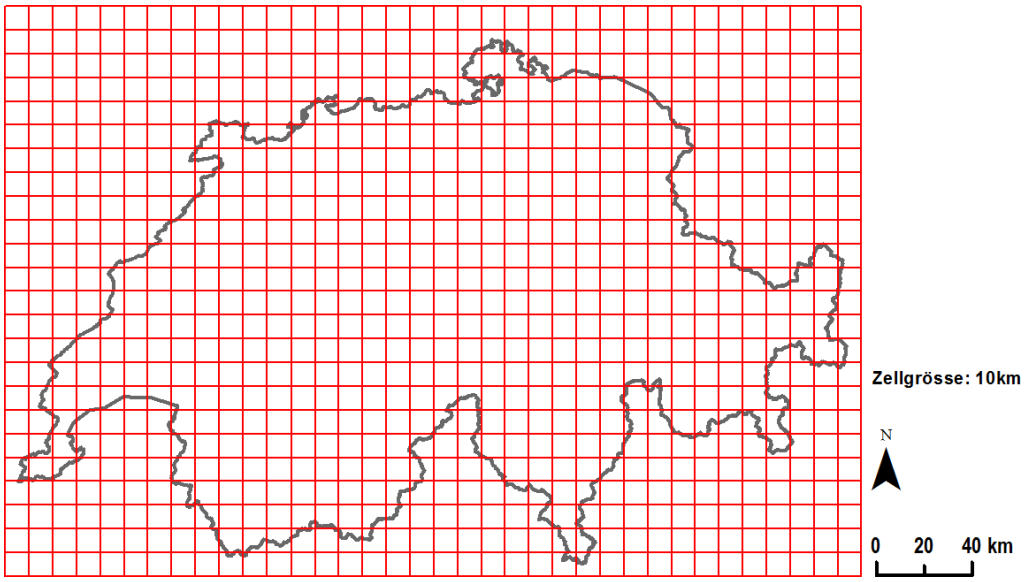


Abb. 4.4: Die Schweiz eingeteilt in ein regelmässiges Gitternetz (Zellgrösse 10 km)
Daten: Swisstopo

Es werden Zellgrössen von 1 km, 5 km und 10 km verwendet um damit auch das Problem der willkürlichen Aggregierung von Photos mit Punktkoordinaten in Gitterzellen als räumliche Analyseeinheiten (*„modifiable areal unit problem – MAUP“* Openshaw, 1984) zu untersuchen.

Bei der Georeferenzierung wird für jedes Photo im Validationsdatensatz eine Gitterzelle als möglicher Standort festgelegt. Die Lagegenauigkeit der Georeferenzierung berechnet sich als euklidische Fehlerdistanz zwischen dem Mittelpunkt der geschätzten Zelle und dem Mittelpunkt der korrekten Zelle, die das Photo enthält.

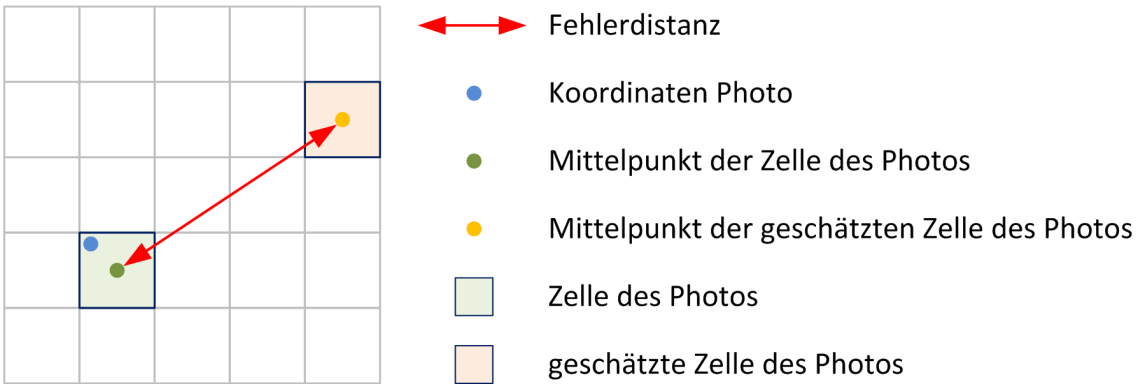


Abb. 4.5: Berechnung der Fehlerdistanz zwischen den Mittelpunkten zweier Zellen

Die Berechnung des globalen Distanzfehlers einer Methode wird aus den Fehlerdistanzen aller Photos berechnet. Da der arithmetische Mittelwert stark durch Ausreisser beeinflusst werden kann, ist der Medianwert aussagekräftiger zur Beurteilung des Ergebnisses. Dieser gibt den maximalen Distanzfehler für die 50 % am besten geschätzten Standorte an. Eine Mediandistanzfehler von beispielsweise 5 km bedeutet somit, dass der Distanzfehler für die Hälfte aller positionierten Photos höchstens 5 km beträgt. Die mögliche Präzision der Georeferenzierung ist, bedingt durch das Gitternetz, direkt von der Zellgrösse abhängig (vgl. Abschnitt 6.1.4).

Neben der Fehlerdistanz ist auch der Anteil der Photos von Interesse, für welche die korrekte Zelle oder eine benachbarte Zelle bestimmt wurde (*„Trefferquote“*). Abbildung 4.6 zeigt die Nachbarschaft einer korrekten Zelle `n0` mit den direkten Nachbarzellen (`n1`) sowie Nachbarzellen zweiten (`n2`) und dritten (`n3`) Grades.

n3	n3	n3	n3	n3	n3	n3
n3	n2	n2	n2	n2	n2	n3
n3	n2	n1	n1	n1	n2	n3
n3	n2	n1	n0	n1	n2	n3
n3	n2	n1	n1	n1	n2	n3
n3	n2	n2	n2	n2	n2	n3
n3	n3	n3	n3	n3	n3	n3

Abb. 4.6: Definiton der Nachbarzellen `n1`, `n2` und `n3` einer Zelle `n0`

4.5 Georeferenzierung mit Referenzmethoden

In den folgenden beiden Abschnitten wird die Methodik der beiden Referenzmethoden NB und TOPO zur Georeferenzierung von Photos erklärt. Ihre Resultate dienen zum Vergleich und der Beurteilung der Georeferenzierung mit den beiden Methoden GEODIS und TSCORE (vgl. Abschnitt 4.6).

4.5.1 NB – Statistisches Sprachmodell mit maschinellem Lernen

Eine häufig verwendete Methode zur Georeferenzierung ist das maschinelle Lernen (ML) mit einem Sprachmodell (z.B. in Serdyukov et al., 2009 oder Wing & Baldridge, 2011). Beim ML wird mit einem Trainingsdatensatz aus Dokumenten mit bekannten

Kategorien ein Klassifikator trainiert. Dieser kann anhand der inhärenten Charakteristiken des Datensatzes die Zugehörigkeit eines unbekannten Dokuments zu einer Kategorie schätzen. Aufgrund der im Trainingsdatensatz definierten Kategorien spricht man auch von beaufsichtigtem ML (Sebastiani, 2002).

Ursprünglich aus dem Bereich des IR und der Textklassifikation stammend (Manning et al., 2009), kann ein multinomialer *Naive Bayes*-Klassifikator (NB) auch für räumliche Anwendungen, wie beispielsweise zur Georeferenzierung verwendet werden (Van Laere et al., 2010). Der in dieser Arbeit verwendete NB-Klassifikator entspricht exakt demjenigen in Manning et al. (2009). Der Ansatz kann zur Vorhersage des Standorts von Flickr-Photos unverändert übernommen werden, wenn die Kategorien durch Zellen eines Gitternetzes, die Dokumente durch Photos und die Terme durch Tags ersetzt werden.

Der NB-Klassifikator geht von der Annahme des unabhängigen Auftretens und der Reihenfolge der Tags eines Photos aus. Ausserdem nehmen ML-Methoden auch implizit an, dass die Trainings- und Validationsdaten aus einer ähnlichen Verteilung stammen. Diese Annahmen treffen in der Realität allerdings selten zu. Im Gegenteil, Tags treten in der Beschreibung von Photos nicht unabhängig voneinander auf. Die in Textdokumenten bedeutende Reihenfolge von Wörtern ist für die Tags aufgrund ihrer unstrukturierten Natur nicht von Bedeutung (Abschnitt 2.4.2). Obwohl NB-Klassifikation durch die genannten Verkürzungen ein „naiver“ Ansatz ist, werden damit gute Ergebnisse erzielt. Zwar schätzt NB-Klassifikation die absoluten Wahrscheinlichkeiten für die Zugehörigkeit eines Photos zu verschiedenen Zellen ungenau, allerdings entscheidet aufgrund von MLE nur die grösste Wahrscheinlichkeit für die Vorhersage der richtigen Zelle. Gleichzeitig ist der Ansatz aufgrund der Annahmen simpel, effizient und auch robust gegen Ausreisser in den Daten.

Der NB-Klassifikator nach Manning et al. (2009) ordnet in der räumlichen Anwendung einem Photo p jene Zelle c zu, für die basierend auf den Häufigkeiten der Photos und Tags in den Trainingsdaten und den Tags des Photos p die grösste Wahrscheinlichkeit (MLE) geschätzt wird. Die Wahrscheinlichkeit $P(c|p)$ der Zugehörigkeit eines Photos aus den Validationsdaten zu einer bestimmten Zelle, berechnet sich aus der allgemeinen Wahrscheinlichkeitsverteilung aller Photos in den Trainingsdaten *(prior probability)* $P(c)$ und der bedingten Vorkommenswahrscheinlichkeit $P(t_k|c)$ eines Photos in einer bestimmten Zelle *(conditional probability)* aufgrund der darin enthaltenen Tags t_k (Gleichung 4.1).

$$P(c|p) \propto P(c) \prod_{1 \leq k \leq n_p} P(t_k|c) \tag{4.1}$$

Die wahrscheinlichste Zelle c eines Photos p lässt sich mit folgender Formel schätzen:

$$c = \arg\max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_p} \log \hat{P}(t_k|c)] \tag{4.2}$$

Im Anhang (Abschnitt A.2) sind die einzelnen Elemente der obigen Gleichung 4.2 eingehender erklärt. Abbildung 4.7 zeigt ein fiktives Beispiel zur Bestimmung der wahrscheinlichsten Zelle eines Photos mit der NB-Methode.

Abb. 4.7: Vorhersage der Zelle für ein Photo mit den Tags x und y mit NB

4.5.2 Toponyme aus offiziellen Ortsverzeichnissen

Zum Vergleich der Resultate werden die Photos auch mit den offiziellen Ortsverzeichnissen von Ordnance Survey und Swisstopo georeferenziert (Abschnitt 3.3). Zur Identifikation von Toponymen in Tags – auch als *Geoparsing* bezeichnet – wird die von Hill (2006) vorgeschlagene Vorgehensweise angewendet:

- Identifizieren von Toponymen durch Nachschlagen der Tags in einem Ortsverzeichnis
- Eine erfolgreiche Suche liefert eine, teilweise auch mehrere mögliche Orte zurück
- Falls mehrere mögliche Orte vorhanden sind, wird der passendste Ort durch Disambiguierung gewählt
- Bei erfolgreicher Suche beziehungsweise Disambiguierung wird das Photo in die zum gewählten Toponym gehördende Zelle georeferenziert

Falls die Suche mit allen Tags genau einen möglichen Ort ergibt, werden dessen Koordinaten dem Photo zugewiesen. Falls für ein Photo mehrere mögliche Orte vorgeschlagen werden, muss der passendste Ort durch Disambiguierung bestimmt werden.

Dazu werden für jede mögliche Kombination von zwei Orten deren Distanz bestimmt. Für jene Ortskombination mit der kürzesten Distanz dazwischen wird nun jenes Toponym ausgewählt, welches aufgrund seiner im Ortsverzeichnis beschriebenen Objektart als höherwertig eingestuft werden kann. Dazu werden die verschiedenen Objektarten im Ortsverzeichnis in eine Hierarchie eingeteilt. Da das Ortsverzeichnis für Grossbritannien nicht die gleichen Typenklassen aufweist wie jener der Schweiz, ergeben sich unterschiedliche Rangierungen. Eine Auflistung der Hierarchien findet sich in Tabelle A.2 im Anhang.

4.6 Georeferenzierung mit ortsrelevanten Tags

Die Georeferenzierung mit ortsrelevanten Tags wird in diesem Abschnitt gezeigt. Bevor mit den beiden Methoden GEODIS (geometrische Disambiguierung) und TSCORE (normierte TF-IDF-Tagscores) Photos georeferenziert werden können, wird aus den Tags der Photos in den Trainingsdaten ein umgangssprachliches Ortsverzeichnis erstellt. Im Gegensatz zum NB-Klassifikator, verwenden die Methoden GEODIS und TSCORE nur ortsrelevante Tags zur Bestimmung der Position eines Photos.

4.6.1 Extraktion ortsrelevanter Tags

Für die Identifizierung von ortsrelevanten Tags wird die räumliche TF-IDF-Methode von Rattenbury & Naaman (2009) verwendet. Allerdings werden die räumlichen Analyseeinheiten nicht mittels Clustering, sondern analog zu Serdyukov et al. (2009) und Wing & Baldridge (2011) durch ein regelmässiges Gitternetz bestehend aus quadratischen Zellen definiert (Abschnitt 4.4). Diese sind einfach zu implementieren, können mit verschiedenen Zellgrössen erstellt werden und bieten einen potentiell flächendeckenden Ansatz.

TF-IDF (`tf`: *term frequency*, `idf`: *inverse document frequency*) ist ursprünglich eine Methode aus dem IR zur Beurteilung der Relevanz von Begriffen in Dokumenten einer Kollektion (Salton & Buckley, 1988). Ähnlich dem NB-Klassifikator (Abschnitt 4.5.1) kann auch die TF-IDF-Methode zur Bewertung von räumlicher Relevanz aus georeferenzierten Flickr-Photos verwendet werden. Je häufiger ein Tag in einer Zelle vorkommt, desto grösser ist seine Relevanz zur Beschreibung dieser Zelle, was mit dem `tf`-Faktor repräsentiert wird. Mit `tf` alleine ist unter Umständen keine genügende Unterscheidung einer Zelle in einem Gitternetz möglich – beispielsweise bei häufigem Auftreten eines Tags in vielen Zellen gleichzeitig. Deshalb wird mit `idf` ein zusätzlicher Faktor verwendet, der Tags priorisiert, die in wenigen Zellen eines Gitternetzes konzentriert sind. Die Relevanz kann mit dem Produkt aus `tf` und `idf` gemessen werden (Salton & Buckley, 1988).

Somit weist die TF-IDF-Methode Tags mit hoher Frequenz an einem Ort bei gleichzeitiger räumlicher Konzentration an wenigen Orten hohe Werte zu, während Tags mit diffuser räumlicher Verteilung oder geringer Frequenz tiefe Werte zugewiesen bekommen. Im Idealfall werden so Tags, die spezifische Orte repräsentieren, durch ihren hohen Wert hervorgehoben.

Die Vorkommenshäufigkeit `tf` eines Tags t in einer Zelle c entspricht der Anzahl Photos in $\mathbb{P}$ in c mit t, also wieviele Photos mit dem Tag t in einer Zelle gezählt werden:

$$tf(c,t) \triangleq |\mathbb{P}_{c,t}| \tag{4.3}$$

Die inverse Dokumenthäufigkeit `idf` eines Tags t entspricht dem Verhältnis der Anzahl Photos in $\mathbb{P}$ in einem Raster zur Anzahl Photos in $\mathbb{P}_t$ mit Tag t.

$$idf(t) \triangleq |\mathbb{P}| \,/\, |\mathbb{P}_t| \tag{4.4}$$

Obwohl Bulk Uploads herausgefiltert wurden, kann ein einzelner Nutzer durch lokal häufige Verwendung eines Tags Verzerrungen herbeiführen. Deshalb erweitern Rattenbury & Naaman (2009) die TF-IDF um eine zusätzliche Nutzer-Variable,

welche annimmt, dass ein Tags umso relevanter ist, je mehr Nutzer es in einer Zelle verwenden. Dieser Nutzerfaktor `uf` *(user frequency)* ist der relative Anteil aller Nutzer in $\mathbb{U}$ in einer Zelle c, die das Tag t benutzen.

$$uf(c,t) \triangleq |\mathbb{U}_{c,t}| \,/\, |\mathbb{U}_c| \tag{4.5}$$

Der Wert `ts` *(Tagscore)*, den ein Tag t in einer Rasterzelle c schliesslich durch die Methode zugewiesen bekommt, ist das Produkt aller drei Faktoren in den Gleichungen 4.3, 4.4 und 4.5:

$$ts(c,t) = tf(c,t) \cdot idf(t) \cdot uf(c,t) \tag{4.6}$$

Abbildung 4.8 zeigt ein fiktives Beispiel zur Berechnung der TF-IDF-Werte für ein Tag x mit der Methode gemäss Rattenbury & Naaman (2009).

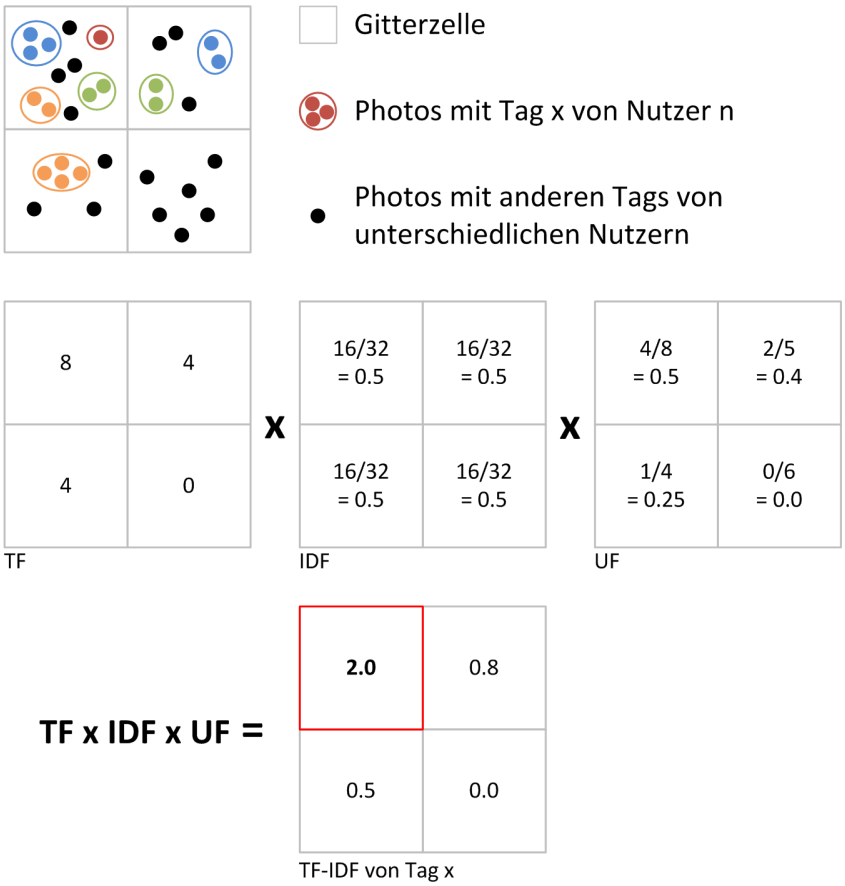


Abb. 4.8: Berechnung der TF-IDF-Werte für ein Tag x

Nach der Berechnung der Tagscores aller Tags werden pro Zelle jene 30 % mit den höchsten `ts`-Werten als *ortsrelevante Tags* isoliert. Die extrahierten ortsrelevanten Tags werden zusammen mit ihrer Tagscore und den geographischen Koordinaten des Mittelpunkts der entsprechenden Rasterzelle in einem umgangssprachlichen Ortsverzeichnis gespeichert.

4.6.2 GEODIS – Geometrische Disambiguierung

Smith & Crane (2001) wenden eine einfache heuristische Methode zur Auflösung von mehrdeutigen Toponymen in Dokumenten einer digitalen historischen Datenbank an. Dabei wird für alle möglichen Orte eines Dokuments der geometrische Schwerpunkt (Zentroid) berechnet und Orte ausgeschlossen, welche weiter als eine definierte Distanz vom Zentroiden entfernt liegen. Für die verbleibenden möglichen Orte wird ein neuer Zentroid und die zugehörigen Distanzen berechnet. Danach werden die Toponyme anhand ihrer Entfernung zu anderen Toponymen, der Entfernung zum Zentroiden sowie ihrer Bedeutung bewertet und das passendste zugeordnet.

Diese Methode wird fast unverändert zur Bestimmung der Position von den zu georeferenzierenden Photos in einem Gitternetz angewendet:

- Mit dem in Abschnitt 4.6.1 erstellten umgangssprachlichen Ortsverzeichnis werden anhand der Tags eines Photos alle potentiellen Gitterzellen identifziert.
- Nach der Berechnung des Zentroidpunktes und dem Ausschluss von Zellen mit einer Entfernung grösser als die doppelte Standardabweichung aller Distanzen vom Zentroidpunkt entfernt, wird für die verbleibenden Zellen ein neuer Zentroid berechnet.
- Zur Bestimmung der passendsten Zelle werden alle verbleibenden Zellen mit einem Gewicht g, berechnet aus ihrer Distanz zum Zentroiden, ihrer Tagscore-Summe und der Anzahl mit dem Photo übereinstimmenden Tags, bewertet.

Die drei Faktoren berechnen sich dabei folgendermassen:

- Der Distanzfaktor f_{Distanz} ist das Verhältnis zwischen der Standarddistanz aller Zellen zum Zentroiden und der aktuellen Distanz zum Zentroiden.
- Für alle in einer Gitterzelle vorkommenden Tags eines Photos werden die normalisierten Tagscores zum Faktor $f_{\text{Tagscores}}$ aufsummiert. Die Tagscore eines Tags t in einer Zelle c wird mit den im umgangssprachlichen Ortsverzeichnis vorkommenden minimalen und maximalen Tagscores von t auf Werte im Bereich $[0 \ldots 1]$ normalisiert (siehe Gleichung 4.9).
- Die Anzahl übereinstimmender Tags zwischen Photo und Gitterzelle fliesst mit dem Faktor f_{Tags} ein.

Für eine Zelle c berechnet sich das Gewicht g aus dem Produkt der drei oben genannten Faktoren:

$$g(c) = f_{\text{Distanz}} \cdot f_{\text{Tagscores}} \cdot f_{\text{Tags}} \tag{4.7}$$

Die Koordinaten des Mittelpunktes der Zelle mit dem höchsten Gewicht $g(c)$ werden dem Photo zugewiesen. Abbildung 4.9 zeigt beispielhaft die Vorhersage der Zelle für ein Photo mit der Methode GEODIS.

