

Geographische Migrationsforschung auf Basis von Fernerkundung und Social-Media-Daten

Inaugural-Dissertation zur Erlangung der Doktorwürde der
Philosophischen Fakultät
der
Julius-Maximilians-Universität Würzburg

vorgelegt von

Johannes Mast

aus
Würzburg

Würzburg

2026

Erstgutachter: Professor Dr. Hannes Taubenböck

Zweitgutachterin: Professorin Dr. Carolin Biewer

Tag des Kolloquiums: 29. April 2026

Danksagungen

Diese Arbeit, mit der sich diese Dissertation befasst, wurde im Laufe mehrerer Jahre durchgeführt, die für mich sehr glückliche und erfüllte Jahre waren. Die Anstrengungen und Herausforderungen, die es durchaus gab, wurden nie überwältigend, da ich das Glück hatte, auf den Rückhalt vieler beeindruckende Menschen zählen zu können. An erster Stelle danke ich Hannes Taubenböck, meinem Doktorvater und Mentor. Er hat mich nicht nur wissenschaftlich begleitet, sondern mir von Anfang an die Freiheit gegeben, meine eigenen Ideen in interessante und relevante Projekte einzubringen. Besonders dankbar bin ich ihm für das Vertrauen, den Respekt und die Menschlichkeit, die er mir entgegengebracht hat, sowie für seine ansteckende Begeisterung für unsere Forschungsrichtung. Carolin Biewer danke ich herzlich für ihre fachkundige Anleitung bei der Nutzung sprachwissenschaftlicher Methoden und für die inspirierenden Gespräche, bei denen die Zeit immerzu verflogen ist. Andreas Hotho danke ich für seine stets treffenden und konstruktiven Kommentare während unserer lehrreichen Diskussionen, die ich sicher und gerne noch viele Stunden hätte führen können. Sie haben die Dissertation deutlich geschärft und verbessert.

Eine herausragende Mentorin im Rahmen der Dissertation war Marta Sapena. Ihre professionelle und strukturierte Herangehensweise an die wissenschaftliche Arbeit war mir während unserer Zusammenarbeit ein Vorbild und wird es auch in Zukunft bleiben. Sie und Elisabeth Schöpfer gaben mir durch Rat und Feedback immer wieder Orientierung in den Prozessen des akademischen Lebens, und sie beeindruckten mich durch ihre Energie und Freundlichkeit auch in schwierigen Zeiten.

Für die Leitung unserer Forschungsgruppe danke ich Christian Geiß, der mir durch Ermutigung, Rat und Unterstützung während der letzten fünf Jahre den Rücken gestärkt hat. Die gute Atmosphäre innerhalb unserer Gruppe war eine wichtige Grundlage für gute Zusammenarbeit und gegenseitige Unterstützung im Forschungsalltag. Ich danke auch den anderen Mitgliedern unseres Teams sowie der gesamten Abteilung Georisiken und Zivile Sicherheit und ihrem langjährigen Leiter, Günter Strunz, unter dessen Leitung ich 2018 meine Tätigkeit am Deutschen Fernerkundungsdatenzentrum aufgenommen habe. Fabian Henkel, Martin Mühlbauer und Ralph Kiefl unterstützten mich in meiner Arbeit mit der essenziellen technischen Infrastruktur sowie wichtigen Daten. Danke an Mihaela Blaskovic, welche durch ihre Unterstützung in administrativen Angelegenheiten grundlegend zur Entstehung dieser Arbeit beigetragen hat, sowie den Mitarbeitenden der Nebenkostenstelle für ihre Geduld bei der Bearbeitung meiner Reisekostenabrechnungen. Allen Mitgliedern des Deutschen Fernerkundungsdatenzentrums gebührt mein Dank für ihren Einsatz zu vielfältigsten Themenbereichen. Bis heute inspiriert und begeistert mich die Bandbreite der Themen und Methoden, die wir innerhalb des Instituts verbinden und auch über den Tellerrand der Forschung hinaus zur Anwendung bringen.

Ich danke dem terrabyte-Team des Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften und des DLR für die Bereitstellung der erstklassigen Recheninfrastruktur. Zudem danke ich Twitter Inc. (seit 2023 X Corp.) für die Bereitschaft, mir die Nutzung der API für diese Forschung zu gewähren. Ermöglicht wurde diese Arbeit durch die dankenswerte Unterstützung der Europäischen Kommission über die Finanzierung des FPCUP Projektes

und des BMBF für die Finanzierung des MIGRAWARE Projektes. Ich danke Justice Nana Inkoom für die Organisation dieses Projektes und des Austauschs unter den erschwerten Bedingungen der Covid-19 Pandemie. Herzlichen Dank auch an Stefan Voigt, der mir die Möglichkeit eröffnete, die Ergebnisse und Kenntnisse meiner Dissertation auch in das OpenWebSearch Projekt einzubringen.

Besonderen Dank für ihre sehr konkreten Ratschläge zu meiner Dissertation gilt Johannes Löw, Lisa Lehnert und Vanessa Rittlinger.

Neben dem DLR war mein zweiter Arbeitsort das Earth Observation Research Cluster der Universität Würzburg. Ich danke allen, die die Räumlichkeiten dort durch ihre großen und kleinen Beiträge nicht nur mit stetiger Forschungsaktivität, sondern auch mit Energie und Leben erfüllen. Eine besonders wichtige Gruppe waren für mich die anderen Studierenden des EAGLE-Programms sowie die anderen Doktoranden des EORC. Ich konnte stets darauf zählen, bei ihnen eine vertraute und positive Atmosphäre vorzufinden, sowie die Bereitschaft, Unterhaltungen über faszinierende Forschungsthemen und andere Aspekte des Lebens zu führen. Dieser Austausch erlaubte es mir, meine eigene Arbeit in Perspektive zu sehen, und Herausforderungen gelassener anzugehen. Danke, dass ich ein Teil eurer Gruppe sein konnte.

Wichtige Partner bei der Entwicklung und Umsetzung von Forschungsideen waren insbesondere Richard Lemoine Rodríguez und sein Bürokollege Henri Debray, welche regelmäßig nicht nur große wissenschaftliche Expertise, sondern vermutlich auch gewisse Geduld demonstrierten, wenn ich ungefragt in ihrem Büro auftauchte und sie mit Diskussionen und Einfällen von ihrer Arbeit ablenkte. Auch über unsere akademische Zusammenarbeit hinaus gab mir ihre Freundschaft Gelassenheit, Freude und Inspiration.

Das wichtigste Fundament für Ruhe und Zuversicht war für mich immer meine Familie: Mein Papa Gebhard, meine Mama Elisabeth und mein Bruder Benedikt. Ich zweifle nie daran, dass ich mich auf sie verlassen kann. Sie haben Verständnis, wenn ich länger nicht zuhause bin, und sind gleichzeitig immer für mich da, wenn ich etwas brauche oder sprechen möchte.

Ich möchte auch Livia Koenitz nicht vergessen. Von ganzem Herzen danke ich ihr für ihren Rückhalt und das Glück, das sie mir in diesen Jahren geschenkt hat. Sie bereicherte meine Dissertation durch ihre Expertise in persönlichen Aspekten menschlicher Migration, mein Leben durch ihre Herzlichkeit, und meine Welt durch die bewundernswerte Fähigkeit, hinter das Offensichtliche zu blicken. Für jemanden, der beruflich die Oberfläche erforscht, ist das von unschätzbarem Wert.

Ich danke allen diesen Menschen, die mich direkt unterstützt haben. Mein Dank gilt zudem all jenen, die unsere Forschung lesen, und der Gesellschaft, die diese Arbeit ermöglicht und trägt. Ich empfinde es als großes Privileg, mich der Erforschung solch spannender und zukunftsweisender Themen widmen zu dürfen. Vielen Dank!

Zusammenfassung

Menschliche Migration ist ein komplexes Phänomen das unsere Welt nachhaltig prägt. Es beschreibt die längerfristige Bewegung von Menschen von einem Ort zum anderen, oft auf der Suche nach besseren Lebensbedingungen oder aufgrund von Umweltveränderungen. Sowohl unsere Umwelt als auch die Gesellschaft interagieren permanent und nachhaltig mit den Bewegungen von Menschen. Städte spielen eine zentrale Rolle in diesem Prozess. Sie stellen nicht nur das Ziel der meisten Migrierten dar, auch die Menschheit als Ganzes befindet sich auf einer langfristigen Wanderung von vorwiegend ländlichen in vorwiegend urbane Regionen. Im Jahr 2025 sind die Ausmaße der Migration größer denn je, und insbesondere in Afrika schreitet das urbane Bevölkerungswachstum in großem Tempo voran.

Für das Verständnis unserer Welt und ihres globalen Wandels ist es essenziell, diese wichtigen Prozesse zu ergründen. Die dafür notwendige geographische Forschung erfordert die großflächige und konsistente Erfassung von Migrationsfaktoren und -Effekten wie Umweltveränderungen, Naturgefahren, Städtewachstum, politischer Instabilität, Sicherheitsempfinden und wirtschaftlicher Leistungsfähigkeit. Fernerkundungsdaten von modernen satellitengestützten Erdbeobachtungssystemen können bereits viele physische Migrationsfaktoren oder -effekte in großem Stil und konsistent messen. Jedoch sind die gleichsam wichtigen soziokulturellen und individuellen Migrationsfaktoren nicht oder nur indirekt mit Bilddaten messbar. In dieser Hinsicht besteht eine Datenlücke, die u.a. durch Webdaten und speziell durch georeferenzierte Daten Sozialer Medien (GSM) reduziert werden könnte. Diese Daten sind größtenteils von Nutzenden auf öffentlichen Plattformen erstellte Textdaten, die mittels Geotags oder durch Geoparsing verortet und damit räumlich als Geodaten ausgewertet werden können. Eine gemeinsame Auswertung mit Fernerkundungsdaten im Kontext menschlicher Migrationsfaktoren ist daher vielversprechend, aber bislang noch nicht direkt erprobt.

Vor diesem Hintergrund befasst sich diese Dissertation damit, den Beweis dafür zu erbringen, dass eine solche Integration der Datenquellen möglich und sinnvoll ist. Dabei werden Potenziale von GSM erörtert und kritisch in Bezug auf Abdeckung und Bias hinterfragt. Aufgrund der Dynamik des Forschungsgegenstandes und der interdisziplinären Natur der Migrationsforschung werden Methoden entwickelt, welche flexibel und modular sind. In dieser Hinsicht setzt sich die Dissertation von anderen Forschungssträngen ab, welche die Optimierung von Prozessketten für die Erfassung spezifischer Zielgrößen wie beispielsweise Bevölkerungsdichte oder Wirtschaftsleistung anstreben.

Im Rahmen der Dissertation werden sechs Studien durchgeführt, welche sich drei thematischen Teilen zuweisen lassen:

Teil A befasst sich mit der Verbindung von Fernerkundung und GSM im urbanen Raum. Studie A1 verbindet Baujahrkartierungen mit georeferenzierten Texten der Plattform Twitter. Dadurch kann belegt werden, dass in vielen Städten Afrikas eine bislang unbekannte digitale Disparität vorliegt, welche sich in einer höheren Dichte von Geotweets in älteren Stadtteilen äußert. Ferner kann belegt werden, dass diese Disparität im Zusammenhang mit dem Grad der Geplantheit und der Struktur der Siedlungen liegt, auch wenn die unterliegenden Kausalitäten in tiefergehenden Studien noch herausgearbeitet werden müssen. Für solch tiefergehende Studien im städtischen Raum leistet Studie A2 einen Beitrag, indem sie eine neuartige Methode zur Unterteilung von städtischem Raum in interpretierbare Zonen demonstriert. Durch Umwandlung von Text-Embeddings in ein Rasterformat können GSM gemeinsam mit Fernerkundungsdaten in räumlichen Clusterungs-Verfahren genutzt werden, um intra-urbane Raumeinheiten zu erstellen, welche sowohl physische als auch textinhaltliche Kohärenz aufweisen, und ähnlich wie Blocks, Nachbarschaften oder Bezirke interpretiert werden. Durch Aggregation von Daten auf Ebene

dieser Raumeinheiten kann die Visualisierung von Daten und interaktive Anwendungen in Mixed-Method Studien unterstützt, und die Kommunikation von Inhalten und Ergebnissen verbessert werden.

Teil B testet die Ableitung von Mobilitätscharakteristika aus GSM und demonstriert deren Anwendbarkeit im Zusammenhang mit Textinhalten. Zahlreiche Studien befassen sich mit der Ableitung von Mobilitätsinformationen aus digitalen Spuren von Internetnutzenden. Die gemeinsame Auswertung von Mobilitätsinformationen und Textinhalten lässt im Migrationskontext jedoch noch bislang unausgeschöpftes Potenzial. Daher erbringt Studie B1 den Beweis, dass aus GSM abgeleitete Mobilitätstypen von Nutzenden in Bezug zu den von ihnen produzierten Textinhalten gesetzt werden können. Für Twitter-Nutzende mit nigerianischem Hintergrund zeigt sich, dass insbesondere internationale Mobilität mit einem höheren Interesse an bestimmten Themen wie beispielsweise Unterhaltung und Technologie einhergeht. Ferner zeigt Studie B2, dass auch lexikalische Merkmale lokaler Sprache algorithmisch identifiziert und in Bezug zu Mobilitätsinformationen der GSM-Nutzenden gesetzt werden können. Hierbei zeigt sich insbesondere eine Tendenz, dass die Nutzung nigerianisch-lokaler lexikaler Elemente des Englischen bei Emigrierten seltener ist als bei Nigerianern in Nigeria. Sowohl Mobilität als auch Sprachvariation stehen in Verbindung mit den Themen des Textes. Diese Ergebnisse bieten neue Einblicke in die Beziehung zwischen Sprache, Mobilität und Kultur, welche künftig mit Methoden der Sprachwissenschaft und Kulturwissenschaften vertieft werden können.

Teil C ergründet aufgrund der Verbreitung von Themen als Analysestruktur mögliche inhaltliche Verzerrungen, welchen GSM bei der Nutzung von Themen unterliegen. Die Tendenzen von Themen, geographische Referenzen zu enthalten, wird in Studie C1 ergründet. Über einen großen Web-Korpus von Twitter, Reddit, Nairaland, Stackexchange, Web-Nachrichten und amerikanischen Journalen hinweg kann belegt werden, dass das Thema der Texte einen Einfluss auf die Wahrscheinlichkeit der Texte hat, verortbare Ortserwähnungen zu enthalten. Studie C2 demonstriert ferner, dass die Einflüsse von Themen sich unterscheiden können, je nachdem ob Geotags oder Geoparsing-Ansätze zur Verortung der Textinhalte genutzt werden.

Übergreifend zeigt sich, dass zahlreiche Methoden zur algorithmischen Auswertung von digitalen Inhalten auch im Migrationskontext eingesetzt werden können. Insbesondere moderne Methoden des maschinellen Lernens ermöglichen die algorithmische Auswertung von großen Mengen an Textinhalten. Die hierdurch erzeugten Ergebnisse können räumlich in Verbindung zu Mobilitätscharakteristika und zu Fernerkundungsdaten gesetzt werden, um einen Mehrwert für geographische Forschung zu generieren. Es ist jedoch wichtig, die Limitationen dieser Methoden und der Daten zu berücksichtigen. Die geographische Auswertung von Textinhalten ist, wie Studien C1 und C2 belegen, von spezifischen Verzerrungen geprägt, die zusätzlich zu den gut dokumentierten demographischen Verzerrungen sozialer Medien auftreten. Die Repräsentativität von GSM-Daten muss daher kritisch hinterfragt werden und stellt, zusammen mit Problemen des Datenzugangs die wesentliche Limitation der in der Dissertation ergründeten Ansätze dar. Aus diesen Gründen können traditionelle Methoden wie in-situ Umfragen, Zensusdaten und qualitative Evaluationen keineswegs durch die in dieser Dissertation demonstrierten Methoden ersetzt werden. Vielmehr kann durch das Zusammenspiel dieser Methoden eine ergänzende und damit umfassendere und vollständigere Erfassung von Migration und Migrationsfaktoren gelingen. Die Perspektiven von Fernerkundung und Sozialen Medien erweisen sich in dieser Hinsicht als hochgradig komplementär. Ihre Verbindung kann, trotz der Herausforderungen, wertvolle Beiträge zur Erforschung von globaler Migration und Urbanisierung leisten.

Abstract

Human migration is a phenomenon of great significance for our world. Both our environment and society interact constantly with the movements of people. Among environments, cities are especially relevant as they are the destination of most migrants. Further, humanity as a whole is on a long-term shift from predominantly rural to predominantly urban regions. In 2025, the scale of migration is greater than ever, and urban population growth in Africa in particular is progressing at a rapid pace.

When researching our world and its global change, it is essential to take these important processes into account. The necessary geographical research requires the large-scale measurement of migration factors. While remote sensing data can measure many physical aspects on a large scale and consistently, the socio-economic, cultural and individual migration factors can only be measured indirectly with image data. In this respect, there is a data gap that could be filled by web data and specifically by geo-referenced social media data (GSM). This data is largely text data that can be localized using geotags or geoparsing and thus spatially analyzed as geodata. A joint analysis with remote sensing data for the purpose of mapping of human migration factors has not yet been carried out.

Against this background, this dissertation aims to provide evidence that such a joint analysis is possible and useful. The potential of GSM is discussed and critically scrutinized with regard to coverage and bias. Due to the dynamic nature of the research subject and the interdisciplinary nature of migration research, methods are developed that are flexible and modular. In this respect, the dissertation sets itself apart from the optimization of processes for inferring specific characteristics. Six studies are carried out as part of the dissertation, which can be divided into three thematic parts:

Part A deals with the combination of remote sensing and GSM in urban areas. Study A1 combines remotely sensed mappings of building construction years with geo-referenced texts from the platform Twitter. This reveals a hitherto unknown digital disparity in many African cities, which manifests itself in a higher density of geotweets in older neighborhoods. Furthermore, it can be shown that this disparity is related to the degree of planning and the structure of the settlements, even if the underlying causalities still need to be worked out in more in-depth studies. Study A2 contributes to such in-depth studies in urban space by demonstrating a novel method for subdividing urban space into study areas. By rasterizing text embeddings, GSM can be used together with remote sensing data in spatial clustering procedures to create spatial units that show both physical and textual coherence and can be interpreted as a variation of neighborhoods, blocks, or urban districts. Aggregate statistics at the level of these units can support data visualization and interactive applications in mixed-method studies and improve the communication of content and results.

Part B tests the derivation of mobility characteristics from GSM and demonstrates their applicability in the context of text content. Numerous studies deal with the derivation of mobility information from digital traces of Internet users. However, the joint evaluation of mobility information and text content still leaves great and as yet untapped potential in the context of migration. Study B1 therefore proves that mobility types derived from GSM can be linked to information about the text content. For Twitter users with a Nigerian background, this combination shows that international mobility in particular is associated with a higher interest in certain topics such as entertainment and technology. Furthermore, study B2 demonstrates that lexical features of local language can also be identified algorithmically and related to mobility information of GSM users. In particular, it reveals a tendency that the use of Nigerian-local lexical items of English is less frequent among emigrants than among Nigerians in Nigeria, and that both mobility and language variation are related to the topics of the text.

Part C explores possible content biases that GSMs are subject to when using topics, due to the usefulness of themes as a structure for analysis. The tendency of topics to contain geographical references is explored in study C1. Across a large web corpus of data from Twitter, Reddit, Nairaland, Stackexchange, web news and American journals, it can be shown that the topics of the texts influence their probability to contain localizable place mentions. Study C2 further demonstrates that the influence of topics can differ depending on whether geotags or geoparsing approaches are used to localize the text content.

Overall, the studies show that there are numerous methods for the algorithmic analysis of digital content that can also be used in the context of migration. Modern machine learning methods in particular enable algorithmic analysis of large volumes of text content. The results generated in this way can be spatially linked to mobility characteristics and remote sensing data, providing considerable added value for geographical research. As studies C1 and C2 demonstrate, there are additional distortions in the geographical evaluation of text content, which must be considered in addition to the frequently documented demographic distortions. The representativeness of GSM data must therefore be critically scrutinized and, together with limitations of data access, represents the main limitation of the approaches explored in the dissertation. For these reasons, traditional methods such as in-situ surveys, census data, and qualitative evaluations can by no means be replaced by the methods demonstrated in this dissertation. Rather, the interaction of these methods can lead to a more comprehensive and complete understanding of migration and migration factors. The perspectives of remote sensing and social media prove to be highly complementary in this respect. Despite the challenges, their combination can make valuable and unique contributions to the study of global migration and urbanization.

Inhaltsverzeichnis

Zusammenfassung	iv
Abstract	vi
Inhaltsverzeichnis.....	viii
Abbildungsverzeichnis	xiii
Tabellenverzeichnis.....	xv
Abkürzungsverzeichnis.....	xvi
1. Einführung	1
1.1. Hintergrund: Globale Migration und Urbanisierung.....	1
1.2. Die Herausforderung: Großflächige Erfassung von Migrationsfaktoren	3
1.3. Eine mögliche Lösung: Fernerkundung und Georeferenzierte Social-Media-Daten	6
1.4. Forschungsschwerpunkte und Ziele der Arbeit.....	9
1.5. Aufbau der Arbeit.....	12
2. Grundlagen.....	19
2.1. Methodische Grundlagen	19
2.1.1. Modelle und Maschinelles Lernen für Fernerkundung und Textprozessierung	19
2.1.2. Embeddings und Sprachprozessierung.....	22
2.1.3. Gemeinsame Auswertung von GSM und FED.....	25
2.1.4. Geoparsing.....	26
2.1.5. Domänenadaption	28
2.1.6. Trajektorienanalysen mittels GSM	29
2.2. GSM-Spezifische Bias-Faktoren	32
2.2.1. Digitale Disparitäten in urbanen Räumen.....	33
2.2.2. Räumlichkeitseffekte von Textinhalten	34
2.3. Anwendungen an der räumlichen Intersektion von GSM & FED	35
2.3.1. Nutzung von GSM und FED zur Analyse und Strukturierung urbaner Räume.....	35
2.3.2. Nutzung von GSM zur Charakterisierung der Perspektive von Migrierten	37
2.3.3. Nutzung von GSM zur Analyse sprachlicher Elemente in Nigeria	38
3. Konzipierung der Studien	39
3.1. Studie A1: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Erfassung digitaler Disparitäten in Afrikas expandierenden städtischen Räumen	39

3.2.	Studie A2: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Ableitung von Raumeinheiten für die explorative Analyse von physischen und thematischen Charakteristika urbaner Räume	43
3.3.	Studie B1: Differenzierung textinhaltlicher Präferenzen von Nigerianischen Twitter-Nutzenden nach Mobilitätscharakteristika	45
3.4.	Studie B2: Analyse der Verbreitung von Sprachmerkmalen nach Mobilitätscharakteristika der Autoren am Beispiel westafrikanischer Twitter-Nutzender	48
3.5.	Studie C1: Analyse der Räumlichkeit von Themen in englischen Webtextdaten ...	50
3.6.	Studie C2: Differenzierung der Räumlichkeit von Themen zwischen Geotag- und Geoparsing-basierten Ansätzen	54
4.	Untersuchungsgebiet und Daten	55
4.1.	Untersuchungsregion Afrika	55
4.2.	Urbane Testgebiete (UTG)	55
4.3.	Fokusregion Nigeria	57
4.4.	Fernerkundungsdaten	59
4.4.1.	World Settlement Footprint Suite	59
4.4.2.	DIY-BU Maps	60
4.4.3.	Quantifizierung der Geplantheitsintensität	61
4.4.4.	Räumliche Metriken	62
4.5.	Webtextdaten	63
4.5.1.	Twitter	63
4.5.2.	Nairaland	66
4.5.3.	Reddit	67
4.5.4.	Stackexchange	67
4.5.5.	GDELT-GKG	68
4.5.6.	GDELT- IA-Americana	69
5.	Studienspezifische Methoden	70
5.1.	Studie A1: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Erfassung digitaler Ungleichheiten in Afrikas expandierenden städtischen Räumen	70
5.1.1.	Berechnung der Geotweet-Dichte	70
5.1.2.	Quantifizierung der Dichtedisparität	72
5.1.3.	Permutation	74
5.1.4.	Statistische Auswertung	75

5.2.	Studie A2: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Ableitung von Raumeinheiten für die explorative Analyse von physischen und thematischen Charakteristika urbaner Räume	76
5.2.1.	Umwandlung der GSM-Daten in Rasterformat	77
5.2.2.	Clustering.....	78
5.2.3.	Auswertung.....	79
5.3.	Studie B1: Ableitung inhaltlicher Präferenzen von nigerianischen Twitter-Nutzenden nach Mobilitätscharakteristika	81
5.3.1.	Initiale Datenakquise und Trajektorienanalyse	82
5.3.2.	Textklassifizierung mittels Domain-Adaptation.....	83
5.3.3.	Plausibilitätsanalysen	85
5.3.4.	Statistische Auswertung.....	87
5.4.	Studie B2: Analyse der Verbreitung von Sprachmerkmalen nach Mobilitätscharakteristika der Autoren am Beispiel westafrikanischer Twitter-Nutzender.....	87
5.4.1.	Datenakquise	88
5.4.2.	Quantifizierung der Lokalen Form des Englischen in Nigeria	89
5.4.3.	Kontextualisierung basierend auf Ort, Nutzermobilität und Themen.....	90
5.5.	Studie C1: Analyse der Räumlichkeit von Themen in englischen Webtextdaten ...	92
5.5.1.	Vorbereitung des Korpus.....	92
5.5.2.	Themen-Taxonomie	95
5.5.3.	Geoinformation	96
5.5.4.	Validierung	97
5.5.5.	Statistische Auswertung.....	97
5.6.	Studie C2: Differenzen der Räumlichkeit von Themen zwischen Geotag- und Geoparsing-basierten Ansätzen.....	98
5.6.1.	Analyse von Unterschieden in der Häufigkeit von Ortsnennungen zwischen Geotweets und Tweets ohne Geotags.....	98
5.6.2.	Analyse von Differenzen in Räumlichkeit	99
6.	Ergebnisse der Studien.....	100
6.1.	Studie A1: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Erfassung digitaler Ungleichheiten in Afrikas expandierenden städtischen Räumen.....	100
6.1.1.	Afrikaweit	102
6.1.2.	Differenziert auf der Ebene von Georegionen Afrikas	103
6.1.3.	Differenziert nach Graden von Geplantheit (IoP)	104

6.1.4.	Modelliert im Zusammenhang mit Räumlichen Metriken	106
6.1.5.	Interpretation.....	107
6.2.	Studie A2: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Ableitung von Raumeinheiten für die explorative Analyse von physischen und thematischen Charakteristika urbaner Räume	110
6.2.1.	Clustererstellung und Interpretation	110
6.2.2.	Statistische Analyse der Clustergüte.....	112
6.2.3.	Interpretation.....	113
6.3.	Studie B1: Ableitung inhaltlicher Präferenzen von nigerianischen Twitter- Nutzenden nach Mobilitätscharakteristika	114
6.3.1.	Mobilitätscharakteristika	115
6.3.2.	Inhaltliche Präferenzen	116
6.3.3.	Interpretation.....	118
6.4.	Studie B2: Analyse der Verbreitung von Sprachmerkmalen nach Mobilitätscharakteristika der Autoren am Beispiel westafrikanischer Twitter Nutzender	121
6.4.1.	Verbreitung von LEN Allgemein.....	121
6.4.2.	Verbreitung nach Ort.....	122
6.4.3.	Verbreitung nach Mobilitätscharakteristika.....	124
6.4.4.	Verbreitung nach Themen	125
6.4.5.	Interpretation.....	127
6.5.	Studie C1: Analyse der Räumlichkeit von Themen in englischen Webtextdaten .	129
6.5.1.	Korpusstatistiken	129
6.5.2.	Ergebnisse der Modellierung	131
6.5.3.	Validierung	134
6.5.4.	Interpretation.....	134
6.6.	Studie C2: Differenzen der Räumlichkeit von Themen zwischen Geotag- und Geoparsing-basierten Ansätzen.....	138
6.6.1.	Unterschiede in der Häufigkeit von Ortsnennungen zwischen Tweets mit und ohne Geotags	138
6.6.2.	Differenzen in Räumlichkeit.....	138
6.6.3.	Interpretation.....	139
7.	Diskussion.....	141
7.1.	Praktisches Potenzial	141
7.1.1.	Technische Möglichkeiten der Nutzung von GSM.....	141
7.1.2.	Abdeckung von GSM	142

7.1.3.	Verknüpfbarkeit von Fernerkundung und GSM	145
7.2.	Validität der Daten.....	146
7.3.	Zugang zu Daten.....	148
7.4.	Ethische Überlegungen hinsichtlich der Datennutzung	149
8.	Synthese.....	152
8.1.	Fazit der Studien	153
8.2.	Schlussfolgerung	156
8.3.	Einordnung und Ausblick	158
9.	Literatur.....	159
	Anhang.....	180
A1.	Beschreibung des Nairaland-Datensatzes.....	180
A2.	Vergleich von Divergenzmetriken.....	181
A3.	Robustheitstests in Studie A1	183
A4.	Variationen des in Studie 2A demonstrierten Clusterings	188
A5.	Ableitung von Raumeinheiten für Studie B1	189
A6.	Ergebnisse der Analyse von Themen nach Mobilitätsgrad	191
A7.	Plausibilitätschecks in Studie B1	192
A8.	Codes der Qualitativen Analyse in Studie B1	193
A9.	Vereinigung von Themen in Experiment in Studie B1	196
A10.	Themen der in Studien C1 und C2 genutzten Taxonomie	197
A11.	Konfusionsmatrizen der Validierung der Themenklassifizierung in Studie C1....	198
A12.	Modelloutputs der Studie C1	199

Abbildungsverzeichnis

Abbildung 1: Datenkonzept der Dissertation.....	7
Abbildung 2: Beitrag der Studie zum Thema der Urbanisierung und Migration in Afrika....	10
Abbildung 3: Disziplinäre Perspektive der Dissertation	11
Abbildung 4: Konzept und Struktur der Dissertation	13
Abbildung 5: Überblick über die in in der Dissertation genutzten maschinellen Lernverfahren	20
Abbildung 6: Illustration von Embeddings	22
Abbildung 7: Illustration des Geoparsings, nach Wang und Hu (2020).....	27
Abbildung 8: Illustration des domänen-adversarialen Trainings nach Ganin et al. (2016)	28
Abbildung 9: Extraktion von Trajektorien aus Digitalen Spuren (2020)	29
Abbildung 10: Segmentbasierter Ansatz zur Identifikation von Migrationsereignissen in Trajektorien, reproduziert nach Chi et al. (2020)	30
Abbildung 11: Anzahl angerufener Geotweets aus Nigeria zwischen 2011 und 2022, eingefärbt nach Art der Georeferenz.....	46
Abbildung 12: Raumeinheiten in Studie B1.....	46
Abbildung 13 Arbeitsablauf von Studie C1	52
Abbildung 14: Schematische Darstellung der Selektion der UTG (-Zentren).....	57
Abbildung 15: Durch Kombination von WSF-Evolution und DIY-BU Kartierungen erstellte Karten des Bebauungsjahres für fünf ausgewählte UTG.	61
Abbildung 16: Beispiele für UTG verschiedener Geplantheitsintensität (IoP).....	62
Abbildung 17: Ableitung des Bebauungsalters der Tweets	71
Abbildung 18: Ableitung der Geotweet-Dichte (f) für ein Jahr (2014) in einem UTG	72
Abbildung 19: Berechnung der Schiefe der Digitalen Dichte (DDS).	73
Abbildung 20: Einige fiktive Verteilungen von Geotweet-Dichte und deren Quantifizierung durch die DDS-Metrik	74
Abbildung 21: Illustration des Permutationsansatzes..	75
Abbildung 22: Arbeitsablauf von Studie B1.	81
Abbildung 23: Illustration der Mobilitätstypen anhand fiktiver Trajektorien.	84
Abbildung 24: Wöchentliche Aktivitätsmuster verschiedener Themen.....	86
Abbildung 25: Geografische Untersuchungsregion der Studie B2.....	88
Abbildung 26: Arbeitsablauf von Studie B2..	91
Abbildung 27: Entwicklung der Thementaxonomie..	95
Abbildung 28: Ableitung des für Studie C2 genutzten Datensatzes, welcher aus Geotweets und Tweets ohne Geotags besteht	99

Abbildung 29: Illustration der Geotweet-Verteilungen und Ergebnisse der Dichte für vier ausgewählte UTGs.	101
Abbildung 30: Vergleich der Disparität zwischen UN-Regionen	103
Abbildung 31: Vergleich der Disparität nach IoP _B	104
Abbildung 32: Vergleich der Disparität nach IoP _B	105
Abbildung 33: Stärken der linearen Beziehung zwischen Disparität (DDS _{UTG})	106
Abbildung 34: Vergleich von beobachteten DDS _{UTG} und durch das multiple lineare Regressionsmodell geschätzten DDS _{UTG}	107
Abbildung 35: Darstellung der räumlichen Cluster für UTG A	111
Abbildung 36: Verteilung der statistischen Evaluationskriterien über UTG und die vier verglichenen Clusterparametrisierungen hinweg.	113
Abbildung 37: Diagramm der detektierten MOVES und M _{INTS}	115
Abbildung 38: Histogramm des Zehnerlogarithmus von Gyrationsradien der Nutzenden, eingefärbt nach Mobilitätstyp	116
Abbildung 39: Gemittelte Themenanteile der Mobilitätstypen.	117
Abbildung 40: LOESS Kurve der Anteile von Themen an Tweets der Nutzenden gegen den Zehnerlogarithmus ihrer Gyrationenradien.	117
Abbildung 41: Identifizierte Codes und ihre Anzahl innerhalb der Themen <i>gaming</i> , <i>religion</i> und <i>travel</i>	118
Abbildung 42: Verbreitung von LEN in MUAs und nichtmetropolischen Zonen im Untersuchungsgebiet.	123
Abbildung 43: Durchschnittlicher LI pro Land.....	124
Abbildung 44: Median des LI von Tweets nach Mobilitätstypen	125
Abbildung 45: Ergebnisse des Topic Modeling in Verbindung mit LEN	126
Abbildung 46: Anteil der Tweets mit LEN-Begriffen verglichen mit dem Anteil an Tweets mit morphosyntaktischen oder pragmatischen Sprachmerkmalen.	127
Abbildung 47: Frac _{geo} der Themen, strukturiert nach Datensatz.....	130
Abbildung 48: Gemessene Frac _{geo} nach Zeitintervallen (a), Quellen (b), und Textlänge (c)..	132
Abbildung 49: Durch Modellierung geschätzte Geospatiality-Effekte jedes Themas	133
Abbildung 50: Vergleich von Frac _{Geo} zwischen Geotweets und Tweets ohne Geotags	138
Abbildung 51: Durch Modellierung geschätzte Geospatiality-Effekte jedes Themas	139

Tabellenverzeichnis

Tabelle 1: Liste der räumlichen Metriken	63
Tabelle 2: Beispiele für Metadatenfelder des GDELT Global Knowledge Graph.....	68
Tabelle 3: Beschreibung der in Studie A2 genutzten UTG	77
Tabelle 4: Beispiel für die Struktur des für Studien C1 und C2 kompilierten Korpus.....	93
Tabelle 5: Regressionskoeffizienten zwischen DDS_{UTG} und räumlichen Metriken.....	107
Tabelle 6: Statistiken der Nutzenden im Satz U_{AN} , nach Mobilitätstyp.....	116
Tabelle 7: Überblick über der für Studie B2 akquirierten Daten.	121
Tabelle 8: Beschreibung und Anzahl der in Studie B2 analysierten detaillierten Mobilitätstypen.....	124
Tabelle 9: Überblick über den für Studie C1 und C2 generierten Korpus, welcher aus sieben verschiedenen Datensätzen besteht.	129
Tabelle 10: Deskriptive Statistiken für Fracgeo über die Datensätze hinweg.....	131
Tabelle 11: Spearman-Rangkorrelation zwischen den Topic-Rankings jedes Datensatzes ...	134
Tabelle 12: Validationsgenauigkeiten in Studie C1.	134

Abkürzungsverzeichnis

API	Application Programming Interface
COVID-19	Corona Virus Disease 2019
DDS	Digital Density Skew
DDS _{UTG}	Digital Density Skew eines urbanen Testgebietes
DLR	Deutsches Zentrum für Luft- und Raumfahrt
ECOWAS	Economic Community of West African States
GSM	Georeferenced Social Media
IDMC	Internal Displacement Monitoring Centre
IDP	Internally Displaced Persons (Binnenvertriebene)
IOM	International Organization for Migration
IPCC	Intergovernmental Panel on Climate Change
ITU	International Telecommunication Union
MIGRAWARE	Rural-urban and cross-border MIGRAtion in West-Africa - an integrated assessment framework of drivers, processes and sustainable REsponses
NLP	Natural Language Processing
OECD	Organization for Economic Co-operation and Development
UN	United Nations
UN-DESA	United Nations Department of Economic and Social Affairs
UNHCR	Office of the United Nations High Commissioner for Refugees
URL	Uniform Resource Locator
UTG	Urbane Testgebiete
WASCAL	West African Science Service Center on Climate Change and Adapted Land Use

1. Einführung

1.1. Hintergrund: Globale Migration und Urbanisierung

Die Suche nach besseren Lebensbedingungen ist wahrscheinlich seit jeher ein grundlegender Bestandteil menschlicher Existenz. Mit diesem Ziel haben Menschen ihre Mobilitätsentscheidungen immer auch an ihrer Umwelt ausgerichtet, und wurden so gesellschaftlich, wirtschaftlich und kulturell von ihr geprägt (Anthony 2023). Gleichzeitig beeinflussen die Menschen in diesem Bestreben auch ihre Umwelt dadurch, wie und wo sie handeln. Im dritten Jahrtausend ist dieser menschliche Einfluss auf die Erde spürbar wie nie zuvor. Mit der zunehmenden Anzahl an Menschen auf der Erde und ihren wachsenden technischen Möglichkeiten wird auch der Einfluss menschlicher Migration zunehmend sichtbarer, greifbarer und messbarer (Canagarajah 2017). Nicht nur sind mehr Menschen als jemals zuvor Migrierte, die Menschheit als Ganzes ist durch den Prozess der globalen Urbanisierung auf einer Wanderung vom Land in die Stadt –einmal mehr angetrieben durch die Suche nach einer besseren Umgebung für ihr Leben und besseren Aussichten für die Zukunft (Saunders 2010). Wissen darüber, wie Menschen dabei ihre Umgebung wählen und gestalten, ist essenziell zum Verständnis unserer Umwelt –und umgekehrt ist diese Umwelt weiterhin wichtig zum Verständnis von menschlichem Migrationsverhalten.

Dies gilt insbesondere für die städtische Umwelt, welche inzwischen den meisten Menschen als Lebensraum dient, und die Hauptdestination für die meisten Migrierten darstellt (IOM 2015; Duncan und Popp 2018). Die Internationale Organisation für Migration (IOM) widmete den 2015 World Migration Report explizit der Interaktion von Migration und Urbanisierung, und hob in der Einleitung die Bedeutung von Städten für Migration hervor: „Migration is essentially an urban affair“ (IOM 2015), oder, wie Duncan und Popp (2018) im darauffolgenden World Migration Report ausführten: „Migration has become nearly synonymous with urbanisation, given the dominance of the city as the destination of most migrants“. Anders gesagt: Globale Migration ist ohne globale Urbanisierung kaum denkbar.

Besonders deutlich wird die Interaktion von Migration und der städtischen Umwelt in Afrika, einem Kontinent von erheblicher Dynamik in Bezug auf Urbanisierung und Migration (McAuliffe 2024), auf welchem das Wachstum der Stadtbevölkerung in hohem Tempo voranschreitet. In den vergangenen Jahrzehnten verzeichnete Afrika, zusammen mit Asien, sowohl die höchsten Zuwachsraten an städtischer Bevölkerung (Kamana et al. 2024) als auch an städtischer Fläche (Taubenböck et al. 2024). Und es wird erwartet, dass sich dieser Trend in Zukunft fortsetzt. Für die drei Jahrzehnte zwischen 2020 und 2050 prognostiziert die OECD¹ einen Anstieg der afrikanischen Stadtbevölkerung von 704 Millionen auf 1 400 Millionen Menschen, was 80% des gesamten Bevölkerungswachstums des Kontinents ausmacht. Damit einhergehend sei ein Anstieg des städtischen Fußabdrucks von 175 000 km² auf 450 000 km² zu erwarten (OECD et al. 2025). Anders als in vielen Regionen der Erde werden die meisten Länder Afrikas ihr Bevölkerungsmaximum erst nach 2100 erreichen (UN DESA 2024). Angesichts dieser Prozesse kann Afrika als der Ort betrachtet werden, wo sich die heutige „Urban Frontier“ befindet, ein städtischer Raum, welcher in vielen Aspekten noch entwickelt werden muss, und der sich durch hohes Potenzial für zukünftige Bevölkerungs- und Stadtentwicklung auszeichnet (Goodfellow 2022; Mast et al. 2025), kurz: Urbanes Grenzland.

¹ Organisation for Economic Co-operation and Development

Migration spielt eine wesentliche Rolle darin, dieses Potenzial zu realisieren. Umzug vom Land in die Städte ist nach dem natürlichen Bevölkerungswachstum meistens der wichtigste Faktor zum städtischen Bevölkerungswachstum². Tacoli et al. (2015) schätzen, dass in Subsahara-Afrika etwa ein Drittel des zwischen 2000 und 2010 erfolgten städtischen Bevölkerungswachstums auf Migration zurückzuführen ist. Migration spielt dabei nicht nur eine wichtige Rolle im Prozess der Urbanisierung (Tacoli et al. 2015) – wobei die Zielorte meist wirtschaftlich dominante Metropolen sind, die Migrierende auf der Suche nach Existenzsicherung und Beschäftigung anziehen – sondern auch in den Wachstums- und Transformationsprozessen der Städte (Bloch et al. 2015; Duncan und Popp 2018).

Doch Migration ist bei weitem nicht der einzige Faktor, welcher das Schicksal der Städte bestimmen wird. Ihr Wachstum erfolgt nicht isoliert wie das Wachstum einer Zellkultur in einer räumlich abgegrenzten Petrischale, sondern eingebettet in eine dynamische Welt und im Kontext zahlreicher globaler Transformationsprozesse wie dem Klimawandel (McAuliffe 2024; Dodman et al. 2022) und damit einhergehend zunehmend häufigeren Naturkatastrophen (IPCC 2022; Taubenböck et al. 2024), Digitalisierung und künstlicher Intelligenz (McAuliffe 2021; Nchake und Shuaibu 2022; McAuliffe et al. 2022), dem demographischen Wandel (Findlay und Wahba 2013; Fargues 2011; Geis-Thöne 2020) und vielen weiteren Faktoren. Die Resilienzen der afrikanischen Siedlungsräume werden angesichts zahlreicher Herausforderungen auf die Probe gestellt (Kamana et al. 2024). Für die negativen Auswirkungen dieser Prozesse, insbesondere des Klimawandels und der damit einhergehen häufigeren Naturkatastrophen, sind Migrierte besonders vulnerabel (Tacoli et al. 2015; IPCC 2022). Die Gestaltung der Migration stellt die Gesellschaft und auch die Städte vor erhebliche weitere Herausforderungen (Becker 2008; Tacoli et al. 2015; Kamana et al. 2024).

Gleichzeitig bereichert Migration die beteiligten Gesellschaften, beispielsweise durch den Austausch von Arbeitskräften (Geis-Thöne 2020; Lee 2015), Ideen (Curry et al. 2024) und Kapital (Becker 2008; World Bank 2022; McAuliffe 2024). Dabei profitieren insbesondere Städte trotz der Herausforderungen des Wachstums von dessen zahlreichen wirtschaftlichen und sozialen Vorteilen (IOM 2015; Tacoli et al. 2015; Glaeser 2011). Und auch auf individueller Ebene ist Migration selbst häufig eine erfolgreiche Anpassungsstrategie an die Veränderungen der physischen und sozialen Umwelt (Dodman et al. 2022; Black et al. 2011; Chazalnoël und Randall 2022). Um einmal mehr die Perspektive der IOM aufzugreifen: Migration stellt in Hinsicht auf die Herausforderungen einer sich schnell wandelnden Welt „einen Teil der Lösung“³ dar (McAuliffe 2024). Es bestehen also sowohl Herausforderungen als auch Chancen hinsichtlich Migration, und damit besteht ein großes gesellschaftliches Interesse an der Erforschung dieses Aspektes der Mensch-Umwelt-Interaktion.

Die hohe Relevanz von Migration wird durch die Vielzahl an Wissenschaften verdeutlicht, welche sich mit ihr befassen (Scholten et al. 2022; Brettell und Hollifield 2015). Die Migrationslinguistik theorisiert beispielsweise die Interaktion von Sprache und Migration bei Spracherwerb und Sprachnutzung von Migrierten (Borlongan 2023). Ansätze der Kulturwissenschaft erforschen unter anderem Konzepte und Begriffe von Migration. Diese reichen von Grenzen und Kolonialismus, von interkultureller Anerkennung, sowie postmigrantischen Diskursen und Praktiken des Migrationsalltags bis hin zu Darstellungen

² Die Ausnahme bilden Länder mit erheblichen Fertilitätsunterschieden zwischen städtischen und ländlichen Regionen, welche relativ selten sind (Becker 2008; Tacoli et al. 2015)

³ „part of the solution in a rapidly changing world“

und Narrativen von Migration in Politik, Kunst und Medien (Nesselhauf 2018; Ezli 2022; Martiniello 2022). Sie verbinden dabei theoretische Ansätze mit literarischen oder kuratorischen Fallbeispielen und ethnologischer Feldforschung (Nesselhauf 2018), thematisieren Aspekte der Identität und Gruppenzugehörigkeiten von Migrierten oder vernetzen Migration in (raum-)theoretischen Modellen mit dem Effekt von Migration auf beteiligte Individuen und Gesellschaften (Rapoport et al. 2021).

Neben diesen auf Theorie fokussierten Forschungsansätzen besteht auch eine lange Tradition empirischer Migrationsforschung. Bereits die frühen „Gesetze der Wanderung“ von Ravenstein (1876), welche als Startpunkt der modernen Migrationsforschung gesehen werden können, wurden auf Basis von Volkszählungen formuliert. Empirische Ansätze zur Migration finden sich unter anderem in der Geographie (Tobler 1995), Demographie (Bean und Brown 2014), Stadtplanung (Bernt 2019; Saunders 2010), in den Sprachwissenschaften (Esser 2006) und den Wirtschaftswissenschaften (Martin 2014). Hinzu kommt die praktische Notwendigkeit des Migrationsmanagements durch Verwaltung und Politik. Ohne Zweifel besteht ein breites wissenschaftliches Interesse daran, die Faktoren, welche Migration und Mobilität beeinflussen, empirisch zu untersuchen (Tjaden 2021) und damit zum Verständnis unserer Welt, dem erfolgreichen Umgang mit Migration und der Entwicklung unserer Städte beizutragen (International Organization for Migration 2015).

Unter den oben genannten Disziplinen ist die Geographie besonders qualifiziert, die Bedeutung von Migration als Prozess zu analysieren, welcher im Raum stattfindet, Räume prägt, und von Räumen geprägt wird (King 2012). Hierfür benötigt sie Daten mit räumlichem Bezug: **Georeferenzierte Daten**. Jedoch ist insbesondere die großflächige Erfassung dieser Daten keinesfalls unkompliziert –in erster Linie aufgrund der Komplexität von Migration und der sie beeinflussenden Faktoren.

1.2. Die Herausforderung: Großflächige Erfassung von Migrationsfaktoren

Wie die Vielfalt an möglichen Ansätzen zur Migration bereits andeutet, ist Migration ein vielschichtiger Forschungsgegenstand, welcher mit vielen Faktoren von Umwelt, Gesellschaft und individuellen Präferenzen in Beziehung steht (Dodman et al. 2022).

Die theoretische Migrationsforschung hat im Laufe des 20. Jahrhunderts verschiedene Modelle zur Erklärung von Migrationsentscheidungen entwickelt. Neoklassische Modelle der Migration führen Migrationsentscheidungen auf einzelne Push- und Pull-Faktoren (Lee 1966) und Kosten-Nutzen-Analysen zurück. Jüngere Modelle wie die Netzwerktheorie der Migration rücken soziale Netzwerke wie Familien und den Fluss von Informationen stärker in den Fokus (Czaika und Reinprecht 2022). Konsens der meisten diskutierten Theorien ist die Einordnung von Migration als hochgradig kontextabhängiges Phänomen. Dieser Kontext definiert sich durch ein Umfeld von zahlreichen Faktoren, welche zu einem bestimmten Zeitpunkt an einem bestimmten Ort die Bereitschaft und Fähigkeit von Menschen zur Migration bedingen. Individuen entscheiden dabei zwischen ihren oft begrenzten Optionen (McAuliffe 2024) im Kontext ihrer persönlichen Lebenssituation und oft unter Druck, Stress sowie mit unvollständigen Informationen (Czaika und Reinprecht 2022).

Die Überführung dieser Theorien in empirische Ansätze ist herausfordernd (Carling et al. 2020). Für die empirische Erforschung von Migration muss eine große Bandbreite möglicher Faktoren und Effekte erfasst und zusammen analysiert werden. Hierbei sind sowohl objektive

Informationen über die physische und soziale⁴ Umwelt relevant als auch die subjektive Wahrnehmung derer durch Individuen und individuelle Präferenzen. Aus der Vielfalt der möglichen Faktoren ergeben sich große Herausforderungen für die Erfassung in geographischen Maßstäben. Eine weithin genutzte Methode zur Kartierung der physischen Umwelt stellt die Fernerkundung dar. In den letzten Jahren haben Fernerkundungssatelliten immense Mengen an frei zugänglichen Fernerkundungsdaten (FED) generiert (Gomes et al. 2020). Im Migrationskontext wurden FED in der Vergangenheit insbesondere während Krisen und Naturkatastrophen eingesetzt. Mittels FED können beispielsweise von Naturkatastrophen betroffene Flächen (Schumann et al. 2018) oder bedrohte Siedlungsflächen (Kühnl et al. 2023; Rentschler et al. 2023), Flüchtlingslager (Weigand et al. 2023) oder informelle Siedlungen (Wurm et al. 2019) kartiert werden, was zur Abschätzung betroffener oder bedrohter Bevölkerung dienen kann und damit die Koordination von humanitärer Hilfe unterstützt (Bircan 2023). Zusätzlich können die physischen Manifestationen von politischen Prozessen mittels Fernerkundung untersucht werden. Hierzu gehören beispielsweise die Geplantheit von Siedlungen (Debray et al. 2023) und nationalstaatliche Grenzen (Taubenböck et al. 2023).

Auch langfristige umweltbedingte Migrationsfaktoren können mittels FED erfasst werden. Schürmann et al. (2024) identifizierten mittels Experteninterviews in Ghana saliente Migrationsfaktoren und bildeten viele von ihnen direkt oder mittels fernerkundlich erfasster Proxy-Variablen ab. Hierzu gehörten beispielsweise Extremwetterereignisse sowie die Fertilität und Qualität von Böden (Schürmann et al. 2024). Während die Fernerkundung also ein mögliches Mittel zur großflächigen Erfassung der physischen Umwelt sein kann, können soziale Informationen aus FED nur dann mittels Proxy-Variablen abgeleitet werden, wenn sie eindeutige physische Ausprägungen haben, wie es bei soziopolitischen Faktoren wie Geplantheit oder Grenzen der Fall ist (Debray et al. 2023; Taubenböck et al. 2023). Insbesondere für die Ableitung personenbezogener und subjektiver Faktoren wie Sprache, Interessen und individueller Mobilität sind die Möglichkeiten der Fernerkundung begrenzt. Um eine größere Bandbreite von sozialen Informationen für die großflächige Analyse messbar zu machen, sind daher weitere Datenquellen notwendig (Schürmann et al. 2024).

Wünschenswert wäre eine Technik, welche auch soziokulturelle und personenbezogene Daten großflächig, regulär und grenzüberschreitend zu erfassen vermag, ähnlich wie es die Fernerkundung für die physische Umwelt erlaubt. Traditionelle Datenerfassungen sind in dieser Hinsicht limitiert. In-situ-Umfragen können sehr detaillierte Informationen über Individuen und deren Migrationsentscheidungen und Absichten sammeln, sind jedoch sehr aufwendig und teuer (Kraler und Reichel 2022). Auch Zensusdaten können teilweise sehr ausführliche sozioökonomische Daten beinhalten (beispielsweise für Ghana, siehe Ghana Statistical Service 2023). Sie sind jedoch auf Ländergrenzen beschränkt, werden nur unregelmäßig erhoben, und unterscheiden sich in der Methodologie teilweise erheblich zwischen Ländern (Tjaden 2021; Willekens 2019; Rampazzo et al. 2021; Sirbu et al. 2021).

Damit ist die Erfassung von soziokulturellen und individuellen Migrationsfaktoren eine erhebliche Herausforderung für die empirische Erforschung von Migration als geographischem Phänomen.

⁴ Beispielsweise spielen soziokulturelle, wirtschaftliche und politische Aspekte der Umwelt eine Rolle.

Eine zweite Herausforderung ergibt sich in der Vielfalt an Mobilitäts**formen**, welche als Migration betrachtet werden können. Denn Migration ist kein klares und alleinstehendes Konzept, sondern ein Teil des weiten Spektrums menschlicher Mobilität⁵ (Urry 2007), auf welchem sie nicht eindeutig von anderen Mobilitätsformen abzugrenzen ist (Kraler und Reichel 2022).

Die Vielfalt von Migrationsformen wird deutlich bei der Erfassung von Migrationsströmen⁶ und -beständen⁷, welche die beiden grundlegenden Statistiken sind, mit welchen das Ausmaß von Migration quantifiziert wird (Kraler und Reichel 2022). Eine der in der Kommunikation von zwischenstaatlichen Organisationen weit verbreitete Unterscheidung ist die in Migranten, Flüchtlinge und Binnenvertriebene.

Für den häufig genutzten Begriff „**Migrant**“⁸ existiert keine universell anerkannte Definition. Als weithin genutzte Grundlage dient die UN⁹ Definition von 1998, welche von der Hauptabteilung für Wirtschaftliche und Soziale Angelegenheiten der Vereinten Nationen (engl. *United Nations Department of Economic and Social Affairs*, UN DESA) wie folgt formuliert wird: „jede Person, die ihr Land des üblichen Aufenthaltsorts ändert“ (UN DESA, 1998). Hierbei wird bei einer beabsichtigten Aufenthaltsdauer von 3-12 Monaten von „temporären Migrantinnen und Migranten“ und bei länger als 12 Monaten von „dauerhaften Migrantinnen und Migranten“ gesprochen (Migration Data Portal 2021). Diese Kriterien werden jedoch von manchen Ländern auch in abgeänderter Weise benutzt (McAuliffe 2024).

Die Definition von **Flüchtlingen** des für Flüchtlinge verantwortlichen Hohen Flüchtlingskommissars der Vereinten Nationen (engl. *United Nations High Commissioner for Refugees*, UNHCR) bezieht sich auf die Definition der Genfer Flüchtlingskonvention von 1951. Danach ist ein Flüchtling eine „Person, die [...] aus der begründeten Furcht vor Verfolgung wegen ihrer Rasse, Religion, Nationalität, Zugehörigkeit zu einer bestimmten sozialen Gruppe oder wegen ihrer politischen Überzeugung sich außerhalb des Landes befindet, dessen Staatsangehörigkeit sie besitzt, und den Schutz dieses Landes nicht in Anspruch nehmen kann oder wegen dieser Befürchtungen nicht in Anspruch nehmen will [...]“.

Davon unterschieden werden **Binnenvertriebene** (engl. *Internally Displaced Persons*, IDPs): Menschen, die zwar ihre Heimatregion verlassen, aber keine international anerkannte Grenze überschreiten. Der UNHCR und das Internal Displacement Monitoring Centre (IDMC 2025) definieren Binnenvertriebene als „Personen oder Personengruppen, die zur Flucht gezwungen oder verpflichtet wurden oder ihre Häuser oder üblichen Wohnsitze verlassen mussten, insbesondere infolge von oder zum Zwecke der Vermeidung der Auswirkungen von bewaffneten Konflikten, Situationen allgemeiner Gewalt, Menschenrechtsverletzungen oder

⁵ Der Begriff „Mobilität“ kann im weiteren Sinne auch als Bewegungen von Gütern oder von Menschen in anderen Dimensionen als der räumlichen (vertikale & soziale Mobilität) verstanden werden (Canagarajah 2017). In dieser Dissertation wird der Begriff jedoch nur für räumliche Bewegungen von Menschen genutzt.

⁶ Die Anzahl an Menschen, welche innerhalb eines bestimmten Zeitraums (oft ein Kalenderjahr) eine Grenze überqueren

⁷ Die Anzahl an Migranten, welche sich zu einem Zeitpunkt an einem Ort (oft in einem Land) aufhalten

⁸ In dieser Theses wird der geschlechtsneutrale Begriff „Migrierte“ synonym zu „Migranten“ verwendet.

⁹ United Nations (deutsch: Vereinte Nationen)

natürlichen oder von Menschen verursachten Katastrophen, und die keine international anerkannte Staatsgrenze überquert haben“ (United Nations Commission on Human Rights 1998).

Nach diesen Definitionen berichtet die IOM für 2024 von 304 Millionen internationalen Migrierten, davon 169 Millionen Arbeitsmigrierte, 35,3 Millionen Flüchtlingen und 71,2 Binnenvertriebenen. Migrierte machen damit 3,6% der Weltbevölkerung aus (McAuliffe 2024).

Die oben genannten Institutionen gehören auch zu den wichtigsten Datenquellen für ihren Zuständigkeitsbereich. Das Büro des UNHCR stellt Informationen über Flüchtlinge zusammen, während die IOM die prominenteste Institution im Bereich der Migration ist (Kraler und Reichel 2022). Auch die Statistiken von UN DESA werden weithin als zentrale Datenquelle betrachtet, welche auch zwischen Arbeitsmigranten, Vermissten, Asylsuchenden, Binnenvertriebenen und Flüchtlingen differenziert (McAuliffe 2024). Neben den oben genannten Unterscheidungen, welche sich nach Migrationsgrund, Freiwilligkeit und Überschreitung nationalstaatlicher Grenzen als Kriterien orientieren, gibt es, auch innerhalb dieser Gruppen, zahlreiche mögliche andere Abstufungen nach Kriterien der Saisonalität (beispielsweise in saisonale Migration oder Wanderweidewirtschaft), Dauer (Kurzzeit- oder Langzeitmigration, siehe Tjaden 2021; Kraler und Reichel 2022) oder nach Lebensphase oder Familienstatus (King 2012).

Aus dieser Vielfalt an Mobilitätsformen und möglichen Unterscheidungen resultiert eine große Herausforderung für die statistische Erfassung und großflächige Analyse von Migration (Trask 2022). Trotz internationaler Zusammenarbeit in der Zusammenstellung von Datenquellen (McAuliffe 2024) ist nicht davon auszugehen, dass alle Datenbedarfe mit bisherigen Mitteln zu decken sind. Aktuell stellen nur 45 Regierungen Daten zu Migrationsströmen zur Verfügung, was zumindest teilweise auf die extreme Schwierigkeit der Datenerhebung zurückzuführen ist (Chi et al. 2025). Neben den Kosten ist die Harmonisierung unterschiedlicher, gleichwohl valider, Definitionen und Bedarfe in der Datenerhebung eine große Herausforderung. Eine Alternative bestünde darin, großflächig und grenzübergreifend Mobilitätsdaten mit sehr feiner Auflösung zu erheben, aus welchen sich dann flexibel unterschiedliche Definitionen von Migration anwenden und Migrationsströme und -bestände ableiten ließen.

Es lässt sich also zusammenfassen, dass die geographische Untersuchung von Migration einen erheblichen Wert aus flächendeckend ableitbaren Datenquellen zu verschiedenen Migrationsprozessen, -faktoren und -effekten ziehen könnte. Für manche Faktoren ist dies bereits möglich.

1.3. Eine mögliche Lösung: Fernerkundung und Georeferenzierte Social-Media-Daten

Fernerkundungsdaten haben das Potenzial, Informationen über die physische Umwelt flächendeckend und systematisch zu liefern. Diese sind als Proxyinformationen für Auslöser (z.B. Naturkatastrophen) und Effekte von Migration (z.B. Flüchtlingslager) zu verstehen. Es fehlt jedoch an gleichsam flächendeckenden Datenquellen für soziokulturelle und subjektive Aspekte von Migration.

Textdaten aus dem Internet könnten diesen Bedarf möglicherweise in Teilen decken (Sirbu et al. 2021). In den letzten Jahrzehnten haben die Verbreitung und Nutzung des Internets weltweit stark zugenommen. Von 1 Milliarde (16% der Weltbevölkerung) im Jahr 2005 ist die

Zahl der Internetnutzenden auf 5,5 Milliarden (68%) im Jahr 2024 angestiegen (ITU 2024). Nach einem Marktforschungsbericht der Meltwater Group¹⁰ (Kemp 2025) verbringen die 5,24 Milliarden Social-Media-Nutzenden täglich durchschnittlich über zwei Stunden auf Social-Media-Plattformen und generieren dabei –wenn auch nicht mit dieser Absicht– gigantische Datenmengen für die Forschung (Martí et al. 2019; Zhu et al. 2022). Motiviert durch das Potenzial der sozialen Medien wurden in den letzten Jahren mit großer Innovationskraft zahlreiche Datenquellen für die Migrationsforschung erschlossen. Hierzu gehören Social-Media-Plattformen, Mobilfunkdaten, Digitale Nachrichtenartikel und Suchmaschinendaten (Tjaden 2021).

Manche dieser Datenquellen beinhalten Daten in Textform. Im Rahmen dieser Dissertation werden unter dem Begriff *Webtextdaten* alle diese Textdaten verstanden, welche im Internet verfügbar sind. Innerhalb der Webtextdaten stellen Inhalte von textbasierten Social Media eine Untergruppe dar, welche sich dadurch auszeichnen, dass sie primär von Menschen und für Menschen generiert werden (Abbildung 1). Inhalte auf diesen Plattformen sind teilweise öffentlich, teilweise privat.

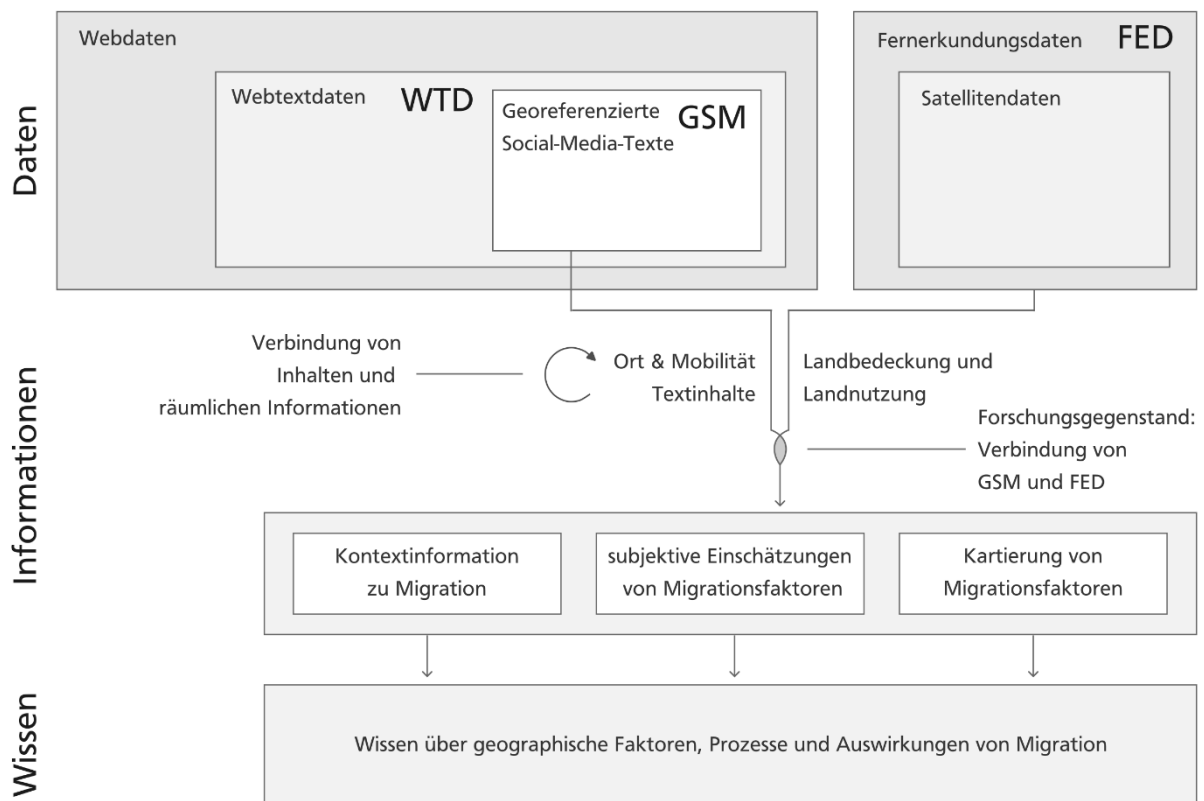


Abbildung 1: Datenkonzept der Dissertation

Unter diesen Social-Media-Daten sind Microblogs (Java et al. 2007) besonders vielversprechend für die Ergänzung der Fernerkundung in der Forschung. Als Blogs sind Posts auf diesen Plattformen in der Regel öffentlich und können mittels Programmierschnittstellen (engl. *Application Programming Interfaces*, APIs, siehe Lomborg und

¹⁰ <https://www.meltwater.com/en>

Bechmann 2014) in großem Stil abgefragt werden, wobei sowohl die Texte der Kommentare als auch detaillierte Metadaten abgerufen werden können, von denen vor allem die optionalen Geotags relevant sind (Zhu et al. 2022). Geotags sind Metadaten, welche einem Inhalt eine geographische Referenz (Georeferenz) zu einem bestimmten Ort auf der Erde zuordnen, in der Regel in Form eines Koordinatenpaares. Das Setzen eines Geotags durch Nutzende oder deren Geräte wird als *Geotagging* bezeichnet (Kruspe et al. 2021).

In den letzten Jahren nutzten zahlreiche räumliche Analysen Daten der Microblog Plattform Twitter (heute: X), welche eine Geotagging-Funktion anbietet. Posts dieser Plattform sind vergleichsweise kurze Texte mit bis zu 280 Zeichen. Sie werden weithin als *Tweets* bezeichnet, und mit Geotags versehene Tweets als *Geotweets*. Diese Geotweets machen auf Twitter zwar nur einen kleinen Anteil¹¹ aller Tweets aus. Ihre Georeferenzen, welche verschiedenen Präzisionsgraden verortet werden können, lassen sich aber vielfältig auswerten. Auf dem Level der individuellen Tweets können sie genutzt werden, um Inhalte zu verorten, beispielsweise im Kontext von Naturgefahren (Kersten und Klan 2020) oder Stadtforschung (Lansley und Longley 2016). Sie werden aber auch –auf Accountebene aggregiert– genutzt, um vergangene Aufenthaltsorte und Mobilität von Nutzenden abzuschätzen (Hawelka et al. 2014; Zagheni et al. 2014).

(Geo-)Tweets sind eine Quelle für geographische Informationen, welche als Teil einer der populärsten Microblog-Plattformen weit verbreitet und teils öffentlich zugänglich sind. Dies erleichtert ihre Akquise und macht sie ethisch weniger problematisch, was positiv in Bezug auf Reproduzierbarkeit¹² und Transparenz bewertet werden kann (Lomborg und Bechmann 2014).

Auch Texte, welche nicht durch Geotags annotiert sind, können geographisch verortet werden, indem für in dem Text erwähnte Orte die geographischen Koordinaten algorithmisch nachgeschlagen werden –ein Prozess, welcher als *Geoparsing* bezeichnet wird (Hu et al. 2023b). Georeferenzen können also sowohl explizit (in Form von Geotags) als auch implizit (Ortserwähnungen) auftreten. In dieser Dissertation werden Social-Media-Inhalte, welche über einen der beiden Ansätze geographisch verortet wurden, als georeferenzierte Soziale Medien (GSM) bezeichnet.

In der geographischen Nutzung von Tweets waren bislang hauptsächlich Geotweets als weithin genutzte Quelle in der Migrationsforschung verbreitet (McAuliffe et al. 2018; Dinbabo 2018). Im Kontext von Migration wurden mittels Geotweets beispielsweise Migrierte identifiziert (Chi et al. 2020) und damit Migrationsmuster erfasst (Zagheni et al. 2014). Geotweets und vergleichbare Inhalte anderer Plattformen werden dabei primär als Einzelsignal gesehen, welche für einen Ort und Zeitpunkt die Präsenz der Nutzenden andeutet.

Ein paralleler Forschungsstrang bezieht sich nicht auf die geographische Information von (Geo-)Tweets, sondern auf deren oft sehr inhaltsreiche Textinformation, welche für die Nutzenden die primäre Information darstellt. Dank moderner Techniken aus dem Bereich der Verarbeitung natürlicher Sprache (engl. *Natural Language Processing*, NLP), welche vielfach auf

¹¹ Die genauen Anteile variieren nach Studie, aber bewegen sich typischerweise im niedrigen einstelligen Prozentbereich. Beispielsweise berichten Kruspe et al. (2021) von 1-2%, Pfeffer et al. (2023) von 0,5%, mit erheblichen geographischen Variationen.

¹² Die Reproduzierbarkeit von Daten sozialer Medien unterliegt verschiedenen Einschränkungen, siehe Kinder-Kurlanda et al. (2017) und Kapitel 7.4 dieser Dissertation.

Ansätzen des maschinellen Lernens basieren, können Textdaten heute auf vielfältige Weise algorithmisch ausgewertet werden. Weit verbreitet für Twitter-Daten sind Methoden für die Aufgabe der Textklassifizierung, bei welcher ganze Texte und Dokumente zu einer Klasse, beispielsweise einem Thema oder einer Stimmung, zugewiesen werden. Die voranschreitende Entwicklung von Techniken für diese Aufgaben stellt in Aussicht, dass vielfältige Merkmale der Texte, auch in großem Stil, algorithmisch für die Forschung quantifizierbar gemacht werden könnten.

Über die Existenz von Georeferenzen, ob als Geotag oder Ortserwähnung, lässt sich zumindest ein Teil dieser Inhalte im geographischen Raum verorten. Dort können sie mit anderen Geodaten wie FED in Bezug gesetzt werden, um eine Vielfalt an Auswertungsmethoden quantitativer und qualitativer Natur zu ermöglichen (siehe Kelle 2014).

Die gemeinsame Nutzung von Textinhalten und Geotags in der Migrationsforschung ist jedoch weit weniger verbreitet als die alleinige Nutzung von Geotags als Einzelsignal. Vielversprechende Beispiele zeigen Studien, welche die Verbreitung von Hate-Speech räumlich messen (Arcila-Calderón et al. 2022), die Bindung von Migrierten zu ihren Ziel und Herkunftsländern quantifizieren (Kim, Sîrbu, Giannotti, et al. 2022), Präferenzen in Themen und Sprache in Bezug zum Aufenthaltsort setzen (Lemoine-Rodríguez et al. 2024) und die Fähigkeit von Städten, Sprecher verschiedener Sprachen räumlich zu integrieren, zu vergleichen (Lamanna et al. 2018). Diese Studien belegen zwar keine flächendeckende Kartierung von Migrationsfaktoren (wie es mittels FED möglich ist), belegen aber die Plausibilität der geographischen Auswertung von GSM-Texten im Migrationskontext.

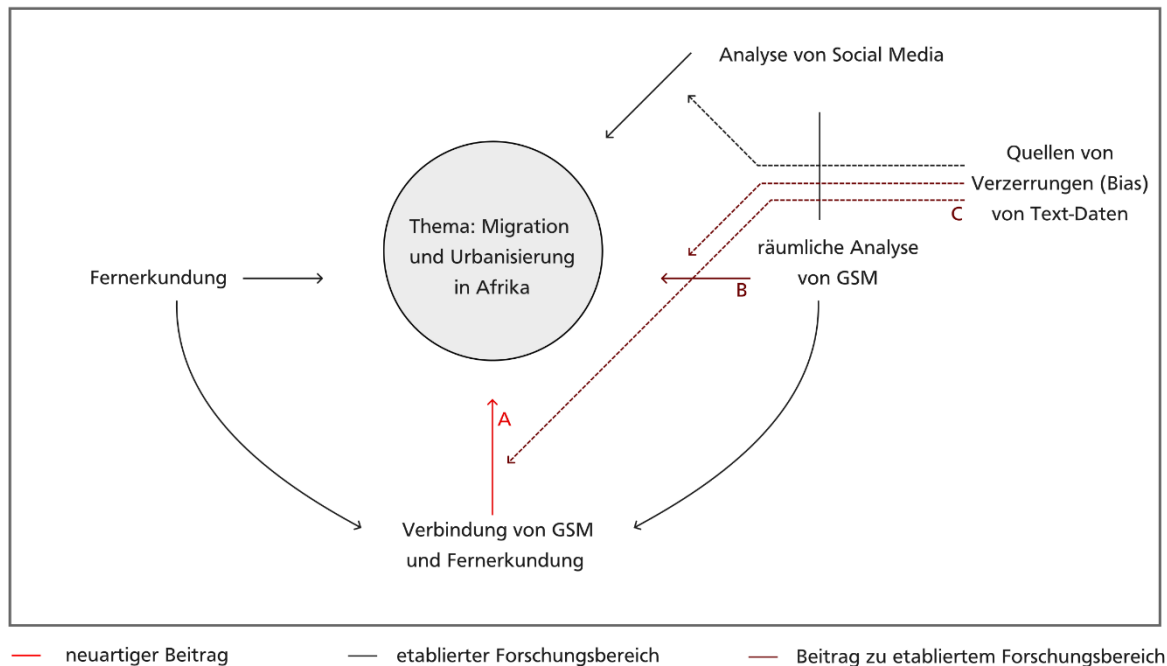
Es ist also vorstellbar, dass die geographische Auswertung von GSM-Texten, welche grundsätzlich als komplementär zu FED eingeschätzt werden können (Zhu et al. 2022), dazu beitragen kann, jene Datenbedarfe in der Kartierung von Migrationsfaktoren zu stillen, welche nicht durch FED gedeckt werden können. Angesichts dieses Potenzials erlaubt der bisherige Stand der Forschung allerdings keine ausreichende Einschätzung der Plausibilität eines solchen Ansatzes. Diese Einschätzung zu ermöglichen, ist die primäre Motivation dieser Dissertation, die damit einen Beitrag zur Schließung dieser Forschungslücke erbringen soll.

1.4. Forschungsschwerpunkte und Ziele der Arbeit

Aufgrund der vielversprechenden Möglichkeiten von GSM beschäftigt sich diese Dissertation also mit der Frage, inwieweit die fernerkundliche Messung von Migrationsfaktoren durch GSM sinnvoll ergänzt werden kann, um damit großflächige Datenbedarfe in der geographischen Erforschung von Migration zu erfüllen. Um diese Frage zu beantworten, sollen Potenziale der Nutzung von GSM in der räumlichen Analyse von Migration und der gemeinsamen Nutzung mit FED ausgelotet werden.

Die Verbindung von GSM und FED wurde bereits im Bereich der Gebäudeklassifikation (Häberle et al. 2022), des Krisenmanagements (Reuter und Kaufhold 2018), Stadtgeographie (Taubenböck et al. 2018) und der Erforschung von Landschaftswahrnehmungen (Zhu et al. 2022) demonstriert, ist jedoch im Kontext menschlicher Migration neuartig. Um Plausibilität und Herausforderungen in diesem Ansatz zu ermöglichen, gilt es, nicht nur die Verbindung von FED und GSM in diesem Kontext zu demonstrieren (Abbildung 2A), sondern auch für diesen Anwendungsbereich spezifische Potenziale von GSM (Abbildung 2B) und potenzielle Verzerrungs-Faktoren (Abbildung 2C) in ihrer räumlichen Nutzung zu identifizieren. Statt jedoch einen dieser Aspekte exklusiv zu vertiefen, zielt diese Dissertation darauf ab, einen Überblick über Herausforderungen und Möglichkeiten zu generieren, welcher perspektivisch

als Ausgangspunkt und Orientierungspunkt für vertiefende Forschung dienen kann.



Migration wird in dieser Arbeit aus geographischer Perspektive (Abbildung 3) als ein räumliches Phänomen mit sowohl physischem als auch sozialem Charakter begriffen, dessen Messbarmachung unabdinglich ist, um ein Verständnis der Welt und deren Wandels zu erlangen. Für die geographische Erforschung von Migration ist beispielsweise von Interesse, wo durch Migration beeinflusste Räume liegen, welche physischen und sozialen Charakteristika diese ausmachen und welche Funktionen sie erfüllen, welche Eigenschaften Orte für Menschen zu (un-)attraktiven Zielen machen¹⁴ und wo Orte mit diesen Eigenschaften zu finden sind. Mit dem Vorhaben, diese Aspekte zu kartieren, und schlussendlich zur Bildung geographischen Wissens beizutragen (Abbildung 1), zielt die Dissertation also auf die räumliche Messbarmachung ab und wählt dafür einen geographisch-empirischen Ansatz. Daher werden in den Studien Interaktionen zwischen Textinformationen und dem Wissen über die physischen Objekte und Raummuster der Landoberfläche sowie die Interoperabilität von GSM mit FED in den Vordergrund gerückt.

¹⁴ Faktoren, welche Immigration anregen, werden gemeinhin als Pull-Faktoren bezeichnet, während Faktoren, die Emigration begünstigen, als Push-Faktoren bezeichnet werden (Barwińska-Małajowicz and David 2014).

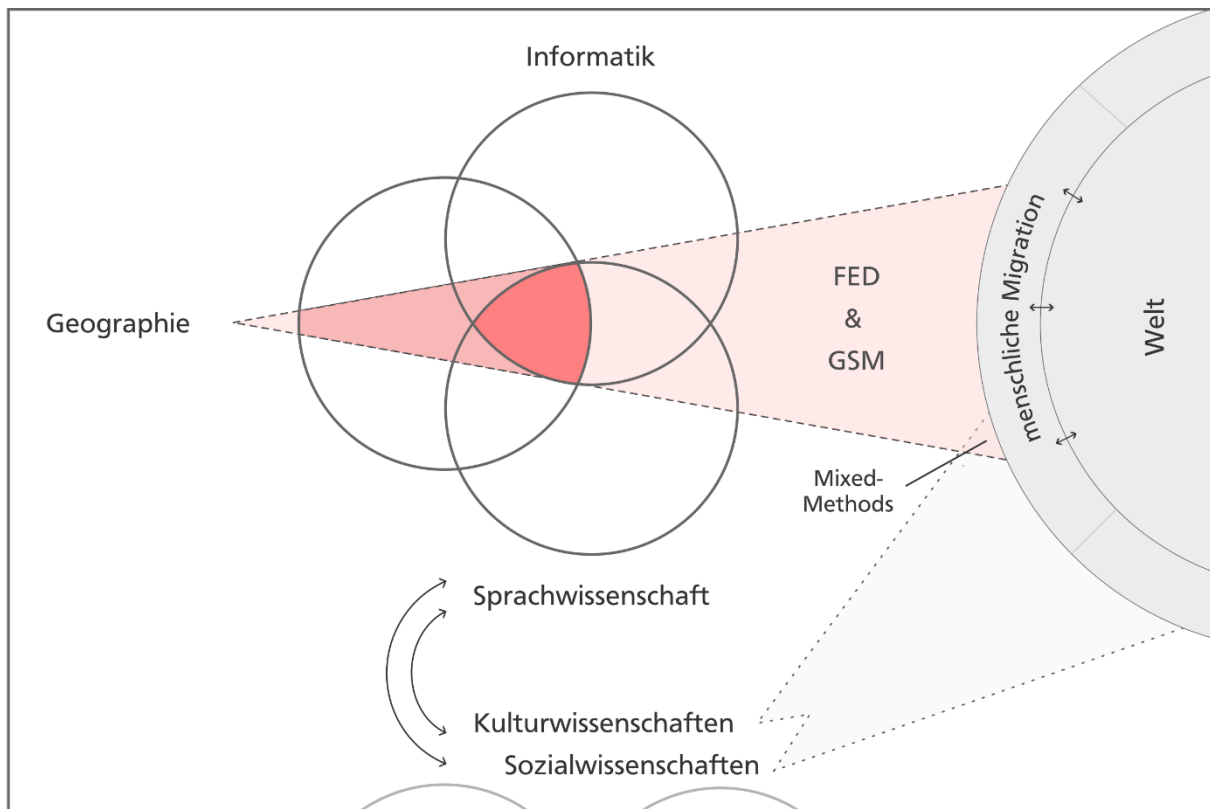


Abbildung 3: Disziplinäre Perspektive der Dissertation (hellrot). Eine Schlüsselfunktion hat die Verbindung von Methoden aus Geographie, Sprachwissenschaft und Informatik (dunkelrot). Mit diesen Methoden werden große FED und GSM-Datensätze verarbeitet, um aus geographischer Perspektive die menschliche Migration und deren Wechselwirkungen mit der Welt zu betrachten. Die Sprachwissenschaft bietet eine Brückenfunktion zu Kulturwissenschaften und Sozialwissenschaften, welche Migration mit denselben Daten aus anderen Blickwinkeln beleuchten, und perspektivisch in multidisziplinären Studien eingebunden werden können.

Die Big-Data Analysen, welche mit GSM und FED durchgeführt werden können, sind zur großflächigen Feststellung von Sachverhalten eignet, deren Bewertung jedoch außerhalb dieser Dissertation durch andere Disziplinen erarbeitet werden muss (vergleiche Albertz 2001). Um dies zu fördern, wird speziell die Verbindung zu sozialwissenschaftlichen und kulturwissenschaftlichen Ansätzen zur Migration, obwohl in dieser Dissertation nicht explizit behandelt, dennoch peripher berücksichtigt (Abbildung 3), da sie außerhalb dieser Dissertation zur Einordnung der Sachverhalte und einem holistischen Verständnis von Migration beitragen können. Beispielsweise sind zur Einordnung der quantitativen Ergebnisse der Dissertation auch Ansätze aus den Sozialwissenschaften bedeutsam, welche aus Migrationsdaten individuelle Migrationsfaktoren herausarbeiten und in Bezug zu sozialen Entwicklungen setzen können. Kulturwissenschaftliche Ansätze zu Migrationsbegriffen, Migrationstheorien und Migrationsnarrativen können den kulturellen Rahmen erfassen, in welchem Diskurse um Migration stattfinden, und welcher für ein Gesamtverständnis von Migration auch eingebunden werden muss. Es ist zu konstatieren, dass keine Disziplin allein zu einer vollständigen Beschreibung menschlicher Migration in der Lage ist, und dass Migrationsforschung daher interdisziplinär sein muss (Brettell und Hollifield 2015).

In Hinblick hierauf werden die in dieser Dissertation entwickelten und getesteten Methoden flexibel und modular konzipiert, um die spätere Integration mit sozialwissenschaftlichen und kulturwissenschaftlichen Methoden zu ermöglichen. Die Sprachwissenschaft übernimmt hierbei eine Brückenfunktion, welche die empirischen Beobachtungen der Dissertation mit den kulturwissenschaftlichen Rahmenwerken und den sozialwissenschaftlichen Ansätzen verbinden kann.

Trotz dieser Anknüpfungspunkte ist die Analyse von Migration nach kulturwissenschaftlichen oder sozialwissenschaftlichen Ansätzen kein Bestandteil der Dissertation, welcher auch keine übergreifende Migrationstheorie oder Migrationsdefinition zugrunde liegt. Stattdessen sollen, wo immer möglich, technisch flexible Ansätze erprobt werden, welche sich auf verschiedene Arten von Migration und Mobilität anwenden lassen. Die angewandten Studien der Dissertation orientieren sich zwar an den weithin akzeptierten Definitionen der Vereinten Nationen, jedoch sollen die Methoden nicht auf diese beschränkt sein.

Zur inhaltlichen Analyse der Textdaten sollen Methoden der Sprachverarbeitung zum Einsatz kommen, während Fernerkundungsdaten und geographische Positionen mit geographischen und geostatistischen Methoden verarbeitet werden. Die Arbeit vereint damit Methoden aus den Disziplinen der Geographie, der Sprachwissenschaft und der Informatik zum Zweck der Migrationsforschung und ist daher –wie die Migrationsforschung selbst– eindeutig interdisziplinär (Scholten et al. 2022).

Die Dissertation zielt dabei nicht darauf ab, alle sich aus diesem interdisziplinären Rahmen ergebenden technischen Möglichkeiten zu dokumentieren oder gar zu testen, was auch aufgrund der fortschreitenden Entwicklungen innerhalb dieser Disziplinen in gegebenen Rahmen nicht möglich ist. Vielmehr sollen grundsätzliche Prinzipien abgeleitet werden, welche die praktischen Möglichkeiten dieses Forschungsansatzes und Potenziale und Limitationen desselben bedingen. Um diese zu ergründen, werden Studien durchgeführt, welche methodisch in verschiedene Richtungen zielen, und dabei als Sondierungen und nicht als vollumfängliche Erhebungen konzipiert sind.

Neben den Potenzialen soll auch die Praktikabilität der Verbindung von GSM und FED eingeschätzt werden. Dazu gehört, dass neben der Entwicklung und Demonstration von Textdatenauswertungen auch die Verfügbarkeit dieser Daten über Raum und Themenbereiche hinweg explizit erforscht wird, sowie die Möglichkeiten, die GSM mit FED zu verknüpfen. Innerhalb GSM soll ein besonderer Fokus auf Daten von textbasierten Plattformen und speziell der Plattform Twitter (jetzt umbenannt in „X“) liegen, welche bereits für zahlreiche Studien im geographischen Kontext genutzt wurden.

1.5. Aufbau der Arbeit

Vor diesem Hintergrund werden sechs Studien durchgeführt, welche jeweils eine primäre Leitfrage adressieren, und zusätzlich aus verschiedenen Perspektiven zum möglichen Einsatz von GSM in den adressierten Themenfeldern beitragen. Abbildung 4 illustriert vor diesem Kontext die Strukturierung der Studien in die drei Blöcke

- A. Verbindung von FED und GSM im urbanen Raum,
- B. Ableitung und Einbindung von Mobilitätscharakteristika, und
- C. Identifikation inhaltlicher Verzerrungen und Potenzial des Geoparsings von Webtextdaten und speziell GSM.

Der verbindende Kontext aller Studien ist dabei immer die räumliche Analyse von GSM und FED in der geographischen Migrationsforschung, mit einem Schwerpunkt auf urbanen Räumen, welche als Ziel der meisten Migrierten besondere Relevanz haben.

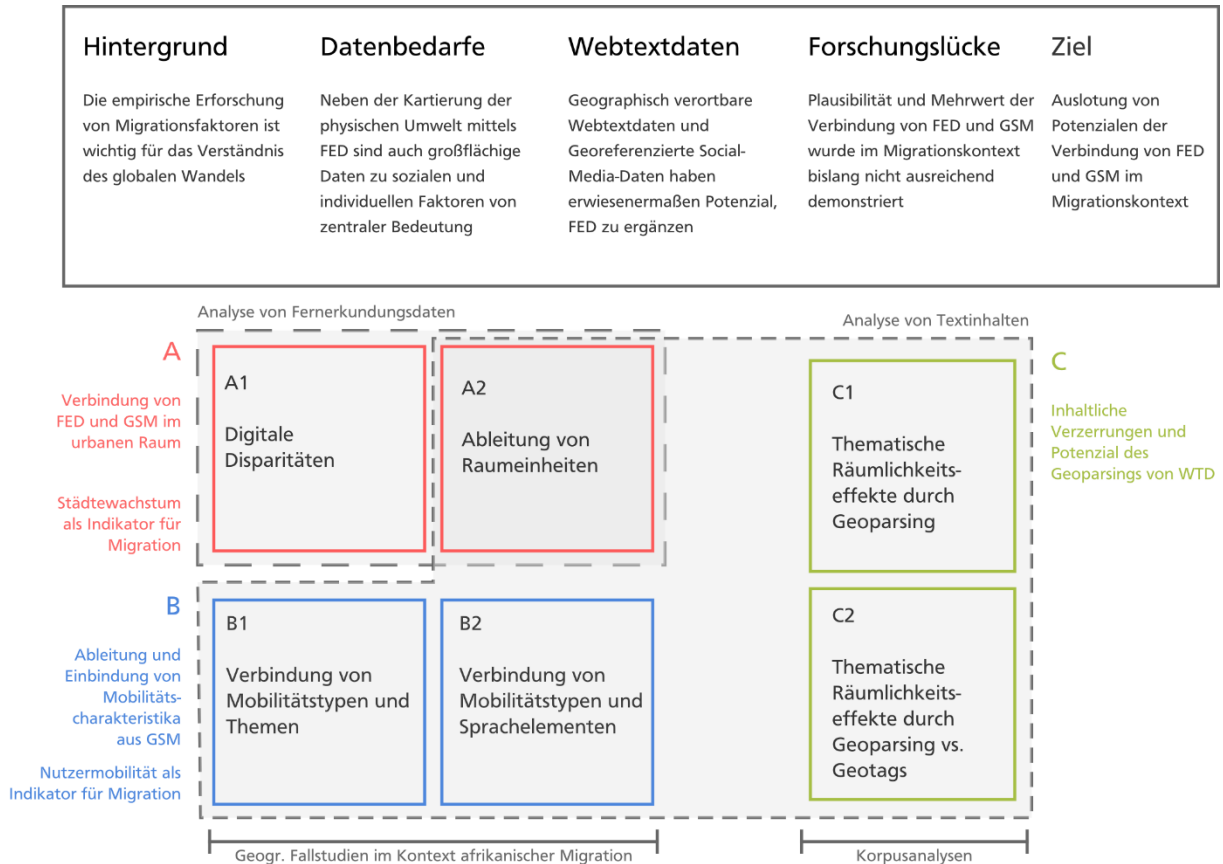


Abbildung 4: Konzept und Struktur der Dissertation

Studie A1: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Erfassung digitaler Disparitäten in Afrikas expandierenden städtischen Räumen.

In einer ersten Studie werden GSM mit FED verbunden zur Anwendung gebracht, um die gut dokumentierte baustrukturelle Expansion afrikanischer Städte um eine digitale Komponente zu erweitern. Ziel ist es dabei, mögliche digitale Disparitäten aus Verhältnis zwischen physischer und digitaler Expansion zwischen neu entstandenen und älteren Teilen der Städte zu identifizieren.

Aufgrund des Potenzials für weitere Urbanisierung kann Afrika, wie zu Beginn dieses Kapitels dargestellt, als Ort einer *Urban Frontier* betrachtet werden (Goodfellow 2022). FED zeigen, wie im Rahmen des hohen Bevölkerungswachstums afrikanischer Städte riesige Flächen bebaut und besiedelt werden (Taubenböck et al. 2024). Wie diese neuen Siedlungsflächen, welche für Zuwanderung mutmaßlich eine besonders große Rolle spielen, im digitalen Raum repräsentiert sind, ist von entscheidender Bedeutung für ihre Attraktivität und die Wettbewerbsfähigkeit (Törnberg 2023). Deshalb ist die digitale Komponente des urbanen Wachstums, also die Ausprägung von Online-Repräsentation innerhalb der neuen Siedlungsflächen, bedeutsam für das Verständnis der Wachstumsprozesse.

Es soll daher durch die Verschneidung von Twitter-Dichte und Bebauungsalter getestet werden, ob das Wachstum dieser digitalen Komponente mit dem physischen Wachstum Schritt hält, oder stattdessen eine Ungleichheit oder Disparität besteht, bei welcher sich die digitale Repräsentation der Städte auf die bereits länger bestehenden Kernzonen beschränkt.

Die Studie fokussiert dabei räumlich die am schnellsten gewachsenen Siedlungsgebiete in Afrika. Sie soll damit auch einen Einblick in die Datenverfügbarkeit für diese migrationsrelevanten Siedlungsgebiete erlauben.

***Leitfrage LA1:** Lässt sich durch die Verschneidung von FED und GSM eine Aussage über Muster digitaler Ungleichheiten in Afrikas expandierenden städtischen Räumen treffen?*

Studie A2: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Ableitung von Raumeinheiten für die explorative Analyse von physischen und thematischen Charakteristika urbaner Räume

Für das Verständnis urbaner Orte und Räume ist nicht nur die Präsenz von dort verorteten digitalen Beiträgen relevant, sondern auch deren Inhalte. Diese können mit den sozialen Aktivitäten und den Funktionen von Stadtteilen und Nachbarschaften in enger Verbindung stehen (Jheng et al. 2025). Eine Charakterisierung dieser Raumeinheiten kann sowohl von GSM als auch von FED profitieren. Die Vereinigung von Textinhalten dieser GSM mit FED ist also vielversprechend, aber aufgrund ihrer unterschiedlichen Formate herausfordernd.

Daher soll in Studie A2 ein praktikabler und modularer Ansatz entwickelt werden, welcher flexibel den Anforderungen der Migrationsforschung angepasst werden, und als Grundlage für Studien multi-modaler Daten dienen kann.

Der erste Schritt in einer datengetriebenen Analyse ist häufig die Darstellung für explorative Zwecke, welche bei großen geographischen Datensätzen die Aggregation zu Raumeinheiten miteinschließt. Sinnvolle Raumeinheiten, welche die untersuchten Phänomene darstellen, sind dabei essenziell (Weigand et al. 2019). Für die großflächige Analyse sollten diese Raumeinheiten auch großflächig und über Staaten hinweg konsistent verfügbar sein. Die Nutzung von administrativen Grenzen ist in Bezug auf diese Kriterien oft nicht optimal (Taubenböck et al. 2019). Wünschenswert wären stattdessen Raumeinheiten, welche sich an der jeweiligen Forschungsfrage und insbesondere den dafür relevanten Daten orientieren. Für stadtgeographische Fragestellungen könnten dies beispielsweise Stadtteile, Nachbarschaften, Blöcke oder Bezirke sein, welche sich sowohl durch physische als auch kulturelle Faktoren von der Umgebung abgrenzen –etwa durch das Alter der Gebäude und die lokalen sozialen Aktivitäten. Diese Aspekte, welche indirekt in Bezug zu Migrationsfaktoren stehen, lassen sich jeweils durch FED beziehungsweise Webtextdaten abbilden.

Anhand dieser Datenquellen wird daher eine Methode getestet, welche Raumeinheiten direkt aus den multimodalen zu analysierenden Daten generiert: Die Überführung von Texten in numerische Repräsentationen im Rasterformat ermöglicht die Gruppierung der Pixel in räumliche Cluster basierend auf ihrer geographischen und textinhaltlichen Distanz zueinander (Chavent et al. 2018). Dies soll zu einer Unterteilung des urbanen Raumes basierend auf räumlicher, physischer und textinhaltlicher Ähnlichkeit und damit zu nützlichen Raumeinheiten führen.

Zur Bewertung der Methode wird statistisch untersucht, inwiefern die so generierten Cluster eine ausreichende Kompaktheit und innere Homogenität in Erbauungsjahr und Textinhalten (speziell Themen) aufweisen. Zusätzlich werden die Cluster qualitativ interpretiert, indem

eine explorative Analyse mit Unterstützung von Webrecherche und öffentlichen Kartendaten durchgeführt wird. Dieser Analyseschritt erlaubt eine Einschätzung der Nutzbarkeit und Interpretierbarkeit der Cluster in der praktischen Anwendung.

Leitfrage LA2: *Ermöglicht die Umwandlung von Textinhalten in ein Rasterformat und geographisch-textinhaltliches Clustering die Generierung von interpretierbaren urbanen Raumeinheiten für die Analyse mittels FED und GSM?*

Studie B1: Differenzierung textinhaltlicher Präferenzen von nigerianischen Twitter-Nutzenden nach Mobilitätscharakteristika

Die in Studien A1 und A2 untersuchten urbanen Flächen haben insofern Migrationsbezug, als sie vor Bebauung mutmaßlich nicht bewohnt wurden¹⁵, ihre späteren Bewohner also davor noch nicht dort gelebt haben. Darüber hinaus kann allerdings nichts über Art und Herkunft der Zugezogenen gesagt werden, beispielsweise ob sie aus nahen Siedlungen oder aus anderen Ländern stammen. Angesichts der vielen möglichen Migrationsarten bestünde ein erheblicher Vorteil darin, Mobilitätsmerkmale der Nutzenden zur Einordnung der von ihnen geposteten GSM-Inhalte heranziehen zu können. Hierdurch könnten Textinhalte nicht nur räumlich, sondern auch demographisch-räumlich kontextualisiert werden.

In Hinsicht auf die Ableitung von Mobilitätsmerkmalen ist der FED-Ansatz der Gebäudeflächenkartierung jedoch limitiert. GSM sind in dieser Hinsicht vielversprechender. Die Möglichkeit, Mobilitätscharakteristika wie Migrationstypen von Nutzenden aus deren GSM-Historien abzuleiten, wurde bereits in mehreren Studien unter Beweis gestellt (Zaghene et al. 2014; Kim, Sirbu, Rossetti, et al. 2022). Studie B1 erprobt daher eine Methode zur Verbindung solcher Mobilitätscharakteristika mit Textinhalten. Hierbei werden speziell thematische Bereiche (engl. *Topics*) als Textinhalte genutzt (Antypas et al. 2022). Als Untersuchungsregion dient dabei das westafrikanische Land Nigeria.

Bei der Analyse von Inhalten in einer neuen Untersuchungsregion sind inhaltliche Kategorisierungen zwar hilfreich, aber anfällig für Subjektivität (Mekaoui et al. 2025). Insbesondere bei der großflächigen Anwendung ist es herausfordernd, Kategorien zu finden, welche die Perspektive der Menschen korrekt repräsentieren. Eine offene Klassifikation auf Basis von Mustern innerhalb der Daten ist zwar möglich, erschwert jedoch die Integration von fachspezifischem und sozio-kulturellem Wissen (Schofield et al. 2025). Daher wird in dieser Studie ein Domänenadaptionsansatz eingesetzt, um Themen aus Textdaten eines nigerianischen Online-Forums zu lernen, die einem ähnlichen geographisch-sozialen Umfeld (nigerianische Twitter-Nutzende) wie dem der zu analysierenden Tweets entsprechen.

Die Studie implementiert weiter einen Ansatz zur Datenakquise, welcher sowohl Geotweets als auch Tweets ohne Geotags einbezieht. Dies ist motiviert durch die Annahme, dass die beiden Tweet-Typen sich inhaltlich unterscheiden, was die Studien C1 und C2 spezifisch analysieren. Hier wird zunächst ein Ansatz demonstriert, welcher Datensätze mit sowohl Geotweets als auch Tweets ohne Geotags erstellen kann.

¹⁵ Und nach Bebauung zumindest teilweise bewohnt werden, was nicht notwendigerweise überall der Fall ist. Aufgrund der zunehmenden Bevölkerung Afrikas ist es jedoch wahrscheinlich, dass die betrachteten Gebiete keine Geisterstädte sind (Zheng et al. 2017) und die Gebäude tatsächlich als Wohnraum genutzt werden (Shi et al. 2025).

Die Studie zeigt in der thematischen Analyse dieser Tweets, welche der Themen überproportionale Anteile der Kommunikation von Migrierten ausmachen, was Rückschlüsse auf Migrationsfaktoren erlaubt. Sie demonstriert dabei, dass die Verbindung von Mobilitätscharakteristika und semantischen Textinhalten möglich und im Migrationskontext sinnvoll ist.

Leitfrage LB1: Können thematische Präferenzen von Twitter-Nutzenden aus deren Texten abgeleitet und in Verbindung zu Mobilitätscharakteristika der Nutzenden gesetzt werden?

Studie B2: Analyse der Verbreitung von Sprachmerkmalen nach Mobilitätscharakteristika der Autoren am Beispiel westafrikanischer Twitter Nutzender.

Neben den semantischen Inhalten der Texte sind auch sprachliche Merkmale von Bedeutung für Migration. Sprache und Mobilität stehen in einer starken wechselseitigen Beziehung (Canagarajah 2017). Einerseits ist Sprache ein wichtiges Element bei Migrationsentscheidungen, umgekehrt hat menschliche Mobilität auch einen Einfluss auf die Entwicklung und Verbreitung von Sprachmerkmalen (Borlongan 2023).

Um auszuloten, inwiefern die großflächige Ableitung von Sprachmerkmalen in den genutzten GSM plausibel ist, werden lexikalische Merkmale einer lokalen Form des Englischen in Nigeria (fortan *LEN*) aus Web-Wörterbüchern abgeleitet und algorithmisch in einem großen Twitter-Datensatz von westafrikanischen Twitter-Nutzenden identifiziert. Textcharakteristika wie Themen und Nutzercharakteristika zu Mobilität und Aufenthaltsorten werden algorithmisch identifiziert und in Beziehung zur Verbreitung der detektierten Sprachmerkmale gesetzt.

Die Studie soll demonstrieren, welche Informationen über Mobilität der Nutzenden durch Analysen von geschriebener Sprache (wie in Studie B1) abgeleitet werden können, und gleichermaßen die Herausforderungen herausarbeiten, die in einem algorithmischen Ansatz zu sprachwissenschaftlichen Analysen liegen.

Leitfrage LB2: Kann die Verwendung von charakteristischen sprachlichen Merkmalen aus Twitter-Texten abgeleitet und in Verbindung zu Mobilität und Ort der Nutzenden gesetzt werden?

Studie C1: Analyse der Räumlichkeit von Themen in englischen Webtextdaten

Die Klassifizierung von Textdatensätzen in semantische *Themen* ist eine weit verbreitete Methode zur Analyse von Webtextdaten und stellt auch innerhalb dieser Dissertation einen wichtigen Bestandteil dar. Ihre räumliche Analyse ist jedoch ein noch emergentes Forschungsfeld und muss kritisch betrachtet werden. Themen werden sowohl mittels unüberwachter (Studie A2 und B2), als auch überwachter Ansätze (Studie B1) des maschinellen Lernens erstellt. Sie können selbst große Datensätze gut strukturieren und zugänglich machen und sind eine weithin genutzte Methode des Text-Minings (Egger und Yu 2022). Jedoch kann nicht davon ausgegangen werden, dass alle Themen gleichsam in georeferenzierten Daten verfügbar sind, beziehungsweise, dass der Anteil georeferenzierter Inhalte über Themen hinweg konstant ist.

So haben Texte mit dem Thema Naturgefahren in der Regel einen klaren Bezugsort, an dem sich die Naturkatastrophe beispielsweise ereignet hat oder die Reaktion darauf koordiniert wird (Kersten und Klan 2020). Demgegenüber haben Texte, die sich mit Epistemologie¹⁶

¹⁶ einem Feld der Philosophie, welches sich mit Wissensgewinnung beschäftigt (Steup und Neta 2005)

befassen, weniger klare Bezugspunkte. Auch für zahlreiche andere Themen lässt sich, zumindest intuitiv, eine mehr oder weniger ausgeprägte Neigung diesbezüglich feststellen. Dieser Effekt, welcher in dieser Dissertation als Heterogenität von Themen in geographischer Räumlichkeit (engl. *topical geospatiality*) betrachtet wird, kann bei der geographischen Analyse eine Quelle von Verzerrungen darstellen, die beim Blick durch die Linse einer georeferenzierten Stichprobe zur inkorrekten Darstellung von Webdaten führen könnten.

Um diese Intuition empirisch zu prüfen, werden zusätzlich zu den Twitter-Daten und weitere Textkorpora von Nairaland (Basmus et al. 2019), Reddit (Völske et al. 2017), Stackexchange (Stack Exchange Inc 2024) und GDELT (Leetaru und Schrodt 2013) gesammelt und mittels Geoparsing (Hu et al. 2023a) auf Erwähnungen von geographisch lokalisierbaren Orten getestet. Die Verbindung zwischen Themen und der Häufigkeit der Rauminformation wird daraufhin in einem gemischten Modell (Gries 2015) statistisch modelliert, womit der Einfluss von Textlänge, Zeitpunkt und autorspezifischen Neigungen kontrolliert werden kann. Die themenspezifischen Koeffizienten des Modells erlauben schlussendlich Aufschluss über inhärente Neigungen von Themen, räumliche Informationen zu enthalten, und damit auf thematische Geospatiality-Effekte. Zudem wird die Ausprägung der Koeffizienten über die verschiedenen Datenquellen und damit die verschiedenen Textsorten verglichen.

***Leitfrage Lc1:** Hat das Thema eines Webtextes einen Einfluss auf dessen Wahrscheinlichkeit, eine geographisch verortbare Ortsreferenz zu enthalten, und wie unterscheidet sich dies nach Textsorte?*

Studie C2: Differenzierung der Räumlichkeit von Themen zwischen Geotag- und Geoparsing-basierten Ansätzen

Geotag-Funktionen sind wertvoll zur Generierung georeferenzierter Textdatensätze, werden jedoch nur von wenigen Quellen angeboten. Die Nutzung von Geoparsing, also der Verortung von textuellen Ortserwähnungen wie in Studie C1, stellt dazu eine generisch anwendbare Alternative und Ergänzung dar. Insbesondere im Hinblick auf die Volatilität in der Nutzung und Verfügbarkeit von Geotags (Kruspe et al. 2021) ist eine Diversifizierung von Georeferenzmethoden sinnvoll. Zudem können größere Anteile der Datensätze und auch diversere Datenquellen für die räumlichen Analysen nutzbar gemacht werden, wenn neben Geotag-basierten Ansätzen auch Geoparsing-basierte Ansätze (Hu et al. 2023b) genutzt werden.

Es ist jedoch nicht klar, ob die Repräsentation von Textinhalten über die beiden Ansätze hinweg gleich ist. Möglicherweise interagieren Inhalte wie Themen anders mit Ortserwähnungen als mit Geotags. Basierend auf Studie C1, welche die grundsätzliche Existenz von themenspezifischen Effekten analysiert, ist also von Interesse, ob sich diese Effekte unterscheiden, je nachdem welcher der Ansätze genutzt wird.

Daher wird in dieser weiteren Studie für einen Twitter-Datensatz, welcher beide Ansätze ermöglicht, die themenspezifische Geospatiality analog zur Studie C1, aber separat für Geotag- und Geoparsing-basierte Georeferenz untersucht.

***Leitfrage Lc2:** Unterscheiden sich Geoparsing- und Geotagging-basierte Ansätze zur Verortung von Texten hinsichtlich der themenspezifischen Einflüsse auf die Verortbarkeit?*

Die Inhalte dieser Arbeit sind weiter wie folgt gegliedert:

Kapitel 2 fasst die für die Dissertation relevante Literatur zusammen, wobei ein Fokus auf dem Forschungsstand zu den in den jeweiligen Studien genutzten Methoden liegt.

Kapitel 3 konzeptualisiert die im Rahmen der Dissertation durchgeführten Studien. Dabei werden die Überlegungen bezüglich der Daten und Methoden dargestellt.

Kapitel 4 führt das Untersuchungsgebiet Afrika ein und beschreibt die genutzten Daten. Diese Daten werden oft in mehreren der Studien verwendet, jedoch studienspezifisch unterschiedlich verarbeitet.

Kapitel 5 beschreibt die studienspezifischen Methoden im Detail, inklusive der angewandten Datenverarbeitung.

Kapitel 6 präsentiert und interpretiert die Ergebnisse der einzelnen Studien.

Kapitel 7 diskutiert die Ergebnisse über Studien hinweg und interpretiert sie in Bezug auf identifizierte Potenziale und Limitationen, die sich im Rahmen der Dissertation gezeigt haben.

Kapitel 8 fasst die Erkenntnisse zusammen und zieht ein übergeordnetes Fazit der Dissertation. Zudem wird eine Einschätzung der Relevanz der Dissertation sowie ein Ausblick auf künftige Entwicklungen gegeben.

Die im Rahmen dieser Dissertation durchgeführten Studien wurden in drei begutachteten Zeitschriftenbeiträgen (Studien A1, B1, C1), zwei Konferenzbeiträgen (Studien A2 und C2) und einem Buchkapitel (Studie B2) veröffentlicht. Die Inhalte dieser Dissertation basieren daher in Teilen auf folgenden Publikationen:

1. Mast, Johannes, Marta Sapena, Henri Debray, Justice Nana Inkoom, Richard Lemoine-Rodríguez, Christian Geiß, und Hannes Taubenböck. 2025. „The Digital Urban Frontier: Disparities in Social Media Activity between Consolidated and Newly Urbanized Areas in Africa“. *Applied Geography* 181 (August):103687. <https://doi.org/10.1016/j.apgeog.2025.103687>.
2. Mast, Johannes, Christian Geiß, und Hannes Taubenböck. 2025. „Spatial delineations of cities based on web content and built-up age – a case study on North African growth poles from 2010 to 2020“. Konferenzartikel präsentiert auf der 17th International Conference on Joint Urban Remote Sensing (JURSE 2025), Tunis.
3. Mast, Johannes, Marta Sapena, Martin Mühlbauer, Carolin Biewer, und Hannes Taubenböck. 2024. „The Migrant Perspective: Measuring Migrants’ Movements and Interests Using Geolocated Tweets“. *Population, Space and Place* 30 (2): e2732. <https://doi.org/10.1002/psp.2732>.
4. Mast, Johannes, Dominik Heps, und Marta Sapena. 2026. „Using geolocated social media data for understanding the relationships between language, topic and mobility: A case study on a local use of English in Nigeria.“. In: *Geolinguistic Studies on Urban Space: Global Identities and Local Connectedness*, in *New Directions in World Englishes Research*. Edinburgh: Edinburgh University Press.
5. Mast, Johannes, Richard Lemoine-Rodríguez, Vanessa Rittlinger, Martin Mühlbauer, Carolin Biewer, Christian Geiß, und Hannes Taubenböck. 2025. „Geospatiality: The Effect of Topics on the Presence of Geolocation in English Text Data“. *International Journal of Geographical Information Science*. <http://dx.doi.org/10.1080/13658816.2025.2460051>.
6. Mast, Johannes, Richard Lemoine-Rodríguez, Vanessa Rittlinger, Christian Geiß, und Hannes Taubenböck. 2025. „Comparing Geospatiality of Topics between Geotag- and Geoparsing-based Geolocations“. In *3rd International Workshop on Geographic Information Extraction from Texts Workshop, GeoExT 2025*, herausgegeben von C CEUR-WS.org, Bd. 3969. CEUR Workshop Proceedings“. CEUR-WS. <https://elib.dlr.de/214374/>.

2. Grundlagen

Die Dissertation fokussiert sich im Kontext menschlicher Migration auf die neuartige Kombination und Anwendung etablierter Methoden und Daten. In diesem Kapitel werden die hierfür relevanten Methoden grundlegend dargestellt.

Zunächst werden die Methoden, auf welchen die Prozessierungen der Daten beruhen, grundlegend eingeführt (2.1). Sodann wird der Stand der Forschung zu verzerrenden Einflüssen der Daten, welche speziell bei den konzipierten Anwendungen im geographischen Bereich berücksichtigt werden müssen, zusammengefasst (2.2). Zuletzt präsentiert Sektion (2.3) den Forschungshintergrund zu den Anwendungen an der räumlichen Intersektion von GSM und FED im Migrationskontext, welche im Rahmen der Dissertation erprobt werden.

2.1. Methodische Grundlagen

2.1.1. Modelle und Maschinelles Lernen für Fernerkundung und Textprozessierung

In der Datenanalyse von Fernerkundungsdaten und Textdaten nehmen Modelle eine entscheidende Rolle ein. Statistische Modelle kommen in der Inferenz zum Einsatz, um innerhalb von Daten auf die Beziehung zwischen Variablen zu schlussfolgern. Komplexere statistische Modelle, insbesondere die des maschinellen Lernens, werden auch in der Prozessierung von komplexen Daten genutzt, beispielsweise um Datenpunkte wie Pixel, Texte, oder Worte in verschiedene Kategorien einzuordnen.

Grundsätzlich ist ein statistisches Modell eine informative, normalerweise vereinfachte Darstellung eines realen Systems oder Zusammenhanges mittels funktionaler Gleichungen. Ein vergleichsweise einfaches statistisches Mittel sind lineare Regressionsmodelle, welche einen linearen Zusammenhang zwischen einer unabhängigen und einer abhängigen Variable darstellen, aber auch auf multivariate Probleme erweitert werden können (Russell und Norvig 2023). Die Parameter, welche den Einfluss der Variablen beschreiben, können aus empirischen Daten abgeleitet (geschätzt) werden. Die Methode, mit welcher beispielsweise die Parameter einer linearen Regression aus den Daten geschätzt werden, ist die Methode der kleinsten Quadrate (Gauss 1823; Freedman 2009).

Für die Schätzung komplexerer Modelle existieren andere Ansätze, welche dem Teilgebiet des Maschinellen Lernens angehören. Der Begriff Maschinelles Lernen beschreibt eine große Bandbreite von Techniken zur Kalibrierung von Modellen, bei welcher die Parameter aus den Daten abgeleitet werden (Russell und Norvig 2023). Unter den verschiedenen Teilgebieten des Feldes spielen in dieser Dissertation vor allem die Paradigmen des überwachten, unüberwachten und selbstüberwachten Lernens eine Rolle (Abbildung 5).

Bei **überwachtem Lernen** werden die Zielkategorien aus vorgegebenen Beispielen gelernt. Das Anpassen eines Modells auf die Beispiele wird als *Trainieren* des Modells bezeichnet. Hierfür müssen die Zielkategorien bereits im Voraus definiert, sowie in den Beispielen (den Trainingsdaten) annotiert sein. **Unüberwachtes Lernen** hingegen benötigt weder vordefinierte Zielkategorien noch annotierte Daten. Datenpunkte werden nach Ähnlichkeit ihrer Merkmale gruppiert, sodass Gruppen entstehen, welche in Bezug auf diese Merkmale möglichst homogen sind. Die Ergebnisoffenheit dieses auch als *Clustering* oder *Cluster-Analyse* (Albertz, 2001) bezeichneten Ansatzes kann als Vorteil betrachtet werden, eine Herausforderung besteht jedoch in der Interpretation der entstehenden Gruppen (engl. *Cluster*), welche durch die Anwender oder Experten geschehen muss (Schofield et al. 2025).

Selbstüberwachtes Lernen ist eine Trainingsmethode, bei welcher die Trainingsdaten durch einfache Manipulation nicht-annotierter Daten erzeugt werden. Dies geschieht oft im Rahmen einer Voraufgabe, welche nur dem Training und nicht der eigentlichen Anwendung dient. Beispielsweise kann eine Voraufgabe darin bestehen, absichtlich generierte Lücken in Texten oder Bildern zu füllen, und dabei den ursprünglichen Texten oder Bildern möglichst nahe zu kommen. Auch wenn die Voraufgabe nicht notwendigerweise nützlich ist, lernt das Modell hierbei die Zusammenhänge innerhalb der Daten, was eine gute Basis für den späterem Transfer auf Folgeaufgaben (engl. *downstream-tasks*) darstellt.

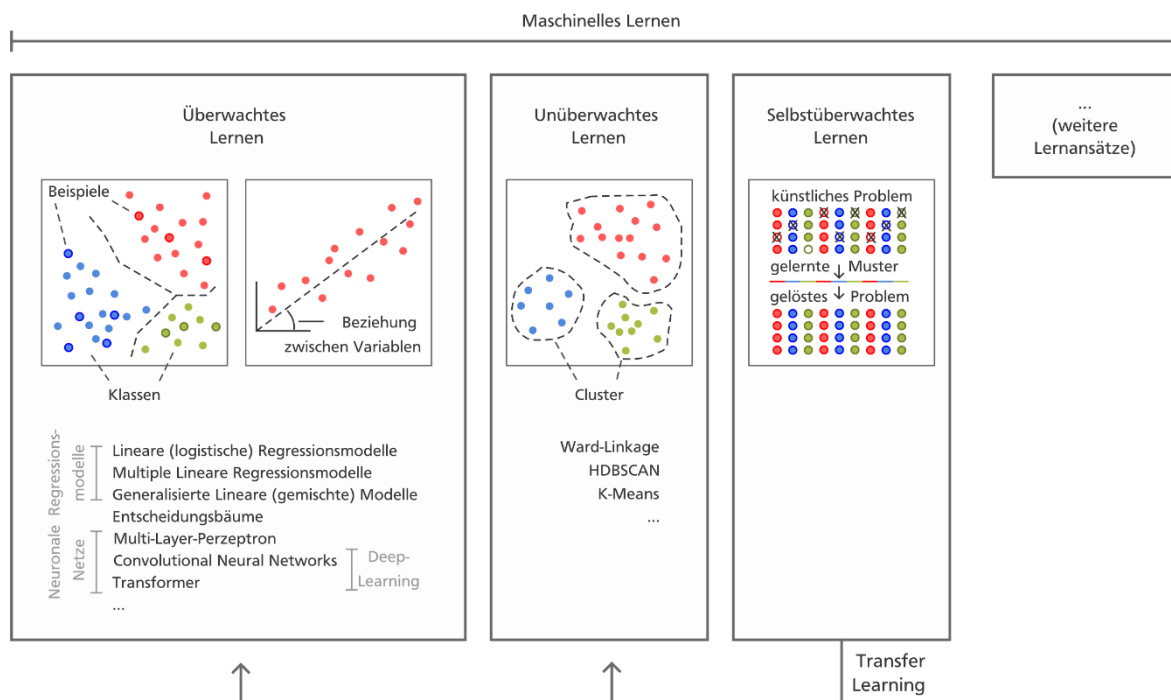


Abbildung 5: Überblick über die in der Dissertation genutzten maschinellen Lernverfahren

Es existieren zahlreiche Modellarchitekturen unterschiedlicher Komplexität und zahlreiche verschiedene Lernansätze. Für eine ausführliche Einführung zu maschinellem Lernen in der Fernerkundung wird auf die Übersichtsarbeit von Russell und Norvig (2023) verwiesen. Hier sollen nur jene Methoden genannt werden, welche im Rahmen dieser Dissertation Anwendung finden:

Die bereits erwähnte lineare Regression stellt einen Ansatz für Inferenz dar, welcher sich für die Modellierung linearer Zusammenhänge eignet und diese auf transparente Weise illustriert. Die Nutzung von verschiedenen Transformationen und Funktionen erlaubt es Regressionen auch, nichtlineare Zusammenhänge darzustellen und neben kontinuierlichen Werten auch diskrete Klassifizierungen zu modellieren. Gemischte Modelle (engl. Mixed Models) erlauben die Schätzung fester Effekte unter Berücksichtigung zufälliger Effekte, die den zusätzlichen Einfluss von Gruppen schätzen. Sie eignen sich daher beispielsweise für die Inferenz unter Kontrolle studienspezifischer Effekte in Metaanalysen oder personenspezifischer Effekte in sozialwissenschaftlichen Studien.

Für die Herausforderungen der Text- und Bildprozessierung werden einfache lineare und gemischte Modelle nur begrenzt eingesetzt. Für diese Anwendungen hat sich in letzten Jahren eindeutig der Ansatz des tiefen maschinellen Lernens (engl. *deep learning*) mittels neuronaler Netze (engl. *neural networks*) als leistungsfähigste Methode etabliert und dabei zu bemerkenswerten Verbesserungen geführt (Russell und Norvig 2023). Der Begriff umfasst eine breite Familie von Techniken, welche darauf basieren, dass mehrere Ebenen (engl. *layer*) verhältnismäßig einfacher Rechenelemente hintereinander geschaltet und dabei miteinander verbunden werden, um tiefe neuronale Netze zu generieren. Die inneren Layer dieser Netze können im Laufe des Trainings auch komplexe Merkmale (engl. *features*) aus den Daten lernen, die für die jeweilige Aufgabe nützlich sind. Hierdurch reduziert sich die Notwendigkeit, Merkmale im Voraus manuell zu berechnen (Russell und Norvig 2023). Die Leistung und Vielseitigkeit von tiefen Modellen gehen aber auch mit Herausforderungen einher. Denn die komplexen Modelle sind verhältnismäßig schwer interpretierbar und ihr Training erfordert eine große Menge an annotierten Daten (Russell und Norvig 2023). Aufgrund der großen Datenmengen, welche von Satelliten bereitgestellt werden, ist die Anwendung von tiefen maschinellen Lernverfahren in der Fernerkundung weit verbreitet (F. Li et al. 2023).

Während sich für einfache lineare Zusammenhänge optimale Lösungen direkt berechnen lassen (Gauss 1823), ist dies bei komplexeren Modellen und Zusammenhängen nicht mehr der Fall. Als weit verbreitete Lösung hat sich stochastischer Gradientenabstieg (engl. *stochastic gradient descent*) etabliert, bei welchem das Modell schrittweise durch inkrementelle Änderung der Parameter angepasst wird, bis der Fehler des Modells ausreichend niedrig ist (das Modell konvergiert). Dieses Verfahren lässt sich auf multivariate Probleme und komplexe Modelle wie neuronale Netze übertragen (Russell und Norvig 2023).

Während des Trainings auf einer bestimmten Aufgabe erlernen Modelle oft Merkmale, welche auch bei der Lösung ähnlicher Probleme hilfreich sind. Ein Modell für die Bildverarbeitung, welches darauf trainiert wurde, Fahrräder von Motorrädern zu unterscheiden, könnte auch für die Erkennung von Einrädern eine gute Basis darstellen. Zur Anpassung auf diese neue Aufgabe wären weniger Beispiele von Einrädern notwendig, als wenn mit einem komplett neuen Modell begonnen würde (Russell und Norvig 2023). Die Übertragung von Modellen von einer Aufgabe auf die nächste wird als Transfer-Learning bezeichnet und ist eine weit verbreitete Methode insbesondere in der modernen Sprachprozessierung. Häufig wird bei NLP-Projekten das Sprachmodell nicht von Grund auf trainiert, sondern auf vortrainierte Modelle zurückgegriffen, welche auf großen Datenmengen trainiert wurden. Üblicherweise beginnen NLP-Projekte heute mit einem vortrainierten Modell (Russell und Norvig 2023). Auch in der Fernerkundung ist die Nutzung von vortrainierten Modellen zunehmend verbreitet (Ma et al. 2024). Sehr große Modelle werden in der Regel als Basismodelle (engl. *foundation models*) bezeichnet. Diese werden typischerweise mittels selbstüberwachter Trainingsmethoden auf sehr großen Datensätzen vortrainiert. Liegen für die anvisierte Aufgabe annotierte Daten vor, kann das vortrainierte Modell dann auf diese Aufgabe optimiert werden (engl. *finetuning*). Das finetuning von vortrainierten Modellen auf bestimmte Anwendungen erlaubt vergleichbare oder bessere Leistung bei geringeren Kosten an Zeit und Daten (Russell und Norvig 2023).

Bezüglich der Verarbeitung von GSM und FED ist bemerkenswert, dass Methoden des maschinellen Lernens selten auf einen Anwendungsbereich beschränkt sind, und sich verhältnismäßig gut transferieren lassen. Einige Beispiele hierfür: Frühe Modelle zur Klassifizierung von Texten bedienten sich der Wahrscheinlichkeiten einzelner Wörter (Bag-of-

Words-Modelle (Russell und Norvig 2023). Dieses Konzept wurde als *Bag of Visual Words* auch in der Verarbeitung von Fernerkundungsbildern angewendet, wobei Bildmerkmale die Rolle der Wörter einnahmen (Yang et al. 2007). *Convolutional Neural Networks* wurden für Anwendungen in maschinellern Sehen (engl. Computer Vision) entwickelt. Ihre Fähigkeit, effektiv Nachbarschaftskontext einzubinden qualifiziert sie jedoch auch für bestimmte Textanalysen (Russell und Norvig 2023; Yin et al. 2017). Die Transformer-Architektur, welche zu enormen Entwicklungen im Bereich der künstlichen Intelligenz geführt hat, wurde primär für die Verarbeitung natürlicher Sprache entwickelt, stellt inzwischen aber als Vision Transformer bezeichnet auch einen Teil des State-of-the-Art in der Fernerkundungsdatenprozessierung dar (Wang et al. 2024).

2.1.2. Embeddings und Sprachprozessierung

Moderne Methoden der Sprachprozessierung machen Texte mathematisch verarbeitbar, indem sie diese in einen numerischen kontinuierlichen Vektorraum überführen, in welchem der Inhalt von Texten oder Textteilen als Vektor von Zahlen repräsentiert wird (Havas et al. 2021). Diese Vektoren sind sich bei syntaktisch, semantisch oder thematisch ähnlichen Worten (beispielsweise „König“ und „Königin“) ähnlicher als bei unverwandten Worten (beispielsweise „König“ und „automatisch“). Diese Vektoren können auch sinnvoll miteinander verrechnet werden. So würde die Operation $\vec{König} - \vec{Mann} + \vec{Frau}$

beispielsweise den Vektor für $\vec{Königin}$ ergeben (Abbildung 6). Diese Operationen können theoretisch als numerische Repräsentation semantischer Beziehungen betrachtet werden (beispielsweise *Geschlechterzugehörigkeit* oder *Königlichkeit* in Abbildung 6), wobei nicht notwendigerweise davon ausgegangen werden kann, dass die Vektoren direkt menschlich interpretierbar sind.

Bei der Erstellung dieser Vektoren können Wörter einzeln oder im Kontext des sie umgebenden Textes berücksichtigt werden. Auch ganze Sätze können in Vektoren *eingebettet* werden (Reimers und Gurevych 2019b). Diese Einbettungen (engl. *Embeddings*) genannten Vektoren ermöglichen die Anwendung von statistischen Methoden auf Textdaten und sind die Grundlage für viele Methoden in der Prozessierung natürlicher Sprache (engl. Natural Language Processing, NLP). Aufgrund des großen Potenzials ist die Erstellung von sinnvollen Embeddings ein bedeutender Forschungsfokus im Bereich der künstlichen Intelligenz.

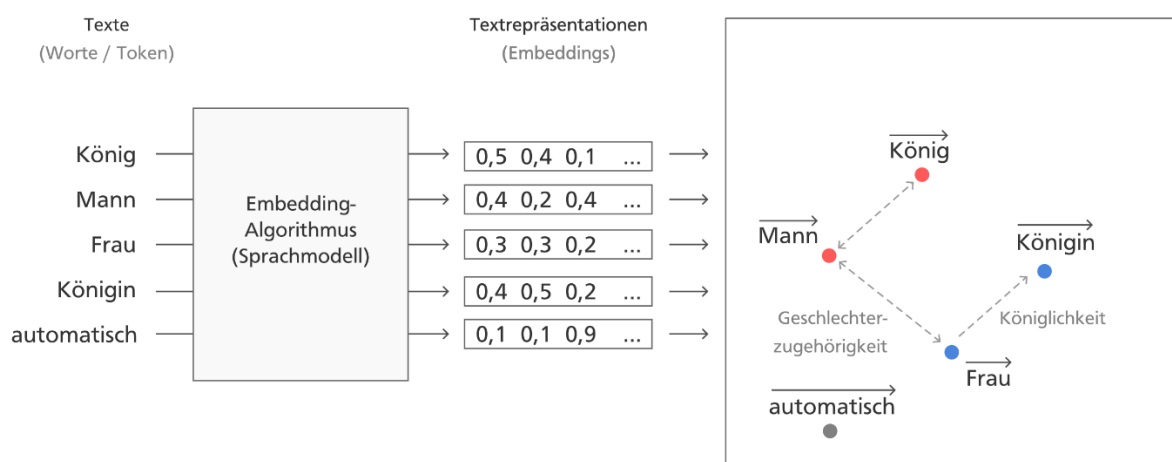


Abbildung 6: Illustration von Embeddings

Ursprünglich wurden Embeddings mittels regelbasierter Methoden erstellt, doch entwickelten sich im Laufe der Jahre eine Vielzahl von Methoden, welche zunehmend in der Lage waren, den Kontext der umgebenden Texte in die Einbettung der Wörter mit einzubeziehen. Worthäufigkeiten oder relative Worthäufigkeiten wie die TF-IDF¹⁷ Metrik wurden dafür genutzt, welche die Häufigkeit eines Begriffes in der Raumeinheit in Bezug zu seiner Häufigkeit im Datensatz allgemein setzt (Patil et al. 2023).

Ein Meilenstein stellte das Training von Wort-Vektoren mittels neuronalen Netzen dar, beispielsweise durch die *Word2Vec* (Mikolov et al. 2013) und *GloVe* (Pennington et al. 2014) Modelle. In den letzten Jahren wurden erhebliche Fortschritte durch die Nutzung von neuronalen Netzen gemacht, welche mittels selbstüberwachtem Lernen trainiert wurden (Patil et al. 2023). Hierdurch kann das Embedding der Worte (oder Token¹⁸) in ihrem Kontext gesehen und angepasst werden. Ein weiterer Meilenstein war die Entwicklung von *Transformer* Architekturen auf Basis des Aufmerksamkeitsmechanismus (engl. *attention mechanism*, Vaswani et al. 2017). *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al. 2018) und darauf aufbauende Architekturen sind in der Lage bidirektional zu arbeiten, also ganze Sätze als Input zu nutzen, und wurden in zahlreichen Anwendungsbereichen zur Nutzung gebracht (Gardazi et al. 2025). Die ebenfalls auf Attention-basierten *Generative Pretrained Transformers* (GPT) sind darauf trainiert, den vorausgegangenen Text zu berücksichtigen, um durch die Prädiktion des jeweils nächsten Tokens einen Text zu generieren. GPTs, und die davon abgeleiteten *ChatGPT* Chatbots (openAI 2022) wurden in der Öffentlichkeit rasch stark rezipiert und sind durch ihre Fähigkeit, menschliche Anweisungen korrekt zu interpretieren, sehr anwenderfreundlich (Zhou et al. 2024).

Weitere entscheidende Entwicklungen beziehen sich auf die Generierung von Embeddings über Sprachen hinweg (Patil et al. 2023) und multimodalen Embeddings, welche beispielsweise Text, Sprache und Bilder in einem gemeinsamen Embedding darstellen können (Patil et al. 2023).

Im Fernerkundungsbereich erschlossen jüngere Entwicklungen die Möglichkeit, die gesamte Erdoberfläche mittels vorberechnete Embeddings von großen Mengen von Satellitendaten zusammenzufassen, welche sowohl auf Pixelbasis (Feng et al. 2025) oder in einem Kontextfenster (Brown et al. 2025) berechnet werden und in einer großen Spanne von Downstream-Anwendungen zum Einsatz kommen können.

Embeddings sind also zunehmend flexibel und mächtig, und stellen die Grundlage für große Fortschritte in zahlreichen Anwendungsbereichen dar, in welchen Bilder und Texte verarbeitet werden (Aggarwal 2018).

Eine ihrer wichtigsten Anwendungsbereiche ist ihre Schlüsselfunktion in modernen Sprachmodellen. Diese Modelle, welche Grundlage vieler NLP-Methoden bilden, können als komplexe Wahrscheinlichkeitsverteilungen gesehen werden, welche die Wahrscheinlichkeit

¹⁷ Der Name ist abgeleitet von der englischen Bezeichnung *term frequency – inverse document frequency* (Sparck Jones 1972).

¹⁸ Worte oder Wortbestandteile werden im Bereich der NLP als Token bezeichnet. Die Aufspaltung eines Textes in Tokens erfolgt algorithmisch nach Regeln, die in einem *Tokenizer* implementiert sind. Die Anwendung dieses Tokenizers ist ein wichtiger Schritt in der Verarbeitung der Texte. Für viele Sprachmodelle wird ein für sie trainierter Tokenizer bereitgestellt.

(engl. *likelihood*) einer Zeichenkette beschreiben können (Russell und Norvig 2023). Diese abstrakte Kernkompetenz ermöglicht die Lösung zahlreicher praktischer Probleme im Bereich der Textverarbeitung, unter anderem die Klassifizierung von Texten oder Textteilen¹⁹, die Eigennamenerkennung²⁰ (engl. *named entity recognition*, NER), die Beantwortung von Fragen oder Übersetzungen. Heutzutage sind Sprachmodelle oft große Basismodelle, welche in der Regel auf der Transformer Architektur basieren. Große Sprachmodelle werden gemeinhin als Large Language Models (LLMs) bezeichnet, wobei keine klare Definition der dafür notwendigen Größe gemeinhin akzeptiert ist. Eine wichtige Funktion von Sprachmodellen ist die Umwandlung von Texten in Embeddings, welche entweder intern vom Sprachmodell genutzt oder für weitere Anwendungen herausgegeben werden. Sprachmodelle können mittels Transfer-learning Ansätzen für bestimmte Anwendungen optimiert werden (Patil et al. 2023).

Innerhalb dieser Dissertation dienen Sprachmodelle primär der Ableitung von inhaltlichen Informationen aus den Textelementen, wobei die Aufgabe als Textklassifizierung konzipiert wird. Die Klassifizierung von Textinhalten erfolgt oft in der Form, dass alle Texte in inhaltlich ähnliche Klassen eingeteilt werden, welche oft als Themenbereiche (engl. *Topics*) bezeichnet werden.

Es existieren zahlreiche Ansätze der Textklassifizierung, von denen die Ansätze des überwachten und unüberwachten **maschinellen Lernens** in dieser Arbeit zur Anwendung kommen. In beiden Fällen spielt die Nutzung von Text-Embeddings eine essenzielle Rolle. Unüberwachtes Maschinelles Lernen zur Textklassifizierung nimmt in NLP oft die Form der Themenmodellierung (engl. Topic Modeling) an. Auch hierfür existieren verschiedene technische Methoden. Die *Latent Dirichlet Allocation* (Blei et al. 2003) basiert auf probabilistischen Textmodellen, welche nicht auf dem semantischen Inhalt einzelner Wörter, sondern auf dem gemeinsame Auftreten von Wörtern, Themen und Text (Dokumenten) beruht. Hierbei werden Textdokumente als Mischung von Themen betrachtet, die wiederum jeweils eine eigene Verteilung von Wörtern haben (Russel & Norveig 2023). Aus der Schätzung dieser Verteilungen ergibt sich eine Zuordnung von Dokumenten und Themen. Alternative Ansätze basieren auf der Nutzung von Text-Embeddings und der Gruppierung von Dokumenten in Themen mittels Clusteringmethoden. Ein Beispiel hierfür ist der BERTopic Ansatz (Grootendorst 2022), welcher Text-Embeddings mittels BERT-basierten Sprachmodellen erzeugt, diese dann mittels des UMAP-Algorithmus dimensional reduziert (McInnes et al. 2018) und schlussendlich mittels des hierarchischen Clustering-Algorithmus HDBSCAN gruppiert (McInnes et al. 2017). Ergebnisse der Topic Models sind immer Gruppen von Dokumenten, welche zwar als Themen (Topics) bezeichnet werden, aber nicht notwendigerweise einem menschlichen Verständnis von Themen entsprechen und in der Regel erst post-hoc benannt und interpretiert werden.

¹⁹ Beispielsweise von Wörtern in lexikalische Kategorien wie Substantiv, Verb, Adjektiv was auch als Part-of-Speech-Tagging bezeichnet wird (Russell & Solveig 2023) oder Stimmungsanalyse (engl. *sentiment analysis*, Pang and Lee 2008), für welche viele Methoden speziell für die Anwendung auf Twitter-Inhalte entwickelt wurden (Giachanou and Crestani 2016).

²⁰ Bei der Eigennamenerkennung (engl. Named Entity Recognition, NER, Esmaail et al. 2024) werden Eigennamen (engl. Entities) identifiziert und bestimmten Typen wie Person, Ort oder Organisation zugeordnet.

Bei der Textklassifizierung mittels überwachtem Lernen werden Beispiele von Texten bereitgestellt, welche mit den Zielklassen annotiert sind. Für den klassifizierenden Text beziehungsweise dessen Elemente werden –beispielsweise durch Embeddings– Merkmale identifiziert, welche dann als Inputs für einen Klassifizierer genutzt werden, der anhand der Beispiele auf die Lösung der Klassifizierungsaufgabe trainiert wird. Dient ein Sprachmodell als Grundlage für die Generierung der Merkmale (Embeddings), kann es in einem Transfer-Learning Ansatz mittrainiert werden. Dies kann auch als Feinabstimmung des Modells betrachtet werden.

2.1.3. Gemeinsame Auswertung von GSM und FED

Die Idee, Fernerkundungsdaten und Textdaten gemeinsam auszuwerten wurde schon vor mehreren Jahrzehnten entwickelt (Jiang 2003). Dennoch befassten sich verhältnismäßig wenige Studien damit, einen Bezug zwischen FED und GSM herzustellen (Zhu et al. 2022).

Die beiden Datenquellen sind also inhaltlich komplementär, unterscheiden sich jedoch stark in ihren Eigenschaften (Hoffmann 2022). FED entspringen einem hochgradig kontrollierten Prozess und liegt typischerweise im Rasterformat vor, bei welchem Datenpunkte regelmäßig im Raum verteilt sind. GSM sind weit weniger systematisch und standardisiert, auch da die Nutzung als Daten kein primäres Ziel der Autoren ist. Im Fall von Tweets bestehen sie aus Texten, welche teilweise Geokoordinaten enthalten, die aber unregelmäßig im Raum verteilt sind (Steiger et al. 2015).

Aufgrund der anwachsenden Datenmengen und verbesserter technischer Methoden wird die Verbindung dieser komplementären Datentypen zunehmend plausibler und attraktiver (Zhu et al. 2022). Die verschiedenen Datentypen können dabei auf zwei verschiedene Arten zusammengeführt werden.

Zum einen können die Daten mit separaten Methoden verarbeitet und gemeinsam interpretiert werden. Dies kam in der Vergangenheit beispielsweise bei Studien von (Kultur-)Landschaften zur Anwendung, welche qualitative und quantitative Ansätze aus Geographie und Literaturwissenschaften kombinierten (Jiang 2003) und dabei sowohl auf FED als auch auf Ethnographien zurückgriffen (Guyer und Lambin 1993). Die Integration gemischter Methoden wird auch mit dem englischen Begriff *Mixed-Methods* bezeichnet²¹ und ist seit dem Ende der 1990er Jahre in zahlreichen Humanwissenschaften eine weithin geübte Praxis (Kelle 2014). Die unterschiedlichen Methoden können sich hierbei gegenseitig validieren, wenn sie auf dieselben Aspekte des Forschungsgegenstandes zielen, oder komplementieren, wenn sie verschiedene Ebenen desselben betrachten (Kelle 2014). Die Integration qualitativer und quantitativer Methoden kann bereits in der Methodenanwendung erfolgen oder erst in der Integration der Ergebnisse. Die Verbindung von GSM und FED mittels *Mixed-Methods* könnte im Migrationsbereich einzigartige Perspektiven bringen, wobei eine Herausforderung darin besteht, Konzepte (wie Migration) zu harmonisieren (Bircan 2023).

²¹ Der nicht klar abzugrenzende Begriff „Multiple Methods“ wird teilweise synonym zu „Mixed Methods“ genutzt, aber auch um alternativ die Nutzung mehrerer qualitativer **oder** quantitativer Methoden zu beschreiben (Anguera et al. 2018). In dieser Studie wird der Begriff „Mixed Methods“ genutzt, um Ansätze zu beschreiben, welche sowohl qualitative als auch quantitative Daten und Methoden integrieren.

Eine technische Alternative zu Mixed-Methods bieten technische Lösungen, bei welchen die verschiedenen Datenquellen angeglichen, gemeinsam verarbeitet oder ausgewertet werden. Dies fällt in den Bereich der multimodalen Datenfusion. Für die Datenfusion von GSM und FED bieten Zhu et al. (2022) eine Übersicht. Auch diese technische Fusion kann auf verschiedenen Ebenen erfolgen: Auf der Ebene der Eingangsdaten (data-level-fusion), der extrahierten Merkmale (feature-level-fusion) innerhalb der Modelle, und im letzten Schritt der Klassifizierung (decision-level-fusion) (Tao et al. 2025). Alle diese Ansätze wurden bereits in Studien im urbanen Raum umgesetzt, wobei die Fusion in den späteren Verarbeitungsschritten häufiger war als die Fusion der Eingangsdaten (Zhu et al. 2022).

Auch hier fokussiert ein Forschungsstrang auf das Training anwendungsagnostischer Basismodelle, welche auf großen Datenmengen vortrainiert werden. Multimodale Basismodelle zeigen Potenzial für eine Vielzahl von Anwendungen (Xing et al. 2024; Kuzu et al. 2025). *Alpha Earth* von Google beispielsweise ist ein auf Geodaten trainiertes Basismodell (Brown et al. 2025), welches Embeddings für ein 10-Meter breites und weltumspannendes Raster generiert, und dafür neben Satellitendaten auch Texte von Wikipedia miteinbezieht (Google, o.D.). Hier werden Informationen aus Fernerkundung und Text also innerhalb des Embedding-generierenden Modells zusammengeführt.

Neben diesem Ansatz, welcher als eine generische und anwendungsagnostische Form der Datenfusion betrachtet werden kann, zielt ein weiterer Forschungsstrang darauf ab, die Kombination unterschiedlicher Datensätze mittels überwachten Lernens für spezielle Anwendungen zu optimieren. Bei dieser Richtung der Datenfusion, werden die Modelle, die die Fusion durchführen, typischerweise spezifisch auf eine Aufgabe trainiert. Angewendet wurde dies beispielsweise von Häberle et al. (2022) für die Aufgabe der Gebäudeklassifizierung. Hierbei wurde ein Decision-Level-Fusionsmodell genutzt, welches die Fernerkundungsdaten mittels eines CNN verarbeitet, Geotweet-Texte separat in Text-Embeddings konvertiert, unabhängig Klassifizierungen durchführt und deren Wahrscheinlichkeitsvektoren dann auf Gebäudelevel fusioniert. Neben Texten kamen auch andere Metadaten zu Anwendung. Beispielsweise nutzten Leichter et al. (2018) Tweet-Metadaten²² auf Pixellevel, aggregieren diese und extrapolieren räumlich zur Klassifizierung von Landnutzung. Für einen umfassenden Überblick über die Fusion von Social-Media und FED wird auf Zhu et al. (2022) verwiesen.

Insgesamt lässt sich sagen, dass die Fusion mittels tiefer neuronaler Netze gleichzeitig vielversprechend und nicht trivial ist. Zahlreiche Studien haben Methoden entwickelt, die Vorzüge und Beiträge beider Datensätze zu kombinieren (Leichter et al. 2018; Häberle et al. 2022; Hoffmann 2022; Li et al. 2024; Xing et al. 2024). Die Entwicklungen der letzten Jahre haben deutliche Verbesserungen in Deep-Learning Ansätzen zu multimodalen Datenfusionstechniken geführt (Zhao et al. 2024). Diese können zusätzlich zu anwendungsbezogenen Kombination von Fernerkundungsdaten, Textdaten, und potenziell weiteren Datentypen beitragen (Xing et al. 2024). Die multimodale Fusion von GSM und Fernerkundungsdaten bleibt ein aktives Forschungsfeld (Zhao et al. 2024; Wieland et al. 2025).

2.1.4. Geoparsing

Selbst auf Plattformen mit Geotag-Funktion machen Inhalte mit Geotags nur einen kleinen Teil der Daten aus. Kruspe et al. (2021) fanden in einer Stichprobe von Tweets, dass

²² beispielsweise der durchschnittlichen Textlänge oder dem Zeitpunkt des Postens.

geographische Eigennamen in 6% der Tweets enthalten sind, von denen nur 3.5% mit einem Geotag annotiert waren. Darüber hinaus machen auch Plattformen mit Geotag-Funktion nur einen Teil aller Webtextdaten aus. Es liegt also vermutlich ein großes Potenzial in der Auswertung von allen textbasierten Geoinformationen (Kruspe et al. 2021).

Die Ableitung von Geokoordinaten aus im Text genannten Toponymen wird als Geoparsing bezeichnet (Hu et al. 2023b). Abbildung 7 illustriert diesen Ansatz, bei welchem zunächst geographische Eigennamen (in der Regel Toponyme) identifiziert werden, was als Toponymerkennung (engl. *toponym recognition*) bezeichnet wird und als ein Teilbereich der NER (Esmaail et al. 2024) darstellt.

In einem zweiten Schritt, der Geocodierung (engl. *geocoding*)²³, werden die Toponyme disambiguiert und mittels eines Ortsverzeichnisses verortet (Gritta et al. 2018). Ein weithin genutztes Ortsverzeichnis ist GeoNames ('GeoNames', o.D.), welches auch eine öffentlich zugängliche API bereitstellt. Auch wenn das Durchsuchen von Ortsverzeichnissen sehr rechenaufwendig ist, ermöglichen technische Fortschritte das Geoparsing zunehmend auch für große Korpora von Webtexten (Farzana und Hecking 2024).

Die Einbindung von generativen²⁴ LLMs in Geoparsing-Prozesse kann dabei zu Verbesserungen in der Performance führen. Generative LLMs sind zwar in der Anwendung verhältnismäßig langsam und ressourcenintensiv, was die großflächige Anwendung einschränkt. GPT-basierte LLMs zeigten aber in Tests eine bessere Genauigkeit als traditionelle Ansätze oder BERT-basierte LLMs und können insbesondere mit informellen Texten gut umgehen (Hu et al. 2024b). Hierdurch steht zu erwarten, dass in Zukunft ein größerer Teil der in informeller und unstrukturierter Sprache verfassten (Hua et al. 2012) Social-Media-Inhalte für geographische Anwendungen erschlossen werden könnte.

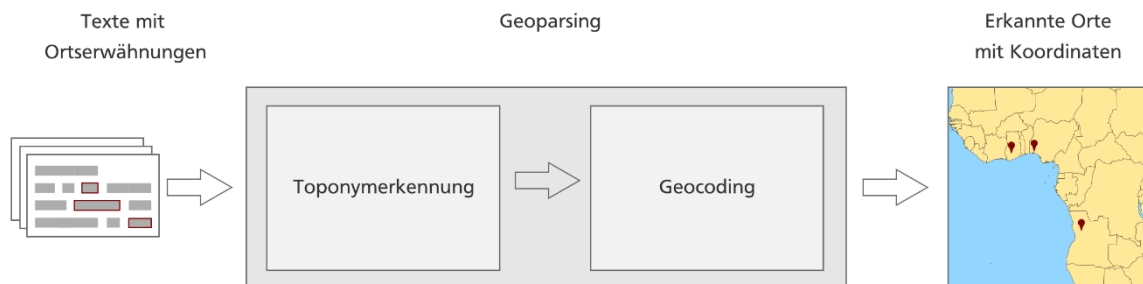


Abbildung 7: Illustration des Geoparsings, nach Wang und Hu (2020)

Bei der Analyse von mit Geoparsing verorteten Texten ist es wichtig, zu beachten, dass eine Ortserwähnung verschiedene Funktionen erfüllen kann. Sie kann sich beispielsweise auf den Standort der Autoren beziehen, oder aber auf den Ort der im Text erwähnten Objekte, Ereignisse oder Personen. Bei Geotags auf Twitter besteht aufgrund der technischen Implementierung der Software eine hohe Wahrscheinlichkeit, dass sie sich auf den Standort des Autors beziehen. Hingegen besteht bei der Auswertung von Texten mittels Geoparsings eine weitaus größere Bandbreite. Ihre Auswertung ist daher gleichzeitig herausfordernder und potenziell informationsreicher.

²³ Alternativ auch als *toponym resolution* bezeichnet

²⁴ Sprachmodelle wie GPT, welche auf die Generierung/Produktion von Texten ausgelegt sind.

2.1.5. Domänenadaption

Im Rahmen der Dissertation sollen überwachte maschinelle Lernverfahren zum Einsatz kommen, um Modelle für die Klassifizierung von GSM-Texten zu trainieren.

Die für das Training von Modellen genutzten annotierten Daten sollten aus einem ähnlichen Hintergrund stammen, wie der, in welchem das trainierte Modell angewendet wird. Dieser Hintergrund des Textes wird im NLP-Bereich als Domäne (engl. Domain) bezeichnet und umfasst eine nicht klar abgegrenzte Menge an Eigenschaften, zu welchen beispielsweise Sprache, Typ und Erstellungsdatum der Inhalte sowie deren Quellen gehören (Farahani et al. 2021)²⁵. Wünschenswert ist, dass Trainingsdaten für ein Modell aus derselben Domäne stammen, in welcher das Modell angewendet wird.

Alternativ können in Transfer-Learning-Ansätzen auch Daten aus anderen (Quell-)Domänen genutzt werden, um Modelle für eine Aufgabe in der anvisierten (Ziel-)Domäne zu trainieren oder zumindest vorzubereiten. Dieser Schritt, sowie Methoden, um ihn umzusetzen werden als Domänenadaption (engl. *domain adaptation*) bezeichnet. Eine Vielzahl von Methoden existiert, von welchen manche unüberwacht funktionieren, also lediglich ungelabelte Beispiele aus der Zieldomäne erfordern (Farahani et al. 2021).

Zu den unüberwachten Methoden gehört beispielsweise der von Ganin et al. (2016) entwickelte domänen-adversariale Ansatz, bei welchem die Identifikation der Domäne als zusätzliche Aufgabe im Training genutzt, der zugehörige Fehlergradient jedoch invertiert wird (Abbildung 8). Hierdurch wird das Lernen von Merkmalen begünstigt, welche zwar hilfreich für die Aufgabe der Textklassifizierung, aber nicht hilfreich für die Unterscheidung von Domänen sind. Dies fördert das Lernen von Features, welche in beiden Domänen auftauchen. Diese domänen-adversariale Trainingsmethode lässt sich mit wenig Aufwand in jede existierende Modellarchitektur implementieren und kann auch die Generalisierungsfähigkeit des Modells verbessern.

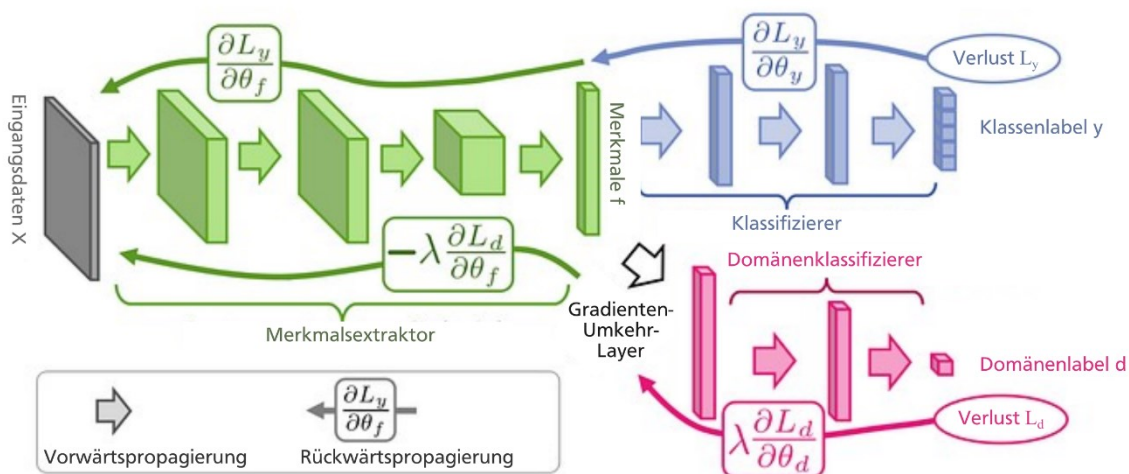


Abbildung 8: Illustration des domänen-adversarialen Trainings nach Ganin et al. (2016)²⁶

²⁵ Dies unterscheidet sich von dem sprachwissenschaftlichen Konzept der Domäne, welche den sozialen Kontext einer Aussage oder Unterhaltung meint (Holmes & Wilson, 2022).

²⁶ Quelle: <https://medium.com/one-minute-machine-learning/domain-adversarial-training-of-neural-networks-2016-one-minute-summary-15bfe81af128>

2.1.6. Trajektorienanalysen mittels GSM

Bei der zunehmend verbreiteten Nutzung von Telekommunikationstechnologien hinterlassen Menschen digitale Spuren (engl. *digital traces*). Wenn diese Aufzeichnungen Georeferenzen und Zeitstempel beinhalten, können sich aus ihnen raumzeitliche Pfade ableiten lassen, welche als Trajektorien (engl. *trajectories*) bezeichnet werden (Abbildung 9). Sie geben Aufschluss über die Bewegungen der Nutzenden. Ihre Analyse erlaubt die Identifizierung der Migrierten als solche im engeren Sinn oder allgemeiner die Erfassung von Mobilitätscharakteristika von Accounts auf Social Media Plattformen.

Diese Ableitung von Mobilitätsinformationen aus den Trajektorien von Internetnutzenden ist ein aktives Forschungsfeld (Becker 2018; Lemoine-Rodríguez et al. 2024; Chi et al. 2025), was einerseits durch das Potenzial motiviert ist, traditionelle Datenquellen zu ergänzen, andererseits durch Fortschritte in verteilter und Cloud-basierter Datenverarbeitung, welche die Anwendung statistischer Methoden auf sehr große Datensätze erlauben (Chi et al. 2025; Pappalardo et al. 2023).

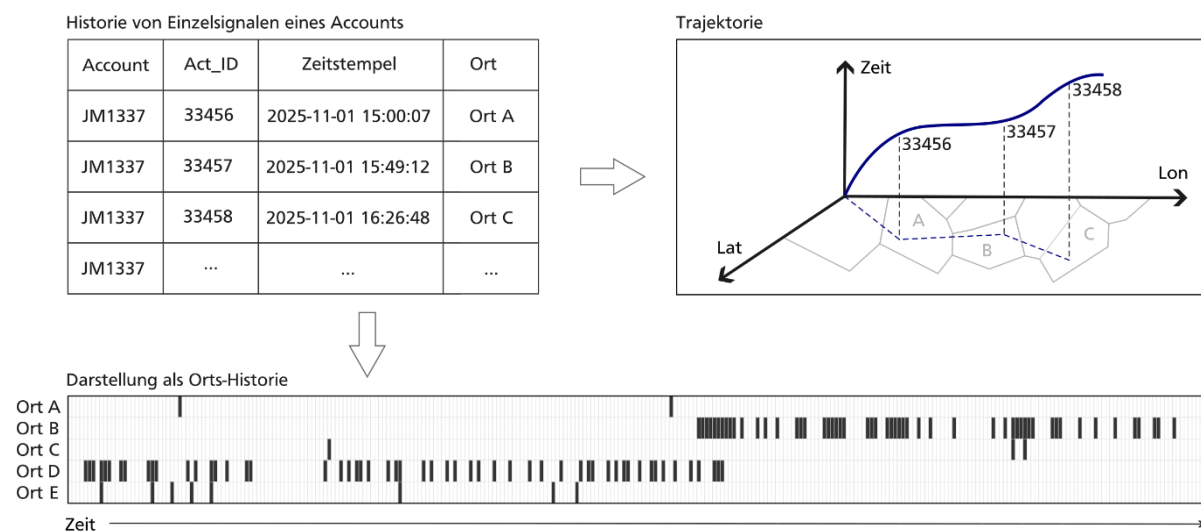


Abbildung 9: Extraktion von Trajektorien aus Digitalen Spuren, reproduziert nach Chi et al. (2020)

Innerhalb dieses weiten Feldes fokussiert sich diese Dissertation auf das Themenfeld der Migration und speziell auf langfristige Migration, weswegen deren Identifizierung im Folgenden detailliert beschrieben wird. Den verschiedenen Definitionen von Migration ist gemein, dass sie Migration als einen Wechsel von Aufenthaltsorten betrachten, bei welchen nationalstaatliche oder administrative Grenzen überschritten werden (Chi et al. 2020). Ein Kernelement der Detektierung von langfristiger Migration ist daher die Identifizierung von Aufenthaltsorten²⁷ der Nutzenden. Wie Chi et al. (2020) in seinem Überblick zeigt, gibt es in der Forschung viele verschiedene Methoden, den Aufenthaltsort zu definieren, was zum Teil daran liegt, dass die Validierung der Methoden herausfordernd ist (Pappalardo et al. 2020). Einfache Ansätze folgen dem Prinzip des modalen Ortes (engl. *modal location*), welches als Aufenthaltsort die Raumeinheit (oft ein Land, Stadt, oder eine Provinz) betrachtet, in welcher die Nutzenden während eines bestimmten Zeitraumes die meisten Spuren generieren. In der Praxis muss dafür also sowohl eine zeitliche als auch eine räumliche Dimension festgelegt

²⁷ Alternativ als *Wohnort*, *Heimat* oder mit den englischen Begriffen *Home* und *Residence* bezeichnet

werden, welche sich an der Art der Migration reflektiert, die betrachtet werden soll (Chi et al. 2020).

So aggregieren Kim et al. (2020) bei der Betrachtung langfristiger Migration auf Kalenderjahren, Zagheni et al. (2014) auf der Ebene von Tertialen (vier Kalendermonate) und Mazzoli et al. (2020) bei der Betrachtung einer humanitären Krise auf Kalendermonate. Diese Aggregation auf feste Zeitintervalle ist einfach umzusetzen und bietet eine heuristische Lösung der Identifizierung des Aufenthaltsortes. Ruths und Armstrong (2020) demonstrieren mit dem auf dynamischer Programmierung basierenden *Residence Inference Algorithm* (ResIn), dass neben heuristischen auch optimale Lösungen dieser Aufgabe möglich sind.

Als weitere Alternative können Zeiträume von Aufenthalten auch dynamisch in den Trajektorien der Nutzenden erkannt werden. Chi et al. (2020) entwickelten einen solchen generischen Algorithmus für welchen flexible Zeitspannen genutzt werden können. Dieser Algorithmus wurde unter anderem für Tweets getestet, um optimierte Parameter für diese Datenquellen zu schätzen. Er besteht aus drei Schritten (Abbildung 10):

(1) Zunächst werden kontinuierliche Segmente identifiziert, bei welchen Einzelsignale über einen längeren Zeitraum an einem Ort registriert werden. Diese Segmente werden als Aufenthalt der Nutzenden an dem Ort betrachtet. Zweitens (2) werden konsekutive Segmente am selben Ort verschmolzen, in der Annahme, dass Lücken zwischen Segmenten auf Datenlücken beruhen und keine tatsächliche Abwesenheit darstellen. Zuletzt (3) werden Zeiträume, in welchen Überschneidungen zwischen Segmenten bestehen, aus den Segmenten entfernt. Der Algorithmus von Chi ist als Software in der Programmiersprache Python öffentlich implementiert²⁸.