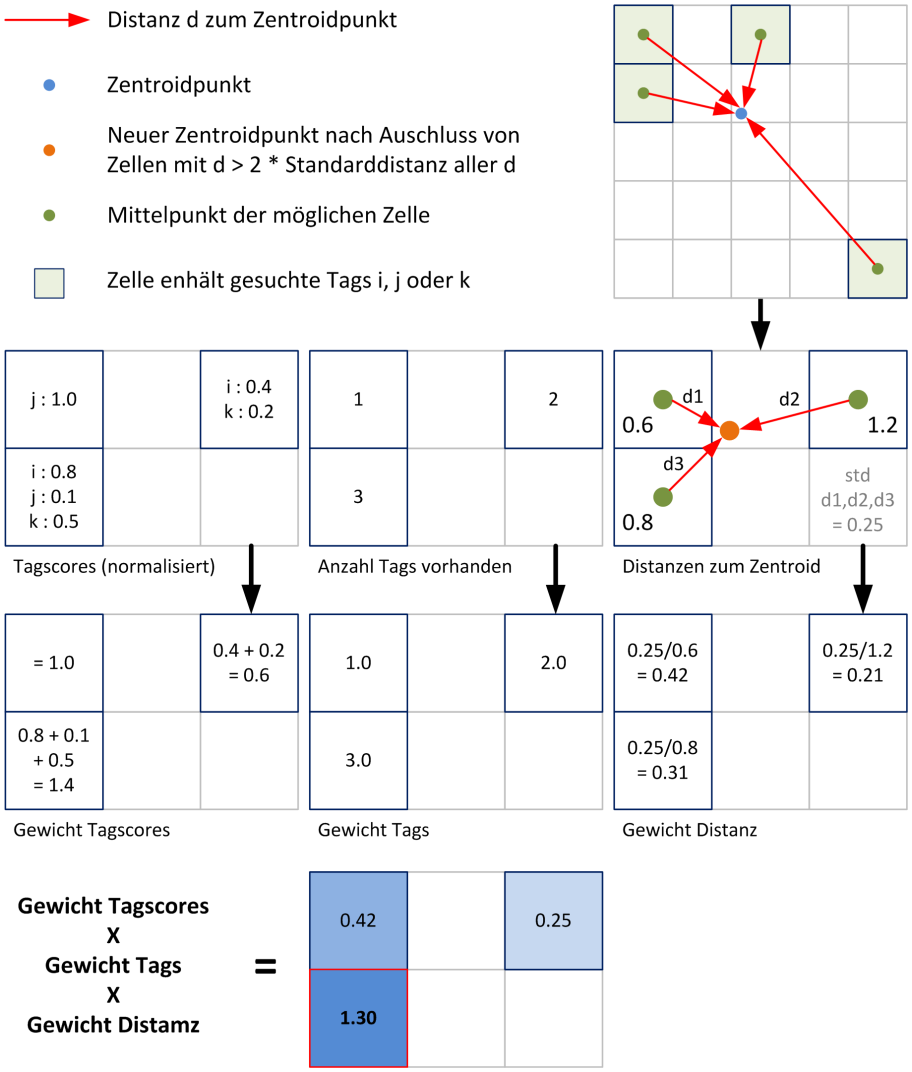


Abb. 4.9: Vorhersage der Zelle für ein Photo mit den Tags i, j und k mit GEODIS

4.6.3 TSCORE – normalisierte TF-IDF-Summen

Im Gegensatz zur vorherigen Variante mit geometrischer Disambiguierung wird mit der TSCORE-Methode der Standort eines Photos in einem Gitternetz aufgrund der pro Zelle aufsummierten Tagscores der ortsrelevanten Tags aus dem umgangssprachlichen Ortsverzeichnis georeferenziert.

Da die Tagscore-Werte der ortsrelevanten Tags eine Aussage über die lokale Relevanz der Tags machen (vgl. Abschnitt 4.6.1), kann damit anhand der Tags eines Photos der mögliche Standort in einem Gitternetz berechnet werden. Die Zelle eines Photos wird in den folgenden Schritten bestimmt:

- Aus dem umgangssprachlichen Ortsverzeichnis wird für alle Tags des Photos jenes Tag t_{max} bestimmt, welches über das gesamte Gitternetz betrachtet die grösste Tagscore in einer Gitterzelle aufweist.
- Die Suche des Standorts wird auf die Zellen $\mathbb{C}$, die t_{max} enthalten eingeschränkt.
- Für die verbleibenden Tags t_n werden anschliessend für jede Gitterzelle c, in der ein Tag t_i vorkommt, die Summe $s(c)$ der normalisierten Tagscores ts gebildet (Gleichung 4.8).

$$s(c) = \sum \text{norm}(ts(c, t_i)) \text{ wobei } t_i \in t_n, t_i \neq t_{\max}, c \in \mathbb{C} \tag{4.8}$$

Wie bereits bei der GEODIS-Methode, werden alle Tagscores mithilfe der der minimalen und maximalen vorkommenden Tagscore eines Tags auf einen Wertebereich von $[0 \ldots 1]$ normalisiert. Die Funktion *norm* normalisiert die Tagscore $ts(t_i, c)$ eines Tags t_i in einer Zelle c wie folgt:

$$norm(ts(t_i, c)) = \frac{ts(t_i, c) - ts_{min}(t)}{ts_{max}(t) - ts_{min}(t)} \tag{4.9}$$

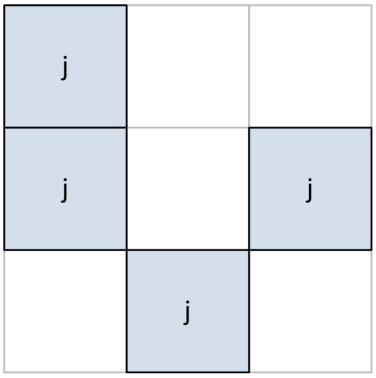
Dem Photo werden die Koordinaten des Mittelpunkts der Zelle mit der höchsten Summe $s(c)_{t_{\max} \in c}$ zugewiesen (vgl. Abbildung 4.10).

Es hat sich gezeigt, dass der Einbezug der „Nachbarschaft“ des relevantesten Tags t_{max} zu besseren Resultaten führt, weil Tags, die räumlich konzentriert von vielen Leuten oft benutzt werden, sich teilweise sehr stark auf einzelne Zellen konzentrieren und so überproportionalen Einfluss haben. Beobachtet werden kann dieser Effekt zum Beispiel in London, an Orten mit hoher touristischer Attraktivität. Dieser Schritt wird übersprungen, falls für die Tags eines Photos nur ein einziges im umgangssprachlichen Ortsverzeichnis vorhanden ist.

Suche Tag mit der global grössten Tagscore

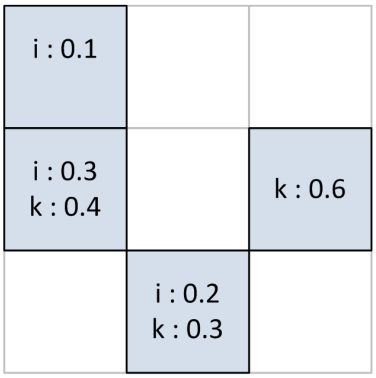
Maximum Tagscore(Tag i) = 10
Maximum Tagscore(Tag j) = 15
Maximum Tagscore(Tag k) = 8

Einschränkung Suche auf Zellen mit Tag j



Alle Zellen mit Tag j

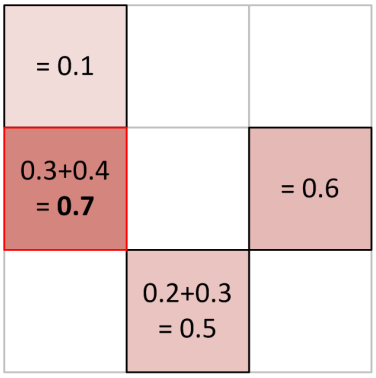
Durchsuche Zellen mit Tag j nach den übrigen Tags i und k



Normalisierte Tagscores von i und k in den Zellen mit Tag j

Summiere normalisierte Tagscores für die übrigen Tags i und k in Zellen mit Tag j

Mittelpunkt der Zelle mit der grössten Summe der Tagscores wird als Standort zugewiesen.



Summe der normalisierten Tagscores von i und k

Abb. 4.10: Vorhersage der Zelle für ein Photo mit den Tags i, j und k mit TSCORE

4.7 Einfluss der Datenfilterung

Bei der Verwendung von UGC-Daten können einzelne Nutzer potentielle Verzerrungen in den Ergebnissen bewirken (vgl. Abschnitt 2.4.3). Zur Untersuchung des Einflusses der Datenfilterung auf die Ergebnisse der Georeferenzierung werden verschiedene, unterschiedlich gefilterte Datensätze verwendet.

Für den Datensatz von Grossbritannien werden die folgenden Filtervarianten mit den Methoden GEODIS und TSCORE georeferenziert:

F0 alle Filterschritte wie in Abschnitt 4.2 ausgeführt
F1 keine Filterung ausser der Entfernung von Duplikaten und Photos ohne Tags
F2 alle Filterschritte ausser der Entfernung von Bulk Uploads, Tagfilter und Nutzerfilter
F3 alle Filterschritte ausser der Entfernung von Stoppwörtern

4.8 Eigenschaften von Photos mit und ohne Geotag

Die in diesem Kapitel beschriebenen Methoden unterliegen alle der Annahme, dass die Flickr-Photos im Trainingsdatensatz zur Extraktion von ortsrelevanten Tags oder zum Training des NB-Klassifikators aus einer ähnlichen Verteilung stammen wie die Photos aus dem Validationsdatensatz. Für die georeferenzierten Photos in den Trainings- und Validationsdaten trifft dies zu, da sie aus dem gleichen Datensatz stammen.

Wenn der Standort von Flickr-Photos ohne Geotag mit ortsrelevanten Tags bestimmt werden soll, die aus Flickr-Photos mit Geotag extrahiert wurden, sollte nachgewiesen werden, dass sich die Charakteristiken der beiden Datensätze ähnlich genug sind.

Falls die Photos einer ähnlichen Verteilung folgen, ist es zumindest potentiell möglich, dass Photos ohne Geotags mit ähnlicher Genauigkeit georeferenziert werden können, wie die Photos in den Validationsdaten mit Geotag.

Dazu wird die Charakteristik der statistischen Kennwerte zwischen zwei Datensätzen von Flickr-Photos mit und ohne Geotag analysiert. Diese umfassen Photos aus Grossbritannien aus dem Jahr 2011. Zudem wird die Verwendung von Toponymen als Tags näher untersucht.

Die Photos mit Geotag für das Jahr 2011 sind eine Untermenge aus dem gesamten Datensatzes von georeferenzierten Photos aus Grossbritannien. Da sich nach Photos ohne Geotag nicht räumlich eingeschränkt suchen lässt, wurden aus den häufigsten Tags der Photos mit Geotag möglichst generische Suchwörter (vgl. Abschnitt 4.1) mit Bezug zu Grossbritannien gesucht. Diese wurden dann als Suchbegriffe für Photos ohne Geotag verwendet.

Vor der Berechnung der statistischen Kennwerte werden die für beide Datensätze noch die folgenden Filter angewendet (vgl. Abschnitt 4.2):

- Filterung von Duplikaten und Photos ohne Tags
- Filterung von Bulk Uploads
- Entfernen von Stoppwörtern und Photographie-Fachbegriffen

5 Resultate

In diesem Kapitel werden statistische Kennwerte der aufbereiteten Datensätze und die Resultate der verschiedenen Methoden zur Georeferenzierung aus Kapitel 4 gezeigt. Es werden der Einfluss der Filterung von Nutzerverzerrungen analysiert und auf die Unterschiede zwischen Photos mit und ohne Geotag eingegangen.

5.1 Statistische Kennwerte der Datensätze

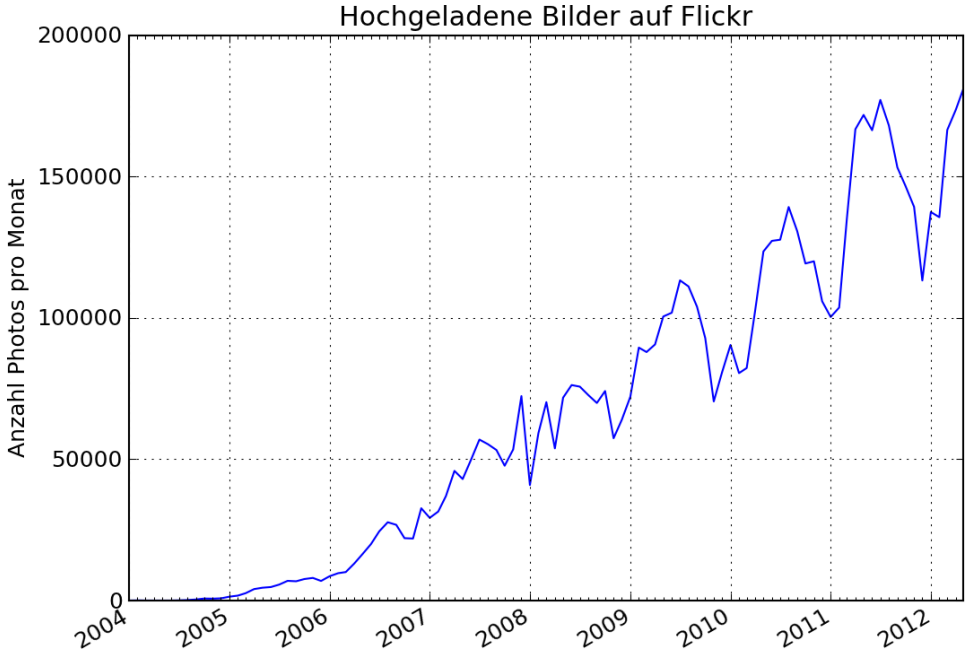
Die beiden Datensätze mit Metadaten von *Flickr*-Photos aus Grossbritannien und der Schweiz weisen, nach der Entfernung von Duplikaten, die in Tabelle 5.1 gezeigten statistischen Kennwerte auf. Für Grossbritannien wurden über 6 Mio., für die Schweiz über 800'000 georeferenzierte Photos extrahiert. Auffallend ist die fast doppelt so grosse mittlere Anzahl Bilder pro Nutzer in Grossbritannien. Sonsten unterscheiden sich die beiden Datensätze nur unwesentlich.

Tab. 5.1: Statistische Kennwerte der extrahierten Flickr-Daten

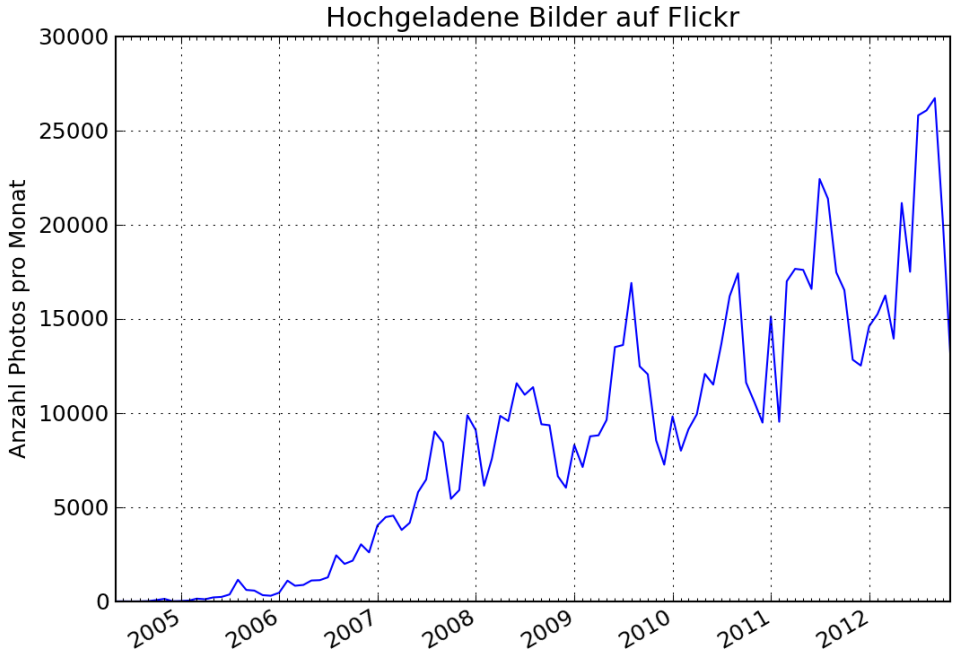
	Grossbritannien	Schweiz
Total Photos	6'663'046	876'182
Photos mit Tags	81.7 %	77.7 %
Photos ohne Tags	18.3 %	22.3 %
Tags pro Photo (Mittelwert)	**8.5**	**8.7**
Total Nutzer	92'662	18'395
Total eindeutige Tags	1'484'355	228'957
Photos pro Nutzer (Mittelwert)	**58.7**	**37**
Photos pro Nutzer (Medianwert)	5	4
Photos pro Nutzer (Standardabweichung)	**501.2**	**243.4**

5.2 Zeitliche und räumliche Verteilung der Datensätze

Die zeitliche Verteilung in Abbildung 5.1 zeigt dass in beiden Ländern ein zunehmender Trend beim Upload von Photos mit Geotag zu erkennen ist. Als Gründe sind neben der wachsenden Popularität von Flickr (Abschnitt 3.2) auch die steigende Verbreitung von GPS-tauglichen Digitalkameras und Smartphones anzuführen. Es lassen sich zudem starke saisonale Schwankungen feststellen.



(a) Grossbritannien (1.1.2004 – 31.5.2012)



(b) Schweiz (1.1.2004 – 24.11.2012)

Abb. 5.1: Monatlich hochgeladene Photos mit Geotag

Wie schon in Purves et al. (2011) gezeigt, sind die georeferenzierten Photos räumlich heterogen verteilt. Sowohl in Grossbritannien als auch in der Schweiz befindet sich der grösste Anteil der Photos in urbanen Regionen und Städten, während sich in ländlichen Regionen nur wenige oder gar keine Photos befinden (Abbildung 5.2 und Abbildung 5.3). In beiden Ländern lassen sich die metropolitanen Regionen gut erkennen. In der Schweiz sind auch touristisch populäre Destinationen wie zum Beispiel Zermatt oder das Berner Oberland durch eine hohe Photodichte deutlich erkennbar.

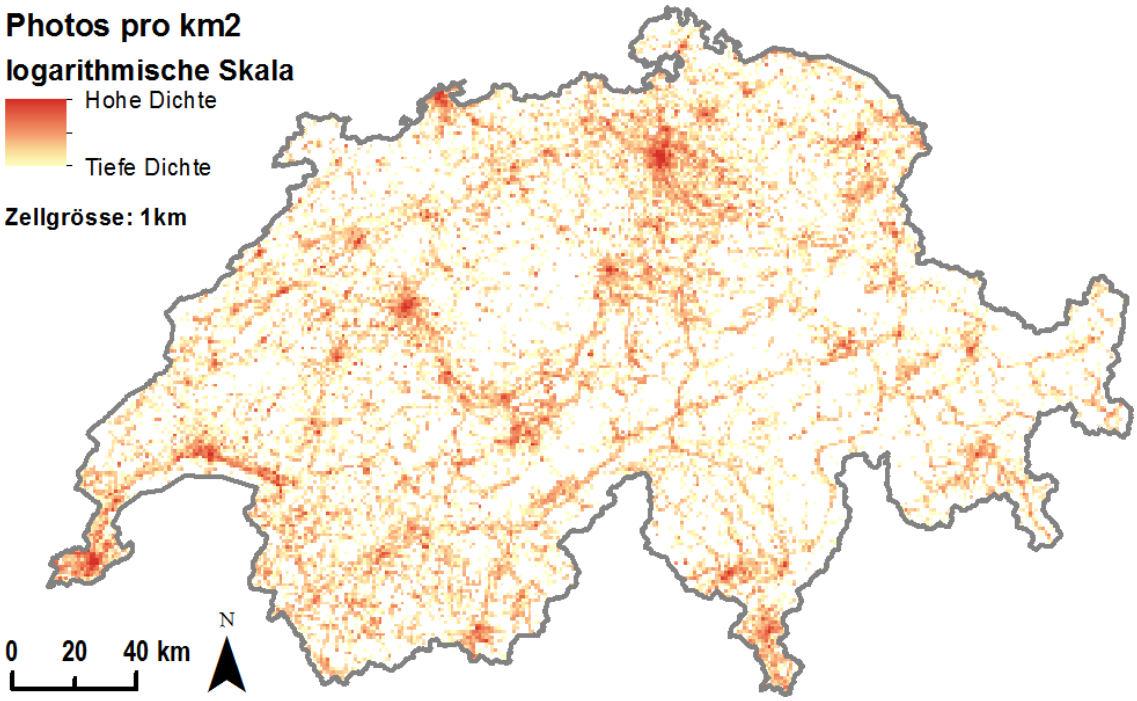


Abb. 5.2: Räumliche Verteilung der georeferenzierten Photos in der Schweiz
Daten: Flickr, Swisstopo

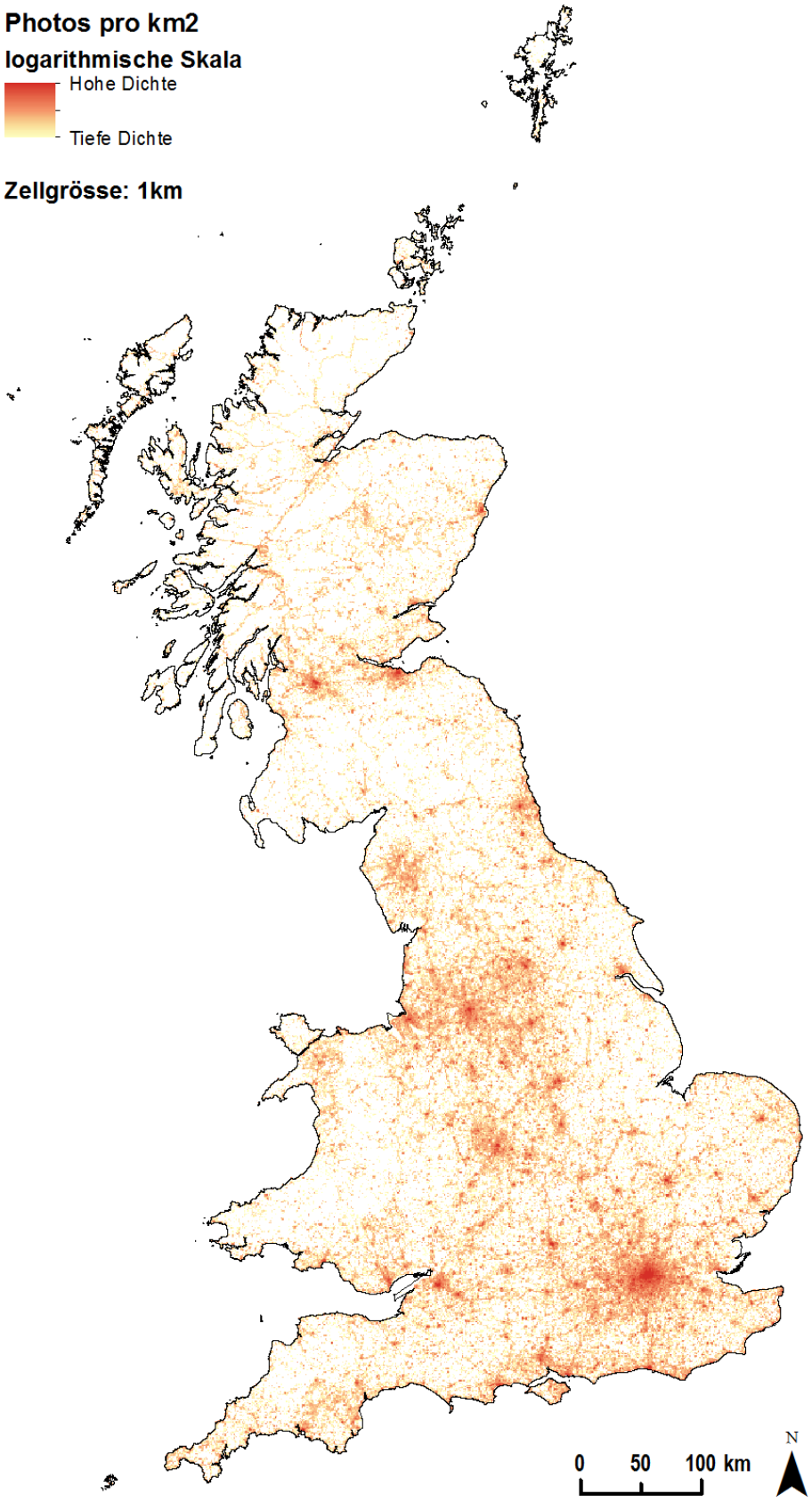


Abb. 5.3: Räumliche Verteilung der georeferenzierten Photos in Grossbritannien
Daten: Flickr, Ordnance Survey

5.3 Filterung und Aufteilung der Daten

Die statistischen Kennwerte der beiden Datensätze sind in Tabelle 5.2 für jeden Filterschritt einzeln aufgeführt. Die Filterschritte wurden, wie in Abschnitt 4.2 beschrieben, für beide Datensätze angewendet.

Tab. 5.2: Statistische Kennwerte der Datensätze nach jeweiligem Filterschritt

Grossbritannien

Filterschritt	P	PT	TP	N	T	PN1	PN2	PN3
Duplikate	6'663'046	5'441'406	8.5	92'662	1'484'355	58.7	5	501.2
ohne Tags	5'441'545	5'441'406	8.5	92'662	1'484'355	58.7	5	501.2
Genauigkeit Geotag < 16	5'441'213	5'441'213	8.5	92'658	1'484'320	58.7	5	501.2
Datumsfehler	5'382'741	5'382'741	8.5	90'855	1'467'222	59.3	5	505.2
Landesgrenzen	5'059'778	5'059'778	8.5	84'214	1'397'564	60.1	5	516.2
Stoppwörter	4'989'792	4'989'792	7.9	83'373	1'029'164	59.9	5	516.3
Normalisierung	**4'989'792**	4'989'778	**7.9**	83'373	985'779	**59.9**	5	516.3
Bulk Uploads	**1'620'253**	1'620'248	**9.6**	**74'369**	897'208	**21.8**	3	153.3
Nutzer mit < 2 oder > 10'000 Photos	1'544'410	1'544'405	9.7	**48'930**	855'001	**31.6**	6	149.6
Tags mit Häufigkeit < 10	1'529'504	1'529'504	8.8	48'789	94'620	31.4	6	149.4
Anzahl Nutzer pro Tag < 2	**1'529'504**	1'529'502	8.8	48'789	94'620	31.4	6	149.4

Schweiz

Filterschritt	P	PT	TP	N	T	PN1	PN2	PN3
Duplikate	876'182	680'612	8.7	18'395	228'957	37.0	4	243.4
ohne Tags	680'618	680'612	8.7	18'395	228'957	37.0	4	243.4
Genauigkeit Geotag < 16	680'577	680'577	8.7	18'395	228'947	37.0	4	243.4
Datumsfehler	635'421	635'421	8.6	17'169	212'289	37.0	4	245.0
Landesgrenzen	514'652	514'652	8.5	13'193	172'583	39.0	4	256.8
Stoppwörter	509'936	509'936	7.9	13'029	135'098	39.1	4	257.9
Normalisierung	**509'936**	509'933	**7.9**	13'027	129'446	**39.1**	4	257.9
Bulk Uploads	**119'670**	119'667	**11.2**	**11'108**	116'038	**10.8**	2	47.5
Nutzer mit < 2 oder > 1'000 Photos	108'853	108'852	11.4	**6'533**	109'429	**16.7**	5	50.5
Tags mit Häufigkeit < 10	106'617	106'617	9.8	6'472	12'884	16.5	4	50.4
Anzahl Nutzer pro Tag < 2	**106'617**	106'617	9.8	6'472	12'884	16.5	4	50.4

P – Anzahl Photos total
PT – Anzahl Photos mit Tags
TP – Anzahl Tags pro Photo (Mittelwert)
N – Anzahl eindeutige Nutzer
T – Anzahl eindeutige Tags
PN – Anzahl Photos pro Nutzer (PN1 : Mittelwert, PN2 : Medianwert, PN3 : Standardabweichung)

Bei beiden Datensätzen fällt die grosse Zahl an Bulk Uploads auf, die entfernt werden. In Grossbritannien sind über zwei Drittel aller verbleibenden Photos Bulk Uploads, in der Schweiz sogar drei Viertel. Offensichtlich haben Photos als Teil von Bulk Uploads deutlich weniger Tags als andere Photos, wie die durchschnittliche Anzahl Tags pro Photo (TP) zeigt.

Viele Nutzer sind nur mit einem einzigen Photo in der Stichprobe vertreten. Diese werden zusammen mit den überdurchschnittlich beitragenden Nutzern entfernt und machen in beiden Datensätzen rund einen Drittel aller Nutzer aus. Die Entfernung

der Nutzer mit nur einem Photo bewirkt eine steigende mittlere Anzahl Photos pro Nutzer (PN1).

Nach der Anwendung aller Filterschritte bleiben im Datensatz von Grossbritannien rund 1.5 Millionen, in der Schweiz etwas mehr als 100'000 Photos für die weitere Analyse übrig.

Anschliessend werden die Daten gemäss Abschnitt 4.3 in Trainings- und Validationsdaten aufgeteilt. Falls nicht explizit angegeben, basieren die folgenden Resultate immer auf einem Trainings-Validations-Verhältnis von 75 zu 25 und zufälliger Aufteilung pro Photo (RP) (vgl. Abschnitt 5.8).

5.4 Georeferenzierung mit den Referenzmethoden

Zum Vergleich der Ergebnisse der beiden Methoden GEODIS und TSCORE werden die Photos in den Validationsdaten mit den beiden Methoden TOPO und NB georeferenziert.

5.4.1 NB – Statistisches Sprachmodell mit maschinellem Lernen

In den Tabellen 5.3 und 5.4 sind die statistischen Kennwerte und die Vorhersagegenauigkeit der Georeferenzierung mit einem NB-Klassifikator (Abschnitt 4.5.1) aufgeführt. Die Diagramme in Abbildung 5.4 zeigen die sortierte Fehlerdistanz pro Photo für verschiedene Zellgrössen mit der Methode NB. Da diese Methode äusserst zeitintensiv ist, wurden die Ergebnisse für verschiedene Zellgrössen mit einer Stichprobe von 1'000 zufälligen Photos aus dem Validationsdatensatz durchgeführt.

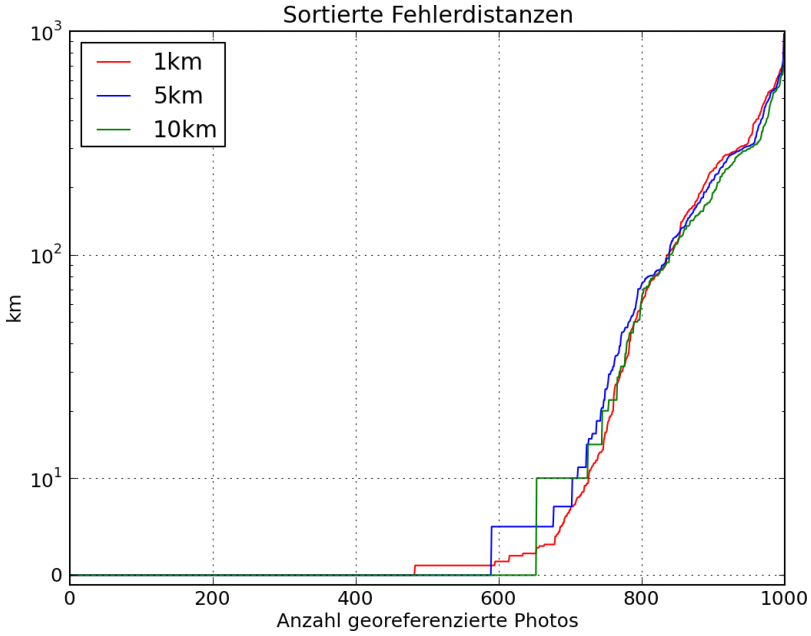
Sowohl in Grossbritannien als auch in der Schweiz schätzt der NB-Klassifikator für eine grossen Teil der Photos die korrekte oder eine benachbarte Zelle.

Tab. 5.3: NB – Resultate in Grossbritannien

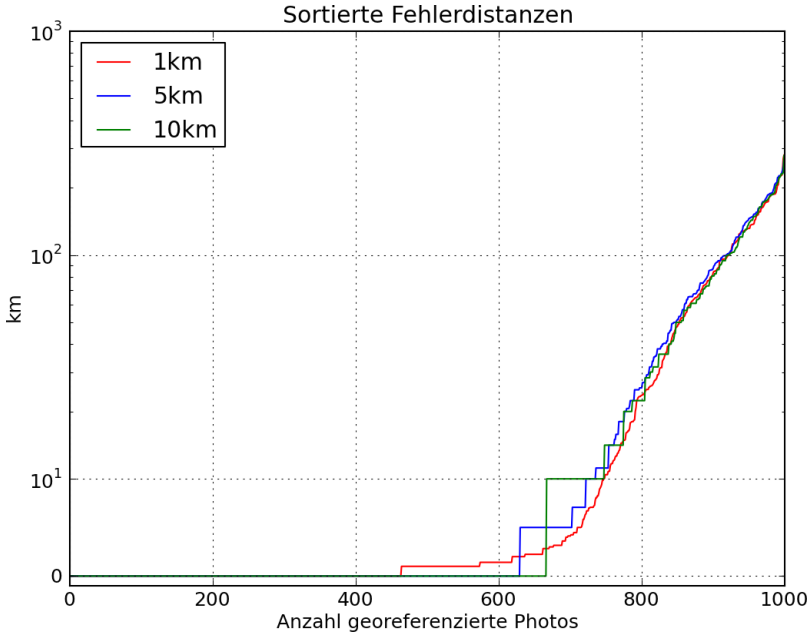
		Einheit	1km	5km	10km
Validationsdaten	Total Photos	Anz.	**1000**	**1000**	**1000**
	Georeferenzierte Photos	Anz.	1000	1000	1000
		%	100.0	100.0	100.0
Fehlerdistanz	Mittelwert	km	55.4	53.7	48.7
	Medianwert	km	**1.0**	**0.0**	**0.0**
	Standardabweichung	km	129.8	123.0	112.1
Vorhersage in	korrekter Zelle (n0)	%	**48.3**	**59.0**	**65.3**
	direkter Nachbarzelle (n1)	%	13.2	11.3	9.2
	Nachbarzelle 2. Grades (n2)	%	4.2	2.3	2.4
	Nachbarzelle 3. Grades (n3)	%	2.4	1.7	1.1
	n0, n1, n2, n3 (Total)	%	**68.1**	**74.3**	**78.0**

Tab. 5.4: NB – Resultate in der Schweiz

		Einheit	1km	5km	10km
Validationsdaten	Total Photos	Anz.	**1000**	**1000**	**1000**
	Georeferenzierte Photos	Anz.	1000	1000	1000
		%	100.0	100.0	100.0
Fehlerdistanz	Mittelwert	km	20.6	21.9	20.8
	Medianwert	km	**1.0**	**0.0**	**0.0**
	Standardabweichung	km	45.6	47.4	45.6
Vorhersage in	korrekter Zelle (n0)	%	**46.4**	**63.0**	**66.7**
	direkter Nachbarzelle (n1)	%	15.5	9.2	10.8
	Nachbarzelle 2. Grades (n2)	%	5.2	4.0	3.7
	Nachbarzelle 3. Grades (n3)	%	2.7	1.5	2.7
	n0, n1, n2, n3 (Total)	%	**69.8**	**77.7**	**83.9**



(a) Grossbritannien



(b) Schweiz

Abb. 5.4: NB – Fehlerdistanzen bei verschiedenen Zellgrössen

5.4.2 TOPO – Toponyme aus offiziellen Ortsverzeichnissen

Die Tabellen 5.5 und 5.6 zeigen die statistischen Kennwerte und die Vorhersagegenauigkeit der Georeferenzierung mit TOPO (Abschnitt 4.5.2). Abbildung 5.5 zeigt die Diagramme mit der sortierten Fehlerdistanz pro Photo für verschiedene Zellgrössen mit der Methode TOPO.

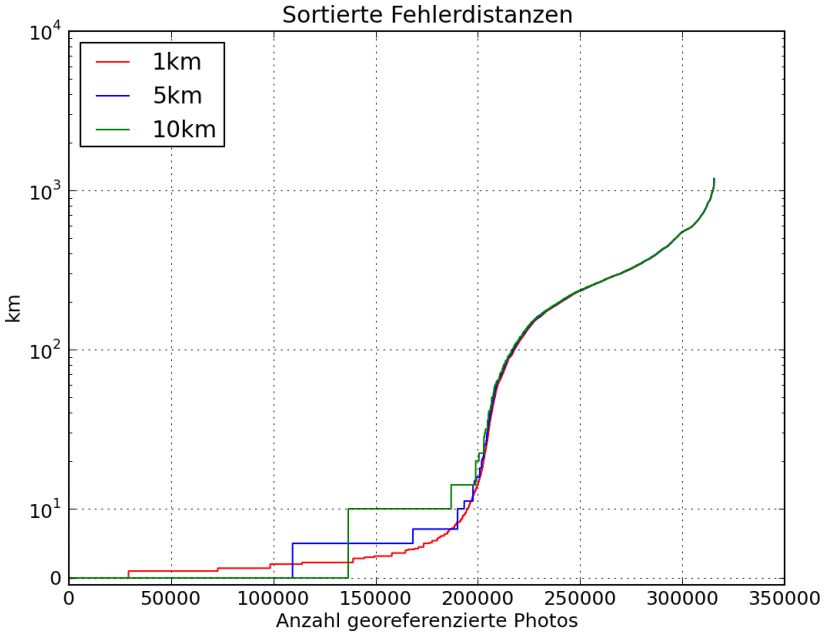
Mit dem Ortsverzeichnis der Schweiz lassen sich die Zellen der Photos mit Toponymen in den Tags besser vorhersagen als in Grossbritannien. Da aber nicht alle Photos ein Toponym in den Tags enthalten, können in der Schweiz nur rund drei Viertel, in Grossbritannien vier Fünftel der Bilder positioniert werden.

Tab. 5.5: TOPO – Resultate in Grossbritannien

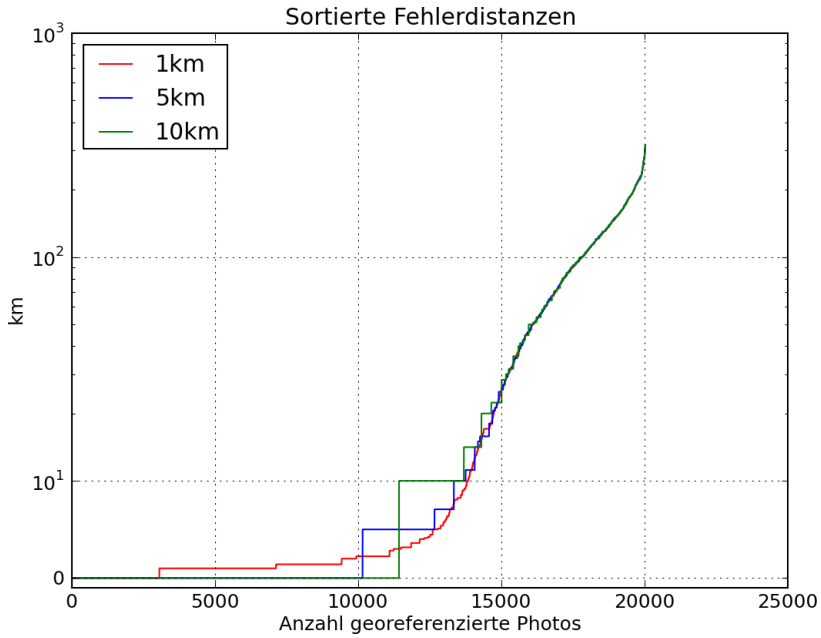
		Einheit	1km	5km	10km
Validationsdaten	Total Photos	Anz.	382625	382625	382625
	Georeferenzierte Photos	Anz.	315517	315517	315517
		%	**82.5**	**82.5**	**82.5**
Fehlerdistanz	Mittelwert	km	110.7	111.5	112.5
	Medianwert	km	**3.2**	**5.0**	**10.0**
	Standardabweichung	km	188.8	188.9	189.2
Vorhersage in	korrekter Zelle (n0)	%	**7.6**	**28.6**	**35.7**
	direkter Nachbarzelle (n1)	%	18.1	21.1	16.3
	Nachbarzelle 2. Grades (n2)	%	12.0	2.2	1.1
	Nachbarzelle 3. Grades (n3)	%	5.8	1.0	0.5
	n0, n1, n2, n3 (Total)	%	**43.5**	**52.8**	**53.7**

Tab. 5.6: TOPO – Resultate in der Schweiz

Validationsdaten	Total Photos	Anz.	26800	26800	26800
	Georeferenzierte Photos	Anz.	20027	20027	20027
		%	**74.7**	**74.7**	**74.7**
Fehlerdistanz	Mittelwert	km	27.8	27.8	28.0
	Medianwert	km	**2.2**	**0.0**	**0.0**
	Standardabweichung	km	52.3	52.6	52.5
Vorhersage in	korrekter Zelle (n0)	%	**15.2**	**50.7**	**57.0**
	direkter Nachbarzelle (n1)	%	31.8	15.9	14.4
	Nachbarzelle 2. Grades (n2)	%	9.1	4.2	4.3
	Nachbarzelle 3. Grades (n3)	%	4.9	2.8	2.4
	n0, n1, n2, n3 (Total)	%	**61.0**	**73.6**	**78.1**



(a) Grossbritannien



(b) Schweiz

Abb. 5.5: TOPO – Fehlerdistanzen bei verschiedenen Zellgrössen

5.5 Georeferenzierung mit ortsrelevanten Tags

Mit den in Abschnitt 4.6.1 gezeigten ortsrelevanten Tags wird für jedes Photo im Validationsdatensatz die Position in einem Gitternetz geschätzt. Die Validierung der geschätzten Position erfolgt mit der Fehlerdistanz zwischen dem Mittelpunkt der geschätzen Zelle und dem Mittelpunkt der das Photo beinhaltenden Zelle (Abschnitt 4.4). Im folgenden werden die Resultate der beiden Methoden GEODIS und TSCORE für verschiedene Zellgrössen gezeigt.

5.5.1 GEODIS – Geometrische Disambiguierung

Die statistischen Kennwerte der Fehlerdistanzen und die Vorhersagegenauigkeit der Methode sind in den Tabellen 5.7 und 5.8 aufgeführt. Die sortierte Fehlerdistanz pro Photo für verschiedene Zellgrössen mit der Methode GEODIS zeigen die Diagramme in Abbildung 5.6.

Sowohl in Grossbritannien als auch in der Schweiz nimmt der Medianwert der Fehlerdistanz mit der Zellgrösse zu. Allerdings stimmt er in Grossbritannien mit der Zellgrösse überein, während die Photos der Schweiz mit der GEODIS-Methode grössere Fehlerdistanzen aufweisen.

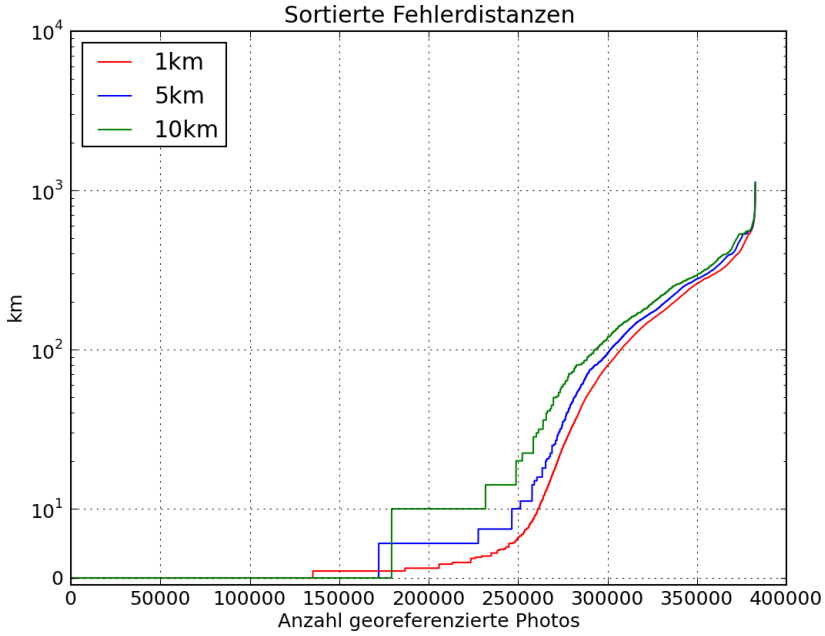
Ebenfalls besser als in der Schweiz ist in Grossbritannien für alle Zellgrössen die Vorhersage der korrekten Zelle und der Nachbarzellen.

Tab. 5.7: GEODIS – Resultate in Grossbritannien

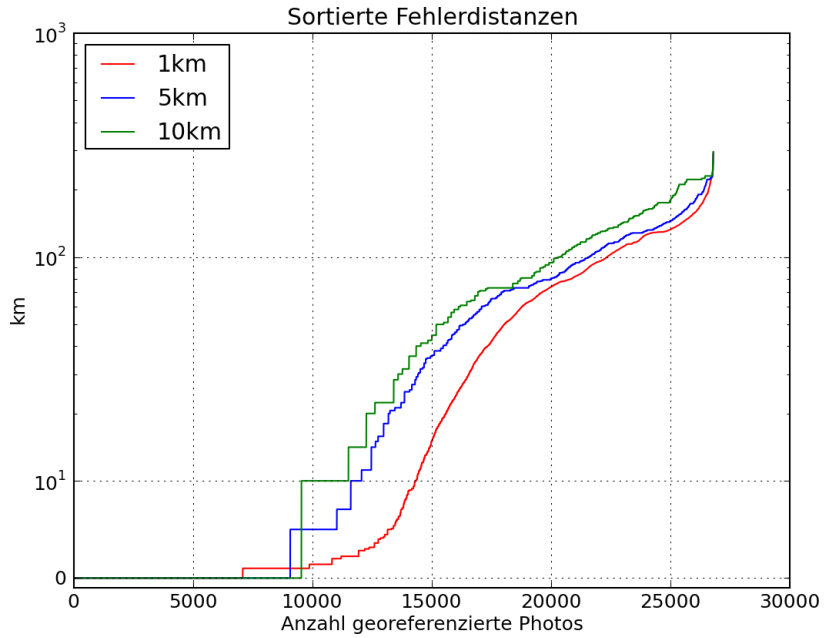
		Einheit	1km	5km	10km
Validationsdaten	Total Photos	Anz.	382625	382625	382625
	Georeferenzierte Photos	Anz.	382624	382623	382605
		%	100.0	100.0	100.0
Fehlerdistanz	Mittelwert	km	57.3	64.6	73.2
	Medianwert	km	**1.4**	**5.0**	**10.0**
	Standardabweichung	km	116.0	125.7	133.6
Vorhersage in	korrekter Zelle (n0)	%	**35.4**	**45.0**	**46.9**
	direkter Nachbarzelle (n1)	%	18.4	19.5	18.2
	Nachbarzelle 2. Grades (n2)	%	5.3	3.3	3.0
	Nachbarzelle 3. Grades (n3)	%	3.5	1.9	1.6
	n0, n1, n2, n3 (Total)	%	**62.6**	**69.6**	**69.6**

Tab. 5.8: GEODIS – Resultate in der Schweiz

		Einheit	1km	5km	10km
Validationsdaten	Total Photos	Anz.	26800	26800	26800
	Georeferenzierte Photos	Anz.	26799	26800	26799
		%	100.0	100.0	100.0
Fehlerdistanz	Mittelwert	km	39.2	47.6	57.1
	Medianwert	km	**5.4**	**20.6**	**22.4**
	Standardabweichung	km	52.4	56.5	67.2
Vorhersage in	korrekter Zelle (n0)	%	**26.4**	**33.8**	**35.6**
	direkter Nachbarzelle (n1)	%	14.0	9.5	10.2
	Nachbarzelle 2. Grades (n2)	%	5.1	3.9	5.0
	Nachbarzelle 3. Grades (n3)	%	2.3	3.0	3.4
	n0, n1, n2, n3 (Total)	%	**47.8**	**50.1**	**54.1**



(a) Grossbritannien



(b) Schweiz

Abb. 5.6: GEODIS – Fehlerdistanzen bei verschiedenen Zellgrössen

5.5.2 TSCORE – normalisierte TF-IDF-Summen

In den Tabellen 5.9 und 5.10 sind die statistischen Kennwerte und die Vorhersagegenauigkeit der Georeferenzierung mit TSCORE aufgeführt. In den Diagrammen in Abbildung 5.7 sind die sortierte Fehlerdistanz pro Photo für verschiedene Zellgrössen mit der Methode TSCORE abgebildet.