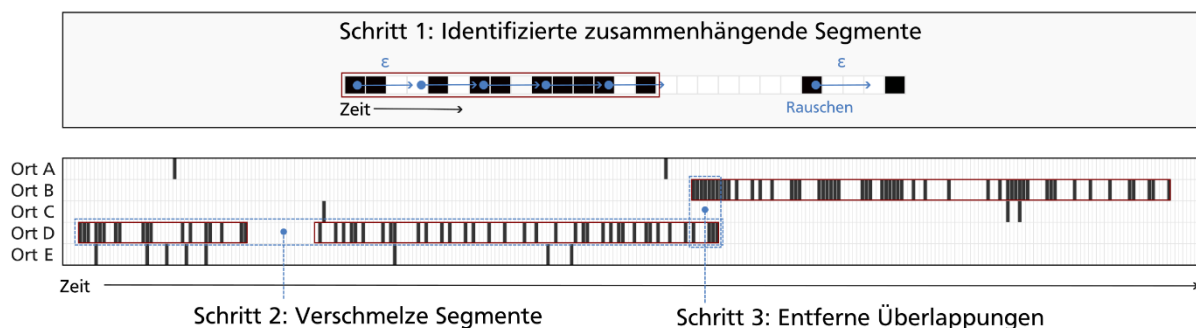


Abbildung 10: Segmentbasierter Ansatz zur Identifikation von Migrationsereignissen in Trajektorien, reproduziert nach Chi et al. (2020)

Bei sozialen Medien mit Netzwerk-Funktionen wie Freundschafts- oder Follow-Funktionen können die Beziehungs-Netzwerke der Nutzenden in die Klassifizierung miteinbezogen werden (Xu et al. 2014; Kim et al. 2020).

Neben kategorischen Einteilungen in Mobilitätstypen können auch kontinuierliche Maße für Mobilität berechnet werden. Der Gyrationradius (R_{GYR} , siehe beispielsweise Zagheni et al. 2014) ist die mittlere Distanz der Spuren eines Accounts zu ihrem gemeinsamen Schwerpunkt (Zagheni et al. 2014). Er kann bei stationären Nutzenden gegen 0, für häufig reisende internationale Nutzende mehrere tausend Kilometer betragen.

Es existieren also zahlreiche Methoden zur Trajektorienanalyse, von denen sich zum aktuellen

²⁸ https://github.com/g-chi/migration_detector

Zeitpunkt kein Ansatz klar durchgesetzt hat. Auch heuristische Methoden liefern eine gute Approximation der tatsächlichen Aufenthaltshistorie (Armstrong et al. 2021), weswegen die Wahl des Algorithmus vermutlich weniger wichtig ist als die Eignung der Methoden und Daten für die Forschungsfrage.

Als Datengrundlage für die Extraktion von Trajektorien kommen verschiedenste Quellen infrage. Eine breite Abdeckung erreichen Telekommunikationsdaten in Form von Call Detail Records (CDR), extended Detail Records (XDRs) oder Control Plane Records (CPR). Diese werden bei der Kommunikation von Mobiltelefonen und dem Netzwerk registriert (Pappalardo et al. 2020) und enthalten Information über die Antenne, welche für die Übertragung genutzt wurde, woraus der Ort der Nutzenden auf wenige Kilometer genau bestimmt werden kann (Haraguchi et al. 2022). Diese von den Netzwerkanbietern gesammelten Informationen decken weite Geographien und Demographien ab und wurden unter anderem für Studien von urbaner Mobilität (Mamei et al. 2019), Schätzungen von durch Naturkatastrophen ausgelöste Bevölkerungsbewegungen (Wilson et al. 2016) oder für die Unterscheidung von grundlegenden Mobilitätstypen (Pappalardo et al. 2015) genutzt. Sie enthalten aber neben den raumzeitlichen Einzelsignalen keine Informationen wie Textinformationen, solange diese nicht zusätzlich über die Analyse von Nutzenden erfasst und bereitgestellt werden. Zudem sind die Daten zum Schutz der Privatsphäre der Nutzenden in der Regel nicht öffentlich (Cecaj et al. 2014).

Demgegenüber enthalten Inhalte von GSM-Plattformen, welche Geotag-Funktionen bereitstellen, diese und weitere Informationen in öffentlicher Form²⁹. Ein Geotweet mit einem Zeitstempel kann als raumzeitliches Einzelsignal interpretiert werden, womit längere Aufzeichnungen (Historien) von Geotweets eines Accounts in Aggregation als Trajektorien interpretiert und analysiert werden können (Armstrong et al. 2021).

Zahlreiche Studien belegen die Plausibilität der Ableitung von Mobilitätscharakteristika aus GSM. Hawelka et al. (2014) analysierten über das Jahr 2012 hinweg Reisebewegungen auf internationaler Ebene, bei welchem Accounts von anderen Ländern als dem modalen Land dieses Jahres posteten. Kim, Sîrbu, Rossetti, et al. (2022) identifizierten Migrierte als Accounts, deren Hauptaufenthaltort nicht der ihres Netzwerkes ist, und kontrastierten ihre Netzwerke und die von ihnen genutzten Hashtags mit denen stationärer Accounts. Hübl et al. (2017) und Petutschnig et al. (2020) nutzten Twitter-Daten, um Flüchtlingsströme auf der Balkanroute zu messen, fanden aber, dass relativ wenige Trajektorien plausiblen Mustern von Flucht entsprachen, und mutmaßten, dass stattdessen Geschäftsreisende oder Touristen erfasst werden konnten. Demgegenüber fanden Lemoine-Rodríguez et al. (2024) eine gute Übereinstimmung zwischen den Flüchtlingsmustern der durch den Russischen Angriffskrieg gegen die Ukraine geflohenen (und zurückgekehrten) Ukrainern und den offiziellen Statistiken des UNHCR. Havas et al. (2021) nutzten Twitter-Daten zusammen mit offiziellen Statistiken zur Prognose von Migrationsströmen zwischen Türkiye und Zentraleuropa.

Auch Daten anderer Plattformen können zur Anwendung kommen. Der auf Facebook-Daten berechnete und öffentlich zugänglich gemachte *Facebook Social Connectedness Index* erwies sich in einer Studie als guter Prädiktor für aus dem Russischen Angriffskrieg gegen die Ukraine resultierende ukrainische Diaspora und damit die konfliktbedingten Migrationsbewegungen (Minora et al. 2022). Auch auf Basis von Facebook-Daten erreichten Chi et al. (2025) auf

²⁹ Auch wenn dies Bedenken der Privatsphäre reduziert (Hawelka et al. 2014), sollte trotzdem berücksichtigt werden. Für eine Diskussion siehe Sektion 7.4 dieser Thesis.

annähernd globaler Ebene³⁰ eine Schätzung internationaler Migrationsflüsse auf monatlicher Basis, welche eine gute Übereinstimmung mit administrativen Statistiken erzielten³¹. Diese Berechnungen basieren auf nicht-öffentlichen Daten und stehen nur den Plattformbetreibern³² zur Verfügung. Jedoch konnten auf interne Herkunftsklassifikationen von Facebook mittels der Advertising API zugegriffen werden. Spyratos et al. (2019) fanden, dass hierauf basierte Schätzungen von Migrationsbeständen stark ($R^2 = 0.735$, $p < 0.001$) mit offiziellen Statistiken korreliert waren.

Auch weitere Plattformen und Quellen konnten und können für die Mobilitätsanalyse genutzt werden (Spyratos et al. 2019), wobei die Plattform entsprechend der Forschungsfrage gewählt werden kann. So wurden *LinkedIn*-Daten genutzt, um Trends in der Migration von Fachkräften zu erfassen (State et al. 2014) und bibliometrische Daten aus wissenschaftlichen Veröffentlichungen ausgewertet, um Migration von Forschenden zu beobachten (Moed und Halevi 2014; Czaika und Orazbayev 2018).

Dieser Überblick soll für die Anwendung von GSM im Rahmen dieser Dissertation ausreichen, jedoch keine ausführliche Zusammenfassung darstellen. Über Quellen, Algorithmen, Anwendungen und Herausforderungen von digitalen Spuren in Mobilitätsstudien im weiteren Sinn bieten die Arbeiten von Chi (2020) und Pappalardo et al. (2023) einen Überblick.

2.2. GSM-Spezifische Bias-Faktoren

Die Analyse von GSM muss unter Berücksichtigung ihrer Repräsentativität erfolgen. Obwohl die Nutzung von Internettechnologien inzwischen unter weiten Teilen der Weltbevölkerung verbreitet ist, bestehen immer noch Ungleichheiten (ITU 2021), welche die Repräsentativität von Webdaten beeinträchtigen können. Derartige Ungleichheiten in den Daten, welche die Repräsentativität von Daten beeinflussen, werden allgemein als Verzerrungen (engl. *Biases*³³) bezeichnet (Olteanu et al. 2019).

Als demographische Dimensionen, in welchen Verzerrungen auftreten können, wurden beispielsweise bereits Geschlecht, Ethnizität, Bildung, Einkommen und Alter identifiziert, und dies sowohl bei der Nutzung von Social Media allgemein als auch für einzelne Social-Media-Plattformen wie Facebook, Reddit, oder Twitter (Hargittai 2020). Allgemein geht die Internetnutzung mit Wohlstand einher, wie eine Studie von Wentrup et al. (2016) auf Länderebene zeigt. Zudem ist die Internetnutzung in städtischen Räumen deutlich höher als in ländlichen Räumen, wobei diese Disparität unterschiedlich stark ist und beispielsweise in weniger wohlhabenderen Ländern besonders stark ist (ITU 2024).

Für einen Überblick über Verzerrungen welche Social-Media allgemein beeinflussen, wird auf Olteanu et al. (2019) verwiesen. Hier soll der Fokus auf Faktoren liegen, welche die Genese speziell von GSM-Daten beeinflussen.

GSM-Daten unterliegen zusätzlichen Verzerrungen, welche bei ihrer Analyse berücksichtigt werden müssen. So zeigten Studien, dass die Nutzung der Geotagging-Funktion auf Twitter zwischen Ländern, Bezirken und Bevölkerungsgruppen nicht konstant ist (Malik et al. 2015; Jiang et al. 2019; Huang und Carley 2019), weswegen nicht davon ausgegangen werden kann,

³⁰ 181 Länder

³¹ Die Korrelation mit Daten aus Neuseeland betrug 0,98, der eingeschränkt mögliche Vergleich mit Statistiken der Europäischen Union (Eurostat) betrug 0,94

³² In diesem Fall die Meta Platforms, Inc.

³³ Singular: *Bias*, Plural: *Biases*

dass diese ein repräsentatives Abbild der Plattformnutzenden darstellen.

Genaue Informationen über die Repräsentation verschiedener Demographien in GSM wären wünschenswert, sind jedoch nur sehr eingeschränkt ableitbar. Oben genannte Studien berufen sich auf administrative Daten oder Zensus-Daten, welche über den Ort des Geotags eingebunden wurden. Eine Alternative kann beispielsweise der Abgleich von Accountnamen mit offiziellen Namensregistern sein (Longley et al. 2015).

Im Migrationskontext sind die demographischen Informationen oft nicht ausreichend, um eine Korrektur der Messungen vorzunehmen. Bei der Interpretation muss daher berücksichtigt werden, dass bestimmte Bevölkerungsgruppen in geo-getaggtten Daten über- und unterrepräsentiert sind. Einige der im Bereich der erweiterten Migrationsforschung mit Trajektorienanalysen von GSM durchgeführte Studien führten zu der Schlussfolgerung, dass sich die Ansätze weniger für die Analyse von Flüchtlingen eignen als für die Mobilität von Fachkräften, Touristen und Geschäftsreisenden (Hübl et al. 2017; Petutschnig et al. 2020; Armstrong et al. 2021). Hecht und Stephens (2014) fanden, dass in Städten deutlich mehr Informationen pro-Kopf vorlagen als in ländlichen Gebieten. Doch auch innerhalb urbaner Räume sind GSM keinesfalls gleichmäßig dicht verteilt. Dies soll speziell im nächsten Kapitel ausgeführt werden.

2.2.1. Digitale Disparitäten in urbanen Räumen

Als Teil des von Rabari und Storper (2015) als „Digitale Haut“ (engl.: „*Digital Skin*“) bezeichneten Systems aus sozialen Medien, Telekommunikationsgeräten und Sensoren stellen GSM eine wesentliche Komponente der virtuellen Repräsentation von Stadt im Internet dar. Mehrere Studien belegen, dass diese digitale Haut nicht überall in Städten gleich stark ist (Robertson und Feick 2016).

Robertson und Feick (2016) untersuchten die Verbreitung von Fotos mit Geotags auf der Plattform Flickr in US-amerikanischen Städten und modellierten deren Zusammenhang mit ausgewählten sozio-ökonomischen Indikatoren aus offiziellen Erhebungen auf der Ebene von Zensusseinheiten (engl. *census tracts*) sowie der Fläche der Blocks. Hiermit identifizierten sie Städte, welche im Vergleich zu ihrer sozioökonomischen Größe über- und unterrepräsentiert waren, sowie erhebliche Ungleichheiten innerhalb von Städten und führten dies auf sozioökonomische Variablen und kulturelle Faktoren wie Tourismus zurück.

Jiang et al. (2016) identifizierten einen Gradienten in der Geotweet-Verteilung von Stadtkern hin zur Peripherie. Cai et al. (2017) und Lv et al. (2021) nutzten Daten von Weibo³⁴ bzw. Baidu Maps³⁵, um urbane Zentren und Subzentren zu identifizieren. Um die digitalen Disparitäten im urbanen Raum zu erkennen, wurden auch andere Geodatentypen mit einbezogen. Lang et al. (2022) nutzten in einer Studie in Oslo neben OpenStreetMap auch Facebook Daten, um Lebendigkeit (engl. *urban vibrancy*) verschiedener Nachbarschaften zu bewerten. Taubenböck et al. (2018) verbanden Geotweets mit Fernerkundung, um die Unterschiede in GSM-Aktivität zwischen formellen und informellen Siedlungen zu kartieren und statistisch zu analysieren. In dieser Anwendung wurde somit auch demonstriert, dass die Verbindung von FED und GSM im urbanen Kontext plausibel ist. Speziell ist bedeutsam, dass FED als Proxy für sozioökonomischen Status dienen kann.

³⁴ Eine auf den chinesischen Markt fokussierte Microblog-Plattform analog zu Twitter

³⁵ Eine auf den chinesischen Markt fokussierte Kartensoftware analog zu Google Maps

Zusammenfassend belegt der Stand der Forschung zum aktuellen Zeitpunkt, dass sozioökonomische Faktoren die GSM-Nutzung bedingen. Die daraus resultierenden räumlichen Disparitäten wurden teilweise als Ausgangspunkt für Studien genutzt, dienten aber auch selbst als Forschungsgegenstand der unter anderem als digitale Spaltung (engl. *Digital Divide*) interpretiert wurde. Eine Dimension dieser Disparitäten, die noch nicht untersucht wurde, obwohl sie angesichts des rapiden Wachstums vieler Städte zunehmend an Bedeutung gewinnt, ist die Disparität zwischen alten und neuen Siedlungsbereichen.

2.2.2. Räumlichkeitseffekte von Textinhalten

Wie in (Sektion 2.2) dargelegt, wurden demographische Verzerrungen mit räumlicher Ausprägung bereits mehrfach untersucht, wobei unter anderem auch Unterschiede in der Häufigkeit der Nutzung von Geotags identifiziert wurden (Malik et al. 2015; Huang und Carley 2019). Diese demographischen Studien wurden auf der Ebene von geographischen Raumeinheiten, Bevölkerungsgruppen oder Accounts durchgeführt, und lassen darauf schließen, dass bestimmte Bevölkerungsgruppen oder Typen überrepräsentiert beziehungsweise unterrepräsentiert sind.

Demgegenüber ist nur selten untersucht worden, inwiefern bestimmte **Inhalte** in Texten mit geographischem Bezug überrepräsentiert sind, auch wenn die Vorstellung, dass der Inhalt einen Einfluss darauf hat, wie raum-bezogen ein Text ist, vielen Menschen intuitiv erscheint³⁶. Studien im Bereich der Suchmaschinenoptimierung und geographischem Informationsabruf (engl. *geographic information retrieval*) zeigten, dass Suchbegriffe unterschiedlicher Wissensbereiche auch unterschiedlich häufig in Verbindung mit Ortsnamen gesucht wurden (Gan et al. 2008; Jiang et al. 2019). Neben diesen technisch-fokussierten Studien existieren nur Wenige, die einen Bezug zwischen Inhalt von Texten und ihrem geographischen Fokus als Verzerrung betrachten.

Jiang et al. (2023) verglichen Inhalte von Geotweets und Tweets ohne Geotags und fanden, dass Geotweets unter anderem eine positivere Stimmung aufweisen und mehr plural-Pronomen nutzen, was als ein Bias zu positiven und gemeinschaftlichen Inhalten gesehen werden kann. Die Autoren der Studie spekulieren ferner, dass dies an den Themen der Inhalte liegen könnte. Über dies hinaus wurde jedoch, trotz der weiten Verbreitung von Themen (Topics) als Kategorisierungsebene in der Textanalyse, der Effekt von Themen als möglicher Bias-Faktor für geographischen Daten noch nicht erforscht. Stock (2018) fand in einer Übersichtsarbeit, dass der Fokus auf Geotweets die Vielfalt an Textinhalten und räumlicher Information nur unzureichend abbildet. Aus dieser Hinsicht ist eine Erweiterung der Studie von Jiang et al. (2023), welche sich auf Twitter-Daten und Geotags fokussiert, auf weitere Datenquellen und Georeferenzarten wünschenswert.

Bezüglich der Georeferenzarten gibt es Gründe, Unterschiede zwischen Geotag und Geoparsing anzunehmen. Tear (2020) machte die vielleicht überraschende Beobachtung, dass Twitter-Nutzende, welche die Geotagging-Funktion verwendeten, tendenziell weniger Orte in ihren Texten erwähnen. Auch dies ist wieder eine Beobachtung aus demographischer Sicht, welche eine Analyse auf Ebene von Texten, und die Interaktion mit thematischen Inhalten wünschen lässt. Ein erster Ansatz könnte hier dadurch gewonnen werden, die Räumlichkeit von Themen zu vergleichen.

³⁶ Im Rahmen der Dissertation wurde mir von zahlreichen Menschen in informellen Gesprächen bestätigt, dass sie von einer solchen Beziehung intuitiv ausgehen würden.

Auch die räumliche Analyse von Posts mittels ihrer Geotags, insbesondere auf Twitter und Flickr, war in den letzten Jahren weit verbreitet. Dies ist zum Teil auf die Zugänglichkeit und Verfügbarkeit für die Forschung zurückzuführen. Tatsächlich existiert eine Vielzahl an möglichen Alternativen (Stock 2018), welche mittels Geoparsing verortet werden könnten. Inwiefern die Nutzung von Geoparsing zu ähnlichen Informationen führt wie die Nutzung von Geotags wurde bislang nur unzureichend adressiert. Stock (2018) identifiziert in ihrer Übersichtsarbeit, dass die Vergleichbarkeit von Studien über Plattformen und Geoinformationsansätze sehr herausfordernd ist.

Durch alternative Georeferenzierungsansätze ließe sich eine potenziell sehr große Datenmenge erschließen. Gleichzeitig wandelt sich der Raum der technischen Möglichkeiten durch technische (Hu et al. 2023b) und wirtschaftliche (Mimizuka et al. 2025) Entwicklungen stetig, was den Wechsel von Ansätzen zukünftig zu einer attraktiven Option machen könnte. Aus diesen Gründen soll herausgefunden werden, inwiefern die beiden unterschiedlichen Ansätze zur geographischen Verortung zu ähnlichen Ergebnissen führen.

2.3. Anwendungen an der räumlichen Intersektion von GSM & FED

2.3.1. Nutzung von GSM und FED zur Analyse und Strukturierung urbaner Räume

Wie in der Einleitung dargestellt, ist es für die geographische Erforschung von Migration wünschenswert, physische und soziale Charakteristika von durch Migration geprägten Orten und insbesondere städtischen Räumen auszumachen.

Die Verbindung von GSM und FED im urbanen Bereich wurde bereits erprobt, wenn auch nicht im Kontext von menschlicher Migration (Zhu et al. 2022). Speziell für die Fusion von FED und GSM identifizieren Zhu et al. (2022) als Anwendungsgebiete die Bestimmung von Gebäudenutzungen, Kartierung von Bevölkerungsdichten, sozioökonomischen und gesundheitlichen Variablen und gesundheitlicher Auswirkung von Grünfläche, kulturellen Ökosystemdienstleistungen und Immobilienpreisen. Die Anwendung mit Bezug auf menschliche Migration ist also vielversprechend, doch noch zu ergründen.

Die starke interdisziplinäre Natur der Migrationsforschung lässt erwarten, dass die beteiligten Fachbereiche eine Vielzahl von sowohl quantitativen als auch qualitativen Methoden in die Analyse einbringen. Zusammen mit dem unausgereiften Stadium der Technik lässt dies erwarten, dass anwendungsbezogene Studien zumindest teilweise einen explorativen Fokus haben werden, bei welchem im Rahmen der Studie erst anhand der Dateninhalte deren Auswertungsmöglichkeiten ausgelotet und, zusammen mit Informationen aus anderen Quellen, zur Theoretisierung und Bildung von Frameworks genutzt werden.

In diesem explorativen Status besteht ein vielversprechender und relativ leicht durchzuführender Schritt in der Strukturierung des Untersuchungsgebietes in interpretierbare räumliche Einheiten wie Bezirke, Blöcke oder Nachbarschaften. Für jeden dieser Bezirke können sowohl die FED als auch die GSM dann sowohl mittels deskriptiver Statistiken als auch mit qualitativen Methoden³⁷ untersucht werden. Die Ergebnisse dieser Analysen können dann gemeinsam visualisiert werden. Diese Art der Analyse, bei welcher die quantitative Analyse von Fernerkundungsdaten mit qualitativen Methoden verbunden wird, ist nicht neu (Jiang 2003) aber flexibel und mächtig, und ermöglicht die Einbindung von Experten, Stakeholdern und Menschen mit persönlichen Erfahrungen, was für das

³⁷ beispielsweise der qualitativen Diskursanalyse, der Ethnographie oder der Psychologie

Verständnis von Migration bedeutsam ist (McAuliffe et al. 2018). Die Datentypen werden bei diesem Ansatz also erst in bei der Interpretation durch Experten zusammengebracht, welchen eine hohe Verantwortung zukommt.

Eine wichtige Grundlage für diesen auf räumliche Aggregation basierenden Ansatz ist die Strukturierung des Raumes in räumliche Einheiten (Bezirke oder Regionen), welche für die Forschungsfrage sinnvoll sind. Die Auswahl der Raumeinheiten ist dabei nicht trivial, denn sie kann das Ergebnis der Analyse maßgeblich beeinflussen, und eine objektiv beste Raumeinheit gibt es nicht (Openshaw 1984; Weigand et al. 2019). Reguläre Raster repräsentieren geographische Unterschiede unzureichend. Demgegenüber sind administrative Einheiten oft von historischen anstatt gegenwärtigen Umständen abgeleitet, nicht konsistent definiert und unterscheiden sich teils massiv von Land zu Land (Taubenböck et al. 2019).

Daher zielt ein Forschungsstrang darauf ab, Raumeinheiten nach den zu nutzenden Phänomenen und dafür gesammelten Daten zu definieren. Da es oft wünschenswert ist, Raumeinheiten mit hoher interner Homogenität zu erlangen, bieten sich Clustering-Ansätze, also Methoden des unüberwachten maschinellen Lernens an (siehe 2.1.1). Grundsätzlich ist es möglich, unüberwachte Lernmethoden wie K-Means (Lloyd 1982) und DBScan (Ester et al. 1996) auch auf räumliche Daten anzuwenden (Varghese et al. 2013), allerdings ergeben sich dabei nicht notwendigerweise räumlich kongruente und zusammenhängende Cluster. Wenn, aber, wie in der konzipierten Anwendung, räumliche Konsistenz wünschenswert ist, kann die räumliche Distanz als zusätzliches Kriterium in das Clustering aufgenommen werden (Distefano et al. 2020), wodurch sich Cluster generieren lassen welche kompakt und konzentriert sind (Chavent et al. 2018). Chavent et al. (2018) implementieren in dem R Paket ClustGeo eine hierarchische Methode basierend auf dem Ward-Minimalvarianz Kriterium. Bei dieser werden die inhaltlichen Distanzen D_0 und die räumlichen Distanzen D_1 mit einem Mischparameter α gewichtet und addiert

$$\alpha = (1 - \alpha) D_0 + \alpha D_1$$

Derartige Clustering-Ansätze basieren auf Distanzmatrizen, welche für numerische Daten wie die meisten Fernerkundungsdaten auf der Ebene von Pixeln berechenbar sind.

Mittels der kontextabhängige Quantifizierung von Textinhalten mit Text-Embeddings können auch Textinhalte statistisch verarbeitet, und somit auch Distanzmatrizen für die Nutzung in Clustering-algorithmen berechnet werden, was jedoch noch nicht zur Anwendung gekommen ist. Der von Chavent et al. (2018) für eine Eingangsvariable implementierte Algorithmus ist prinzipiell auf beliebig viele Variablen erweiterbar, was die Einbindung von Daten aus mehreren Quellen und Datentypen ermöglicht, solange sie numerisch quantifizierbar sind. Es besteht also hier noch unausgeschöpftes Potenzial darin, diesen Ansatz zu nutzen, um FED-Inputs und GSM-Inputs zu kombinieren und für die Generierung von Clustern als Raumeinheiten nutzbar zu machen.

Ein weiterer Forschungsstrang beschäftigt sich damit, Embeddings für räumliche Regionen zu generieren, oft mittels selbstüberwachten Lernens (Li et al. 2023) und einer potenziell großen Vielfalt an Geodaten, wobei die Nutzung von Microblog Daten noch nicht sehr verbreitet ist (siehe Chen et al., 2024, für eine Überblicksarbeit, welche auch andere Arten von Geodaten mit einbezieht). Diese generalisierbaren Repräsentationen können dann als Inputs für eine Vielzahl von Anwendungen genutzt werden, unter anderem um möglichst unterschiedliche Regionen zu identifizieren.

Während alle diese Ansätze valide sind, unterscheiden sich in der Menge an benötigten Daten und Rechenleistung, Transparenz und Anwendungsspezifität. Die verhältnismäßig einfachen räumlich-construierten Clustering Ansätze bestehen durch Transparenz und Flexibilität, da sie prinzipiell nur Distanzmatrizen benötigen, welche durch vortrainierte Embeddingsmodelle generiert werden können, und die Clustering Algorithmen verhältnismäßig wenig Rechenaufwand erfordern. Dass die Distanzmatrizen auch direkt ausgelesen und interpretiert werden können, stellt ein zusätzliches Potenzial für algorithmische Transparenz dar.

Die Anwendung von Datenfusion von GSM und FED in der Migrationsforschung ist basierend auf der Vielzahl an demonstrierten Anwendungen plausibel. Gleichzeitig ist noch nicht ausführlich erwiesen, inwiefern auch Textdaten aus GSM auf diese Weise für die Migrationsforschung ausgewertet werden können.

2.3.2. Nutzung von GSM zur Charakterisierung der Perspektive von Migrierten

Während zahlreiche wissenschaftliche Studien (beispielsweise Heidenreich et al. 2020; Lee und Nerges 2018; Rowe et al. 2021) sich mit dem Diskurs über Migration und Migrierte beschäftigen, ist die Perspektive der Migrierten selbst vergleichsweise wenig erforscht worden (Khatua und Nejd 2021).

Dabei ist es für die Einschätzung von Migrationsfaktoren essenziell, deren Wahrnehmung durch Menschen aus deren Perspektive zu berücksichtigen, und deren Neigungen und Interessen zu verstehen. GSM stellt in dieser Hinsicht eine reiche und große Datenquelle dar, in welcher sich Menschen frei über eine große Bandbreite von Themen äußern (Wang et al. 2019). Zusätzlich können Soziale Medien selbst eine Rolle in Migrationsentscheidungen spielen, indem sie zum Networking und Informationsaustausch in Netzwerken ermöglichen (Dekker und Engbersen 2014; Akanle et al. 2021). Die Studie von Daten der sozialen Medien kann daher Einblicke in die Interessen und Verhalten von Migrierten liefern (Reips und Buffardi 2012).

Khatua und Nejd (2021) nutzten Sprachmodelle zur Erkennung von Selbst-Identifizierung³⁸ von Migrierten in Tweet-Texten und weiter, um in deren Tweets deren Ansichten und Berichte zu ihren Herausforderungen und Eingewöhnungsprozesse zu identifizieren. Damit demonstrierten sie die Plausibilität der Analyse von Migriertenperspektiven auf Twitter, ergründeten aber auch die Herausforderungen. Denn mit dem Ansatz, Selbstidentifikation von Migrierten zu nutzen, ist wahrscheinlich nur eine sehr kleiner und nicht repräsentativer Anteil (Olteanu et al. 2019) der Migrierten erfassbar.

Für georeferenzierte Soziale Medien besteht eine alternative Möglichkeit, Migrierte zu identifizieren. Dabei werden Trajektorien analysiert (siehe 2.1.6 für eine Beschreibung der Methoden). Über die letzten Jahre wurden in der Forschung zahlreichen Methoden entwickelt, welche zur Ableitung von Mobilitätsinformationen und -Kategorien in der Lage sind. Einige dieser Methoden wurden für Twitter-Daten erprobt, und kommen im Rahmen dieser Dissertation zur Anwendung.

Diese Methoden wurden primär entwickelt, um Mobilitätsströme und -Bestände zu messen. Das Potenzial, auf Basis der Georeferenzen auch die von Migrierten erstellten Inhalte zu betrachten, wurde in den Arbeiten von Kim et al. (2022a) und Kim et al. (2022b) demonstriert,

³⁸ Beispiel: "I'm an immigrant [...]"

bei welchen Migrierte und Stationäre (im *Original Natives*)³⁹ über die Metadaten und Hashtags ihrer Tweets unterschieden wurden. Obwohl Metadaten und Hashtags relativ leicht auszuwerten und inhaltsreich sind, könnte ein noch größeres Potenzial in den Textinhalten liegen. Fortschritte in NLP-Methoden ermöglichen zunehmend die algorithmische Extraktion von inhaltlichen Informationen aus großen Datenmengen von auch nicht-formellen und unstrukturierten Textdaten wie GSM. Die Kombination von Trajektorienanalysen und Textprozessierung birgt also noch unausgeschöpftes Potenzial.

2.3.3. Nutzung von GSM zur Analyse sprachlicher Elemente in Nigeria

Die Sprachwissenschaft bedient sich schon lange der Anwendung korpuslinguistischer Methoden auf Webtextdaten. Dabei können linguistische Hypothesen *corpus-based* empirisch überprüft oder aufgrund der Korpora⁴⁰ *corpus-driven* entwickelt werden (Tognini-Bonelli 2001). Trotz der Unstrukturiertheit von Webtextdatensätzen zeigte sich, dass die hierdurch ermöglichte Analyse großer Datenmengen für viele sprachwissenschaftliche Anwendungen ein geeignetes Mittel darstellt (Hundt et al. 2007; Lüdeling et al. 2007).

Fortschritte in Prozessierungsressourcen und -algorithmen ermöglichen die zunehmende Skalierung von korpuslinguistischen Methoden auf große, mit vielen Metadaten versehene Korpora. Hierbei spielen Frequenzanalysen eine große Rolle, welche die Häufigkeit von Sprachelementen und -mustern erfassen (Mukherjee 2008). Diese empirischen Messungen können über soziale Prozesse Aufschluss geben.

Innerhalb der Sprachwissenschaft befasst sich primär die Soziolinguistik damit, die Interaktion zwischen Gesellschaft und Sprache zu untersuchen. Im Migrationskontext geht es dabei beispielsweise um die Entstehung von Kontaktsprachen, die Sprachwahl und -akquise von Migrierten in ihren neuen (und alten) Umgebungen und dem gesellschaftlichen Status von Minderheitensprachen (Holmes & Wilson 2022). Social-Media-Inhalte stellen hierbei eine reiche Datenquelle dar, da sie oft mit reichen Metadaten annotiert sind, welche über den Kontext des Textes Aufschluss geben können (van der Lee und van den Bosch 2017).

Auch geographische Merkmale können hierfür analysiert werden. Die Verortung von Texten und die Analyse verorteter Texte mit Unterstützung Geographischer Informationssysteme ist in der Sprach- und Literaturwissenschaft eine erprobte Praxis (Gregory et al. 2015). Für Anwendungen im Bereich der Migrationsforschung ist bedeutsam, dass auch geographische Informationen auf Ebene der nutzenden Personen ausgewertet werden können. So können Textinhalte in Bezug zur Herkunft der Textautoren, deren aktuellen und früheren Aufenthaltsorten und Migrationsstatus gesetzt und nach soziolinguistischen Ansätzen ausgewertet werden. Neben der händischen Kodierung von geographischen Informationen durch Lesen der Nutzerbeiträge und Profilinformatoren können auch algorithmische Ansätze genutzt werden (Heyd 2014). Mehrere Studien im erweiterten Kontext westafrikanischer Migration nutzten diese geographischen Informationen, um von mit Nigeria assoziierten Menschen genutzte sprachliche Merkmale einzuordnen und damit beispielsweise die gegenseitige Beziehung zwischen deren Sprache und Migrationserfahrung zu analysieren (Heyd und Honkanen 2015; Honkanen und Müller 2021; Mair 2021) oder

³⁹ Welche unterschieden wurden, indem über längere Zeiträume die Georeferenzen von Accounts mit denen ihrer Netzwerke verglichen wurden

⁴⁰ Als „Korpus“ wird in dieser Thesis ein gut annotierter, authentischer, maschinenlesbarer und für den Forschungszweck repräsentativ zusammengestellter Datensatz bezeichnet (Mukherjee 2008).

Repräsentationen von ethischer Zugehörigkeit und Narrative von Heimat zu identifizieren (Heyd 2014).

Informationen zu Ort und Mobilität können daher eine wichtige Rolle in der Analyse von nigerianischen Sprachen und Migration spielen. Dabei können sowohl qualitative als auch quantitative Ansätze genutzt werden, wie die Studien von Heyd (2014) und Honkanen und Müller (2021) zeigen. Sowohl Sprachmerkmale als auch geographische Informationen können prinzipiell algorithmisch abgeleitet werden.

Modelle wie der *Compact Language Detector 3* (CLC3) von Google (2016) sind auf die Klassifizierung von Texten in Sprachen⁴¹ trainiert, und von Twitter bereitgestellte Tweets waren mit algorithmisch annotierten⁴² Sprachcodes versehen. Sprachmodelle sind jedoch oft nicht auf die Erkennung von Sprachvarietäten Afrikas trainiert, für welche oft nur relativ (im Vergleich zur Anzahl der sie sprechenden Menschen) wenige Textdaten vorliegen.

Zusammenfassend lässt sich sagen, dass die algorithmische Ableitung von Merkmalen nigerianischer Sprachen bereits demonstriert und angewendet wurde, und auch der Bezug von Geographie und Sprache bereits in mehreren Studien im Kontext nigerianischer Sprachen hergestellt wurde. Diese Informationen wurden jedoch noch nicht in Bezug zu Mobilitätsinformationen über die Nutzenden gesetzt, wie sie beispielsweise aus der Analyse von Trajektorien gewonnen werden können.

3. Konzipierung der Studien

Die Arbeiten dieser Dissertation bestehen aus sechs Studien, welche sich in drei grundlegende Richtungen strukturieren lassen. In diesem Kapitel werden für jede der Studien der konzeptionelle Rahmen erläutert, in welchem der methodische Ansatz entwickelt wird, der die in der Einleitung definierten Leitfragen beantworten soll.

3.1. Studie A1: Verbindung von Fernerkundungsdaten und Twitter-Daten zur

Erfassung digitaler Disparitäten in Afrikas expandierenden städtischen Räumen

Im 21. Jahrhundert sind Städte durchdrungen von Sensoren und ihre Bevölkerung verbringt einen nennenswerten Teil ihrer Zeit sowohl in digitalen als auch physischen Räumen (Kemp 2025). Konzepte wie *Internet of Things*, *Smart Cities* und *Digital Citizenship* verdeutlichen, dass Städte zunehmend eine parallele Repräsentation im digitalen Raum aufbauen (Robertson und Feick 2016).

Dieses digitale Abbild, welche sich zum Teil aus nutzergenerierten Informationen (beispielsweise in Social Media oder Onlinekartendiensten) besteht, kann als wichtige Datenquelle zur Erforschung städtischer Nachhaltigkeit dienen. Insbesondere in den Bereichen ökologische Nachhaltigkeit, Gesundheit, Soziale Gerechtigkeit, Mobilität, Wirtschaftsentwicklung (Ilieva und McPhearson 2018) können digitale geographische Daten zur Forschung beitragen. Für die Bewohner der Stadt kann die digitale Repräsentation der Städte deren digitale Demokratie, Administration, Wirtschaft, und Autonomie fördern (Robertson und Feick 2016). Da die Macht, das Stadtbild zu gestalten, historisch in den Händen von Autoritäten und Eliten konzentriert war, wird die digitale Entwicklung von manchen als einen grundsätzlichen Wandel betrachtet, bei welchem diese Macht dezentralisiert wird (Boy

⁴¹ Das CLD3 Modell unterscheidet 107 Sprachen, siehe <https://github.com/google/cld3>.

⁴² Die Methode, welche Twitter hierfür nutzt, ist dem Autor nicht bekannt.

and Uitermark 2017).

Doch auch die räumlich-digitale Repräsentation ist weder im Raum noch über Bevölkerungsgruppen gleichmäßig verteilt (Zhu und Lerman 2016; Boy und Uitermark 2017). In der Praxis scheinen soziale Medien zudem oft Prozesse der Stratifikation und Ungleichheit aufrechtzuerhalten und zu verstärken, sowie Selbstsegregation und Tribalisierung zu fördern (Törnberg 2018). Ob Ungleichheiten der Gestaltungsmacht in Städten durch die digitale Entwicklung tatsächlich beseitigt oder aber gar verstärkt werden, muss daher noch erforscht werden (Boy und Uitermark 2017). Wie die Literatur (2.2.1) zeigt, wurden digitale Disparitäten von nutzergenerierten Inhalten in der Stadt bisher aus zwei Perspektiven untersucht: Die sozialwissenschaftliche Perspektive thematisiert die soziale Bedeutung der Verteilung der Inhalte. Die datenorientierte Perspektive demgegenüber zielt auf das Ermessen von Verfügbarkeit, Qualität, Konsistenz und Repräsentativität der Daten ab (Robertson und Feick 2016). Für das Thema dieser Dissertation sind beide Perspektiven unmittelbar relevant. Einerseits kann die Verteilung von GSM auf sozioökonomische Ungleichheiten hindeuten, welche Migration bedingen oder aus ihr resultieren. Andererseits bedingen Datenverfügbarkeit und -Qualität in urbanen Untersuchungsregionen die Möglichkeiten, dort Studien mittels GSM durchzuführen.

Aus dem eingangs dargelegten Zusammenhang zwischen Städtewachstum und Migration ergibt sich eine Notwendigkeit, die digitale Repräsentation der neuen Stadtflächen ins Verhältnis zu dem der älteren Stadtflächen zu setzen. Entwickelt sich die online-Repräsentation dieser Flächen in ähnlichem Tempo wie ihre physische Expansion? Kann die digitale *urban frontier* Schritt mit der physischen *urban frontier* Schritt halten, oder konzentrieren sich digitale Inhalte weiterhin auf die älteren Zonen der Städte?

Die erste Studie hat daher zum Ziel, den Zusammenhang zwischen Dichte der Twitter-Repräsentation und dem Alter der Bebauung in afrikanischen Siedlungsräumen zu ermitteln. Mittels fernerkundlicher Methoden wird das Bebauungsjahr von Siedlungsflächen abgeleitet. Innerhalb dieser Siedlungsflächen wird die Verteilung von Geotweets gemessen.

Um eine große und robuste Datenbasis zu generieren, wird die Analyse nicht für ein einzelnes Stichdatum durchgeführt, sondern über einen **Analysezeitraum** von 10 Jahren von 2010 bis 2019 für jedes Kalenderjahr separat durchgeführt und zuletzt über den Analysezeitraum gemittelt. In diesem Analysezeitraum wird die Verfügbarkeit von Twitter-Daten als gut eingeschätzt. Geographisch fokussiert sich die Analyse auf 135 urbane Testgebiete (UTG, Sektion 4.2) mit 5km Radius in welchen besonders starkes Siedlungswachstum gemessen wird. Die Auswahl dieser **Testgebiete** erfolgt nach Kriterien des Siedlungswachstums, welches mittels Erdbeobachtungsdaten gemessen wird, und der geographischen Verteilung: Für jede der fünf UN-Regionen Afrikas werden die Orte, um die es in dem analysierten Jahrzehnt das höchste Wachstum gab, ausgewählt.

Innerhalb dieser Testregionen werden sodann alle verfügbaren Geotweets gesammelt und verortet. Nur Geotweets mit hoher Präzision werden für die Analyse beibehalten, welche räumlich auf der feinstmöglichen durch FED erlaubten Auflösung von 10m durchgeführt wird.

Vor dem Hintergrund der Analyse, welche sich mit der Sichtbarkeit von Stadtteilen befasst, wird eine **Gewichtung** der Tweets als sinnvoll erachtet, da nicht jeder Tweet dieselbe

Sichtbarkeit generiert. Beispielsweise wird davon ausgegangen, dass Retweets⁴³ eines Geotweets die Sichtbarkeit eines Ortes in ähnlichen Maße erhöhen wie der ursprüngliche (verlinkte) Geotweet, auch wenn die Retweets selbst keine Geotags enthalten. Umgekehrt ist es wahrscheinlich, dass keine erhöhte Sichtbarkeit für einen Ort zu erwarten ist, wenn ein Account innerhalb eines kurzen Zeitraumes diesen Ort wiederholt geotaggt⁴⁴. Da mit der Messung die digitale Repräsentation gemessen werden soll, werden Tweets mit mehreren Retweets höher gewichtet. Auch wenn der genaue Algorithmus, welchen Twitter für die Auswahl von angezeigten Tweets nutzt, nicht bekannt ist⁴⁵, werden in der Annahme, dass Retweets und Antworten auf einen Tweet dessen Sichtbarkeit und damit die Repräsentation des getaggtten Ortes erhöhen, Tweets nach Retweets und Antworten höher gewichtet, und wiederholte Tweets eines Accounts zum selben Ort geringer gewichtet.

Das Bebauungsjahr der bebauten Flächen wird mittels Satellitenfernerkundung kartiert. Hierbei werden über den Analysezeitraum 2010-2019 hinaus auch frühere Bebauungsjahre erfasst, um die breitmöglichste Spanne von Bebauungsalter abzudecken. Durch eine Kombination von Fernerkundungsprodukten kann eine Unterscheidung von Bebauungsjahren im Zeitraum 1985-2020 vorgenommen werden. Frühere Baujahre können aufgrund der Datenverfügbarkeit nicht differenziert werden, und werden auf 1985 begrenzt.

Um die Beziehung zwischen Geotweets und Bebauungsalter herzustellen, wird jedes UTG für jedes Kalenderjahr des Untersuchungszeitraums in Zonen unterteilt, die dem Jahr der Bebauung entsprechen. Für jede dieser Zonen wird die bebaute Fläche gemessen sowie die gewichtete Anzahl an Geotweets summiert. Aus dem Verhältnis der beiden Werte ergibt sich die gewichtete Geotweet-Dichte pro Bebauungsjahr. Die Beziehung zwischen der Dichte und dem Bebauungsjahr (und damit Alter der Bebauung) ist visuell als **Kurve** interpretierbar. Die Ausgewogenheit dieser Dichtekurve ist die grundsätzliche Kernfrage der Analyse. Wäre die GSM-Aktivität unabhängig vom Baujahr, ergäbe sich hier theoretisch eine diskrete Gleichverteilung. Ein systematischer Unterschied, bei welchem sich Alter auf die GSM-Aktivität auswirkt, wäre durch eine schiefe Verteilung sichtbar.

Um dies statistisch messbar und vergleichbar zu machen, wird die Schiefe der Verteilung systematisch für jedes Testgebiet und Analysejahr quantifiziert. Hierfür wird eine Metrik analog zu typischen Schiefemaßen entwickelt, welche als Schiefe der Digitalen Dichte bezeichnet wird (engl. *Digital Density Skew, DDS*). Bestehende Maße, wie Divergenzmetriken (Lin 1991) oder lineare Regressionen der Verteilungskurve erfüllen die besonderen Anforderungen dieser Studie nicht, welche darin bestehen, die Schiefe der Verteilungen zwischen verschiedenen Testgebieten vergleichbar und interpretierbar zu machen. Hierfür muss eine Metrik klar begrenzt und sinnvoll zwischen Sequenzen unterschiedlicher Länge vergleichbar sein, sowie ein interpretierbares Minimum und Maximum aufweisen. Der DDS erfüllt diese speziellen Anforderungen der Studie, soll aber nicht über die Studie hinaus Gültigkeit haben, sondern lediglich innerhalb der Studie für den Vergleich genutzt werden. Nachdem der DDS-Wert für jedes Analysejahr gemittelt wird, erlaubt er eine Einschätzung, inwiefern in einem Testgebiet die Dichte der GSM-Aktivität mit dem Alter der Bebauung

⁴³ Das Weiterleiten eines Tweets durch andere Benutzer wird als „Retweet“ bezeichnet. Retweets enthalten dieselben Inhalte wie der ursprüngliche Tweet, jedoch keine Geotags.

⁴⁴ Was auch als Anzeichen für Automatisierung sein kann

⁴⁵ Bekannt ist, dass der Algorithmus zumindest in Teilen auf tiefem maschinellen Lernen basiert (Bandy and Diakopoulos 2021; Twitter Team 2023).

zusammenhängt. Besteht ein solcher Zusammenhang wird dies im Rahmen der Dissertation als Digitale Disparität bezeichnet. Es stellt in erster Linie eine geographische datentechnische Beobachtung dar und ist keine Bewertung im Sinne sozialer Ungleichheiten oder ethischer Ungerechtigkeiten.

Um die Signifikanz der gemessenen DDS-Werte zu bewerten wird ein **Permutationstest** genutzt. Es ist bekannt, dass GSM nicht komplett zufällig im Raum verteilt sind, sondern zur Häufung neigen (Steiger et al. 2015). Um dieser Eigenschaft Rechnung zu tragen, wird daher für jede Region und für jedes Analysejahr ein Clustermodell auf die wahre räumliche Verteilung der Tweets geschätzt. Basierend auf diesem Modell werden zufällige räumliche Verteilungen generiert, welche eine ähnliche Häufung aufweisen wie die realen Daten. Die Berechnung des DDS aufgrund dieser zufälligen Permutationen ergibt eine Referenzkurve, vor welcher die Signifikanz des tatsächlich gemessenen DDS evaluiert werden kann.

Aufgrund der Existenz von demographischen Verzerrungen und technischen Fehlern in sozialen Medien (Sektion 32 und Olteanu et al. 2019; Merry und Bettinger 2019) wird in dieser Studie ein besonderer Fokus darauf gelegt, die Robustheit der Ergebnisse zu testen. Es werden Tests durchgeführt, um den Einfluss besonders aktiver User, niedrigen Datenvolumens, und möglicher GPS-Ungenauigkeiten auf die berechneten GSM zu evaluieren.

Für jedes UTG ergibt sich ein mit Signifikanz versehener DDS-Wert. Diese werden dann in verschiedenen Schritten **analysiert**. Zunächst erfolgt dies individuell für ausgewählte Untersuchungsgebiete, um ein Verständnis zu entwickeln, wie DDS-Werte die räumlichen Geotweet-Verteilungen in ihrer geographischen Umgebung quantifizieren. Hierbei wird eine visuelle Interpretation der kartierten Geotweet-Verteilungen durch eine Webrecherche unterstützt, welche OpenStreetMap⁴⁶ und Google Maps⁴⁷ als Quellen heranzieht.

Die Verteilungen der Twitter-Dichte werden vermutlich von zahlreichen demographischen, kulturellen, politischen und wirtschaftlichen Rahmenbedingungen sowie der lokalen Planungskultur und -geschichte beeinflusst. Da für diese Faktoren keine Daten afrikaweit verfügbar sind, können deren Einflüsse nicht analytisch ergründet werden. Jedoch können weitere Geodaten genutzt werden, um diese Faktoren zu approximieren, und eine grundlegende Richtungsweisung zu ermöglichen. Dies fußt auf dem Grundgedanken, dass wenn kein Zusammenhang zwischen sozioökonomischen Faktoren und der digitalen Disparität besteht, ein statistischer Zusammenhang mit abgeleiteten Proxy-Variablen unwahrscheinlich ist. Auch wenn die unterliegenden Prozesse sehr komplex sind, und Disparitäten nicht auf bestimmte Faktoren zurückgeführt werden können, ist hierdurch doch eine erste Einschätzung möglich.

In einem ersten Ansatz dienen geographische Regionen als Proxy-Kategorien für kulturelle oder sozioökonomische Räume. Die Form und Ausprägung des Urbanisierungs- und Migrationsgeschehens kann zwischen Teilen des afrikanischen Kontinents sehr unterschiedlich sein, und viele der oben genannten Faktoren unterscheiden sich auf geographisch regionaler Ebene (McAuliffe 2024; Kamana et al. 2024). Daher wird die Verteilung des DDS zwischen den fünf durch die Vereinten Nationen definierten Georegionen Afrikas (United Nations Statistical Office 1982) verglichen.

⁴⁶ <https://www.openstreetmap.org/>

⁴⁷ <https://www.google.com/maps>

Zusätzlich kann die Ausprägung des Städtewachstums, welche sich mit Satellitendaten großflächig erfassen lässt (siehe beispielsweise Taubenböck et al. 2020), eingesetzt werden, um soziale Prozesse zu approximieren. Ein Ansatz hierzu basiert auf der Geplantheitsintensität (engl. *intensity of plannedness*, IoP) der Untersuchungsregion, welche sich nach Debray et al. (2023) in fünf diskrete Klassen einteilen lässt. Diese Einteilung fasst die historischen Prozesse, die das innerstädtische Layout beeinflusst haben, zusammen, wie demographisches Wachstum, Kolonisierung (Şen 2024), Wirtschaftswachstum und Stadtplanungskultur (Taylor 2013) und viele weitere, zusammen. Die IoP wird durch Expertenanalyse mittels visueller Interpretation des Straßen- und Blocklayouts eingeschätzt (4.4.3).

Der dritte Ansatz besteht in der Modellierung des Zusammenhangs zwischen Disparität und räumlichen Metriken der lokalen Bebauungsstruktur. Diese Metriken, welche aus FED abgeleitet und pro UTG berechnet werden können, quantifizieren die morphologische Konfiguration des bebauten Gebietes. Sapena et al. (2020) zeigten, dass aus FED abgeleitete räumliche Metriken in der Lage sind, sozioökonomische Indikatoren zu approximieren. Es wird daher ein Satz von räumlichen Metriken auf Basis der kartierten Bebauung zum Ende des Analysezeitraums berechnet (4.4.4). Sodann wird ein Test auf einen linearen Zusammenhang zwischen den Metriken und dem DDS durchgeführt, sowohl für jede Metrik in Isolation (einfache lineare Regression) als auch für mehrere Metriken zusammen (multiple lineare Regression).

Durch diese Studie soll insgesamt gezeigt werden, dass FED gemeinsam mit GSM ausgewertet werden können, um die physische und die digitale Repräsentation der Städte zu vergleichen, und eine bislang unerforschte Form von digitalen Disparitäten aufzudecken. Zusätzlich kann die Datenverfügbarkeit in den für die Migrationsforschung sehr relevanten Regionen eingeschätzt werden.

3.2. Studie A2: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Ableitung von Raumeinheiten für die explorative Analyse von physischen und thematischen Charakteristika urbaner Räume

Geotweets enthalten neben den in Studie A1 genutzten Georeferenzen auch Textinhalte, welche im Kontext dieser Dissertation von primärem Interesse sind, da sie eine Vielzahl von Anwendungen ermöglichen.

Trotz der unterschiedlichen Charakteristika von GSM und FED ist ihre Kombination technisch möglich. Wie die Literaturanalyse zeigt, bietet der aktuelle Stand der Forschung verschiedene Methoden zur Kombination von verschiedenen Formaten und Modalitäten. Dazu gehört die Nutzung von Mixed-Method Ansätzen, bei welchen die Daten in ihren ursprünglichen Formaten verbleiben, sowie die Nutzung von maschinellem Lernen und Embeddings für die Datenfusion auf verschiedenen Ebenen.

Der hier entwickelte Ansatz bedient sich Elemente beider Richtungen, indem mittels maschinellen Lernens der Raum und Inhalte der Tweets in sinnvolle Einheiten strukturiert werden, auf deren Basis qualitative und quantitative Datenexploration durchgeführt werden kann. Konkret ist das Ziel, Mixed-Methods für FED und GSM zu verbessern, indem mittels maschinellen Lernens eine sinnvolle Unterteilung des zu untersuchenden urbanen Raumes in Raumeinheiten erstellt wird. Diese Raumeinheiten sollen beispielsweise Blöcken, Nachbarschaften und Bauprojekten entsprechen, welche sich in den zu untersuchenden Eigenschaften intern ähnlich sind und sich als Gruppe in diesen Eigenschaften von ihrer

Umgebung abgrenzen.

Die Strukturierung des zu untersuchenden Raumes auf Basis der für diese Untersuchung herangezogenen Daten unterstützt die Analyse dieser Daten mittels verschiedener Analysewerkzeuge und Mixed-Method Ansätze. Hierdurch soll eine Herangehensweise für die Arbeit mit Fernerkundungsdaten und Textdaten gefördert werden, welche die unterschiedlichen Eigenschaften beider Datenquellen als Mehrwert begreift, und sowohl die Anwendung von qualitativen als auch quantitativen Methoden aus verschiedenen Forschungsfeldern ermöglicht. Beispielsweise können Texte aus einer Nachbarschaft in Diskursanalysen ausgewertet, und repräsentative Texte aus ihrer räumlichen Nachbarschaft dargestellt werden, während zusätzlich quantitative Metriken aus FED-Informationen über die Umwelt liefern, wie Begrünung oder Alter der Gebäude. Angesichts der Bedeutung von Interdisziplinarität (Brettell und Hollifield 2015) und Stakeholder-Engagement in der Migrationsforschung (McAuliffe et al. 2018) kann diese Anwendung gemischter Methoden einen großen Mehrwert bringen, welcher mit rein statistischen Analysen nicht erreichbar ist.

Zur Gruppierung von Textdaten ist in der Textanalyse die Methode des Topic-Modeling weit verbreitet. Für die Gruppierung von Inhalten im geographischen Raum soll in dieser Studie ein möglicher Ansatz entwickelt und demonstriert werden, welcher sich für heterogene Datentypen, also sowohl für FED als auch GSM-Texte eignet.

In dem hier genutzten Ansatz werden die verschiedenen Eingangsdatentypen, FED und GSM, in ein Rasterformat überführt und dann parallel in einem unüberwachten maschinellen Lernverfahren zur Erstellung von Gruppen (engl. Cluster) merkmalsähnlicher Rasterzellen genutzt. Dies kann aus technischer Sicht als eine feature-level Datenfusion betrachtet werden (vergleiche 2.1.3), ist jedoch für einen modularen und interpretierbaren Aufbau in explorative Herangehensweisen und visuelle Darstellungen konzipiert und hebt sich damit praktisch von Datenfusionstechniken ab.

Der Ansatz wird für acht der in Studie A1 genutzten urbanen Testgebiete angewandt, wobei anders als in Studie A1 keine Strukturierung nach Jahren erfolgt. Ein Clusteringverfahren auf Pixelbasis wird angewendet, welches auf einer Erweiterung des als *ClustGeo* publizierten Ansatzes von Chavent et al. (2018) basiert. Das Verfahren verwendet in dieser Studie sowohl die Twitter-Daten als auch FED, welche in Studie A1 zur Messung der Digitalen Disparität genutzt werden.

Grundlage für ein Clustering ist eine Matrix von Distanzen im Merkmalsraum, welche von den Eingangsdaten abgeleitet ist. Für die FED ist das der Altersunterschied der bebauten Pixel, welcher sich einfach als Differenz der Pixel-Werte des Rasters berechnen lässt. Für Texte ist die Erstellung einer numerischen Distanzmatrix nicht direkt möglich. Um die semantischen Inhalte der Twitter Texte für ein Clustering nutzbar zu machen, werden sie mittels eines LLM in Embeddings (siehe 2.1.2) umgewandelt, welche mittels statistischer Verfahren auf wenige (5) Dimensionen reduziert und dann mittels geographischer Extrapolation in ein reguläres räumliches Raster von Embeddings konvertiert werden.

In der veröffentlichten Version der Software *ClustGeo* ist grundsätzlich nur das Clustering basierend auf einer einzelnen Variable sowie der geographische Distanz berücksichtigt, effektiv also von zwei Distanzmatrizen, welche mit einem Mischparameter α verbunden werden. Um sowohl FED als auch die GSM-Texte zu integrieren, wird der Ansatz insofern erweitert, dass eine finale Distanzmatrix aus beliebig vielen input-Matrizen zusammengesetzt wird. Die dabei genutzten Gewichte können als Parameter genutzt verwendet werden, um

bestimmte Inputs höher zu bewerten. Beispielsweise kann, wenn kompaktere Cluster gewünscht werden, das Gewicht der geographischen Distanz erhöht werden. Wird dieses Gewicht auf 0 gesetzt, entspricht das Ergebnis dem eines nicht-geographischen Clusterings.

Es besteht dadurch ein hohes Maß an Subjektivität in der Wahl der Parameter. Die Studie zielt daher explizit nicht darauf ab, objektiv hochwertigere Raumeinheiten für Analysezwecke zu generieren, sondern flexible und plausible Raumeinheiten für explorative Zwecke und Visualisierungen.

Die Bewertung der Raumeinheiten erfolgt zunächst durch die qualitative Interpretation der Cluster, welche mit einer Mischung aus Webrecherche, GIS-Analyse und Textanalyse mittels Topic-Modeling unterstützt wird. Für das Topic-Modeling wird die von Grootendorst (2022) entwickelte Abfolge von BERT-basiertem Text-Embedding, Dimensionsreduktion und Clustering genutzt. Da die ersten beiden Schritte auch für das räumliche Clustering genutzt werden, können für das Topic-Modeling dieselben Embeddings genutzt werden, wodurch eine hohe Angleichung zwischen Clustern und Themen erreicht wird.

Zusätzlich wird auf einer statistischen Ebene untersucht, inwiefern die generierten Cluster eine Homogenität in Bezug auf die Inputdaten aufweisen.

Insgesamt soll also ein praktikabler Ansatz gezeigt werden, welcher einen Beitrag zur interdisziplinären Forschung leisten kann, indem er FED und GSM auf modulare, transparente und interpretierbare Weise kombiniert.

3.3. Studie B1: Differenzierung textinhaltlicher Präferenzen von Nigerianischen Twitter-Nutzenden nach Mobilitätscharakteristika

In der dritten Studie sollen Mobilitätsmerkmale von Twitter-Nutzenden sowie thematische Schwerpunkte ihrer Tweets abgeleitet und verbunden werden. Zusammengeführt ermöglicht dies die Differenzierung von inhaltlichen Präferenzen nach Mobilitätsgraden. Konkrete Zielsetzung dieser Studie soll die Identifikation von thematischen Schwerpunkten von Migrierten sein, konkret: Themen, welche bei nigerianischen Migrierten einen höheren Anteil der Tweets ausmachen als bei nigerianischen nicht-Migrierten.

Der methodische Ansatz, der dies realisieren soll, erfordert zunächst die Ableitung von Mobilitätscharakteristika, welche hier primär die Form einer Einteilung von Nutzenden in Mobilitätstypen annimmt. Aufgrund der Vielfalt von Migrationsformen und unterschiedlichen Definitionen soll diese Einteilung flexibel und parametrisierbar implementiert sein.

Zunächst wird hierfür eine Datengrundlage geschaffen, welche aus Geotweet-Timelines für eine große Anzahl an nigerianischen Nutzenden besteht. Als Analysezeitraum, für welchen Daten akquiriert werden, wird die Periode zwischen dem 01. Januar 2015 und dem 01. Januar 2019 gewählt, was einem Zeitraum von vier Kalenderjahren entspricht. Die Wahl dieses Zeitraums soll bestimmte Ereignisse, welche die Demographie und Verfügbarkeit von Twitter in Nigeria beeinflussen, ausklammern, und damit temporäre Biases minimieren (Olteanu et al. 2019). Die spätere Grenze des Zeitraums ist bedingt durch Änderungen in der Geotagging Funktionalität durch den Plattformbetreiber Twitter im Juni 2019 (Kruspe et al. 2021), und die #endsars Proteste in Nigeria, welche zu einem starken Anstieg der Twitter-Nutzung führten (Ojedokun et al. 2021), sowie dem zeitweisen Verbot von Twitter durch die nigerianische Regierung im Jahr 2021 (Princewill et al. 2021). Die frühere Grenze der Zeitspanne wurde anhand vorangegangener Experimente zur Datenverfügbarkeit gewählt, welche nahelegten,

dass die Geotag-Funktion ab Mitte des Jahres 2014 weitreichend in Nigeria genutzt wurde (Abbildung 11)

Für die so definierte Zeitspanne werden zunächst Geotweets für die Untersuchungsregion Nigeria gesammelt.

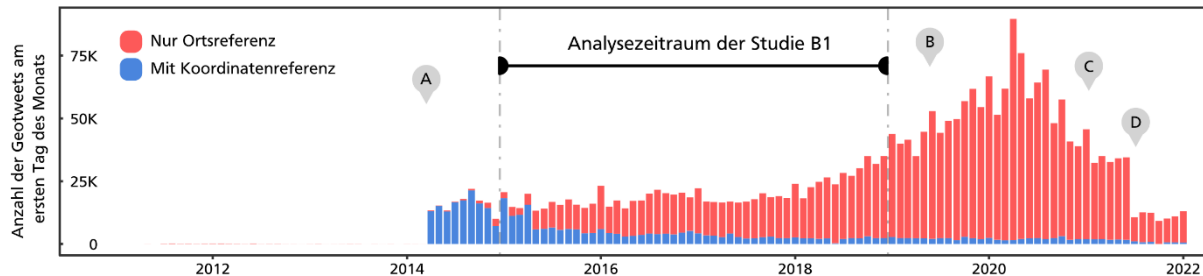


Abbildung 11: Anzahl abgerufener Geotweets aus Nigeria zwischen 2011 und 2022, eingefärbt nach Art der Georeferenz. Einige Ereignisse sind angedeutet: (A) Erstes nennenswertes Auftreten von Geotweets in Nigeria; (B) Änderungen in der Geotagging-Funktion (Kruspe et al. 2021); (C) #endsars Protestbewegung, in welcher Twitter eine wichtige Rolle spielte (Ojedokun et al. 2021); (D) zeitweises Verbot von Twitter durch die nigerianische Regierung (Princewill et al. 2021).

In dem Geotweet-Datensatz, welcher sich nach Accounts aufschlüsseln lässt, werden Accounts identifiziert, aus welchen für die weitere Analyse geeignete Nutzende identifiziert werden. Primär werden in diesem Schritt automatisierte Accounts (Bots) anhand ihrer Aktivitätsmerkmale erkannt und entfernt. Aus dem Geotweet-Datensatz werden dann für jeden geeigneten Account Trajektorien für eine Mobilitätsanalyse abgeleitet. Um die vollständigen globalen Trajektorien der Accounts zu konstruieren, werden für diese Accounts auch alle Geotweets außerhalb Nigerias abgerufen.

Alle diese Geotweets werden dann räumlich auf Raumeinheiten aggregiert. Diese Aggregation erfolgt außerhalb Nigerias auf Länderebene. Um regionale Mobilität zu erfassen, erfolgt sie innerhalb Nigerias auf der Ebene von hexagonalen Rasterzellen, in welche die Fläche des Staatsgebiets unterteilt wird. Um zu verhindern, dass bei dieser Rasterung Städte durch die Ränder von Rasterzellen geteilt werden, werden Metropolen identifiziert und als separate kreisrunde Zonen von ähnlicher Größe wie die der Rasterzellen abgegrenzt (Abbildung 12).

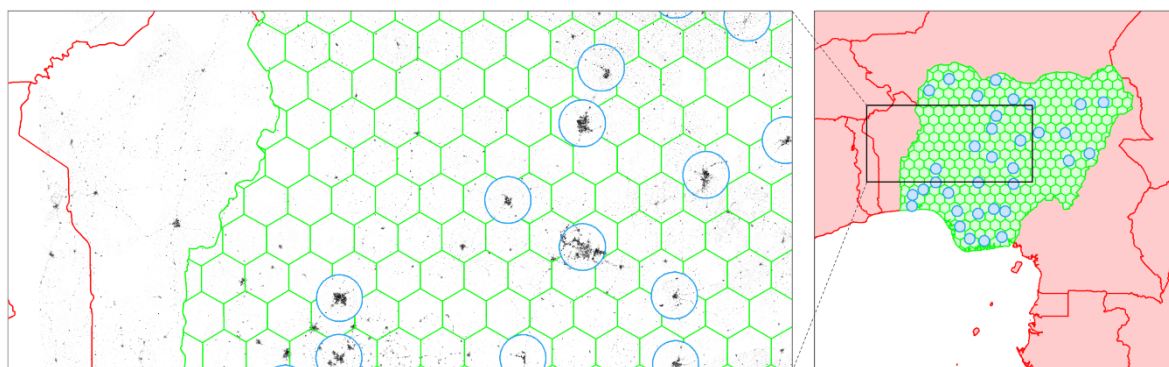


Abbildung 12: Raumeinheiten in Studie B1. Rot: Internationale Grenzen. Grün: Hexagonale Rasterzellen innerhalb Nigerias. Blau: Zonen um nigerianische Metropolen.

Sodann wird der von Chi et al. (2020) entwickelte Algorithmus, welcher bereits auf Tweets getestet wurde, in leicht angepasster Form eingesetzt, um die Accounts anhand ihrer

Trajektorien in drei Mobilitätstypen einzuteilen: (1) In Nigeria Stationäre; (2) Mobile; (3) International Migrierte.

Die Ableitung von thematischen Indikatoren aus den Texten erfolgt sodann nur für jene Accounts, welche sich mit großer Sicherheit einer dieser Gruppen zuordnen lassen.

Für diese Accounts liegen bereits die für die Trajektorien gesammelten Geotweets vor, welche aber für die Textanalyse möglicherweise nicht geeignet sind. Denn es ist möglich, dass Geotweets aufgrund thematischer Geospatiality (siehe Sektion 2.2.2) inhaltlich nicht repräsentativ für die Gesamtmenge der Tweets der Nutzenden sind –ein Effekt, welcher in Studie C2 genauer ergründet wird. Um in dieser Studie eine bestmögliche Abdeckung der Gesamtmenge an Tweets zu erhalten, werden für die analysierten Accounts auch Tweets ohne Geotags abgerufen. Da die Twitter API keine dezidierte Funktion für Stichproben bereitstellt, ist die Abfrage einer repräsentativen Stichprobe durchaus herausfordernd. Die gesamte Tweet-Historie aller Accounts abzurufen ist basierend auf den Beschränkungen des gewährten API-Zugang nicht plausibel. Der Abruf einer geringeren Menge würde jedoch nur die neuesten Tweets zurückgeben, was zu zeitlichen Verzerrungen führen und frühere Zeiträume des Analysezeitraums stark benachteiligen würde. Daher wird ein Stichprobenansatz entwickelt, welcher über den Analysezeitraums hinweg Tweets aus zufällig verteilten 7-Tage Perioden ausliest. Hierdurch werden zeitliche Verzerrungen minimiert, welche den Inhalt beeinflussen könnten (Olteanu et al. 2019). Beispielsweise könnten, wenn Tweets immer vom Ende des Kalenderjahres abgerufen würden, überwiegend positive, religiöse und feiertägliche Tweets in den Datensatz gelangen.

Der mittels dieses Stichprobenansatzes zusammengestellte Datensatz wird als Datenbasis für die Textanalyse genutzt. Ziel dieser Analyse ist es, eine inhaltliche Annotation vorzunehmen, bei welcher jedem Tweet genau ein Thema zugewiesen wird.

Solche inhaltliche Klassifikationen sind inhärent subjektiv und ihre Erstellung erfordert Verständnis der Lebensumstände der Autoren und deren Sprachnutzung, mit welcher diese ihre Perspektiven in Texten ausdrücken. Insbesondere im Kontext von Migration und interkultureller Mobilität sind diese Merkmale nicht immer ausreichend dokumentiert. Das erschwert die Annotation von großen Textdatensätzen insbesondere für Forschende ohne extensive Kenntnisse des geographischen und sprachlichen Kontexts der Autoren.

In dieser Studie wird daher ein Ansatz des domänenadaptiven maschinellen Lernens erprobt, bei welchem die Annotationen nicht händisch erstellt, sondern aus Daten eines Web-Community Forums abgeleitet werden, dessen Demographie der der Tweet-Autoren ähnelt.

Dass ein Transfer von Modellen von Foren auf Tweets mittels maschinellern Lernen möglich ist, wurde bereits von Fiallos und Jimenes (2019) gezeigt. Jedoch reduzieren Unterschiede zwischen Quelldomäne (Nairaland) und Zieldomäne (Twitter) die Genauigkeiten bei der Anwendung in der Zieldomäne. Daher wird in dieser Studie ein unüberwachter domänenadversarialer Ansatz (siehe 2.1.5) in das überwachte Lernverfahren integriert. Auf diese Weise sollen die Annotationen aus dem nigerianischen Webforum Nairaland gelernt und auf Tweets übertragen werden, ohne annotierte Twitter-Daten zu benötigen.

Die Klassen (Themen) ergeben sich aus der Forenstruktur Nairalands, welches flach hierarchisch in verschiedene thematische Bereiche wie Außenpolitik (*foreign affairs*), Sport (*sports*), Technologie (*technology*) und Religion (*religion*) gegliedert ist. Auch wenn die Nutzenden keinen direkten Einfluss auf diese Forenstruktur haben, entscheiden sie durch die

Wahl des Subforums, was das treffendste Thema für ihren Beitrag ist. Dies kann als eine Art der impliziten inhaltlichen Einordnung analog zu einer Annotation betrachtet werden. Es wird aufgrund der Konzepts des Forums und bisheriger Studien ('Nairaland Forum', o.D.; Basmus et al. 2019; Nelson et al. 2018) davon ausgegangen, dass die Nutzenden Nairalands größtenteils nigerianische Menschen oder Migrierte mit Beziehungen zu Nigeria sind, und deren Kommunikation größtenteils in Nigerianischem Englisch oder Varianten von Nigerianischem Pidgin English erfolgt. Damit stellen die Textdaten aus diesem Forum eine Quelle an inhaltlich annotierten Texten dar, welche als Trainingsdaten für ein Klassifizierungsmodell genutzt werden können.

Die Klassifizierung erfolgt mittels Feinabstimmung des vortrainierten tiefen Sprachmodells BERTweet (Nguyen et al. 2020), welches auf einem großen Twitter Datensatz vortrainiert wurde, und lediglich um ein Klassifizierungs-Layer für die Themen und den Domänenklassifizierer (Ganin et al. 2016) ergänzt wird.

Die Evaluation des Klassifizierungsmodells erfolgt aufgrund der Abwesenheit gelabelter Daten in der Zieldomäne anhand eines Ensembles von drei Kriterien, welche für jedes Thema separat berechnet werden können. (1) Trainingsgenauigkeit in der Quelldomäne, basierend auf dem F1-Score welcher auf einer zurückgehaltenen Partition des Nairaland Datensatzes berechnet wird. Diese bedingt zwar keine ähnliche Genauigkeit in der Zieldomäne, wird aber als Voraussetzung dafür angesehen. (2) Ähnliches Auftreten von Neologismen in Ziel- und Quelldomäne. In speziell dafür akquirierten Nairaland- und Twitter-Datensätzen aus einer späteren Periode (2021-2022) werden Neologismen identifiziert, welche in dem für die Analyse genutzten Datensatz nicht auftreten. Das Auftreten dieser Worte über Themen hinweg wird als Feature-Vektor genutzt, um die Qualität des Domänentransfers für jedes Thema zu quantifizieren. Zum Beispiel: Tritt der neuartige Begriff Covid-19 sowohl in Quell- als auch in Zieldomäne hauptsächlich im Thema („Gesundheit“) auf, unterstützt dies die Plausibilität des Themas „Gesundheit“. (3) Wöchentliche Aktivitätsmuster, welche die meisten Themen aufweisen, wurden zwischen Quell- und Zieldomäne mittels des Pearson-Korrelationskoeffizienten verglichen. Die Annahme ist, dass bei einer erfolgreichen Domänenadaption auch die wöchentlichen Muster übereinstimmen.

Zusätzlich zu diesen quantitativen Plausibilitätskriterien werden die klassifizierten Themen, für welche die größten Unterschiede zwischen Mobilitätsklassen gefunden wurden, qualitativ anhand einer Stichprobe untersucht und codiert, um eine feinere Klassifikation der Themen und damit ein besseres Verständnis derselben zu erhalten.

3.4. Studie B2: Analyse der Verbreitung von Sprachmerkmalen nach Mobilitätscharakteristika der Autoren am Beispiel westafrikanischer Twitter-Nutzender

Menschliche Migrationsentscheidungen und Sprache beeinflussen einander. Aus diesem Grund befasst sich Studie B2 damit, Informationen über Mobilität und Informationen über Sprache zusammenzubringen, und testet hierfür, inwiefern in den kompilierten Twitter-Datensätzen auch migrationsrelevante Sprachaspekte, wie das Entstehen von Kontaktsprachen und Code-Switching, algorithmisch identifiziert werden können.

Ziel ist hierbei nicht eine sprachwissenschaftliche Schlussfolgerung, sondern die Erweiterung der Textanalysen in Richtung korpuslinguistischer und soziolinguistischer Methoden, welche, wie in der Einleitung dargelegt, perspektivisch die Brücke zwischen den empirischen Messungen der Dissertation und den Sozial-, Literatur- und Humanwissenschaften darstellen könnten.

Wie Studie B1 bereits zeigte, können für Social-Media-Nutzende mit ausreichend dichten Geotweet-Zeitreihen sowohl Mobilitätscharakteristika als auch Textinhalte algorithmisch abgeleitet werden. Informationen über die Sprache, in welcher die Texte erfasst sind, können abgeleitet werden. Tweets sind, wie in Kapitel 2.3.3 angemerkt, bereits mit Sprachcodes annotiert, welche ca. 60 Sprachen⁴⁸ abdecken.

Natürliche Sprache ist jedoch deutlich vielfältiger und mehrdimensionaler als eine Einteilung in Kategorien abbilden kann. So gibt es zum einen deutlich mehr als 60 Sprachen, und innerhalb von Sprachen typischerweise viele Varietäten wie Regiolekte und Soziolekte. Zum Anderen lassen sich verschiedene Aspekte von Sprache unterscheiden, wie Aussprache, Lexis, Grammatik und Pragmatik (Bieswanger und Becker 2021). Sozio-historische Faktoren wirken auf diese unterschiedlichen Aspekte von Sprache unterschiedlich ein. Die Entwicklung von Sprachvariation ist also nicht linear und Aspekte von Sprache werden in verschiedenem Maß durch Migration beeinflusst. Im Migrationskontext bestünde daher ein erheblicher Mehrwert durch die Erkennung sprachlicher Variation. In dieser Studie soll exemplarisch eine lokale Nutzung des Englischen in Nigeria (*LEN*)⁴⁹ mittels lexikaler Merkmale erfasst, und deren Häufigkeit in den Tweets westafrikanischer Twitter-Nutzenden quantifiziert werden. Die Datengrundlage hierfür liefert der in Studie B1 genutzte Ansatz der Trajektorienanalyse von Twitter-Daten, welche wieder einen Analysezeitraum von vier Kalenderjahren zwischen dem 01. Januar 2015 und dem 01. Januar 2019 abdecken. In dieser Studie B2 wird die Analyse jedoch von Nigeria auf mehrere der benachbarten Länder Westafrikas ausgeweitet: Ghana, Burkina Faso, Benin, Togo. Hierdurch wird ermöglicht, die Verbreitung der Sprachmerkmale auch über Landesgrenzen hinweg zu analysieren, und in Bezug zu Mobilitätscharakteristika zu setzen.

Zur Erkennung der Sprachmerkmale wird eine von Honkanen und Müller (2021) genutzte Methodik in leicht abgewandelter Form genutzt. Hierbei werden lokale lexikalische Merkmale in einem abstimmungsbasierten Auswahlverfahren innerhalb mehrerer Web-Wörterbücher identifiziert und durch Experten geprüft und ergänzt. Die Auswahl der Merkmale erhebt keinen Anspruch, lokale lexikalische Einflüsse im Englischen vollständig abzubilden. Für die Ziele der Studie ist es jedoch ausreichend, deren Verteilung zu messen. Die Nutzung von freiwillig beigetragenen Ressourcen erlaubt hier die Einbindung der Web-Community in den Prozess, und damit eine Einbringung von Know-how aus dem die Tweets generierenden sozialen Milieu Nigerianischer Web-Nutzender, wie es auch durch die Nutzung von Nairaland-Daten in Studie B1 angestrebt wurde.

Das Vorkommen der LEN-Merkmale wird algorithmisch detektiert und als Ratio LEN-Index

⁴⁸ <https://github.com/igorbrigadir/twitter-advanced-search?tab=readme-ov-file#supported-languages>

⁴⁹ Die hier zu erfassende lokale Form LEN ist nicht identisch mit nigerianischem Pidgin (NigPE) oder nigerianischem Englisch (NigE), welche gut dokumentierte Sprachvarietäten darstellen. Demgegenüber ist die hier analysierte LEN ein offenes Konzept, welches sich teilweise mit NigPE und NigE überlappt, aber nicht als eine Sprache gesehen werden soll.

(LI) quantifiziert. Hierbei werden Begriffe, welche außer in Nigerianischem Englisch auch in anderen Sprachen vorkommen, als mehrdeutig betrachtet, und nur dann als LEN gewertet, wenn sie innerhalb eines Suchfensters um eindeutige LEN-Textelemente auftauchen. Der LEN-Index wird in Relation zum Ursprungsort der Tweets, dem nutzerspezifischen Hauptaufenthaltort und der nutzerspezifischen Mobilität gesetzt.

Um **Nutzerspezifische Mobilität** abzuleiten, werden die obersten räumlichen Einheiten der jeweiligen Länder genutzt. In Nigeria sind dies beispielsweise die Bundesstaaten (engl. States) und das Federal Capital Territory der Hauptstadt Abuja.

Nutzerspezifische Mobilität soll in folgende Klassen unterschieden werden:

Stationäre, welche sich über den Analysezeitraum in Nigeria am selben Ort aufhalten;

Mobile, welche sich in Nigeria an verschiedenen Orten aufgehalten haben;

Interregionale, welche sich in Nigeria und an anderen Orten der Region aufhalten, und

Emigrierte, welche sich zum Teil in Nigeria, zum Großteil jedoch in anderen Ländern aufhalten.

Der **Nutzerspezifische Hauptaufenthaltort** hingegen wird auf Länderebene berechnet. Hierbei ist zu erwarten, dass aufgrund der auf Westafrika fokussierten Datenakquise für Länder außerhalb Westafrikas vorwiegend Migrierte aus Westafrika erfasst werden (beispielsweise nigerianische Emigrierte in den Vereinigten Staaten), während Länder innerhalb Westafrikas höhere Anteile von lokalen Nutzenden aufweisen (beispielsweise Ghanaer und Ghanaerinnen in Ghana). Der Anspruch der Studie ist damit, die Verwendung nigerianischer Sprachmerkmale durch Nutzende mit westafrikanischem Hintergrund zu erfassen, und nicht die Verbreitung dieser Merkmale im Allgemeinen.

Um als zusätzliches Element Kategorien der semantischen Inhalte in die Analyse mit einzubringen, wird zudem ein weiteres Mal die unüberwachte und weit-verbreitete Topic-Modeling-Methode BERTopic (Grootendorst 2022) genutzt. Durch die Kombination der Sprachmerkmale und Mobilitätscharakteristika mit dieser zusätzlichen inhaltlichen Kategorie kann das multidimensionale Potenzial ausgelotet werden, welches sich bei der gemeinsamen räumlichen und inhaltlichen Analyse von GSM ergibt.

Die Wahl bestimmter lexikalischer Merkmale stellt nur eine Art von Sprachvariation dar; welche sich jedoch mit dem beschriebenen Ansatz algorithmisch erfassen und skalieren lässt. Um zu prüfen, inwiefern diese gemessenen lexikalischen Sprachmerkmale die praktische Nutzung anderer Sprachmerkmale approximieren können, wird eine Stichprobe an Tweets manuell analysiert. Durch die Kodierung pragmatischer sowie morphosyntaktischer Merkmale sollen Potenziale für weitere Entwicklungen in der sprachwissenschaftlichen Auswertung von GSM aufgezeigt werden.

3.5. Studie C1: Analyse der Räumlichkeit von Themen in englischen Webtextdaten

Die Strukturierung von Textinhalten in semantisch ähnliche Themen ist eine Kernkomponente der Dissertation und ein häufig eingesetztes Mittel in der Prozessierung von Texten. Es steht zu erwarten, dass Themen bei der Ableitung von Migrationsfaktoren aus räumlichen GSM eine wichtige Rolle spielen werden. Aus diesem Grund muss untersucht werden, inwiefern sich Themen und der Raumbezug von Texten bedingen.

Diesbezüglich ist geographische Räumlichkeit (engl. *Geospatiality*) das zentrale, aber bisher nicht ausreichend erforschte Bindeglied zwischen Raum und Textinhalt (siehe 2.2.2). In dieser Dissertation wird mit thematischer *Geospatiality* eine hypothetische Eigenschaft von Themen bezeichnet, welche sich innerhalb von GSM-Daten durch das Aufkommen von geographisch verortbaren Georeferenzen äußert.

Nach diesem Konzept besitzen Themen diese Eigenschaft in unterschiedlich starker Ausprägung. So hätten Tweets über Sport eine höhere Wahrscheinlichkeit, mit Geotags annotiert zu sein, als Tweets über Mathematik, und Forumsbeiträge über Reisen würden mehr Orte erwähnen als Forumsbeiträge über Technologie.

Die Analyse von GSM-Daten mittels Themen, die ein wichtiges Mittel in dieser Dissertation ist, wäre von der Existenz eines thematischen Räumlichkeitseffekts stark beeinflusst. Themen mit stark ausgeprägter Räumlichkeit würden in georeferenzierten Datensätzen, also beispielsweise Datensätze, welche nach bestimmten Orten gefiltert werden, einen höheren Anteil ausmachen als in den ungefilterten Datensätzen. Damit könnten Verzerrungen der thematischen Profile entstehen.

Als Beispiel ist eine Bibliothek vorstellbar, welche von vielen Menschen besucht wird, die sich dort mit Literatur auseinandersetzen und auf Social Media Erfahrungen und Inhalte wie Zitate oder Gedanken über Bücher teilen, ohne jedoch Bezug auf die Bibliothek als Ort zu nehmen. Durch die Linse der GSM würde diese Bibliothek im Vergleich zu Touristenattraktionen als ein geradezu verlassener Ort erscheinen, obwohl sie tatsächlich ein Hotspot von Aktivität wäre, welche lediglich keinen starken Raumbezug aufweist.

Verschiedene Konzepte von Räumlichkeit sind vorstellbar, welche sich dadurch unterscheiden, was sie als valide Ortserwähnungen betrachten. Dem Rahmen dieser Dissertation als geographisch-methodisches Werk entsprechend wird hier eine geographisch-methodische Definition von *Geospatiality* gewählt, welche sich auf Erwähnungen von Orten bezieht, welche sich mit aktuellen algorithmischen Methoden des Geoparsings (siehe 2.1.4) auf der Erde verorten lassen. Sie beschränkt sich also auf bekannte und registrierte Toponyme, welche in Ortsverzeichnissen aufgeführt sind. Erwähnungen von Demonymen, Metonymen und fiktiven Orten sind in diesem geographisch-methodischen Konzept damit keine validen Ortserwähnungen, auch wenn sie unter anderen Räumlichkeitskonzepten durchaus als solche gesehen werden können. Beispielsweise könnten Ortserwähnungen auch syntaktisch (nach Wortart) definiert werden.

In Studie C1 soll thematische *Geospatiality* statistisch messbar gemacht werden, indem in einem großen Korpus an Webtextdaten (1) die Häufigkeit von Ortserwähnungen mittels Geoparsing und (2) die Themen der Texte quantifiziert und in Relation gesetzt werden sollen. Es wird davon ausgegangen, dass sowohl bei Geoparsing als auch bei Geotagging Räumlichkeitseffekte auftreten, diese bei Geoparsing jedoch stärker ausgeprägt sind. Aufgrund dieser Vermutung, welche in Studie C2 untersucht wird, soll für die Untersuchung eines Räumlichkeitseffektes über Datensätze hinweg zunächst ein Geoparsing-Ansatz genutzt werden, welcher auf jede Art von Webtextdaten angewandt werden kann, nicht nur auf Texte von Plattformen mit Geotagging-Funktionen.

Aufgrund eines möglichen Einflusses der Textsorte auf Räumlichkeitseffekte, analysiert die Studie verschiedene Textdatensätze unterschiedlicher Typen, welche bereits in anderen Studien (zum Teil in dieser Dissertation) zur wissenschaftlichen Anwendung kamen. Obwohl auch diese kein repräsentatives Set von allen Webdaten abbilden, erhöht die Nutzung

verschiedener Textsorten auch die Robustheit und Generalisierbarkeit der Analyse. Die primäre Sprache dieser Datensätze ist Englisch, welches die häufigste Sprache im Internet ist (Common Crawl 2024). Es werden jedoch abgesehen von der Auswahl der Datensätze keine Schritte unternommen, um Texte anderer Sprachen auszusortieren. Für jeden Datensatz werden Autor der Texte, Zeitpunkt der Veröffentlichung, und Textlänge als zusätzliche Variablen extrahiert.

Die Analyse von thematischen Räumlichkeitseffekten innerhalb der so zusammengestellten Textdatensammlung geschieht wie folgt (Abbildung 13). Zunächst wird identifiziert, ob sie eine verortbare Ortserwähnung beinhalten. Zweitens werden alle Texte einem Thema zugewiesen. Bei ersterer Aufgabe wird Geoparsing angewendet.

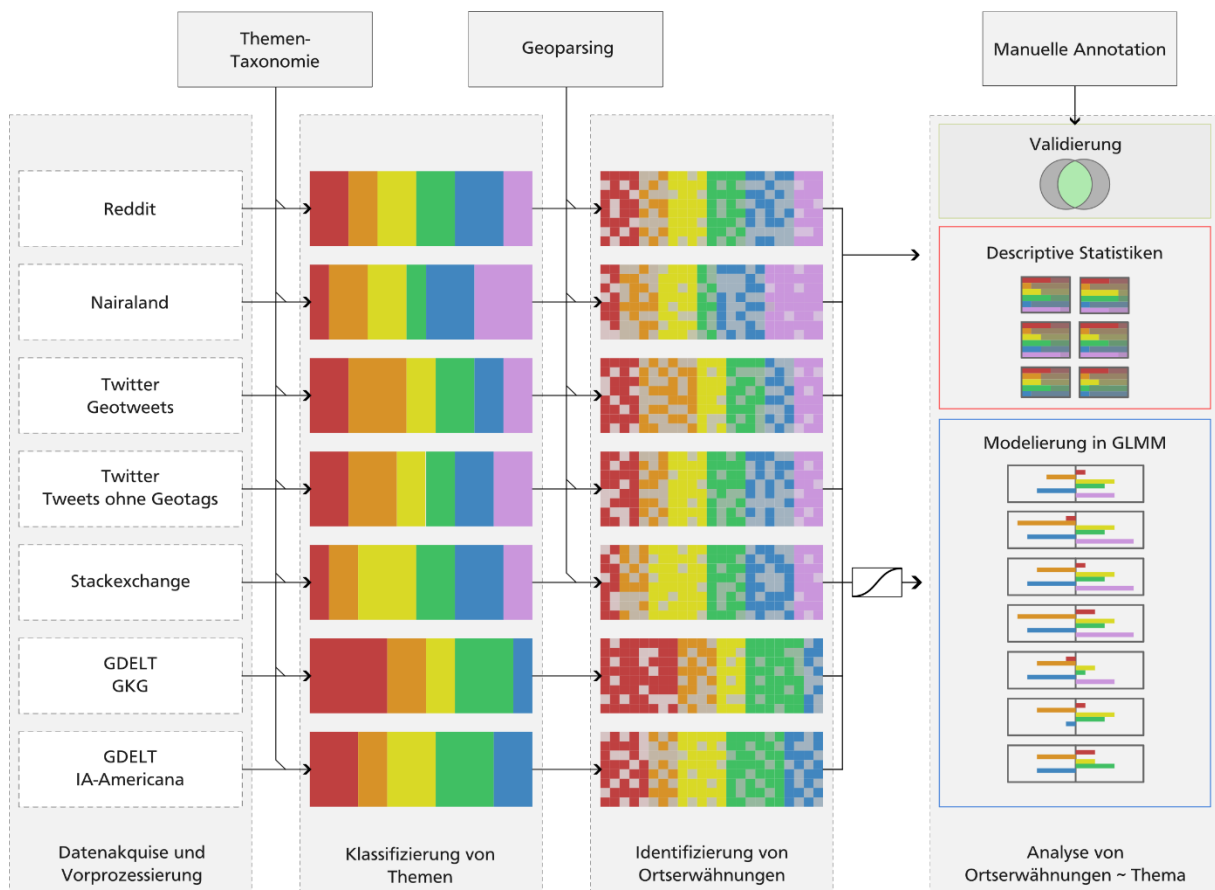


Abbildung 13 Arbeitsablauf von Studie C1

Geoparsing besteht aus zwei grundsätzlichen Schritten: (1) Der Erkennung von Ortserwähnungen und (2) deren Verortung. Um die Robustheit der Geoparsings zu erhöhen, wird zur Erkennung von Ortserwähnungen ein Ensemble aus vier wissenschaftlich publizierten NER-Modellen eingesetzt. Die durch diese im Text erkannten Ortserwähnungen werden dann mittels der API des Geonames-Dienstes verortet. Für die Analyse werden nur die Ortserwähnungen als valide betrachtet, für welche auch der zweite Schritt erfolgreich ist, da diese Orte mit Koordinaten in der praktischen geographischen Nutzung (beispielsweise im Migrationskontext) in einer weitaus größeren Bandbreite an räumlichen Analysen ausgewertet werden könnten.

Datensätze der Nachrichtendatenbank GDELT müssen hierbei gesondert behandelt werden. Da diese in den Metadaten bereits mit Ortserkennungen annotiert sind, wird kein Geoparsing

betrieben, da dies auch in der praktischen Nutzung dieser Datenbank nicht notwendig wäre. Auch diese Entscheidung folgt dem Anspruch der Dissertation, die anwendungsbezogene Realität der Nutzung von GSM abzubilden und kein theoretisches Konstrukt.

Neben der Geoinformation werden auch alle Posts einem Thema zugewiesen. Hierfür wird zunächst mittels eines induktiven Ansatzes eine Thementaxonomie entwickelt, welche an die Inhalte anknüpft und sie bestmöglich abdeckt. Die Zuweisung von Inhalten erfolgt basierend auf bestehenden Strukturen der genutzten Datensätze und baut auf ihnen auf. So besteht Stackexchange aus diversen Seiten für *Travel* oder *Astronomy*, und jeder Kommentar ist eindeutig einer dieser Seiten zugewiesen. Diese Seiten werden durch den Autor als Ganzes einem der Themen zugewiesen (beispielsweise *Travel* zu *Reisen*, *Tourismus* und *Migration*). Daraufhin wird diese Zuordnung von drei weiteren Forschenden auf Plausibilität geprüft und basierend auf Übereinstimmung und Diskussion angepasst.

Nachdem alle Texte einem Thema zugewiesen sind und identifiziert ist, ob sie verortbare Ortserwähnungen enthalten, wird der Zusammenhang zwischen der Räumlichkeit von Texten und deren Themen durch (1) deskriptive Statistiken und (2) Modellierung untersucht. Abhängige Variable ist dabei nicht die Häufigkeit von Ortserwähnungen, sondern die Häufigkeit von Texten (innerhalb von Datensätzen) welche mindestens eine verortbare Ortserwähnung enthalten. Dies wird als $Frac_{geo}$ innerhalb jedes Themas und Datensatzes berechnet, und über Datensätze, Zeit, Textlänge und Autoren hinweg verglichen. Dieser deskriptive Ansatz ermöglicht einen Einblick in die tatsächlich in den Daten beobachteten Unterschiede.

$$Frac_{geo} = \frac{\text{Anzahl der Tweets mit mindestens einer textualen Ortserwähnung}}{\text{Anzahl der Tweets}}$$

Da hierbei beobachtete Unterschiede durch verschiedene Faktoren bedingt sind, können sie nicht klar auf den Effekt von Themen selbst zurückgeführt werden. Um diesen herauszuarbeiten, wird in einem zweiten Schritt ein Modellierungsansatz genutzt, welcher für die messbaren anderen Einflussfaktoren wie Zeit, Autor, und Textlänge kompensiert, und den tatsächlichen Einfluss des Themas greifbar macht. Der Effekt von Themen auf die Gegenwart von Ortserwähnungen wird für jeden Datensatz separat modelliert und dann verglichen. Ein gemischtes Modell wird genutzt, um zeit- und autorspezifische Effekte als zufällige Effekte (engl. *random effects*) zu kontrollieren. Die themenspezifischen Koeffizienten können dann als Proxy-Indikator für die Geospatiality des jeweiligen Themas in dem jeweiligen Datensatz interpretiert werden.

Von Interesse ist in der Auswertung dieser Koeffizienten sowohl die Existenz von Unterschieden zwischen Themen allgemein als auch die Konsistenz von themenspezifischen Effekten über Datensätze hinweg. Diese Konsistenz, als Rangkorrelation quantifizierbar, gibt Aufschluss darüber, inwiefern thematische Geospatiality von Datentyp definiert ist oder eine konsistente, möglicherweise universelle Eigenschaft von Themen vermutet werden kann.

3.6. Studie C2: Differenzierung der Räumlichkeit von Themen zwischen Geotag- und Geoparsing-basierten Ansätzen

Während Studie C1 ergründen soll, inwiefern themenspezifische Räumlichkeitseffekte existieren, soll Studie C2 untersuchen, inwiefern diese davon abhängen, welche Art der Geoinformation genutzt wird. Konkret sollen in dieser Hinsicht Geoparsing und Geotagging-Ansätze verglichen werden.

Als Daten hierfür dienen die accountspezifischen Tweet-Zeitreihen nigerianischer Twitter-Accounts, welche bereits in Studien 3,4, und 5 genutzt werden. Dieser Datensatz enthält sowohl Geotweets als auch Tweets ohne Geotags. Auf beide Typen wird das in Studie C1 beschriebene Geoparsing angewandt. Hierdurch ergeben sich vier Typen von Tweets, je nachdem, ob sie Georeferenzen (1) in Form von Geotags enthalten, (2) in Form verortbarer Ortserwähnungen, (3) in beiden Formen oder (4) in keiner Form.

Die Datenakquise und Vorverarbeitung der Daten entspricht der in Studie C1, inklusive der angewandten Themen-Taxonomie. Die Auswertung erfolgt anders als bei Studie C1 nicht den Vergleich zwischen Datensätzen, sondern durch den Vergleich zweier Ansätze bei Anwendung auf denselben Datensatz.

Auch hier wird mit einer deskriptiven Analyse begonnen, welche zunächst die Häufigkeit von Tweets mit Ortserwähnungen innerhalb von Themen zwischen beiden Datensätzen vergleicht. Wenn die Erwähnung von Orten unabhängig von der Nutzung von Geotags erfolgt, sollten die Häufigkeiten in beiden Datensätzen identisch sein. Themen, bei welchen dies nicht der Fall ist, deuten eine pragmatische Beziehung an, in welcher die Nutzung von Geotags mit einem höheren räumlichen Bezug, anderen Unterthemen oder anderen Konversationsformen und -absichten einhergeht.

Für die zweite Analyse wird, wie in Studie C1, der Zusammenhang zwischen Themen und Räumlichkeit in zwei separaten Modellen modelliert, wobei in einem Fall die Präsenz von Geotags als abhängige Variable genutzt wird, und in einem zweiten Fall die Präsenz von verortbaren Ortserwähnungen. Der Vergleich der themenspezifischen Koeffizienten zwischen den beiden Ansätzen erlaubt eine Einschätzung, inwiefern die thematische Geospatiality von dem genutzten Geoinformationstyp abhängt.

Auch wenn die Stichprobe nur eine sehr kleine Bandbreite des Volumens und der Vielfalt von Webtextdaten abdeckt, ermöglicht sie durch ihre Abdeckung von Tweets mit und ohne Geotags aus einem klar definierten zeitlichen und geographischen Umfeld diese erste Analyse eines wenig erforschten Phänomens, und kann Hinweise darauf liefern, inwiefern die Art der genutzten Geoinformation die thematische Zusammensetzung des damit akquirierten Daten beeinflusst. Zusätzlich wird ein Einblick in die innere Komplexität der thematischen Räumlichkeit ermöglicht, und Beobachtungen für die Erforschung von Beziehungen zwischen Geotag-Nutzung und Ortserwähnungen beigetragen.

4. Untersuchungsgebiet und Daten

4.1. Untersuchungsregion Afrika

Die Studien der Dissertation werden für Untersuchungsgebiete auf dem Afrikanischen Kontinent durchgeführt. Die Wahl dieser Studienregion Afrika ergibt sich aus drei Überlegungen.

1. Afrika ist von hoher Relevanz für internationale Migration. So titeln Foster und Hall (2025) beispielsweise einen Kommentar mit: „*Die demographische Zukunft der Migration ist Afrikanisch*“.
2. Umgekehrt ist menschliche Migration auch von hoher Relevanz für viele afrikanische Länder, in welchen eine hohe ökonomische Bedeutung von Migration und lange Traditionen von (besonders intra-regionalen) Migrationsmustern bestehen, beispielsweise durch Rücksendungen (Iheke 2012; Fayissa und Nsiah 2010). Angesichts der hohen Bevölkerungsentwicklungen und der heterogenen Entwicklung afrikanischer Staaten kann erwartet werden, dass Migrationsbewegungen auch in Zukunft eine große Bedeutung für afrikanische Staaten haben werden.
3. Gleichzeitig ist die Darstellung von afrikanischer Migration in Literatur und öffentlichem Diskurs oft übersimplifiziert und mit Stereotypen behaftet (Flahaux und De Haas 2016). Es ist tatsächlich herausfordernd, die Komplexität von Migrationsprozessen in Afrika adäquat zu repräsentieren, denn Migrationsmuster und -Trends sind in stetigem Wandel. Gleichzeitig bestehen auch in Afrika Inkonsistenzen in der Datenlage zur Migration (Flahaux und De Haas 2016). Aufgrund dieser Komplexität und gleichzeitig schlechten Datenlage ist es eine Region, deren Untersuchung besonders von empirischen und grenzübergreifenden Ansätzen profitieren kann, wie sie in dieser Dissertation angestrebt werden.

Eine Unterteilung Afrikas in kulturelle Zonen ist herausfordernd. In dieser Dissertation werden für geographische Gruppierungen von Ländern die von den Vereinten Nationen für statistische Zwecke definierten Georegionen genutzt (United Nations Statistical Office 1982). Diese können und sollen nicht alle historischen, geographischen und kulturellen Grenzen abbilden, sondern lediglich als eine gemeinhin akzeptierte statistische Basis zur räumlichen Gruppierung von Ländern dienen.

Die räumlichen Studien der Dissertation werden nicht flächendeckend Afrika-weit durchgeführt. Sie fokussieren einerseits auf eine Auswahl von urbanen Testgebieten in Afrikanischen Städten (Studien A1 und A2) und andererseits die Region des Landes Nigeria und dessen Nachbarländer (Studien B1 und B2).

4.2. Urbane Testgebiete (UTG)

Da die flächenhafte Akquise von GSD für alle urbanen Flächen des Kontinentes die zur Verfügung stehenden Ressourcen überschreitet, werden die Studien für eine Auswahl von räumlich verteilten Urbanen Testgebieten (UTG) durchgeführt.

Als Standorte dieser UTG werden dabei die Flächen gewählt, in welchen in den vergangenen Jahrzehnten das stärkste Siedlungsflächenwachstum gemessen wurde. Dieses Siedlungsflächenwachstum wird aus fernerkundlichen Produkten abgeleitet.

Um statistische Vergleichbarkeit zwischen UTGs zu gewährleisten, werden mindestens 100 UTGs mit identischer Gesamtfläche definiert. Als Breite der UTGs wird 10km gewählt, was aufgrund von ersten Schätzungen mit einer durch das Datenbudget realisierbaren Menge von Twitter-Daten (4.5.1) einhergeht.

Um zu gewährleisten, dass alle geographischen Regionen Afrikas abgedeckt sind, wird eine stratifizierte Stichprob aller infrage kommenden UTG-Kandidaten generiert. Der Algorithmus wird, mit anderen Worten, für jede der fünf durch die UN definierten Georegionen separat durchgeführt. Innerhalb dieser Regionen ist das oberste Auswahlkriterium die Maximierung der innerhalb des Radius liegenden, in jüngerer Zeit neu entstandenen bebauten Fläche.

Diesen Kriterien folgend wird ein Algorithmus zur Identifikation der UTGs auf der Cloud-Computing Plattform Google Earth Engine implementiert (Abbildung 14). Zunächst wird die neu hinzugekommene Siedlungsfläche (Wachstumsfläche) als Raster mit einer Auflösung von 100m afrikaweit erfasst. Hierfür wurden die vom Deutschen Zentrum für Luft- und Raumfahrt bereitgestellten World Settlement Footprint Produkte (Marconcini et al. 2020, 2021) genutzt. In einer fokalen Operation mit einem 10km*10km großen Fenster wird für jeden Pixel die ihn umgebende Wachstumsfläche summiert (Abbildung 14c). Aus dem sich ergebenden Raster, welches die lokale Wachstumsintensität interpretiert werden kann, werden lokale Maxima als Kandidaten für die Mittelpunkte der UTGs gewählt. Diese Mittelpunkte liegen aufgrund der räumlichen Autokorrelation (Tobler 1970) oft sehr nahe beieinander, was zu überlappenden UTGs führen würde. Daher wird die Methode der Nicht-Maxima-Unterdrückung (engl. *non-maximum Suppression*) angewendet, welche unter benachbarten Kandidaten jene mit dem niedrigsten Wachstum eliminiert, bis alle Kandidaten mindestens R km voneinander entfernt sind (Abbildung 14f).

Innerhalb der fünf Regionen Afrikas wurden dann jeweils die 20 Kandidaten mit dem höchsten Wachstum als Zentren für UTGs ausgewählt. Da für den Parameter R kein natürlich geeigneter Wert besteht, und er in Experimenten Einfluss auf die Verteilung der UTG zeigte, wird der Algorithmus mehrfach angewendet, mit Werten für R von 5 km, 20 km, und 80 km. In jedem Fall wurden hierdurch 20 Kandidaten identifiziert, welche sich größtenteils zwischen Durchführungen überlappen. Somit wurden insgesamt 135 UTGs identifiziert, welche alle 5 Regionen Afrikas repräsentieren, innerhalb diesen räumlich dispers sind, jedoch gleichzeitig die wachstumsstarken Gebiete abdecken, welche für die urbane Studien im Migrationskontext (Studie A1 und Studie A2) von hoher Relevanz sind.

Die UTGs, welche als Kreis mit Radius 5 km oder als Quadrat mit Kantenlänge 10 km (100 km²) realisiert werden können, sind konzeptionell in keiner Form am historischen oder administrativen Zentrum von Stadtregionen orientiert und liegen oft am Rand von oder zwischen traditionellen Siedlungsgebieten. Eine klare Zuordnung von UTG zu Stadt ist daher weder möglich noch nötig. Stattdessen wird jedes UTG über eine ID identifiziert und einer Bezeichnung zugewiesen, welche entweder die Lage zu einem eindeutig identifizierbaren Siedlungsgebiet ausdrückt (e.g. Nordwest-Port Harcourt) oder den geographisch nächsten Eintrag aus der Geonames Datenbank ('GeoNames', o.D.).

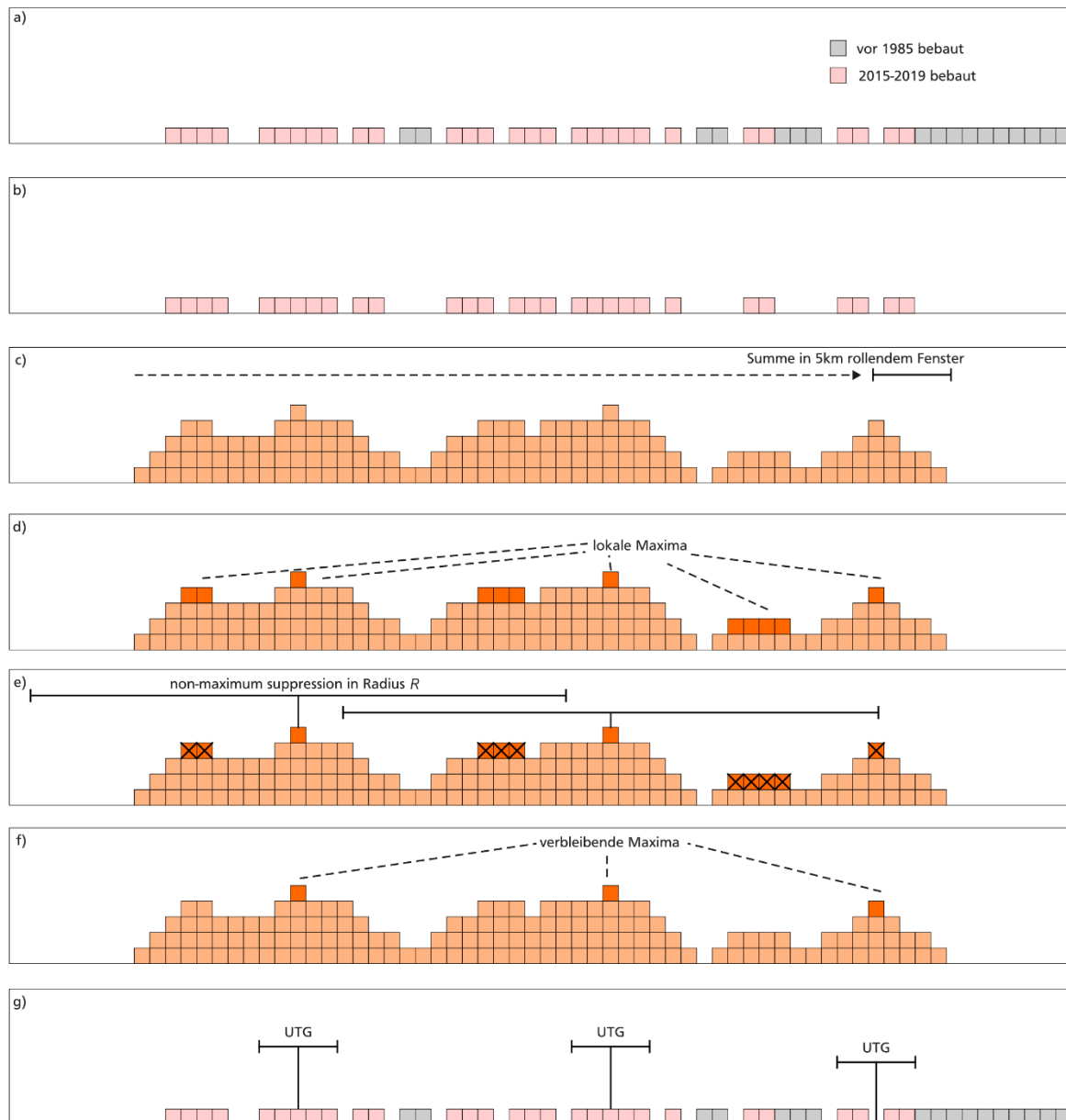


Abbildung 14: Schematische Darstellung der Selektion der UTG (-Zentren). Was hier als Querschnitt durch eine polyzentrische Agglomeration dargestellt ist, wurde tatsächlich in einem zweidimensionalen Raster durchgeführt.

4.3. Fokusregion Nigeria

Westafrika ist eine der Weltregionen mit dem höchsten Potenzial an regionaler und städtischer Migration und städtischem Bevölkerungswachstum. Die meisten intraregionalen Migrierten Afrikas leben und bleiben in Westafrika, welches viele Bevölkerungsgruppen mit langen Traditionen von Migration beheimatet (Schürmann et al. 2022). In Westafrika findet sich eine junge und dynamische Bevölkerung (McAuliffe 2024), große Städte und ein weiterhin hohes Urbanisierungspotenzial. Dabei ist besonders Nigeria hervorzuheben, für welches der nach China und Indien global größte Zuwachs an Stadtbewohnern erwartet wird (UN DESA 2019).

Gleichzeitig sieht sich Westafrika auch multiplen Herausforderungen ausgesetzt, welche im Zusammenhang mit Migration stehen. Zu diesen gehören beispielsweise Klimawandel, Konflikte, Bodenqualität und Mangel an finanziellem Kapital (Schürmann et al. 2022).

Das *Protocol on Free Movement of Persons, Residence and Establishment* (Protocol A/P.1/5/79) der ECOWAS⁵⁰-Staatengemeinschaft erlaubt visafreie Migration zwischen den teilnehmenden Ländern (Flahaux und De Haas 2016). Trotz Herausforderungen in der Implementierung und der gesellschaftlichen Akzeptanz von Migrierten gehören westafrikanische Bürger nach Statistiken als zu den weltweit mobilsten (Aniche 2022). Unter migrationswilligen Westafrikanern machen Menschen Nigerias den weitaus größten Teil aus (Kirwin und Anderson 2018).

Nigeria zeichnet sich durch die mit über 236 Millionen Menschen höchste Bevölkerung in Afrika aus (CIA 2025c), wodurch eine hohe Anzahl an potenziellen Social-Media Nutzenden zu erwarten ist. Nigeria ist kulturell einflussreich, intern kulturell divers und wie eine vorausgegangene Studie von Blanchard et al. (2020) fand, besteht erhebliche Mobilität unter den betrachteten nigerianischen Nutzenden von Mobiltelekommunikationsgeräten.

Das urbane System Nigerias weist eine gestreute Verteilung von signifikanten Bevölkerungszentren auf (CIA 2025c) und wird von Bloch et al. (2015) als polyzentrisch eingestuft. Zu den bedeutendsten Städten gehören die Küstenmetropole Lagos, die seit den 1970er Jahren als Planstadt entstandene Hauptstadt Abuja, sowie Kano, die größte Stadt im Norden Nigerias. Besonders im Norden des Landes um Kano und in den küstennahen Regionen haben sich stark urbanisierte Zonen herausgebildet (Seto et al. 2012). Geschichtlich, kulturell und funktional sind diese Siedlungsgebiete sehr unterschiedlich. Nigerianische Städte haben eine Vielzahl von Formen und Funktionen und lassen sich nicht eindeutig einer bestimmten urbanen Form zuordnen (Bloch et al. 2015).

Die Verbreitung von Internettechnologien ist in Nigeria im internationalen Vergleich nicht hoch, dennoch ist aufgrund der großen Bevölkerung des Landes mit einer hohen Anzahl an potenziellen Nutzenden zu rechnen. So nutzten 2015-2019 etwa 100 Millionen Menschen in Nigeria das Internet, was etwa der Hälfte der Bevölkerung entspricht (Adeleke 2021; Kemp 2015; 2019b). Auf Social Media aktiv ist ein Teil dieser Menschen, welcher um Januar 2015 auf 13,6 Millionen (7% der Landesbevölkerung) und im Januar 2019 auf 24 Millionen (12% der Gesamtbevölkerung) geschätzt wurde. (Kemp 2015; 2019b). Die Social-Media-Nutzenden Nigerias sind überwiegend jung (Kemp 2019b) und leben in Städten, wo Telekommunikation und Internetnutzung um ein Mehrfaches höher sind als in den ländlichen Regionen (Blanchard et al. 2020; Adeleke 2021). Viele dieser Menschen in Nigeria benutzen Social Media im Rahmen von kommerziellen Aktivitäten und als Nachrichtenquelle (Kemp 2025).

Während Nigeria die Fokusregion darstellt, muss bei Betrachtung der Internationalen Migration auch berücksichtigt werden, dass ganz Westafrika eine Region starker Mobilität ist, welche, wie Aniche (2022) formuliert, ethnische, sprachliche und kulturelle Grenzen überschreitet. Daher werden in Studie B2 auch die folgenden weiteren westafrikanischen Länder miteinbezogen: Benin, Burkina Faso, Ghana und Togo. Unter diesen ist Ghana, nach Nigeria, das zweitgrößte und zweitbevölkerungsreichste Land in der Untersuchungsregion. Die Internetpenetrierungsrate Ghanas wird auf 35% geschätzt, die Social-Media-Nutzung auf

⁵⁰ Economic Community of West African States (Deutsch: Westafrikanische Wirtschaftsgemeinschaft)

19% und damit deutlich höher als in Nigeria (Kemp 2019a). In Nigeria und Ghana ist Englisch die Amtssprache und wird von weiten Teilen der Bevölkerung gesprochen. Demgegenüber ist Französisch die Amtssprache in Togo, Benin und Burkina Faso (CIA 2025d; 2025a; 2025b), sowie auch in den östlicheren Nachbarländern Nigerias (Niger, Tschad und Kamerun), womit Nigeria von frankophonen Ländern umgeben ist (Omotuyi 2022).

Die meisten Menschen der Region sprechen als Muttersprache nicht Englisch oder Französisch, sondern eine der Sprachen lokalen Ursprunges, beispielsweise Yoruba, Igbo, Fon, Akan und Hausa. Die Sprachsituation Westafrikas ist dynamisch und in stetigem Wandel (Connell und Zeitlyn 2023). Durch den Kontakt zwischen Sprachen haben sich in vielen Regionen auch Kreolsprachen herausgebildet (Ofulue 2012). Pidgin-Englisch erfüllt insbesondere in Nigeria zunehmend die Funktion einer Verkehrssprache. Lokale Sprachmerkmale des Englischen in Nigeria sollen in Studie B2 genauer untersucht werden.

4.4. Fernerkundungsdaten

Für die flächenhafte Kartierung geografischer Regionen nehmen Fernerkundungsdaten (FED) eine zunehmend bedeutende Rolle ein. Eine wachsende Anzahl von satellitengestützten Systemen erfasst mit zunehmender Genauigkeit die Erdoberfläche. Bei dieser Form der Fernerkundung wird die Erdoberfläche regulär und regelmäßig erfasst, wobei der Erfassungsprozess in der Regel gut dokumentiert und nachvollziehbar ist (Albertz, 2001). Zivile Fernerkundungsdaten, wie die der Sentinel-Satelliten des Copernicus-Programms der Europäischen Union oder der Landsat-Satelliten der NASA, sind frei zugänglich. Für die Auswertung dieser Daten sind sowohl manuelle visuelle Bildinterpretation durch Fachexperten als auch die algorithmische Bildanalyse relevant (Albertz, 2001). Für viele Aspekte der Erdoberfläche stehen bereits algorithmisch abgeleitete Produkte oder etablierte Algorithmen zur freien Verfügung, was die Nutzung dieser Daten für eine breite Anwenderbasis ohne Expertenkenntnisse ermöglicht. In dieser Arbeit, welche sich mit dem urbanen Raum beschäftigt, werden diese existierenden fernerkundlichen Lösungen zur Messung und Charakterisierung des urbanen Raumes genutzt.

Konkret besteht ihre Anwendung in der Erfassung des Bebauungsjahres bebauter Flächen (4.4.2) der Geplantheitsintensität (4.4.3) und der Siedlungsstruktur (4.4.4). Aufgrund der zentralen Rolle dieser Daten werden sie im Folgenden im Detail vorgestellt.

4.4.1. World Settlement Footprint Suite

Die Kartierung der von Gebäuden bedeckten Fläche stellt eine wichtige Grundlage für zahlreiche Analysen dar, welche sich beispielsweise mit dem Einfluss des Menschen auf die Landoberfläche befassen. Für diesen Zweck wurden seit den 1980er Jahren zunehmend von Satelliten erfasste FED genutzt (Marconcini et al. 2020). Durch technische Fortschritte hat sich die Genauigkeit und räumliche Auflösung im Laufe der Jahre stetig verbessert. Während zwischen 2000 und 2010 noch Auflösungen von 500m-1km die Regel waren (Potere et al. 2009), erreichen heutige Produkte Auflösungen von 10m-30m (Esch et al. 2017; Marconcini et al. 2021).

Zahlreiche auf fernerkundlicher Basis erstellte Produkte kartieren den bebauten Raum und können sowohl im Vektorformat als auch im Rasterformat vorliegen. Im Vektorformat werden Gebäudegrundflächen (engl. *building footprints*) von den Unternehmen Google und Microsoft kostenfrei angeboten, erreichen dabei aktuell jedoch keine globale Abdeckung (Gonzales 2023; Sapena et al. 2024).

Im Rasterformat existieren Produkte welche die gesamte Siedlungsfläche des Globus als binäre Maske (bebaut oder unbebaut) zur Verfügung stellen. Zu diesen gehören der Global Urban Footprint mit einer Auflösung von 12m, welcher auf Radardaten der *TerraSAR-X/TanDEM-X* Mission beruht (Esch et al. 2017; Krieger et al. 2007). Das *WorldCover v200* von ist eine aus Sentinel-1 und Sentinel-2 abgeleitete Landbedeckung mit 10m Auflösung und 11 thematischen Klassen, wovon eine urbane Klasse der Siedlungsfläche entspricht (Zanaga et al. 2022). Das *Global Human Settlement* Produktes des *Joint Research Center der Europäischen Kommission* beinhaltet mit dem *GHS-BUILT-S R2023A* ein aus Microsoft Gebäudegrundflächendaten, Sentinel-2-Daten und Landsat-Daten abgeleitetes Produkt des Gebäudebedeckungsgrades mit 100m Auflösung, welches für 5-Jahres Intervalle zwischen 1975 und 2035 multitemporal verfügbar ist (Pesaresi und Politis 2023). Die Suite von *World Settlement Footprint* Produkten (*WSF*) (Marconcini et al. 2020; 2021) stellt eines der neuesten und hochauflösetesten Produkte dar.

Der *WSF2015* kartiert alle Siedlungsflächen im Jahr 2015 mit 10m räumlicher Auflösung, der *WSF2019* alle Siedlungsflächen gleichsam für das Jahr 2019. Hierfür wurden multitemporale optische Aufnahmen des Landsat-8 Satelliten und des RADAR Aufnahmen des Sentinel-1 Satelliten zusammen ausgewertet. Der *WSF-Evolution* kartiert zusätzlich das Bebauungsjahr rückwirkend bis 1985 auf jährlicher Basis mit 30m Auflösung, mittels Daten der Landsat-Satelliten (Marconcini et al. 2021). Zusammen erlauben die *WSF*-Produkte also eine hochaufgelöste Siedlungsflächenkartierung für die Stichjahre 2015 und 2019 sowie eine Abschätzung des Bebauungsjahres für zwischen 1985 und 2015 erbaute Flächen. Für die Zeit ab 2015 ist diese jährliche Differenzierung leider momentan nicht enthalten.

4.4.2. DIY-BU Maps

Diese Datenlücke kann durch die von Sapena et al. (2024) entwickelte Methode (*DIY-BU Mapping*) gefüllt werden. Diese Methode, welche als frei zugängliches Werkzeug in Google Earth Engine⁵¹ implementiert und frei zugänglich ist, erlaubt die Kartierung des Bebauungsjahres mit einer Auflösung von 10m für flexible Zeiträume ab 2015 und optimiert für lokale Anwendungen. Die auf maschinellem Lernen basierende Methode nutzt die bereits erwähnten Gebäudegrundflächendatensätze als Ausgangsdaten und kartiert dann retroaktiv jährlich auf Basis von Sentinel-1 und Sentinel-2 Daten mittels eines Random-Forest-Klassifizierers die Landoberfläche binär in bebaute und unbebaute Flächen.

Im Vergleich zu den globalen Produkten wird der Klassifizierers in diesem Ansatz für lokale geographische Kontexte trainiert und erreicht damit höhere Genauigkeiten welche insbesondere im globalen Süden eine deutliche Verbesserung gegenüber globalen Produkten darstellen (für Details, siehe Sapena et al. 2024).

In dieser Arbeit wird dieses Werkzeug angewendet, um die Datenlücke der *WSF*-Suite zu schließen und für den Zeitraum 1985-2020 das Bebauungsjahr innerhalb der UTG zu kartieren (Abbildung 15).

⁵¹ <https://earthengine.google.com/>

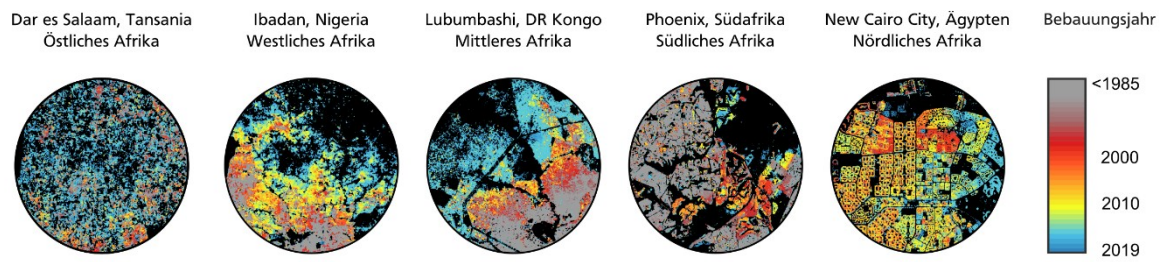


Abbildung 15: Durch Kombination von WSF-Evolution und DIY-BU Kartierungen erstellte Karten des Bebauungsjahres für fünf ausgewählte UTG.

4.4.3. Quantifizierung der Geplantheitsintensität

Oben genannte Produkte erlauben es, die Intensität und das Ausmaß und die globale Vielfalt des Städtewachstums zu erfassen (Taubenböck et al. 2020; 2024; 2025). Diese Vielfalt wird in erheblichem Maße durch die Umstände und Prozesse geprägt, welche die Bebauung der Stadt unmittelbar beeinflussen. Dies können kolonialistische Einflüsse sein (Şen 2024), oder demographische und wirtschaftliche Entwicklungen, sowie die Kultur der Stadtplanung (Taylor 2013). Um diese diversen Prozesse zusammenzufassen, entwickelten Debray et al. (2023) das Konzept der Geplantheitsintensität (engl. *Intensity of Plannedness*, IoP) welches die Verbindung zwischen diesen vielfältigen gesellschaftlichen Prozessen und der sich ergebenden Siedlungsstruktur greifbar macht. IoP fasst die historischen Prozesse in einen im aktuellen Zustand der Stadt interpretierbaren Planungsgrad zusammen, welcher anzeigt inwiefern die Siedlungsform das Resultat geplanter Prozesse ist. Dieser Planungsgrad lässt sich auf Basis aktueller Karten manuell ableiten und in eine Skala von fünf Graden einteilen (Abbildung 16), wobei der niedrigste Grad 1 einer durchgehend ungeplanten (spontanen) Siedlungsform entspricht, der höchste Grad 5 einer komplett geplanten Siedlung (Debray et al. 2023).

Dieser Geplantheitsintensitäts-Ontologie folgend, wird die IoP jedes UTG manuell erfasst. Als Datengrundlage hierfür dienen aktuelle Luft- und Satellitenbilder, OpenStreetMap-Daten und Google Street View Bilder. Durch Interpretation der Bilder wurde die Geplantheitsintensität jeder UTG in eine der fünf Grade von *extrem spontan* bis *komplett geplant* eingestuft. Dies erfolgte sowohl für die Geplantheitsintensität auf der Ebene von Straßennetzwerken (*IoPs*) und die Geplantheitsintensität auf der Ebene von Gebäudeblocks (*IoPs_B*). Zwei Experten führten diese Interpretation unabhängig voneinander durch. Daraufhin wurden die Interpretationen, insbesondere an Stellen, bei denen die Einschätzungen sich unterschieden, ausführlich zwischen den Experten und unter Zuhilfenahme von Webrecherchen diskutiert. Geplantheitsintensität wurde im Rahmen dieser Arbeit in Studie A1 genutzt.

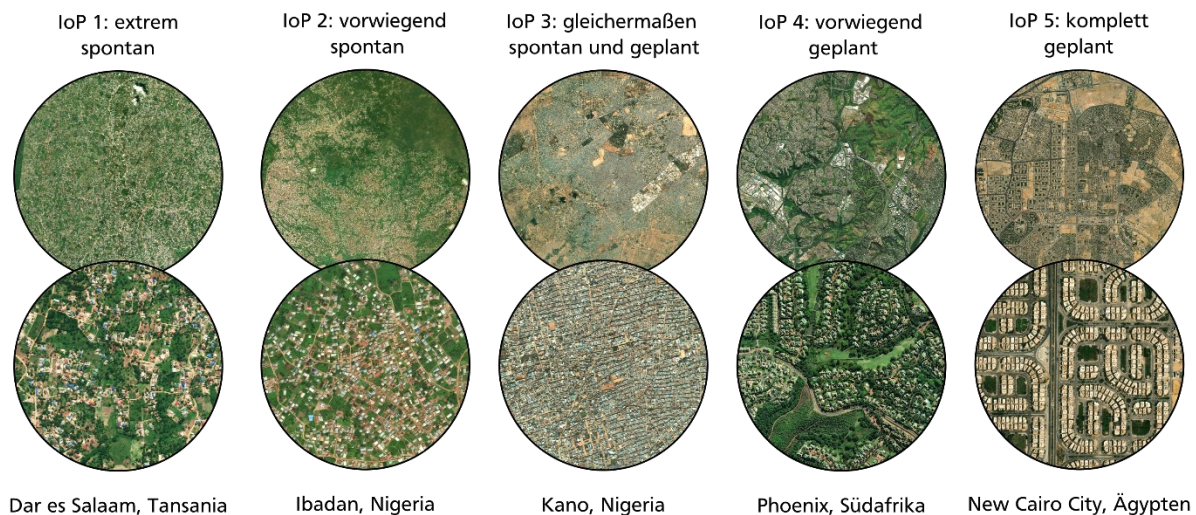


Abbildung 16: Beispiele für UTG verschiedener Geplantheitsintensität (IoP). Datenquelle: Microsoft (Bing Maps).

4.4.4. Räumliche Metriken

Räumliche Metriken stellen eine weitere Möglichkeit dar, die Struktur von Siedlungen zu erfassen und sie als Proxy für sozioökonomische Situation nutzbar zu machen, wenn diese aufgrund von Datenknappheit nicht direkt zu erfassen ist. Die Berechnung derartiger Metriken erfolgt auf Grundlage von zusammenhängenden Gruppen von Pixeln derselben Klasse (i.d.R. mit dem englischen Begriff *patches* bezeichnet siehe Blaschke, 2000). Die Form dieser *patches* und ihre Größe und Lage zueinander wird mittels der räumlichen Metriken quantifiziert und jeweils innerhalb des urbanen Testgebietes (UTG) aggregiert. Im Kontext von Siedlungen wären Beispiele für räumliche Metriken die Kompaktheit und Geschlossenheit der Siedlungsteile (entsprechend den *patches*), die durchschnittliche Größe der Siedlungsteile oder ihre Dispersion. Anders als die in fünf diskrete Grade unterteilte IoP stellen diese Metriken ein stetiges Maß dar. Sapena et al. (2020) zeigten, dass die ausgewählten Metriken die Siedlungsmorphologie ausreichend quantifizieren können, um sie als Proxy für sozioökonomische Indikatoren nutzbar zu machen.

Basierend auf den Studien von Altieri et al. (2019) und Sapena et al. (2020; 2021) wird eine Reihe von Metriken ausgewählt, um innerhalb der UTG die Siedlungsstruktur zu quantifizieren.

Die von Altieri et al. (2019) entwickelte räumliche Transinformation der Entropie (engl. *spatial partial information, SPI*) stellt hierbei ein Sonderfall dar, da sie für verschiedene Distanzklassen berechnet wurde, und inhärent zwischen 0 und 1 skaliert ist. Hierdurch erlaubt sie es, die Kompaktheit der Bebauung auf verschiedenen räumlichen Skalen zu quantifizieren (Altieri und Cocchi 2022) und zu berücksichtigen, dass Kompaktheit auf Gebäudeebene sich von Kompaktheit auf Blockebene unterscheiden kann. Die SPI wurde mittels des R Pakets *SpatEntropy* (Altieri et al. 2021) für folgende Distanzklassen berechnet: <100m; 100-500m; 500-2000m; und >2000m.

Die räumlichen Metriken wurden auf Basis der DIY-BU Daten von 2019 für die UTG berechnet, also für das Ende des in Studie A1 genutzten Analysezeitraums. Hierbei wurde das genutzte Bebauungsraster um 5 km über den Rand des UTG hinaus erweitert, wodurch Randeffekte vermieden werden, welche die Metriken beeinflussen könnten.

Tabelle 1 beschreibt die genutzten Metriken. Diese morphologischen räumlichen Metriken wurden im Rahmen dieser Arbeit in Studie A1 als unabhängige Variablen in einer multiplen linearen Regression genutzt.

Tabelle 1: Liste der räumlichen Metriken

Abkürzung	Metrik	Beschreibung
MPS	Mittlere Patchgröße (engl. mean patch size)	Mittlere Größe der bebauten Patches.
SI	Formindex (engl. shape index)	Quantifiziert die Abweichung der Formkomplexität eines bebauten Patches von einem Quadrat derselben Fläche, gemittelt über alle bebauten Patches.
ENND	Euklidische Nächste-Nachbar Distanz (engl. euclidean nearest-neighbor distance)	Mittlere euklidische Distanz zwischen bebauten Patches und ihren nächsten Nachbarn.
PI	Porositätsindex (engl. porosity index)	Verhältnis der Fläche von unbebauten Pixeln innerhalb eines bebauten Patches zu dessen bebauter Fläche
UD	Urbane Dichte (engl. urban density)	Anteil der bebauten Fläche an der Gesamtfläche der UTG.
UC	Urbane Kompaktheit (engl. urban compactness)	Quantifiziert die Abweichung der Kompaktheit eines bebauten Patches von einem kompakten Kreis derselben Fläche, gemittelt über alle bebauten Patches.
EMS	Effektive Netzgröße (engl. effective mesh size)	Wahrscheinlichkeit, dass zwei zufällig innerhalb des bebauten Fläche gewählte Punkte innerhalb desselben Patches liegen.
OD	Objektdichte (engl. object density)	Anzahl der bebauten Patches pro Fläche.
DI	Dispersionsindex (engl. dispersion index)	Verhältnis zwischen der Anzahl von bebauten Patches und dem Anteil des größten bebauten Patches an der bebauten Gesamtfläche.
AWSD	Flächengewichtete Standarddistanz (engl. area-weighted standard distance)	Durchschnittliche (flächen-gewichtete) Distanz zwischen Patches und dem räumlichen Schwerpunkt aller Patches.
SPI (<100m; 100–500m; 500–2000m; >2000m)	Räumliche Transinformation (engl. spatial partial information, SPI)	Räumlicher Anteil der Entropie (Transinformation) an der gesamten Entropie, wenn sie unter Berücksichtigung von Paaren von Pixeln im Raum berechnet wird. Dieses Maß für Kompaktheit wird für verschiedene Distanzen von Paaren berechnet.

4.5. Webtextdaten

Das Internet enthält große Datenmengen von öffentlich zugänglichen Texten. Für die Analyse von sozialen Prozessen wie Migration sind insbesondere Social-Media-Daten relevant, welche sich durch eine hohe Individualität und Subjektivität auszeichnen. Georeferenzierte Social-Media-Daten (GSM) werden, wie oben beschrieben, als Teil der Webtextdaten betrachtet und stellen, primär repräsentiert durch die Microblog Plattform Twitter, die primäre Textdatenquelle in den Studien dieser Dissertation dar. Neben Social-Media-Daten werden in der Dissertation auch weitere verwandte Quellen analysiert, wie Webnachrichten oder Q&A Seiten, welche nicht im engeren Sinn soziale Medien, in vielen Aspekten aber ähnlich sind und ähnliche Potenziale für die wissenschaftliche Nutzung aufweisen.

In diesem Kapitel werden alle in den Studien als Primärdaten genutzten Textdatenquellen erläutert.

4.5.1. Twitter

Twitter (seit 2023 als „X“ bezeichnet) ist eine öffentliche Microblog Plattform (Java et al. 2007).

Über die Jahre nach seiner Gründung wurde Twitter eine der meistgenutzten Social-Media-Plattformen und erreichte durch die öffentliche Natur und zeitweisen liberalen Datenzugang eine hohe Sichtbarkeit in Journalismus und Forschung (Williams et al. 2017). 2023 wurde im

Rahmen einer weiten Umstrukturierung des Unternehmens durch den neuen Eigentümer Elon Musk die Marke Twitter in „X“ umbenannt. Im Rahmen dieser Dissertation wird weiterhin die Bezeichnung Twitter benutzt, um die Plattform in der Zeit vor 2023 zu bezeichnen, für welche auch die Daten dieser Dissertation gesammelt wurden (Nutakki et al. 2025).

Auch wenn beispielsweise Facebook oder Instagram insgesamt höhere Nutzungszahlen aufweisen (Adika 2023; Kemp 2025), ist Twitter in der Forschung und Öffentlichkeit stark repräsentiert. Die Gründe dafür liegen wahrscheinlich in der öffentlichen Natur der Plattform und den ursprünglich leicht zugänglichen Informationen.

Dieser Zugang wird durch die von Twitter bereitgestellte Programmierschnittstelle (API) ermöglicht. Über diese API können die primären Inhalte der Tweets abgefragt werden. Die Nutzung der Twitter API wurde über lange Zeiträume für wissenschaftliche Zwecke frei angeboten. Zeitweise war sie für Nutzung innerhalb ausgewählter wissenschaftlicher Prozesse sogar kostenlos, wenngleich mit Einschränkungen bei Zugang und Datenvolumen (Twitter Inc. 2022). Über diesen sogenannten *Academic Research Track* (Twitter Inc, 2022) wurde für die Zwecke der Untersuchung von Migration Zugriff auf die API gewährt. Hierdurch bestand für die Umsetzung der Dissertation die Möglichkeit, bis zu 10 Millionen Tweets pro Monat kostenfrei abzurufen.

Die API erlaubt auch die Abfrage von Metadaten von Tweets und Accounts. Dies beinhaltet öffentlich einsehbare Metadaten, welche auch für Nutzende der offiziellen Twitter-Apps angezeigt werden. Hierzu gehören beispielsweise für jeden Tweet die Nutzernamen, Anzahl der Retweets, und Zeitstempel. Zusätzlich können über die API auch Metadaten abgerufen werden, die nicht in allen Apps öffentlich sichtbar sind, wie der Zeitpunkt der Erstellung des Accounts und die App, mit welcher der Tweet gesendet wurde. Von besonderer Relevanz für diese Dissertation sind unter den Metadaten die Georeferenzen (Geotags). Diese können auf Twitter in zwei Formen registriert sein:

Als **Koordinaten** (fortan: Koordinatenreferenz). Hierbei ergänzen Nutzende ihren Textbeitrag durch ein Paar von Längengrad- und Breitengradkoordinaten, welches in der Regel dem Standort des Gerätes entspricht, und von dessen Hardware bereitgestellt wird. Die Nutzung dieses Features wurde im Laufe des Jahres 2019 deaktiviert, kann aber noch vorkommen wenn Inhalte von anderen Apps (beispielsweise Instagram) verlinkt werden (Kruspe et al. 2021).

Als **Orte** (fortan: Ortsreferenz). Hierbei wählen Nutzende einen Ort aus einer begrenzten Liste von möglichen Ortsnamen aus. Die genaue Quelle dieser Ortsnamen und der Algorithmus, der die Vorschläge bedingt, ist nicht bekannt, berücksichtigt aber i.d.R. den Standort des Gerätes des Nutzenden. Diese Orte entsprechen, anders als die Koordinatenreferenzen, nicht notwendigerweise einem Punkt, sondern einem Gebiet wie einer Nachbarschaft, Stadt, Verwaltungseinheit oder einem Land. Diese sind in ihrer Ausdehnung datentechnisch durch vier Eckkoordinaten definiert. Der Typ des Ortes (Punkt, Nachbarschaft, /Stadt, Verwaltungseinheit, Land)⁵² ist als Teil der Metadaten angegeben.

Unabhängig von der Form, in der sie auftreten, sind Geotags eine große Bereicherung des Datenwerts von Tweets, da sie über den räumlichen Bezug des jeweiligen Texts informieren. Mit Geotags versehene Tweets werden in dieser Dissertation als *Geotweets* bezeichnet (Hu et al. 2024a; Lewis und Kakkar 2025). Zusätzlich zu dieser räumlichen Referenz definiert der

⁵² Im Original jeweils bezeichnet als: *poi, neighborhood, city, admin, country*

Zeitstempel eine zeitliche Referenz. Zusammen können sie, auf Nutzerebene als digitale Spur (engl. *digital trace*) gesammelt, über die Mobilität von Nutzenden Auskunft geben (Armstrong et al. 2021). Aus diesen Gründen stellt Twitter für räumliche Studien eine vielseitige Datenquelle dar (Kemp 2019b).

Auch die weiteren Metadaten können für Studien über das Verhalten von Nutzenden und die Verbreitung von Inhalten sehr hilfreich sein, um demographische oder technische Filterungen vorzunehmen. In dieser Arbeit wurde beispielsweise nur mit Tweets gearbeitet, welche von offiziellen Apps gesendet wurden, um den Einfluss von Automatisierung und Software zu reduzieren. Die Anzahl der Retweets und Antworten auf einen Tweet ist auch in dessen Metadaten enthalten.

Bei der Nutzung von Twitter-Daten muss berücksichtigt werden, dass Twitter im Laufe der Zeit stets Einflüssen unterlag, welche die Nutzung und damit die Datenverfügbarkeit beeinflussen. So nahm Twitter im Jahr 2019 technische Änderungen in der Geotagging-Funktionalität vor, welche das Taggen von Koordinaten als Option entfernte, worauf sich die Verfügbarkeit von Koordinatenreferenzen in Twitter-Datensätzen erheblich reduzierte (Kruspe et al. 2021). Auch die Übernahme von Twitter durch Elon Musk und das Rebranding als „X“ im Jahr 2023 gingen mit erheblichen Änderungen in der Moderation der Plattform und Demographie von Nutzenden einher (Hickey et al. 2023; Bisbee und Munger 2025). Hinzu kommen externe Ereignisse, wie beispielsweise die nigerianischen #endsars Proteste, in denen Twitter eine wichtige Rolle als Kommunikationsplattform spielte, was eine erhebliche Erhöhung des lokalen Datenvolumens bewirkte (Ojedokun et al. 2021).

In den Studien dieser Dissertation wurden die Analysezeiträume so gewählt, dass sie solche bekannten Ereignisse nicht beinhalten. Hierdurch sollen zeitliche Verzerrungen (Olteanu et al. 2019) minimiert werden. Neben zeitlichen Verzerrungen ist Twitter selbstverständlich auch nicht über alle Demographien hinweg gleich verteilt (Spyratos et al. 2018; Taubenböck et al. 2018). Es ist gemeinhin akzeptiert, dass junge und urbane Bevölkerungsgruppen auf Social Media stärker repräsentiert sind (Hughes et al. 2016; Blanchard et al. 2020). Dass auch die Nutzung von Geotags nicht homogen ist, wurde von Malik et al. (2015); Armstrong et al. (2021) aufgezeigt. Speziell in Bezug auf Migration attestierten Armstrong et al. (2021) und Petutschnig et al. (2020), dass Twitter weniger zum Verständnis von Flüchtlingen und ländlichen Migrationsbewegungen zu informieren, sondern eher städtische Migration, Geschäftsreisende, Touristen und transnationale Gruppen abbildet – Mobilitätsformen, welche weitverbreitet, aber in der Forschung unterrepräsentiert (Armstrong et al. 2021) und schwer zu messen sind (KC 2025).

Zusätzlich muss berücksichtigt werden, dass Twitter-Accounts teilweise oder vollständig automatisiert werden können. Die Methodiken, welche in dieser Arbeit zur Entfernung von automatisierten Inhalten genutzt wird, variieren von Studie zu Studie basierend auf den jeweiligen Forschungsansätzen und werden in dem jeweiligen methodischen Kapitel detailliert.

Twitter-Daten wurden in allen Studien genutzt. Da die Datenakquise eine essenzielle Komponente der Methodik darstellt und an die Zielsetzung angepasst werden muss, werden die in den Studien kompilierten Twitter-Datensätze für jede Studie detailliert in Kapitel 5 beschrieben.

4.5.2.Nairaland

Die Webseite Nairaland ist ein Webforum mit dezidiert nigerianischem Fokus. Nairaland ist ein Ort des Austauschs für Nigerianer in Nigeria, aber auch nigerianische Emigrierte und Nicht-Nigerianer, welche sich für Themen des Landes interessieren (Heyd 2014). Bei der Gründung des Forums war das Ziel des Gründers, eine Diskussionsplattform zu schaffen, welche Nigerianern zuhause eine Stimme gab, anders als die von Nigerianern in UK und US betriebenen Plattformen, welche Themen behandelten, die für Nigerianer im Ausland von Interesse waren (Chiluwa und Odebunmi 2016). Es entwickelte sich in ein großes öffentliches Forum für eine Vielzahl an Themenbereichen und mit einem starken Fokus auf Inhalte, die für Nigerianer in Nigeria von Interesse sind (Chiluwa und Odebunmi 2016).

Das Forum ist hierarchisch in mehrere Unterforen strukturiert, welche in manchen Fällen selbst weitere Unterforen haben. Jeder der sich so ergebenden (mehreren Dutzend) Unterforen trägt einen Titel und eine kurze Beschreibung, beispielsweise:

- **“Family:** Marriage/Family issues including husband to wife, parent to child, sibling and extended family relationships”⁵³
- **„Politics:** Our country Nigeria is the giant of Africa!”⁵⁴
- **“Programming:** Software programming, development of applications”⁵⁵
- **“Agriculture:** The business and science of crop and livestock production, Agric Science, Agric Economics, etc “⁵⁶

Innerhalb dieser Foren können Nutzende frei Threads erstellen und betiteln. In diesen Threads können Nutzende in einer linearen Abfolge von hauptsächlich textbasierten⁵⁷ Kommentaren interagieren. Diese Kommentare sind, mit wenigen Ausnahmen, auch ohne Account öffentlich einsehbar. Aus datenkonzeptioneller Sicht stellt Nairaland also eine große, öffentliche, vorstrukturierte Datenquelle dar, welche diverse Textinhalte einer grob abgegrenzten Nutzerdemographie enthält. Da die Nutzenden frei entscheiden können, in welchem Unterforum sie kommentieren, kann es auch als ein durch die Nutzenden selbst gelabelte Datenquelle betrachtet werden. Dabei werden die Kategorien von den Plattformbetreibern benannt und dann von den Nutzenden implizit inhaltlich ausgefüllt.

Hierbei kann selbstverständlich nicht davon ausgegangen werden kann, dass alle Posts im *Politics* Forum auch ein ähnliches Verständnis von Politik haben. Zudem ist insbesondere bei langen Themen eine Vermischung von Themen und ein Abdriften zu erwarten (Osborne 1998). Dennoch stellt die Interpretation der Forenstruktur als Annotation im Vergleich zu manuellen Annotationen eine Alternative dar, welche den Arbeitsaufwand reduziert und die Online-Community in den Annotationsprozess implizit einbindet.

Es wird keine API bereitgestellt (Oyebode und Orji 2019a), aber die Webseite erlaubt automatisierte Zugriffe zur Auslesung⁵⁸ (sogenanntes *Web-Scraping*, siehe Krotov et al. 2020 & Honkanen und Müller 2021). Auf diese Weise bezogene Nairaland-Daten wurden in verschiedenen Studien zu verschiedensten Themen genutzt. Beispielsweise Politik (Nelson et

⁵³ <https://www.nairaland.com/family>

⁵⁴ <https://www.nairaland.com/politics>

⁵⁵ <https://www.nairaland.com/programming>

⁵⁶ <https://www.nairaland.com/agriculture>

⁵⁷ Auch die Einbettung von Emojis und Bildern in den Text ist möglich

⁵⁸ <https://www.nairaland.com/robots.txt>

al. 2018; Oyeboade und Orji 2019b), Humor (Lamidi 2016), Terrorismus und Konflikt (Chiluwa und Odebunmi 2016), Gesundheit (Oyeboade und Orji 2019a), sprachliche Ausdrücke von Zugehörigkeit (Heyd 2014) und politisches Interesse von Diaspora-Angehörigen (Basmus et al. 2019). Mittels randomisiertem Scrapings werden Nairaland Daten aus einer zufälligen, aber großen, Auswahl von Threads innerhalb der Unterforen bezogen⁵⁹. Anhang A1 beschreibt diesen Datensatz genauer. Diese Daten werden in Studie B1 als Trainingsdaten und in Studie C1 als eine der primären Datenquellen verwendet.

4.5.3.Reddit

Reddit ist ein Webforum und Nachrichtenaggregator. Die 2005 gegründete Plattform ist in eine Vielzahl an untergeordneten Foren strukturiert, welche (anders als beispielsweise bei Nairaland) nicht hierarchisch angeordnet, sondern konzeptionell gleichrangig sind. Jedes dieser Unterforen (*subreddits* genannt) ist typischerweise einem bestimmten Thema, Inhaltstyp oder Community gewidmet.

Auf Reddit angelegte Accounts sind nicht an bestimmte Subreddits gebunden und haben die Möglichkeit, von anderen Accounts erstellte Inhalte zu bewerten und damit deren Sichtbarkeit zu erhöhen. Aufgrund seiner inhaltliche Vielfalt und Größe wurde Reddit bereits als Forschungsgegenstand und Datenquelle genutzt, unter anderem für geographische Informationen (Fox et al. 2021; Berragan et al. 2022). Im Vergleich zu Twitter ermöglicht Reddit auch deutlich längere Texte, was einen Vorteil für sprachwissenschaftliche Studien darstellt.

Reddit ist eine öffentliche Plattform, deren Inhalte auch ohne Accounts eingesehen werden können. Eine API ermöglicht zusätzlich den strukturierten Zugang zu Daten. Diese API wurde in den letzten Jahren kommerzialisiert und ist nicht mehr kostenfrei zugänglich (Davidson et al. 2023). Diese Dissertation nutzt einen vorkompilierten Reddit-Datensatz, den WEBISTldr-17 Reddit Korpus welcher von Völske et al. (2017) für das Training von Modellen zur Textzusammenfassung kompiliert wurde. Dieser bereits vorprozessierte Datensatz enthält relativ lange und inhaltsreiche Kommentare, was für die Anwendung in Studie C1 einen Vorteil darstellt. Die Kommentare entstammen mehreren tausend verschiedenen Subreddits. Um diese zusammenzufassen, wurde auf die von Partridge et al. (2024) erstellten *Reddit Community Embeddings* zurückgegriffen. Diese gruppieren Subreddits basierend auf Ähnlichkeit in Nutzernetzwerken.

Reddit-Daten werden in Studie C1 als eine der primären Datenquellen verwendet. In dieser Studie wurden die Reddit Community Embeddings-Gruppen weiter zu Themen gruppiert. Dieser Prozess ist in Kapitel 5.5 detailliert.

4.5.4.Stackexchange

Stackexchange ist ein Q&A⁶⁰ Netzwerk von über 170 Webseiten, welche alle einem Themenfeld gewidmet sind, beispielsweise Kochen⁶¹, Musik⁶², Robotik⁶³, und Reisen⁶⁴, und

⁵⁹ Einige Unterforen wurden nicht berücksichtigt, da sie wenige Beiträge aufwiesen oder nur kurze Zeiträume aktiv waren.

⁶⁰ Questions & Answers (Deutsch: Fragen & Antworten)

⁶¹ <https://cooking.stackexchange.com/>

⁶² <https://music.stackexchange.com/>

⁶³ <https://robotics.stackexchange.com/>

⁶⁴ <https://travel.stackexchange.com/>

das Stellen & Beantworten von Fragen zu den jeweiligen Themen ermöglichen.

In diesem Prozess erstellen Nutzende Threads mit einer Frage zu einem Thema, auf welches mehrere andere Nutzende in Kommentaren antworten können. Alle Nutzende haben die Möglichkeit, die Konversation insgesamt oder auch einzelne Antworten zu bewerten und dadurch die Sichtbarkeit zu beeinflussen. Die Fragen und Kommentare sind auch ohne Stackexchange-Account öffentlich sichtbar. Das Netzwerk stellt keine API bereit, veröffentlicht diese Daten aber in monatlichen Abständen in der Form von sogenannten Data-Dumps. Diese Daten wurden bereits beispielsweise für Netzwerkanalysen und das Training von Sprachmodellen genutzt (Firouzjaei 2024; Stack Exchange Inc. 2024).

Im Rahmen dieser Arbeit wurden Daten von Stackexchange in Studie C1 als eine der primären Datenquellen verwendet. Hierfür wurden aus den Webseiten manuell eine Auswahl getroffen und zu Themen zusammengefasst. Dieser Prozess ist in Kapitel 5.5 detailliert.

4.5.5.GDELТ-GKG

Die Datenbank Global Knowledge Graph (GKG) des Global Database of Events, Language, and Tone Project (GDELТ) ist ein stetig aktualisiertes Repositorium von Web-Nachrichtenartikeln. Daten aus einer Vielzahl von Nachrichtenquellen in vielen Ländern und mehr als einhundert Sprachen werden regelmäßig ausgelesen, prozessiert und in die Datenbank eingepflegt. Dabei werden über Sprachprozessierung auch inhaltliche Charakteristika der Texte extrahiert und als Metadaten in die Datenbank eingetragen. Hierzu gehören beispielsweise Themen, Akteure und Orte, die in den Texten erwähnt wurden, einschließlich der Position im Text (Leetaru und Schrodt 2013). Der GKG besteht aus einem Nachrichteneintrag pro Zeile. Die Texte selbst werden nicht als Teil des GKG gespeichert, die Metadaten sind jedoch öffentlich einsehbar. Tabelle 2 zeigt Beispiele für Metadaten des GKG. Eine ausführliche Beschreibung ist im GDELТ Codebuch einsehbar (GDELТ Project 2015).

Tabelle 2: Beispiele für Metadatenfelder des GDELТ Global Knowledge Graph

Feldname	Beschreibung
GKGRECORDID	Ein String mit einer global einzigartigen Zeichenfolge, welche zur Identifikation dient
SourceCommonName	Ein interpretierbarer Name des Autors oder der Quelle (z.B. „BBC-Monitoring“)
DATE	Das Publikationsdatum des Artikels, Nummer im YYYYMMDDHHMMSS Format.
V2Counts	Ein Semikolon-getrennter String mit allen im Text detektierten Zahlenmengen sowie ihrer Position (in Zeichen) im Text
V2Themes	Ein Semikolon-getrennter String mit allen im Text detektierten Themen sowie die Position (in Zeichen) der Person im Text
V2Locations	Ein Semikolon-getrennter String mit allen im Text detektierten Orten, einschließlich Ortstyp, Ortsname und Koordinaten des Ortes sowie die Position (in Zeichen) der Ortserwähnung im Text
V2Persons	Ein Semikolon-getrennter String mit allen im Text detektierten Personen sowie die Position (in Zeichen) der Person im Text
...	...

Unter diesen Metadaten sind insbesondere die von GDELТ identifizierten Themen (V2Themes) und Orte (V2Locations) für die Dissertation von Interesse. Die Themen sind kurze thematische Annotationen für Stellen im Text. Die Taxonomie dieser Themen ist teilweise von GDELТ definiert, zum Teil aber auch aus anderen Taxonomien entlehnt, beispielsweise der

Thementaxonomie der Weltbank (World Bank 2015), CrisisLex (Olteanu et al. 2014) oder des Codebuchs der Conflict and Mediation Event Observations (CAMEO). GDELT legt nicht offen, mit welchem Algorithmus die Themen und Orte identifiziert wurden, und stellt auch keine umfassende Beschreibung derselben bereit (Williams 2020). Jedoch haben viele der Themen aussagekräftige Namen und Präfixe, welche Einblicke in die Quelle und Art des Themas gestatten. Als Beispiel sei hier das Thema *WB 470 EDUCATION* genannt, welches sich vermutlich aus der Weltbank (WB) Taxonomie ableitet und mit Bildung (engl. Education) befasst.

Die vom GDELT-Projekt bereitgestellten Daten sind frei zugänglich und unbeschränkt für wissenschaftliche Zwecke nutzbar. GDELT bietet zwar keine Programmierschnittstelle für den Download an, stellt die Daten aber auf der Google Cloud Plattform zur Verfügung. Von dort können die Rohdaten direkt heruntergeladen werden oder über den BigQuery Dienst von Google selektiert und vorprozessiert werden (Luber 2021).

Von GDELT im GKG gesammelte Daten sind eine mögliche Ressource für den rechnergestützten Journalismus (El Ouadi und Beskow 2024) und wurden auch wissenschaftlich für vielseitige Anwendungen ausgewertet, beispielsweise für Konfliktforschung (Qiao et al. 2017; Blanqué et al. 2022; Senaratne et al. 2023), Stimmungsanalysen (Bodas-Sagi und Labeaga 2016), städtisches Marketing (Zheng 2020), Naturkatastrophen (Owuor et al. 2020), und viele weitere Themen.

Im Rahmen dieser Arbeit wurden GDELT-GKG Daten in Studie C1 als eine der primären Datenquellen verwendet. Hierbei wurden aus den von GDELT identifizierten Themen händisch eine Anzahl ausgewählt und in generelle Themenbereiche zusammengefasst.

4.5.6. GDELT- IA-Americana

Zusätzlich zu dem Global Knowledge Graph (GKG) stellt GDELT auch mehrere Spezialkollektionen bereit, welche mit denselben Algorithmen prozessiert wurden wie der GKG. Eine dieser Spezialkollektionen sind die amerikanischen Bücher und Dokumente der American Libraries Collection⁶⁵ des Internet Archive für welches englische Volltexte verfügbar waren. Dies sind größtenteils literarische Werke in der Public Domain⁶⁶, frei zugängliche Zeitschriften und offizielle Regierungsdokumente. Diese Texte wurden von GDELT mit denselben Metadatenstruktur annotiert wie die Nachrichtentexte des GKG und sind auch gleichsam über Google Cloud verfügbar (Leetaru 2015). Anders als bei GDELT ist für diese Daten auch der Volltext bereitgestellt, wenn auch mit Digitalisierungsfehlern und Abstrichen bei der Genauigkeit der Metadaten (Leetaru 2015).

Der Datensatz enthält Texte, welche über 215 Jahre alt sind, und deckt damit auch einen deutlich längeren Zeitraum ab als der GKG, wobei die Anzahl an Büchern für Jahre nach 1922 stark abfällt, aufgrund des Copyrights (GDELT Project, o.D.), welches zum Zeitpunkt der Prozessierung noch für später publizierte Werke in Kraft war. Im Rahmen dieser Arbeit werden Daten aus dieser Sammlung als GDELT-IA-Americana bezeichnet und in Studie C1 als eine der primären Datenquellen verwendet. Hierbei werden alle verfügbaren Einträge der Jahre 2005 bis 2014 genutzt, und, wie auch für die GDELT Daten (4.5.5), aus den von GDELT identifizierten Themen eine Anzahl ausgewählt und in Themenbereiche zusammengefasst.

⁶⁵ <https://archive.org/details/americana>

⁶⁶ Frei von Urheberrechten, grob vergleichbar, aber nicht identisch mit dem deutschen Begriff der Gemeinfreiheit.

5. Studienspezifische Methoden

5.1. Studie A1: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Erfassung digitaler Ungleichheiten in Afrikas expandierenden städtischen Räumen

In dieser ersten Studie werden FED und GSM räumlich verschnitten, um digitale Disparitäten zwischen älteren und neueren Siedlungsgebieten in afrikanischen Städten zu quantifizieren.

Hierzu wird aus FED und Twitter-Daten zunächst die Dichte von Twitter Aktivität in verschiedenen Bebauungs-Altersklassen berechnet (5.1.1). Innerhalb der sich ergebenden Verteilung wird Ungleichheit zwischen alt und neu mittels eines speziell definierten Maßes quantifiziert (5.1.2) und mittels Permutation auf Signifikanz getestet (5.1.3). Die Ergebnisse werden in Bezug zu Georegionen der UN, Geplantheitsintensität und morphologischen Metriken gesetzt und statistisch ausgewertet (5.1.4).

5.1.1. Berechnung der Geotweet-Dichte

Für jedes UTG wird ein kreisrundes Untersuchungsgebiet mit 5km Radius um das UTG-Zentrum definiert. Ein kreisförmiges Untersuchungsgebiet ist für die vergleichende Analyse wünschenswert, da eine Kreisform unabhängig gegenüber Effekten der geographischen Ausrichtung ist.

Zunächst wird mittels Fernerkundung (4.4) für jedes UTG eine Kartierung der Bebauungsjahre vorgenommen. Der WSFEvolution Datensatz (4.4.1) deckt hierbei den Zeitraum 1985-2015 ab. Um die Jahre 2016-2020 abzudecken, wird mittels des DIY-BU Tools (4.4.2) eine Kartierung durchgeführt.