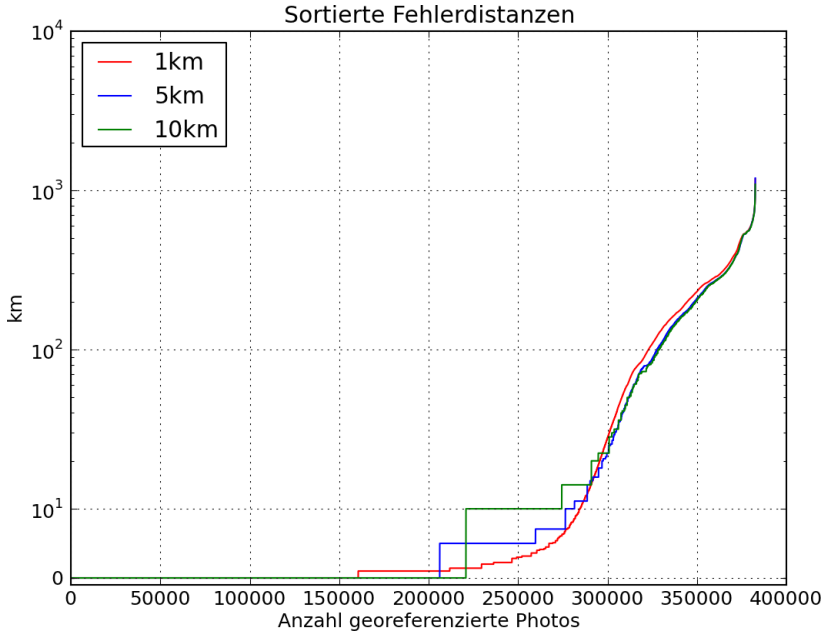
In beiden Ländern erreicht die TSCORE-Methode für alle Auflösungen einen Mediandistanzfehler von 1.0 km oder weniger. Da für die Zellgrössen 5 km und 10 km für mehr als die Hälfte der Photos im Validationsdatensatz die richtige Zelle vorhergesagt werden kann, beträgt der Mediandistanzfehler folglich 0.0 km.

Tab. 5.9: TSCORE – Resultate in Grossbritannien

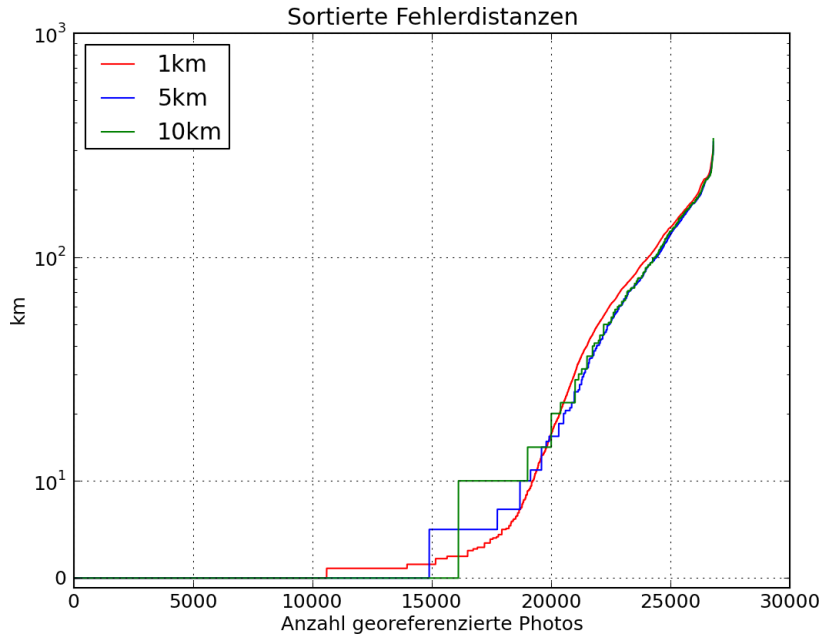
		Einheit	1km	5km	10km
Validationsdaten	Total Photos	Anz.	382625	382625	382625
	Georeferenzierte Photos	Anz.	382624	382623	382605
		%	100.0	100.0	100.0
Fehlerdistanz	Mittelwert	km	49.4	46.2	46.3
	Medianwert	km	**1.0**	**0.0**	**0.0**
	Standardabweichung	km	118.1	113.1	112.1
Vorhersage in	korrekter Zelle (n0)	%	**42.0**	**53.9**	**57.7**
	direkter Nachbarzelle (n1)	%	18.0	18.4	18.3
	Nachbarzelle 2. Grades (n2)	%	5.1	3.6	3.0
	Nachbarzelle 3. Grades (n3)	%	3.2	1.9	1.5
	n0, n1, n2, n3 (Total)	%	**68.3**	**77.8**	**80.5**

Tab. 5.10: TSCORE – Resultate in der Schweiz

		Einheit	1km	5km	10km
Validationsdaten	Total Photos	Anz.	26800	26800	26800
	Georeferenzierte Photos	Anz.	26799	26800	26799
		%	100.0	100.0	100.0
Fehlerdistanz	Mittelwert	km	25.8	23.8	24.5
	Medianwert	km	**1.0**	**0.0**	**0.0**
	Standardabweichung	km	52.6	49.3	49.9
Vorhersage in	korrekter Zelle (n0)	%	**39.5**	**55.6**	**60.1**
	direkter Nachbarzelle (n1)	%	17.0	14.2	14.5
	Nachbarzelle 2. Grades (n2)	%	5.9	4.1	4.3
	Nachbarzelle 3. Grades (n3)	%	2.9	3.0	2.4
	n0, n1, n2, n3 (Total)	%	**65.4**	**76.8**	**81.3**



(a) Grossbritannien



(b) Schweiz

Abb. 5.7: TSCORE – Fehlerdistanzen bei verschiedenen Zellgrössen

5.6 Vergleich der Methoden

Die Ergebnisse aller verwendeten Methoden zur Georeferenzierung der Bilder sind in den Tabellen 5.11 und 5.12 für die Zellgrösse 1 km dargestellt. Zur direkten Vergleichbarkeit der Resultate der Methoden GEODIS, TSCORE und TOPO wurden diese mit der zufälligen Stichprobe ($n = 1000$) der Methode NB wiederholt. Als zusätzliche Referenz sind die Kennwerte bei einer zufälligen Vorhersage (RAND) der Zellen angegeben.

Die Methoden TSCORE und NB weisen für beide untersuchten Gebiete den gleichen Mediandistanzfehler von 1.0 km auf. Somit sind mehr als die Hälfte aller Photos entweder in der korrekten oder in einer direkt angrenzenden Zelle geschätzt werden.

In Grossbritannien erreicht die GEODIS-Methode eine ähnliche Vorhersagegenauigkeit wie TSCORE und NB, in der Schweiz ist der Mediandistanzfehler dagegen deutlich höher, was sich auch in Abbildung 5.8 deutlich erkennen lässt.

Die TOPO-Methode liefert für beide Untersuchungsgebiete eine vergleichbare Vorhersage der korrekten oder benachbarten Zellen, in der Schweiz ist aber der Mediandistanzfehler kleiner.

Obowhl die TOPO-Methode im Vergleich mit den anderen Methoden in beiden Datensätzen für weniger Photos die korrekte Zelle vorhersagen, haben sie bei Betrachtung der Nachbarzellen eine bessere Vorhersagegenauigkeit. In der Schweiz ist der Mediandistanzfehler der TOPO-Methode deshalb tiefer als der GEODIS-Methode.

Die Kurve der zufälligen Georeferenzierung (RAND) bestätigt, dass alle verwendeten Methoden deutlich besser sind, als als die Bestimmung der Position durch Zufall.

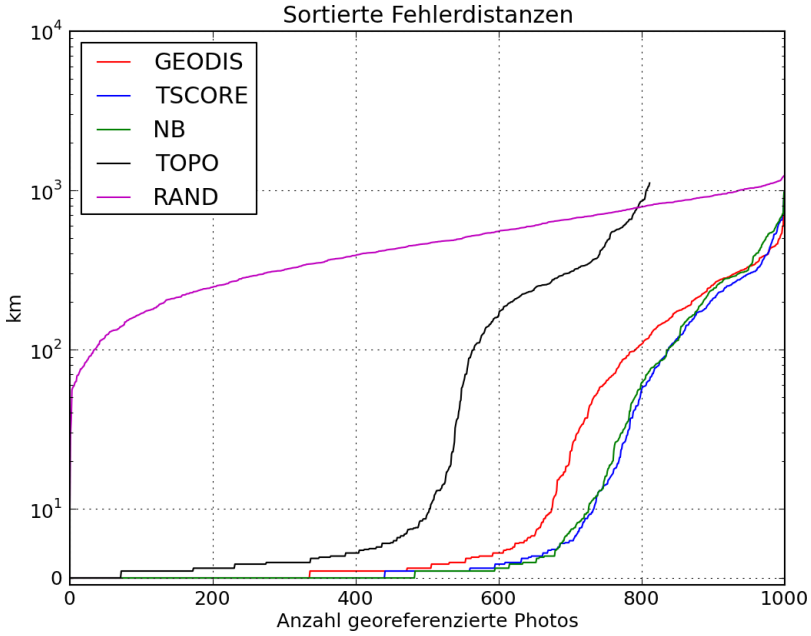
Beim Vergleich zwischen Grossbritannien und der Schweiz fällt auf, dass Mittelwerte und die Standardabweichungen bei allen Methoden in der Schweiz systematisch deutlich tiefer sind. Die Geometrie des Untersuchungsgebietes, das für Grossbritannien theoretisch viel grössere Fehlerdistanzen erlaubt als in der Schweiz, spielt hier wohl eine entscheidene Rolle.

Tab. 5.11: GB: Resultate aller Methoden mit Zellgrösse 1 km

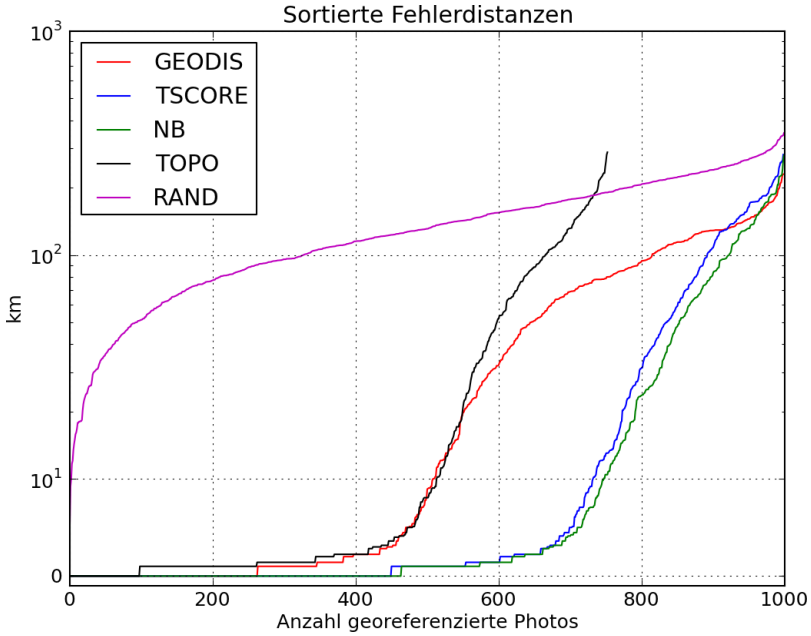
			GEODIS	TSCORE	NB	TOPO	RAND
Validationsdaten		Einheit					
	Total Photos	Anz.	1000	1000	1000	1000	1000
	Georeferenzierte Photos	Anz.	1000	1000	1000	812	1000
		%	100.0	100.0	100.0	**81.2**	100.0
Fehlerdistanz	Mittelwert	km	60.8	**49.5**	55.4	114.2	510.5
	Medianwert	km	1.4	**1.0**	**1.0**	4.0	462.1
	Standardabweichung	km	117.3	**116.0**	129.8	203.8	280.4
Vorhersage in	korrekter Zelle (n0)	%	33.6	44.1	**48.3**	7.2	0.0
	direkter Nachbarzelle (n1)	%	17.0	15.4	13.2	15.9	0.0
	Nachbarzelle 2. Grades (n2)	%	5.6	4.7	4.2	11.8	0.0
	Nachbarzelle 3. Grades (n3)	%	4.8	2.8	2.4	6.6	0.0
	n0, n1, n2, n3 (Total)	%	61.0	**67.0**	**68.1**	41.5	0.0

Tab. 5.12: CH: Resultate aller Methoden mit Zellgrösse 1 km

			GEODIS	TSCORE	NB	TOPO	RAND
Validationsdaten		Einheit					
	Total Photos	Anz.	1000	1000	1000	1000	1000
	Georeferenzierte Photos	Anz.	999	999	1000	753	1000
		%	99.9	99.9	100.0	**75.3**	100.0
Fehlerdistanz	Mittelwert	km	42.0	24.8	**20.6**	28.8	141.1
	Medianwert	km	8.1	**1.0**	**1.0**	2.2	131.0
	Standardabweichung	km	53.3	52.9	**45.6**	52.9	71.1
Vorhersage in	korrekter Zelle (n0)	%	26.3	**45.0**	**46.4**	9.8	0.0
	direkter Nachbarzelle (n1)	%	12.0	15.2	15.5	24.6	0.0
	Nachbarzelle 2. Grades (n2)	%	6.1	6.6	5.2	8.0	0.0
	Nachbarzelle 3. Grades (n3)	%	1.8	1.8	2.7	3.2	0.0
	n0, n1, n2, n3 (Total)	%	46.2	**68.6**	**69.8**	45.6	0.0



(a) Grossbritannien



(b) Schweiz

Abb. 5.8: Fehlerdistanzen aller Methoden mit Zellgrösse 1km

5.7 Einfluss der Datenfilterung

Tabelle 5.13 zeigt die Resultate der Georeferenzierung mit den Methoden GEODIS und TSCORE für verschiedene Filtervarianten des Datensatzes in Grossbritannien mit Zellgrösse 1 km. Folgende Filtervarianten wurden verwendet:

- F0 alle Filterschritte wie in Abschnitt 4.2 ausgeführt
- F1 keine Filterung ausser der Entfernung von Duplikaten und Photos ohne Tags
- F2 alle Filterschritte ausser der Entfernung von Bulk Uploads, Tagfilter und Nutzerfilter
- F3 alle Filterschritte ausser der Entfernung von Stoppwörtern

Die Resultate unterscheiden sich für die einzelnen Filtervarianten nur geringfügig. Die Variante F0 und jene ohne Filterung der Stoppwörter (F3) sind nahezu identisch. Die Varianten F1 und F2 erreichen ohne die Entfernung von Nutzerverzerrungen eine bessere Vorhersage der korrekten Zelle für beide gezeigten Methoden. Auch der Mittelwert der Fehlerdistanz sinkt erkennbar bei allen Methoden.

Tab. 5.13: Vergleich der Methoden bei unterschiedlicher Filterung der Daten

		GEODIS				TSCORE			
	Einheit	F0	F1	F2	F3	F0	F1	F2	F3
Validationsdaten									
Total Photos	Anz.	382625	1246442	1248472	382107	382625	1246442	1248472	382107
Georeferenzierte Photos	Anz.	382624	1243557	1245614	382106	382624	1243557	1245614	382106
	%	100.0	99.8	99.8	100.0	100.0	99.8	99.8	100.0
Fehlerdistanz									
Mittelwert	km	57.3	**54.2**	**54.1**	56.7	49.4	**45.8**	**44.4**	48.9
Medianwert	km	1.4	**1.0**	**1.0**	1.4	1.0	1.0	1.0	1.0
Standardabweichung	km	116.0	117.1	117.1	115.2	118.1	114.2	111.4	117.0
Vorhersage in									
korrekter Zelle (n0)	%	35.4	**39.1**	**39.1**	35.4	42.0	**49.2**	**49.5**	42.0
direkter Nachbarzelle (n1)	%	18.4	16.7	16.9	18.4	18.0	14.7	14.8	18.0
Nachbarzelle 2. Grades (n2)	%	5.3	5.4	5.4	5.4	5.1	4.6	4.5	5.1
Nachbarzelle 3. Grades (n3)	%	3.5	3.8	3.7	3.5	3.2	2.8	2.8	3.2
n0, n1, n2, n3 (Total)	%	62.6	65.0	65.1	62.7	68.3	71.3	71.6	68.3

F0 – alle Filterungschritte
F1 – keine Filterung (aber ohne Duplikate und Photos ohne Tags)
F2 – alle Filterschritte ausser Bulk Uploads, Nutzerfilter und Tagfilter
F3 – alle Filterschritte ausser der Entfernung von Stoppwörtern

5.8 Einfluss der Aufteilung in Trainings- und Validationsdaten

In den beiden folgenden Unterabschnitten wird der Einfluss der Trainingsdaten auf die Resultate der Georeferenzierung gezeigt.

5.8.1 Zufallsverfahren

In den vorangegangen Resultaten wurde die Aufteilung in Trainings- und Validationsdaten mit einem Zufallsgenerator für jedes Photo einzeln bestimmt (RP). Die Tabelle 5.14 vergleicht diese Resultate für die Methoden GEODIS, TSCORE und NB in Grossbritannien für die Zellgrösse 1 km mit den Resultaten der in Abschnitt 4.3 erwähnten zufälligen Zuteilung aller Photos eines Nutzer (RU) entweder in den Trainings- oder Validationsdatensatz.

Die Vorhersage der korrekten Zellen nimmt mit RU für alle Methoden markant ab, beziehungsweise der Mediandistanzfehler zu.

Tab. 5.14: Vergleich der Methoden bei unterschiedlichen Zufallsverfahren

		GEODIS		TSCORE		NB	
Validationsdaten	Einheit	RP[1]	RU[2]	RP[1]	RU[2]	RP[1]	RU[2]
Total Photos	Anz.	382625	388669	382625	388669	1000	1000
Georeferenzierte Photos	Anz.	382624	388044	382624	388044	1000	1000
	%	100.0	99.8	100.0	99.8	100.0	100.0
Fehlerdistanz							
Mittelwert	km	57.3	66.2	49.4	78.3	55.4	93.8
Medianwert	km	**1.4**	**2.2**	**1.0**	**2.2**	**1.0**	**8.1**
Standardabweichung	km	116.0	115.8	118.1	143.7	129.8	145.8
Vorhersage in							
korrekter Zelle (n0)	%	35.4	27.8	**42.0**	27.9	**48.3**	24.9
direkter Nachbarzelle (n1)	%	18.4	19.4	18.0	17.8	13.2	12.6
Nachbarzelle 2. Grades (n2)	%	5.3	6.0	5.1	5.7	4.2	4.7
Nachbarzelle 3. Grades (n3)	%	3.5	4.2	3.2	3.7	2.4	3.0
n0, n1, n2, n3 (Total)	%	**62.6**	57.3	**68.3**	55.0	**68.1**	45.2

[1] RP – zufällige Aufteilung pro Photo
[2] RU – zufällige Aufteilung pro Nutzer

5.8.2 Robustheit der Methoden bei weniger Trainingsdaten

Die bisher gezeigten Resultate basierten alle auf einem Verhältnis von 75 zu 25 zwischen Trainings- und Validationsdaten. Wie gross der Einfluss von weniger Trainingsdaten auf die Vorhersagegenauigkeit der Methoden GEODIS, TSCORE und NB ist, zeigt Tabelle 5.15.

Alle Methoden zeigen eine nur leicht schlechtere Vorhersagegenauigkeit bei einem Verhältnis von 25 zu 75 zwischen Trainings- und Validationsdaten.

Tab. 5.15: Vergleich mit anderen Trainings-Validations-Verhältnissen

		GEODIS		TSCORE		NB	
Validationsdaten	Einheit	T75V25	T25V75	T75V25	T25V75	T75V25	T25V75
Total Photos	Anz.	**382625**	**1147456**	**382625**	**1147456**	1000	1000
Georeferenzierte Photos	Anz.	382624	1147328	382624	1147328	1000	1000
	%	100.0	100.0	100.0	100.0	100.0	100.0
Fehlerdistanz							
Mittelwert	km	57.3	58.0	49.4	60.0	55.4	64.1
Medianwert	km	**1.4**	**1.4**	**1.0**	**1.0**	**1.0**	**1.0**
Standardabweichung	km	116.0	114.6	118.1	128.5	129.8	128.5
Vorhersage in							
korrekter Zelle (n0)	%	**35.4**	**33.1**	**42.0**	**38.2**	**48.3**	**40.2**
direkter Nachbarzelle (n1)	%	18.4	18.5	18.0	17.8	13.2	15.0
Nachbarzelle 2. Grades (n2)	%	5.3	5.5	5.1	4.9	4.2	3.6
Nachbarzelle 3. Grades (n3)	%	3.5	3.6	3.2	3.0	2.4	2.2
n0, n1, n2, n3 (Total)	%	**62.6**	**60.7**	**68.3**	**63.9**	**68.1**	**61.0**

T75V25 – Verhältnis von Trainings- zu Validationsdaten 75 zu 25
T25V75 – Verhältnis von Trainings- zu Validationsdaten 25 zu 75

5.9 Unterschiede von Photos mit und ohne Geotag

Für das Untersuchungsgebiet von Grossbritannien wurden für das Jahr 2011 je ein Datensatz bestehend aus Photos mit Geotags und einer mit Photos ohne Geotags verglichen (Abschnitt 4.8). Tabelle A.3 im Anhang zeigt eine Gegenüberstellung der 100 häufigsten Tags in den beiden Datensätzen.

5.9.1 Statistische Kennwerte

Tabelle 5.16 zeigt die statistischen Kennwerte der Datensätzte nach Ausführung des jeweiligen Filterschrittes. Photos ohne Geotag haben im Durchschnitt eine signifikant höhere Anzahl Tags als Photos mit Geotag. Ansonsten unterscheiden sich die Kennwerte der Datensätze nur unmerklich.

Tab. 5.16: Vergleich zwischen Photos mit und ohne Geotags

Filterschritt	PMT	TPP	N	T	PPN1	PPN2	PPN3
Photos mit Geotags							
Duplikate & Photos ohne Tags	**1334112**	8.8	40993	492372	32.5	4	184.0
Bulk uploads	**476216**	10.4	37137	450889	12.8	2	68.8
Stoppwörter	468773	**9.5**	36731	387402	**12.8**	2	68.9
Photos ohne Geotags							
Duplikate & Photos ohne Tags	**2627899**	12.6	58934	580220	44.6	5	301.8
Bulk uploads	**679099**	14.2	49311	519219	13.8	2	72.0
Stoppwörter	679098	**13.5**	49311	484999	**13.8**	2	72.0

P – Anzahl Photos total
PMT – Anzahl Photos mit Tags
TPP – Anzahl Tags pro Photo (Mittelwert)
N – Anzahl eindeutige Nutzer
PPN – Anzahl Photos pro Nutzer (PPN1 : Mittelwert, PPN2 : Medianwert, PPN3 : Standardabweichung)

5.9.2 Verwendung von Toponymen in den Tags

Mit dem offiziellen Ortsverzeichnis für Grossbritannien (vgl. Abschnitt 3.3) wurde die Verwendung von Toponymen in den Tags der beiden Datensätze aus dem vorherigen Abschnitt 5.9.1 untersucht. Tabelle 5.17 vergleicht die Ergebnisse zwischen den beiden Datensätzen.