Alle verfügbaren Geotweets werden mittels der R-Bibliothek *academictwitterR* (Barrie und Ho 2021) für den Analysezeitraum (1. Januar 2010 – 1. Januar 2020) und die 135 UTG über die Twitter-API bezogen.

Von den durch die API bereitgestellten Geotweets werden nur diejenigen genutzt, welche Koordinatenreferenzen oder Ortsreferenzen der Ortsklasse *Punkt* beinhalten, da die Analyse eine hohe räumliche Präzision erfordert. Ausgeschlossen werden Geotweets von Accounts, welche eine große Anzahl von Geotweets mit deutlichen Anzeichen von Automatisierung produzierten. Als Anzeichen für Automatisierung werden ungewöhnlich hohe Aktivität, Geschwindigkeit und Gleichzeitigkeit gesehen, also beispielsweise das fast gleichzeitige Absenden von Tweets mit Geotags an weit voneinander entfernten Orten. Diese Kriterien wurden auch beispielsweise von Petutschnig et al. (2020) und Lemoine-Rodríguez et al. (2024) angewendet.

Für jeden der Geotweets dieser geeigneten Accounts wird aus den FED das Bebauungsjahr des im Tweet getaggtten Ortes abgeleitet. Somit ergibt sich für jeden Geotweet (1) ein Kalenderjahr, in welchem er gesendet wurde, (2) ein Baujahr, in welchem der dortige Ort erbaut wurde, sowie (3) ein Bebauungs-Alter aus der Differenz der beiden Jahre (Abbildung 17).

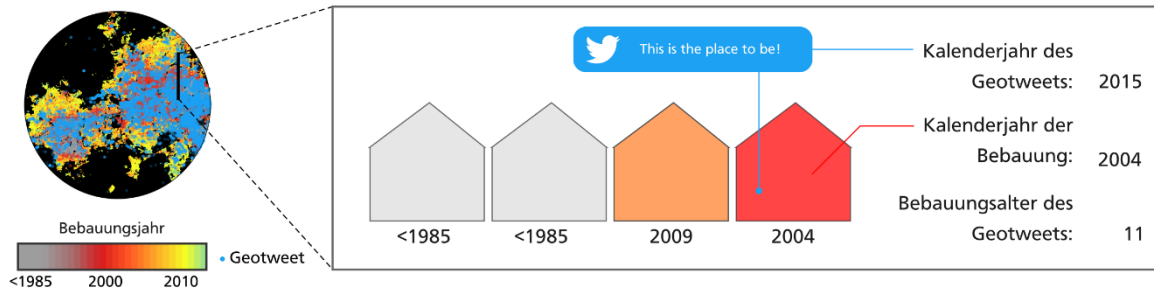


Abbildung 17: Ableitung des Baualters der Tweets aus der Differenz des Jahres der Veröffentlichung des Tweets und dem Jahr der Bebauung des Ortes, an welchen der Geotweet getaggt wurde.

Zusätzlich wird eine Gewichtung der Tweets wie folgt berechnet, um durch andere Accounts rezipierte Tweets positiv, und wiederholte Tweets desselben Accounts am selben Ort negativ zu gewichten. Jedes UTG wird dafür in ein Raster von 100-m Zellen und Kalendermonaten aufgeteilt, woraufhin sich das Gewicht für jeden Tweet so berechnet:

$$W_{tweet} = \frac{1}{n_{pro\ Nutzer}} + n_{Retweets} + n_{Antworten}$$

Hier entspricht $n_{pro\ Nutzer}$ der Anzahl an Geotweets des Autoren-Accounts in derselben Zelle während desselben Monats, $n_{Retweets}$ der Anzahl an Retweets, welche der Geotweet erhalten hat und $n_{Antworten}$ der Anzahl der Antworten auf den Geotweet.

Zur Berechnung der Dichte wird daraufhin folgendermaßen vorgegangen: Zunächst wird für jedes der zehn Kalenderjahre im Untersuchungszeitraum die Gewichte aller Geotweets mit demselben Baualter summiert. Es ergibt sich für jedes der zehn Kalenderjahre die gewichtete Anzahl an Geotweets pro Baualters-Klasse, wobei die Anzahl an Klassen von den verstrichenen Jahren seit 1985 abhängt, beispielsweise gäbe es im Kalenderjahr 2014 maximal $2014 - 1985 = 29$ Baualters-Klassen, denn zwischen Bauzeiträumen vor 1985 kann nicht differenziert werden. In Abbildung 18 ist dieser Vektor als e) gekennzeichnet. Dem entsprechend ist die Formel zur Berechnung der Geotweet-Dichte für eine Baualters-Klasse in einem Jahr X wie folgt:

$$density_X^{Alter} = \frac{Ntw * X^{Alter}}{Fläche_X^{Alter}}$$

Wobei $Ntw * X^{Alter}$ der gewichteten Anzahl Geotweets im Jahr X entspricht, deren Sendeorte in diesem Jahr ein gewisses *Alter* aufweisen. Gleichsam ist die $Fläche_X^{Alter}$ die bebaute Fläche des *Alter* im Kalenderjahr X, in welchem der Geotweet gepostet wurde.

Aus den FED wird die Fläche einer jeden Baualtersklasse abgeleitet (Abbildung 18 c). Aus dem Verhältnis von gewichteter Anzahl und Fläche ergibt sich dann die Geotweet-Dichte (in Geotweets pro km²), welche wie in f) als Kurve dargestellt werden kann.

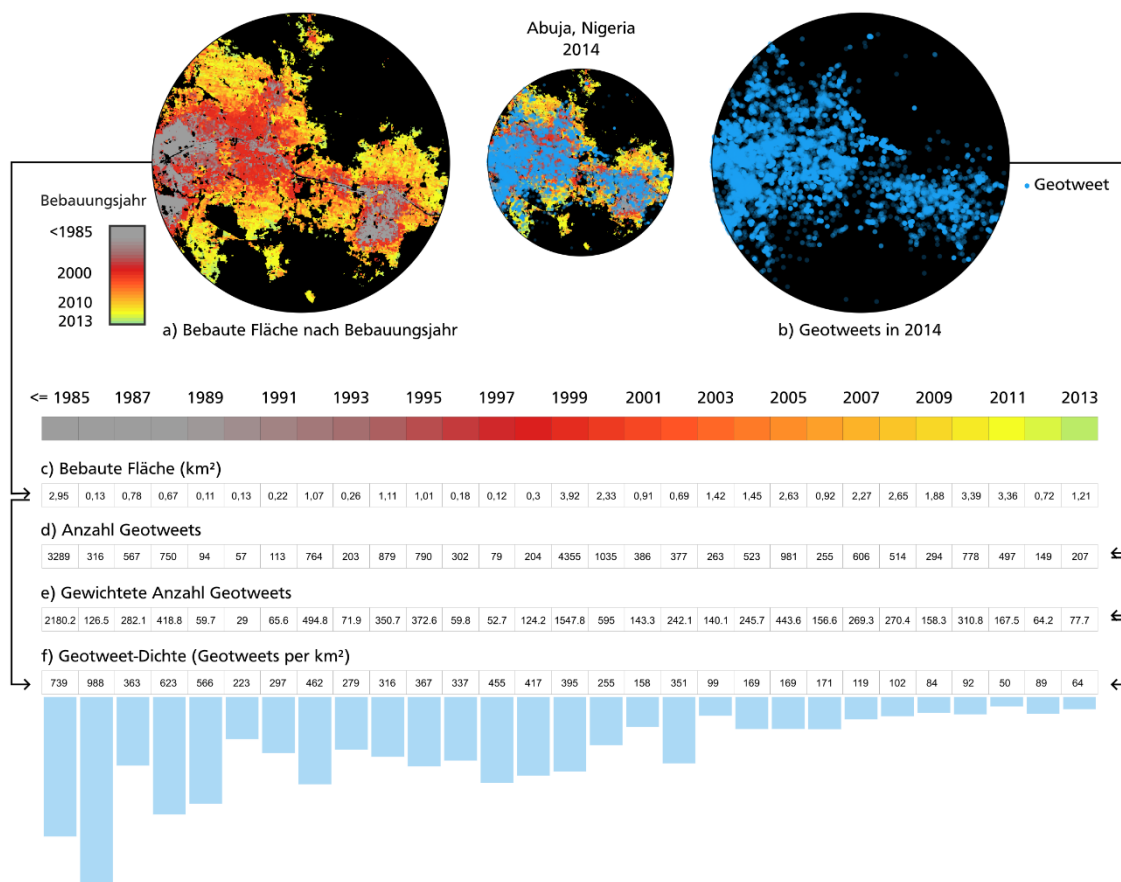


Abbildung 18: Ableitung der Geotweet-Dichte (f) für ein Jahr (2014) in einem UTG (östliches Abuja, Nigeria). Aus den Eingangsdaten (a) und (b) wurde jeweils die in jedem Jahr bebaute Fläche (c) und die Anzahl der Geotweets (d, gewichtet e) berechnet. Aus diesen ergibt sich die Geotweet-Dichte pro Baujahr (Alter). In diesem Fall ist klar ersichtlich, dass die Geotweet-Dichte in älteren Gebieten höher ist als in neueren.

5.1.2. Quantifizierung der Dichtedisparität

Bei der Betrachtung dieser Kurve ist ein mögliches Ungleichgewicht in der Dichte zwischen älteren (links in Abbildung 19) und neueren (rechts in Abbildung 19) Bereichen ein Indikator für die zu untersuchende digitale Disparität.

Zwecks einer quantitativen Analyse über alle UTG hinweg soll dieses Ungleichgewicht quantifiziert werden. Die dafür genutzte Metrik sollte gerichtet sein, sodass sie einen hohen Wert annimmt, wenn die Dichte in älteren Teil höher ist, und umgekehrt. Sie soll zusätzlich vergleichbar zwischen Sequenzen unterschiedlicher Länge sein.

Zum Vergleich von Verteilungen kommen Divergenzmetriken wie die Kullback-Leibler Divergenz infrage. Alternativ könnte eine Korrelation oder Regression genutzt werden, um den Zusammenhang zwischen Baujahr und Twitter-Dichte zu analysieren. Diese Ansätze erfüllen jedoch die oben genannten Anforderungen der Gerichtetheit und Längeninvarianz nicht (für eine Illustration, siehe Appendix A2). Daher wird eine neue Metrik definiert.

Sie wird in dieser Arbeit aufgrund ihrer Ähnlichkeit zu Schiefemaßen (Doane und Seward 2011) als Schiefe der Digitalen Dichte (engl. *Digital Density Skew*, DDS) bezeichnet.

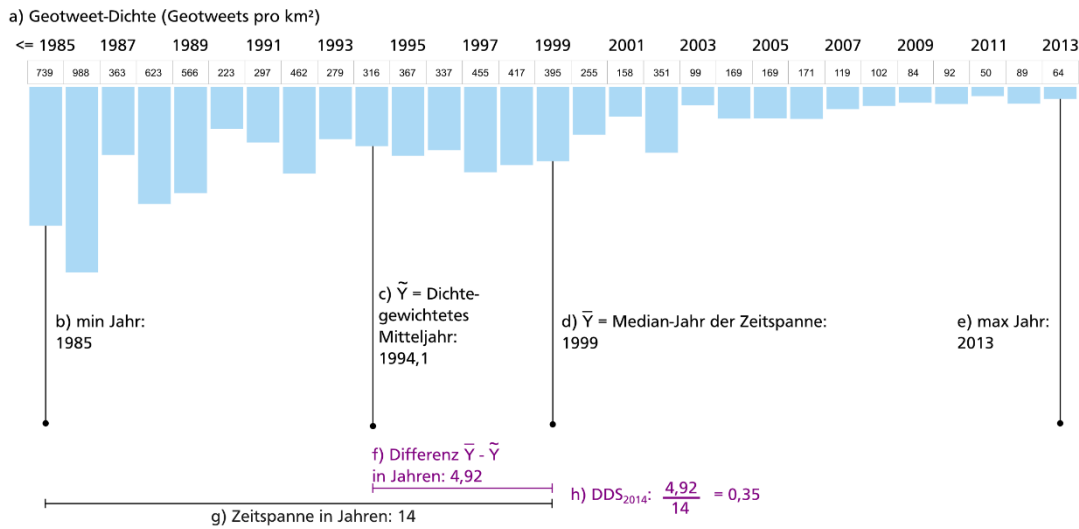


Abbildung 19: Berechnung der Schiefe der Digitalen Dichte (DDS).

Für Fälle, in welchen Lücken in der Kurve existieren (also keine Tweets in einem Jahr in einer Bebauungsaltersklasse gesendet wurden), wird ein Dichte-Wert linear interpoliert, in der Annahme, dass die Sichtbarkeit von Stadtteilen eine zeitliche Kontinuität aufweist.

Basierend auf der in 5.1.1 beschriebenen und in Abbildung 19 illustrierten Dichteverteilung ergibt sich der DDS für ein UTG in einem Jahr X dann wie folgt:

$$DDS_X = \frac{\tilde{Y} - \bar{Y}}{Y_{min} - \bar{Y}}$$

Hierbei stellt $\bar{Y}$ den Median der Spanne an möglichen Bebauungsjahre dar (also beispielsweise 1985, 1986, ..., X) und Y_{min} das früheste Jahr dieser Spanne⁶⁷. $\tilde{Y}$ ist das durch die Geotweet-Dichte gewichtete Mitteljahr, welches die Dichteverteilung sozusagen in zwei gleichwertige Hälften trennt:

$$\tilde{Y} = \frac{\sum_{i=1985}^{n=X} Y_i * Dichte_X^{Alter}}{\sum_{i=1985}^X Dichte_X^{Alter}}$$

Der Zähler in der DDS-Berechnung ist also die Differenz zwischen Median und Mittelwert, welche bei schiefen Verteilungen besonders groß wird. Der Zähler ($Y_{min} - \bar{Y}$) ist die maximal mögliche Verschiebung, welche sich ergeben würde, wären alle Tweets in der ältesten Bebauungsaltersklasse gesendet worden. Der DDS liegt infolge immer in einer Spanne von -1 bis 1. Abbildung 20 illustriert die DDS-Metrik für einige fiktive Verteilungen. Der DDS nimmt -1 an, wenn alle Tweets in der ältest-möglichen Bebauungsaltersklasse gesendet wurden, und 1, wenn alle Tweets in der newest-möglichen Bebauungsaltersklasse gesendet wurden. Ein Wert von 0 deutet an, dass ein Gleichgewicht (nicht notwendigerweise eine Gleichverteilung) in der Verteilung besteht. Ein Wert von 0,5 kann als ein Ungleichgewicht zugunsten der älteren Gebiete interpretiert werden, welches 50% der theoretisch möglichen Ungleichgewichts entspricht.

⁶⁷ In der Regel beginnt die Spanne mit dem Baujahr 1985. In seltenen Fällen führen Datenlücken dazu, dass die Spanne mit einem späteren Jahr beginnt.

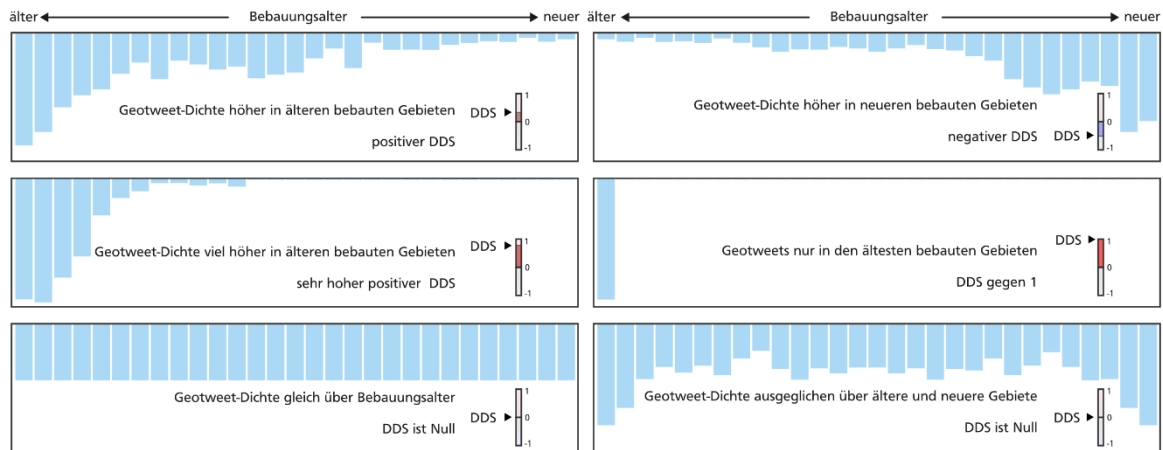


Abbildung 20: Einige fiktive Verteilungen von Geotweet-Dichte und wie die DDS-Metrik sie quantifiziert

Der DDS wird auf diese Weise separat für jedes UTG und alle 10 Kalenderjahre des Untersuchungszeitraumes berechnet. Sodann werden diese DDS-Werte für jedes UTG gemittelt, um einen einzigen DDS_{UTG} zu generieren.

Jahre im Analysezeitraum, für welche in dem betrachteten UTG nicht mindestens 100 Geotweets von mindestens 10 verschiedenen Accounts vorliegen, werden von dieser Aggregation ausgeschlossen. Ebenso werden UTG in Gänze ausgeschlossen, in welchen weniger als 100 verschiedene Accounts identifiziert werden.

5.1.3. Permutation

Die DDS-Metrik liefert eine vergleichbare und interpretierbare Quantifizierung des Ungleichgewichts der Tweet-Dichte in einem UTG, gibt jedoch keinen Einblick in die Signifikanz des betrachteten Ungleichgewichts

Um diese einzuschätzen, wird eine Nullverteilung benötigt, gegen welche der gemessene DDS_{UTG} verglichen werden kann. Aufgrund des genutzten Gewichtungsansatzes und topologischer Effekte kann nicht davon ausgegangen werden, dass die Wahrscheinlichkeitsverteilung des DDS symmetrisch und auf 0 zentriert ist.

Daher wird ein Permutationstest genutzt, um die Nullverteilung zu generieren. Permutationstests basieren im Vergleich zu parametrischen Methoden weniger Annahmen bezüglich des Prozesses der die Daten generiert (Pesarin und Salmaso 2010). Ein solcher Permutationstest kann durch zufällige Neuordnung der Tweets im Raum und Wiederholung des in 5.1.1 und 5.1.2 beschriebenen DDS-Prozesses eine Nullverteilung von DDS-Werten generieren, gegen welche die tatsächlich gemessenen Werte verglichen werden, um die Wahrscheinlichkeit einzuschätzen, dass diese zufällig entstanden sind.

Eine komplett zufällige Umverteilung der Geotweets im Raum entspricht jedoch nicht der in der empirischen Beobachtung, dass Geotweets räumlich gehäuft (engl. *clustered*) aufzutreten (Steiger et al. 2015). Um diese Häufung in der Permutation zu berücksichtigen, wird für jedes Kalenderjahr des Analysezeitraums und jedes UTG separat ein Punktprozess-Cluster-Modell (Schmidt 2013) auf die Geotweets geschätzt, welches deren Verteilung im Raum als einen geclusterten Punktprozess mit einer Cauchy Verteilung modelliert (Baddeley und Turner 2005). Dieses generiert dann räumlich zufällige Neuordnungen der Tweets, welche eine vergleichbare Häufung aufweisen, wie die tatsächlich Beobachteten (Abbildung 21).

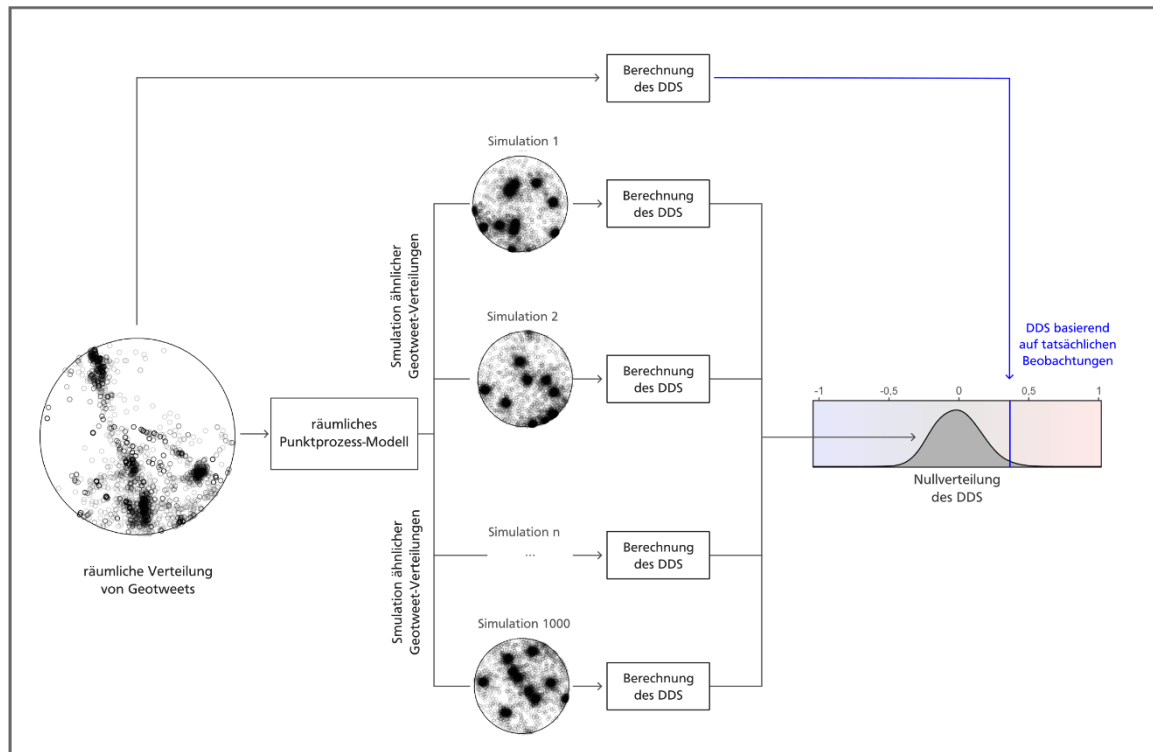


Abbildung 21: Illustration des Permutationsansatzes. Das beobachteten räumliche Geotweet-Muster wird in einem Punktprozessmodell geschätzt, welches daraufhin 1000 ähnliche aber durch Zufall beeinflusste räumliche Verteilungen generiert. Auf alle diese wird der DDS berechnet, was eine Nullverteilung ergibt anhand derer der tatsächlich gemessene DDS bewertet werden kann.

Es werden für jedes Jahr und UTG 1 000 Permutationen durchgeführt, bei welcher jeweils für jedes UTG der DDS_{UTG} neu berechnet wird. Hierdurch ergibt sich eine Nullverteilung für jedes UTG, welche über einen p-Wert die Signifikanz des gemessenen DDS_{UTG} quantifizierbar macht. So wird sowohl ein rechtsseitiger p-Wert (Anteil der Permutationen mit einem höheren DDS_{UTG} als dem tatsächlich gemessenen) und ein linksseitiger p-Wert (Anteil der Permutationen mit einem niedrigeren DDS_{UTG} als dem tatsächlich gemessenen) berechnet. Zur Einschätzung des p-Werts wird ein Signifikanzniveau von 95% angelegt.

Es werden zusätzlich Sensitivitäts-Tests durchgeführt, welche die Robustheit gegenüber (simulierten) GPS-Ungenauigkeiten, niedriger Verfügbarkeit von Tweets, niedriger Anzahl von Accounts und dem Einfluss von außergewöhnlich aktiven Accounts mit stark überdurchschnittlichen Tweet-Anzahl quantifizieren. Diese sind im Appendix A3 dokumentiert.

5.1.4. Statistische Auswertung

Ein positiver DDS_{UTG} -Wert deutet an, dass Geotweets in älteren Gebieten der Stadt dichter vorliegen als in neueren. Ein gleichzeitig signifikanter rechtsseitiger Test deutet an, dass dieser Wert höher ist, als es durch Zufall hätte sein können. Um diese statistischen Messungen einschätzen zu können, werden zunächst Beispiele für UTG mit signifikanten und nicht-signifikanten DDS_{UTG} -Werten identifiziert und deren Geotweet-Verteilungen mittels Webrecherche und weiteren Geodaten kontextualisiert und interpretiert.

Daraufhin werden DDS_{UTG} und Signifikanz statistisch auf vier verschiedene Weisen ausgewertet:

Afrikaweit. Die Verteilung von DDS_{UTG} über alle UTG hinweg wird analysiert, um die grundsätzliche Frage zu beantworten, ob ein Ungleichgewicht erkennbar ist, wie verbreitet es ist und ob es in seiner Ausprägung klar zugunsten von entweder älteren oder neueren Siedlungsbereichen tendiert.

Differenziert auf der Ebene von Georegionen Afrikas. Das UN-Geoschema (United Nations Statistical Office 1982) teilt Afrika in fünf verschiedene Regionen ein, wie in 4.1 beschrieben. Aufgrund der weiten Nutzung für statistische Zwecke ist dieser Standard der Vereinten Nationen eine geeignete Wahl, um geographische Unterschiede herauszuarbeiten. Hierfür werden die UTG nach geographischer Lage in die fünf Regionen eingeteilt und sodann wird eine Varianzanalyse (engl. Analysis of Variance, ANOVA, siehe Cohen 1988) nach Regionen über die gemessenen DDS_{UTG} angewendet.

Dabei werden sowohl eine parametrische als auch eine nicht-parametrische ANOVA (Kruskal-Wallis-Test, siehe Hollander et al. 1973) verwendet; Die nichtparametrische Version beruht nicht auf Annahmen von Normalität und Heteroskedastizität und ist daher geeignet wenn der Test auf Normalität (hier wird ein Shapiro-Wilk-Test genutzt, siehe Royston 1982) oder Homoskedastizität (hier wird ein Levene-Test genutzt, siehe Levene 1960) negativ ausfällt. Durch die Nutzung verschiedener Tests kann eine abgestufte und damit robustere Einschätzung erfolgen. P-Werte und ω^2 beziehungsweise η^2 wurden als Effektstärken für die parametrische bzw. nichtparametrische ANOVA genutzt (Cohen 1988). Zusätzlich wird mittels post-hoc durchgeführten t-Tests identifiziert, zwischen welchen Regionen Afrikas Unterschiede auftraten.

Differenziert nach Graden von Geplantheit (IoP). Die für UTG durch Expertenanalyse erfasste Geplantheitsintensität (4.4.3) ist in fünf Grade nach Debray et al. (2023) eingestuft. Hierbei kann zwischen IoP_s und IoP_b unterschieden werden. In beiden Fällen wird sowohl eine parametrische als auch eine nichtparametrische ANOVA des DDS_{UTG} nach IoP_s beziehungsweise IoP_b durchgeführt.

Modelliert im Zusammenhang mit räumlichen Metriken. Um den Zusammenhang zwischen den DDS_{UTG} und den räumlichen Metriken, welche die Siedlungsstruktur quantifizieren (4.4.4), zu testen, wird eine lineare Regression angewendet. Um auch hier eine differenzierte Einschätzung zu ermöglichen, werden Regressionen auf drei verschiedene Weisen durchgeführt. Erstens, mit Nutzung aller UTG als Beobachtungen. Zweitens, nur unter Nutzung von den UTG, welche eine signifikantes Ungleichgewicht aufweisen. Drittens wird, da eine einzelne Metrik allein die Siedlungsform nur unzureichend abbilden kann, eine multiple lineare Regression in einer schrittweisen Vorwärtsselektion von Metriken durchgeführt, bei welchem nur unkorrelierte Metriken beibehalten werden (Kuhn und Johnson 2019). Für alle Regressionsmodelle werden die Koeffizienten der Metriken, die zugehörige Signifikanz sowie der Determinationskoeffizient R^2 ausgewertet.

5.2. Studie A2: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Ableitung von Raumeinheiten für die explorative Analyse von physischen und thematischen Charakteristika urbaner Räume

Anschließend an die erste Studie wird für eine Auswahl von acht UTG durchgeführt, welche hauptsächlich in Nordafrika liegen. Zum Vergleich wurden auch zwei UTG außerhalb Nordafrikas herangezogen: Ein UTG im Süden von Abuja, Nigeria (F in Tabelle 3) und ein UTG im Nordwesten Pretorias in Südafrika (H in Tabelle 3).

Tabelle 3: Beschreibung der in Studie A2 genutzten UTG

UTG	Land	Bebaute Fläche [km²]		Anzahl der Tweets
		1985	2020	
A) Sheikh Zayed Stadt	Ägypten	0,9	46,8	93 028
B) West-Tunis	Tunesien	33,9	52,8	34 931
C) Al-Shorouk	Ägypten	0,6	39,5	28 621
D) Nador	Marokko	13,7	28,8	7 719
E) Südwest-Algiers	Algerien	8,7	42,4	9 969
F) Süd-Abuja	Nigeria	5,2	49,1	100 000
G) New Cairo	Ägypten	2,3	58,2	100 000
H) Nordwest-Pretoria	Südafrika	22,0	51,6	49 489

Ziel dieser Studie ist die algorithmische Unterteilung der UTG in räumliche Zonen, welche sich sinnvoll als Nachbarschaften oder Blocks interpretieren lassen. Diese werden für diese Studie als Teile bebauter Flächen definiert, welche zu ähnlichen Zeiten erbaut wurden, und ähnliche Funktionen aufweisen.

Um diese Aufteilung zu erreichen, werden zunächst die Geotweets räumlich verortet, mittels Sprachmodellen in numerische Embeddings transformiert und in Rasterformat überführt (5.2.1). Fernerkundungs-Raster und Textinhalts-Raster werden daraufhin zusammen in einem räumlichen Clustering genutzt, um die bebaute Fläche des UTG in diskrete Zonen zu unterteilen (5.2.2). Die Eignung der Cluster wird dann sowohl statistisch und durch Inspektion analysiert (5.2.3).

Als Fernerkundungsdaten dienen hier, wie in Studie A1, die durch Kombination von WSF und DIY-Werkzeug erstellten Kartierungen der Bebauungsjahre. Als GSM-Daten dienen hier wieder die Geotweets, deren Akquise bereits in Studie A1 beschrieben ist (5.1.1). Anders als in Studie A1 werden die UTG hier nicht kreisförmig, sondern quadratisch um das Zentrum der UTG (4.2) definiert, mit einer Seitenlänge von 10km. Hierdurch mussten keine Geotweets ausgeschlossen werden. Da anders als in Studie A1 kein quantitativer Vergleich zwischen UTG angestrebt wird, spielen hier Effekte von Ausrichtung des UG keine Rolle, weswegen die Reduktion auf ein kreisförmiges Gebiet nicht notwendig ist.

5.2.1. Umwandlung der GSM-Daten in Rasterformat

Geotweets werden wie in Studie A1 (Kapitel 5.1.1) beschreiben für den die Jahre 2010 bis 2020⁶⁸ spannenden Analysezeitraum akquiriert und vorbereitet. Nur Geotweets von offiziellen Apps werden genutzt. Für jede UTG wurden maximal 100 Tsd. Geotweets zufällig ausgewählt. Tabelle 3 zeigt die Anzahl der Geotweets in jeder der analysierten UTG.

Die Textinhalte dieser Geotweets werden in drei Schritten in ein Rasterformat überführt: (1) Projektion der Textinhalte einen numerischen Merkmalsraum mittels feature-Embeddings, (2) Dimensionsreduktion dieses Merkmalsraums auf wenige Variablen und (3) räumliche Extrapolation dieser Variablen in Raster.

⁶⁸ 1. Januar 2010 – 1. Januar 2020

Zunächst werden die Textdaten mittels Embeddings in einen numerischen Merkmalsraum überführt. Hierfür wurde ein multilinguales Modell ⁶⁹ aus der sentence-transformer Familie genutzt, welches auch im weithin genutzten BERTopic Ansatz für mehrsprachige Texte empfohlen wird⁷⁰. Dieses Modell ist darauf trainiert, in mehr als 50 Sprachen Texte in von 768-dimensionale Embedding-Vektoren zu konvertieren.

Daraufhin wird Dimensionsreduktion angewendet um diese Vektoren von 768 auf 5 Dimensionen zu reduzieren, in welchen die größtmögliche Varianz in wenigen Variablen eingefangen wird (McInnes et al. 2018). Dieser Schritt ist theoretisch optional, praktisch aber notwendig, da die hohe Anzahl an Dimensionen die anschließende Prozessierung prohibitiv verlangsamen würde. Als Algorithmus für die Dimensionsreduktion wird UMAP⁷¹ genutzt, welches auch im weit verbreiteten BERTopic Ansatz zur Themenmodellierung genutzt wird, der später auch für die Interpretation der Daten eingesetzt wird.

In dem dritten Schritt werden die nun punktförmig vorliegenden numerischen Variablen räumlich extrapoliert, um ein flächenhaftes Raster zu generieren. Wie in Studie 1 werden hierfür nur Geotweets mit der nötigen räumlichen Präzision genutzt, also Koordinatenreferenzen oder Ortsreferenzen deren *place_type* dem Punkt entspricht.

Die Extrapolation erfolgte mittels inverser Distanzwichtung⁷² (engl. *inverse distance weighting*, IDW) über die in R implementierte *terra* Bibliothek (Hijmans 2025).

5.2.2. Clustering

Daraufhin folgt der zentrale Schritt des Ansatzes, welcher die bebauten Pixel anhand der Merkmale in ähnliche Gruppen (Cluster) zusammenfasst. Für das Clustering wird die von Chavent et al. (2018) entwickelte Methode, welche das räumlich restringierte Clustern mit einer Variable erlaubt, auf drei Variablen ausgeweitet. Die maßgebliche Anpassung in der Implementierung ist, dass nicht zwei Distanzmatrizen mit einem Mischparameter α gemischt, sondern drei Distanzmatrizen mittels individueller Gewichte β addiert werden.

Aufgrund der Größe der Raster ist eine Aggregation im Raum notwendig, da sonst die Prozessierung prohibitiv anspruchsvoll ist. Benachbarte bebaute Pixel werden daher zusammengefasst, bis sich Blocks von etwa 100m Kantenlänge ergeben. Dabei werden die Embedding-Vektoren und das Bebauungsalter der bebauten Pixel gemittelt.

Hieraus werden drei Distanzmatrizen generiert:

- Dist_{FE} , welche dem Unterschied im Alter der Bebauung entspricht.
- Dist_{TW} , welche der Distanz im 5-dimensionalen Merkmalsraum der reduzierten Text-Embeddings entspricht.
- Dist_{GEO} , welche der räumlichen Distanz entspricht.

⁶⁹ sentence-transformers/paraphrase-multilingual-mpnet-base-v2

⁷⁰ https://www.sbert.net/docs/sentence_transformer/pretrained_models.html#semantic-similarity-models

⁷¹ UMAP wird über das Python Paket UMAP aufgerufen, und mit einer Anzahl von 15 Nachbarn, ohne Minimumdistanz und einer Kosinusmetrik parametrisiert.

⁷² Mittels der `terra::interpIDW` Funktion. Als Parameter wurden `maxPoints=200` und `radius=10` gewählt.

Diese Distanzmatrizen werden auf 0-1 skaliert, gewichtet, und dann durch Addition in eine $\text{Dist}_{\text{Gesamt}}$ Distanzmatrix zusammengefasst:

$$\text{Dist}_{\text{Gesamt}} = \beta_{\text{FE}} * \text{Dist}_{\text{FE}} + \beta_{\text{TW}} * \text{Dist}_{\text{TW}} + \beta_{\text{GEO}} * \text{Dist}_{\text{GEO}}$$

Hierauf wird dann hierarchisches Clustering mittels der in R implementierten Funktion `hclust`⁷³ (R Core Team 2024) angewendet um eine hierarchische Struktur in Form eines Dendrogramms zu generieren. Die Anzahl der Cluster k , in welche dieses Dendrogramm unterteilt wird⁷⁴, kann frei gewählt werden. In den hier präsentierten Auswertungen wurde nach vorangegangenen Experimenten die Anzahl $k = 12$ gewählt, welche die lokalen Kontexte gut repräsentiert und sich gut visualisieren lässt.

Gleichsam lässt sich über die Gewichtung steuern, welchen Einfluss auf die Raumstrukturierung die verschiedenen Komponenten haben sollen. Eine höhere Gewichtung von Dist_{GEO} führt zu kompakteren Clustern, also Clustern, welche intern homogen in Bezug auf ihren Ort sind. Eine höhere Gewichtung von Dist_{TW} und Dist_{GEO} führt dazu, dass sich die Raumstrukturierung an die Verteilung der respektiven Eingangsdaten annähert. So würde ein Clustering, bei welchem die volle Gewichtung auf den Textinhalten läge ($\beta_{\text{GEO}} = 0\%$, $\beta_{\text{FE}} = 0\%$, $\beta_{\text{TW}} = 100\%$) die interne Homogenität der Textinhalte innerhalb jedes Clusters maximieren, die räumliche Distanz und Unterschiede im Baujahr jedoch unberücksichtigt lassen.

5.2.3. Auswertung

Ziel des Ansatzes ist es nicht, Kompaktheit oder Homogenität einer einzelnen der Eingangsdaten zu maximieren, sondern einen Kompromiss zwischen den verschiedenen Eingangsdaten und der geographischen Kompaktheit herzustellen –in der Annahme, dass die hieraus entstehenden räumlichen Cluster eine sowohl physische als auch textinhaltliche Aspekte aufweisen, und damit interpretierbar sind.

Um auszuwerten, inwiefern dieses Ziel erreicht wird, wird auf zwei verschiedene Methoden vorgegangen: zunächst wird eine qualitative Interpretation versucht, zweitens wird die Güte der Cluster statistisch bewertet.

Erstens wird eine explorative Interpretation simuliert, welche zum Ziel hat, Cluster zu erstellen und ihre Funktion kurz zu beschreiben. Dies soll der in 3.2 konzipierten Anwendung nahekommen. Wie in 3.2 dargelegt, sollen dabei einfache explorative Statistiken auf Clusterbasis einfließen. Zunächst werden die Gewichtungen experimentell variiert, bis ein Mischverhältnis gefunden wird, dessen Outputs (Cluster) einem für die Autoren intuitiven Konzept einer Nachbarschaft, eines Baublocks oder Bauprojekts nahekommen.

Für diese Cluster werden dann aus den Daten eine Reihe von Statistiken abgeleitet und visualisiert, welche sich auf die Bebauung und die darin liegenden Textinhalte beziehen:

Bebauungsjahr: Die Verteilung des Bebauungsjahres innerhalb eines Clusters wird als Kurve dargestellt.

Sprachen: Die Anteile der Sprachen, in welchen die in den Clustern verorteten Geotweets verfasst sind. Hierfür wird auf die Metadaten der Tweets zurückgegriffen, welche im Feld *lang* einen Sprachcode enthalten.

Themen: Basierend auf dem weit verbreitete Topic-Modeling Ansatz BERTopic (Grootendorst

⁷³ Die Funktion `hclust()` implementiert den Ward 1-Algorithmus (Murtagh and Legendre 2014).

⁷⁴ Zur Teilung des Dendrogramms with die Funktion `cutree()` genutzt.

2022) werden die Tweets in Themen eingeteilt. Eine typische Anwendung von BERTopic basiert auf der Abfolge von Embedding, Dimensionsreduktion und Clustering. Embedding und Dimensionsreduktion werden bereits an vorhergegangener Stelle in der Prozessierung durchgeführt und sind in 5.2.1 beschrieben. Die Tweet-Texte werden dann durch Anwendung des Clustering-Algorithmus HDBSCAN⁷⁵ auf die dimensionsreduzierten Text-Embeddings in Themen gruppiert. Hierdurch werden alle Tweets einem Thema zugewiesen. Themen mit ähnlichen Inhalten werden daraufhin durch manuelle Vereinigung verschmolzen, und die Verteilungen der resultierenden Überthemen (*Supertopics*) innerhalb der Cluster aggregiert.

Zusätzlich wird mittels einer Webrecherche nach ortsspezifischen Textquellen gesucht, welche Information über die Orte in den Clustern liefern. Von Interesse sind hierbei beispielsweise Projektpläne, Nachrichten, oder Reviews von lokalen Unternehmen.

Falls mit diesen Mitteln eine qualitative Unterscheidung der Cluster vorgenommen werden kann, wird angenommen, dass die Methode zumindest grundsätzlich funktionsfähig ist. Dies stellt selbstverständlich nur eine von vielen Arten dar, wie Cluster interpretiert werden können. In der praktischen Projektarbeit könnten auch Umfragen, sozioökonomische Variablen und andere vergleichbare Quellen mit integriert werden.

Zweitens wird die Güte der Cluster statistisch bewertet: Für jede der drei Dimensionen quantifiziert, inwiefern die erstellten Cluster intern homogen in Bezug auf die respektiven Inputs sind. Hierfür werden folgende Maße genutzt:

Bebauungsjahr: Die Standardabweichung des Bebauungsjahres innerhalb des Clusters wird für jedes Cluster berechnet und über alle Cluster gemittelt, wobei Cluster nach ihrer Fläche gewichtet werden.

Raum: Die räumliche Metrik *Landscape Shape Index (LSI)* aus dem *landscapemetrics* R Bibliothek genutzt, um die Aggregation der Cluster zu beschreiben. LSI erfasst die Aggregation mittels der Ratio zwischen der Kantenlänge eines Clusters und der theoretisch minimalen Kantenlänge des Clusters bei maximaler Kompaktheit (in der Regel ein einzelner kreisförmiger Patch ohne Löcher, Hesselbarth et al. 2019)

Textinhalte: Mittels der Shannon-Entropie (Shannon 1948) wird quantifiziert, inwiefern die Cluster homogen in Bezug auf die oben genannten Text-Themen sind. Eine große Homogenität kann so interpretiert werden, dass in den Clustern keine große Mischung von Themen auftaucht, sondern nur wenige Themen pro Cluster und diese dafür jeweils mit großer Häufigkeit. Dies lässt sich mittels der Shannon-Entropie quantifizieren. Die mit H bezeichnete Entropie wird über alle Cluster hinweg gemittelt, wobei Cluster nach ihrer Fläche gewichtet werden. Dieser Wert wird in Verhältnis zur theoretischen Entropie gesetzt, welche sich ergeben würde, wenn kein Clustering vorgenommen würde. Dieses Verhältnis dient als Kriterium für die Güte des Clusterings in Bezug auf die Homogenität der Themenverteilung.

⁷⁵ HDBSCAN wird über das in Python implementierte Paket HDBSCAN aufgerufen und wie folgt parametrisiert: `min_cluster_size=Anzahl der Tweets / 250`, `min_samples=15`, `metric = „euclidean“`, `clster_selection_method=“leaf“`.

Anhand dieser Kriterien wird dann verglichen:

- Das oben beschriebene Clustering, welches die vorgeschlagene vereinigende Lösung darstellt, die wie oben beschrieben auch qualitativ interpretiert wird.
- Ein Clustering, das rein auf Textinhalten basiert ($\beta_{\text{GEO}} = 0$, $\text{Dist}_{\text{FE}} = 0$, $\text{Dist}_{\text{TW}} = 1$).
- Ein Clustering, das rein auf Baujahren basiert ($\beta_{\text{GEO}} = 0$, $\text{Dist}_{\text{FE}} = 1$, $\text{Dist}_{\text{TW}} = 0$).
- Ein Clustering, das rein auf räumlicher Distanz basiert ($\beta_{\text{GEO}} = 1$, $\text{Dist}_{\text{FE}} = 0$, $\text{Dist}_{\text{TW}} = 0$).

Wenn Clustering (a) in allen drei Kriterien mindestens den zweithöchsten Rang erreicht, wird dies als Verbesserung gegenüber der Fokussierung auf ein einzelne Art von Distanz gesehen. Dies kann als die Fähigkeit des Ansatzes zur Generierung räumlicher Einheiten interpretiert werden, die zur Aggregation verschiedener Datentypen im urbanen Raum sinnvoll sind.

5.3. Studie B1: Ableitung inhaltlicher Präferenzen von nigerianischen Twitter-Nutzenden nach Mobilitätscharakteristika

Studie B1 hat zum Ziel, sowohl die Mobilität der Twitter-Nutzenden als auch ihre Mobilität kategorisch einzuordnen und in Bezug zu setzen. Hierfür wird ein Ansatz entwickelt, welcher zunächst Mobilitätscharakteristika aus Geotweet-Trajektorien der Nutzenden ableitet, und daraufhin auch die Anteile verschiedener Themen an den von den Nutzenden erstellten Textinhalten misst.

Abbildung 22 illustriert den Arbeitsablauf der Studie, bei welchem die Textanalyse auf der räumlichen Analyse aufbaut. Zunächst wird auf einem Datensatz georeferenzierter Tweets eine Trajektorienanalyse durchgeführt, um für die Analyse geeignete Accounts zu erkennen, für welche weitere Daten abgefragt werden. Um die Textinhalte dieser Accounts zu klassifizieren, wird ein Klassifizierungsmodell auf Daten des Nairaland Webforums trainiert und auf die Tweets adaptiert (5.3.2). Die Klassifizierung wird mittels dreier Plausibilitätschecks evaluiert (5.3.3) und statistisch ausgewertet (5.3.4).

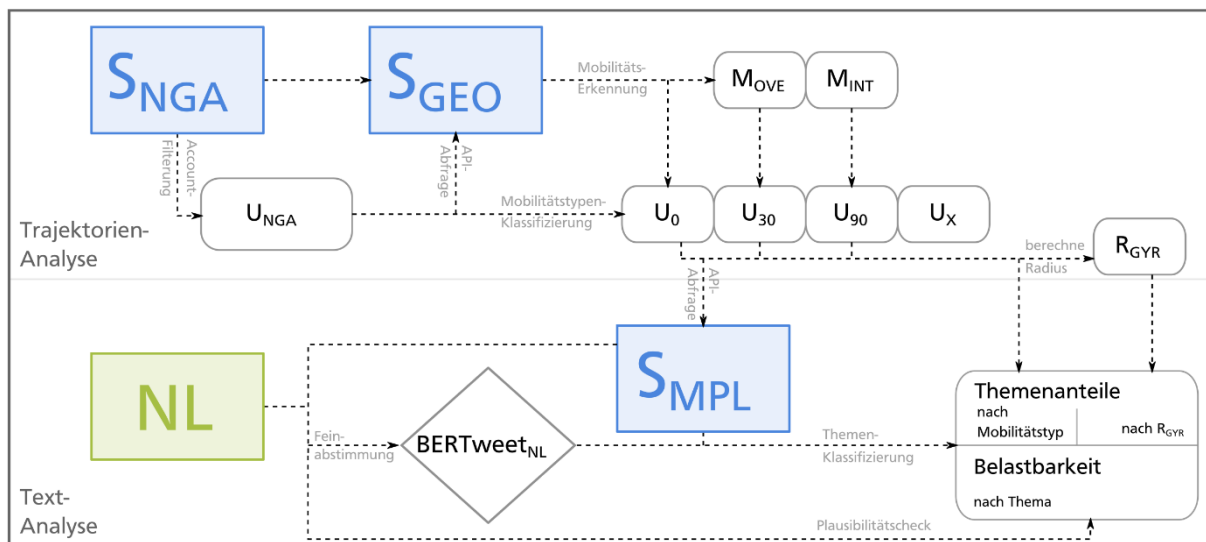


Abbildung 22: Arbeitsablauf von Studie B1. Oben die räumliche Verarbeitung und Trajektorienanalyse, unten die Textverarbeitung und Textanalyse. S_{NGA} , S_{GEO} und S_{MPL} sind Twitter-Datensätze. NL ist ein Nairaland-Datensatz (4.5.2). M_{OVE} und M_{INT} sind beobachtete Änderungen im Aufenthaltsort. U_{NGA} ist ein Satz von Accounts, von denen U_0 , U_{30} , U_{90} und U_X Teile (Mobilitätstypen) darstellen. R_{GYR} ist der Gyrationsradius auf Account-Ebene.

5.3.1. Initiale Datenakquise und Trajektorienanalyse

Alle für den Analysezeitraum (01.01.2015 – 31.12.2018) in der Fokusregion Nigeria verfügbaren Geotweets werden mit der API abgefragt. Aus diesem Set S_{NGA} werden Nutzende identifiziert, welche für die weitere Analyse infrage kommen. Zunächst wird eine Reihe von Filtern angewendet, um untypische Accounts zu entfernen, ähnlich wie in den Werken von Hübl et al. (2017), Petutschnig et al., (2020) und Lemoine-Rodríguez et al. (2024):

1. Accounts welche zwischen mehr als 25% ihrer Tweets unrealistische Geschwindigkeiten ($>150\text{km/h}$) zwischen zwei aufeinanderfolgenden Orten aufwiesen. Hierdurch wurden Accounts entfernt, welche die Geotagging-Funktion nicht nutzen, um ihre eigene Position anzugeben, sondern beispielsweise einen Ort, den sie kommentieren oder zu besuchen planen.
2. Accounts welche mehr als 150 Geotweets an einem Tag (Hübl et al. 2017) oder durchschnittlich mehr als 15 pro Tag posten (Petutschnig et al. 2020). Hierdurch wurden automatisierte Accounts oder außergewöhnlich aktive Accounts von der Analyse ausgeschlossen.
3. Zuletzt werden zur Verbesserung der Konsistenz der Datenbasis nur die Accounts ausgewählt, welche bereits zu Beginn des Untersuchungszeitraums existierten.

Für die sich ergebenden nigerianischen Accounts U_{NGA} werden zusätzlich alle außerhalb Nigeria liegenden Geotweets abgerufen, um deren räumliche Trajektorien S_{GEO} zu vervollständigen. Diese werden dann für die Trajektorienanalyse genutzt. Da kleinräumige Mobilität nicht relevant für die konzipierte Anwendung (3.3) ist, werden Tweets räumlich größeren Raumeinheiten (RE, siehe Abbildung 12) zugewiesen.

Die Definition dieser RE erfolgte mit Rücksicht auf das konzeptionellen Ziel der zu unterscheidenden Mobilitätstypen: In Nigeria Stationäre, Mobile und International Migrierte. Die Wahl und Definition dieser Typen basiert auf den UN-Definitionen für Kurzzeit- und Langzeitmigration und dem Fokus der Studie auf Nigeria. So war Mobilität von Emigrierten außerhalb Nigerias und innerhalb ihres Ziellandes nicht relevant, Mobilität zwischen Städten innerhalb Nigerias schon. Die RE wurden daher außerhalb Nigerias auf Landesebene definiert und innerhalb Nigerias als hexagonales Raster, wobei um bedeutende Städte jeweils ein kreisförmiges RE ausgespart wurde. Details für die Ableitung der RE finden sich im Anhang (A5).

Tweets von S_{GEO} wurden also mittels ihrer Geotags in die RE verortet. Hierbei wurden Ortsreferenzen der Typen Punkt, Nachbarschaft und Stadt genutzt, Koordinatenreferenzen jedoch bevorzugt, falls vorhanden (siehe Kapitel 4.5.1). Für Ortsreferenzen nach Punkt und Nachbarschaft wurde jeweils der Mittelpunkt des annotierten Ortes genutzt. Daraufhin wurde eine Trajektorienanalyse durchgeführt, um Accounts in verschiedene Mobilitätstypen zu unterscheiden. Hierfür wurde der von Chi et al. (2020) entwickelte segmentbasierte Ansatz genutzt, da dieser unter anderem auf Twitter-Trajektorien kalibriert und validiert wurde. Für die hier beschriebene Studie wurde der Algorithmus leicht adaptiert, sodass von mehreren einander überschneidenden Segmenten eines beibehalten wird, wenn es mehr als 70% der Tweets im Zeitraum der Überschneidung abdeckt.

Die Anwendung auf S_{GEO} führt zu kontinuierlichen, einander zeitlich nicht überschneidenden Segmenten, welche die Aufenthaltshistorie des Accounts approximiert. Innerhalb dieser segmentbasierten Historie jedes Accounts werden zwei Arten von Migrationsereignissen identifiziert (Abbildung 23). Diese Migrationsereignisse werden wie folgt definiert:

1. Internationale Migration (M_{INT}) welche zwischen zwei Aufenthalten von mindestens 90 Tagen liegen, von welchen nur eines an einem Ort innerhalb Nigerias liegt. Die Wahl von 90 Tagen entspricht der typischerweise genutzten Unterscheidung zwischen Besuchern und längeren Aufenthalten, wie sie beispielsweise in Visaabkommen festgelegt ist (Armstrong et al. 2021) und entspricht der UN-Definition von Kurzzeitmigration mit mindestens 3 Monaten Aufenthalt⁷⁶ (United Nations Department of Economic and Social Affairs 1998).
2. Generelle Mobilität ($MOVE$) welche zwischen zwei beliebigen Aufenthaltsorten mit jeweils mindestens 30 Tagen Aufenthaltsdauer erfolgt. Dieser Typ von Migrationsereignissen soll eine Spanne verschiedener regionaler Mobilitätsformen abdecken.

Anhand dieser Ereignisse werden Accounts in die drei Klassen eingeteilt, je nachdem, welche Migrationsereignisse in ihrer Historie vorkommen. Der UN-Definition von Migration (United Nations Department of Economic and Social Affairs 1998) folgend entspricht dies den Definitionen von Kurzzeitmigration (3–12 Monate Aufenthalt) und Langzeitmigration (mehr als 12 Monate Aufenthalt).

1. **International Migrierte** (U_{90}): Accounts, deren Historie mindestens ein M_{INT} aufweist.
2. **Mobile** (U_{30}): Accounts, deren Historie kein M_{INT} aber mindestens ein $MOVE$ aufweist.
3. **Innerhalb Nigeria Stationäre** (U_0): Accounts, deren Historie kein M_{INT} und kein $MOVE$ aufweist, aber mehrere Aufenthalte von mindestens 30 Tagen ($ResLength = 30$) welche alle an demselben Ort innerhalb Nigerias liegen. Um zu vermeiden, dass Inaktivität als Immobilität interpretiert wird (Schewel 2020), müssen diese Accounts zusätzlich mindestens so viele Tweets aufweisen, wie das Minimum über alle Accounts in den U_{90} und U_{30} Gruppen.
4. **Undefinierte** (U_x): Accounts, welche keine dieser Kriterien erfüllen.

5.3.2. Textklassifizierung mittels Domain-Adaptation

Für Accounts in den Gruppen U_{90} , U_{30} und U_0 wird eine Textanalyse durchgeführt um die Textinhalte in Kategorien eingeordnet analysieren zu können. Wie in Kapitel 2.2.2 dargelegt, sind Geotweets allein hierfür möglicheres nicht sinnvoll, da sie nur einen Ausschnitt darstellen, welcher wahrscheinlich nicht repräsentativ für die Gesamtmenge der von den Accounts geposteten Tweets ist. Der vollständige Abruf dieser Gesamtmenge ist technisch möglich, aber aufgrund der großen Menge nicht innerhalb des zur Verfügung stehenden API-Zugangs realisierbar. Es soll daher eine möglichst repräsentative Stichprobe der gesamten Twitter-Historie generiert werden.

Die zur Verfügung stehende Version der Twitter-API unterstützt das zufällige Ziehen von Stichproben nicht. Bei Abfragen werden immer deterministisch die neuesten Tweets, welche die Kriterien erfüllen, zurückgegeben. Daher wird ein pseudo-Stichprobenansatz entwickelt. Über den Analysezeitraum werden 48 zufällig verteilte Stichproben-Zeitspannen von 7 Tagen Länge definiert, was im Durchschnitt einer Zeitspanne pro Monat entspricht. Innerhalb dieser Zeitspannen werden bis zu 10 Tweets pro Account abgefragt. Auch wenn die Tweets auch dann nicht komplett zufällig in der Zeit verteilt sind, werden mit diesem pseudo-zufälligen Stichprobe zeitliche Verzerrungen minimiert, welche den Inhalt der auszuwertenden Tweet

⁷⁶ Die Vereinten Nation sprechen bei Aufenthalt von 3-12 Monaten von Kurzzeitmigration und bei mehr als 12 Monaten von Langzeitmigration.

Texte beeinflussen könnten. Der Satz der sich ergebenden Tweets S_{MPL} wird für alle folgenden Textanalysen von Studie B1 genutzt.

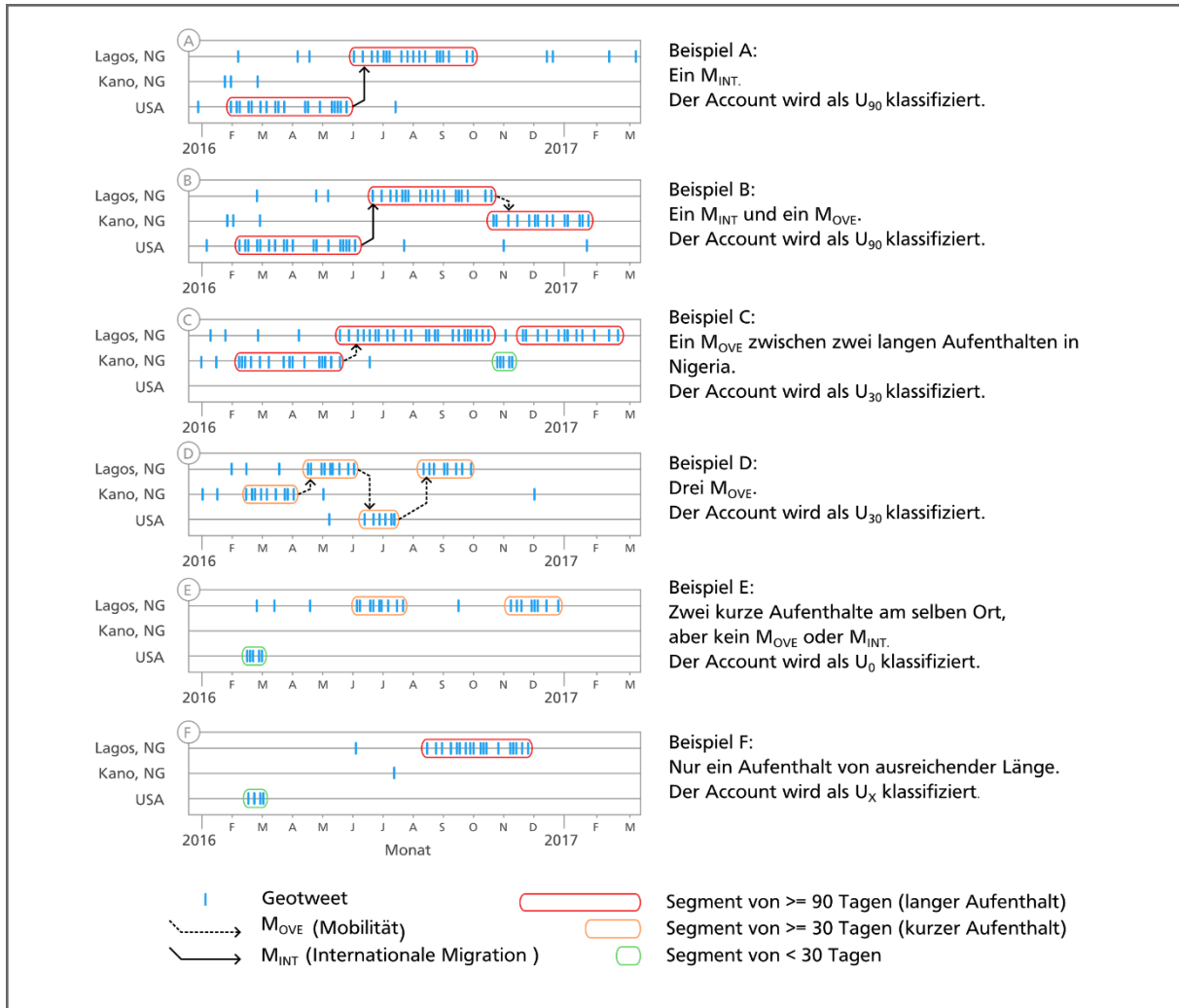


Abbildung 23: Illustration der identifizierten Mobilitätstypen anhand sechs fiktiver Trajektorien. Nachdem Segmente identifiziert wurden, werden Abfolgen von Segmenten an unterschiedlichen Orten als Migration gewertet. Danach werden Accounts entweder als U_0 , U_{30} , U_{90} oder U_x eingestuft. NG: Nigeria; US: Vereinigte Staaten von Amerika.

Um diese Tweets in Themen zu klassifizieren, wird maschinelles Lernen genutzt, um einen Textklassifizierer auf Daten des Nairaland-Webforums (4.5.2) zu trainieren, welche bereits in Themen strukturiert sind. Der Klassifizierer wird daraufhin auf S_{MPL} angewendet, um auch die Tweets in Themen zu kategorisieren. Aus dem Nairaland-Forum werden für diese Studie 40 Unterforen ausgewählt (siehe Appendix A1). Die Anzahl der Nairaland Kommentare beträgt 2 091 491. Aufgrund von technischen Unterschieden der Plattformen kann von einem Domänenunterschied (engl. domain shift) ausgegangen werden. Die Aufgabe besteht daher in der Multi-label Klassifizierung mit der Herausforderung eines Domänenunterschieds zwischen der Quelldomäne Nairaland und der Zieldomäne S_{MPL} .

Um diesen Domänenunterschied zu reduzieren, werden Texte in S_{MPL} und Nairaland identisch vorprozessiert, wobei plattformspezifische Elemente wie der retweet-Marker RT und hochspezifische Textelemente wie URLs⁷⁷ und Telefonnummern entfernt wurden. Wie im

⁷⁷ Uniform Resource Locators (URLs) werden in fast allen Fällen als Links zu Webseiten benutzt

ursprünglichen Training von BERTweet werden Texte, welche kürzer als 10 Tokens sind, von der Analyse ausgeschlossen.

Architektur des Modells: Als Basis (engl. *backbone*) des Modells diene BERTweet (Nguyen et al. 2020), ein auf der Transformer-Architektur (Vaswani et al. 2017) basierendes tiefes Sprachmodell, welches auf einem großen Twitter-Korpus vortrainiert wurde. BERTweet konvertiert die Textinputs in einen Embedding-Vektor der Länge 768. Die Architektur wird durch einen Themenklassifizierer abgeschlossen, welcher diese Embeddings akzeptiert und in einer vollständig verbundenen Ebene (engl. *fully connected layer*) die Klassifizierung in die 40 Klassen vornimmt.

Zur Implementierung des domänenadversarialen Ansatzes wird dieses Klassifizierungsmodell durch einen sekundären Domänenklassifizierer ergänzt. Dieser akzeptiert gleichsam die Embedding-Vektoren und klassifiziert sie mittels zweier vollständig verbundener Layer in die zwei Domänen Twitter und Nairaland. Dabei ist diesem Domänenklassifizierer ein umkehrendes Layer (engl. *reversal layer*) vorgeschaltet, welches bei der Rückwärtspropagierung⁷⁸ eine Inversion des Fehlergradienten vornimmt. Bei jedem Trainingsschritt werden neben den zu klassifizierenden Texten der Quelldomäne auch Texte der Zieldomäne eingegeben, welche nur in diesem Domänenklassifizierer verarbeitet werden. So wird erreicht, dass domänenspezifische Features, welche für die Domänenklassifizierung sinnvoll, für die Anwendung jedoch unerwünscht sind, beim Training des Backbone-Modells benachteiligt werden.

Das gesamte Netzwerk wird mittels der Nairaland-Daten und Twitter-Daten trainiert, wobei 80% der Nairaland Kommentare zum Training und 10% zur Validierung genutzt werden, sowie 10% für eine finale Evaluation zurückgehalten werden. Alle Twitter-Daten in *SMPL* werden als Beispiele für die Zieldomäne genutzt. Training erfolgt mit einer Dropout-Rate von 0,5. Da die Anzahl von Kommentaren zwischen den Nairaland-Themen unausgeglichen war, wird klassengewichtete Kreuzentropie als Verlustfunktion genutzt (Torch Contributors 2019).

Das Training erfolgt, bis keine weitere Verbesserung der Validierungsgenauigkeit in der Themenklassifizierungsaufgabe stattfindet. Dann wird das Modell auf alle Tweets in *SMPL* angewendet, um diese den 40 Themen zuzuordnen.

5.3.3. Plausibilitätsanalysen

Da keine annotierten Daten für die Zieldomäne *SMPL* vorliegen, kann das Modell nicht anhand von traditionellen quantitativen Validierungsmetriken evaluiert werden. Stattdessen wird in dieser Studie ein aus drei Komponenten bestehender Plausibilitätscheck durchgeführt, welcher für jedes der 40 Themen die Güte der Klassifikation in Bezug auf Übereinstimmung zwischen den Inhalten eines Themas im algorithmisch annotierten Twitter-Datensatz und den Inhalten des korrespondierenden Nairaland-Unterforums bewertet.

Die erste Komponente ist die themenspezifische Genauigkeit in der Quelldomäne Nairaland, welche mittels des F1-Wertes auf den zurückgehaltenen Nairaland Testdaten berechnet wurde. Die Annahme ist, dass wenn ein Thema bereits in der Quelldomäne nicht akkurat bestimmt werden kann, dies auch in der Zieldomäne unwahrscheinlich ist.

⁷⁸ Hiermit wird das Verfahren bezeichnet, durch das das Modell die während des Trainings gemachten Fehler auf seine Parameter zurückverfolgt und damit sozusagen aus seinen Fehlern lernt. Durch das umkehrende Layer wird dieser Lernprozess ins Gegenteil verkehrt.

Der F1-Wert wurde wie folgt berechnet:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

wo TP = richtig Positive, FP = falsch Positive,
 FN = falsch Negative.

Die zweite Komponente ist die Übereinstimmung zwischen Wochentags-typischen Aktivitätsmustern, welche viele der Themen aufweisen (Abbildung 24). Die Annahme ist, dass wenn ein Thema in der Zieldomäne ein ähnliches wochentag-typisches Aktivitätsmuster aufweist wie in der Quelldomäne, es in der Zieldomäne akkurat repräsentiert ist. Zur Quantifizierung dieser Übereinstimmung wurde der Pearson Korrelationskoeffizient berechnet (Leydesdorff 2005). Themen, für welche ein Variationskoeffizient von $<0,5$ andeutete, dass keine Wochentags-typischen Aktivitätsmuster existieren, wurden ausgespart (Beispiel: *Gaming* in Abbildung 24)

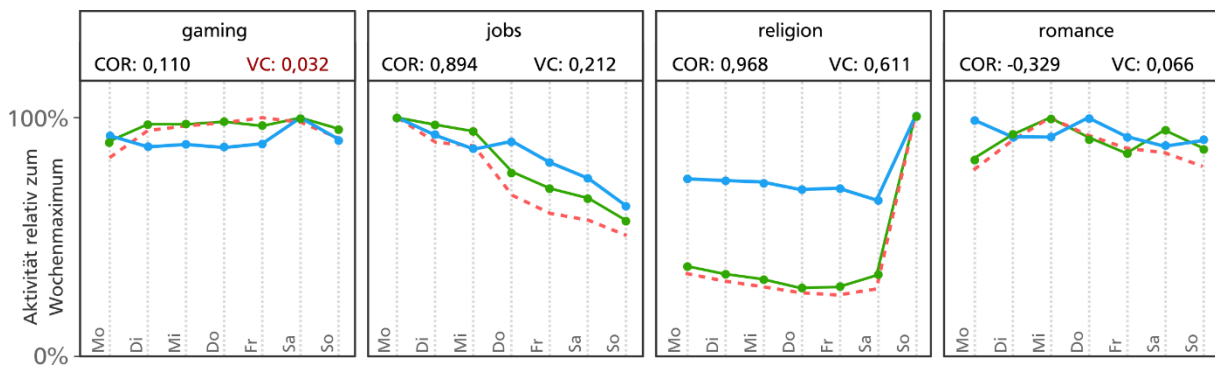


Abbildung 24: Wöchentliche Aktivitätsmuster verschiedener Themen. Grün: Tatsächliche Label aller Nairaland-Posts. Rot: Inferierte Label auf Nairaland Testdaten. Blau: Inferierte Label auf Twitter Daten. Die Kurven sind jeweils auf das Wochenmaximum skaliert. CV: Variationskoeffizient. COR: Korrelation zwischen Aktivitätsmustern von Nairaland (Grün) und Twitter (Blau).

Die dritte Komponente des Plausibilitätschecks basiert auf Übereinstimmung der Quell- und Zieldomänen in der Verteilung von Neologismen. Die Annahme ist, dass im Laufe der Zeit neue Begriffe und Konzepte auftauchen (wie *covid-19*), welche in den Trainingsdaten nicht auftauchen, aber klar einem Thema angehören. Es werden kleinere Hilfs-Datensätze von Nairaland (N_{LNEU}) und Twitter (S_{NEU}) aus einer auf den Analysezeitraum folgenden Zeit (2021-2022) akquiriert und S_{NEU} wird mittels des auf älteren Daten trainierten Modells in Themen klassifiziert. Aus N_{LNEU} werden Neologismen identifiziert, welche in N_{LNEU} auftraten, im älteren Nairaland-Datensatz jedoch nicht. Tauchen diese Neologismen (wie *covid-19*) in N_{LNEU} und S_{NEU} in denselben Themen auf, ist das ein Anzeichen dafür, dass die Themen über die Domänen hinweg übereinstimmen. Um die Ähnlichkeit für jedes Thema zu quantifizieren, werden die Themenverteilungen der Neologismen vektorisiert und als Basis benutzt, um die Kosinus-Ähnlichkeit zwischen den Vektoren zu berechnen (Egghe und Leydesdorff 2009).

Um diese drei Komponenten interpretierbar zu machen, werden sie in vier diskrete Plausibilitätsgrade von A (belastbar) bis D (wenig belastbar) eingestuft, je nachdem ob sie in allen drei Kriterien die Schwellwerte von 0,5, 0,4, 0,3 oder nur das Minimum von -1 übertreffen.

Zusätzlich zu dieser quantitativen Analyse soll eine qualitative Textanalyse für eine Auswahl von Themen durchgeführt werden. Ziel hierbei ist, Plausibilität einzuschätzen, Fehler zu identifizieren und die Vielfalt innerhalb der Themen zu verstehen. Die Themen, für die dies durchgeführt wird, sollen das mutmaßlich migrationsrelevante Thema *Travel* sein, sowie die beiden Themen, welche größte Unterschiede zwischen Nutzergruppen aufweisen. Für jedes dieser drei Themen werden 100 Tweets pro Mobilitätstyp zufällig aus S_{MPL} ausgelesen. Diese 900 Tweets werden basierend auf den Textinhalten und ggf. irgendwelchen verlinkten Inhalten manuell kodiert. Der Forschungsstil der *grounded theory* folgend (Charmaz 2006; Strübing 2019) wird hier nicht von vordefinierten Hypothesen ausgegangen, sondern Codes in einer explorativen Weise von den Dateninhalten aus entwickelt, wobei Migration und Mobilität immer einen Fokus in der Interpretation darstellten. In anderen Worten ist das Ziel keine universelle Rahmen von Themen, Emotionen, Konzepten oder Stilen über alle Themen hinweg, sondern ein spezifisches Codeset für jedes Thema, welches die Variation innerhalb des Themas bestmöglich beschreibt.

5.3.4. Statistische Auswertung

Die sich aus der Mobilitätsanalyse ergebenden $MOVE$ und $MINT$ werden zusammengefasst, ihre Anzahl und geographische Verteilung wird umrissen. Auch die drei Nutzergruppen U_{90} , U_{30} und U_0 werden mittels deskriptiver Statistiken charakterisiert.

Der Fokus der Auswertung liegt jedoch auf der Analyse der Textinhalte über diese Mobilitätstypen hinweg. Für jeden der Accounts U_{AN} wird der Anteil eines Themas an seinen gesamten Tweets berechnet, wodurch sich eine diskrete Wahrscheinlichkeitsverteilung ergibt. Diese Verteilungen werden dann über alle Nutzende einer Mobilitätsklasse hinweg aggregiert.

Es ergibt sich erstens für jede der drei Mobilitätsklassen und jedes Thema eine diskrete Häufigkeitsverteilung, welche als Indikator des Interesses oder der Relevanz verschiedener Themen für diese Gruppe betrachtet werden kann. Von Interesse sind aber weniger die absoluten Anteile der Themen als die Unterschiede zwischen den Mobilitätstypen. Diese werden mittels eines Kruskal-Wallis Tests erfasst (Kruskal und Wallis 1952). Ein signifikanter Test kann andeuten, dass ein Thema von besonderer Relevanz für einen bestimmten Mobilitätstyp ist, und somit als Grundlage für weiterführende Analysen genommen werden.

5.4. Studie B2: Analyse der Verbreitung von Sprachmerkmalen nach Mobilitätscharakteristika der Autoren am Beispiel westafrikanischer Twitter-Nutzender

In Studie B2 werden lexikalische Sprachmerkmale, welche ortstypisch für Nigeria sind, quantifiziert und in Relation zu Ort, Themen und Nutzermobilität gesetzt. Hierzu werden zunächst zwei Datensätze von Tweets westafrikanischer Twitter-Nutzenden zusammengestellt (5.4.1). Eine Liste von Sprachmerkmalen wird erstellt, deren Verbreitung in Tweets dann algorithmisch erkannt und quantifiziert wird (5.4.2). Ähnlich wie in Studie B1 werden dann Informationen über Ort und Mobilität der Nutzenden sowie Themen der Tweets genutzt um diesmal Verbreitung der Sprachmerkmale zu kontextualisieren (5.4.3).

5.4.1. Datenakquise

Die Datenakquise erfolgt analog zu der in Studie B1 (Kapitel 5.3.1) beschriebenen Methode in mehreren Stufen: Wieder wird für den Analysezeitraum von 1. Januar 2015 bis 31. Dezember 2019 ein Datensatz (T_{ROI}) aller für die Untersuchungsregion verfügbarer Geotweets zusammengestellt, in welchem Kandidaten-Accounts (U_c) identifiziert werden. Für diese werden dann für den Analysezeitraum vollständige Geotrajektorien abgerufen (T_{GEO}) und in einer Trajektorienanalyse genutzt, um die Accounts in Mobilitätstypen zu klassifizieren.

Wesentliche Unterschiede bestehen in der Definition des Untersuchungsgebietes und dessen Raumaufteilung sowie den identifizierten Mobilitätstypen. Diese werden angepasst, um die Verteilung von Sprachmerkmalen supranational untersuchen zu können. Dem entsprechend wurden Daten nicht nur für Nigeria erhoben, sondern für eine erweiterte Untersuchungsregion (ROI), welche vier weitere westafrikanische Länder miteinschließt (Abbildung 25).

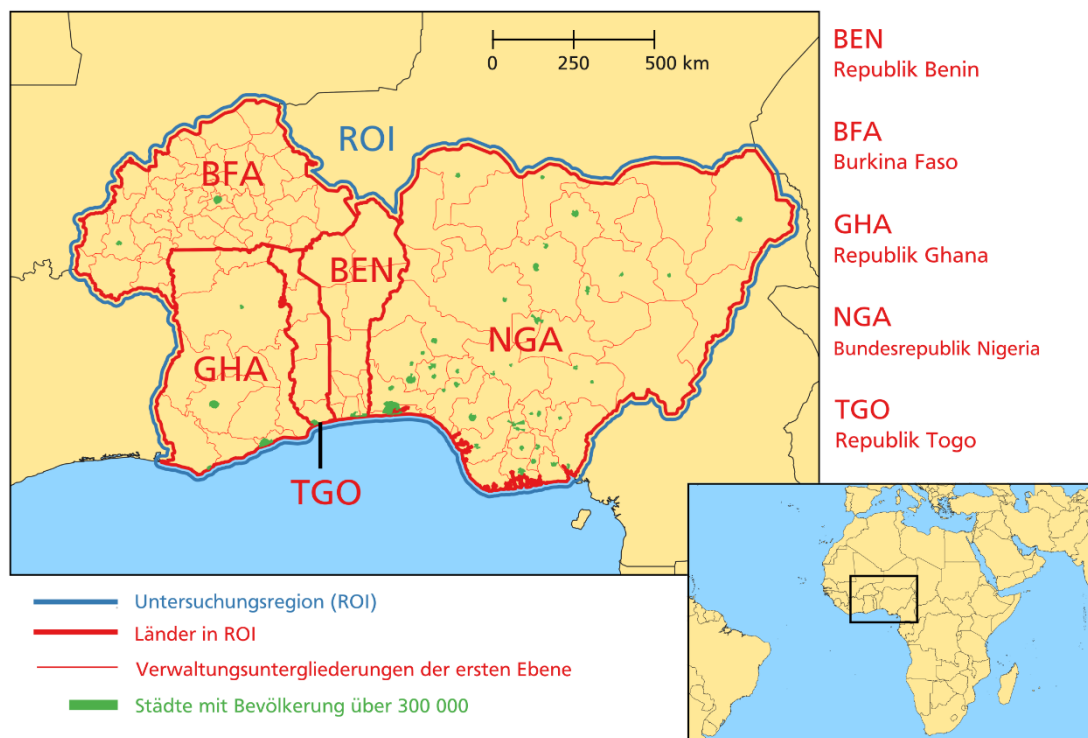


Abbildung 25: Geographische Untersuchungsregion der Studie B2

Als Raumeinheiten (RE) dienen in dieser Studie die höchstgelegenen Verwaltungsebenen der Länder⁷⁹. Außerhalb der Region werden Positionen von Geotweets auf Länderebene aggregiert. Nach Anwendung des segmentbasierten Algorithmus (siehe Kapitel 5.3.1) werden zunächst zwei Mobilitätsgrade unterschieden:

Mobile Accounts, welche Aufenthalte von 30 Tagen in verschiedenen RE aufweisen, und Stationäre Accounts, welche mehrere Aufenthalte von 30 Tagen alle in derselben RE und dazu mindestens 20 Tweets aufweisen.

⁷⁹ Staaten in Nigeria, Regionen in Ghana, Départements in Benin, Régions in Togo und Burkina Faso

Für die Accounts U_{AN} , welche klar einem dieser Mobilitätsgrade zugewiesen werden konnten, werden analog zu Studie B1 wieder Tweets in einem pseudo-zufälligen Stichprobenverfahren von der V2 der Twitter API abgerufen. Dies ergibt für alle Nutzende U_{AN} eine Reihe von Tweets T_{AN} , deren Texte analysiert werden.

Zusätzlich werden restriktivere Kriterien angewendet, um aus den Accounts in U_{AN} fünf speziellere Mobilitätstypen auszuwählen, welche sich auf Aufenthalte an bestimmten Orten beziehen:

NGA-Stationäre, deren Trajektorien nur Aufenthalte am selben Ort in Nigeria aufweisen,

NGA-Mobile, deren Trajektorien mehrere Aufenthalte an verschiedenen Orten in Nigeria aufweisen,

ROI-Mobile, deren Trajektorien mindestens einen Aufenthalt in Nigeria, und mindestens einen in einem anderen Land des ROI aufweisen,

GBR-Emigrierte, deren Trajektorien mindestens einen Aufenthalt in Nigeria, aber den Großteil der Aufenthaltszeit im Vereinigten Königreich aufweisen, und

NA-Emigrierte, deren Trajektorien mindestens einen Aufenthalt in Nigeria, aber den Großteil der Aufenthaltszeit in den Vereinigten Staaten haben.