In beiden Datensätze finden sich hohe Anteile von Photos mit Toponymen, wobei Photos ohne Geotag häufiger Toponyme als Tags verwenden. In beiden Datensätzen sind rund ein Fünftel aller benutzen Tags Toponyme.

Umgekehrt werden nur ein kleiner Teil aller im Ortsverzeichnis vorhanden Toponyme in den Tags gefunden. Photos ohne Geotags nutzen im Vergleich zu Photos mit Geotags einen leicht höheren Anteil der verfübaren Topoynme zur Beschreibung der Photos mit Tags.

Tab. 5.17: Verwendung von Toponymen als Tags

		mit Geotags	ohne Geotags
Total Photos	Anz.	468781	679098
mit Toponymen	Anz.	362867	635749
	%	**77.4**	**93.6**
Total Tags	Anz.	4469771	9158151
Toponyme	Anz.	849916	1853867
	%	**19.0**	**20.2**
Total eindeutige Tags	Anz.	440888	358277
Toponyme	Anz.	23390	19473
	%	5.3	5.4
Total Toponyme im Gazetteer	Anz.	188600	188600
als Tags benutzt	Anz.	19473	23390
	%	**10.3**	**12.4**

6 Diskussion

Die in Kapitel 5 gezeigten Resultate der Georeferenzierung von Flickr-Photos werden in diesem Kapitel analysiert und interpretiert. Die verschiedenen Methoden werden miteinander verglichen und verschiedene Faktoren mit Einfluss auf die Resultate gezeigt.

6.1 Georeferenzierung von Flickr-Photos

6.1.1 Georeferenzierung mit Sprachmodell und Toponymen

Mit der NB-Methode (Abschnitt 4.5.1) wird die Position von Photos mittels eines statistischen Sprachmodells – erstellt aus Tags der Photos in den Trainingsdaten – geschätzt. Sowohl in Grossbritannien als auch in der Schweiz werden mit NB ein Mediandistanzfehler von maximal 1.0 km erreicht. Dies entspricht bei einer Zellgrösse von 1.0 km einer Vorhersagegenauigkeit der korrekten Zelle für 48.3 % aller Photos in der Stichprobe in Grossbritannien und 46.4 % in der Schweiz. Im direkten Vergleich mit der TSCORE-Methode ist die Vorhersagegenauigkeit der korrekten Zelle mit NB minim besser, obwohl zumindest in GB der Mittelwert und die Standardabweichung leicht höher sind als bei der TSCORE-Methode.

Die TOPO-Methode (Abschnitt 4.5.2) versucht die Position von Photos aufgrund von in den Tags vorkommenden Toponymen zu bestimmen. Die Toponyme stammen aus den offiziellen Ortsverzeichnissen von Ordnance Survey und Swisstopo (Abschnitt 3.3). Da nicht alle Nutzer zur Beschreibung der Photos Toponyme verwenden, können nur rund drei Viertel der Photos georeferenziert werden (Tabelle 5.17). In der feinsten Gitterauflösung von 1.0 km beträgt der Mediandistanzfehler in Grossbritannen 3.2 km und in der Schweiz 2.2 km. Dies entspricht einer Vorhersagegenauigkeit der korrekten Zelle von 7.6 % (GB) beziehungsweise 15.2 % (CH). Damit ist die Georeferenzierung mit Toponymen deutlich weniger genau, als die TSCORE-Methode. Der kleinere Mediandistanzfehler in der Schweiz kann teilweise mit der höheren Anzahl Toponymen pro km^2 im Ortsverzeichnis erklärt werden (Tabelle 3.2).

6.1.2 Georeferenzierung mit ortsrelevanten Tags

Mit der in Abschnitt 4.6.1 gezeigten TF-IDF-Methode wurden für jede Zelle eines regelmässigen Gitternetzes die lokal relevantesten Tags aus den Flickr-Photos des Trainingsdatensatzes extrahiert.

Die beiden Methoden GEODIS und TSCORE (Abschnitt 4.6.2) nutzen diese ortsrelevanten Tags zur Vorhersage der Positionen der Photos im Validierungsdatensatz. Die TSCORE-Methode wählt für ein Photo jene Zelle als Standort, die für die vorhandenen Tags des Photos die höchste Summe der normalisierten TF-IDF-Werte aufweist. Im Unterschied dazu versucht die GEODIS-Methode die Position möglicher Zellen durch geometrische Disambiguierung einzuschränken. Aus den davon verbleibenden Zellen wählt die Methode dann ähnlich der TSCORE-Methode jene mit der höchsten Summe der normalisierten TF-IDF-Werte.

Die Resultate im Abschnitt 5.5 zeigen, dass die Vorhersage der korrekten Zelle, der Mittel- und der Mediandistanzfehler der Georeferenzierung mit TSCORE für alle berechneten Zellgrössen besser ist als mit GEODIS. Mit der TSCORE-Methode beträgt der Mediandistanzfehler (Abbildung 4.5) zwischen den Mittelpunkten der vorhergesagten und der das Photo enthaltenden Zelle sowohl in Grossbritannien als auch in der Schweiz maximal 1.0 km und nimmt mit zunehmender Zellgrösse ab. Somit können mit dieser Methode bei einer Zellgrösse von 1 km für rund 40 % aller georeferenzierten Bildern die korrekte Zelle vorhergesagt werden. Bei einer Zellgrösse von 5 km oder 10 km steigt die Vorhersagegenauigkeit auf über 50 %. Bei der Betrachtung der Vorhersagegenauigkeit unter Berücksichtigung der benachbarten Zellen (Abbildung 4.6) kann die geographische Region für mindestens 65 % der Photos korrekt bestimmt werden.

Mit der GEODIS-Methode nimmt der Mediandistanzfehler im Gegensatz zur TSCORE mit steigender Zellgrösse zu. In Grossbritannien wird mit 1.0 km Zellgrösse ein Mediandistanzfehler von 1.4 km erreicht. Dieser steigt bei Zellgrössen von 5 km und 10 km auf 5.0 km beziehungsweise 10.0 km. In der Schweiz liegt der Mediandistanzfehler bei einer Zellgrösse von 1.0 km mit dieser Methode bei 5.4 km und ist damit deutlich höher als in Grossbritannien. Damit verbunden ist auch der Anteil an Photos mit einer korrekt vorhergesagten Zelle geringer als mit der TSCORE-Methode. Bei 1.0 km Zellgrösse können in Grossbritannien mit der GEODIS-Methode für 35.4 %, in der Schweiz für 26.4 % der Photos die korrekte Zelle vorausgesagt werden.

6.1.3 Beurteilung und Kritik der Methoden

Die vorherigen beiden Unterabschnitte und die Resultate in Abschnitt 5.6 zeigen, dass die Methode TSCORE für die beiden untersuchten Gebiete keine besseren Ergebnisse erzielt als die Referenzmethode NB. Allerdings sind die Vorhersagen der beiden Methoden in etwa vergleichbar.

TF-IDF (Abschnitt 4.6.1) ist den Resultaten nach zu urteilen ein geeignetes Mittel, um aus den Beschreibungen von georeferenzierten Photos in Zellen ortsrelevante Tags zu extrahieren. Im Gegensatz zur Verwendung von Toponymen kann mit

ortsrelevanten Tags auf lokaler Massstabsebene besser auf die räumliche Variation von Tags eingegangen werden. Mit den offiziellen Ortsverzeichnissen zeigt sich das Problem der geometrischen Repräsentation der enthaltenen Ortsnamen als Punkte. In Abbildung 6.1 zeigt sich am Beispiel des Tags *London* der grosse Unterschied der räumlichen Ausdehnung bei der Verwendung von Toponymen und ortsrelevanter Tags. Da das Toponym *London* als Punkt vorliegt, entspricht seine Ausdehnung mangels Informationen der es enthaltenden Zelle, obwohl sich das Gebiet der Stadt weit über diese eine Zelle hinaus erstreckt. Im Gegensatz dazu umfasst die Ausdehnung des ortsrelevanten Tags *London* eine grössere Fläche des Stadtzentrums, wobei die Relevanz des Tag in den entsprechenden Zellen variiert.

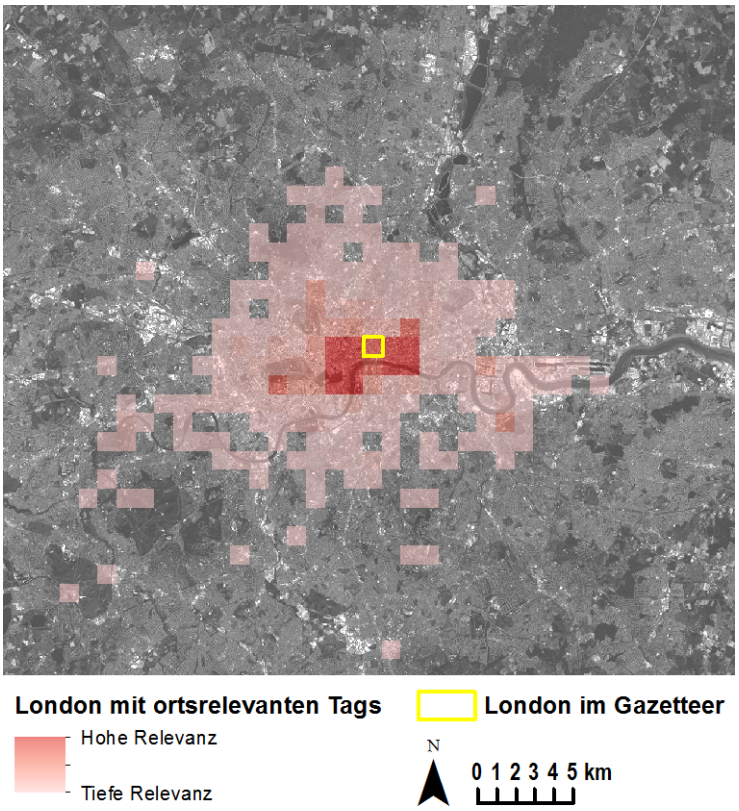


Abb. 6.1: Ausdehnung des Tags *London* mit TF-IDF (Hintergrund: NASA)

Zum Vergleich zeigt Abbildung 6.2 für den gleichen Bildausschnitt die Ausdehnung des Tags *London* mit einem NB-Klassifikator. Die Ausdehnung ist im Vergleich zu TF-IDF grösser, da jede Zelle, die im Trainingsdatensatz ein Photo mit dem Tag 'London' enthält eine gewisse Wahrscheinlichkeit zugeordnet bekommt. Im Unterschied dazu kommt *London* bei TF-IDF-Methode nur in Zellen vor, für die das Tag laut Algorithmus ortsrelevant ist.

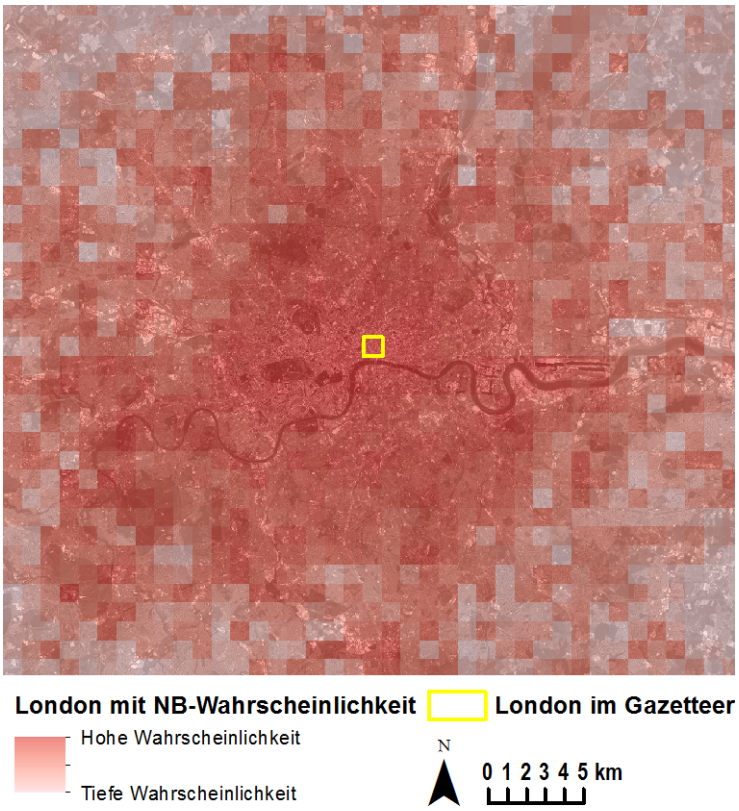


Abb. 6.2: Ausdehnung des Tags *London* mit Naive Bayes (Hintergrund: NASA)

Allerdings haben diese Methoden auch Nachteile und Limitierungen. Falls für eine Zelle im Trainingsdatensatz keine Photos enthalten sind, ist diese bei der späteren Georeferenzierung auch nicht als mögliche Position für Photos aus dem Validationsdatensatz vorhanden. Dieses Problem besteht für alle Methoden, welche Trainingsdaten zur Vorhersage des Standorts nutzen (GEODIS, TSCORE, NB). In Abschnitt 5.2 ist ersichtlich, dass vor allem in ländlichen und weniger dicht besiedelten Gebieten wenige oder keine Photos vorhanden sind.

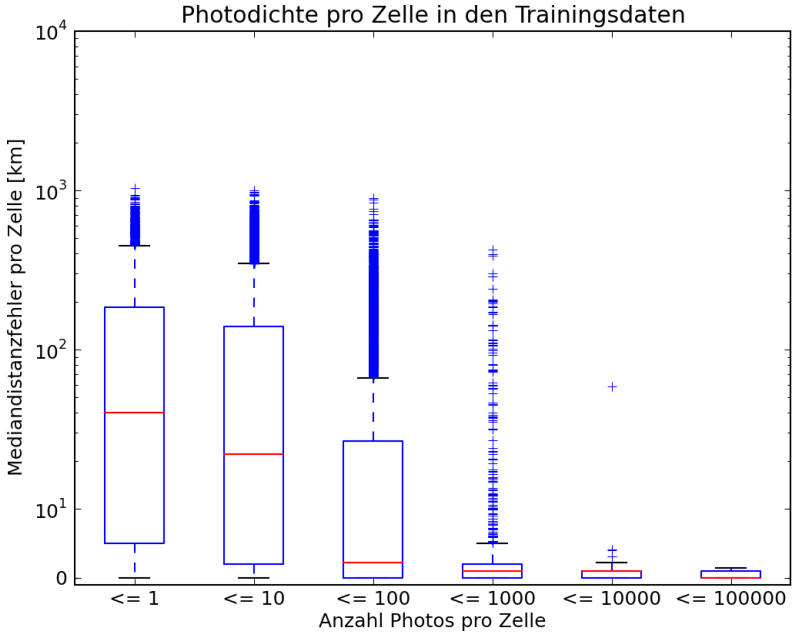
Grundsätzlich besteht das Problem auch bei der TOPO-Methode, die mit Toponymen die Position eines Photos ermittelt. Allerdings ist auf den Abbildungen im Abschnitt A.4 erkennbar, dass die Toponyme in den Ortsverzeichnissen räumlich wesentlich gleichmässiger verteilt sind. In der Schweiz besteht bei einer Zellgrösse von 1 km eine praktisch lückenlose Abdeckung, hingegen sind in Grossbritannien mehr Lücken zu erkennen.

Für die Methoden GEODIS, TSCORE und NB ist die Anzahl Photos pro Zelle in den Trainingsdaten ein wichtiger Einflussfaktor für die Vorhersagegenauigkeit. In Grossbritannien sind im Durchschnitt nach allen Filterschritten mit 6.7 Photos pro

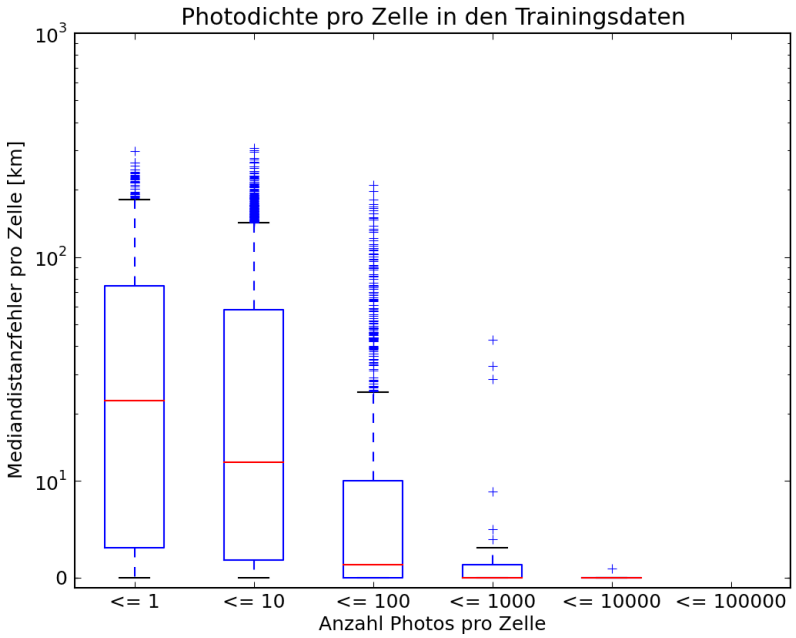
km^2 mehr als doppelt so viele Daten vorhanden als in der Schweiz mit 2.6 Photos pro km^2. Abbildung 6.3 zeigt den Einfluss der Anzahl Photos pro Zelle auf den Mediandistanzfehler am Beispiel der Methode TSCORE. Je mehr Photos in einer Zelle in den Trainingsdaten vorhanden sind, desto kleiner ist der Mediandistanzfehler von Photos in dieser Zelle.

Wie Abbildung 5.3 und 5.2 im vorhergehenden Kapitel 5 zeigen, konzentriert sich ein grosser Teil der Photos auf die urbanen Räume oder Tourismusdestinationen. In Abbildung 6.4 sind in Grossbritannien der Anteil Anteil 1 km-Zellen pro 10 km-Zelle mit einem Mediandistanzfehler kleiner als 5 km dargestellt (Methode TSCORE). In urbanen Ballungsgebieten wie zum Beispiel London oder Liverpool-Leeds lässt sich ein höherer Anteil an Zellen mit kleinem Mediandistanzfehler feststellen als in eher ländlichen Gebieten wie etwa Wales. Ebenfalls feststellen lässt sich dieser Effekt in touristischen Regionen an der Südküste der Hauptinsel, in Nationalparks wie beispielsweise Lake District oder Snowdonia sowie sonstigen geographisch prägnanten Punkten wie Kaps an den Küsten.

Als weitere Verkürzung und Vereinfachung betrachten sowohl die Methode TSCORE als auch NB die einzelnen Tags der Photos unabhängig voneinander. Bei der Extraktion der ortsrelevanten Tags beziehungsweise beim Training des NB-Klassifikators wird ein gemeinsames Auftreten von Tags in der Beschreibung eines Photos oder zusammen in einer Zelle genau so wenig berücksichtigt, wie bei der TF-IDF-basierten Extraktion ortsrelevanter Tags. Wie in Abschnitt 4.5.1 beschrieben, geht der Algorithmus von Naive Bayes sogar explizit von der Unabhängigkeit der Tags aus. Diese vereinfachende Annahme ist bei beiden erwähnten Methoden sowohl für eine Stärke – ihre einfache Implementation und Nachvollziehbarkeit der Ergebnisse – als auch für die bereits erwähnte Schwäche – das Ignorieren von gemeinsam auftretenden Tags – verantwortlich.



(a) Grossbritannien



(b) Schweiz

Abb. 6.3: Photodichte und Mediandistanzfehler mit TSCORE (Zellgrösse 1 km)

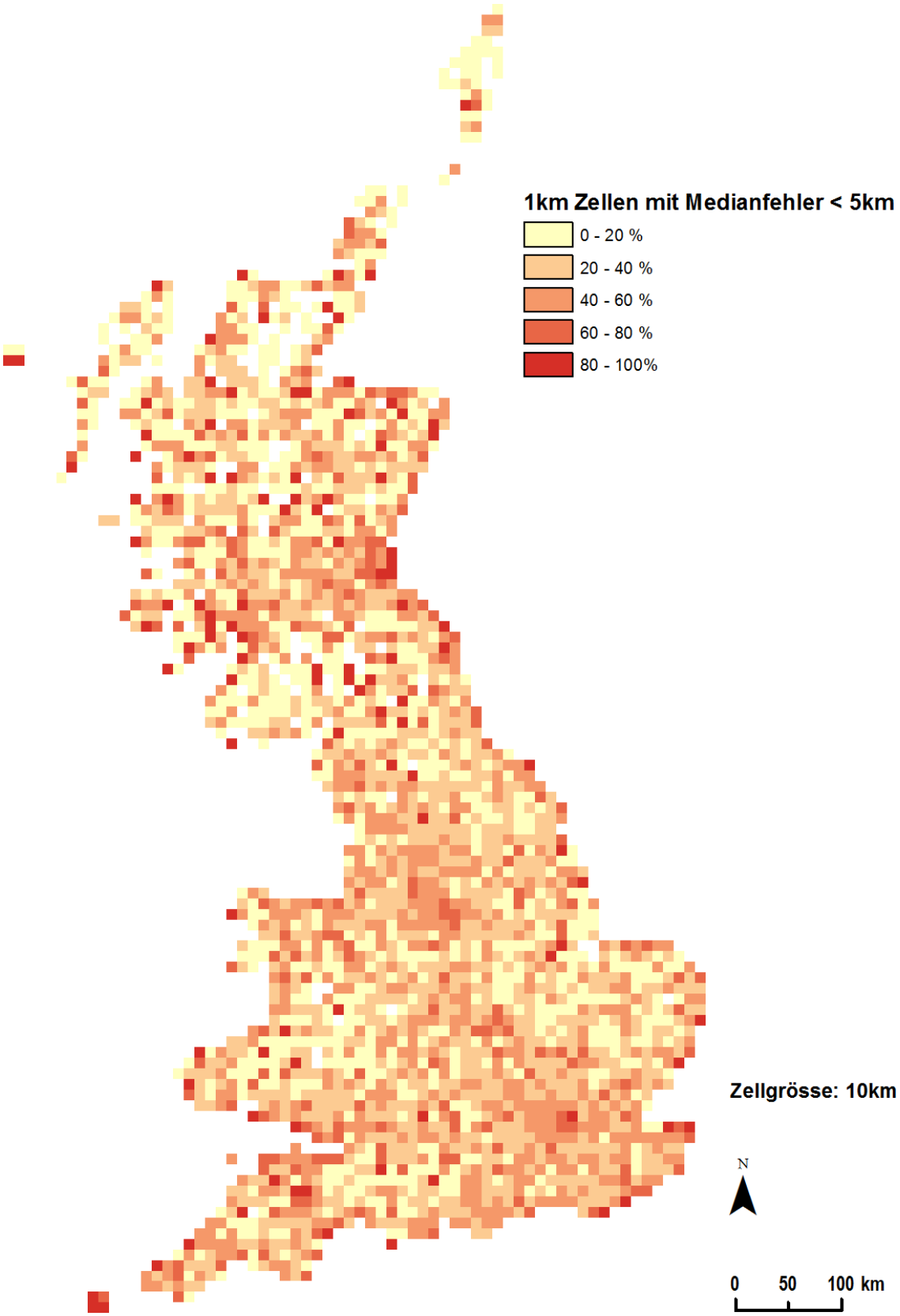


Abb. 6.4: Anteil 1 km-Zellen pro 10 km-Zelle mit Mediandistanzfehler < 5 km mit TSCORE in Grossbritannien

6.1.4 Einfluss der Zellgrösse auf die Resultate

Die Zellgrösse definiert die Seitenlänge der räumlichen Analyseeinheiten in der Form von regelmässigen Gitterzellen. Der Einfluss von unterschiedlichen Zellgrössen manifestiert sich in den Resultaten (Kapitel 5) in verschiedener Hinsicht.