5.4.2. Quantifizierung der Lokalen Form des Englischen in Nigeria

Kernkomponente der Studie ist die Identifikation von Sprachmerkmalen auf Tweet-Basis und aggregiert auf Account-Basis. Um eine lokale Färbung oder Form des Englischen in Nigeria (LEN) zu erkennen, wird ein lexikalischer Ansatz genutzt, welcher der Methodik folgt, die von Honkanen und Müller (2021) zur quantitativen Erfassung des Nigerianischen Pidgin English, einer lokalen Sprachvarietät, in Nairaland-Daten genutzt wurden.

Zunächst werden hierbei Begriffe aus den für Nigeria-spezifischen Web-Wörterbüchern *naijalingo* (Naija Lingo 2023) und *Babawilly's Dictionary of Pidgin English Words and Phrases* (Babawilly 2023) zusammengestellt. Zusätzlich wurde in dieser Studie das *Dictionary of Nigerian English* von Blench und Dendo (2005) herangezogen. Begriffe mit einer Länge von bis zu drei Worten werden in dieser Studie als Merkmale von LEN betrachtet. Auch wenn die Wörterbücher einige längere Begriffe enthalten, sind diese sehr spezifisch und werden daher ausgeschlossen.

Um für die Quantifizierung von LEN ein robustes Wörterbuch W_L zu erstellen, werden diese Eingangs-Wörterbücher durch Abstimmung kombiniert, wodurch nur Begriffe übernommen wurden, welche in mindestens 2 der 3 Wörterbücher vorkamen, oder durch manuelle Analyse in 400 Tweets identifiziert wurden.

Wie in der Arbeit von Honkanen und Müller (2021) werden Begriffe als bivalent markiert, welche auch in anderen Sprachen als der zu analysierenden LEN vorkommen. Neben Englisch wird in dieser Studie mit Französisch, Spanisch und Portugiesisch abgeglichen. Um diese Sprachen zu repräsentieren wurde der *OpenSubtitle Corpus* genutzt, welcher Filmuntertitel in zahlreichen Sprachen enthält (Hermit 2016).

Begriffe in W_L , die starken Bezug zu bestimmten Orten haben, wurden manuell identifiziert, als ortsspezifisch markiert und später in der Analyse von Orten (6.4.2) ausgeschlossen. Beispiele hierfür sind *unilag*, was die Universität von Lagos bezeichnet, oder *jankara*, was sich auf gefälschte Produkte oder einen bestimmten Markt in Lagos beziehen kann.

Gleichsam werden einige Begriffe als themenspezifisch markiert. Ein Beispiel ist *mallam*, was eine muslimische oder Hausa Person aber auch eine gebildete Autorität im Bereich der Wissenschaft oder Religion bezeichnen kann. Dies steht in Kontrast zu Begriffen wie *wahala*, was Probleme oder Ärgernisse im Allgemeinen und unabhängig vom Thema bezeichnet. Vorwiegend themenspezifische Begriffe wurden manuell identifiziert, entsprechend markiert und später in der Analyse von Themen besonders berücksichtigt.

Nachdem das kombinierte Wörterbuch W_L erstellt ist, werden seine Einträge mit den Texten der Tweets abgeglichen, um für jeden Tweet den LEN-Index (LI) zu bestimmen. Der LI ist der Anteil von Wörtern⁸⁰ in jedem Tweet-Text, welche den lokalen Begriffen in W_L entsprechen. Bivalente Begriffe wurden hierbei nur dann als lokale Begriffe betrachtet, wenn sie im jeweiligen Tweet in unmittelbarer Nähe (bis zu drei Wörter Abstand) zu einem eindeutigen lokalen Begriff auftauchen.

Das Wörterbuch hat keinen Anspruch auf Vollständigkeit, und daher kann der LI die wahre Häufigkeit von lokalen lexikalischen Merkmalen nur unterschätzen. Als eine Proxy-Variable kann er jedoch für Vergleiche zwischen Orten, Nutzenden und Themen genutzt werden.

Um die Verbreitung lexikalischer Merkmale mit der anderer Sprachmerkmale zu vergleichen, werden in einer manuellen Auslesung von 900 Tweets morphosyntaktische und pragmatische Features codiert. Dies stellt keinen Teil der Analyse dar, sondern soll angesichts der Komplexität von Sprachvariation der Einordnung der lexikalischen Ergebnisse dienen.

5.4.3. Kontextualisierung basierend auf Ort, Nutzermobilität und Themen

Der LI wird zunächst für jeden Tweet einzeln berechnet. Für die Auswertung wird er innerhalb von Gruppen aggregiert als Grundlage genutzt, die Verbreitung von LEN-Merkmalen zwischen Gruppen zu vergleichen. Diese Vergleiche erfolgen auf drei Ebenen: Orte, Mobilitätstypen der Nutzenden, und Themen. Abbildung 26 illustriert den Workflow der Studie.

Ort: Für eine Auswahl der bedeutendsten Städte der Region werden Geotweets aus der Sammlung von Geotweets T_{GEO} extrahiert. Zur Verortung dieser Tweets werden Koordinatenreferenzen sowie Ortsreferenzen der Typen Nachbarschaft und Punkt genutzt. Um die Städte zu definieren werden die morphologischen Stadtgebiete (engl. *morphologic urban areas*, MUAs) eingesetzt, welche von Taubenböck et al. (2019) als räumlich zusammenhängende städtische Siedlungen definiert wurden, deren Kernstadt 2016 mindestens 300 Tsd. Einwohner hatte. Alle der von Taubenböck et al. (2019) definierten 47 MUAs, welche innerhalb der Untersuchungsregion liegen, werden genutzt, um die Geotweets den Städten zuzuordnen. Für jede MUA werden 100 Tsd. Geotweets für die Analyse auf Ortsebene gesammelt. Zum Vergleich werden auch für jedes Land 100 Tsd. Geotweets gesammelt, welche nicht in MUAs fallen, sondern in die nicht-metropolischen Gebiete.

Für diese MUA und nicht-metropolischen Gebiete werden folgende Statistiken berechnet: Der Mittelwert des LI der dort verorteten Geotweets, die für die Raumeinheit spezifischsten⁸¹ Begriffe aus dem Wörterbuch und für die MUA ihre Bevölkerungszahl im Jahr 2016 nach Taubenböck et al. (2019).

⁸⁰ Wörter sind alle durch Leerzeichen und Satzzeichen getrennten Zeichenfolgen, ausgenommen URLs oder plattformspezifischen Elemente wie Hashtags.

⁸¹ Diese werden mittels des TF-IDF-Maßes bestimmt (Sparrck Jones 1972)

Mobilitätstyp: Zunächst wird das primäre Aufenthaltsland der Accounts als das Land definiert, für welches die Accounts die insgesamt höchste Länge an Aufenthalten aufweisen.

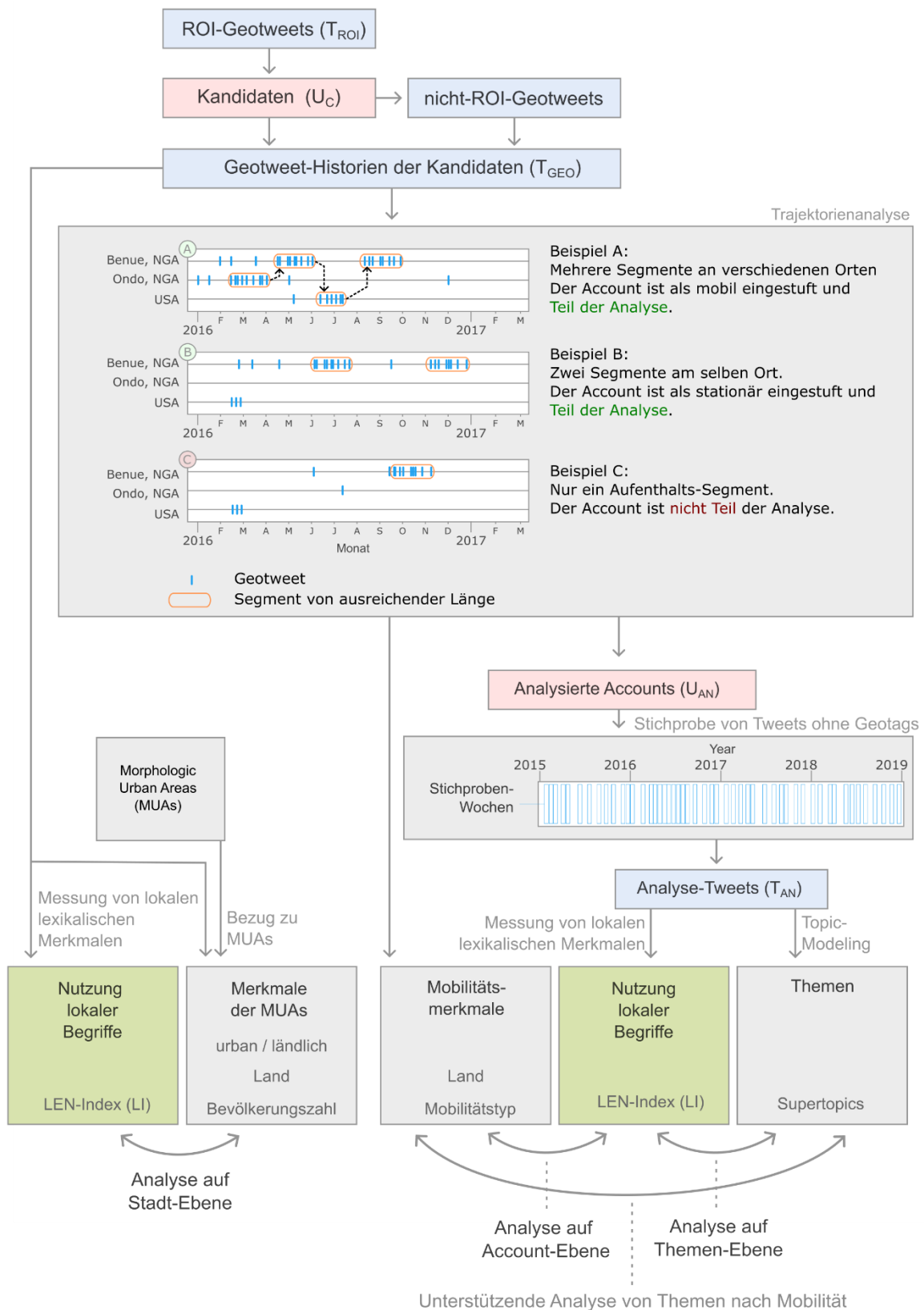


Abbildung 26: Arbeitsablauf von Studie B2. Sammlungen von Tweets sind blau markiert, Sammlungen von Accounts in Rot, Schätzungen von LEN in Grün.

Das primäre Aufenthaltsland dient als Proxy für den geographischen Hintergrundes der Nutzenden und damit der geographischen Verbreitung von LEN. Damit stellt es eine Alternative zu der Analyse auf Stadtbasis dar. Der arithmetische Mittelwert des LI wird pro primärem Aufenthaltsland berechnet. Zur Einstufung der Nutzermobilität dienen die in 5.4.1 eingeführten fünf Mobilitätstypen. Der arithmetische Mittelwert des LI wird pro Mobilitätstyp berechnet. Zusätzlich wird als kontinuierliches Mobilitätsmaß der Gyrationradius (R_{GYR}) genutzt (siehe auch 2.1.6). Der Gyrationradius der Accounts wird in einer linearen Regression gegen den durchschnittlichen LI der Accounts modelliert

$$LI = a \times \log(R_{GYR}) + b$$

Hierbei stellt der Koeffizient a den Einfluss von R_{GYR} auf den LI dar. In dieser Regression tritt der Gyrationradius in seiner logarithmischen Form auf, da die Mobilität von Menschen oft exponentiellen Verteilungen folgt, was in einer halblogarithmischen Darstellung einer geraden Linie entspricht (Kang et al. 2012; Jurdak et al. 2015).

Themen: Zur Klassifizierung der Themen dient eine Kombination von Topic-Modeling mittels BERTopic (bereits in Studie A2 genutzt, siehe 5.2.3) und manueller Fusion von Themen. Für die dafür nötigen Embeddings wird das Modell *paraphrase-multilingual-MiniLM-L12-v2* eingesetzt. Auch wenn dieses Modell nicht auf lokale Sprachen Nigerias trainiert ist (Reimers und Gurevych 2019), wird doch erwartet, dass es für Texte des westafrikanischen Sprachraums, welche auch Lehnwörter aus anderen Sprachen enthalten, bessere Text-Embeddings generiert als monolinguale englische Modelle. BERTopic wird in der empfohlenen Standardparametrisierung mittels UMAP und HDBSCAN (McInnes et al. 2018; 2017) angewendet, um Themen zu generieren. Aus diesen werden manuell fünf übergreifende fortan als *Supertopics* bezeichnete, Themenbereiche identifiziert, in welche passende Topics gruppiert wurden. Für diese fünf Supertopics ergibt sich der LI als der Mittelwert aller ihnen zugehörigen Tweets. Auch wenn mehr als fünf Supertopics identifiziert werden könnten, soll sich die Analyse in dieser Studie auf diese konzentrieren, welche als kohärent und divers betrachtet werden. Abschließend wird der Mobilitätsgrad der Nutzenden als weitere Dimension integriert und in Bezug zu Themen und LI gesetzt.

Insgesamt soll durch den Vergleich des LI über Orte, Mobilitätstypen und Themen hinweg getestet werden, ob GSM die Grundlage für eine multivariate Analyse räumlicher Merkmale (Mobilitätstypen) mit Textinhalten (Themen) und der lokalen sprachlichen Färbung (LI) bilden können.

5.5. Studie C1: Analyse der Räumlichkeit von Themen in englischen Webtextdaten

Die Räumlichkeit von Themen wird in Studie C1 durch Quantifizierung der Häufigkeit von Ortsreferenzen in Webtextdaten untersucht. Hierzu wird ein großer Korpus von Webtextdaten verschiedener Plattformen zusammengestellt (5.5.1) und inhaltlich in 18 Themen strukturiert (5.5.2). Für alle Texte wird dann durch die Anwendung von Geoparsing geprüft, ob sie eine verortbare Ortsreferenz enthalten (5.5.3). Nachdem die Qualität der Eingangsdaten anhand einer Stichprobe geprüft wurde (5.5.4), wird das Verhältnis von Thema und Ortsreferenzen durch statistische Modellierung analysiert (5.5.5).

5.5.1. Vorbereitung des Korpus

Der für diese Studie als Datengrundlage dienende Korpus besteht aus Datensätzen des Webforums Reddit (Reddit, 4.5.3), des Q&A Netzwerkes Stackexchange (Stackexchange, 4.5.4), der Nachrichtendatenbank GKG des GDELT Projektes (GDELT-GKG, 4.5.5), von

GDELT kompilierte Zeitschriftenartikel aus dem Internet Archive (GDELT-IA-Americana, 4.5.6), des Nigerianischen Webforums Nairaland (Nairaland, 4.5.2) und Tweets nigerianischer Twitter-Nutzender (Twitter, 4.5.1). Die Twitter Daten entsprechen den Tweets der stationären nigerianischen Nutzenden aus Studie B1 (5.3).

Gemeinsam ist den Datensätzen, dass sie größtenteils in englischer Sprache verfasst sind, und aus Web-Quellen bezogen wurden, was sie zu Webtextdaten macht. Innerhalb dieser bilden sie speziell soziale Medien ab, mit Ausnahme von GDELT und IA-Americana, welche als Web-Versionen klassischer Medien gesehen werden können.

Die Datensätze unterscheiden sich in Aspekten der Demographie der Nutzenden, den abgedeckten Zeiträumen, der Formate der Plattformen und technischer Möglichkeiten wie der Länge der Texte, welche beispielsweise für Tweets vergleichsweise kurz ist. Diese Unterschiede sollen im Rahmen des Studiendesigns adressiert und kontrolliert werden. Da längere Texte für die Erwähnung von Orten mehr Raum bieten, wird für alle Texte die Textlänge in Zeichen berechnet, um in der Modellierung für den Effekt der Länge zu kontrollieren. Zusätzlich unterscheiden sich die Datensätze auch in Vorhandensein einer bereits definierten thematischen Struktur beziehungsweise Annotation. Manche der Datensätze sind bereits inhärent vorstrukturiert, beispielsweise Nairaland und Reddit, welche in Unterforen aufgeteilt sind, wobei die Anzahl an Unterforen in Nairaland weniger als Hundert und auf Reddit viele Tausend beträgt. Diese Unterforen werden als Basis für eine thematische Gruppierung ihrer Inhalte genutzt, wobei jeder Text klar einem einzigen Unterforum zugeordnet werden kann. Demgegenüber können Texte in GDELT-GKG und GDELT-IA-Americana mehrere bis keine Themen-Tags enthalten. Twitter ist grundsätzlich nicht thematisch annotiert, jedoch können die Tweets wie in Studie B1 beschrieben mittels eines Klassifizierers in die von Nairaland übernommenen Themen eingeteilt werden.

Um einen aus diesen Datensätzen vereinten Korpus für die Analyse vorzubereiten, werden alle Datensätze in eine vergleichbare Struktur gebracht, welche durch Tabelle 4 illustriert wird:

Tabelle 4: Beispiel für die Struktur des für Studien C1 und C2 kompilierten Korpus (Einträge fiktiv)

ID	Datensatz	AutorenID	Zeitstempel	Textlänge	Geolokalisiert	Thema
GD27	GDELT-GKG	abcd.com	month 2017 08	271	FALSE	Gesundheit
GD28	GDELT-GKG	abcd.com	month 2017 08	824	FALSE	Gesundheit
GD29	GDELT-GKG	abcd.com	month 2017 09	434	TRUE	Ereignisse
GD30	GDELT-GKG	abcd.com	month 2017 09	3 006	FALSE	Religion
GD31	GDELT-GKG	zyxw.de	month 2015 01	579	TRUE	Religion
...	...	...	...	...	...	...
TW731	Twitter_ohne_Geotag	JohnDoe1990	month 2016 02	131	FALSE	Ereignisse
...	...	...	...	...	...	...

Für alle Datensätze wird zunächst die Akquise vollzogen und AutorenID, Zeitstempel und Textlänge jedes Posts bestimmt. Die Methoden der Akquise und Vorprozessierung der Datensätze unterscheiden sich aufgrund der Differenzen der Datenquellen stark.

Stackexchange: Für diese Studie wird der Stackexchange Data-Dump vom April 2024 genutzt, welcher Posts von August 2008 bis März 2024 enthält. Um die große Datenmenge zu reduzieren, werden Beiträge von Accounts entfernt, welche weniger als fünf oder mehr als 10 000 Beiträge insgesamt aufweisen, oder nur zu einem Thema posteten. Die Stackexchange-

Kommentare enthalten Zeitstempel, welche auf Kalendermonate aggregiert werden. Die Textlänge in Zeichen wird aus dem Haupttext abgeleitet.

Reddit: Der *WEBISldr-17 Reddit* Korpus (Völske et al. 2017) enthält Beiträge mehrerer Tausend Subreddits. Diese werden mit den von Partridge et al. (2024) erstellten Reddit Community Embeddings verknüpft, welche Subreddits basierend auf Übereinstimmung ihrer Nutzernetzwerke hierarchisch gruppieren. Die Reddit-Kommentare enthalten keine Zeitstempel. Die Textlänge in Zeichen wird aus dem Text des Kommentars abgeleitet. Der Accountname wird als AutorenID genutzt.

Nairaland: Daten für zahlreiche Unterforen von Nairaland werden in einem randomisierten Web-Scraping Ansatz heruntergeladen (4.5.2). Diese Daten, welche einen Zeitraum von 2014 bis 2021 abdecken, enthalten mehrere Millionen Kommentare inklusive Zeitstempel, welche auf Kalendermonate aggregiert werden. Die Textlänge in Zeichen wird aus dem Text des Kommentars abgeleitet. Der Accountname wird als AutorenID genutzt.

Twitter: Die Tweets entsprechen den Tweets der stationären nigerianischen Nutzenden (U_0), welche in Studie B1 identifiziert werden. Für diese Accounts werden sowohl Tweets mit als auch ohne Geotags in einem pseudo-stratifizierten Ansatz über 48 zufällig verteilte 7-Tage Intervalle zwischen 2015 und 2019 über die Twitter API bezogen (vergleiche Kapitel 6.3.1). Geotweets und Tweets ohne Geotags werden separat analysiert, um ein differenzierteres Bild zu erhalten. Die detaillierten Metadaten der Tweets enthalten sekundengenaue Zeitstempel, welche auf den jeweiligen Kalendermonat reduziert werden. Die Textlänge in Zeichen wurde aus dem Text des Tweets⁸² berechnet. Der Accountname wird als AutorenID genutzt.

GDELT-GKG: Aufgrund der immensen Größe der GDELT Datenbank werden nur Einträge akquiriert, welche in jene 7-Tage Intervalle fallen, für die auch Twitter-Daten akquiriert werden. Hierdurch wird eine zeitliche Übereinstimmung der beiden Datensätze erreicht. GDELT-GKG Daten für diesen Zeitraum werden über die Google BigQuery Plattform ausgewählt und heruntergeladen. Zusätzlich werden aus allen vorhandenen Quellen (typischerweise Web-Nachrichtenseiten) 10 000 Quellen ausgewählt, für welche mindestens 10 unterschiedliche Artikel in 10 verschiedenen Kalendermonaten vorhanden waren. Der Name der Quellen wird als AutorenID genutzt. Die GDELT-GKG-Einträge enthalten Zeitstempel, welche auf den jeweiligen Kalendermonat reduziert werden. Da die Einträge keinen Volltext enthalten, wird die Textlänge durch die höchste Zeichenposition approximiert, welche in den Feldern der GDELT-GKG Metadaten enthalten ist (beispielsweise die Zeichenposition von Personen, Mengen, oder Orten).

GDELT-IA-Americana: Alle verfügbaren Dokumente für die Jahre 2005 bis 2014 werden über die Google BigQuery Plattform ausgewählt und heruntergeladen. In dieser Zeitspanne enthalten sie hauptsächlich Magazine, regelmäßige Blog-Posts und wenige offizielle Dokumente. Der Name der Quellen wird als AutorenID genutzt, wobei hervorzuheben ist, dass dies nicht notwendigerweise einem einzelnen Autor entspricht, da auch mehrere Autoren vorliegen können (beispielsweise „Marx, Karl, 1818–1883; Engels, Friedrich, 1820–1895“). Die GDELT-IA-Americana Daten enthalten Zeitstempel auf Jahresebene. Aufgrund der großen Textlänge werden die Dokumente zwecks Analyse in Fenster aufgeteilt. Dazu wird um die Position jeder detektierte Ortsreferenz ein 100-Wort Fenster extrahiert und für die Analyse genutzt. Die Textlänge dieser Fenster liegt also fest bei 201-Zeichen.

⁸² Alle Zeichen inklusive URLs, Hashtags oder Erwähnungen anderer Accounts (Mentions)

5.5.2. Themen-Taxonomie

Um alle diese verschiedenen Datensätze einer vergleichbaren inhaltliche Struktur zuzuführen, wird eine generische Themen-Taxonomie erstellt. Sie kann und soll keinen Anspruch auf Universalität haben, sondern eine von vielen Möglichkeiten darstellen, die thematische Vielfalt der genutzten Datensätze auf eine Weise einfangen, die gleichzeitig konsistent und menschlich interpretierbar ist.

Hierfür wird zunächst eine Liste von mehr als 5 aber weniger als 25 Themen erstellt, welche die existierenden Kategorien (wie Unterforen) nach Einschätzung der Autoren angemessen repräsentieren soll. Eine Kategorie *Sonstiges* wird zur Sammlung von generischen oder mehrdeutigen Inhalten wie Literatur oder Witze definiert, und dient im Modellierungsansatz als Referenzkategorie. Nachdem die Themen der Taxonomie festgelegt sind, werden die bestehenden Themenstrukturen der Datensätze dieser Taxonomie zugeordnet (Abbildung 27). Die Erstellung der Taxonomie und die Zuordnung der Datensatz-spezifischen Kategorien zu den Themen der Taxonomie wird von dem Autor der Studie im *grounded-theory* Stil, also auf Basis des empirischen Materials und für dieses entwickelt (Strübing 2019). Daraufhin prüfen drei Forscher aus verschiedenen Wissenschaftsbereichen⁸³ sie auf Plausibilität und Konsistenz und schlagen Änderungen vor. Bei Übereinstimmung von mindestens zwei der beteiligten Forschenden wird die Taxonomie angepasst.

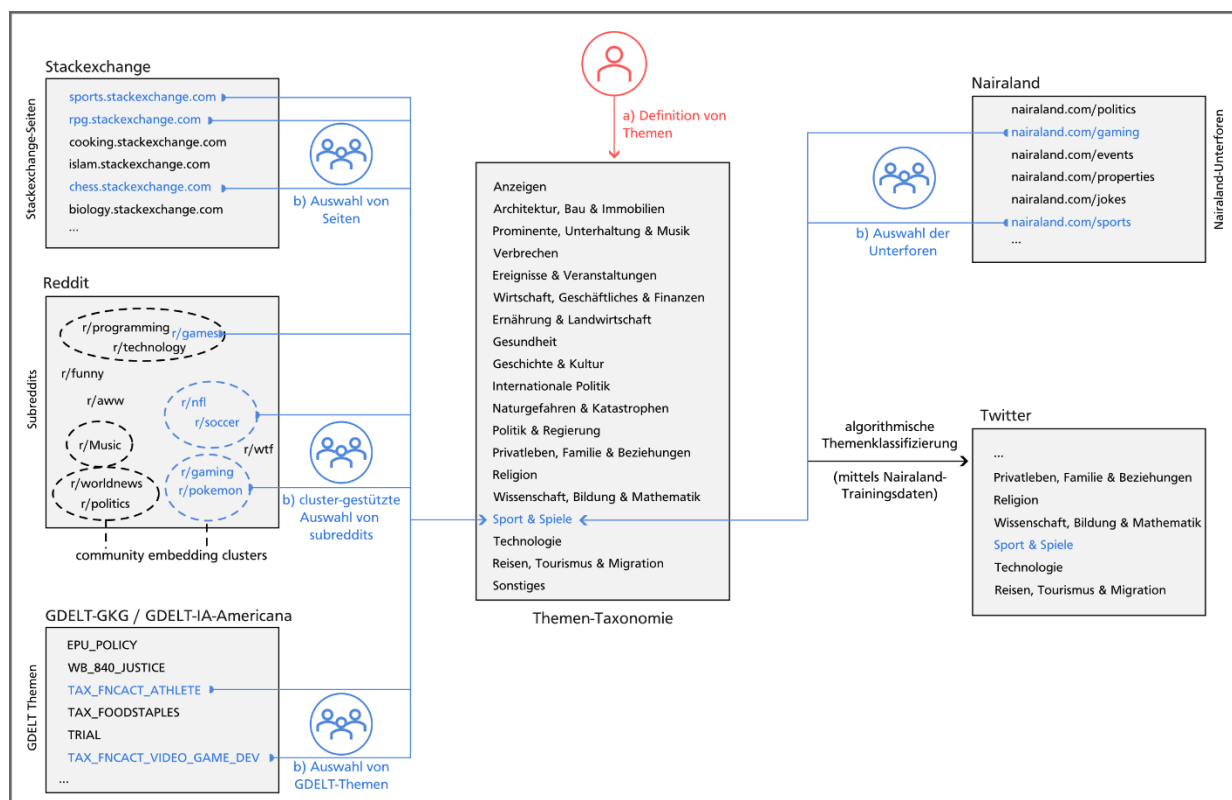


Abbildung 27: Entwicklung der Thementaxonomie. Themen wurden zunächst definiert (a). Daraufhin wurden datenquellenspezifische Kategorien (beispielsweise GDELT Themen auf dem GDELT-GKG) den Themen zugewiesen (b).

⁸³ Der akademische Hintergrund der Forscher lag jeweils in Geographie, Digital Humanities und Englischer Sprachwissenschaft. Ich danke ihnen aufrichtig für ihren Beitrag zu dieser Arbeit.

Es ist nicht zu erwarten, dass alle Themen in allen Datensätzen enthalten sind. Zum Beispiel existiert keine auf Werbung fokussierte Stackexchange-Seite⁸⁴ und es ist aufgrund von internen Filterungen des GDELТ Projekts nicht damit zu rechnen, dass eindeutig kommerzielle Inhalte oder Unterhaltungsinhalte in den GDELТ-GKG und GDELТ-IA-Americana Daten auftauchen.

Die Zuordnung zu den Themen der Taxonomie erfolgt für alle Datensätze unterschiedlich:

Stackexchange: Stackexchange-Seiten können aufgrund ihrer relativ geringen Anzahl manuell den Themen zugeordnet werden.

Reddit: Aufgrund der großen Anzahl von Subreddits wurden die Community Embedding Cluster als Grundlage für eine manuelle Zuordnung genommen. Cluster von Subreddits werden zunächst als Ganzes manuell den Themen zugeordnet. Daraufhin werden die Cluster geprüft und Anpassungen zur Verbesserung der inhaltlichen Konsistenz vorgenommen, bei welchem einzelne Subreddits Themen zugeordnet oder aus diesen entfernt werden.

GDELТ und GDELТ-IA-Americana: Die große Anzahl an GDELТ-Themen⁸⁵ macht eine manuelle Zuordnung höchst aufwendig. Stattdessen werden für jedes Thema der Taxonomie aus einer Auflistung der im Datensatz enthaltenen GDELТ-Themen (bezogen via BigQuery) passende GDELТ-Themen ausgewählt, bis alle Themen durch mehrere GDELТ-Themen repräsentiert sind. Hierbei werden für Themen, die in GDELТ-GKG unterrepräsentiert sind, mehr GDELТ-Themen ausgewählt, um die Unterschiede im Datenvolumen zu verringern.

Nairaland: Nairaland-Themen können aufgrund ihrer relativ geringen Anzahl manuell den Themen zugeordnet werden.

Twitter: Um Tweets zu Themen zuzuweisen, wurde wie in Studie B1 ein domänenadaptives Klassifizierungsverfahren genutzt (vgl. 5.3.2) welches einen Klassifizierer auf Nairaland-Daten trainiert und auf Tweets anwendet. Anders als in Studie B1 wird der Klassifizierer hier direkt auf die Erkennung der Themen der Taxonomie trainiert und nicht die Erkennung von Nairaland-Unterforen. Zudem wird das neuere *twhin-bert* Modell als Backbone des Klassifizierers genutzt, welches auf Tweets in verschiedenen Sprachen trainiert wurde (Zhang et al. 2023).

5.5.3. Geoinformation

Für alle Einträge wird identifiziert, ob eine Geoinformation enthalten ist. Auch dies erfolgt unterschiedlich je nach Datensatz. Für Stackexchange, Reddit, Nairaland, Twitter wird Geoparsing angewandt. Hierbei sollen nicht genaue Geokoordinaten identifiziert werden, sondern lediglich, ob verortbare Geokoordinaten vorhanden sind. Zunächst wird ein Ensemble von NER-Modellen genutzt, um erwähnte Eigennamen von Orten im Text zu erkennen. Ein Ensemble ist gegenüber einem einzelnen NER-Modell robuster. Folgende Modelle werden genutzt: *Spacy-NER* (Honnibal et al. 2020), *flair-ner-english-ontonotes-large* (Version für 18 Entitätsklassen) (Schweter und Akbik 2021), *bert-base-NER* (Devlin et al. 2018) und *MasakhaNER* (Adelani et al. 2021), ein Modell welches auf afrikanische Texte optimiert ist.

⁸⁴ Auch wenn Eigenpromotion in gewissem Rahmen erlaubt ist, wenn es für die zu beantwortende Frage relevant ist, siehe meta.stackexchange.com/questions/7931/

⁸⁵ 59 315 unterschiedliche Themen (Stand 2021), siehe <https://blog.gdelproject.org/new-november-2021-gkg-2-0-themes-lookup/>

Nachdem Entitäten der Typen GPE⁸⁶, LOC⁸⁷ und FAC⁸⁸ identifiziert sind, wird getestet, ob die Geonames-API⁸⁹ für die Entität mindestens einen Ortseintrag zurückgibt. Wenn ja, dann wird der Text als *geolokalisiert* betrachtet.

GDELT-GKG: Die von GDELT detektierten Locations sind in einem Metadatenfeld (V2_Locations) in Form von Textstrings eingetragen. Enthält dieses Feld einen Eintrag, wird der Text als *geolokalisiert* betrachtet.

GDELT-IA-Americana: Auch hier wird das Metadatenfeld V2_Locations durchsucht. Für GDELT-Themen, welche einem Thema der Taxonomie entsprechen, wird die Position des Themas im Text extrahiert. Wenn in einem Fenster um diese Position (100 Zeichen vor und 100 Zeichen nach der Erwähnung des Themas) auch eine Ortserwähnung detektiert wurde, wird der Text als *geolokalisiert* betrachtet.

5.5.4. Validierung⁹⁰

Um die Ergebnisse der Themenklassifizierung und des Geoparsings mit menschlicher Einschätzung zu vergleichen, werden 10 Texte zufällig für jede Kombination von Datensatz und Thema ausgewählt⁹¹ und händisch durch den Autor annotiert, wobei Webrecherche die manuelle Interpretation des Textes unterstützt. Für jeden Text wurde sowohl das Thema als auch Ortserwähnungen annotiert. Für GDELT-IA-Americana kann dies nicht durchgeführt werden. Obwohl für diesen Satz Volltexte verfügbar sind, sind die von GDELT extrahierten Themen nicht in diesen Texten verortbar, da sie von einer variierenden und nicht dokumentierten Zeichenverschiebung betroffen sind. Daher konnte der Fenster-basierte Text eines GDELT-Themas nicht extrahiert werden.

Zum Vergleich der menschlichen Einschätzung und der algorithmisch unterstützten Themenklassifizierung wurde Cohens Kappa (κ) berechnet (Cohen 1960).

5.5.5. Statistische Auswertung

Die Häufigkeit von verortbaren Ortserwähnungen wird zunächst mittels deskriptiver Statistiken analysiert. Danach wird der eigentliche Fokus der Analyse realisiert, indem diese Beziehung zwischen der Häufigkeit von verortbaren Georeferenzen und der Themen der Texte statistisch modelliert wird.

In der deskriptiven Analyse wird diese Häufigkeit als der Anteil von Dokumenten mit verortbaren Georeferenzen an allen Dokumenten berechnet und als $Frac_{geo}$ bezeichnet:

$$Frac_{geo} = \frac{N_{geolocated}}{N_{geolocated} + N_{notgeolocated}}$$

⁸⁶ Geopolitische Entitäten (engl. *Geopolitical Entity*)

⁸⁷ Orte, die keine Geopolitischen Entitäten sind

⁸⁸ Einrichtungen (engl. *Facility*)

⁸⁹ <https://www.geonames.org/export/web-services.html>

⁹⁰ Der Begriff Validierung wird hier genutzt, da die Anwendung der (in der Fernerkundung verbreiteten) Konvention entspricht, händisch annotierte Daten als *Ground Truth* zu nutzen. Die hier zum Vergleich genutzten händisch annotierten Daten sollen jedoch keinen Anspruch auf allgemeine Korrektheit haben, und lediglich eine alternative Perspektive bieten.

⁹¹ Insgesamt wurden 1075 Texte annotiert.

Die **deskriptive Auswertung** erfolgt durch die Quantifizierung von Frac_{geo} für (1) jeden Datensatz und Thema, (2) jeden Datensatz und Zeiteinheit, (3) jeden Datensatz und Autor, (4) jeden Datensatz und Textlängengruppe.

Der Modellierungsansatz fokussiert direkt die Beziehung zwischen Themen und Häufigkeit von verortbaren Georeferenzen, wobei die Einflüsse von Zeit, Autoren und Textlänge kontrolliert werden sollen.

Für diese Kontrolle eignet sich ein verallgemeinertes lineares gemischtes Modell (englisch *generalized linear mixed model*, GLMM, für eine Anwendung von GLMM auf sprachliche Analysen siehe Gries, 2015) mit der Präsenz von Geoinformation als der binären abhängigen Variable. Die Themen und die Textlänge gehen als feste Effekte (engl. *fixed effects*) in die Modellierung ein, die Autoren und Zeitschritte als diskrete zufällige Effekte (engl. *random effects*). Für den Reddit Datensatz waren keine Zeitstempel vorhanden und für GDELT-IA-Americana wurde die Textlänge nicht genutzt, da sie stark durch die Wahl der Fenstergröße beeinflusst war.

Zur Schätzung des GLMM dient die Softwarebibliothek *glmmTMB* (Brooks et al. 2017), welche in der Programmiersprache R implementiert ist (R Core Team 2024). Um die Prozessierung eines großen Korpus zu ermöglichen, wurde für jeden Datensatz separat ein GLMM geschätzt. In diesem Ansatz kann jedes Modell als eigenständiges Experiment zur Geospatiality für die jeweilige Textsorte betrachtet werden.

Für jedes Thema wird der geschätzte Koeffizient des Themas in logarithmierte Chancenverhältnis (*Logits* oder engl. *log-odds*) angegeben und als Ausdruck für die Stärke der Geospatiality-Effekte dieses Themas ausgewertet. Dessen Signifikanz wird durch den p-Wert des Koeffizienten repräsentiert. Die Übereinstimmung zwischen den Datensätzen wird mittels des Spearman Rangkorrelationskoeffizienten ρ (Spearman 1904) interpretiert.

Durch die Nutzung von Rangkorrelation wird der Fokus auf die relative Rangfolge der Themen in ihrer Räumlichkeit gelegt und nicht auf die absolute Höhe der Koeffizienten, welche vermutlich stark von der Textsorte des Datensatzes beeinflusst ist.

5.6. Studie C2: Differenzen der Räumlichkeit von Themen zwischen Geotag- und Geoparsing-basierten Ansätzen

Während Studie C1 das Phänomen der thematischen Geospatiality basierend auf Geoparsing analysiert, wird in Studie C2 speziell das Verhältnis von Geotag-basierter und Geoparsing-basierter Geoinformationen in den Fokus genommen. Dies erfolgt für einen bereits in Studie C1 vorgestellten Twitter Datensatz. Die Analyse vergleicht zunächst die Häufigkeit von Ortsnennungen zwischen Geotweets und Tweets ohne Geotags (5.6.1). Zweitens wird die Räumlichkeit modelliert, sowohl –wie in Studie C1– basierend auf Geoparsing, als auch basierend auf Geotags (5.6.2), und die Unterschiede in Räumlichkeit zwischen beiden Ansätzen verglichen.

5.6.1. Analyse von Unterschieden in der Häufigkeit von Ortsnennungen zwischen Geotweets und Tweets ohne Geotags

Die Datenbasis für diese Studie bildet der wie in Studie C1 beschrieben akquirierte Twitter Datensatz (5.5.1). Die Autoren dieser Tweets sind dieselben Accounts, welche als zum Zeitpunkt des Postens stationär in Nigeria eingestuft wurden (vgl. Studie B1 in 5.3). Alle diese Tweets stammen aus 48 zufällig zwischen 2015 und 2019 verteilten Wochen. Der Datensatz

enthält sowohl Tweets mit als auch ohne Geotags (Abbildung 28).

Wie in Kapitel 5.5 beschrieben wurde dieser Datensatz aufbereitet, sodass jeder Tweet (1) einem Thema zugewiesen wurde, und (2) mittels Geoparsing identifiziert wurde, ob er eine verortbare Georeferenz enthält. Diese Tweets können daher mit zwei Ansätzen geographisch verortet werden: Einmal durch Geoparsing, zum anderen über die Nutzung des Geotags. Hierdurch kann die Äußerung themenspezifischer Geospatiality zwischen den beiden Ansätzen verglichen werden.

Zunächst wird statistisch analysiert, ob die Häufigkeit von Ortsnennungen $Frac_{geo}$ für Tweets mit Geotags anders ist als für Tweets ohne Geotags. Dies erfolgt in einem ersten Schritt, indem $Frac_{geo}$ für beide Gruppen von Tweets verglichen wird. In einem zweiten Schritt kann dies auf Ebene der Themen heruntergebrochen werden, um zu identifizieren, ob die Beziehung von Geotags und geographischen Textinhalten sich zwischen Themen unterscheidet.

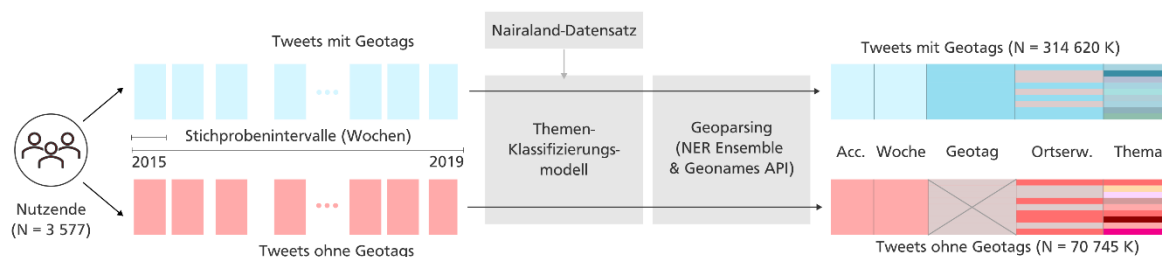


Abbildung 28: Ableitung des für Studie C2 genutzten Datensatzes, welcher aus Geotweets und Tweets ohne Geotags besteht

5.6.2. Analyse von Differenzen in Räumlichkeit

Wie bereits in Studie C1 wird die Geospatiality der Themen auch hier durch Modellierung erfasst. Anders als in Studie C1 wird hierbei der Datensatz nicht in Geotweets und Tweets ohne Geotags geteilt. Stattdessen werden auf den gesamten Twitter-Daten zwei Modelle geschätzt, welche sich lediglich durch die abhängige Variable unterscheiden. Im ersten Modell M_{tag} ist die abhängige Variable das Vorhandensein eines Geotags. Im zweiten Modell M_{txt} ist die abhängige Variable das Vorhandensein einer verortbaren Georeferenz im Text. Abgesehen davon sind Daten und Variablen identisch, und wie auch in Studie C1 wird für Autor, Zeit, und Textlänge kontrolliert. Die Modellierung erfolgt wieder in einem verallgemeinerten linearen gemischten Modell welches mittels glmmTMB geschätzt wird (Brooks et al. 2017).

Durch den Vergleich der Koeffizienten der Themen in M_{tag} und M_{txt} kann erörtert werden, inwiefern thematische Geospatiality von der Art der genutzten Geodatentyps abhängt. Mittels des Spearman-Rangkorrelationskoeffizienten ρ wird geprüft, ob die beiden Ansätze zu ähnlichen Bewertungen der Räumlichkeit von Themen kommen.

6. Ergebnisse der Studien

In diesem Kapitel werden die Ergebnisse der Studien präsentiert und grundsätzlich eingeschätzt.

6.1. Studie A1: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Erfassung digitaler Ungleichheiten in Afrikas expandierenden städtischen Räumen

Die Gesamtzahl an Tweets, welche im Analysezeitraum für die 135 UTG von der API bereitgestellt werden, beläuft sich auf 14 391 264. Nach Ausschluss von ungeeigneten Accounts verbleiben 6 996 346 Geotweets, deren Verteilung innerhalb der UTG die Grundlage der Analyse bildet.

Zur Illustration der Daten sind in Abbildung 29 für vier UTG die Verteilungen der Geotweets räumlich vor dem Hintergrund des kartierten Bebauungsjahres und im Verbund mit den berechneten statistischen Ergebnissen dargestellt.

Ginti, Nigeria (B1) ist das UTG, welches die größte zwischen 1985 und 2020 neu bebaute Fläche aufweist. Das UTG liegt in Ikorodu nordöstlich von Lagos und zeichnet sich durch kontinuierlich und organisch anmutende Siedlungsexpansion auf, welche von den Hauptverkehrsachsen ausstrahlt und von mehreren Grünflächen und Feuchtgebieten eingegrenzt wird, die das UTG durchziehen. Die etwa 62 Tsd. Geotweets sind flächig im bebauten Gebiet verteilt und erreichen ihre höchste Dichte in den zentralen Zonen entlang der Straßen, welche Orte von sozialer und wirtschaftlicher Aktivität sind. Nahe des geographischen Zentrums des UTGs ist eine Konzentration von 17 Tsd. Geotweets auffällig, welche von verschiedenen Accounts stammen, und verschiedene Inhalte aufweisen. Es ist unklar, ob dies auf eine technische Gegebenheit zurückzuführen ist. Ein offensichtlicher Fehler ist nicht zu erkennen. Insgesamt wird aufgrund der Konzentration entlang der Straßen, wo die Bebauung am ältesten ist, eine signifikante, moderat hohe Disparität gemessen ($DDS_{UTG} = 0,257$, $p = 0$).

In **Nador, Marokko (B2)** an der Mittelmeerküste findet sich die höchste Geotweet-Dichte im verhältnismäßig alten Zentrum. Dies führt zu einer verhältnismäßig hohen gemessenen Ungleichheit ($DDS_{UTG} = 0,367$, $p = 0,01$). In dieser Stadt, welche im Osten von einer Lagune und auf allen anderen Seiten von bergigem Gelände begrenzt ist, wurde der Großteil der bebauten Fläche vor 1985 bebaut und formt einen durchgängig besiedelten Stadtkern. Dennoch hat in den letzten Jahrzehnten sichtbare Siedlungsexpansion stattgefunden: Erstens in der Form von dispersen Bebauung in den westlichen Vorstädten und zweitens in einem offenbar hochgradig geplanten Bezirk am südöstlichen Ufer, für welchen aufgrund zahlreicher Hotels und gastronomischer Einrichtungen ein touristischer Fokus der Landnutzung vermutet werden kann. Innerhalb dieses geplanten Gebietes wurden mehrere Hundert der insgesamt 17 Tsd. Geotweets lokalisiert, während die erstgenannten Vorstädte sehr wenige enthalten. Der weitaus größte Teil der Geotweets des UTGs findet sich im Stadtkern, insbesondere entlang der Hauptstraßen. Die Konzentration ähnelt in diesem Aspekt der in Ginti, allerdings sind die Kerngebiete in Nador älter, was zu einem höheren DDS_{UTG} führt.

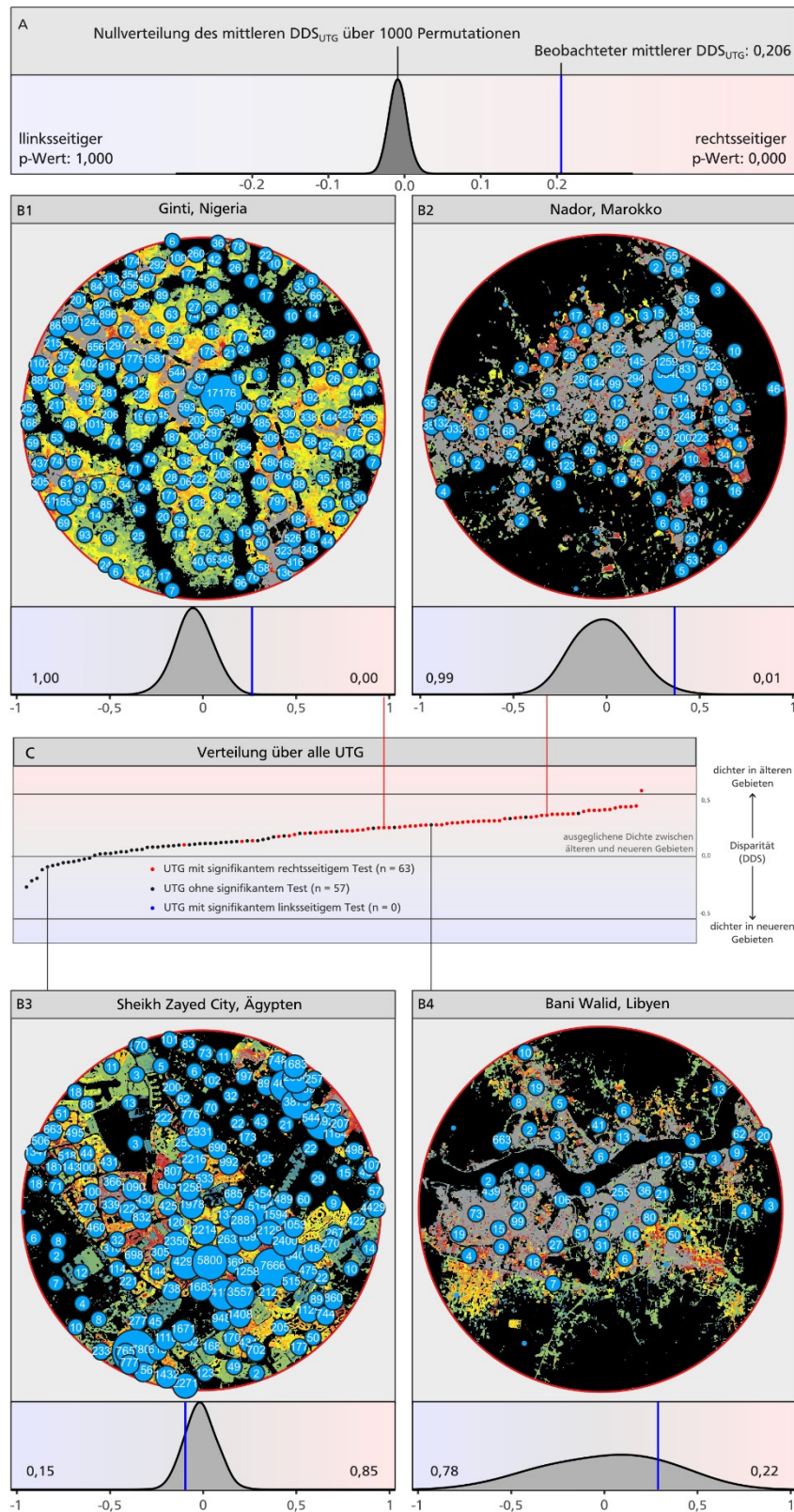


Abbildung 29: Illustration der Geotweet-Verteilungen und Ergebnisse der Dichte für vier ausgewählte UTGs. Oben (A): Mittlerer DDS_{UTG} (blaue vertikale Linie) verglichen mit der simulierten Nullverteilung (grau). B1-B4: Vier UTGs mit deren räumlicher Geotweet-Verteilung. Tweets sind Visualisierungszwecke räumlich aggregiert, die Nummer in blauen Kreisen gibt die Anzahl der zugehörigen Geotweets an. Die Kurve unter jeder Karte zeigt die simulierte Nullverteilung (grau) relative zu der gemessenen Disparität (blaue vertikale Linie). Werte links deuten an, dass die Geotweet-Dichte in neueren Gebieten höher ist und umgekehrt. Die Werte über der Kurve geben jeweils den linksseitigen und rechtsseitigen P-Wert an. Mitte (C): Verteilung des DDS_{UTG} über alle UTG hinweg.

Sheikh Zayed Stadt, Ägypten(B3) ist eine von mehreren hochgradig geplanten Siedlungen, welche im Laufe der Jahrzehnte im Westen der Hauptstadt Kairo entstanden. Die 110 Tsd. Geotweets liegen räumlich verteilt innerhalb des UTG, wobei mehrere bemerkenswerte Konzentrationen herausstechen. Zum einen liegen diese in mehreren Einkaufszentren und Malls, von welchen die zwei Größten um das Jahr 2010 erbaut wurden. Eine weitere Konzentration liegt in einem Hochtechnologiebezirk im Norden, welcher hochtechnologie-Einrichtungen der Wirtschaft, Bildung und Administration beherbergt. Die Erbauungszeiträume dieser Orte besonderer Konzentration sind über die letzten Jahrzehnte verteilt. Aus diesem Grund wurde kein signifikantes Ungleichgewicht gemessen ($DDS_{UTG} = -0,093$, $p = 0,85$).

Ein Tal trennt die in **Libyen** gelegene Oasenstadt **Bani-Walid (B4)** in eine Nord- und Südhälfte. Bani-Walid war im libyschen Bürgerkrieg 2012 ein Schauplatz heftiger Kämpfe (Gumuchian 2012), wuchs dennoch im Laufe der Jahre und hat sowohl in südwestlicher, südöstlicher und nördlicher Richtung entlang von Korridoren expandiert. Insgesamt hat sich die bebaute Fläche zwischen 1985 und 2020 etwa verdoppelt. Trotz der erheblichen zugewachsenen Fläche befinden sich fast alle der 2,5 Tsd. Geotweets in den älteren, vor 1985 bebauten Gebieten. Selbst dort sind sie hochgradig konzentriert und in einigen Fällen teilen sich Dutzende Geotweets dieselben Koordinaten. Wenngleich dieses Phänomen auch an anderer Stelle beobachtet wurde (beispielsweise in Ginti, siehe oben) hatte es aufgrund der geringen Anzahl an Geotweets in Bani Walid großen Einfluss auf die Analyse. Zusammen mit der geringen Gesamtnutzeranzahl führte dies zu einer hohen Varianz des DDS im Permutationstest, wodurch der rechtsseitige Signifikanztest klar negativ ausfiel ($DDS_{UTG} = 0,281$, $p = 0,22$).

6.1.1. Afrikaweit

Die Verteilung von DDS_{UTG} wird über alle UTG hinweg analysiert. Für 63 der 135 UTG zeigt der rechtsseitige Hypothesentest einen signifikant positiven DDS_{UTG} , was auf ein Ungleichgewicht in der Dichteverteilung zugunsten der älteren Gebiete hindeutet.

Demgegenüber gibt es nur wenige UTG, in welchen der DDS_{UTG} negativ ist, und keine dieser Messungen wird durch einen statistisch signifikanten linksseitigen Test unterstützt. Dies ist beispielsweise in Zango in Angola ($-0,272$), Lumpundu im Kongo ($-0,217$) oder Al-Fashir in Sudan ($-0,195$) der Fall. Insgesamt gibt es also keine Fälle, in welchen ein Ungleichgewicht zugunsten der neueren Siedlungsteile besteht. Für 55 UTG ist der Test nicht signifikant. Für 17 UTG waren die Daten als unzureichend für die Auswertung eingestuft. Insgesamt liegt der Mittelwert des DDS_{UTG} bei 0,201 ($\sigma=0,16$). Vor dem Hintergrund des Permutationstests ergab sich ein p-Wert von 0.0. Es ist also als unwahrscheinlich zu betrachten, dass der gemessene DDS_{UTG} -Wert durch zufällige Anordnung der Geotweets entsteht. Beschränkt man die Betrachtung auf UTG mit einem signifikanten rechtsseitigen Test, liegt der Mittelwert des DDS_{UTG} unter ihnen bei 0,308, also bei 30,8% der theoretisch möglichen Disparität.

Die Sensitivitätsanalyse (siehe Appendix A3) legt nahe, dass die hier dargestellten Ergebnisse robust sind. Wenngleich der DDS_{UTG} grundsätzlich von GPS-Fehlern und Datenverfügbarkeit beeinflusst werden kann, ist es sehr unwahrscheinlich, dass dies in einem Ausmaß geschieht, welches die Schlussfolgerung der Studie beeinflusst.

Die Ergebnisse der Analyse unterstützen die Hypothese, dass insgesamt ein Trend existiert, welcher in vielen Fällen zu einem Ungleichgewicht zugunsten von älteren Siedlungsteilen und nicht zugunsten von neueren Siedlungsteilen führt.

6.1.2. Differenziert auf der Ebene von Georegionen Afrikas

Diese Ergebnisse können nach den geographischen Regionen differenziert werden (Abbildung 30). Sowohl parametrische als auch nichtparametrische ANOVA der DDS_{UTG} -Verteilung zwischen Regionen sind signifikant (p-Werte von 0,002 und 0,001) mit einer moderat hohen Effektstärke ($\omega^2 = 0,108$ und $\eta^2 = 0,1295$). Extreme Ausreißer treten nicht auf, und die DDS_{UTG} sind über die UTG hinweg ungefähr normalverteilt (Shapiro-Wilk Test p-Wert > 0,05), jedoch gibt es Hinweise auf Varianzheterogenität (Levene-test p-Wert = 0,012). Daher ist die nichtparametrische ANOVA bevorzugt zu interpretieren. Richtet man den Blick auf die Häufigkeit, mit welcher Disparität auftritt, zeigen sich erhebliche Unterschiede zwischen Regionen. Im Nördlichen und Südlichen Afrika ist die Mehrheit der UTG kein Ort messbarer signifikanter Disparität. Im Westlichen Afrika hingegen wird für die klare Mehrheit (23 von 29 UTG) ein signifikant positiver DDS gemessen. Im Östlichen- und Mittleren Afrika erweist sich die Datenlage als schlecht, hier ist die Datenverfügbarkeit in mehr als einem Fünftel der Fälle unzureichend. Insgesamt ergibt sich das Bild von statistisch messbaren Unterschieden zwischen manchen, aber nicht allen, Regionen Afrikas.

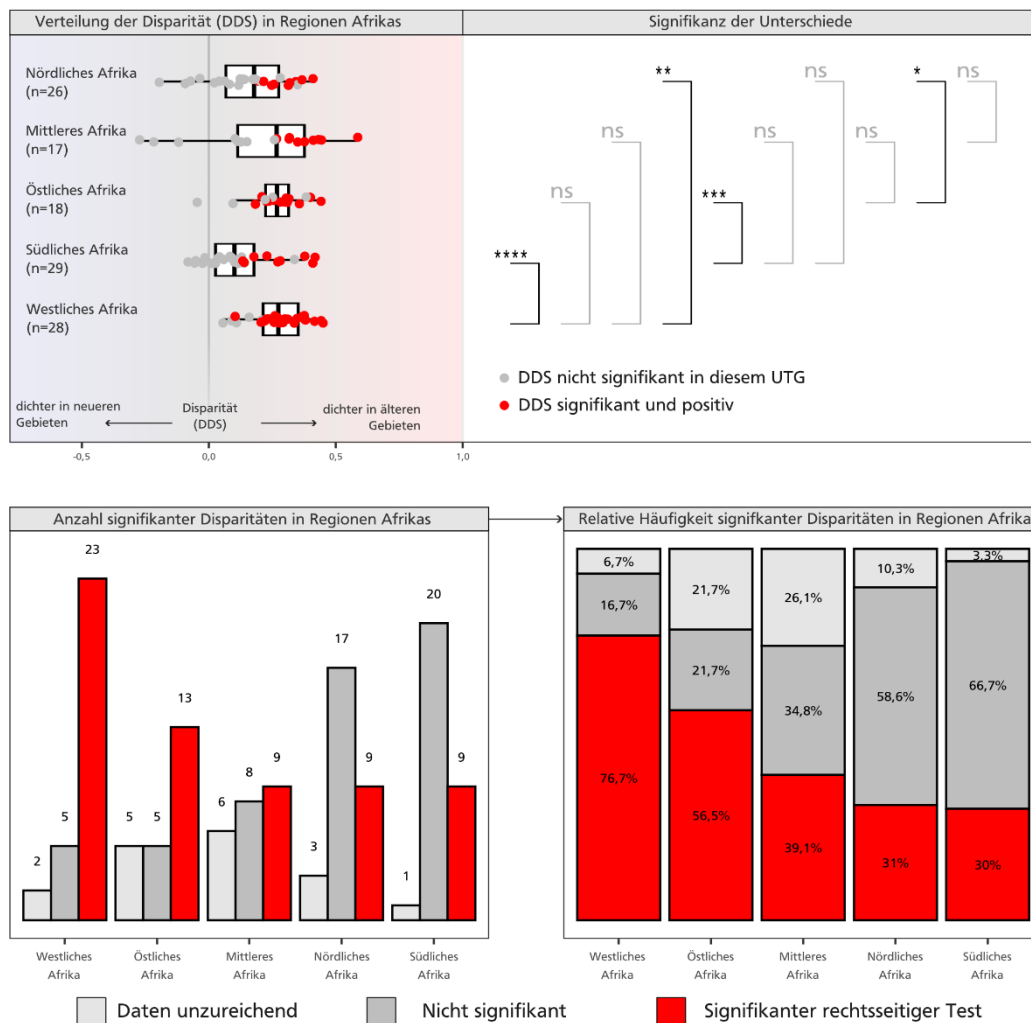


Abbildung 30: Vergleich der Disparität (gemessen durch DDS_{UTG}) zwischen UN-Regionen. Oben: Vergleich des DDS mittels parametrischer ANOVA. DDS_{UTG} -Unterschiede zwischen Regionen wurde mittels t-test erfasst und wie folgt dargestellt: ns: kein signifikanter Unterschied; * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$; **** $p < 0,0001$. Unten: Prävalenz (Anzahl der UTG) und Häufigkeit (Anteil der UTG) von signifikanten DDS_{UTG} innerhalb jeder Gruppe.

6.1.3. Differenziert nach Graden von Geplantheit (IoP)

Der Blick auf die Beziehung der Disparität zur Geplantheitsintensität zeigt folgendes Bild: Sowohl auf der Ebene von Blocks (Abbildung 31) als auch Straßennetzwerken (Abbildung 32) werden signifikante Unterschiede zwischen IoP und DDS_{UTG} gemessen. Auf der Ebene von Blocks zeigen die ANOVA des DDS_{UTG} nach IoP_B ein signifikantes Ergebnis ($p = 0,002$) mit einer mittleren Effektstärke ($\omega^2 = 0,109$). Es gibt keine Hinweise auf Varianzheterogenität (Levene-test p -value = 0,607), jedoch zeigt die Klasse IoP_B 2 einen extremen Ausreißer, und Normalität kann nicht bestätigt werden. Daher sollte in diesem Fall die nichtparametrische ANOVA interpretiert werden, wobei sich ein ähnliches Bild der Unterschiede ergibt (Kruskal-Test p -Wert = 0,001, η^2 war moderat mit 0,1354). Der zweiseitige post-hoc t-test zeigt, dass die stärksten signifikanten Unterschiede zwischen IoP_B 2 und IoP_B 5 bestehen ($p = 0,00096$). Weitere Unterschiede werden zwischen IoP_B 2 und IoP_B 4 gefunden ($p = 0,01670$), zusätzlich zwischen IoP_B 2 und IoP_B 3 ($p = 0,01145$), IoP_B 3 und IoP_B 5 ($p = 0,02260$) und IoP_B 2 und IoP_B 1 ($p = 0,03601$), wobei Letztere aufgrund der geringen Größe von Gruppe IoP_B 1 mit Vorsicht zu interpretieren ist. Auch die Häufigkeit der Disparität variierte zwischen Blocks. Während drei Viertel der mehrheitlich spontanen (IoP_B 2) UTG eine signifikante Disparität aufwiesen, war dies nur bei einer der sieben hochgradig geplanten UTG der Fall.

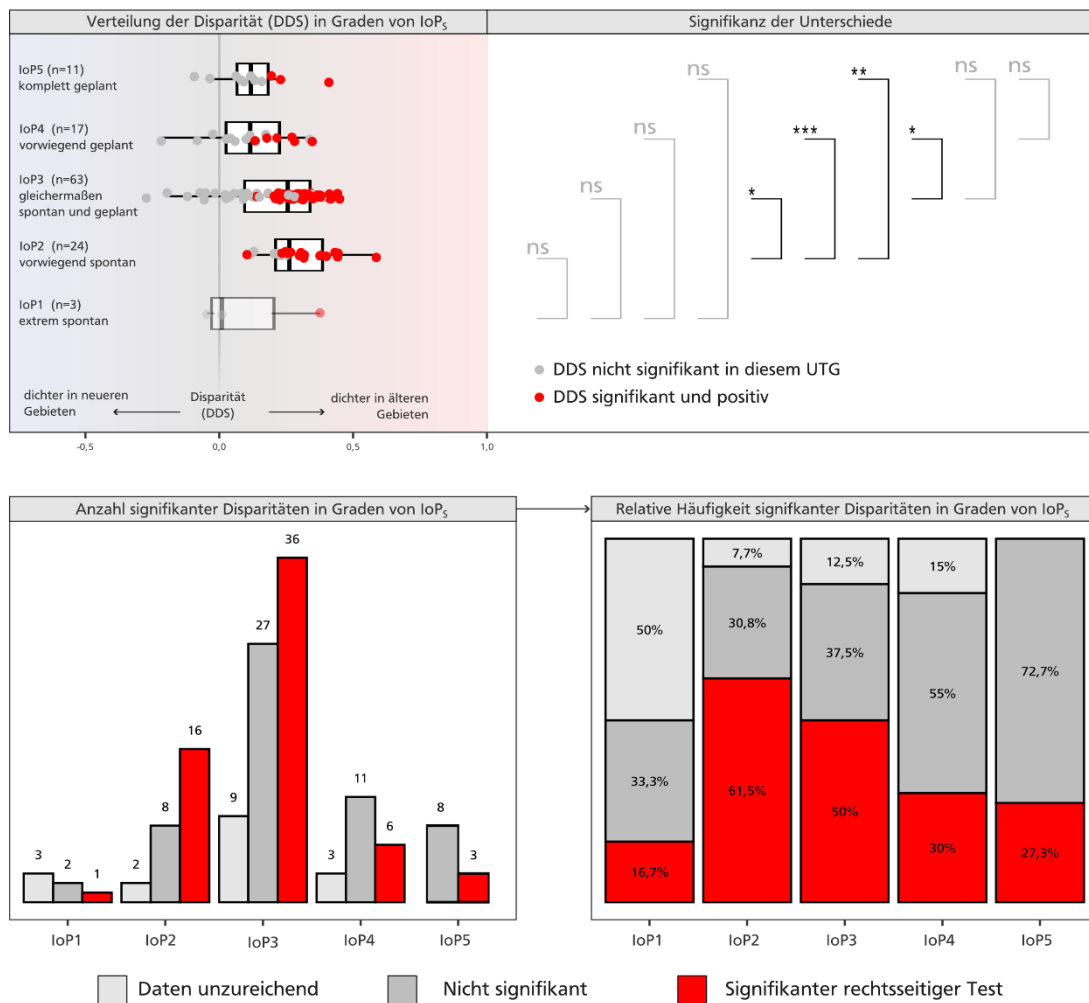


Abbildung 31: Vergleich der Disparität (gemessen durch DDS_{UTG}) nach IoP_B . Oben: Vergleich des DDS_{UTG} mittels parametrischer ANOVA. DDS-Unterschiede zwischen Gruppen wurde mittels t-test erfasst und wie folgt dargestellt: ns: kein signifikanter Unterschied; * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$; **** $p < 0,0001$. Unten: Prävalenz (Anzahl der UTG) und Häufigkeit (Anteil der UTG) von signifikanten DDS_{UTG} innerhalb jeder Gruppe.

Auf der Ebene von Straßennetzwerken ergibt sich ein sehr ähnliches Bild. Die ANOVA über Grade von IoPs hinweg ist signifikant ($p = 0,003$) mit einer moderaten Effektstärke ($\omega^2 = 0,099$). Es gibt keine extremen Ausreißer, keine Hinweise auf Heteroskedastizität (Levene-test p -Wert = 0,593) und die Verteilungen sind ungefähr normalverteilt mit Ausnahme von IoPs 1, einer kleinen Gruppe, von welcher nur die Hälfte der UTG ausreichende Daten aufweist. Auch die nichtparametrische ANOVA bestätigt dieses Bild (Kruskal-Test p -Wert = 0,004, $\eta^2 = 0,1016$). Der zweiseitige post-hoc T-Test zeigt die stärkste Signifikanz für Unterschiede zwischen IoPs 2 und IoPs 4 ($p = 0,0006$) und zwischen IoPs 2 und IoPs 5 ($p = 0,0027$). Die Häufigkeit der Disparität variiert auch auf der Ebene von Straßen zwischen IoPs. Während die hauptsächlich ungeplanten (IoPs 2) UTG größtenteils Disparitäten aufweisen (16 von 24 UTG, also ~67 %) ist dies nur bei 9 der 28 (~32 %) geplanten UTGs (IoPs 4 oder IoPs 5) der Fall.

Insgesamt stützen die statistischen Tests die Hypothese, dass die digitale Disparität in ihrer Ausprägung mit der Geplantheitsintensität zusammenhängt.

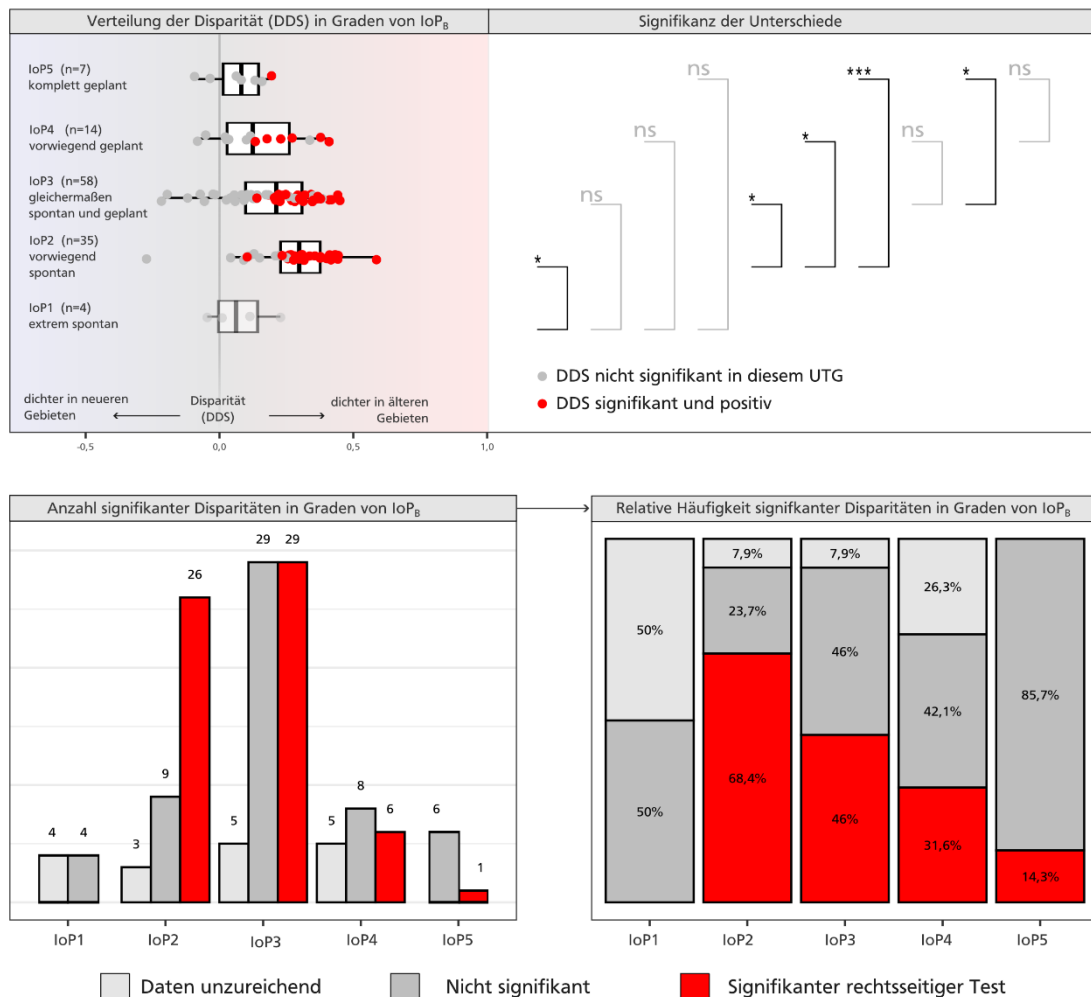


Abbildung 32: Vergleich der Disparität (gemessen durch DDS_{UTG}) nach IoP_B. Oben: Vergleich des DDS_{UTG} mittels parametrischer ANOVA. DDS_{UTG} -Unterschiede zwischen Gruppen wurde mittels t-test erfasst und wie folgt dargestellt: ns: kein signifikanter Unterschied; * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$; **** $p < 0,0001$. Unten: Prävalenz (Anzahl der UTG) und Häufigkeit (Anteil der UTG) von signifikanten DDS_{UTG} innerhalb jeder Gruppe.

6.1.4. Modelliert im Zusammenhang mit Räumlichen Metriken

Zwischen DDS_{UTG} und den 14 räumlichen Metriken werden lineare Modelle geschätzt. Von diesen vierzehn sind elf Beziehungen signifikant bei einem 95% Signifikanzniveau (Abbildung 33). Jedoch ist die Bestimmtheit der Beziehungen relativ gering ($R^2 = 0,09\text{--}0,20$) was nahelegt, dass andere Faktoren die Ausprägung der Disparität beeinflussen.

Bei den Regressionen, welche sich nur auf UTG mit signifikanten Disparitäten beziehen, erweisen sich nur 9 von 14 Modellen als signifikant ($R^2 = 0,04\text{--}0,22$). Im Detail gibt es keine statistisch signifikanten linearen Zusammenhänge zwischen DDS_{UTG} und Urbaner Dichte, Porositätsindex und SPI ($>2\text{km}$). Dies impliziert, dass Dichte und Kontinuität der Bebauung nicht direkt die Ausprägung der Disparität bedingen. Jedoch zeigen andere Metriken durchaus Zusammenhänge: DDS_{UTG} ist signifikant korreliert mit der Flächengewichteten Standarddistanz ($R^2 = 0,17$, $p < 0,01$), dem Dispersionsindex ($R^2 = 0,20$, $p < 0,01$), der Urbanen Kompaktheit ($R^2 = 0,18$, $p < 0,01$), der Objektdichte ($R^2 = 0,09$, $p < 0,01$), der Mittleren Patch Größe ($R^2 = 0,09$, $p < 0,01$), der Effektiven Netzgröße ($R^2 = 0,07$, $p < 0,01$), und dem SPI in Fenstern von $<100\text{ m}$ ($R^2 = 0,12$, $p < 0,01$), $100\text{--}500\text{ m}$ ($R^2 = 0,14$, $p < 0,01$) und $500\text{--}2000\text{ m}$ ($R^2 = 0,12$, $p < 0,01$).

Dies zeichnet ein Bild, in welchem ein höherer Grad an Zentralität, Kompaktheit und größere Patches mit höherem DDS_{UTG} einhergehen, während eine größere Anzahl von kleinen und dispersen Siedlungsteilen mit niedrigerem DDS einhergeht.

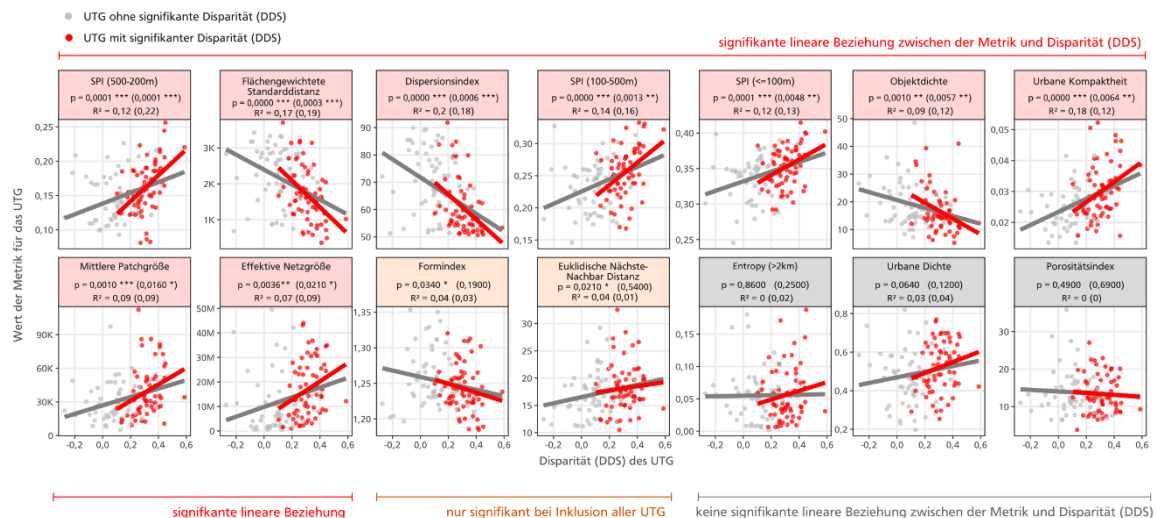


Abbildung 33: Stärken der linearen Beziehung zwischen Disparität (DDS_{UTG}) und räumliche Metriken der Siedlungsstruktur, gemessen mit dem Bestimmtheitsmaß (R^2) und P-Wert (p) der jeweiligen linearen Regression. Signifikanz ist wie folgt angegeben: * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$. Werte in Klammern beziehen sich auf alternative Modelle, die nur UTG mit signifikanten DDS_{UTG} einbeziehen.

Diese Regressionen beziehen sich jeweils auf einfache lineare Zusammenhänge zwischen DDS_{UTG} und einer einzelnen räumlichen Metrik. Die multiple lineare Regression erlaubt die Inklusion von mehreren Metriken (Abbildung 34). Unter diesem Modell erklären die Metriken Dispersionsindex und euklidische Nächste-Nachbar-Distanz zusammen etwa ein Viertel der Varianz in DDS_{UTG} ($p < 0,001$, multiples $R^2 = 0,2911$, korrigiertes $R^2 = 0,2724$, Tabelle 5). Dieser Wert ist deutlich höher als der einer einzelnen Metrik, was zeigt, dass eine Kombination von komplementären räumlichen Metriken stärker mit DDS_{UTG} assoziiert ist.

Insgesamt implizieren die Korrelationen, dass kompaktere Siedlungen (hohe Kompaktheit und SPI, niedriger Dispersionsindex) bei gleichzeitig hohem Abstand zwischen patches (hohe euklidische Nächste-Nachbar-Distanz) und wenigen Öffnungen im Siedlungsbereich (hohe urbane Dichte) mit größerer Ungleichheit korrelieren. So hätte eine typische Siedlung mit einer Disposition zur digitalen Disparität einen großen und dichten urbanen Kern, welcher von mehreren isolierten Subzentren umgeben ist, die weit voneinander entfernt liegen.