Die Einteilung des Untersuchungsgebietes in regelmässige Gitterzellen erfolgt willkürlich und nimmt keine Rücksicht auf die lokalen Gegebenheiten. Durch eine Verschiebung des gesamten Gitternetzes oder verschiedenen Zellengrössen erhält man mit allen verwendeten Methoden unterschiedliche Resultate. Je grösser die Zellen gewählt werden, desto stärker ist auch der Einfluss auf die Ergebnisse (Abbildung 6.5).

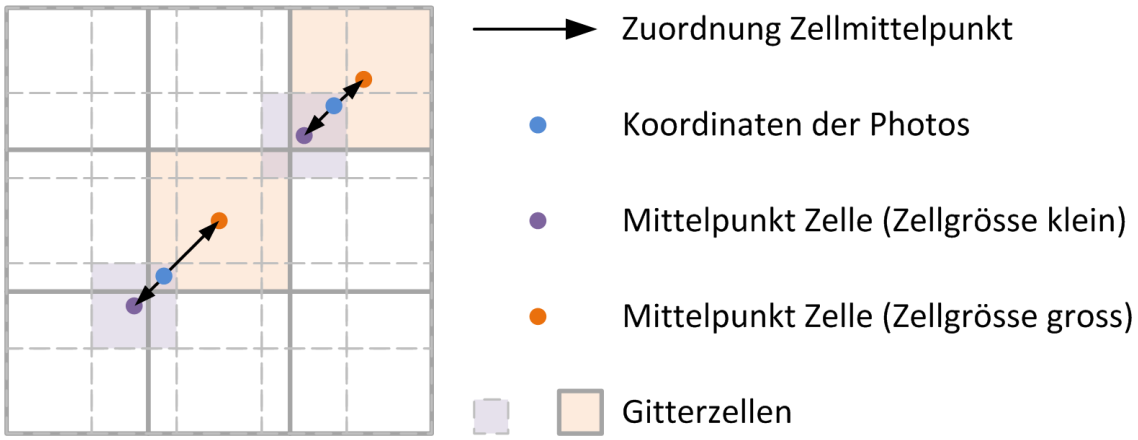


Abb. 6.5: Einfluss der Zellgrösse auf die Zugehörigkeit eines Photos zu einer Zelle

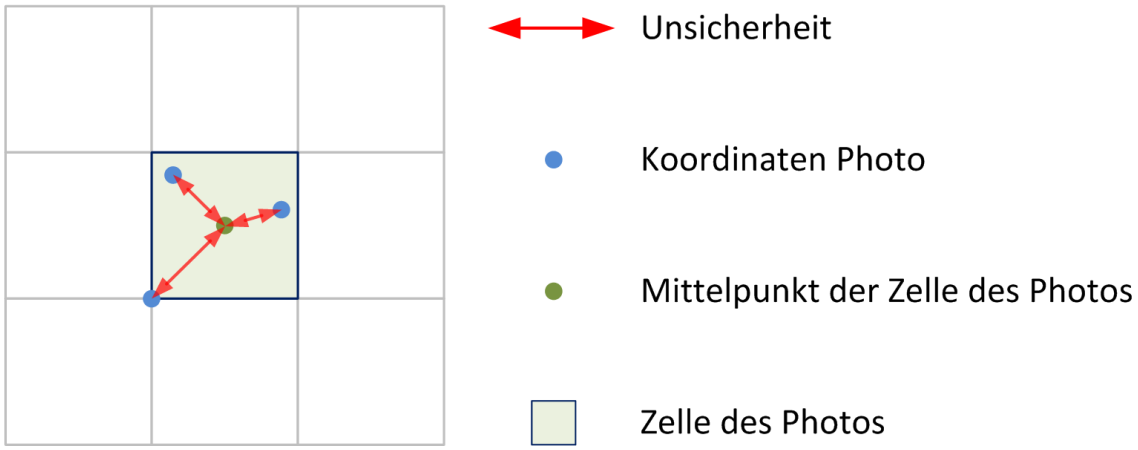


Abb. 6.6: Unsicherheiten verschiedener Positionen eines Photo in einer Zelle

Die Messung der Fehlerdistanz zwischen der vorhergesagten und der effektiven Zelle eines Photos (Abbildung 4.5) erfolgt zwischen den Mittelpunkten der beiden genannten Zellen. Zwischen dem Mittelpunkt der Zelle mit dem Photo und den im Geotag des Photos angegeben Koordinaten besteht somit immer eine zusätzliche Unsicherheit von maximal einer halben Zelldiagonale. Abbildung 6.6 zeigt mögliche Standorte und die dazugehörigen Unsicherheiten eines Photos in einer Zelle. Da die Länge der Zelldiagonalen abhängig von der Zellgrösse ist, ist die Unsicherheit umso geringer, je kleiner die Zellgrösse des Gitternetzes gewählt wird.

Die Zellgrösse beeinflusst auch die Vorhersagegenauigkeit der Methoden GEODIS, TSCORE und NB. Diese nimmt bei der geometrischen Disambiguierung (GEODIS) mit zunehmender Zellgrösse ab. Die Distanzen zwischen den Mittelpunkten der möglichen Zellen eines Photos und dem Zentroidpunkt werden mit zunehmender Zellgrösse ähnlicher. Damit ist es schwieriger, diese als Entscheidungskriterium für die korrekte Vorhersage der Zelle einzusetzen. Bei den Methoden TSCORE und NB verbessert sich die Vorhersagegenauigkeit bei steigender Zellgrösse, weil mit steigender Zellgrösse insgesamt weniger mögliche Zellen für eine korrekte Vorhersage in Frage kommen (Abbildung 6.7). Gleichzeitig steigt aber der bereits erwähnte Distanzfehler (Abbildung 6.6) zwischen dem Mittelpunkt der Zelle mit dem Photo und der gegebenen Position des Photos.

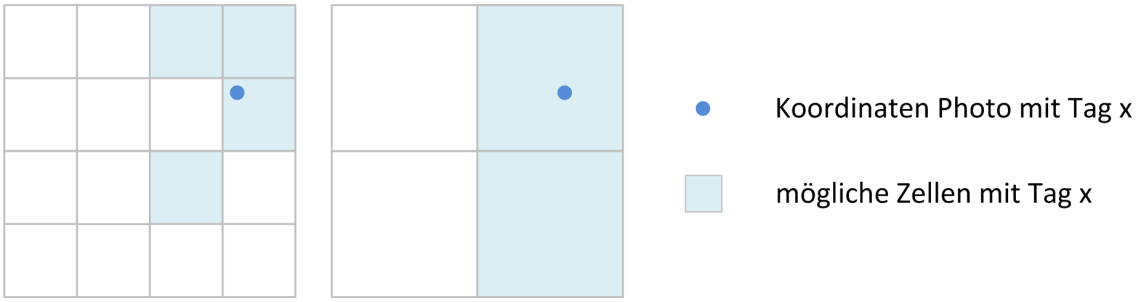


Abb. 6.7: Einfluss der Zellgrösse auf die Anzahl möglicher korrekter Zellen

Bei der Wahl der Zellgrösse spielen somit mehrere Kriterien eine Rolle. Mit den Methoden TSCORE und NB wird mit steigender Zellgrösse auch die Vorhersagegenauigkeit besser. Wenn also in einer Anwendung beispielsweise die Region von UGC auf Stadtebene vorhergesagt werden soll, erzielt man mit einer gröberen Zellgrösse von 5 km bessere Ergebnisse als mit 1 km. Hingegen erfordert eine feiner aufgelöste Georeferenzierung auch eine kleinere Analyseeinheiten und so eine kleinere Zellgrösse. Zudem müssen die Daten auch die geforderte Positionsgenauigkeit aufweisen. Je nach Grösse der Zellen werden die auf der entsprechenden Masstabsebene relevanten Tags relevanter eingestuft, wie das vereinfachende Beispiel in Abbildung 6.8 aufzeigt.

Eine kleinere Zellgrösse hat aber durch die heteregone räumliche Verteilung der Daten

Abb. 6.8: Relevanz von Tags aufgrund ihrer Häufigkeit bei unterschiedlicher Zellgrösse

auch mehr leere Zellen zur Folge. Dieses Problem kann aber durch die Anwendung eines zellbasierten Glättungsverfahrens wie zum Beispiel in Serdyukov et al. (2009) oder in O'Hare & Murdock (2012) gelöst werden.

6.1.5 Vergleich der Ergebnisse mit bisheriger Forschung

Die jüngsten Forschungsergebnisse in der Georeferenzierung von UGC basieren meist auf der Anwendung eines Sprachmodells und ML, ähnlich der NB-Methode in dieser Arbeit. Die in dieser Arbeit entwickelte TSCORE-Methode, basierend auf ortsrelevanten Tags, wird im Folgenden mit diesen Ergebnissen verglichen.

Da die folgenden Arbeiten ihre Stichproben jeweils zufällig für alle Photos eines Nutzers in Trainings- und Validationsdaten aufgeteilt haben, werden die Ergebnisse der TSCORE-Methode mit RU-Aufteilung (Abschnitt 4.3 und 5.8.1) zur Gegenüberstellung verwendet.

Van Laere et al. (2010) haben bei der Georeferenzierung von Flickr-Photos in 55 europäischen Städten auf einen einfaches Sprachmodell mit NB gesetzt. Im Unterschied zu dieser Arbeit verwenden sie kein regelmässiges Gitternetz als räumliche Analyseeinheiten, sondern definieren diese mit einer Clustering-Methode. Abhängig von der vorgängig definierten Anzahl an räumlichen Clustern erreichen sie unterschiedliche Vorhersagegenauigkeiten. Mit 55 Clustern, was einem pro untersuchter Stadt enspricht, erreichen sie eine Vorhersagegenauigkeit von 86.9 %. Mit 1'000 Clustern sinkt der Wert auf 41.2 %. Ein Vergleich mit dem flächendeckenden Ansatz

dieser Arbeit ist durch die Beschränkung der Analyse auf Städte mit durchschnittlich hoher Photodichte (siehe Abschnitt 6.1.3) nur bedingt möglich.

Besser vergleichbar sind die Resultate dagegen mit der Arbeit von Serdyukov et al. (2009) beziehungweise der erweiterten Variante von O'Hare & Murdock (2012), die regelmässige Gitternetze verwenden. Wie bereits bei Van Laere et al. (2010) werden in dieser Arbeit auch ein Sprachmodell zur Georeferenzierung von Flickr-Photos eingesetzt. Im Gegensatz zu dieser Arbeit schränken O'Hare & Murdock (2012) ihr Untersuchungsgebiet nicht ein, sondern versuchen die Vorhersage der Position der Photos weltweit. Für eine Zellgrösse von 1 km können sie für 16.2 % der Photos die richtige Zelle vorhersagen. Tabelle 6.1 zeigt eine Gegenüberstellung der Resultate von O'Hare & Murdock (2012) und der TSCORE-Methode dieser Arbeit für Grossbritannien für eine Zellgrösse von 1 km.

Tab. 6.1: Vergleich der Ergebnisse von O'Hare & Murdock und der TSCORE-Methode

		O'Hare & Murdock	TSCORE
Untersuchungsgebiet		**weltweit**	**Grossbritannien**
Zellgrösse		1km	1km
Vorhersage in			
korrekter Zelle (n0)	%	**16.2**	**27.9**
+ direkter Nachbarzelle (n1)	%	29.3	45.2
+ Nachbarzelle 2. Grades (n2)	%	34.3	51.4
+ Nachbarzelle 3. Grades (n3)	%	**37.5**	**55.0**

Die Resultate der TSCORE-Methode mit ortsrelevanten Tags sind deutlich besser als der aktuellste Stand der Forschung in O'Hare & Murdock (2012) mit einem ML-Ansatz. Da sie aber die ganze Welt abdecken und verglichen mit der Fläche von Grossbritannien und der Schweiz viel weniger Daten zum Training des Sprachmodells zur Verfügung haben, ist der Vergleich in Tabelle 6.1 ebenso nur mit Vorbehalt möglich.

Abschliessend zeigen die Resultate der TSCORE-Methode, auch verglichen mit der NB-Methode in dieser Arbeit, das Potential der Extraktion und Bewertung von örtlicher Relevanz mit TF-IDF aus UGC auf.

6.2 Einfluss der Filterung

In Abschnitt 5.7 wird der Einfluss der Filterung der Daten gezeigt. Durch gezieltes Weglassen von Filterschritten wird deren Einfluss auf die Resultate der Georeferenzierung analysiert.

Zwischen dem Datensatz mit allen Filterschritten (F0) und jenem ohne entfernte Stoppwörter (F3) lässt sich bei den Methoden GEODIS und TSCORE in Grossbritannien kein Unterschied in den Resultaten nachweisen. Da Stoppwörter in einem Fliesstext mutmasslich häufiger vorkommen als in den Tags von Flickr-Photos, ist ihr Einfluss auf die Vorhersagegenauigkeit offenbar minim. Zudem ist ihre räumliche Verteilung eher homogen, was bei der Bewertung mit dem TF-IDF-Algorithmus eher zu kleiner Ortsrelevanz führt. Nichtsdestotrotz sollten Stoppwörter aus Gründen der Effizienz bei der Datenverarbeitung entfernt werden.

Der Effekt von Bulk Uploads (mehrere Photos eines Nutzers mit identischen Tags) ist vor allem bei der TSCORE-Methode deutlich erkennbar. Im Vergleich mit ungefilterten Daten (F1) oder Daten ohne Filterung von Bulk Uploads, Tagfilterung und Nutzerfilterung (F2) erzielen die normal gefilterten Daten eine schlechtere Vorhersagegenauigkeit. Wo in Grossbritannien für die Zellgrösse 1 km mit der Methode TSCORE eine Vorhersagegenauigkeit von 42.0 % für F0 erreicht wird, ist diese mit 49.2 % für F1 beziehungsweise 49.5 % für F2 erkennbar höher. Die Filterung von Bulk Uploads und Nutzereffekten ist deshalb ratsam, da diese sonst zu markanten Verzerrungen in den Ergebnissen führen. Der Effekt auf die Vorhersage der korrekten Zelle ist allerdings moderat, wenn man den Anteil an entfernten Bulk Uploads in F0 betrachtet (Tabelle 5.13). Dies dürfte ein Effekt des zusätzlichen Nutzerfaktors uf (Gleichung 4.5) in der TF-IDF-Methode zur Extraktion von ortsrelevanten Tags sein, welche dem Einfluss der nur von wenigen Nutzern verwendeten Tags in einer Zelle dämpft.

6.3 Einfluss der Aufteilung in Trainings- und Validationsdaten

In Abschnitt 5.8 wurde der Einfluss des Zufallsverfahrens bei der Aufteilung der Stichprobe und des Verhältnisses zwischen der Anzahl Photos in den Trainings- und Validationsdaten aufgezeigt.

Ob die gefilterten Daten zufällig pro Photo (RP) oder zufällig pro Nutzer (RU) in Trainings- und Validationsdaten aufgeteilt werden, hat einen erheblichen Einfluss auf die Vorhersagegenauigkeit der Methoden. Mit der Aufteilung RP sagt die Methode TSCORE mit einer Zellgrösse von 1 km für 42.0 % der Validationsdaten die korrekte Zelle voraus. Dieser Anteil sinkt markant auf 27.9 %, wenn die Zuteilung für alle Photos eines Benutzers entweder zu den Trainings- oder den Validationsdaten erfolgt. Die mit der RP-Aufteilung erzielte Vorhersagegenauigkeit ist also für neue, nicht im Trainingsdatensatz vorhandenen Photos nur bedingt erreichbar, da offensichtlich Verzerrungen durch Photos von Nutzern in beiden Datensätzen zu einer tendenziell zu hohen Vorhersagegenauigkeit führen.

Der Einfluss des Verhältnisses zwischen Trainings- und Validationsdaten ist in Tabelle 5.15 aufgeführt. Bei einem Verhältnis von 25 zu 75 zwischen der Anzahl Trainings- und Validationsdaten lässt sich im Vergleich mit dem Verhältnis 75 zu 25 nur eine leicht geringere Vorhersagegenauigkeit feststellen. Allerdings dürfte die Sensitivität der Methoden auf kleinere Trainingsstichproben bei der vorher erwähnten Aufteilung RU grösser sein.

6.4 Photos mit und ohne Geotag

Die Ähnlichkeiten zwischen Photos mit beziehungsweise ohne Geotag aus Grossbritannien für das Jahr 2011 wurden mittels statischen Kennwerten in Abschnitt 5.9 gezeigt.

Die beiden analysierten Datensätze aus Grossbritannien für das Jahr 2011 sind sich in ihren statistischen Kennwerten mit einer Ausnahmen ähnlich. Nach der Filterung haben Photos ohne Geotag im Durchschnitt mit 13.5 Tags pro Photo 4.0 Tags mehr als Photos mit Geotags (9.5 Tags pro Photo). Zudem findet sich auch in den 100 häufigsten Tags für beide Datensätze eine grössere Übereinstimmung (Tabelle A.3). Damit sind die Voraussetzungen zur Georeferenzierung von Photos ohne Geotags mithilfe von Photos mit Geotag potentiell erfüllt.

Aufgrund der Suchmethodik zum Empfang der Photos ohne Geotag mit Suchwörtern, können sich in dieser Stichprobe auch Photos befinden, die nicht in GB geschossen wurden. Zudem handelt es sich dabei, im Gegensatz zu den Photos mit Geotags, nicht um alle öffentlichen Photos ohne Geotags aus GB im Jahr 2011, da nur Photos mit einem der verwendeten Suchbegriffe in der Stichprobe auftauchen können. Die obigen Resultate sind deshalb auch nur unter Vorbehalt korrekt.

Bei der Verwendung von Toponymen in den Tags der beiden untersuchten Stichproben zeigte sich, dass in den Tags von Photos ohne Geotag häufiger Toponyme vorkommen als bei Photos mit Geotag. 93.2 % aller Photos ohne Geotag verwenden mindestens ein Toponym als Tag, während dies bei 77.4 % der Photos mit Geotag der Fall ist, was ein ähnlicher Anteil ist wie bei der Untersuchung von Hollenstein & Purves (2010). Diese Anteile dürften in der Realität allerdings tiefer sein, wie Tabelle A.4 im Anhang zeigt. In den 500 häufigsten Tags beider Datensätze finden sich viele Toponyme mit ambivalenter Bedeutung. In beiden Datensätzen ist *Park* das am häufigsten auftretende Toponym. In den vielen Fällen ist damit nicht einer der Ortsnamen im Ortsverzeichnis gemeint, sondern der Begriff *Park* als Bezeichnung für ein diskretes Objekt. Weitere Beispiele von ambivalenten Ortsnamen sind *Hill, Lake, Street* oder *Castle*.

Zusammenfassend verwenden Photos ohne Geotags im Mittel sowohl mehr Tags als auch mehr Toponyme zur Beschreibung der Photos, womit eine wichtige Grundvoraussetzung zur Georeferenzierung erfüllt ist.

7 Schlussfolgerungen und Ausblick

Im letzten Kapitel werden die Forschungsfragen beantwortet, die Erkenntnisse der Arbeit rekapituliert sowie offene Fragen und Verbesserungsvorschläge für zukünftige Forschung thematisiert.

7.1 Erreichtes

Ein Ziel dieser Arbeit war es, den Standort von nutzergenerierten Daten nur aufgrund ihrer textuellen Beschreibung vorherzusagen. Adressiert wurden damit zentrale Herausforderungen von GIR – die Erkennung von räumlicher Semantik in Form von Ortsnamen und umgangssprachlichen räumlichen Bezeichnungen in Textdokumenten oder Suchanfragen sowie deren räumliche Interpretation. Eine wichtige Voraussetzung dafür ist neben den bestehenden Konzepten die Implementation von menschlich-subjektiven Raumvorstellungen in GIS, die unter anderem durch die Analyse der natürlichen Sprache in UGC umgesetzt werden kann. In dieser Arbeit wurde versucht, aus den Metadaten von UGC in der Form von georeferenzierten Flickr-Photos räumliche Semantik zu extrahieren, welche dann zur Vorhersage des Standorts von anderen Photos verwendet werden kann. Als Untersuchungsgebiete dienten Grossbritannien und die Schweiz.

Mit einer räumlichen TF-IDF-Variante wurde die örtliche Relevanz der aus den Tags der Photos im Trainingsdatensatz extrahierten Informationen bewertet (Abschnitt 4.6.1) und zur Georeferenzierung der Photos im Validationsdatensatz nur aufgrund ihrer Tags verwendet (Abschnitt 4.4). Die Validation der Ergebnisse erfolgte mit den in den Metadaten der Photos vorhandenen geographischen Koordinaten und durch einen Vergleich mit den Ergebnissen von Referenzmethoden (Kapitel 5).

Der Ansatz von Rattenbury & Naaman (2009) hat sich als robust und geeignet zur Extraktion von ortsrelevanten Tags aus UGC erweisen. Die ortsrelevanten Tags umfassen dabei im Gegensatz zu einem offiziellen Ortsverzeichnis nicht nur Toponyme, sondern auch umgangssprachliche Ortsbezeichnungen. Mit dem TF-IDF-Mass als numerische Grösse für örtliche Relevanz der umgangssprachlichen Ortsbezeichnungen konnten in Grossbritannien und in der Schweiz je nach Zellgrösse für mindestens 40 % der Photos die korrekte Zelle vorhergesagt werden. Für mindestens 65 % der Photos konnte die Region in einer 7×7-Nachbarschaft der korrekten Zelle vorhergesagt werden. Die Grössenordnung der Resultate bewegt sich dabei im Rahmen von aktuellen Forschungsergebnissen, die meist auf einem statistischen Sprachmodell und ML

basieren und übertreffen die Resultate von rein toponym-basierter Georeferenzierung deutlich.

Die Voraussetzungen eines möglichen Einsatzes der Methode zur Georeferenzierung von Photos ohne Geotag wurden mit einem Vergleich von zwei Stichproben von Photos aus Grossbritannien im Jahr 2011 geprüft (Abschnitt 5.9). Photos mit und solche ohne Geotag haben im Mittel mit einer Ausnahme ähnliche statistische Kennwerte. Allerdings beschreiben Nutzer ihre Photos ohne Geotag im Durchschnitt mit mehr Tags als solche mit Geotag. Ausserdem werden zur Beschreibung der Photos ohne Geotag öfters Ortsnamen verwendet. Damit sind die theoretischen Voraussetzungen zur Georeferenzierung von Flickr-Photos ohne Geotag mit räumlicher Semantik extrahiert aus Photos mit Geotag potentiell gegeben.

Weiter wurde der Einfluss der Datenfilterung auf die Qualität der Georeferenzierung untersucht. Bisherige Forschung (zum Beispiel Hollenstein & Purves, 2010 oder Purves et al., 2011) haben das Problem von durch Nutzerbeiträge verursachte Verzerrungen in den Ergebnissen aufgezeigt. Die Untersuchungen in dieser Arbeit haben ergeben, dass vor allem der Einfluss von Bulk Uploads das Resultat verzerrt. Die Verzerrungen fielen aber im Verhältnis zu den in der Stichprobe enthaltenen Bulk Uploads moderat aus, was mitunter auch eine Folge der robusten Methode von Rattenbury & Naaman (2009) ist, da sie eine verbreitete Verwendung eines Tags unter vielen Nutzern berücksichtigt und stärker gewichtet.