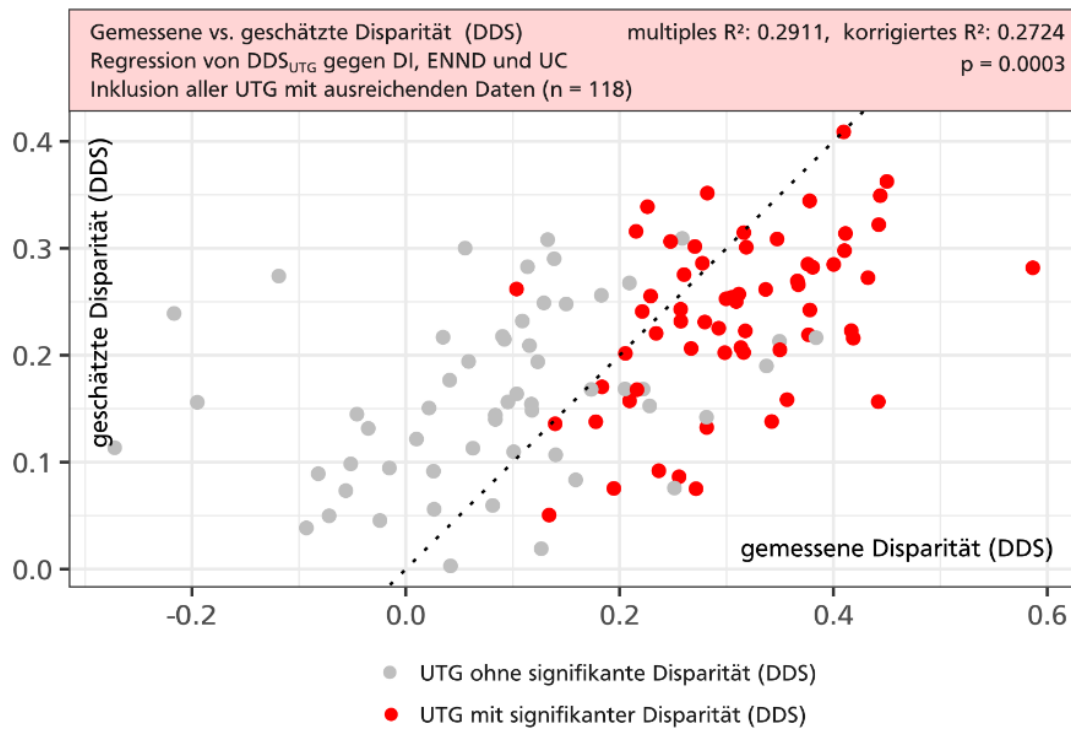


Abbildung 34: Vergleich von beobachteten DDS_{UTG} und durch das multiple lineare Regressionsmodell geschätzten DDS_{UTG}

Tabelle 5: Regressionskoeffizienten zwischen DDS_{UTG} und räumlichen Metriken. Signifikanz ist angegeben als: * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$. Multiples $R^2 = 0,2911$, korrigiertes $R^2 = 0,2724$.

Term	Koeffizient	Standardfehler	t	p (Signifikanz)
Konstante (Intercept)	0,296	0,113	2,626	0,0098 **
Dispersionsindex	-0,005	0,001	-4,077	0,0001 ***
Euklidische Nächste-Nachbar-Distanz	0,007	0,003	2,218	0,0290 *
Urbane Kompaktheit	4,000	1,845	2,168	0,0320 *

6.1.5. Interpretation

Insgesamt zeigt die Verschneidung von fernerkundlich abgeleiteten Bebauungskartierungen und Geotweets deutlich, dass das Alter eines urbanen Raumes im Zusammenhang mit der Dichte der Geotweet Aktivität stehen kann. Diese Beziehung ist nicht überall eindeutig und vergleichbar stark ausgeprägt, aber in fast der Hälfte der untersuchten Gebiete messbar.

Aus einer stadtgeographischen und gesellschaftlichen Sicht implizieren diese Ergebnisse die Existenz einer Art von digitaler Disparität. Neuere Nachbarschaften sind auf Twitter –und

möglicherweise auch auf anderen GSM-Quellen– weniger sichtbar als ältere Nachbarschaften derselben Stadt. Es mag sein, dass diese als eine neue Art von Ungleichheit zu betrachten ist, die sich aus dem rapiden Wachstum dieser Siedlungen ergibt, oder lediglich die räumliche Manifestation von bereits bekannten sozioökonomischen Ungleichheiten ist.

Vermutlich sind sozioökonomische Variablen, neben technischen, die wahrscheinlichsten Faktoren, welche die Dichte von GSM in Siedlungsgebieten beeinflussen. Basierend auf der Inspektion in dieser Studie vorliegenden Daten liegt nahe, dass ein großer Einfluss von der Dichte von Bevölkerung, wirtschaftlicher und kultureller Aktivität und technischer Infrastruktur ausgeht. Es ist plausibel, dass diese sich in neueren Stadtteilen erst entwickeln müssen –dass der physischen Urbanen Frontier eine digitale Urbane Frontier erst mit einiger Verzögerung folgt. Damit einher geht die Frage nach den Funktionen der Räume. Während Zentren oft eine hohe Konzentration von Arbeitsplätzen, Behörden und Einkaufsmöglichkeiten enthalten, sind andere Siedlungsbereiche teilweise fokussiert auf Wohnraum oder Gewerbe. Die Funktionen bedingen die Aktivitäten der Nutzenden dort, welches wiederum das Online-Verhalten der Nutzenden beeinflusst. Es ist wahrscheinlich, dass neue Siedlungsgebiete nicht exakt dieselben funktionalen und kulturellen Profile entwickeln wie die älteren Siedlungsgebiete, was indirekt auch das Potenzial für digitale Sichtbarkeit beeinflusst. Aufgrund der kulturell sehr unterschiedlichen Siedlungsgeographien Afrikas ist dieser Effekt jedoch an dieser Stelle nicht eindeutig zu fassen. Die Analyse von der in den Tweets genannten Themen, wie sie in Studie A2 durchgeführt wurde, könnte hierfür einen wertvollen Beitrag leisten.

Die zugrundeliegenden Kausalitäten genau zu untersuchen, erfordert zudem fein aufgelöste sozioökonomische Daten, welche in dieser Studie nicht zur Verfügung standen. Tatsächlich ist zweifelhaft, ob entsprechende Daten für ganz Afrika erhoben werden könnten, insbesondere bei der feinen Auflösung die nötig wäre, nämlich auf der Ebene von Gebäuden oder Gebäudeblocks. Und, selbst wenn diese verfügbar wären, müssten sie für mehr als den hier genutzten 135 Sichtproben analysiert werden, um die große Diversität von Siedlungsformen und Siedlungswachstum in Afrika abzudecken.

Angesichts dieses Data-Gaps stellt die hier genutzte Ableitung von Proxy-Variablen aus FED eine Möglichkeit dar, einen ersten Eindruck der zugrundeliegenden Faktoren zu gewinnen.

Die hiermit durchgeführten ersten Experimente, welche geographische Regionen und FED als Proxy für sozioökonomische Prozesse nutzen, zeigen, dass diese in Beziehung zur Ausprägung und Häufigkeit der Disparität stehen könnten. Für die Erforschung des Phänomens und seiner kausalen Zusammenhänge können diese ersten Ergebnisse grundsätzlich richtungsweisend sein. Beispielsweise wurden Unterschiede zwischen Regionen Afrikas identifiziert, wobei die Disparität in Westafrika größer und häufiger ist als im Nördlichen und Südlichen Afrika. Es ist möglich, dass dies an der höheren Verbreitung von Internet- und Kommunikationstechnologien in diesen Regionen liegt (Nchake und Shuaibu 2022), oder an dem unterschiedlichem Tempo der Urbanisierung, welches in Nordafrika und Südafrika geringer ist (UN DESA 2019), was eventuell die Ungleichheiten abschwächt.

Auch für die Struktur der Siedlungen legen die Ergebnisse einen Zusammenhang mit der Ausprägung der Digitalen Disparität nahe. Ein höherer Grad von Geplantheit der Siedlung geht mit einer geringen Disparität einher. Auch besteht ein Zusammenhang zur Morphologie der Siedlung, wobei kompakte Siedlungen mit einem dominierenden großen urbanen Kern

zur Disparität neigen. Die Aussagekraft der Modelle ist dabei recht begrenzt (geringes R^2), was nahelegt, dass die Siedlungsstruktur nicht direkt die Ungleichheit bedingt, sondern dass andere Faktoren im Zusammenhang mit Planungsgrad und Siedlungsstruktur die weitaus größere Rolle spielen. Möglicherweise hat die Verfügbarkeit von finanziellem und kulturellem Kapital und die Bevölkerungsdichte einen Einfluss. Mittels geeigneter sozioökonomischer Daten ließen sich diese Einflüsse entwirren. Auch wenn hochaufgelöste sozioökonomische Mikrodaten nicht konsistent auf kontinentaler Ebene erhoben werden können, könnten auf Interviews basierende Umfragen und stichprobenartige Studien die Datenlücken füllen. Diese Methoden werden vermutlich sehr relevant für die Erforschung von Ungleichheiten sein und bleiben. In Bezug auf die digitale Urban Frontier könnten existierende Umfragen und Zensuserhebungen um Fragen zur Internetnutzung erweitert werden

Aus einer Datenperspektive implizieren die Ergebnisse, dass die neu bebauten Gebiete afrikanischer Städte, die von hoher Bedeutung sind, um die Bevölkerungsentwicklung der Städte abzufangen, leider nicht in gleichem Maße mit GSM untersucht werden können wie die älteren Stadtteile. Bezüglich der Datenverfügbarkeit allgemein war festzustellen, dass in 12,6% der UTG keine ausreichende Datenlage für die Analyse vorlag.

Jedoch bedeutet dies nicht, dass die Datenlage grundsätzlich überall schlecht ist: Für 63 der UTG sind mindestens 10 Tsd. Geotweets auch in den neu entstandenen (seit 1985) Gebieten vorhanden, was angesichts der Größe der untersuchten Expansionsgebiete durchaus beachtlich ist. Die neuen Gebiete stellen daher nicht grundsätzlich *Digital Deserts* dar (Taubenböck et al. 2018). Sie könnten aber in Studien, welche sich auf GSM berufen, um Siedlungen zu erforschen, unterrepräsentiert sein. Datenverfügbarkeit sollte also bei der Konzipierung von Studien nicht grundsätzlich angenommen, sondern getestet werden, beispielsweise durch Anfragen an die API oder den Kundenservice der anvisierten Plattform. Zudem sollte, auch wenn ausreichende Daten verfügbar sind, überprüft werden ob Ungleichheiten in der räumlichen Verteilung einen Einfluss auf die gewählte Methode haben.

Die Ergebnisse müssen in Bezug auf einige Limitationen interpretiert werden. Die untersuchten UTG wurden speziell aufgrund ihres ausgeprägt hohen Siedlungsflächenwachstums ausgewählt, weswegen die Ergebnisse der Studie nicht auf langsamer wachsende urbane Regionen extrapoliert und generalisiert werden können. Auch ist die Einstufung eines UTG von fast 80 km² in einen einzelnen IoP-Grad eine starke Vereinfachung der Realität. Ein fixer Radius verbessert zwar die in dieser Studie angestrebte Vergleichbarkeit zwischen Testgebieten, reflektiert nicht die tatsächlichen morphologischen und funktionalen Grenzen in der Stadt.

Die Qualität der Daten ist nicht überall perfekt und variiert zwischen Regionen, was Vergleiche zwischen Regionen beeinflussen könnte. Die FED sind aufgrund technischer Entwicklungen in früheren Jahren von geringerer Qualität und erlauben es nicht, Bebauungszeiträume früher als 1985 zu erfassen. Auch für die GSM ist zu berücksichtigen, dass die Präzision der Geotagging-Funktion nicht perfekt ist. Merry und Bettinger (2019) berichten beispielsweise von einem horizontaler Fehler von 7-13m bei Smartphones in städtischer Umgebung. Die Tests der Studie (siehe Appendix A3) belegen jedoch, dass dies zwar einen kleinen Einfluss auf die Ergebnisse, aber keinen Einfluss auf die insgesamt Schlussfolgerung der Studie hat. Die Datenverfügbarkeit für manche UTG ist teilweise gering, wodurch, wie bereits angesprochen, für manche UTG kein Ergebnis berechnet wurde. Insbesondere gibt es in manchen Orten nur wenige beitragende Accounts. Dies verstärkt das

bekannte Problem der Ungleichverteilung von Inhalten of Twitter, wo eine kleine aktive Mehrheit der Accounts die meisten Tweets verfasst (Wojcik und Hughes 2019; Lemoine-Rodríguez et al. 2024). Da in der Studie ein Bild der Sichtbarkeit gezeichnet werden soll, wurde der Einfluss dieser Accounts mittels eines Gewichtungsansatzes limitiert. Dennoch ist eine wichtige Limitation, dass das Bild der digitalen Aktivität, welches in dieser Studie gezeichnet wird, nur eine von vielen möglichen Perspektiven darstellt, und von den Nutzenden einer bestimmten Plattform geprägt ist, die Internetnutzung im Allgemeinen repräsentieren.

Um weitere Perspektiven abzubilden könnten vielseitigere Daten genutzt werden, aber auch verschiedene Methoden. Über Geotags wird primär die Aufmerksamkeit gemessen, welcher einem Ort von demselben Ort aus gegeben wird. Demgegenüber kann Geoparsing (Kapitel 2.1.4) genutzt werden, um die Rezeption eines Ortes an anderen Orten miteinzuschließen. Geoparsing beruht jedoch auf existierenden Ortsverzeichnissen, deren Verfügbarkeit oft eingeschränkt und regional stark unterschiedlich ist, was sie für diese vergleichende Studie unbrauchbar macht (Graham und De Sabbata 2020). Einige der Unterschiede zwischen Geotweet und Geoparsing Ansätzen werden in Studie C2 ausgeleuchtet.

6.2. Studie A2: Verbindung von Fernerkundungsdaten und Twitter-Daten zur Ableitung von Raumeinheiten für die explorative Analyse von physischen und thematischen Charakteristika urbaner Räume

Für eine Auswahl der UTG wurden die FED und GSM-Daten genutzt, um in Studie A2 mittels eines Clustering-Ansatzes Raumeinheiten für die explorative Analyse urbaner Räume zu erstellen und zu bewerten.

6.2.1. Clustererstellung und Interpretation

Die Cluster werden, Experimenten und visueller Interpretation folgend, mit Gewichtungen von 20% für β_{GEO} , 20% für β_{FE} und 60% für β_{TW} erstellt. Ein Vergleich von verschiedenen Gewichtungen ist dieser Dissertation angehängt (Anhang A4). Durch die gemischte Gewichtung werden Cluster erstellt, welche sowohl räumliche Kompaktheit als auch Varianz in Baujahr und Textinhalten berücksichtigen. Abbildung 35 zeigt anhand von sechs hervorgehobenen Clustern eines UTG, dass sich die Verteilungen von Bebauungsjahren sowie die Verteilung von Themen zwischen den Clustern sichtbar unterscheiden.

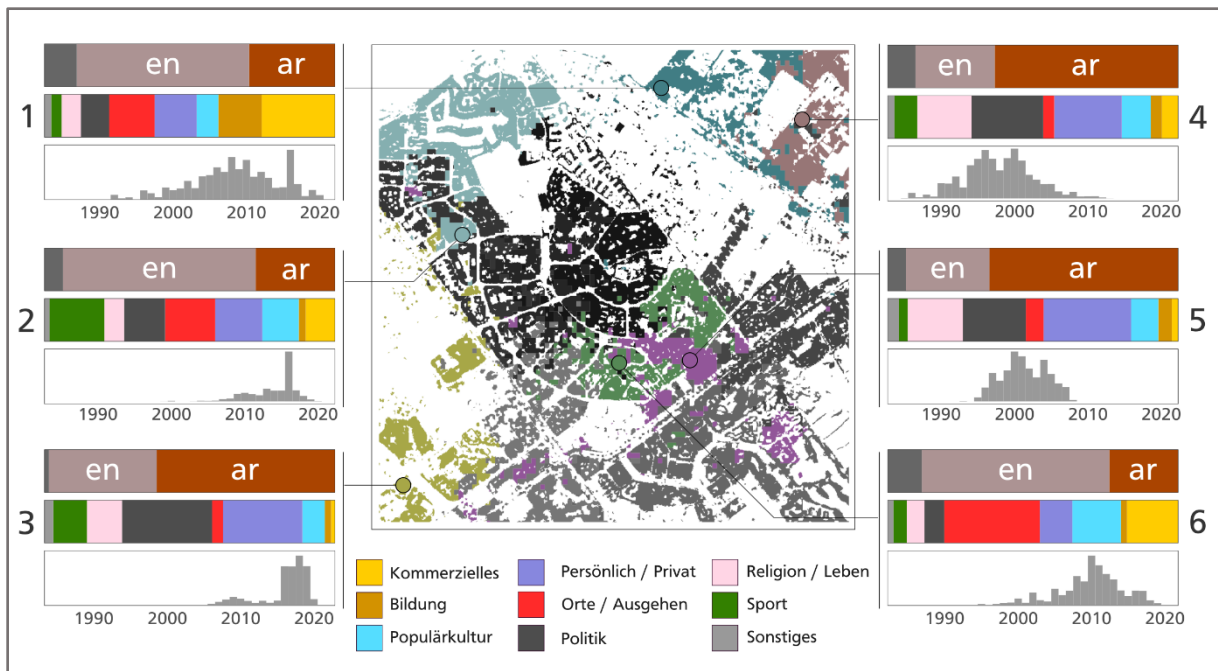


Abbildung 35: Darstellung der räumlichen Cluster für UTG A (Sheikh Zayed Stadt, Ägypten), sowie die Verteilung von Sprachen und Themen für ausgewählte Cluster (1-6), zusammen mit einem Histogramm der Bebauungsjahre der Cluster

Diese Unterschiede erlauben es, Cluster zu charakterisieren und zu kontrastieren. Einige Cluster wurden beispielsweise in relativ kurzen Zeitspannen erbaut (2 und 3) während andere über einen längeren Zeitraum entstanden sind (insbesondere Cluster 1). Bemerkenswert ist, dass sich auch nahe beieinander liegende Cluster gut abgrenzen lassen (1 und 4, 5 und 6). Cluster 5 und 6 sind räumlich fast verwoben, weisen jedoch erhebliche Unterschiede in den kartierten Charakteristika auf. So dominieren in Cluster 5 Tweets über Themen wie Religion, Privatleben und Politik, welche größtenteils auf Arabisch verfasst sind. Demgegenüber sind in dem benachbarten Cluster 6 englischsprachige Tweets über Gastronomie, Ausgehen, Populärkultur und Konsum prävalent. Mittels einer Webrecherche können diese Eindrücke bestätigt werden. Diese zeigt, dass Cluster 6 tatsächlich mehrere bedeutende Freizeit- und Einkaufszentren beherbergt, beispielsweise die *Mall of Arabia* (Aggour 2014) und die Fußgängerzone *Arkan Plaza* (Badreldin Group 2024). Die Bebauung der beiden Cluster erfolgte auch zu unterschiedlichen Zeiten. Während Cluster 5 größtenteils um das Jahr 2000 bebaut wurde, erfolgte die Bebauung von Cluster 6 größtenteils um 2010. Ähnlich starke Unterschiede zeigen sich im Nordosten des UTG. Cluster 1 ist als Standort der *Smart Village* (Smart Village 2022) Ort von englischsprachigen Tweets über Bildung und Geschäftliches, das tendenziell früher bebaute benachbarte Cluster 4 weist hingegen mehr arabischsprachige Tweets über private oder politische Themen auf.

Die Webrecherche bestätigt somit die Plausibilität und qualitative Interpretierbarkeit der Cluster. Es lässt sich so beispielsweise zusammenfassen, dass Cluster 6 ein um das Jahr 2010 erbauter Bereich von öffentlicher und größtenteils kommerzieller Konsum- und Freizeitaktivität ist, welcher zu einem gewissen Grad von internationalem Tourismus geprägt wird.

Die qualitative Analyse zeigt auch, dass einige der durch das Topic-Modeling generierten Themen fast ausschließlich Links und Erwähnungen von Personen oder Accounts enthalten. In der manuellen Inspektion ist es teilweise möglich, auch diese zu interpretieren, wenn

Accountnamen beispielsweise bekannten Personen des öffentlichen Lebens zugeordnet werden können. Daher sind diese Inhalte nicht grundsätzlich von der Analyse ausgeschlossen, sondern manuell inspiziert und in übergeordnete Themen aggregiert. Zur Unterstützung dient ein lokal betriebenes Sprachmodell *Llama-2-7b-chat-hf* welches zur Annotation und zur Übersetzung eingebunden wird (Touvron et al. 2023). Bei der Analyse der den Clustern zugewiesenen Texte ist bemerkenswert, dass dieses Sprachmodell wie auch das auf BERTweet basierte Topic-Modeling keine interpretierbaren Zusammenhänge für Texte herstellen können, welche in Arabizi geschrieben sind, einer Methode, Arabisch mit lateinischen Buchstaben zu schreiben (beispielsweise حبيبي als *7abibi*, Tobaili 2016). Die Sprachmodelle können Tweets in dieser Schreibweise offenbar nicht korrekt in Embeddings überführen, was für einen Teil der Texte zu schlechtem Verständnis und nicht interpretierbaren Themen führt. Für Arabisch, Englisch und Französisch war dies nicht der Fall. Für diese Sprachen sind die generierten Themen über Sprachgrenzen hinweg interpretierbar.

Diese hier für ein UTG dargestellte Einschätzung bestätigt sich in allen getesteten UTG, wobei die Variabilität bei UTG mit weniger Tweets (UTG D und E) ungleich höher war.

6.2.2. Statistische Analyse der Clustergüte

Eine derartige Post-hoc Analyse beweist allein nicht, dass das Clustering einen guten Ausgleich zwischen verschiedenen Dateninputs erreicht. Daher werden zusätzlich die Statistiken für den gewählten (gemischten) Ansatz verglichen mit un-gemischten Ansätzen, die nur auf einer Art von Inputs basieren.

Zunächst zeigt diese statistische Analyse, dass das beste Clustering für jedes Kriterium immer das Clustering ist, welches auf Basis der relevanten Daten durchgeführt wurde. Die Analyse zeigt auch, dass der gemischte Ansatz in 21 von 24 Fällen über UTG hinweg die zweitbeste Clustergüte erreicht. Hierbei gibt es einzelne Ausnahmen. So erreicht der auf geographischer Distanz basierte Ansatz (grün in Abbildung 36) in UTG C eine höhere Reduktion der Themenentropie, ist jedoch weitaus schlechter für UTG B, D und E. Umgekehrt führt der auf Textinhalten basierende Clustering Ansatz (blau in Abbildung 36) für UTG D und E zu kompakteren Clustern als der gemischte Ansatz, funktioniert jedoch deutlich schlechter in UTG A, F, G und H. Der gemischte Ansatz stellt dabei über UTG hinweg einen robusten Kompromiss dar.

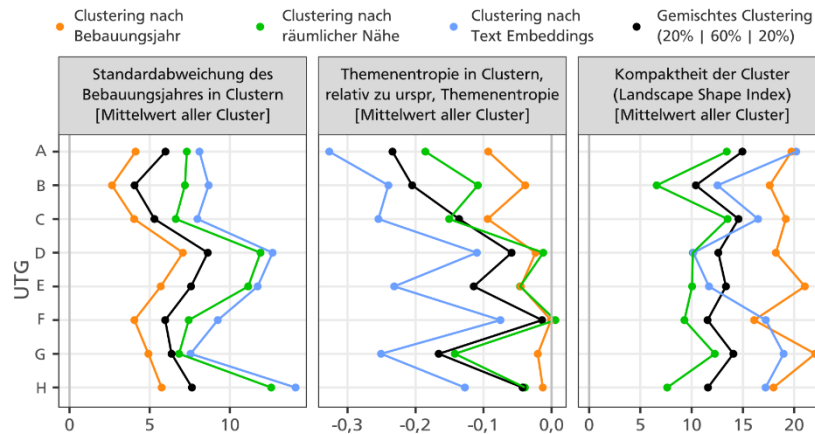


Abbildung 36: Verteilung der statistischen Evaluationskriterien über UTG und die vier verglichenen Clusterparametrisierungen hinweg. Kriterien sind so skaliert, dass ein geringerer Wert (links) ein besseres (homogeneres) Cluster andeutet.

6.2.3. Interpretation

Dass die Verteilungen von Baujahr, Sprachen und Themen so unterschiedlich waren (Abbildung 35) auch wenn sie nicht direkt in die Clusterbildung eingegangen sind, kann als positives Indiz für die Nützlichkeit des Ansatzes gewertet werden. Die entstandenen Raumeinheiten entsprechen zwar keinem eindeutigen Konzept wie Bezirk, Nachbarschaft, oder Bauprojekt, können aber sinnvoll interpretiert werden. Das legt nahe, dass die Nutzung einer Kombination von FED und GSM zu Clustern führen kann, welche auch für andere Anwendungen sinnvoll sein können.

Eine Evaluation des Nutzens in der praktischen Anwendung steht noch aus. Die Diskussion mit Experten des geographischen Kontexts, Stakeholdern und Bewohnern der Stadt wären sinnvolle nächste Schritte, um die Praktikabilität der Methode zu evaluieren. Basierend auf den in dieser Studie erfolgten Tests ist der Ansatz für die praktische Anwendung geeignet. Hierbei müssen Limitationen berücksichtigt werden, wie die Verfügbarkeit von Daten (siehe Studie A1 in 6.1), die Verfügbarkeit von Sprachmodellen, welche für die Sprachen der gewählten Untersuchungsregion geeignet sind und die Variabilität der Clusterergebnisse in Bezug auf die Parametrisierung des Ansatzes.

Die Mehrsprachlichkeit der in diesem Experiment genutzten Sprachmodelle ist dabei bereits durchaus positiv zu bewerten und erlaubt die Integration von Textdaten verschiedener Sprachen und Sprachvarietäten. Dennoch zeigt die Ausnahme der Arabizi-Schreibweise, dass noch Herausforderungen für die Integration von Sprachen und Schreibweisen bestehen. Insbesondere im Migrationskontext ist auch die Repräsentation von auch informellen und datenarmen Sprachen von Bedeutung.

Bei der Anwendung muss auch unterstrichen und gegebenenfalls an Beteiligte kommuniziert werden, dass die Wahl der Parameter eine hohe Auswirkung auf das Bild hat, welches sich von der Stadt ergibt. Diese Anpassbarkeit erlaubt ein hohes Maß an Flexibilität, was im explorativen Kontext von Vorteil sein kann. Der Ansatz lässt sich auf weitere Inputdaten

erweitern, so könnten beispielsweise Bilder⁹² und Videos in Embeddings überführt und genau wie die Text-Embeddings in das Clustering integriert werden.

Auch wenn der Ansatz also vielversprechend für diverse explorative Anwendungen ist, wäre es nicht sinnvoll, die erstellten Cluster für analytische Auswertungen zu nutzen. Hierfür sollten, sobald die Forschungsfragen und Daten klar definiert sind, klar konzipierte Raumeinheiten herbeigezogen oder notfalls vor dem Hintergrund der Forschungsfragen händisch definiert werden. Alternativ könnten Methoden für die optimierte Parameterauswahl, wie sie beispielsweise im R Paket *geomeans* (Jeremy 2023) implementiert sind, die analytische Anwendung unterstützen.

Auch um eine universelle Repräsentation der städtischen Umgebung herzustellen, ist dieser Ansatz nicht das Mittel der Wahl. Vielversprechender hierfür sind die in Kapitel 2.3.1 beschriebenen Ansätze aus dem Bereich des selbstüberwachten Maschinellen Lernens und der räumlichen Basismodelle (engl. *geospatial foundation models*), welche mittels großer und diverser Datenmengen trainiert und auch für die Anwendung der Abgrenzung von Regionen optimiert werden können (Chen et al. 2024).

Von räumlichen Basismodellen generierte Embeddings (Google, o.D.; Brown et al. 2025) könnten auch als Eingangsdaten für das hier demonstrierte Clustern genutzt werden. Damit könnte eventuell eine ähnliche Verbindung erzielt werden wie im hier demonstrierten Ansatz, wobei durch die Auslagerung der Datenfusion in das Training eine geringere Transparenz und Kontrolle, aber deutlich höhere Geschwindigkeit in der Anwendung zu vermuten ist.

Demgegenüber zeichnet sich durch der hier demonstrierte Ansatz durch Flexibilität, Transparenz und Modularität aus. Er basiert auf etablierten Konzepten und out-of-the-box Softwarelösungen und kann aufgrund geringer Ansprüche an Datenumfang und Prozessierungsleistung mit einer handelsüblichen Desktop-GPU durchgeführt werden. Durch die Modularität des Ansatzes können zukünftig Entwicklungen aus verschiedenen benachbarten Forschungsfeldern integriert werden⁹³. Damit soll diese Studie einen Beitrag leisten, die Anforderungen der interdisziplinären Migrationsforschung in der multimodalen Datenanalyse zu unterstützen.

Die Bedeutung für diese Dissertation liegt in der Demonstration eines Ansatzes, durch gemeinsame Nutzung von FED und GSM-Daten eine Grundlage für die Aggregation, Visualisierung und Interpretation verschiedener Datenquellen zu schaffen.

6.3. Studie B1: Ableitung inhaltlicher Präferenzen von nigerianischen Twitter-Nutzenden nach Mobilitätscharakteristika

Anhand der Ergebnisse von Studie B1 kann das Potenzial von Mobilitätsanalysen auf GSM-Basis und in Kombination mit Textinhalten erörtert werden.

⁹² Insbesondere vom Straßenraum aus aufgenommen Bilder könnten im urbanen Kontext sehr vielversprechend sein, siehe Biljecki und Ito (2021).

⁹³ Beispielsweise könnten von räumlichen Basis-Modellen generierte Embeddings (Google, o.D.; Brown et al. 2025) als generische FED-Eingangsdaten für das Clustern genutzt werden.

6.3.1. Mobilitätscharakteristika

Der initiale Datensatz S_{NGA} enthält 325 061 Accounts mit 27 818 148 Geotweets innerhalb Nigerias. Aus diesen eignen sich 116 670 Accounts U_{NGA} mit insgesamt 25 021 363 Geotweets für die Trajektorienanalyse.

Diese erkennt 10 101 M_{OVE} , von welchen 50,9% innerhalb Nigerias und 11,4% außerhalb Nigerias verlaufen, 18,8% aus Nigerias herausführen und 18,4% nach Nigeria hineinführen (Abbildung 37). Von den 1 057 detektierten M_{INT} führen 48,0% nach Nigeria hinein und 52% hinaus. Die meisten dieser Bewegungen knüpfen an Nigerias Hauptstadt Abuja oder die größte Stadt Lagos an. Außerhalb Nigerias sind das Vereinigte Königreich und die Vereinigten Staaten von Amerika die stärksten Anknüpfungspunkte.

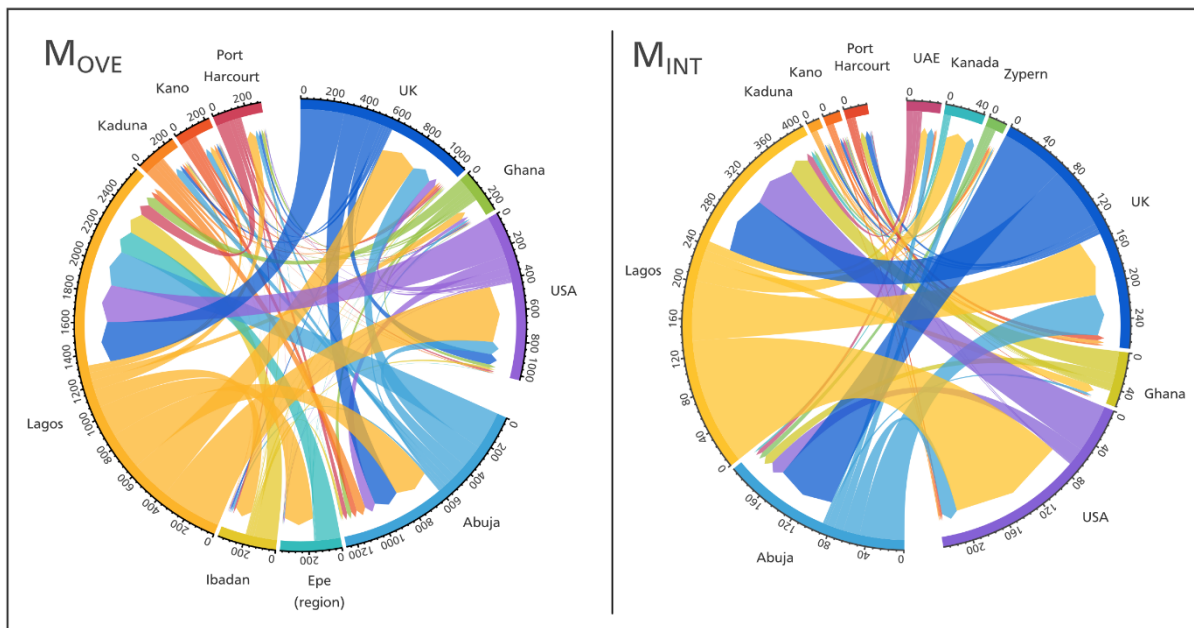


Abbildung 37: Links: M_{OVE} zwischen Raumeinheiten mit mindestens 250 M_{OVE} . Rechts: M_{INT} zwischen Raumeinheiten mit mindestens 25 M_{INT} .

Anhand der M_{OVE} und M_{INT} lassen sich 9 762 Accounts U_{AN} mit hoher Gewissheit einem der drei Mobilitätsklassen zuordnen (Tabelle 6). Von diesen sind 4 794 Mobile (U_{30}), von denen 3 229 (67,4%) nur Mobilität innerhalb Nigerias aufweisen. Für etwa ein Drittel (35,3) der U_{30} lassen sich mehrere M_{OVE} erkennen. Die Zahl der International Migrierten U_{90} betrug 926. Für 111 (12%) von diesen wurden mehrere internationale Bewegungen (M_{INT}) registriert. Die Zahl der stationären Nutzenden (U_0) betrug 3 857. Die meisten dieser Nutzenden haben ihre erfassten Aufenthaltsorte in Lagos (49,2%), Abuja (20,0%), Kaduna (3,5%), Ibadan (3,3%) und Port Harcourt (3,2%).

Tabelle 6: Statistiken der Nutzenden im Satz U_{AN} , nach Mobilitätstyp

	Anzahl der Nutzenden	N _{Tweets}				Gyrationsradius [km]			
		min	max	mittel	σ	min	max	mittel	σ
U_0 (In Nigeria Stationäre)	3 748	13	480	282,4	128,7	0,6	5 879,7	413,2	714,7
U_{30} (Mobile)	4 655	1	480	276,4	128,0	1,5	8 204,4	864,4	1 258,9
U_{90} (International Migrierte)	898	22	480	316,9	117,9	66,4	10 045,5	2 478,4	1 463,7

Als Ergänzung zu dieser kategorischen Beschreibung der Mobilität zeigt Abbildung 38 die Verteilung der Mobilität mittels des Gyrationsradius der Nutzenden. Es zeigt sich, dass sich die Gruppen anhand des Gyrationsradius nicht klar voneinander trennen lassen. Ferner ist interessant, dass die Verteilung klar bimodal ist, und speziell die Gruppe der Mobilten Nutzenden zwei Spitzen hat, welche ungefähr bei 300 km und 3 000 km liegen.

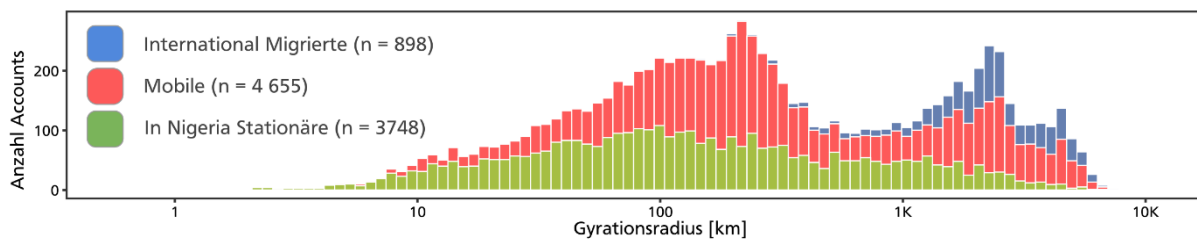


Abbildung 38: Histogramm des Zehnerlogarithmus von Gyrationsradien der Nutzenden, eingefärbt nach Mobilitätstyp

6.3.2. Inhaltliche Präferenzen

U_0 , U_{30} und U_{90} bilden zusammen die Gruppe der U_{AN} , für welche sich mittels des pseudo-zufälligen Stichprobenansatzes 2 831 483 Tweets für das Set S_{MPL} akquirieren lassen, welche mittels des BERT-basierten Modells klassifiziert werden.

Das Training erfolgte über 5 Epochen. Die Validierungsgenauigkeit erreichte 42.5%. Die Ergebnisse des Plausibilitätschecks legen nahe, dass sich die Qualität der Klassifizierung und damit die Belastbarkeit der Ergebnisse klar zwischen Themen unterscheidet. In diesem Kapitel werden nur die als belastbar betrachteten Themen der Grade A, B, und C präsentiert. Die vollen Ergebnisse und die Statistiken des Plausibilitätschecks sind im Anhang der Dissertation dokumentiert (Anhang A7).

Abbildung 39 zeigt für jede Gruppe und jedes Thema den durchschnittlichen Anteil des Themas an Tweets der Nutzenden. Herausragend sind Unterschiede in den Themen *islam*, *religion*, und *gaming*. Unterschiede in *travel* und *music-radio* waren signifikant, aber weniger stark ausgeprägt. Es ist bemerkenswert, dass die Gruppe der U_{30} , welche mobile Nutzende repräsentiert, dem Thema *travel* geringere Anteile zuweist als es U_{90} und U_0 tun.

*Islam*⁹⁴ macht für U_{90} einen Anteil von 3,86% aus, für U_{30} nur 3,68% und für U_0 nur 3,31%, was dafür spricht, dass Mobilität einen positiven Zusammenhang mit Interesse an diesem Thema aufweist. Für *religion* ist das Bild umgekehrt. Hier sind die Anteile 2,64%, 3,44% und 3,77% für jeweils U_{90} , U_{30} und U_0 .

⁹⁴ Die als *Islam* klassifizierten Texte enthalten neben eindeutig religiösen Inhalten auch viele nicht offensichtlich religiöse Texte, welche Phrasen mit religiöser Wortherkunft oder arabische Begriffe enthalten.

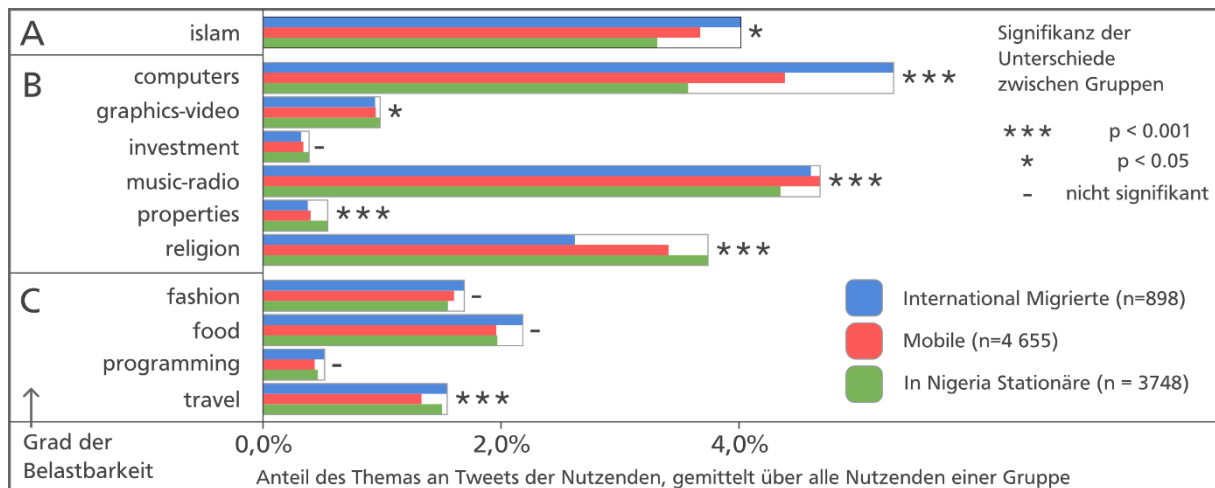


Abbildung 39: Gemittelte Themenanteile der Mobilitätstypen. Themen sind nach Belastbarkeit in Grade (a), (b) und (c) eingeteilt. Ein Test auf signifikante Unterschiede zwischen Mobilitätstypen erfolgte mittels der Kruskal-Wallis Methode. Themen mit geringerer Belastbarkeit sind in Anhang A6 dargestellt.

Als kontinuierliches Maß der Mobilität kann auch hier wieder der Gyrationradius genutzt und in Bezug zu den Anteilen der Themen visualisiert werden (Abbildung 40). Aufgrund der großen Anzahl von Accounts sind hier keine einzelnen Accounts als Punkte dargestellt, sondern geglättete Kurven⁹⁵ für eine Auswahl an Themen. Auch in dieser Perspektive zeigt sich eine negative Beziehung zwischen *religion* und Mobilität und eine positive Beziehung zwischen *gaming* und Mobilität. Auch für *travel* zeigt sich ein Bild, bei dem ab einem gewissen Mobilitätsgrad (etwa bei einem Gyrationradius von 100km) die Relevanz des Themas stetig zunimmt. Für *islam* und *properties* waren die Kurven komplexer und nicht stetig. Für *food* ist keine klare Beziehung erkennbar.

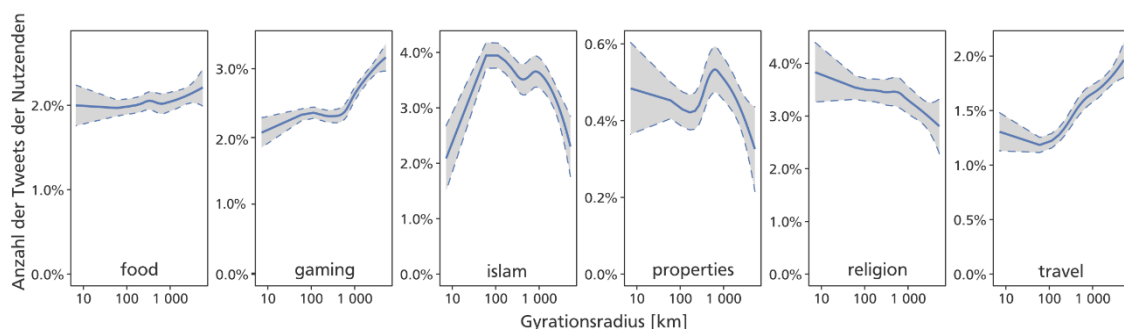


Abbildung 40: LOESS Kurve der Anteile von Themen an Tweets der Nutzenden gegen den Zehnerlogarithmus ihrer Gyrationradien. Aufgrund der großen Anzahl an Beobachtungen (9672 Nutzende) werden nur die Kurven gezeigt. Die Abszisse ist logarithmisch. Die grau schattierten Flächen zeigen das 95% Konfidenzintervall.

Da die stärksten Unterschiede zwischen Mobilitätstypen für die Themen *gaming* und *religion* gemessen werden, werden diese zusammen mit dem Thema *travel* für die qualitative Codierung herangezogen. Die Codes geben eine feinere Aufteilung der Inhalte der Themen an, und sind in Abbildung 41 dargestellt sowie in Anhang A8 beschrieben.

Das bereits angesprochene Thema *travel* enthält vielseitige Formen von Information und Diskussion mit Mobilitätsbezug. Bemerkenswert ist, dass die geographische Ebene dieser Inhalte eine große Spanne aufweist, von lokalen Verkehrsnachrichten bis hin zu

⁹⁵ locally estimated scatterplot smoothing (LOESS, Cleveland and Loader 1996)

internationalen Nachrichten und Diskussion von Visaproblemen und kritischen Kommentaren zum Zustand der Verkehrsinfrastruktur. Relativ selten wird die Absicht zu Reisen oder Migration ausgedrückt (in 2% der Fälle), wohingegen das Teilen von Reiseerfahrungen weitaus häufiger war (~13%). Etwa 25% der evaluierten Tweets hatten keinen offensichtlichen Mobilitätsbezug. *Gaming*, welches bei international Migrierten einen höheren Anteil ausmacht, enthält eine Mischung von diversen Inhalten mit Unterhaltungsbezug im weiteren Sinn, wie Sport (insbesondere Fußball), Spiel und Musik. Öffentlicher Diskurs zu Aspekten und Personen des öffentlichen und kulturellen Lebens machen auch einen großen Anteil (insgesamt etwa 45%) aus. Die Inhalte, welche in die Kategorie *gaming* eingeteilt wurden, sind also divers. *Religion* weist eine vergleichsweise höhere Homogenität auf. Nur etwa 18% der Tweets lassen keinen klaren Bezug zu Religion und Spiritualität erkennen. Jedoch kam der Diskurs zu Religion in einer Vielzahl von Formen, von Bibelversen über Gebete und Glaubensbekenntigungen bis hin zu Diskussionen über das richtige Verhalten und dem Verhältnis von Politik und Religion.

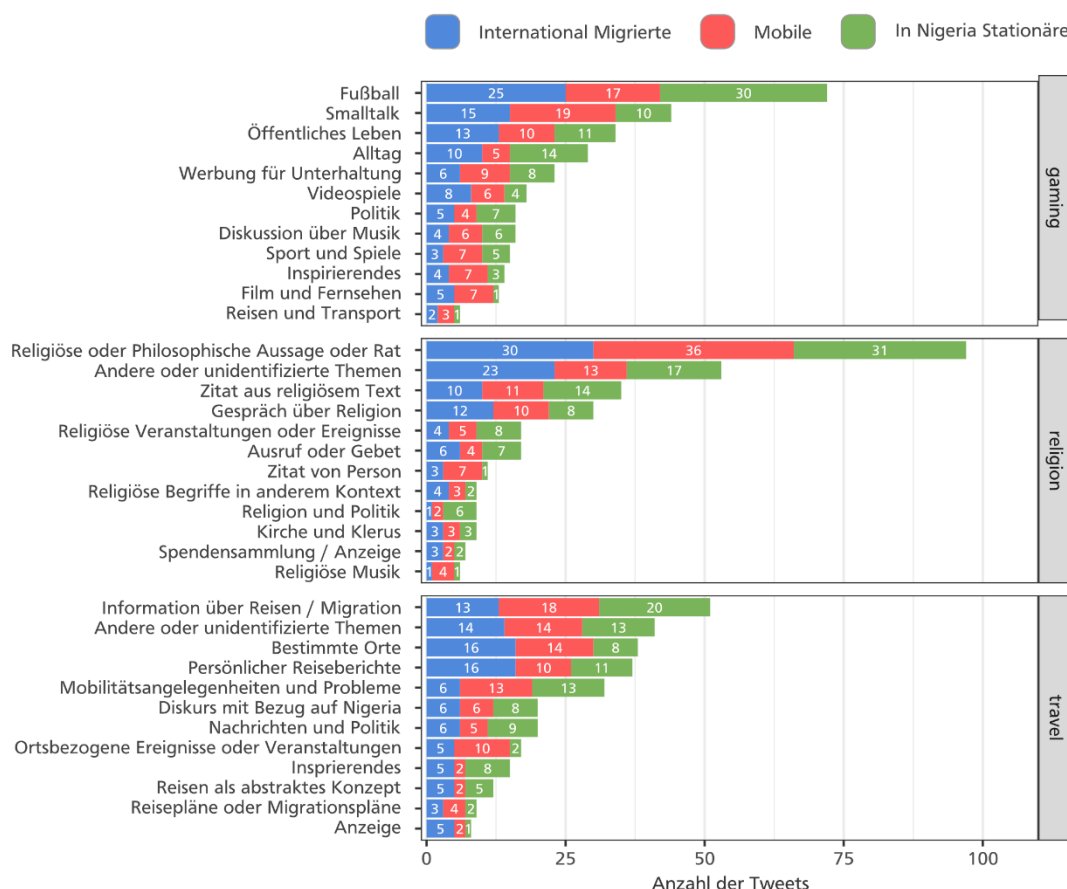


Abbildung 41: Identifizierte Codes und ihre Anzahl innerhalb der Themen *gaming*, *religion* und *travel*

6.3.3. Interpretation

Mittels des gewählten Ansatzes war es möglich, thematische Präferenzen von Twitter Nutzenden aus ihren Online-Texten abzuleiten und in Verbindung zu Charakteristika ihrer Mobilität zu setzen. Die Erweiterung des Ansatzes von Chi et al. (2020) und dessen Implementierung als R-Bibliothek war damit erprobt und konnte daraufhin veröffentlicht werden⁹⁶.

⁹⁶ <https://github.com/dlr-eoc/ukis-migrationdetectr>

Die erfassten Mobilitätsflüsse können als nützliches Zwischenprodukt der Analyse betrachtet werden. Sie weisen auf die Bedeutung von Lagos als Migrationsknotenpunkt hin, was mit Ergebnissen von Kirwin und Anderson (2018) übereinstimmt. Die Prominenz der Länder USA und UK als internationale Kontaktpunkte entspricht den Ergebnissen früherer Umfragen (NOI Polls Ltd 2015). Die gut dokumentierten Strömungen von Binnenflüchtlingen aus dem Nordosten Nigerias (IOM, 2022) sind in diesen Mobilitätsflüssen nicht reflektiert, was in Einklang mit den Ergebnissen früherer Studien zeigt, dass Geflüchtete und Binnenflüchtlinge kaum bis gar nicht auf Twitter aktiv sind (Petutschnig et al. 2020). Es lässt sich also auch schlussfolgern, dass die in dieser Studie genutzten GSM nicht geeignet sind, um Informationen über die Interessen von Flüchtlingen und Binnenvertriebenen zu sammeln. Das größte Potential, welches in dieser Studie demonstriert wird, liegt in der Erfassung von Mobilität auf regionaler bis internationaler Ebene.

Das Format von GSM und die dafür entwickelten Methoden erwiesen sich als geeignet, um hier detailliertere Informationen abzuleiten. Die Unterscheidung von verschiedenen Mobilitätstypen ist algorithmisch möglich und hauptsächlich durch Datenverfügbarkeit limitiert. Damit einher geht eine Abwägung von Repräsentativität, denn nur wenige sehr aktive Accounts haben ausreichend dichte Geotweet-Historien für die Trajektorienanalyse, und es kann nicht davon ausgegangen werden, dass diese Accounts repräsentativ für Twitter-Nutzende oder Migrierte im Allgemeinen sind. Hingegen kann die Unterscheidung zwischen mobilen und stationären Nutzenden für eine deutlich größere Anzahl Nutzenden mit höherer Sicherheit gemacht werden.

Mit 42,5% ist die Genauigkeit des Klassifizierers in der Quelldomäne relativ gering. Dies ergibt sich zumindest teilweise aus der Informalität und Kürze der Texte und der großen Anzahl möglicher Themen (42 Themen), welche ein deutlich schwierigeres Problem darstellt als die Klassifizierung weniger Klassen (Quercia et al. 2012). Aufgrund von inhaltlichen Themenwechseln im Laufe von Konversationen, (Content-drift, siehe Hobbs 1990) und nicht-erkanntem Spam enthalten zudem viele Themen in Nairaland auch Texte, welche inhaltlich keinen Zusammenhang mit dem Unterforum haben, was zu Unschärfen in den Trainings- und Validierungsdaten führt. Angesichts dieser Herausforderungen ist die Genauigkeit der in Studie B1 genutzten Klassifizierers nachvollziehbar. Sie ist vergleichbar mit den 42,1% Genauigkeit des Gewinners der TREC-IS-2018 Challenge, bei welcher 25 Klassen von Inhalten im Kontext von Tweets über Naturkatastrophen unterschieden werden mussten (McCreadie et al. 2019). Es ist daher nicht davon auszugehen, dass aktuelle Sprachmodelle oder der gewählte Ansatz grundsätzlich ungeeignet sind. Darüber hinaus ist eine genauere Evaluierung der Ungenauigkeiten und ihrer Auswirkungen im Rahmen des gewählten Ansatzes nicht möglich, die Evaluierung von Domänenadaptionsansätzen in diesem Anwendungsbereich könnte jedoch ein sinnvolles Ziel weiterer Forschung sein. Angewandte Studien sollten demgegenüber in Betracht ziehen, zur Verbesserung der algorithmischen Genauigkeit eine Reduktion der möglichen Labels vorzunehmen und die Datenqualität zu verbessern⁹⁷.

Trotz der Herausforderungen erweist sich der Ansatz als grundsätzlich erfolgreich, denn in den klassifizierten Themen zeigen sich messbare Unterschiede zwischen den Mobilitätstypen, die Aufschluss auf mögliche Zusammenhänge zwischen Mobilität und Interessen von

⁹⁷ In einem Experiment zeigte sich, dass bereits durch manuelles Gruppieren einiger Klassen und dem Ausschluss besonders kurzer Texte eine Erhöhung der Genauigkeit von 42,5% auf 60% erreicht werden konnte (Siehe Anhang A9).

Nutzenden geben können. Es steht zu vermuten, dass diese bei höherer Klassifikationsgenauigkeit noch stärker ausfallen würden.

Dabei sind *gaming* und *islam* als Themen identifiziert, welche für International Migrierte überproportional häufig sind. Demgegenüber sind die geringen Unterschiede zwischen Mobilitätstypen im Thema *travel* bemerkenswert. Die qualitative Analyse einer Stichprobe dieses Themas zeichnet ein komplexes Bild, bei welchem Mobilität als Thema nicht in direktem Zusammenhang mit Migrationsaktivität steht. Offensichtlich haben auch stationäre Menschen ein bemerkenswertes Interesse an Mobilitäts- und Verkehrsthemen, beispielsweise der lokalen Transportinfrastruktur und Staus, welche in vielen nigerianischen Städten als Problem gesehen werden (Afolabi et al. 2017) und negative Auswirkungen auf Menschen und Wirtschaft haben (Economic Intelligence Unit 2013). Angesichts dessen ist *travel* ein Thema, dessen Relevanz für Menschen verschiedener Mobilitätsgraden ähnlich hoch ist, welches aber auf verschiedenen geographischen Ebenen und Größenordnungen in unterschiedlicher Form auftritt. Die Stichprobe der codierten Tweets ist für statistische Tests nicht ausreichend, zeigt aber, dass die Vielfalt von Mobilitätsformen, auf welche bereits in der Einleitung dieser Dissertation Bezug genommen wurde, auch in den Texten reflektiert ist.

Selbst vermeintlich klar definierte Themen erwiesen sich als inhaltlich komplex. Beispielsweise zeigten die Ergebnisse *islam* als ein Thema von besonderer Bedeutung für International Migrierte, was zunächst im Widerspruch zu Ergebnissen von Bloch et al. (2015) steht. Ein genauere Analyse der Texte zeigt jedoch, dass die als „islam“ annotierten Tweets sich oft nicht primär mit der Religion des Islam beschäftigten, sondern arabische Aussprüche, Äußerungen und Sprichworte mit teilweise religiösem Ursprung enthalten. Möglicherweise durch geringere Leistungsfähigkeit des Sprachmodells bei Arabischer Sprache, oder durch Unterschiede zwischen der Quelldomäne Nairaland und der Zieldomäne Twitter bedingt, entspricht das algorithmisch zugeordnete Label daher nicht notwendigerweise einem intuitiven Verständnis desselben. Dementsprechend kann das Label eine nur begrenzt interpretierbare Repräsentation von möglicherweise komplexen Inhalten darstellen. Auch die LOESS Regression nach Gyrationradius zeigte im Falle von *islam* ein komplexeres Bild, als es die Kategorisierung in Migrationstypen suggerierte. Die Beziehung zwischen Mobilität und Themenanteil war dabei nicht stetig steigend, sondern fiel bei hoher Mobilität wieder ab.

Auch das Thema *gaming* deckt ein weites Spektrum ab, welches über Spiele (engl. *Games*) hinausgeht, und könnte nach Einschätzung des Autors aufgrund der Vielzahl von Teilbereichen als *Allgemeine Unterhaltung* bezeichnet werden. Oder das Thema *religion*, welches sich zwar inhaltlich als vergleichsweise homogen in Bezug auf christliche Religion darstellt, dessen Inhalte sich aber in einer großen Vielfalt von Konversationstypen äußern (beispielsweise Zitate, Debatten, Wünsche, Gebete, Gedanken und Ausrufe).

Diese Betrachtung zeigt auf, dass Themen und Mobilitätsklassen Generalisierungen einer komplexen Realität darstellen, und damit als Ausgangspunkte für tiefere Analysen genutzt werden können. Angesichts der multidimensionalen Komplexitäten ist die Aggregation von Inhalten ein sinnvoller Schritt, um grundlegende Muster zu identifizieren. Die Komplexitäten sollten jedoch nicht verworfen, sondern berücksichtigt und herausgearbeitet werden, um eine korrekte Interpretation der Ergebnisse zu ermöglichen. Auch hierbei können die Mobilitätsinformationen eine Rolle spielen, um die Textinhalte zu kontextualisieren. Das Potenzial der Verbindung von Räumlichen und Textinhaltlichen Informationen ist damit sowohl auf aggregierter als auch auf disaggregierter Ebene zu sehen.

Insgesamt demonstriert die Studie das Potenzial, große GSM-Datensätze sowohl textinhaltlich als auch räumlich zu prozessieren, um die Textinhalte der GSM-Daten mit Kontext zum Ort und Mobilität der Nutzenden zu bereichern. Herausforderungen der Datenqualität und der fehlenden Annotationen können durch technische Lösungen adressiert werden, was letztendlich zu einem komplexen und aufwändigen Ansatz führt, welcher aber neuartige Analysen in einem herausfordernden Anwendungsgebiet ermöglicht.

Die grundsätzliche Limitation des Ansatzes besteht in der Verfügbarkeit und Repräsentativität der genutzten Daten. Die ermöglichte Reduzierung der Komplexität auf Kategorien (wie beispielsweise Themen) und Mobilitätstypen ist hierbei gleichsam als Chance und als Gefahr zu bezeichnen. Denn menschliche Migrationsentscheidungen und deren Effekte auf die Individuen dürfen nicht übersimplifiziert werden, und bedarf weiterer Untersuchungen. Die Ergebnisse der Studie belegen daher einerseits die Macht der Algorithmen und unterstreichen gleichzeitig die Notwendigkeit, verschiedene Methoden in der Migrationsforschung zu kombinieren. In dieser Hinsicht liefert der Ansatz der Studie durch eine Möglichkeit, die Perspektiven von Migrierten empirisch messbar zu machen, einen Beitrag zur nötigen methodischen Vielfalt.

6.4. Studie B2: Analyse der Verbreitung von Sprachmerkmalen nach Mobilitätscharakteristika der Autoren am Beispiel westafrikanischer Twitter Nutzender

Der initiale Datensatz T_{ROI} enthält 33 358 088 Geotweets von 416 712 Accounts innerhalb der ROI im Untersuchungszeitraum. Für 146 312 dieser Accounts ergibt sich nach Vervollständigung ihrer Geotweet-Historie ein Satz von insgesamt 33 538 087 Geotweets, welche in einer Trajektorienanalyse verarbeitet werden. Für 12 692 der Accounts kann hierdurch eine Einteilung in mobil oder stationär vorgenommen werden (Tabelle 7). Die für sie abgefragten 3 681 197 Tweets werden in der Textanalyse untersucht.

Tabelle 7: Überblick über der für Studie B2 akquirierten Daten.

Datensatz	Anzahl Tweets	Anzahl Accounts	Georeferenz	Zweck
T_{ROI}	33 358 088	416 712	Vollständig	Identifikation von Kandidatenaccounts U_C
T_{GEO}	33 538 087	146 312	Vollständig	Trajektorienanalyse (von Accounts U_C)
T_{AN}	3 681 197	12 692	Teilweise	Textanalyse (von Nutzenden U_{AN})

6.4.1. Verbreitung von LEN Allgemein

Das kombinierte Wörterbuch W_L enthält 521 Begriffe, wovon 307 als bivalent identifiziert und 34 Begriffe manuell hinzugefügt wurden.

Der Abgleich von W_L mit den Tweets in T_{AN} identifiziert in 1,7% der Tweets LEN-Merkmale. Innerhalb dieser Tweets ist der häufigste Begriff aus dem Wörterbuch *abeg*⁹⁸ welches in 12,8% dieser Tweets auftaucht, kurz gefolgt von dem bivalenten Begriff *for*⁹⁹ mit 12,7%. Die

⁹⁸ „Bitte“

⁹⁹ Ein flexibles, polysemes Wort mit einer Vielfalt von lokativen Funktionen, beispielsweise „von“ und „bei“ (Corum 2015).

häufigsten aus mehreren Worten bestehenden Begriffe sind *too much*¹⁰⁰ und *na wetin*¹⁰¹ mit Häufigkeiten von 0,6% und 0,3%. Der mittlere LI lag bei 0,003 (0.3%), was bedeutet, dass insgesamt im Durchschnitt eines von 350 getweeteten Wörtern einem Begriff des Wörterbuches entspricht.

6.4.2. Verbreitung nach Ort

Von den 33 538 087 Geotweets in T_{GEO} liegen 19 453 465 innerhalb der ROI und enthalten mindestens ein Wort¹⁰². Mit 17 932 448 Geotweets liegt die weite Mehrheit (ca. 92%) in einer der MUAs und nur 1 521 017 (ca. 8%) in Gebieten, welche in dieser Studie als nicht-metropolitisch betrachtet werden. Dabei zeigt sich kein klarer Megacity-Effekt, welcher die Nutzung von lokalen Merkmalen in sehr großen Städten begünstigt. Die weitaus bevölkerungsreichste MUA Lagos weist zwar einen erhöhten LI im Vergleich zum nicht-metropolitanen Nigeria auf, liegt aber unter den nigerianischen MUAs nur auf Rang 8 hinter kleineren MUA wie Ondo, Sokoto oder Minna.

Es zeigt sich klar (Abbildung 42), dass nigerianische MUAs weitaus höhere Werte des LI aufweisen als die anderen Länder. Die einzigen Ausnahmen bilden die MUAs Okene und Okpogho, für welche vergleichsweise wenige Geotweets verfügbar waren, und deren LI unter dem von Ghana, Togo, und Benin liegt. Die MUA mit dem insgesamt höchsten LI (0.7%) liegt im Norden Nigerias, doch innerhalb des Landes ist kein Unterschied zwischen Norden und Süden Nigerias zu erkennen.

Auch unter den über einen hohen TF-IDF Wert definierten typischen Begriffen zeigt sich keine klare Nord-Süd Dualität: Die aus dem Hausa¹⁰³ stammenden Begriffe *mallam*¹⁰⁴ und *kudi*¹⁰⁵ stellen sich als typisch für nordnigerianische MUAs dar, doch gilt dies auch für die Wörter *oga*¹⁰⁶ und *jare*¹⁰⁷, welche aus dem Yoruba stammen (Blench und Dendo 2005; Unuabonah und Oladipupo 2021). Unter den typischen Begriffen zeigen sich mit bemerkenswerter Häufigkeit kulinarische Begriffe: *banga*, ein Begriff für eine Suppe, ist disproportional häufig in Warri im südlichen Nigerdelta vertreten (65-mal in 79 047 Geotweets) während der Begriff *fufu* für eine Art von stärkehaltigem Brei typisch für ghanaische MUAs ist. Auch typisch für ghanaische MUAs ist der Begriff *yawa*, welcher sich auf ein Problem oder eine unangenehme Situation bezieht und in MUAs anderer Länder nicht besonders genutzt wurde.

¹⁰⁰ „Sehr“, oft als Bekräftigung als Teil von Dankesbekundungen ausgedrückt.

¹⁰¹ „Was ist los?“ „Was ist es dieses Mal?“

¹⁰² Hierbei nicht berücksichtigt sind Emojis, URLs, Zahlen und Benutzernamen.

¹⁰³ Afroasiatische Sprache, welche insbesondere im Norden Nigerias und Süden Nigers verbreitet ist, und vornehmlich vom Volk der Hausa gesprochen wird

¹⁰⁴ Eine den Hausa angehörige oder muslimische Person, oder eine gebildete Autorität im Bereich der Wissenschaft, Bildung oder (insbesondere islamischer) Religion

¹⁰⁵ Geld

¹⁰⁶ Eine Person mit Autorität

¹⁰⁷ Betonungsmarker

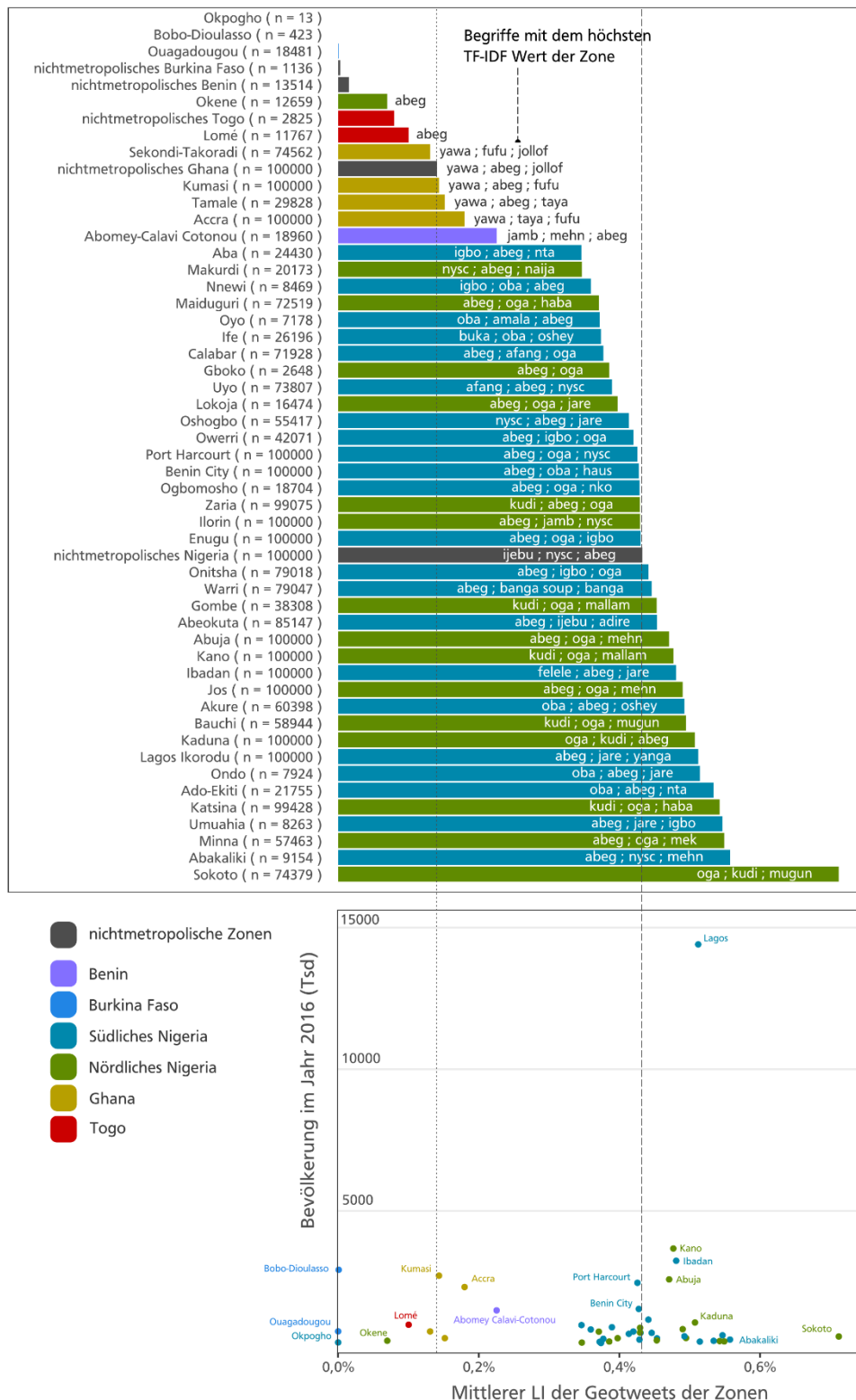


Abbildung 42: Verbreitung von LEN in MUAs und nichtmetropolischen Zonen (außerhalb MUAs) im Untersuchungsgebiet. Oben: Durchschnittlicher LI der Geotweets innerhalb der MUAs und nichtmetropolischen Zonen in den fünf Ländern. Neben den Balken sind die nach der TF-IDF Metrik spezifischsten Begriffe der MUA dargestellt. Unten: Durchschnittlicher LI der Geotweets der MUA gegen deren Bevölkerung im Jahr 2016. Die gepunkteten und gestrichelten Linien zeigen jeweils den durchschnittlichen LI der nichtmetropolischen Gebiete in Ghana und Nigeria an.

6.4.3. Verbreitung nach Mobilitätscharakteristika

Die oben beschriebenen Beobachtungen auf der Ebene des Orts (der MUAs) zeigen also bereits einige Unterschiede in der Nutzung von LEN-Merkmalen zwischen Ländern auf. Dies wird durch die Differenzierung auf der Ebene von Aufenthaltsländern der Nutzenden genauer in den Fokus genommen (Abbildung 43). Anders als die oben genannten Unterschiede auf der MUA-Ebene beruhen diese Statistiken auch auf Tweets ohne Geotags.

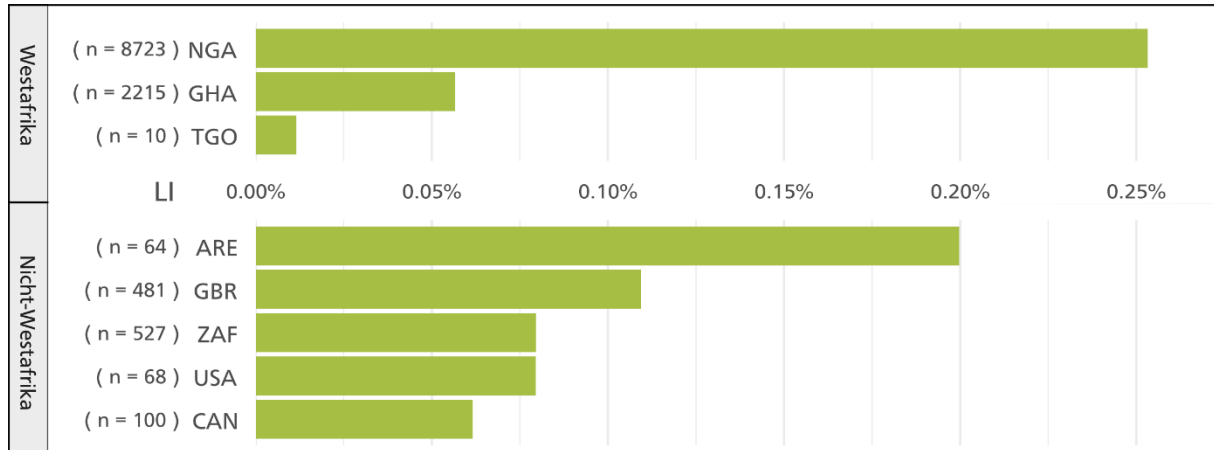


Abbildung 43: Durchschnittlicher LI pro Land (der höchsten Aufenthaltsdauer der Nutzenden). In Klammern die Anzahl (n) der Nutzenden. ARE: Vereinigte Arabische Emirate. GBR: Vereinigtes Königreich Großbritannien und Nordirland. ZAF: Südafrika. USA: Vereinigte Staaten von Amerika. CAN: Kanada. NGA: Nigeria. GHA: Ghana. TGO: Togo.

Aufgrund der für die Datenakquise genutzten Methode ist es sinnvoll, Länder innerhalb Westafrikas und Länder außerhalb Westafrikas getrennt zu interpretieren. Über die betrachteten westafrikanischen Länder hinweg ist zunächst auffällig, dass innerhalb Nigerias eine weitaus höhere Nutzung von LEN-Merkmalen dokumentiert wird als in Ghana und besonders als im französisch geprägten Togo. Für Benin und Burkina Faso sind unzureichende Anzahl an Accounts für die Aggregation vorhanden.

Von den 12 692 Accounts in T_{AN} können 8 008 einem der detaillierten Mobilitätstypen zugeordnet werden (Tabelle 8).

Tabelle 8: Beschreibung und Anzahl der in Studie B2 analysierten detaillierten Mobilitätstypen

Mobilitätstyp	Kriterium	Anzahl der Accounts
NGA Stationäre	Alle Segmente am selben Aufenthaltsort in Nigeria	3 122
NGA Mobile	Mehrere Segmente in unterschiedlichen Aufenthaltsorten in Nigeria	3 984
ROI Mobile	Mindestens ein Segment in Nigeria und mindestens ein weiteres in einem anderen Land des ROIs	226
GBR Emigrierte	Mindestens ein Segment in Nigeria, doch der Großteil der Zeit im Vereinigten Königreich	333
NA Emigrierte	Mindestens ein Segment in Nigeria, doch der Großteil der Zeit in Nordamerika (Kanada und USA)	343
Summe		8 008

Außerhalb Westafrikas ist der LI insbesondere in den Vereinigten Arabischen Emiraten ausgeprägt und weitaus höher als selbst in Ghana. Im Vereinigten Königreich war der LI höher als in den USA. Abgesehen davon sind keine Unterschiede zwischen nicht-westafrikanischen Ländern statistisch signifikant.

Abbildung 44 zeigt den LI für diese Gruppen. Innerhalb Nigerias zeigt sich ein schwacher Mobilitätseffekt: Für die Gruppe der mobilen Nigerianer wird ein höherer LI gemessen als für stationäre Nigerianer. Nutzende innerhalb der westafrikanischen ROI weisen eine signifikant höhere Nutzung der lexikalischen Sprachmerkmale auf als Emigrierte.

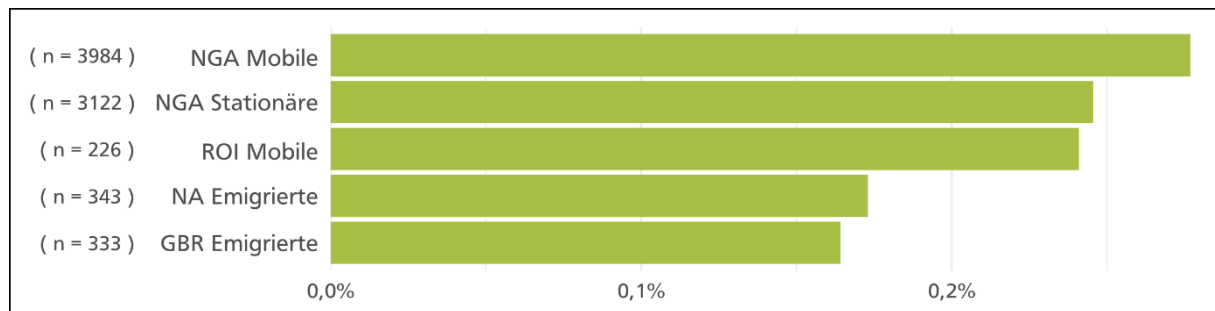


Abbildung 44: Median des LI von Tweets nach Mobilitätstyp

Neben diesem kategorischen Bild erlaubt die lineare Regression mittels des Gyrationradius (R_{GYR}) ein kontinuierliches Bild der Beziehung von LI und Mobilität auf Ebene von Nutzenden. Das Ergebnis ist signifikant ($p\text{-Wert} \approx 0$) und negativ, was als Indiz dafür gesehen werden kann, dass die Nutzung von LEN-Sprachmerkmalen mit zunehmender Mobilität abnimmt. Dabei ist der Regressionskoeffizient a sehr gering ($a = -1,804E-09$ für alle Accounts, und $a = -1,698E-10$ für Accounts, welche mindestens einmal einen LEN-Begriff nutzen), genau wie die Bestimmtheitsmaße ($R^2 = 0,019$ beziehungsweise $R^2 = 0,004$).

6.4.4. Verbreitung nach Themen

Von insgesamt 224 Themen, welche von dem in 5.4.3 spezifizierten BERTopic Modell ausgegeben werden, wurden innerhalb des Twitter-Datensatzes T_{AN} folgende fünf *Supertopics* identifiziert: *Musik*, *Wirtschaft*, *Innenpolitik*, *Globale Nachrichten*, und *Sport*.

Abbildung 45a zeigt die Themen eingebettet in einem zweidimensionalen Merkmalsraum¹⁰⁸ und farblich den Supertopics zugeordnet. Dabei stellt Abbildung 45b die thematischen Präferenzen der fünf Mobilitätstypen dar (wie in Studie B1, vgl. 6.3.2). Hier ist beispielsweise zu sehen, dass *Sport* bei mobilen Nigerianern einen höheren Anteil der Tweets ausmacht, als bei stationären Nutzenden aus Nigeria, was im Einklang mit Ergebnissen aus Studie B1 steht (vgl. 6.3.3).

Abbildung 45c illustriert den LI der Supertopics. Es zeigt sich, dass *Musik* das Supertopic mit dem durchschnittlich höchsten LI ist. Bemerkenswert dabei ist, dass, wie Abbildung 45b zeigt, Musik einen höheren Anteil der Tweets von Emigrierten ausmacht – einer Gruppe, die allgemein einen geringeren LI aufweist (vgl. Abbildung 44). Das legt nahe, dass zumindest in diesem Fall der negative Effekt der Mobilität durch einen positiven Effekt des Themas ausgeglichen wird. Möglicherweise ist Musik als ein kulturelles Artefakt offener für lokale sprachliche Merkmale, oder diese werden genutzt um in Werbung und Bewerbung von Musik

¹⁰⁸ Welcher aus einer weiteren Dimensionsreduktion mittels UMAP der für das Topic Modeling genutzten dimensionsreduzierten Embeddings hervorgeht.

und Künstlern, welche einen nennenswerten Anteil der Musik-Tweets ausmacht, eine breite Zielgruppe anzusprechen.

Auch für *Wirtschaft* wird ein relativ hoher LI gemessen. Die Betrachtung einer Auswahl von zugehörigen Tweets zeigt, dass die dabei verwendeten lokalen Begriffe hauptsächlich genutzt werden, um wirtschaftliche Probleme anzusprechen und Kritik am politischen Management der Wirtschaft auszudrücken und zu verstärken.