7.2 Beantwortung der Forschungsfragen

Die aus den Forschungslücken in Abschnitt 2.5 abgeleiteten Forschungsfragen dieser Arbeit werden im Folgenden beantwortet.

Forschungsfrage 1:

> *Mit welcher Genauigkeit können nutzergenerierte Daten ohne explizite Ortsangaben aufgrund ihrer textuellen Beschreibung georeferenziert werden?*

Als nutzergenerierte Daten wurden in dieser Arbeit Flickr-Photos aus Grossbritannien und der Schweiz verwendet. Aus den Tags (textuelle Beschreibung) und Geotags (geographische Koordinaten) der Photos konnten ortsrelevante Tags erkannt und extrahiert werden. Mit diesen konnten weitere Photos aufgrund ihrer textuellen Beschreibung georeferenziert werden (Kapitel 4).

Die metrische Genauigkeit der Georeferenzierung lässt sich für die Photos in den Validierungsdaten als Fehlerdistanz zwischen den Mittelpunkten der geschätzten und der das Photo enthaltenden Gitterzelle ausdrücken (Abschnitt 4.4). Ausserdem

kann auch der Anteil der Photos, für welche die korrekte Zelle oder eine benachbarte Zelle vorhergesagt wurde, als Qualitätskriterium für die Vorhersagegenauigkeit der Georeferenzierung verwendet werden.

In dieser Arbeit wurden für eine Zellgrösse von 1 km in Grossbritannien für 42.0 %, in der Schweiz für rund 39.5 % der georeferenzierten Photos die korrekte Zelle vorhergesagt. Dies entspricht einem Mediandistanzfehler von 1.0 km. Bei Berücksichtigung der Nachbarzellen innerhalb von 5 km um die korrekte Zelle kann die Region eines Photos in beiden Ländern in über 65 % der Fälle richtig bestimmt werden. Die Resultate in dieser Arbeit erreichen im Vergleich zur Georeferenzierung mit einem Sprachmodell eine vergleichbare Genauigkeit und übertreffen damit die Genauigkeit der Georeferenzierung mit einem Ortsverzeichnis (Abschnitt 5.6).

Wie die Untersuchungen in dieser Arbeit gezeigt haben, hängt die erzielte Genauigkeit aber von mehreren Einflussfaktoren ab. Die Zellgrösse des Gitternetzes hat einen Einfluss auf die mögliche Unsicherheit, welche durch die Distanz zwischen dem effektiven Standort eines Photos und dem Mittelpunkt der Zelle entsteht. Zudem verändert sich mit der Zellgrösse auch die Vorhersagegenauigkeit, der Mediandistanzfehler und die örtliche Relevanz von Tags innerhalb der Zellen (Abschnitt 6.1.4). Ebenso haben das gewählte Untersuchungsgebiet, die Filterung von Nutzerverzerrungen, die Grösse der Stichprobe zur Extraktion der ortsrelevanten Tags und das gewählte Zufallsverfahren zur Aufteilung der Stichprobe in Trainings- und Validationsdaten einen teilweise erheblichen Einfluss auf die Genauigkeit (Abschnitt 6.2 und Abschnitt 6.3).

Eine allgemeine Aussage über die mögliche Genauigkeit der Georeferenzierung von nutzergenerierten Daten ist deshalb schwierig. Obwohl die Verwendung der Methoden in dieser Arbeit problemlos mit anderen Datensätzen möglich ist, kann im Hinblick auf die oben genannten Einflussfaktoren und aufgrund der unterschiedlichen Eigenschaften von Daten aus anderen Quellen keine abschliessende Aussage über die mögliche Genauigkeit einer Georeferenzierung von UGC gemacht werden. Die oben gezeigten Aussagen sind somit auf die Georeferenzierung von Flickr-Daten beschränkt. Deren Verwendung hat aber das Potential der Methoden im Hinblick auf eine mögliche Qualität der Georeferenzierung aufgezeigt. Durch Verfeinerung und Verbesserung können diese zudem noch optimiert werden.

Die Verwendung von georeferenzierten Photos der im Gegensatz zu Flickr moderierten Plattform *Geograph*, könnte durch die homogenere räumliche Verteilung zu weniger leeren Gitterzellen und somit zu einer besseren Abdeckung mit ortsrelevanten Tags führen. Jedoch haben Purves et al. (2011) gezeigt, dass auf Geograph weniger Nomen und Toponyme zur Beschreibung der Bilder verwendet werden, was wiederum einen negativen Effekt auf die Qualität der Georeferenzierung haben dürfte. Für Twitter-Kurznachrichten ist die Bestimmung des Standortes aufgrund ihrer begrenzten Zeichenanzahl und den spärlich enthaltenen räumlichen Hinweisen im Vergleich zu

Flickrdaten eine grössere Herausforderung (Cheng et al., 2010; Wing & Baldridge, 2011; Kinsella et al., 2011).

Forschungsfrage 2:

> *Wie ähnlich sind sich die textuellen Beschreibungen von Flickr-Photos mit Geotag und solchen ohne Geotag?*

Bei der Untersuchung zweier Stichproben mit Flickr-Photos aus Grossbritannien hat sich gezeigt, dass Flickr-Photos ohne Geotags im Durchschnitt mit mehr Tags beschrieben werden als Photos mit Geotags (Abschnitt 6.4). Ein Vergleich der 100 häufigsten Tags in beiden Datensätzen zeigt eine grosse Ähnlichkeit bei den verwendeten Schlüsselwörtern zur Beschreibung der Photos (Abschnitt A.5).

Auch hinsichtlich der Verwendung von Toponymen zeigt sich im Vergleich der beiden Stichproben, dass bei Photos ohne Geotag häufiger Toponyme zur Beschreibung der Photos verwendet werden. Zudem kommt in der Stichprobe in einem höheren Anteil der Photos ohne Geotag mindestens ein Toponym vor.

Eine Ähnlichkeit zwischen Photos mit und solchen ohne Geotag ist in der untersuchten Stichprobe vorhanden. Damit hat diese Arbeit das theoretische Potential zur Vorhersage des Standorts und die damit mögliche Anreicherung der Metadaten von Flickr-Photos ohne Geotag unter der Verwendung von Photos mit Geotag aufgezeigt.

Forschungsfrage 3:

> *Welchen Einfluss hat die Filterung nutzergenerierter Daten auf die Qualität der Georeferenzierung?*

Der Vergleich zwischen den Resultaten der Georeferenzierung von gefilteren und ungefilterten Datensätzen hat durch Nutzerbeiträge verursachte Verzerrungen aufgezeigt (Abschnitt 6.2). Vorallem Bulk Uploads – mehrere Photos eines Nutzers mit identischen Tags – verzerren die Ergebnisse. Die Vorhersagegenauigkeit der korrekten Zelle wird durch Bulk Uploads deutlich besser angegeben, als wenn diese aus der Stichprobe entfernt wurden.

Die Filterung von Tags, die nur von einem Benutzer verwendet werden, der Auschluss von Nutzern mit einer überdurchschnittlich hohen Anzahl Photos in der Stichprobe oder die Entfernung von Stoppwörtern scheinen dagegen verglichen mit den Bulk Uploads einen geringen Einfluss auf die Ergebnisse zu haben. Dies ist teilweise auf die Robustheit der verwendeten Methode von Rattenbury & Naaman (2009) zur Extraktion der ortsrelevanten Tags zurückzuführen, da sie von vielen Nutzern verwendete Tags in einer Zelle höher gewichtet, als von wenigen Nutzern verwendete Tags.

7.3 Erkenntnisse und Fazit

Der Beitag dieser Arbeit mit dem Titel *Georeferenzierung von nutzergenerierten Daten* lässt sich wie folgt zusammenfassen:

- Die räumliche TF-IDF-Methode von Rattenbury & Naaman (2009) ist ein möglicher Ansatz zur Extraktion von ortsrelevanten Tags für die Georeferenzierung und liefert mit statistischen Sprachmodellen vergleichbare Ergebnisse.
- Die Qualität der Georeferenzierung mit ortsrelevanten Tags ist abhängig von der Zellgrösse des Gitternetzes, des Untersuchungsgebietes, der Repräsentativität und räumlichen Verteilung der Stichprobe sowie dem Zufallsverfahren zur Aufteilung in Trainings- und Validationsdaten.
- Die Filterung von Nutzereffekten hat einen nachweisbaren Einfluss auf die Vorhersagegenauigkeit der Georeferenzierung.
- Flickr-Photos ohne Geotag werden im Mittel mit einer höheren Anzahl Tags beschrieben als jene mit Geotag. Zudem finden sich in den Tags von Photos ohne Geotag häufiger Toponyme.

Die Georeferenzierung von Flickr-Photos mit ortsrelevanten Tags hat sich als simpler und robuster Ansatz herausgestellt. Die Resultate sind vergleichbar mit denen von einfachen Sprachmodellen, welche bei Ansätzen mit ML verwendet werden.

Nachteilig bei der Verwendung eines Gitternetzes mit fixer Zellgrösse ist die willkürliche räumliche Aggregation der Photos in Gitterzellen. Zudem wurde in dieser Arbeit das Problem von leeren Zellen nicht adressiert.

Die Extraktion von ortsrelevanten Tags mit TF-IDF basiert zudem auf ähnlichen vereinfachenden Annahmen wie zum Beispiel die NB-Klassifikation. Die örtliche Relevanz der Tags wird unabhängig von einem gemeinsamen Auftreten von anderen Tags am selben Ort bewertet.

Das Zufallsverfahren zur Aufteilung der Stichprobe in Trainings- und Validationsdaten hat einen grossen Einfluss auf die Resultate. Es hängt aber von der geforderten Anwendung der Georeferenzierung ab, welches Verfahren realistischer ist.

Mit der Erstellung von Ortsverzeichnisses mit aus UGC gewonnenen räumlichen Informationen zur Georeferenzierung hat diese Arbeit eine Teilziel des GIR adressiert und einen potentiellen Ansatz zur Beantwortung der formulierten Forschungsfragen aufgezeigt. Oder um nochmals auf die von Goodchild & Hill (2008) in der Einleitung erwähnte Suchanfrage zurückzukommen:

> „*find an orange grove five miles north of Bakersfield*“

Falls Flickr-Photos oder andere nutzergenerierte Daten von besagtem Orangenhain inklusive der nötigen Metadaten vorhanden sind, kann die obige Suchanfrage vielleicht bereits mit den in dieser Arbeit gezeigten Methoden beantwortet werden.

7.4 Ausblick

Im Verlaufe dieser Arbeit sind einige Fragen und Unsicherheiten aufgetaucht, die nach einer vertieften Analyse verlangen.

Die Positionsgenauigkeit der Geotags von Flickr-Photos wurde zum Beispiel von Hochmair & Zielstra (2012) untersucht. Allerdings beschränkten sie ihre Analyse auf eine kleine Stichprobe aus urbanen Regionen. Im Zusammenhang mit der zunehmenden Verbreitung von Mobiltelefonen mit integrierter Kamera stellt sich die Frage, ob diese Geotags dieselbe Positionsgenauigkeit aufweisen. Obwohl die meisten Smartphones einen integrierten GPS-Empfänger haben, werden in der Praxis wohl viele Photos mit den durch Funkmasten-Triangulation bestimmten Koordinaten georeferenziert. Die Genauigkeit dieser Positionsbestimmung ist abhängig von der Funkmastendichte (Zogg, 2007). Zur Vermeidung von pseudogenauen Ergebnissen bei der Verwendung von Flickr-Daten müssten deshalb umfangreichere Untersuchungen zur Positionsgenauigkeit der Geotags von Flickr-Photos gemacht werden, damit Richtwerte für eine minimale Granularität der räumlichen Analyseeinheiten bekannt sind.

Im Zusammenhang mit einer Funktion der Flickr-Plattform, die für Photos mit geograpischen Koordinaten beim Hochladen automatisch Vorschläge von Ortsnamen macht (siehe Abbildung A.3), stellt sich die Frage nach dem Einfluss auf das Tagging-Verhalten der Nutzer. Es sollte im Hinblick auf den Gebrauch von umgangsprachlichen Ortsbezeichnungen untersucht werden, ob die durch Flickr vorgeschlagenen Ortsnamen bei den Nutzern zu einer einheitlicheren und weniger vielfältigen Verwendung von Ortsbezeichnungen führen.

Aus den Schwächen und Nachteilen der Methoden in dieser Arbeit im vorherigen Abschnitt 7.3 können Handlungsempfehlungen für die zukünftige Verbesserung der Methoden abgeleitet werden.

Durch die Anwendung von geeigneten Glättungsverfahren (zum Beispiel in Serdyukov et al., 2009 oder O'Hare & Murdock, 2012) könnte das Problem von nicht vorhandenen Daten und daraus entstehenden Leerzellen gelöst werden. Eine andere Möglichkeit, welche die Repräsentativität der ortsrelevanten Tags potentiell erhöhen könnte, ist der Einbezug von weiteren Datenquellen wie beispielsweise *Geograph*. Mit dem Einbezug des gemeinsamen Auftretens von Tags in einer Zelle könnte zudem die Vorhersagegenauigkeit verbessert werden (Sanderson & Croft, 1999; Overell & Rüger,

2008). Van Laere et al. (2010) schlagen die Nutzung von TF-IDF zur Gewichtung der Tag-Verteilung in den Sprachmodellen vor. Ebenfalls möglich ist die höhere Gewichtung von Toponymen (Serdyukov et al., 2009).

Der heterogenen Verteilung der Daten könnte mit einem der Datendichte angepassten Gitternetz entsprochen werden, welches die Granularität der Analyse den örtlich vorhanden Daten anpasst (Roller et al., 2012). Als mögliche Lösung für den Einfluss der Zellgrösse auf die Ergebnisse bietet sich gemäss Rattenbury & Naaman (2009) die Kombination von Resultaten für verschiedene Zellgrössen an.

Bei der Integration von aus verschiedenen UGC-Datenquellen extrahierten geographischen Information zu einem einheitlichen Ortsverzeichnis taucht eine weitere Herausforderung auf. Es sollten Ansätze entwickelt werden, welche die Bewertung der Relevanz von aus verschiedenen Datenquellen stammenden ortsrelevanten Bezeichnungen vereinheitlichen können. Smart et al. (2010) zeigen eine möglichen Ansatz zur Datenintegration von Toponymen aus verschiedenen Ortsverzeichnissen. Das Problem der geometrischen Repräsentation von Ortsverzeichniseinträgen als Punkte könnte zum Beispiel durch eine geeignete automatische Generierung der räumlichen Ausdehnung mit Methoden, wie sie Grothe & Schaab (2009) oder Hollenstein & Purves (2010) zeigen, gelöst werden.

Es sollte erklärtes Ziel aller Ansätze sein, neben hoher Informationsdichte und Genauigkeit auch eine universelle Anwendung mit Daten aus verschiedenen Quellen zu ermöglichen. Dabei sollte die Verarbeitung von grossen Datenmengen, wie sie bei UGC auftreten können, möglichst maschinell und ohne Eingriffe von aussen durchgeführt werden können.

Literaturverzeichnis

Ahern, S., Naaman, M., Nair, R. & Yang, J.H.I. (2007). World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In: Proceedings of the 2007 conference on Digital libraries - JCDL '07, 1–10. ACM Press, New York, New York, USA.

Ames, M. & Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07, 971–980. ACM Press, New York, New York, USA.

Antin, J., Yee, R., Cheshire, C. & Nov, O. (2011). Gender differences in Wikipedia editing. In: Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11, 11–14. ACM Press, New York, New York, USA.

Bird, S., Klein, E. & Loper, E. (2009). Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. O'Reilly Media. http://nltk.org/book/

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.

Cheng, Z., Caverlee, J. & Lee, K. (2010). You are where you tweet. In: Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10, 759–768. ACM Press, New York, New York, USA.

CIA (2012). The World Factbook. Zugriff: 03.12.2012. http://www.cia.gov/library/publications/the-world-factbook/

Cox, A.M. (2008). Flickr: a case study of Web2.0. Aslib Proceedings, 60(5):493–516.

Crandall, D.J., Backstrom, L., Huttenlocher, D. & Kleinberg, J. (2009). Mapping the world's photos. In: Proceedings of the 18th international conference on World wide web - WWW '09, 761. ACM Press, New York, New York, USA.

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.

Dykes, J., Purves, R., Edwardes, A. & Wood, J. (2008). Exploring Volunteered Geographic Information to Describe Place : Visualization of the 'Geograph British Isles' Collection. In: D. Lambrick (Hg.), Proceedings of the GIS Research UK 16th Annual Conference GISRUK, 256–267.

Edwardes, A.J. & Purves, R.S. (2007). A theoretical grounding for semantic descriptions of place. In: J.M. Ware & G.E. Taylor (Hg.), Web and Wireless Geographical Information Systems - Proceedings of 7th Intl. Symposium on Web and Wireless Geographical Information Systems (W2GIS), Bd. 4857 von *Lecture Notes in Computer Science*, 106–120. Springer Berlin Heidelberg.

Egenhofer, M.J. (2002). Toward the semantic geospatial web. In: Proceedings of the tenth ACM international symposium on Advances in geographic information systems - GIS '02, 1–4. ACM Press, New York, New York, USA.

Egenhofer, M.J. & Mark, D.M. (1995). Naive Geography. In: A. Frank & W. Kuhn (Hg.), Spatial Information Theory: A Theoretical Basis for GIS, Bd. 988 von *Lecture Notes in Computer Science*, 1–15. Springer Berlin Heidelberg.

Fake, C. (2005). Yahoo actually does acquire Flickr. Zugriff: 03.12.2012. http://blog.flickr.net/en/2005/03/20/yahoo-actually-does-acquire-flickr/

Fisher, P. & Unwin, D.J. (2005). Re-presenting Geographical Information Systems. In: P. Fisher & D.J. Unwin (Hg.), Re-presenting GIS, 1–14. John Wiley & Sons.

Flickr (2012). API Documentation. Zugriff: 03.12.2012. http://www.flickr.com/services/api/

Flickr Code Blog (2009). 100,000,000 geotagged photos (plus). Zugriff: 03.12.2012. http://code.flickr.com/2009/02/04/100000000-geotagged-photos-plus/

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

Goodchild, M.F. (2009). NeoGeography and the nature of geographic expertise. Journal of Location Based Services, 3(2):82–96.

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

Grothe, C. & Schaab, J. (2009). Automated Footprint Generation from Geotags with Kernel Density Estimation and Support Vector Machines. Spatial Cognition & Computation, 9(3):195–211.

Gruber, T.R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199–220.

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.

Guillén, R. (2008). GeoParsing Web Queries. In: C. Peters, V. Jijkoun, T. Mandl, H. Müller, D.W. Oard, A. Peñas, V. Petras & D. Santos (Hg.), Advances in Multilingual and Multimodal Information Retrieval, Bd. 5152 von *Lecture Notes in Computer Science*, 781–785. Springer Berlin Heidelberg.

Hart, G. & Dolbear, C. (2007). What's So Special about Spatial? In: A. Scharl & K. Tochtermann (Hg.), The Geospatial Web, Advanced Information and Knowledge Processing, 39–44. Springer London.

Heipke, C. (2010). Crowdsourcing geospatial data. ISPRS Journal of Photogrammetry and Remote Sensing, 65(6):550–557.

Hill, L.L. (2006). Georeferencing: The Geographic Associations of Information. The MIT Press.

Hochmair, H.H. & Zielstra, D. (2012). Positional Accuracy of Flickr and Panoramio Images in Europe. In: T. Jekel, A. Car, J. Strobl & G. Griesebner (Hg.), GI_Forum 2012: Geovizualisation, Society and Learning, 14–23. Wichmann.

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

Hu, Y.H. & Ge, L. (2005). A Supervised Machine Learning Approach to Toponym Disambiguation. In: A. Scharl & K. Tochtermann (Hg.), The Geospatial Web, Advanced Information and Knowledge Processing, 117–128. Springer London.

Janowicz, K., Schade, S., Bröring, A., Kessler, C., Maué, P. & Stasch, C. (2010). Semantic Enablement for Spatial Data Infrastructures. Transactions in GIS, 14(2):111–129.

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. (2008a). Modelling vague places with knowledge from the Web. International Journal of Geographical Information Science, 22(10):1045–1065.

Jones, R., Zhang, W.V., Rey, B., Jhala, P. & Stipp, E. (2008b). Geographic intention and modification in web search. International Journal of Geographical Information Science, 22(3):229–246.

Kessler, C., Maué, P., Heuer, J.T. & Bartoschek, T. (2009). Bottom-Up Gazetteers: Learning from the Implicit Semantics of Geotags. In: K. Janowicz, M. Raubal & S. Levashkin (Hg.), GeoSpatial Semantics, Bd. 5892 von *Lecture Notes in Computer Science*, 83–102. Springer, Berlin, Heidelberg.

Kinsella, S., Murdock, V. & O'Hare, N. (2011). "I'm eating a sandwich in Glasgow": modeling locations with tweets. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents - SMUC '11, June, 61–68. ACM Press, New York, New York, USA.

Koukoletsos, T., Haklay, M. & Ellul, C. (2012). Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. Transactions in GIS, 16(4):477–498.

Kremerskothen, K. (2011). Flickr Blog: 6,000,000,000. Zugriff: 28.11.2012. http://blog.flickr.net/en/2011/08/04/6000000000/

Kuhn, W. (2001). Ontologies in support of activities in geographical space. International Journal of Geographical Information Science, 15(7):613–631.

Larson, R.R. (1996). Geographic information retrieval and spatial browsing. In: L.C. Smith & M. Glueck (Hg.), Geographic information systems and libraries: patrons, maps, and spatial information, 81–124.

Leidner, J.L. (2004). Towards a Reference Corpus for Automatic Toponym Resolution Evaluation. In: Proceedings of the Workshop on Geographic Information Retrieval, SIGIR 2004. ACM Press, Sheffield, United Kingdom.

Leveling, J. & Hartrumpf, S. (2008). On metonymy recognition for geographic information retrieval. International Journal of Geographical Information Science, 22(3):289–299.

MacKay, D. (2003). Information Theory, Inference, and Learning Algorithms. Zugriff: 19.01.2013.
http://www.inference.phy.cam.ac.uk/itprnn/book.pdf

Manning, C.D., Raghavan, P. & Schütze, H. (2009). Introduction to Information Retrieval. Cambridge University Press, Cambridge, England.

Montello, D.R., Goodchild, M.F., Gottsegen, J. & Fohl, P. (2003). Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. Spatial Cognition & Computation, 3(2-3):185–204.

Naughton, J. (2008). How Flickr developed into a classic Web 2.0 success. Zugriff: 03.12.2012.
http://www.guardian.co.uk/media/2008/mar/09/web20.internet

Nielsen, J. (2006). Participation Inequality: Encouraging More Users to Contribute. Zugriff: 15.12.2012.
http://www.useit.com/alertbox/participation_inequality.html

Nov, O., Naaman, M. & Ye, C. (2008). What drives content tagging: the case of photos on Flickr. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '08, 1097–1100. ACM Press, New York, New York, USA.

Nov, O., Naaman, M. & Ye, C. (2010). Analysis of participation in an online photo-sharing community: A multidimensional perspective. Journal of the American Society for Information Science and Technology, 61(3):555–566.

Nov, O. & Ye, C. (2010). Why do people tag?: motivations for photo tagging. Communications of the ACM, 53(7):128–131.

O'Hare, N. & Murdock, V. (2012). Modeling locations with social media. Information Retrieval, (April 2012):1–33.

Openshaw, S. (1984). The modifiable areal unit problem. In: Concepts and Techniques in Modern Geography, 38, 1–40. Geo Books, Norwich, England.

OpenStreetMap Foundation (2012). OpenStreetMap – Die freie Wiki-Weltkarte. Zugriff: 17.12.2012.
http://www.openstreetmap.org/

Ordnance Survey (2012). 1:50 000 Scale Gazetteer. Zugriff: 10.12.2012.
http://www.ordnancesurvey.co.uk/oswebsite/products/50k-gazetteer/index.html

Overell, S. & Rüger, S. (2008). Using co-occurrence models for placename disambiguation. International Journal of Geographical Information Science, 22(3):265–287.

Purves, R.S., Clough, P. & Joho, H. (2005). Identifying imprecise regions for geographic information retrieval using the web. In: J. Drummond, R. Billen, D. Forrest & E.M. Joao (Hg.), Proceedings of the GIS Research UK Conference, 130–135.

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.

Purves, R.S., Edwardes, A. & Wood, J. (2011). Describing place through user generated content. First Monday [online], 16(9).
http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3710/3035

Rattenbury, T., Good, N. & Naaman, M. (2007). Towards automatic extraction of event and place semantics from flickr tags. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07, 103–110. ACM Press, New York, New York, USA.

Rattenbury, T. & Naaman, M. (2009). Methods for extracting place semantics from Flickr tags. ACM Transactions on the Web, 3(1):1–30.

Roller, S., Speriosu, M., Wing, B. & Baldridge, J. (2012). Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 1500–1510. Association for Computational Linguistics.

Rorissa, A. (2010). A comparative study of Flickr tags and index terms in a general image collection. Journal of the American Society for Information Science and Technology, 61(11):2230–2242.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5):513–523.

Sanderson, M. & Croft, B. (1999). Deriving concept hierarchies from text. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99, 206–213. ACM Press, New York, New York, USA.

Sanderson, M. & Kohler, J. (2004). Analyzing geographic queries. In: Proceedings of the 2004 Workshop on Geographic Information Retrieval - SIGIR '04.

Schmitz, P. (2006). Inducing Ontology from Flickr Tags. In: Proceedings of the Collaborative Web Tagging Workshop - WWW '06. Edinburgh, Scotland.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47.

Serdyukov, P., Murdock, V. & van Zwol, R. (2009). Placing flickr photos on a map. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09, 484. ACM Press, New York, New York, USA.

Sigurbjörnsson, B. & van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In: Proceeding of the 17th international conference on World Wide Web - WWW '08, 327. ACM Press, New York, New York, USA.

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. Kreveld & C. Schlieder (Hg.), Geographic Information Science, Bd. 6292 von *Lecture Notes in Computer Science*, 234–248. Springer, Berlin, Heidelberg.

Smeulders, A., Worring, M., Santini, S., Gupta, A. & Jain, R. (2000). Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349–1380.

Smith, B. & Mark, D.M. (2001). Geographical categories: an ontological investigation. International Journal of Geographical Information Science, 15(7):591–612.

Smith, D.A. & Crane, G. (2001). Disambiguating Geographic Names in a Historical Digital Library. In: P. Constantopoulos & I.T. Sø lvberg (Hg.), Research and Advanced Technology for Digital Libraries, Bd. 2163 von *Lecture Notes in Computer Science*, 127–136. Springer Berlin Heidelberg.

Sui, D. & Goodchild, M. (2011). The convergence of GIS and social media: challenges for GIScience. International Journal of Geographical Information Science, 25(11):1737–1748.

Swisstopo (2008). SwissNames 25. Zugriff: 24.12.2012.
http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html

Van Laere, O., Schockaert, S. & Dhoedt, B. (2010). Towards automated georeferencing of Flickr photos. In: Proceedings of the 6th Workshop on Geographic Information Retrieval - GIR '10, 1–7. ACM, New York, New York, USA.

Wing, B. & Baldridge, J. (2011). Simple Supervised Document Geolocation with Geodesic Grids. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 955–694. Association for Computational Linguistics, Portland, Oregon, USA.

Worldatlas (2012). Worldatlas. Zugriff: 13.01.2013.
http://www.worldatlas.com/

Zogg, J.M. (2007). Standortbestimmung im Schweizer Mobilfunknetz. Zugriff: 22.01.2013.
http://www.zogg-jm.ch/Dateien/Standortbestimmung_im_Schweizer_Mobilfunknetz.pdf

A Anhang

A.1 Liste von manuell entfernten Stoppwörtern

Tab. A.1: Manuell entfernte Schlüsselwörter

35mm	macro
50mm	nikon
auto-upload	nikon d90
b&w	panasonic
black and white	photo
bw	photography
cameraphone	photos
canon	portrait
canonpowershots3is	shozu
d90	sony
dslr	square[1]
ef-s18-55mm f/3.5-5.6 is	square format[1]
eos	st
foursquare:	uploaded:by=instagram[1]
fuji hs10	2003
geo:lat	2004
geo:lon	2005
geotagged	2006
hdr	2007
hipstamatic	2008
instagram app[1]	2009
iphone	2010
iphoneography[1]	2011
lumix	2012

[1] *Instagram* fügt allen damit geschossenen Bildern automatisch die gekennzeichneten Tags hinzu. Sie werden nur falls vollständig vorhanden entfernt.

A.2 Naive Bayes-Klassifikator

Die folgenden Erklärungen basieren auf dem NB-Klassifikator in Manning et al. (2009).

Die Gleichung A.1 aus Abschnitt 4.5.1 schätzt für ein Photo p die wahrscheinlichste Zelle c mit der höchsten Wahrscheinlichkeit $P(c|p)$. Diese berechnet sich aus der allgemeinen Wahrscheinlichkeitsverteilung aller Photos in den Trainingsdaten $\hat{P}(c)$ und der bedingten Vorkommenswahrscheinlichkeit $\hat{P}(t_k|c)$ eines Photos in einer bestimmten Zelle aufgrund der darin enthaltenen Tags t_k.

$$c = \arg\max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_p} \log \hat{P}(t_k|c)] \tag{A.1}$$

Der Ausdruck $\hat{P}(c)$ steht für die A-priori-Wahrscheinlichkeit eines Photos in einer Zelle c aufgrund der relativen Häufigkeit der Photos. N_c ist die Anzahl Photos in der Zelle c und N die gesamte Anzahl an Photos im Trainingsdatensatz.

$$\hat{P}(c) = \frac{N_c}{N} \tag{A.2}$$

Jede bedingte Wahrscheinlichkeit $\hat{P}(t_k|c)$ zeigt, wie gut ein Tag t_k die Zugehörigkeit zu einer Zelle c beschreibt und kann aus der relativen Häufigkeit von t_k in allen Photos der Zelle c geschätzt werden.

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \tag{A.3}$$

T_{ct} ist die Häufigkeit eines Tags t in Photos der Zelle c in den Trainingsdaten, während $\sum_{t' \in V} T_{ct'}$ die Summe der Häufigkeiten aller Tags t' in Photos der Zelle c in den Trainingsdaten ist. V ist dabei die Gesamtheit aller Tags.

Problematisch ist das Auftreten von Nullwahrscheinlichkeiten für Tags, die in den Trainingsdaten nicht in einer bestimmten Zelle vorkommen, da die bedingten Wahrscheinlichkeiten in Gleichung A.1 logarithmiert werden und $\log x$ für $x \leq 0$ nicht definiert ist. Deshalb wird durch *Laplace*-Glättung die Gleichung A.3 angepasst, in dem zu allen Häufigkeiten eins addiert wird. Die Formel lautet dann ergänzt:

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} T_{ct'} + 1} \tag{A.4}$$

A.3 Hierarchie der Ortsverzeichnis-Kategorien

Tab. A.2: Hierarchie der Ortsverzeichnis-Kategorien

Schweiz			Grossbritannien	
Rang	Kategorie	Anzahl Einträge	Kategorie	Anzahl Einträge
1	Stadt	10	Stadt	65
2	Grosse Gemeinde	117	Ortschaft	1'255
3	Mittlere Gemeinde	678	Hügel, Berg	14'509
4	Kleine Gemeinde	1993	Andere Orte	41'297
5	Grosse Ortschaft	112	Gewässer	24'425
6	Mittlere Ortschaft	1987	Wald	8'714
7	Kleine Ortschaft	2849	Bauernhof	34'692
8	Weiler	10697	Alle anderen Objekte	128'417
9	Streusiedlung	1424	Antike Objekte (nicht-römisch)	4'890
10	Einzelhaus	42456	Antike Objekte (römisch)	223
11	Hütte	771		
12	Haupttal	178		
13	Nebental	2046		
14	Flurname	54980		
15	Massiv	143		
16	Gipfel	165		
17	Gipfel	866		
18	Gipfel	4414		
19	Grat	1440		
20	Hügel	2543		
21	Graben	2628		
22	Grosser See	53		
23	Kleiner See	817		
24	Stausee	83		
25	Fluss	399		
26	Bach	1004		
27	Bächlein	3960		
28	Weiher	101		
29	Wasserfall	52		
30	Quelle	69		
31	Wald	7093		
32	Sumpf	191		
33	Gebiet	87		
34	Fels	1998		
35	Erratischer Block	185		
36	Gletscher	730		
37	Höhle	80		
38	Strasse	44		
39	Weg	197		
40	Fusspass	1898		
41	Strassenpass	102		
42	Tunnel	60		
43	Bahnhof	464		
44	Flugplatz	68		
45	Industrie	1018		
46	Sportanlage	401		
47	Öffentliche Gebäude	648		
48	Ruine	603		
49	Brücke	309		
50	Historischer Ort	226		
51	Kirche	459		
52	Schloss	401		
53	Staumauer	5		
54	Friedhof	19		
55	Park	39		
56	Hafen	24		
57	Turm	24		
58	Hotel	61		
59	Brunnen	86		
60	Denkmal	59		
61	Zoll	54		
Total		156755		258487

A.4 Räumliche Verteilung der Toponyme in Ortsverzeichnissen

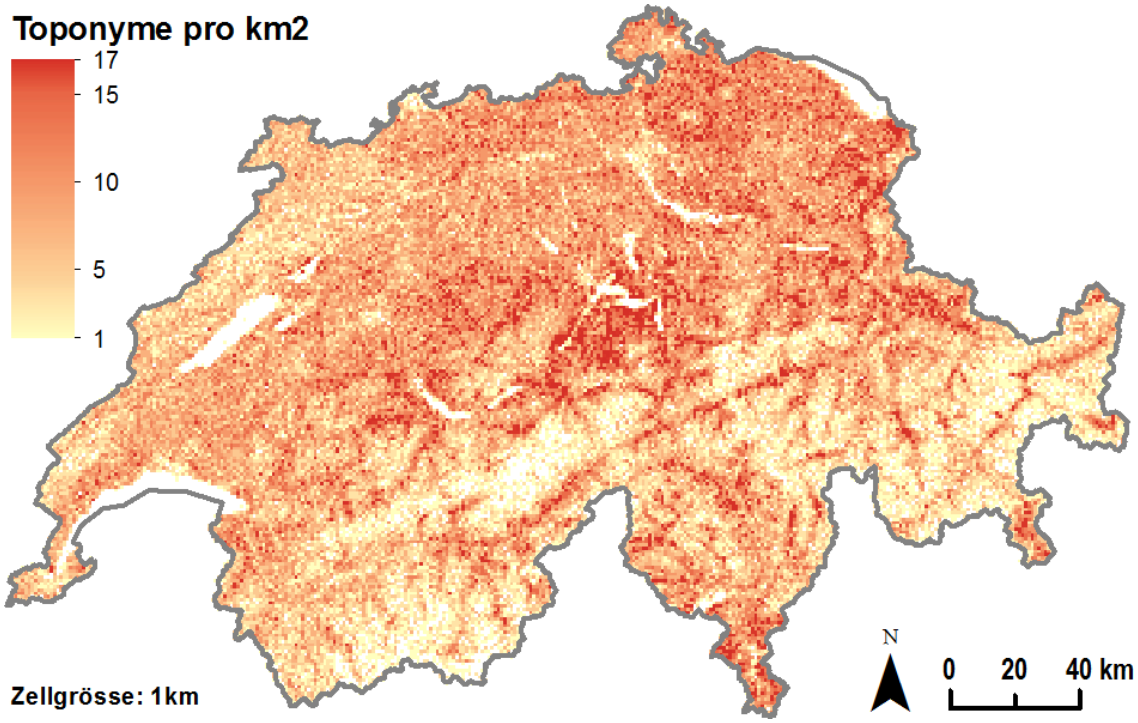


Abb. A.1: Anzahl Toponyme pro km^2 in der Schweiz (Daten: Swisstopo)

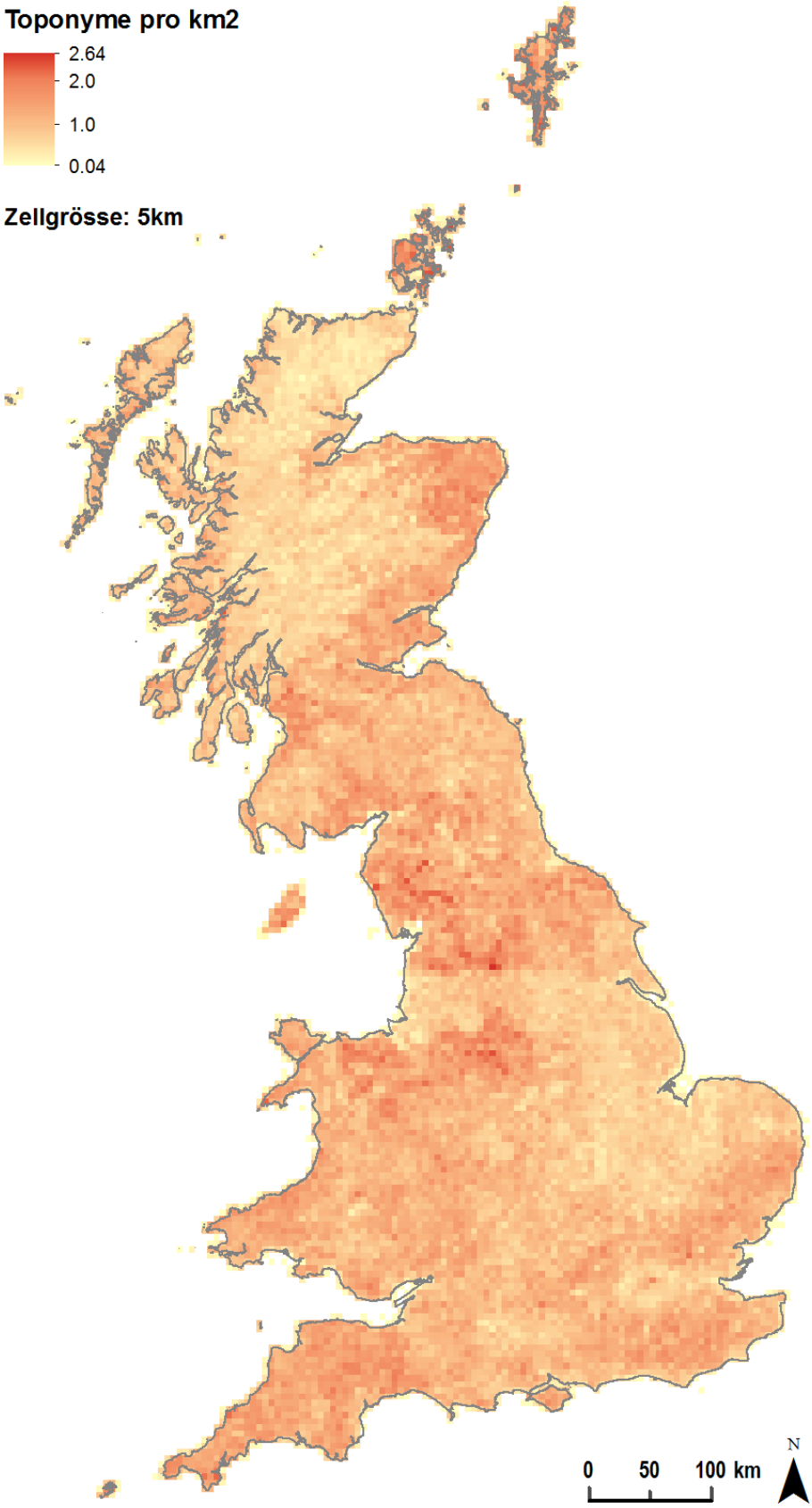


Abb. A.2: Anzahl Toponyme pro km^2 in Grossbritannien (Daten: Ordnance Survey)

A.5 Vergleich der 100 häufigsten Tags

Tab. A.3: Die 100 häufigsten Tags auf Flickr in Grossbritannien 2011

Photos mit Geotag				Photos ohne Geotag			
1	london	51	essex	1	uk	51	film
2	england	52	train	2	england	52	autumn
3	uk	53	car	3	scotland	53	night
4	united kingdom	54	station	4	london	54	bird
5	scotland	55	sign	5	united kingdom	55	bridge
6	britain	56	summer	6	britain	56	flowers
7	street	57	flowers	7	great britain	57	railway
8	water	58	museum	8	nature	58	colour
9	bus	59	castle	9	water	59	old
10	great britain	60	birmingham	10	landscape	60	castle
11	sky	61	road	11	gb	61	flower
12	city	62	kent	12	europe	62	car
13	ireland	63	reflection	13	sky	63	wales
14	architecture	64	bird	14	british	64	scottish
15	gb	65	spring	15	street	65	sun
16	blue	66	old	16	sea	66	train
17	red	67	hampshire	17	edinburgh	67	sussex
18	nature	68	autumn	18	city	68	brighton
19	sea	69	bristol	19	architecture	69	great
20	edinburgh	70	gbr	20	beach	70	graffiti
21	church	71	english	21	blue	71	museum
22	landscape	72	sun	22	summer	72	countryside
23	white	73	yellow	23	white	73	north
24	night	74	brighton	24	green	74	yellow
25	green	75	winter	25	clouds	75	highlands
26	river	76	365	26	english	76	reflection
27	manchester	77	graffiti	27	trees	77	dorset
28	building	78	wildlife	28	river	78	town
29	wales	79	glasgow	29	red	79	east
30	beach	80	colour	30	urban	80	devon
31	europe	81	volvo	31	art	81	grass
32	trees	82	square	32	people	82	cornwall
33	park	83	coast	33	black	83	west
34	railway	84	town	34	coast	84	snow
35	urban	85	window	35	yorkshire	85	house
36	tree	86	tower	36	sunset	86	boat
37	light	87	transport	37	park	87	road
38	black	88	olympus	38	light	88	history
39	dublin	89	united	39	united	89	seaside
40	british	90	man	40	winter	90	birds
41	clouds	91	thames	41	tree	91	londres
42	people	92	liverpool	42	spring	92	manchester
43	art	93	house	43	kingdom	93	hampshire
44	bridge	94	iphoneography	44	wildlife	94	thames
45	sunset	95	grass	45	building	95	gbr
46	pub	96	devon	46	church	96	essex
47	yorkshire	97	travel	47	glasgow	97	girl
48	film	98	sculpture	48	kent	98	station
49	garden	99	candid	49	travel	99	cumbria
50	flower	100	dorset	50	garden	100	man

A.6 Toponyme in den 500 häufigsten Tags

Tab. A.4: Toponyme in den 500 häufigsten Tags in Grossbritannien im Jahr 2011

Photos mit Geotag		Photos ohne Geotag	
park	york	**park**	wales
ford	sheffield	**ford**	seaside
hill	fence	preston	loch
stone	island	**hill**	island
lake	shoreditch	**stone**	york
highlands	palace	**highlands**	bath
green	college	**lake**	fence
rock	steps	green	leeds
street	london	rock	aberdeen
scotland	sea	street	sheffield
cross	edinburgh	scotland	palace
ireland	church	woodland	portsmouth
castle	manchester	cliff	valley
woods	tree	cross	raw
bank	british	castle	london
trees	clouds	woods	british
bridge	sunset	trees	sea
sand	birmingham	bridge	edinburgh
field	kent	sand	clouds
river thames	old	field	coast
beach	bristol	river thames	sunset
garden	glasgow	isle	tree
tower	coast	beach	kingdom
wall	liverpool	garden	church
cambridge	devon	tower	glasgow
wood	travel	wall	kent
beer	kingdom	wood	travel
abbey	westminster	cambridge	old
brighton	rocks	abbey	devon
oxford	bournemouth	beer	manchester
shop	leyland	brighton	birmingham
harbour	headstone	oxford	rocks
march	irish	harbour	bristol
nottingham	southampton	mountain	mountains
greenwich	norwich	march	liverpool
newcastle	gallery	hills	westminster
forest	hall	forest	gallery
mountain	silverstone	shop	tour
midlands	mountains	midlands	hall
eye	isle of wight	highland	waterfall
victoria	bench	shore	fields
water	eyes	bay	london eye
city	fountain	france	outside
river	moon	eye	bournemouth
wales	aldershot	greenwich	dawn
dublin	lost	argyll	lincoln
seaside	coventry	nottingham	bench
portsmouth	rose	newcastle	
bath	london eye	water	
raw	waterfall	city	
leeds		river	

A.7 Automatische Vorschläge von Ortsnamen für Photos

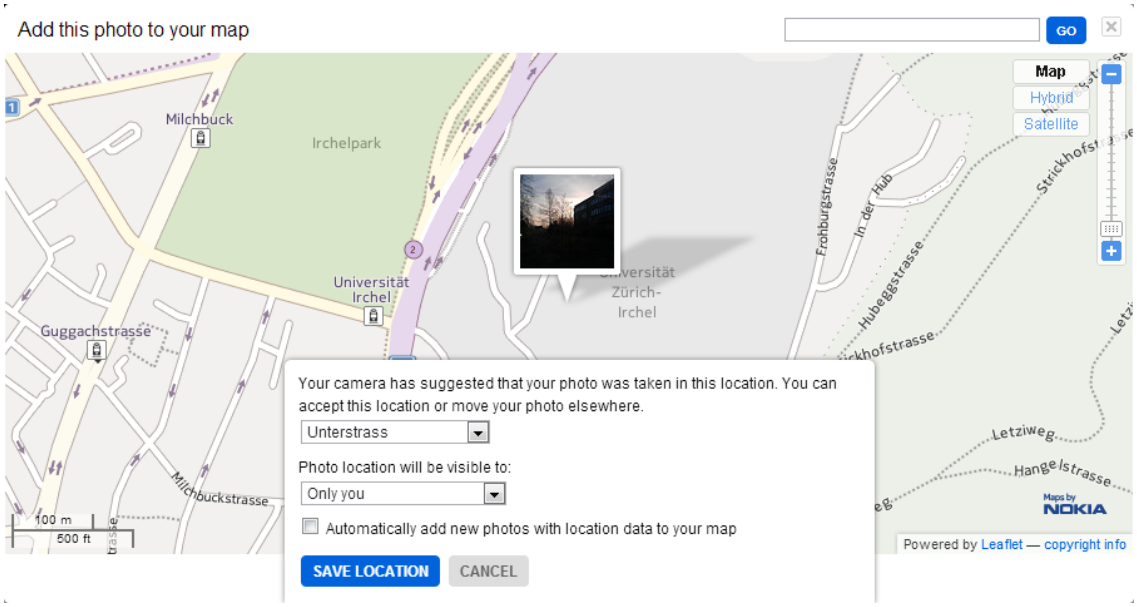


Abb. A.3: Automatische Vorschläge von Ortsnamen beim Upload von Photos mit geographischen Koordinaten (Bild: Flickr)

Persönliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und die den verwendeten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Zürich, 31. Januar 2013

Thomas Wider