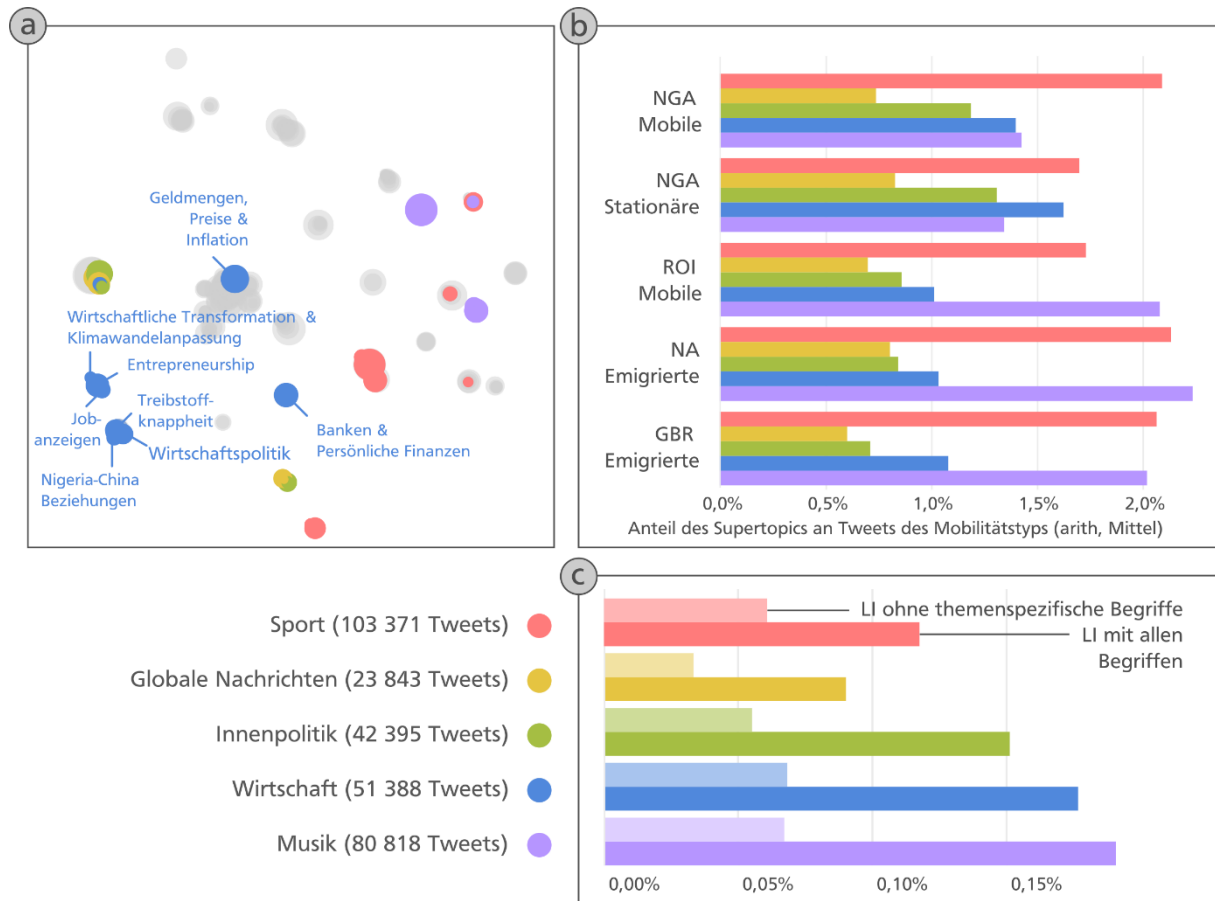


Abbildung 45: Ergebnisse des Topic Modeling in Verbindung mit LEN. (a) Themen in einem dimensionsreduzierten Embedding-Raum. Supertopics sind farblich hervorgehoben. (b): Supertopic-Neigung von Nutzenden nach Mobilitätstyp, Median über Nutzende. (c) Mittlerer LI der Supertopics mit und ohne themenspezifische lokale Begriffe.

Zuletzt wird die Übereinstimmung zwischen den algorithmisch erfassten lexikalischen Merkmalen (LEN-Begriffen) und den manuell kodierten morphosyntaktischen und pragmatischen Merkmalen verglichen. Wie in Abbildung 46 dargestellt, wird für die ausgewählten Gruppen kein Zusammenhang zwischen diesen beiden Merkmalstypen ersichtlich. Anders ausgedrückt würde die Analyse von LEN auf Basis anderer Sprachmerkmale vermutlich zu anderen Ergebnissen führen.

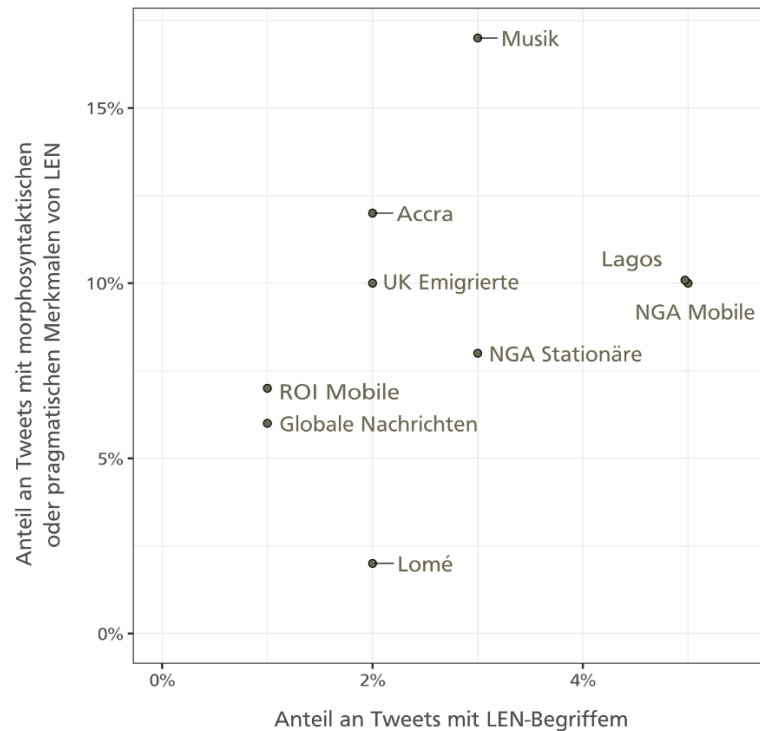


Abbildung 46: Anteil der Tweets mit LEN-Begriffen verglichen mit dem Anteil an Tweets mit lokalen morphosyntaktischen oder pragmatischen Sprachmerkmalen

6.4.5. Interpretation

Insgesamt bestätigt sich, dass Geotweets primär in städtischen Regionen und insbesondere in großen Städten generiert werden. Für Untersuchungen von Sprache kann der demonstrierte Ansatz also insbesondere in diesen Regionen umgesetzt werden, und die Ergebnisse können vornehmlich als repräsentativ für städtische Twitter-Nutzende gesehen werden.

In diesem Rahmen zeigt sich insgesamt ein Bild, bei welchem internationale Mobilität mit längeren Aufenthaltsdauern einen messbar negativen Effekt auf die Nutzung von lokalen lexikalischen Sprachmerkmalen hat. Effekte von regionaler Mobilität sind gering bis nicht messbar. Auch geographische Unterschiede sind nur schwach innerhalb von Ländern ausgeprägt (Nordnigeria vs. Südnigeria) und deutlich stärker auf Länderebene (z.B. zwischen Ghana und Nigeria). Eine mögliche Erklärung ist, dass ausgewanderte Nigerianer mehr mit nicht-Nigerianern interagieren als mit Nigerianern in Nigeria. Doch darf die Sprachwahl von Migrierten auch nicht zu stark vereinfacht werden. Sowohl die Identitätskonstruktion einer migrierten Person und Narrative der Zugehörigkeit oder Abgrenzung zu Heimat- oder Zielland kann die Nutzung von Sprachmerkmalen wie lokalen lexikalischen Merkmalen beeinflussen (Honkanen 2020). Sprachentscheidungen werden von der Selbstwahrnehmung des Verfassers beeinflusst, den konzipierten Rezipienten der Nachricht und der Kommunikationsabsicht. In dieser Vielschichtigkeit kann Potenzial für weitere Forschung gesehen werden. Nicht nur aus sprachwissenschaftlicher, sondern auch aus geographischer Sicht wäre es durchaus wünschenswert, über eine nähere Untersuchung der hier empirisch gemessenen Sprachentscheidungen ein Verständnis des sozialen Umfelds von Migrierten und ihrer Bindung an Ursprungs- und Zielland (siehe Kim, Sirbu, Giannotti, et al. 2022) zu gewinnen. Da jedoch alle oben genannten Faktoren eine Rolle spielen (und wahrscheinlich noch viele weitere) und dynamisch und intersektional interagieren (Levon 2015) ist es

herausfordernd, die empirischen Ergebnisse auf den Einfluss bestimmter Faktoren zurückzuführen. Hierfür sind Informationen über soziale und persönliche Hintergründe der Verfasser notwendig. In Social-Media-Daten ist solch kontextuelle Information zwar oft in verschiedenen Formen vorhanden, aber nicht standardisiert und damit schwierig algorithmisch zu identifizieren. Qualitative Methoden bleiben also weiterhin sinnvoll und notwendig, welche durchaus durch algorithmische Statistiken unterstützt werden können. Ein möglicher Ansatz könnte darin bestehen, Texte zu identifizieren, bei welchen Migrierte mit Nutzenden aus dem Herkunftsland kommunizieren, und darin qualitativ zu analysieren, ob und in welcher Form dies die Nutzung lokaler lexikalischer Merkmale beeinflusst. In dieser Studie wurden bereits verschiedene Statistiken zu Ort, Nutzermobilität und Themen mit qualitativer Analyse verbunden, um plausible, wenngleich vorläufige Erklärungen zu generieren. Dies kann als Indiz dafür gesehen werden, dass ein solcher Mixed-Method Ansatz grundsätzlich sinnvoll ist.

Auch, wenn hier nur die wichtigsten statistischen Ergebnisse vorgestellt werden, demonstrieren sie insgesamt, dass verschiedene Auswertemethoden nicht nur in Isolation vorgenommen, sondern auch zusammengeführt werden können. Für die Integration von Methoden der Geographie und der Sprachwissenschaft ist dies vielversprechend.

Die Quantifizierung lokaler sprachlicher Merkmale mittels Webressourcen und algorithmischer Detektion ist daher innerhalb des hier erörterten Rahmens als erfolgreich zu bewerten. Der dabei genutzte lexikalische Ansatz erwies sich als flexibel, aber limitiert. Da das Wörterbuch kein vollständiges Inventar aller lokalen lexikalischen Merkmale darstellt, ist der geringe Anteil von detektierten lokalen Merkmalen in Tweets nicht überraschend –die tatsächliche Verbreitung von LEN ist vermutlich deutlich höher.

Es ist zudem nicht eindeutig, was die Quellen des Wörterbuches sind oder welche spezifischen regionalen Varietäten des Englischen dort stärker abgebildet sind. Naijalingo beschreibt sich als ein offenes Wörterbuch, welches Nutzende frei editieren können und gibt zu bedenken, dass Wörter im Laufe der Zeit eine Bedeutungsänderung erfahren und neue Wörter entstehen können (Naija Lingo 2023). Babawilly's Dictionary (2023) gibt an, hauptsächlich Begriffe des Lagos-Pidgin zu enthalten, welches aus überproportional vielen Yoruba Wörtern besteht, und verweist zusätzlich darauf, dass Igbo, Yoruba und Hausa die drei indigenen Sprachen mit dem höchsten Einfluss auf Nigerianisches Pidgin seien, was impliziert, dass andere indigene Sprachen weniger repräsentiert sind.

In einer manuellen Auslesung von 900 Tweets zeigte sich, dass das Auftreten morphosyntaktischer und pragmatischer Merkmale nicht durch die Messung lexikalischer Merkmale approximiert wird, und diese daher möglicherweise in anderer Beziehung zu Mobilität und Lokalität stehen. Diese anderen Merkmale sind nicht so leicht algorithmisch erfassbar wie die lexikalischen Merkmale. Gewisse pragmatische Merkmale, wie das Anhängen des Partikels -o¹⁰⁹ könnten jedoch mit erweiterten Wörterbüchern algorithmisch erkannt werden. Zudem könnten LLMs genutzt werden, um eine größere Vielfalt von Sprachmerkmalen zu erkennen (Park et al. 2021). Allerdings sind große Sprachmodelle, welche auf neuronalen Netzen basieren, in der Regel weniger transparent (Castelvecchi 2016). Im Kontrast dazu ist der hier demonstrierte Ansatz transparent und nachvollziehbar. Er ist

¹⁰⁹ Das Anhängen des Satzabschlusspartikels „o“ ist weit verbreitetes Feature des Nigerianischen Pidgin English und erfüllt eine Vielzahl von möglichen Funktionen, siehe Manfredi und Čeplo (2023).

dafür konzipiert, durch eine Mischung von qualitativen Methoden, einfachen algorithmischen Methoden und Sprachmodellen flexibel und nachvollziehbar zu sein, und somit für verschiedenste Forschungsfragen im Migrationskontext einen Beitrag liefern zu können.

Im Rahmen der Dissertation demonstriert die Studie, dass die Analyse sprachlicher Merkmale wie lokaler (und weniger dokumentierter) lexikalischer Merkmale in GSM-Daten möglich ist. Aufgrund der gegenseitig hohen Bedeutung von Sprache und geographischer Mobilität ist hier ein hohes Potenzial für weitere Forschung zu erkennen. Da die algorithmischen Methoden zur Auswertung von Sprache noch nicht alle Aspekte von Gebrauch und Variation gut abdecken, bleiben traditionelle Methoden von Relevanz. Dass Aspekte von Ort (*Wo?*), Zeit (*Wann?*), Mobilitätstyp (*Wer?*), Themen (*Worüber?*) und Sprache (*Wie?*) quantifiziert und in unmittelbaren Zusammenhang gebracht werden können, ist eine große Stärke von GSM-Daten. Zu wissen, wer wann wo wie worüber schreibt, kann Antworten auf zahlreiche Forschungsfragen liefern - insbesondere dann, wenn über die Georeferenz auch weitere Geoinformationen wie FED eingebunden werden können, was beispielsweise in Studie A2 demonstriert wurde.

6.5. Studie C1: Analyse der Räumlichkeit von Themen in englischen Webtextdaten

Um herauszuarbeiten, ob die Nutzung räumlicher Textdaten die Verteilung von Themen in den Daten beeinflusst, werden themenspezifische Geospatiality-Effekte in Studie C1 analysiert.

6.5.1. Korpusstatistiken

Hierdurch wurden die Einträge des Korpus in eine Taxonomie von folgenden 18 Themen strukturiert: Anzeigen; Architektur, Bau & Immobilien; Prominente, Unterhaltung & Musik; Verbrechen; Wirtschaft, Geschäftliches & Finanzen; Ereignisse & Veranstaltungen; Ernährung & Landwirtschaft; Gesundheit; Geschichte & Kultur; Internationale Politik; Naturkatastrophen und Naturgefahren; Politik & Regierung; Privatleben, Familie und Beziehungen; Religion; Wissenschaft, Bildung & Mathematik; Sport & Spiele; Technologie; Reisen, Tourismus & Migration. Eine Beschreibung dieser Themen kann in Anhang A10 eingesehen werden. Tabelle 9 zeigt einen Überblick über den für die Analyse generierten Korpus.

Tabelle 9: Überblick über den für Studie C1 und C2 generierten Korpus, welcher aus sieben verschiedenen Datensätzen besteht.

Dataset	Textsorte	Quelle der Themen	Quelle der Georeferenz	Autoren	Zeiteinheit
WEBIS	Forumkommentar	Subreddit	Ensemble-Geoparser	Accountname	- Kein -
Reddit	(N = 522 353)			(N = 113 512)	
Nairaland	Forumkommentar	Unterforum	Ensemble-Geoparser	Accountname	Monat
	(N = 2 153 680)			(N = 99 045)	(N = 80)
Twitter	Geotweet	Textklassifizierung	Ensemble-Geoparser	Account ID	Monat
(mit Geotags)	(N = 314 620)			(N = 3 577)	(N = 35)
Twitter	Tweet	Textklassifizierung	Ensemble-Geoparser	Account ID	Monat
(ohne Geotags)	(N = 70 745)			(N = 3 577)	(N = 35)
Stackexchange	Forumkommentar	Unterseite	Ensemble-Geoparser	Account ID	Monat
	(N = 5 671 182)			(N = 147 895)	(N = 188)
GDELT-GKG	Forumkommentar	GDELT-tagger	GDELT-Geoparser	Internetadresse	Monat
	(N = 26 018 682)			(N = 10 000)	(N = 35)
GDELT-IA-Americana	Textabschnitt	GDELT-tagger	GDELT-Geoparser	Buchautor	Jahr
	(N = 4 146 089)			(N = 2 726)	(N = 10)

Die mittlere Länge der Tweets ist höher für nach 2018 gepostete Tweets (ungefähr 133 Zeichen) als für vor 2018 gepostete Tweets (ungefähr 105 Zeichen), was angesichts des in 2018 erhöhten Zeichenlimits für Tweets nachvollziehbar ist. Die Anzahl von Geotweets ist höher als die der nicht getaggtten Tweets, was darauf hindeutet, dass die Accounts, für welche Tweets abgefragt wurden, eine Neigung zur Nutzung der Geotagging-Funktion aufweisen.

Um die Beziehung zwischen Themen und Georeferenzen herauszuarbeiten, wird zunächst die Verteilung von $Frac_{geo}$ der Themen innerhalb jedes Datensatzes dargestellt (Abbildung 47). Bereits auf den ersten Blick zeigt sich, dass innerhalb aller Datensätze starke Unterschiede zwischen Themen bestehen, was als erstes Indiz für thematische Geospatiality gewertet werden kann. Diese Unterschiede sind zwischen Datensätzen verschieden stark ausgeprägt. Während $Frac_{geo}$ innerhalb von Stackexchange und Reddit stark variiert, sind Themen auf GDELT-GKG in dieser Hinsicht verhältnismäßig konsistent hoch.

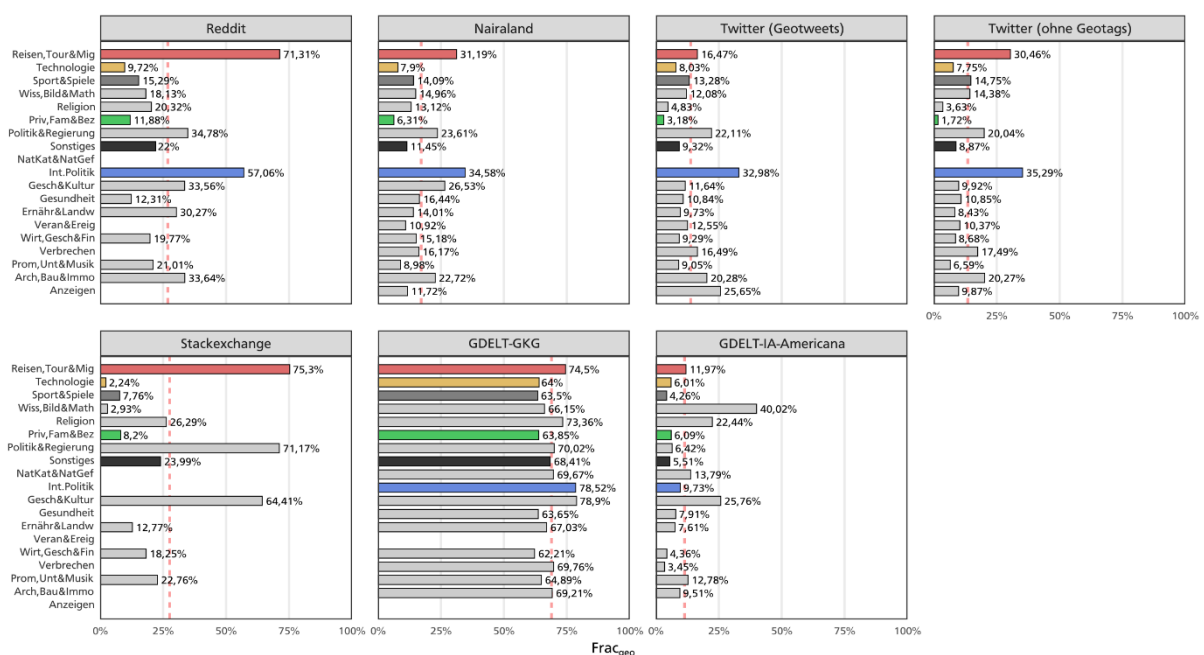


Abbildung 47: $Frac_{geo}$ der Themen, strukturiert nach Datensatz. Deutlich sichtbar sind die Unterschiede in der Häufigkeit von Ortserwähnungen erstens zwischen Plattformen und zweitens zwischen Themen innerhalb der Plattformen. Die gepunktete vertikale Linie zeigt den Mittelwert aller Themen innerhalb des Datensatzes an. Einige Themen sind für Visualisierungszwecke farblich hervorgehoben.

Tabelle 10 verdeutlicht dies durch aggregierte Statistiken für jeden dieser Datensätze. Die größten Unterschiede zwischen Themen bestehen auf Stackexchange, mit einer $Frac_{geo}$ -Spanne von 73,1% und einer Standardabweichung σ von 26,8%. Das Thema *Technologie* wird auf Stackexchange nur mit 2,27% $Frac_{geo}$ gemessen, das Thema *Reisen, Tourismus & Migration* hingegen mit 75,3%. Ähnlich hoch war die Variabilität auf Reddit (Spanne von 61,6%, $\sigma = 17,2\%$). Bei den anderen Datensätzen war sie weitaus geringer. Insgesamt war $Frac_{geo}$ bei GDELT am höchsten (im Mittel bei 69%) und bei GDELT-IA-Americana am geringsten (11%). Die anderen Quellen liegen dazwischen, wobei $Frac_{geo}$ bei Twitter niedriger (im Mittel bei 13% und 14%) und die Webplattformen Reddit (im Mittel 27%) und Stackexchange (im Mittel bei 28%) höher liegt.

Tabelle 10: Deskriptive Statistiken für Frac_{geo} über die Datensätze hinweg

Datensatz	Minimum	Maximum	Mittel	σ	Spanne
GDELT-GKG	0,622	0,789	0,687	0,051	0,167
GDELT-IA-Americana	0,035	0,400	0,116	0,096	0,365
Nairaland	0,063	0,346	0,167	0,079	0,283
Twitter (ohne Geotags)	0,017	0,353	0,133	0,087	0,336
Twitter (mit Geotags)	0,032	0,330	0,138	0,075	0,298
Reddit	0,097	0,713	0,274	0,172	0,616
Stackexchange	0,022	0,753	0,280	0,268	0,731

Neben der Häufigkeit von Ortsreferenzen in Themen ist auch von Interesse, ob diese Häufigkeit Frac_{geo} in einer Beziehung mit der Zeit der Veröffentlichung, den Autoren und der Länge von Texten steht.

In der **zeitlichen** Dimension (Abbildung 48a) sind die Daten von GDELT-GKG und Twitter trotz kurzfristiger Variabilität langfristig stabil in der Frac_{geo} . Bemerkenswerterweise ist in den Twitter Datensätzen ein Anstieg von Ortserwähnungen im Sommer 2018 zu verzeichnen, ungefähr zu der Zeit als die FIFA-Fußballweltmeisterschaft ausgetragen wurde. Bei Stackexchange zeigt sich eine Zunahme der Ortsreferenzen über die Zeit hinweg, was vermutlich in Zusammenhang mit der Aufnahme neuer Seiten und damit möglicherweise neuer Thematiken in das Stackexchange Netzwerk steht. Bei GDELT-IA-Americana nimmt Frac_{geo} ab dem Jahr 2011 stark ab.

Zwischen **Autoren** (Abbildung 48b) bestand eine erhebliche Variabilität in fast allen Datensätzen. Die Standardabweichung über Autoren hinweg war größer als der Mittelwert, also mit Variationskoeffizienten von >1 . Die einzige Ausnahme stellt GDELT-GKG dar, mit einem Variationskoeffizienten von 0,36. Bemerkenswert ist, dass für Reddit und Stackexchange der Median von Frac_{geo} bei 0 lag, also für mehr als die Hälfte der Autoren keine Ortserwähnungen im Datensatz gefunden wurden.

Auch die **Textlänge** (Abbildung 48c) hat einen erheblichen Einfluss auf die Häufigkeit von verortbaren Georeferenzen. Der Zusammenhang ist für Twitter am stärksten ausgeprägt.

6.5.2. Ergebnisse der Modellierung

Der Modellierungsansatz bereinigt diese sichtbaren Einflüsse, um den Effekt der Themen zu herauszuarbeiten. Die Ergebnisse bestätigen in allen Datensätzen einen signifikanten Zusammenhang zwischen Themen und Wahrscheinlichkeit von verortbaren Ortserwähnungen. Die festen Effekte der Themen sind nach Datensätzen differenziert in Abbildung 49 dargestellt, die kompletten Ergebnisse der Modelle finden sich im Appendix A12.

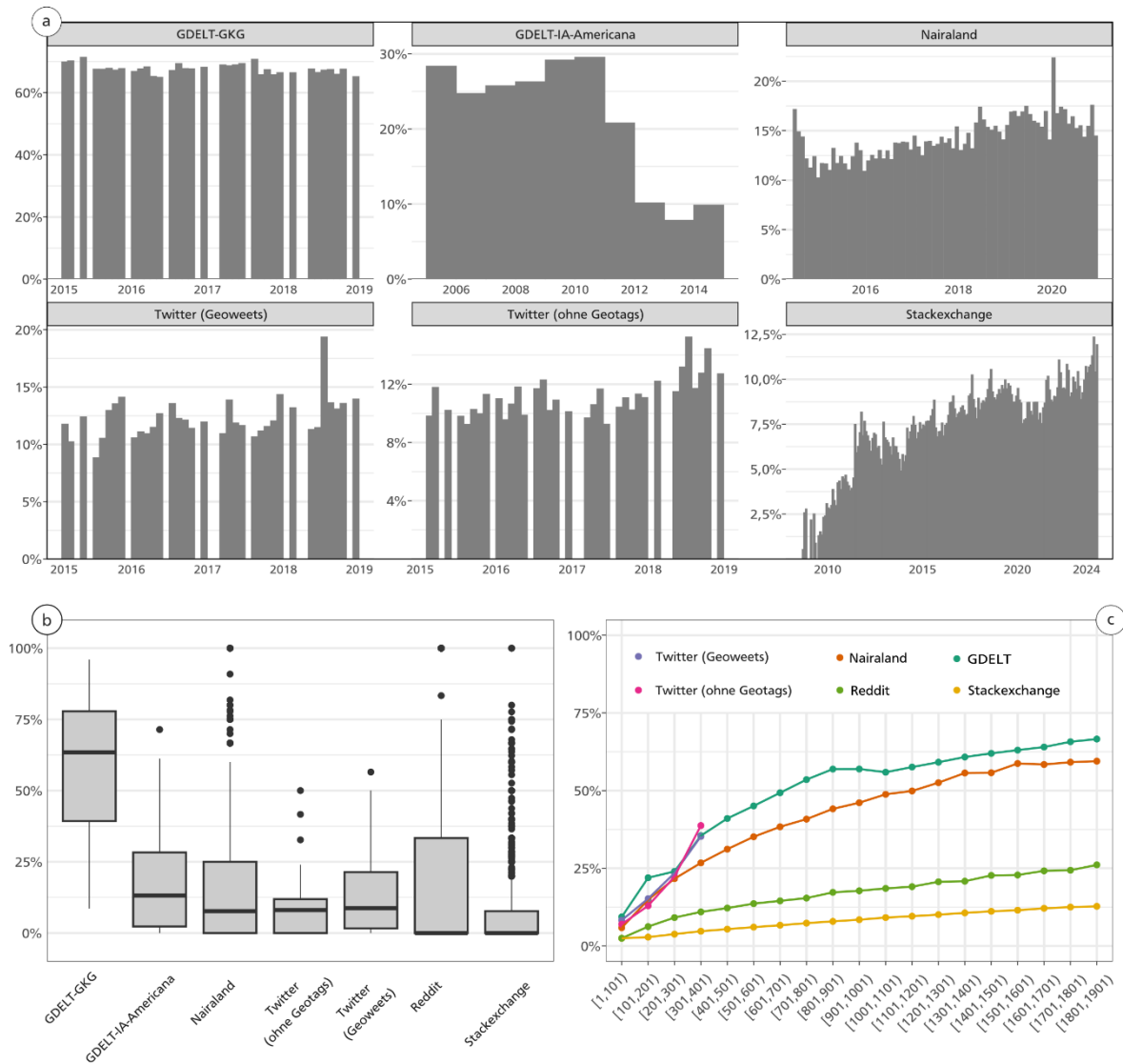


Abbildung 48: Gemessene $Frac_{geo}$ nach Zeitintervallen (a), Quellen (b), und Textlänge (c)

Für alle Datensätze gilt, dass die meisten Themen signifikant ($p < 0,001$) von der Basiskategorie *Sonstiges* abweichen. Der stärkste positive Effekt von 2,52 wird für das Thema *Reisen, Tourismus & Migration* auf Stackexchange geschätzt. Auf Stackexchange findet sich auch der stärkste negative Effekt von -2,47 für das Thema *Technologie*.

Die relative Rangfolge der Themen ist moderat konsistent über Datensätze hinweg. *Reisen, Tourismus & Migration* rangiert in der Regel hoch, genau wie *Internationale Politik*, wohingegen *Technologie* und *Privatleben, Familie & Beziehungen* niedrige Koeffizienten aufweisen. Hierbei finden sich jedoch auch Ausnahmen. In den Geotweet-Daten liegt *Reisen, Tourismus & Migration* nur an fünfter Stelle. Auf Reddit ist *Gesundheit* das am wenigsten räumliche Thema, liegt bei den anderen Datensätzen aber im Mittelfeld. Das Bild zeigt daher keine eindeutige Konsistenz zwischen Datensätzen.

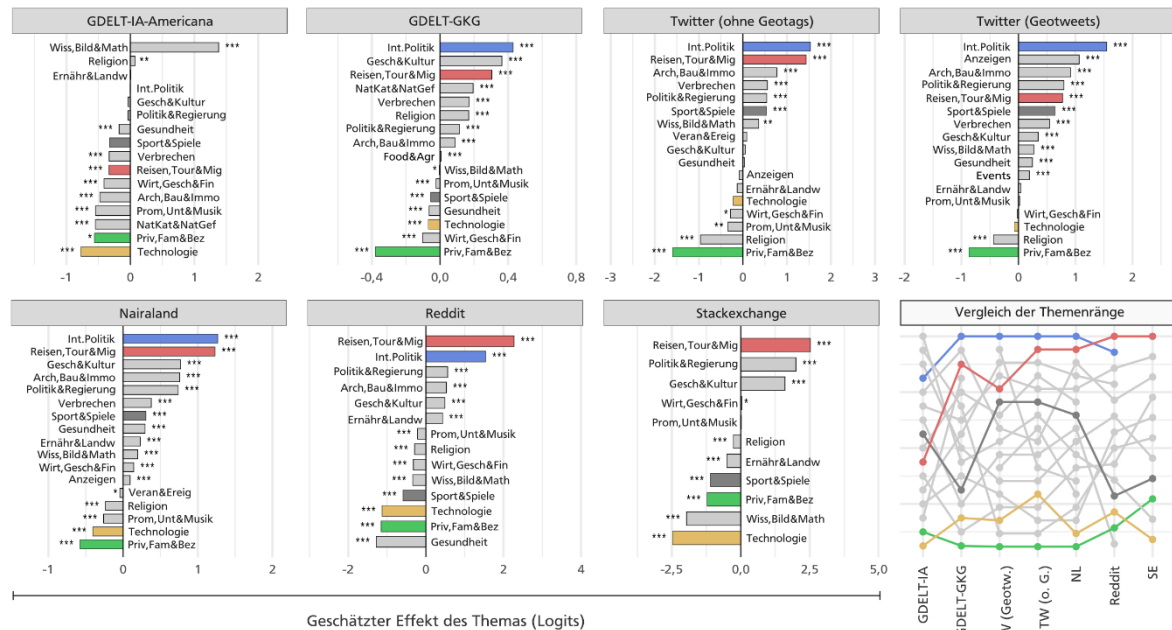


Abbildung 49: Durch Modellierung geschätzte Effekte jedes Themas auf die Wahrscheinlichkeit (Odds) eines Textes, eine verortbare Ortserwähnung zu enthalten

Die Übereinstimmung des Rankings zwischen Datensätzen wird mittels des Spearman Rangkorrelationskoeffizienten ρ quantifiziert (Tabelle 11). Insbesondere zwischen den drei Datensätzen aus dem nigerianischen Kontext ist die Rangkorrelation hoch, wobei zu berücksichtigen ist, dass die Themenklassifikation der Twitter-Daten von Nairaland-Daten abgeleitet wurde, und daher größere Ähnlichkeiten als zu anderen Plattformen nachvollziehbar sind. Doch auch zwischen unterschiedlichen Plattformen gibt es bemerkenswerte Ähnlichkeiten. Das Ranking auf Reddit korreliert mit dem Ranking auf Stackexchange ($\rho = 0,87$ mit $p < 0,001$), sowie auch mit dem auf GDELT-GKG ($\rho = 0,84$ mit $p < 0,001$) und, wenngleich mit geringerer Signifikanz, in den Twitter-Datensätzen ($\rho = 0,64$ mit $p < 0,05$ und $\rho = 0,68$ mit $p < 0,01$). Auffällig ist, dass GDELT-GKG, dessen Themen aus einer Multi-label-Struktur abgeleitet wurden, signifikant mit allen Einzeln gelabelten Datensätzen korreliert ist, jedoch nicht mit GDELT-IA-Americana, dessen Themenklassifizierung mutmaßlich auf derselben Methode basiert wie GDELT-GKG. Das könnte als Indiz dafür gesehen werden, dass weniger der Algorithmus als die Textsorte die thematische Geospatiality beeinflusst.

Jedoch muss bedacht werden, dass GDELT-IA-Americana mit gar keinem der Datensätze korreliert –was eine Ausnahme von der Regel darstellt, denn abgesehen von GDELT-IA-Americana sind die einzigen nicht signifikant korrelierten Rankings zwischen Stackexchange und Twitter. Alle anderen Beziehungen sind signifikant korreliert mit Korrelationskoeffizienten zwischen 0,64 und 0,87, was als moderate bis sehr starke Korrelation interpretiert werden kann (Akoglu 2018).

Tabelle 11: Spearman-Rangkorrelation zwischen den Topic-Rankings jedes Datensatzes

Dataset	Twitter (ohne Geotags)	Twitter (mit Geotags)	Nairaland	Reddit	Stack- exchange	GDEL- T-GKG	GDEL- T- IA- Americana
Twitter (ohne Geotags)		0,85***	0,86***	0,64*	0,39	0,65**	0,21
Twitter (mit Geotags)	0,85***		0,79***	0,68**	0,55	0,64**	0,23
Nairaland	0,86***	0,79***		0,72**	0,67*	0,79***	0,34
Reddit	0,64*	0,68**	0,72**		0,87***	0,84***	0,25
Stackexchange	0,39	0,55	0,67*	0,87***		0,61*	0,11
GDEL-T-GKG	0,65**	0,64**	0,79***	0,84***	0,61*		0,4
GDEL-T-IA-Americana	0,21	0,23	0,34	0,25	0,11	0,4	

6.5.3. Validierung

Tabelle 12 zeigt die Ergebnisse des Vergleichs der erkannten Ortsreferenzen zwischen dem algorithmischen Ansatz und menschlicher Annotation. Für die meisten Datensätze gibt es große Übereinstimmungen mit Gesamtgenauigkeiten von 93 bis 97% und κ von 0,74 bis 0,92. Als Ausnahme ist hier GDEL-T-GKG zu nennen, mit einer Gesamtgenauigkeit von nur 69% und κ von 0,25. Für die Themenklassifikation sind die Übereinstimmungen geringer, mit Gesamtgenauigkeiten zwischen 32% und 67% und κ von 0,28 bis 0,64. Hier stellt der Stackexchange-Datensatz eine positive Ausnahme dar, bei der die Übereinstimmung deutlich höher liegt (Gesamtgenauigkeit 85%, κ 0,84). Bemerkenswert ist, dass die meisten Unterschiede der Annotationen für die Kategorie *Sonstiges* auftreten. Appendix A11 zeigt die vollen Ergebnisse der Validierung als Konfusionsmatrizen.

Tabelle 12: Validationsgenauigkeiten. TP: Richtig Positive. TN: Richtig Negative. FP: Falsch Positive. FN: Falsch Negative.

Datensatz	Gesamt- genauigkeit	κ	Präzision	Sensitivität	FP	FN	TP	TN	n
GDEL-T-GKG	0,70	0,27	0,92	0,70	12	64	146	30	252
Nairaland	0,97	0,88	0,88	0,91	3	2	21	154	180
Twitter (mit Geotags)	0,94	0,76	0,88	0,73	3	8	22	147	180
Twitter (ohne Geotags)	0,93	0,74	0,92	0,68	2	11	23	144	180
Stackexchange	0,94	0,83	0,79	0,96	6	1	22	91	120
WEBIS Reddit	0,97	0,92	0,88	1,00	4	0	29	115	148

6.5.4. Interpretation

Die Häufigkeit von Ortserwähnungen in Texten variiert stark zwischen Texten verschiedener Themen. Die Ergebnisse der Modellierung unterstützen die Hypothese, dass ein messbarer Einfluss hierbei auf die Themen selbst zurückzuführen ist. Mit anderen Worten: Das Thema eines Web-Textes kann einen Einfluss auf dessen Wahrscheinlichkeit haben, eine verortbare räumliche Referenz zu enthalten.

Diese *Geospatiality*-Effekte von Themen können erheblich sein. So ist der Effekt von *Religion* im nicht-geotaggten Twitter Datensatz mit 1 logit geschätzt, was bedeutet, dass Texte dieses Themas etwa 2,7-mal häufiger verortbare Georeferenzen enthalten als Themen der Vergleichskategorie *Sonstiges*. Für viele Themen ist der Effekt noch größer, zudem variiert er stark von Thema zu Thema und zwischen Datensätzen. Es findet sich damit ein erheblicher Einfluss der Datensätze, und damit der Textsorten, Kommunikationsarten und Intentionen der Autoren, welche sich von Microblogs wie Twitter stark zu Nachrichtenartikeln und Q&A

Websites unterscheiden. Auch die Strukturen der Plattformen spielen wahrscheinlich eine Rolle: Dass die Unterschiede zwischen Themen auf SE und Reddit hoch waren, könnte an der Moderation dieser Plattformen liegen, welche inhaltliche Relevanz der Nutzerbeiträge und damit thematische Homogenität fördert. Zudem kann auf manchen Plattformen, wie bereits beschrieben, auch Topic-Drift stattfinden (Hobbs 1990). Trotz dieser Unterschiede weisen die Rankings der Themen signifikante –wenngleich nicht perfekte– Korrelationen auf. Es gibt also Konsistenzen der thematischen Geospatiality über Datensätze hinweg, welche wahrscheinlich das Maß des Zufalls überschreiten. Dies könnte ein Indiz dafür sein, dass Geospatiality eine inhärente Eigenschaft von Themen ist.

Neben den Effekten von Thema und Datentyp werden auch erhebliche Einflüsse von Textlänge, Zeit und Autor gefunden. Dass der Autor großen Effekt hat, bestätigen Ergebnisse von Karami et al. (2021). Aufgrund dieser Variabilität ist die Kontrolle für solche Variablen, wie in dieser Dissertation durch den Einsatz eines GLMMs, empfehlenswert.

Die Bedeutung dieser Ergebnisse im Rahmen der Studie ist zweierlei: **Erstens** deutet sich an, dass beim Vergleich von Themen in georeferenzierten Textdaten Rücksicht auf Geospatiality-Effekte genommen werden sollte. Diese könnten zu einem zusätzlichen Bias führen, wodurch die Ergebnisse nicht repräsentativ für die Gesamtheit der Webtextdaten sind.

Bei der komparativen Analyse von Themen im Raum kann der Vergleich der absoluten Zahlen von Themenerwähnungen zu falschen Interpretationen ihrer Bedeutung führen. Grundsätzlich kann dies durch Normalisierung der Zahlen innerhalb jedes identifizierten Themas korrigiert werden –sofern davon ausgegangen werden kann, dass die thematische Geospatiality konstant ist, sie also nicht zwischen Regionen oder im Laufe der Zeit variiert. Um solche Variationen zu identifizieren sind spezielle auf diese Aspekte adressierte Forschungsarbeiten notwendig. In dieser Studie sind bereits Unterschiede nach Datensatz messbar. Daraus lässt sich die Empfehlung ableiten, dass wenn also eine Analyse über verschiedene Datensätze hinweg durchgeführt wird, eine Normalisierung von Themen innerhalb jedes Datensatzes separat durchgeführt werden sollte.

Zweitens können die Ergebnisse als Ungleichheit in der Verfügbarkeit von Daten interpretiert werden. Anders ausgedrückt, georeferenzierte Webtextdaten sind nicht für alle thematischen Anwendungen gleich ergiebig. Für Forschung zu Privatleben von Individuen, privaten Beziehungen und geistiger Gesundheit, für welche Social Media grundsätzlich eine sinnvolle Datenquelle darstellt (Chancellor und De Choudhury 2020), sind Einschränkungen bei der Erweiterung um eine räumliche Dimension zu erwarten. Demgegenüber sind für Studien von Reisen, Tourismus und weltpolitischem Diskurs verhältnismäßig große Mengen an georeferenzierten Textdaten zu erwarten.

Zusätzlich ist messbar geworden, dass Geotweets und Tweets ohne Geotags in Bezug auf Geospatiality ähnlich, aber nicht identisch sind, was Ergebnisse von Serere und Resch (2024) ergänzt. Dies kann die Übertragbarkeit von Ergebnissen aus Studien, welche bereits mittels Geotweets durchgeführt wurden, maßgeblich beeinflussen. Diese Unterschiede zwischen Geotweets und Tweets ohne Geotags werden in Studie C2 genauer beleuchtet.

Über den geographischen Kontext dieser Dissertation hinaus liefert diese Studie erste empirische Evidenz von themenspezifischer Geospatiality in wissenschaftlich genutzten Webtextdaten und kann damit dazu beitragen, Räumlichkeit als ein sprachliches, kognitives und literarisches Konzept zu verstehen (Tally 2012).

Der Rahmen, in welchem diese Ergebnisse zu verstehen sind, wird durch die spezifischen Limitationen definiert. In dieser Studie wurden Datensätze zusammengeführt, welche in einer Vielzahl von Charakteristika variierten. Zwar erlaubt dies, die Verbreitung der getesteten Effekte besser abzuschätzen. Dennoch sind die genutzten Daten, auch wenn sie mehrere verschiedene Textsorten und verschiedene Zeiten abdecken, kein ausreichendes und repräsentatives Abbild aller Webtextdaten. Die hier dokumentierten thematischen Geospatiality-Effekte sollten also nicht als ein generelles Phänomen von Textdaten gesehen werden und schon gar nicht von Sprache allgemein. Gleichzeitig erschwert es die Verschiedenheit der Datensätze, die Einflüsse der verschiedenen Charakteristika herauszuarbeiten. Es wäre für zukünftige Forschung wünschenswert, sich auf bestimmte Charakteristika zu fokussieren und Datensätze für deren Untersuchung entsprechend zusammenzustellen. Ein diverserer und gut annotierter Korpus von metadaten-reichen globalen Social-Media-Texten wäre hierfür wünschenswert, ist aber aktuell nicht frei verfügbar.

Die Definition von Geospatiality als das Vorkommen einer geographisch verortbaren Georeferenz in einem Text, das zentrales Kriterium und Variable der Analyse ist, stellt grundsätzlich eine Abstraktion dar. Tatsächlich existiert vermutlich keine einzige universelle und absolute Definition von Räumlichkeit, sondern verschiedene Grade und Perspektiven. Diese *Fuzziness* von Räumlichkeit ist kein theoretisches Phänomen von rein sprachwissenschaftlichem Interesse. Eine qualitative Analyse der Validierungsdaten zeigt, dass viele falsch positive Klassifizierungen durch Demyne und Metonyme verursacht werden, sowie durch nicht-geographische Orte wie *Narnia* oder *Jupiter* und durch Ortsnamen in einem nicht-geographischen Kontext wie *Virgin America*. Je nach Räumlichkeitskonzept und Forschungsfrage könnten diese Arten von Ortserwähnungen valide und nützliche Geoinformationen sein. Die technische Implementierung durch algorithmische Differenzierung von Phänomenen wie Metonymie ist jedoch eine technische Herausforderung und stellt auch in dieser Studie eine Limitation dar. Es ist möglich, dass technische Fortschritte im Bereich LLMs und AI es künftig erlauben könnten, derartige Formen zu identifizieren, und verschiedene Konzepte von Geospatiality empirisch zu vergleichen. Es steht zu erwarten, dass dies einen merkbaren Einfluss auf das Ergebnis hätte. Ob beispielsweise Demyne als räumlich betrachtet werden, sollte einen großen Einfluss auf die Geospatiality eines Themas wie *Geschichte & Kultur* haben.

Selbstverständlich ist auch die Klassifizierung der Themen, wie sie hier durchgeführt wurde, nur eine Möglichkeit der Strukturierung von Texten. Geospatiality könnte sich unter anderen Taxonomien anders präsentieren. In dieser Studie wurde eine über Plattformen hinweg konsistente Einteilung von Themen angestrebt. Es zeigte sich, dass diese Konsistenz nur begrenzt realisierbar ist. Auf einem Diskussionsforum wie Nairaland ist das, was als *Politik* bezeichnet wird, interaktiver, diskursiver und geographisch fokussierter als in den globalen politischen Nachrichten, welche durch GDELT registriert werden. Im Referenzrahmen ihrer Plattform können beide Ausprägungen als politisch betrachtet werden, doch selbst bei identischem inhaltlichen Bezug (beispielsweise einem politischen Gipfeltreffen) nimmt der Diskurs eine unterschiedliche Form ein. Zudem ist klar, dass die Einordnung von Texten in klar abgegrenzte thematische Kategorien nicht der Realität entspricht, da Texte gemischte Inhalte enthalten können und eindeutige Abgrenzungen im semantischen Raum nicht immer möglich sind, da beispielsweise bei der Interpretation des Fokus oder der Funktion eines Textes durch Autor und Adressat eine Unschärfe besteht. Die in dieser Studie durchgeführte

Validierung bestätigt dies mit einer geringen Übereinstimmung zwischen dem semialgorithmischen und dem menschlich annotierten Ansatz. Viele der Unstimmigkeiten sind bei näherer Betrachtung nachvollziehbar oder zumindest interpretierbar. Beispielsweise zeigte die menschliche Annotation eine stärkere Neigung dazu, bei Ambiguität die Kategorie „Sonstiges“ zu wählen. Zudem ergab sich eine höhere Übereinstimmung bei den Plattformen Stackexchange und Reddit, was vermutlich auf eine dort höhere thematische Homogenität und geringere inhärente Fuzziness der Textinhalte zurückzuführen ist, die sich durch die von Moderatoren durchgesetzten Regeln dieser Plattformen ergeben. Angesichts dieser Effekte ist trotz der geringen Übereinstimmung die Datengrundlage und damit die insgesamt Schlussfolgerung dieser Studie plausibel. Mehr noch wird die Stärke des Geospatiality-Effektes dadurch unterstrichen, dass trotz der oben diskutierten Unschärfen und Abstraktionen starke Muster zu erkennen sind.

Zusammenfassend bestehen viele Möglichkeiten, thematische Geospatiality zu definieren und messbar zu machen. Der in dieser Studie genutzte Ansatz wurde gewählt, um den aktuellen Stand der Forschung und Software zu reflektieren.

Zukünftige Forschung sollte Konzepte der Geospatiality aus verschiedenen Feldern (Tally 2012) bedienen und empirisch testen. Um dies möglich zu machen, sind Datensätze nötig welche nicht unbedingt größer, aber systematisch vergleichbarer und gleichzeitig repräsentativer sind. Damit könnten auch die Einflüsse von den hier beschriebenen Charakteristika und weiteren Charakteristika untersucht werden. Der Effekt von **Zeit**, hier als zufälliger Effekt nicht stärker berücksichtigt, könnte selbst von Interesse sein. Beispielsweise ist *Gesundheit* nicht grundsätzlich ein sehr räumliches Thema. Doch wurden, während der Covid-19 Pandemie, zahlreiche Gesundheitsupdates, Nachrichten, Maßnahmen und Indikatoren räumlich referenziert gepostet, während andere Aspekte des Lebens durch Ausgehbeschränkungen weniger räumlich wurden. Ein weiteres Beispiel aus dieser Studie ist der Anstieg in Geospatiality des Themas *Sport & Spiele* auf Nairaland und Twitter während der Fußballweltmeisterschaft 2018. Diese anekdotische Beobachtung könnte weiter untersucht werden, um bessere Einschätzungen durch Webdaten zu ermöglichen. Dies könnte einem verbesserten Monitoring von Ereignissen durch Social Media dienen (Senaratne et al. 2014) und indirekt damit auch die Untersuchung migrationsrelevanter Faktoren mittels GSM verbessern.

Auch bei Variationen im Raum besteht Potenzial für tiefergehende Forschung. Beispielsweise ist es möglich, dass Innenpolitik in größeren Ländern mit räumlich ausgebildeten ethnischen Grenzen (beispielsweise Nigeria) ein räumlicheres Thema ist als in kleineren, ethnisch homogenen Ländern (beispielsweise Lesotho).

Insgesamt lässt sich schlussfolgern, dass Geospatiality ein schwer zu definierendes aber gerade deshalb interessantes und vielseitiges Konzept ist. Als Nexus zwischen theoretischen und praktischen Forschungsfeldern hält seine Untersuchung großes Potenzial für die Geographie und benachbarte technische und soziale Forschungsfelder, methodische Verbesserungen und ein besseres Verständnis der Bedeutung von Raum zu ermöglichen.

6.6. Studie C2: Differenzen der Räumlichkeit von Themen zwischen Geotag- und Geoparsing-basierten Ansätzen

Bei der Untersuchung von Geospatiality sowie bei der Nutzung von GSM im Allgemeinen ist von Interesse, inwiefern die Art der genutzten Georeferenz das Ergebnis beeinflusst. Studie C2 untersuchte daher Geotag-basierte und Geoparsing-basierte Ansätze mit einem Fokus auf die beobachteten Geospatiality-Effekte.

6.6.1. Unterschiede in der Häufigkeit von Ortsnennungen zwischen Tweets mit und ohne Geotags

Über den gesamten Datensatz hinweg zeigt sich, dass in 12,5% der Geotweets verortbare Orte erwähnt werden, und damit leicht häufiger als in den Tweets ohne Geotags ($\text{Frac}_{\text{geo}} = 11,2\%$). Wie Abbildung 50 zeigt, gilt dies auch innerhalb der meisten Themen. Es gibt jedoch zwei Ausnahmen. So ist die Differenz in *Anzeigen* deutlich ausgeprägter, mit 25,6% in Geotweets zu 9,9% in Tweets ohne Geotags. In *Reisen, Tourismus & Migration* ist das Verhältnis sogar umgekehrt: Hier sind in 16,5% der Geotweets Ortsreferenzen zu finden, und damit deutlich seltener als in den Tweets ohne Geotags (30,4%).

Insgesamt zeichnet sich ein starker Einfluss des Themas ab, welcher zu Schwankungen von Frac_{geo} von 1,7% bis 35,2% führt.

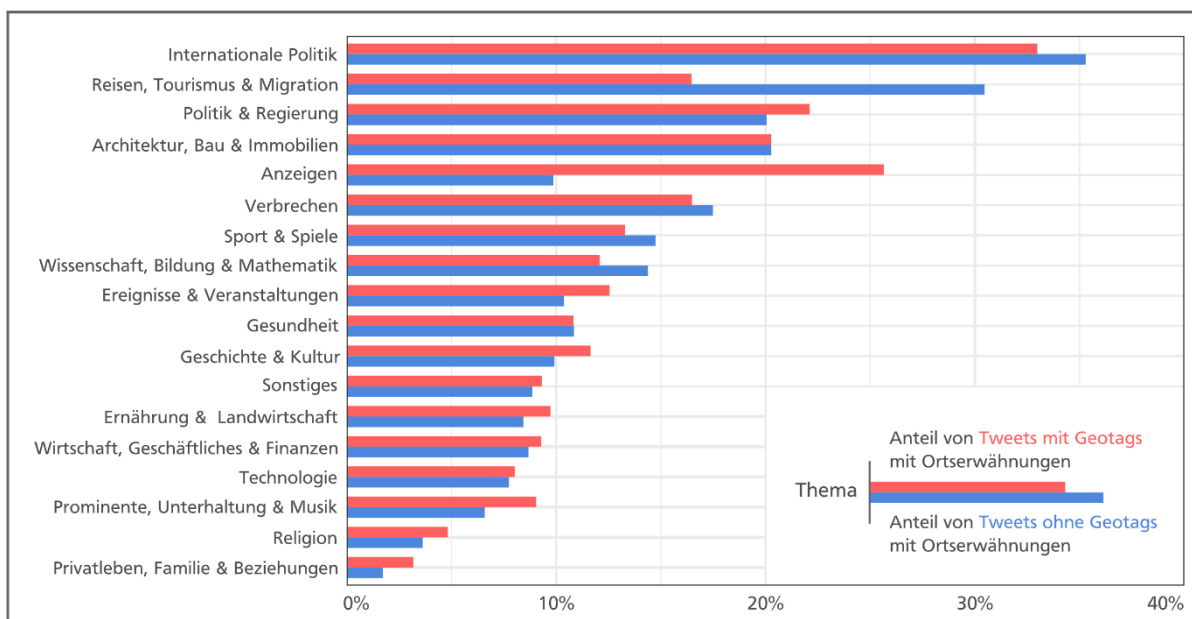


Abbildung 50: Vergleich von Frac_{Geo} zwischen Geotweets (rote Balken, $N=314\,620$) und Tweets ohne Geotags (blaue Balken, $N=70\,745$) für die 17 Themen, jeweils basierend auf Daten von in Nigeria verorteten Nutzenden.

6.6.2. Differenzen in Räumlichkeit

Durch den Modellierungsansatz kann dieser Einfluss herausgearbeitet werden.

Abbildung 51 kontrastiert die Koeffizienten der Themen in M_{tag} und M_{txt} . Koeffizienten in M_{tag} , dem Geotag-basierten Ansatz, sind insgesamt geringer als im auf Ortserwähnungen basierten Ansatz M_{txt} . Für M_{tag} sind für 11 der 17 Themen signifikante Koeffizienten geschätzt, für M_{txt} 12 der 17. In beiden Perspektiven lässt sich also ein Geospatiality Effekt sehen. Die Ausprägung und selbst die Richtung dieses Effektes ist jedoch keinesfalls identisch. Die Logits von *Reisen, Tourismus & Migration* liegen bei M_{tag} etwa doppelt so hoch wie bei M_{tag} . Der Effekt

von *Internationale Politik* ist bei M_{tag} leicht negativ, bei M_{txt} stark positiv im Vergleich zur Referenzkategorie *Sonstiges*.

Damit ergeben sich auch unterschiedliche Themen-Rangfolgen. Diese Rangfolgen der Themen sind zwischen den beiden Ansätzen nicht statistisch korreliert (Spearman $\rho = 0,45$, p-Wert = 0,073)

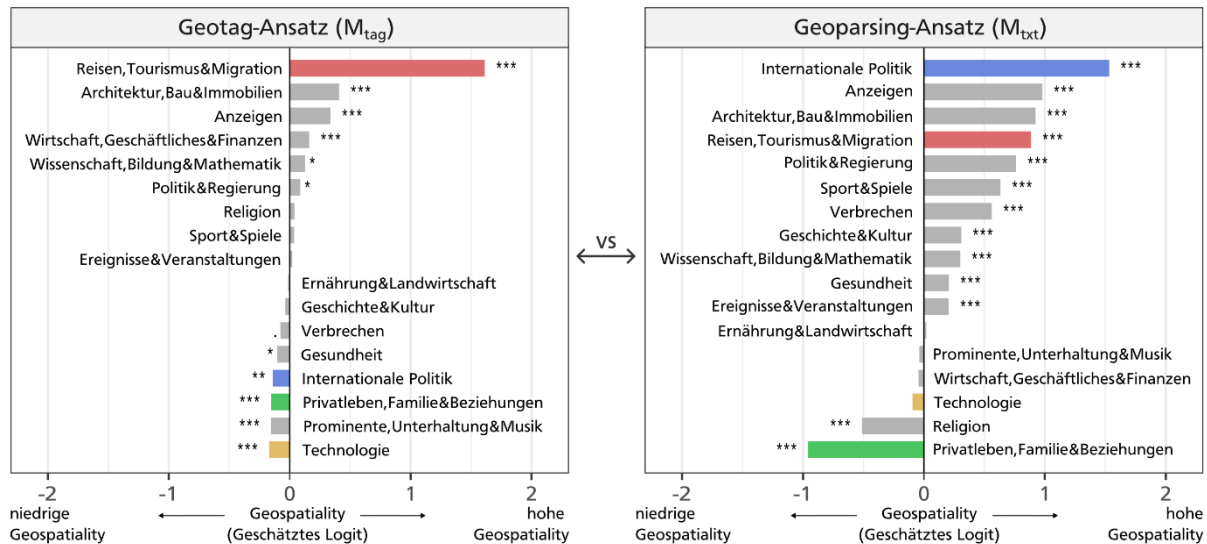


Abbildung 51: Durch Modellierung geschätzte Effekte (log-odds Ratio, Logits) jedes Themas auf die Wahrscheinlichkeit (Odds) eines Tweets, entweder einen Geotag (links) oder eine verortbare Ortserwähnung (rechts) zu enthalten. Positive Werte deuten darauf an, dass das Thema im Vergleich zur Referenzkategorie einen positiven Effekt auf die Wahrscheinlichkeit von Ortserwähnungen hat.

6.6.3. Interpretation

Klar ist, dass innerhalb dieser Daten die Nutzung von Geotag- und Geoparsing zu unterschiedlichen Themenverteilungen in den Daten führt. Die Wahl zwischen Geotags und Geoparsing als Quelle zur Georeferenzierung könnte also die Ergebnisse von Studien maßgeblich beeinflussen. Zwar ist die Repräsentativität dieses Datensatzes begrenzt, es gibt aber doch Anlass, die Vergleichbarkeit der Ergebnisse von Studien, die auf Geotweets basieren, und Studien, die Geoparsing nutzen, kritisch zu hinterfragen.

Dass die modellierten Effekte der Themen bei der Nutzung von Geoparsing stärker sind als bei Geotags ist plausibel, da bei diesem Ansatz sowohl Geoinformation als auch Themeninformation aus demselben Text abgeleitet werden, und vermutlich teilweise durch dieselben Worte definiert werden. Damit ist der starke Effekt von *Internationale Politik* durch das Vorkommen von Toponymen zu erklären, welche als stellvertretend für politische Einheiten dienen (*Deutschland* für die Bundesrepublik Deutschland). Interessant ist im Fall von *Internationale Politik* auch, dass sich dieser Zusammenhang nicht auf das Setzen von Geotags erstreckt. Das Kommentieren internationaler Politik hat also offenbar keinen zusätzlichen Anreiz auf Nutzende, ihre Kommentare durch Geotags zu verorten. Zwar erwähnen in diesem Thema auch viele Tweets mit Geotags einen Ort, doch dies ist genauso für Tweets ohne Geotags der Fall (Abbildung 50).

Dass im Fall von *Reisen, Tourismus & Migration* Geotweets weniger häufig Ortserwähnungen erwähnen als Tweets ohne Geotags ist bemerkenswert. Eventuell stehen der aktiv gesetzte Geotag und der implizite Bezug, welcher sich im Text findet, für die reisenden Menschen in einer substitutiven Funktion zueinander, bei welcher das Geotaggen in manchen Situationen

die Erwähnung eines Ortes überflüssig macht. Alternativ ist es möglich, dass der Ansatz unterkomplex für die Abbildung der Sprachformen und Textinhalte ist. Studie B1 (6.3.3) zeigte, dass algorithmisch klassifizierte Themen intern sehr divers sind und viele Unterthemen beinhalten können. Es ist also vorstellbar, dass Geotweets einen anderen Teilbereich des Themas *Reisen, Tourismus & Migration* (beispielsweise Berichte über Besuche berühmter Sehenswürdigkeiten) als Tweets ohne Geotags (beispielsweise Reisepläne). Auch bei anderen Themen bestehen Abweichungen. Bei *Politik* weisen Geotweets mehr Ortserwähnungen auf, bei *Internationale Politik* jedoch weniger. Dies differenziert eine von Tear (2018) auf demographischer Ebene in einem politisch-fokussierten Twitter Datensatz gemachte Beobachtung, dass Nutzende des Geotaggings weniger oft Orte erwähnen als Nutzende, die kein Geotagging nutzen.

Die Funktion der Ortserwähnung spielt hierbei mit großer Wahrscheinlichkeit eine wichtige Rolle. Selbst innerhalb eines Themas kann diese Funktion sich zwischen Unterthemen deutlich unterscheiden. So kann bei der Teilnahme an politischen Veranstaltungen wie einer Demonstration die Erwähnung des eigenen Standortes und damit des Ortes der Veranstaltung die Funktion von Solidarisierung und Bildung von Gruppenzugehörigkeit erfüllen. Demgegenüber stehen bei der Kommentierung von internationalen politischen Verträgen die Namen der betreffenden Staaten und eventuell des Ortes der Vertragsunterzeichnung im Vordergrund. Auch in dieser Hinsicht sind die Zusammensetzungen der Themen wahrscheinlich essenziell für deren durchschnittliche Räumlichkeit.

Ob der beobachtete Effekt auf einen dieser vorgeschlagenen Mechanismen zurückzuführen ist, könnte durch gezielte Analysen herausgearbeitet werden. Als Datenquelle hierfür ist Twitter eine sinnvolle Plattform, da durch den in dieser Dissertation genutzten Ansatz ein vergleichbarer und Metadaten-reicher Korpus generiert werden kann, welcher sowohl Tweets mit als auch ohne Geotags enthält. Plattformspezifische technische Effekte, wie die Freiwilligkeit und Nutzbarkeit der Geotagging-Funktion, müssen dabei selbstverständlich berücksichtigt werden.

Die wechselseitige Beziehung von Geotagging und Ortserwähnungen ist ein fruchtbares Forschungsfeld für die Webdatenforschung und Sprachwissenschaft. Drücken der aktiv gesetzte Geotag und der implizite Bezug, welcher sich im Text findet, unterschiedliche Konzepte von Bezugsorten aus, stehen sie in Konkurrenz, oder werden sie komplett unabhängig voneinander genutzt? Es könnte ein Kontrast zwischen qualitativen und quantitativen Raumwahrnehmungen vorliegen, wo Ersterer den erlebten Raum des Autors reflektiert, Letzterer den (Anspruch des Autors auf den) objektiven Raum (Dangschat 2019). Die differenziertere Klassifizierung von Ortserwähnungen, beispielsweise in Fokus-Orte und sekundäre Orte (D'Ignazio et al. 2014), könnte zudem tiefergehenden Einblick in unterschiedliche Funktionen von Ortserwähnungen und ihre Beziehung zu Räumlichkeitseffekten geben.

Für die Zwecke dieser Dissertation ist eine Beschreibung des beobachteten Effekts ausreichend. Zusammenfassend kann über die Ergebnisse von Studie C2 gesagt werden, dass (1) Geospatiality-Effekte sowohl für Geoparsing und Geotagging auftreten und (2), dass diese Ansätze nicht zu denselben Ergebnissen führen, wenn die Anwendung sich auf Themen bezieht. Wird dies berücksichtigt und durch angemessene Maßnahmen korrigiert (beispielsweise Normalisierung oder Modellierung durch zufällige Effekte), können beide Ansätze höchst wertvolle Quellen räumlicher Daten sein.

7. Diskussion

7.1. Praktisches Potenzial

Wie ist angesichts dieser Ergebnisse das Potenzial von GSM zu bewerten, FED in der Kartierung von Migrationsfaktoren zu ergänzen? Im Folgenden (Kapitel 7.1.1) werden die Potenziale dieser Anwendung zusammengefasst. Es soll dabei über die theoretischen Möglichkeiten hinaus auch eine Einschätzung der Praktikabilität gegeben werden, welche die Datenverfügbarkeit und ethischen und rechtlichen Überlegungen miteinbezieht.

Die Einschätzung der Praktikabilität erfolgt auf Basis der drei Kriterien der (1) technischen Möglichkeiten, welche sich aus der Nutzung von GSM für die Ableitung von Migrationsfaktoren ergeben, (2) der geographischen, demographischen und inhaltlichen Abdeckung, welche mit GSM erreicht werden kann, und (3) der Verknüpfbarkeit von GSM-basierten Daten und Methoden mit Fernerkundungsansätzen. Die Ermessung dieser Kriterien soll jedoch keinen Anspruch auf Vollständigkeit haben. Eine Dissertation allein kann diese Aspekte nicht ausführlich und vollständig ergründen, zudem bedingen technische und soziale Entwicklungen ein ständig wechselndes Forschungsumfeld. Die Beiträge dieser Dissertation zielen darauf ab, den Stand der Forschung in diesen Aspekten zu vertiefen und für die spezielle Anwendung in der geographischen Erforschung von Migrationen zu testen. Zusätzlich wurden allgemeine Limitationen der Daten sichtbar. Diese Punkte, Limitationen, und die vielversprechendsten Entwicklungen im Bereich dieser Aspekte sollen nun diskutiert werden um abschließend eine Einschätzung zur Praktikabilität zu geben.

7.1.1. Technische Möglichkeiten der Nutzung von GSM

Zunächst erlaubt die in GSM enthaltene Geoinformation es, Textdaten in eine Art von Geodaten zu konvertieren, wodurch eine Bandbreite von geographischen Auswertungsmethoden ermöglicht wird. Lakes (2019) verweist auf topologische und räumliche Analysen wie die Berechnung der Distanz und Nachbarschaft zwischen Texten oder die Erkennung von Häufungen (Hotspots), sowie die Integration von Social-Media-Daten mit anderen Geodaten über ihre Raumkoordinaten sowie die Visualisierung als Karte. Insbesondere die Studien A1 und A2 dieser Dissertation verdeutlichen diese Potenziale durch die Verschneidung von Geotweets und fernerkundlichen Baujahrkartierungen, und die gemeinsame explorative Verarbeitung und Visualisierung.

In Studie A1 konnte durch diese räumliche Verknüpfung von Geotweet-Positionen und Fernerkundungskartierungen empirisch ein Trend zu einer digitalen Ungleichheit von neueren Stadtvierteln in Afrika nachgewiesen werden. Die Daten erlaubten eine großflächige Analyse über eine Fläche von über 10 500 km². Studie A2 zeigte, dass durch die Verknüpfung von Textinhalten und Fernerkundungsdaten und deren gemeinsame Auswertung mittels Mixed-Methods eine plausible Einschätzung dieser dynamischen Siedlungsräume ermöglicht wird, insbesondere in Bezug auf ihre soziale Funktion.

Im Migrationskontext sind Mobilitätsaspekte von besonderer Bedeutung. Wie Studie B1 zeigte, lässt sich für Twitter-Nutzende mit ausreichend häufiger Geotagging Aktivität diese Mobilität quantifizieren und mit Textinhalten verbinden. Was die Studie B1 nicht zeigte, aber in Bezug auf Studien A1 und A2 denkbar erscheint, ist die Integration von Fernerkundungsdaten und Nutzermobilitätsinformationen. Da viele Tweets mit verhältnismäßig geringer räumlicher Präzision verortet sind, funktioniert die Ableitung von Mobilitätscharakteristika insbesondere für regionale und internationale Studien gut. Daher

besteht das größte Potenzial der Methodik vermutlich in vergleichenden Studien von internationaler Migration, wo die Eigenschaften von GSM einen klaren Mehrwert zu traditionelle Datenquellen bieten (European Commission 2022). Hierbei stellen die unterschiedlichen Nutzungsgewohnheiten von GSM in unterschiedlichen Kulturen einerseits eine Herausforderung für internationale Studien dar, bieten gleichzeitig aber ein vielversprechendes Anwendungsfeld. Neben der räumlich weiten Abdeckung durch große Datenmengen besteht Flexibilität in der Definierung von Migrationskonzepten und Auswertemethoden, was Forschern eine hohe Flexibilität in Hypothesentests gestattet und erlaubt, mehrere Kriterien zu vergleichen und in einem Ensemble zu verbinden (wie beispielsweise von Johnson et al. (2016) vorgeschlagen). Hierdurch könnten beispielsweise Beziehungen zwischen Langzeit- und Kurzzeitmigration besser verstanden werden (Mau et al. 2015).

Neben der Auswertung thematischer Inhalte kann eine Vielzahl an mobilitätsrelevanten Textmerkmalen (beispielsweise Sprachmerkmale) abgebildet werden. Wie Studie B2 zeigt, können diese sowohl mit Mobilitätsinformationen als auch untereinander verknüpft werden, was potenziell sehr viele Dimensionen der Analyse ermöglicht.

Für Fernerkundung wurde bereits der Beweis erbracht, dass sie in vielen Bereichen sinnvolle Beiträge zur Migrationsforschung beitragen kann (siehe Kapitel 1.2). Werden zusätzlich aus GSM abgeleitete Informationen hinzugezogen, kann eine umso größere Bandbreite an relevanten Aspekten abgebildet und in Bezug zueinander gesetzt werden. Die Integration von GSM und FED liefert damit ein breiteres und, wenngleich noch immer nicht vollständiges, doch ein deutlich vielseitigeres und robusteres Bild auf menschliche Migration und ihre Beziehung zur Umwelt.

7.1.2. Abdeckung von GSM

Für geographische Fragestellungen sind großflächig verfügbare und grenzüberschreitend konsistente Daten von Vorteil. Zahlreiche FED und davon abgeleitete Produkte sind global und frei verfügbar, wenn auch mit variabler Dichte und Qualität (Sapena et al. 2024). GSM sind, wie in der Einleitung dargelegt, ein vielversprechendes Gegenstück mit ebenso hoher, wenn auch nicht vollständiger oder konsistenter Abdeckung. Die Studien dieser Dissertation bestätigen die Verfügbarkeit von großen GSM-Datenmengen aber auch deren heterogene Verteilung, welche als Einschränkungen der Abdeckung betrachtet werden muss, die mit GSM erreicht werden kann. Im Folgenden soll die Abdeckung auf Ebene von Geographie, Textinhalten und Mobilitätstypen und Demographien diskutiert werden.

Geographie: Die Möglichkeit, mittels GSM grenzübergreifende Analysen durchzuführen, wurde in dieser Dissertation demonstriert und die methodischen sowie inhaltlichen Mehrwerte können positiv bestätigt werden. So wurden in Studien A1 und A2 konsistente Methodologien über viele afrikanische Länder hinweg angewendet. In Studien B1 und B2 konnte die Auswertung von Trajektorien frei definiert und konsistent über Landesgrenzen hinweg erfolgen. Zum Teil ergeben sich hierbei Einschränkungen aus bekannten demographischen und infrastrukturellen Heterogenitäten. Die Nutzung der Geotagging-Funktion variiert beispielsweise von Land zu Land. Die unterschiedliche Verfügbarkeit von Geotweets in Afrika wurde auch in dieser Dissertation in Studien A1, B1 und B2 bestätigt. Studie A1 zeigte erhebliche Unterschiede in der Anzahl von Geotweets zwischen den verschiedenen UTG. Starke Unterschiede bestehen insbesondere zwischen dem mittleren Afrika, welches relativ wenige Geotweets aufweist, und dem in dieser Hinsicht deutlich

datenreicheren nördlichen Afrika. Studie B2 zeigte für die Fokusregion Westafrika, dass für Nigeria und Ghana ausreichend viele Geotweets, für die benachbarten Länder Togo und Burkina Faso aber nicht ausreichend Geotweets für eine tiefergehende Analyse zur Verfügung standen. Diese im Rahmen der Dissertation gemachten Beobachtungen beziehen sich zwar speziell auf die Plattform Twitter, stehen aber in Einklang mit Ergebnissen anderer Erhebungen bezüglich des Anteils der Internetnutzenden allgemein, welcher erheblich von Land zu Land schwankt (Kraler und Reichel 2022, basierend auf Daten der Internationalen Fernmeldeunion).

Zusätzlich zu diesen bekannten Abdeckungsungleichheiten wurden in dieser Dissertation auch neue Abdeckungsungleichheiten identifiziert, welche speziell für die Erforschung von Migration relevant sind. Wie Studie A1 zeigt, sind in schnell wachsenden Städten Afrikas die neu entstandenen Stadtgebiete tendenziell schlechter durch Geotweets repräsentiert. Die Abdeckung von GSM in Afrika ist daher nicht durchgehend als gut zu betrachten. Die Ergebnisse der Dissertation unterstreichen das Potenzial von GSM-basierten Auswertungen insbesondere für wirtschaftlich starke, etablierte oder stark geplante urbane Regionen.

Textinhalte: Textinhalte wurden sowohl thematisch als auch nach Sprachmerkmalen und Sprachen ausgewertet. Auf sprachlicher Ebene zeigte sich, dass Sprachmodelle English und Französisch gut verarbeiten und auch über Sprachen hinweg sinnvolle Texteinbettungen generieren können. Bei weniger repräsentierten Sprachen bestehen hingegen noch Herausforderungen. Ähnliches gilt für alternative Schreibweisen wie Arabizi, Slang oder *leetspeak*. Einen erfolgreichen Ansatz zur Untersuchung dieser Merkmale demonstrierte Studie B2: Hier konnten lexikalische Sprachmerkmale einer lokalen Form des Englischen in Nigeria algorithmisch erfasst und in Bezug zu Mobilitätscharakteristika der Nutzer gesetzt werden. Für andere Sprachmerkmale ist die algorithmische Ableitung herausfordernder. Die unterrepräsentierten Sprachen und Schreibweisen stellen also eine Limitation der algorithmischen Nutzung von GSM dar, gleichzeitig besteht hier ein besonderes Potenzial für sprachwissenschaftliche Anwendungen. Bezüglich der semantischen Inhalte zeigte sich, dass eine Vielzahl von Themen in GSM algorithmisch erkannt werden konnte, wobei sowohl öffentliche Themen wie Politik als auch persönliche Themen wie Beziehungen und technische Themen wie Technologie und Naturwissenschaften messbar waren. Das Vorkommen dieser Themen könnte als Proxy für Interessen der Nutzenden und damit als Migrationsfaktor interpretiert werden (Studie B1). Jedoch sind nicht alle Themen im gleichen Maße in GSM vertreten. Studie C1 zeigt, dass in den Themen, die mit GSM abgedeckt werden können, erhebliche Unterschiede bestehen, sodass Aspekte des Privatlebens und der persönlichen Beziehungen schlechter erfasst werden können als Diskurs über Weltpolitik oder Unterhaltung. Die Abfrage von Tweets ohne Geotags nach Account-ID, wie in Studien B1 und B2, stellt für diese Herausforderung eine Lösung dar, welche für die Plattform Twitter sinnvoll und umsetzbar ist. Künftige Studien auf anderen GSM-Plattformen könnten in Betracht ziehen, ähnliche Methodiken zu entwickeln, um eine breitere Spanne von Themen zu erfassen.

Mobilitätstypen und Demographien: Die in Studien B1 und B2 beobachteten Mobilitätsmuster zeigen primär Ströme zwischen urbanen Räumen. Bewegungen im ländlichen Raum, Flüchtlinge und Binnenvertriebene wurden durch die Ansätze der Studien nicht abgebildet, was mit den Ergebnissen vorausgegangener Studien in Einklang ist, die auch geringe bis keine Aktivität von Flüchtlingen, die eindeutig als solche identifiziert werden konnten, auf Twitter fanden (Petutschnig et al. 2020). Die Anzahl von Tweets mit hoher Präzision (Kruspe et al. 2021) ist zwar für statische Analysen, wie in Studien A1 und A2

ausreichend, Mobilitätsanalysen erfordern jedoch zeitlich dichtere Timelines. Auch innerstädtische Mobilität ist unter der beobachteten Datensituation vermutlich nicht mit ausreichender räumlicher Präzision zu erfassen, oder nur für eine kleine Gruppe von sehr aktiven Accounts. Daher eignen sich die Daten für die Erfassung von langfristigen interregionalen und internationalen Bewegungen.

Insgesamt zeigen sich für die Orte, an welchen Twitter genutzt wird, große Mengen von Accounts und Texten, welche als Big-Data bezeichnet werden können. Die Menge an Daten in einer Region sollte dabei im Verhältnis zu der dort lebenden Bevölkerung gesetzt werden. Für Nigeria, ein Land mit ca. 237 Millionen Einwohnern¹¹⁰ (CIA 2025c), wurden in Studie B1 ca. 300 Tsd. Accounts identifiziert, wovon für ca. 114 Tsd. Accounts eine Trajektorienanalyse durchgeführt wurde und für 9 301 Nutzende mit ausreichender Verlässlichkeit der Mobilitätstyp eingeschätzt werden konnte. Große Haushaltsbefragungen bestehen im Vergleich dazu typischerweise aus 5 – 30 Tsd. Haushalten. Die Abdeckung von Twitter ist also vergleichsweise nicht klein, doch die damit gesammelten Daten stellen eine Stichprobe dar. Um richtig einzuordnen, welche Gesamtbevölkerung diese Stichprobe repräsentiert, müssen Demographien, Geographien und Social-Media-Nutzung verstanden werden (Ilieva und McPhearson 2018). Auf demographischer Seite ist ein Effekt des Alters (engl. *Age-Gap*) in der Internetnutzung dokumentiert, welcher aber über die letzten Jahre geschrumpft ist (ITU 2024). Disparitäten zwischen Geschlechtern (engl. *Gender Gap*) in der Internetnutzung sind nach den Entwicklungen der letzten Jahre fast, aber nicht vollständig und überall, eliminiert (ITU 2024). Im Internetzugang besteht (vermutlich ökonomisch bedingt) ein erheblicher Stadt-Land Gradient. Dieser ist insbesondere in Afrika und bei Ländern mit niedrigem Pro-Kopf Einkommen ausgeprägt, und scheint sich nach Einschätzung der Internationalen Fernmeldeunion auch nicht zu reduzieren (ITU 2024). Insgesamt jedoch nimmt die Verbreitung von Internetnutzung über Demographien und Geographien hinweg zu und Disparitäten verringern sich. Für die Fähigkeit künftiger Studien, mittels GSM eine breite Bevölkerung abzubilden, bietet dies eine positive Perspektive. Momentan jedoch sind Disparitäten bei wirtschaftlich schwächeren Ländern und insbesondere in Afrika stärker ausgeprägt. Zudem kann man nicht von Internetnutzung auf individuelle Plattformen schließen.

Es lässt sich also nur vermuten, dass die in dieser Dissertation betrachteten Ansätze zum aktuellen Zeitpunkt geeignet sind, um junge, gut gebildete und vorwiegend männliche Nutzer von internationaler Mobilität abzubilden, welche in den Kernzonen städtischer Ballungsgebiete leben. Dies stellt nur eine Momentaufnahme und eine Vereinfachung dar, da mit Sicherheit auch andere Gruppen in den Daten repräsentiert sind. Die demographische Zusammensetzung der abgebildeten Nutzenden ist aufgrund der Dynamik von Social Media und Anonymität der Nutzenden nur schwer für eine Analyse zu erfassen. Um hier eine Nutzung von GSM einschätzen zu können, ist es wünschenswert, zeitnah auch in-situ Umfragen durchzuführen. Hierdurch kann abgeschätzt werden, ob die für die Forschungsfrage relevanten Bevölkerungsgruppen online messbare Spuren hinterlassen, und wenn ja, auf welcher Plattform.

Zur Einschätzung des Datenpotenzials ist nicht nur relevant, ob die Plattformen grundsätzlich genutzt werden, sondern auch wie. Die empirischen Methoden dieser Dissertation beziehen sich auf die Messung der Produktion von Inhalten, primär des Postens von Tweets. Es

¹¹⁰ 236 747 130 im Jahr 2024.

bestehen jedoch auch alternative Möglichkeiten der Internetnutzung wie Sehen, Teilen, Diskutieren, und Bewerten von Inhalten. Tatsächlich beschäftigt sich auf vielen Plattformen nur ein kleiner Teil der Nutzenden mit der Produktion von Inhalten. Der überwiegende Teil der beitragenden Nutzenden befasst sich primär mit der Kommentierung und Bearbeitung von Inhalten, und der weitaus größte Teil der Nutzenden konsumiert und sichtet Inhalte nur, ohne selbst Inhalte beizusteuern (oft als 90-9-1 Prinzip formuliert, siehe Sun et al. (2014)). Individuelle Faktoren wie soziale Bedürfnisse, Technologiekenntnisse und Charaktereigenschaften beeinflussen dabei das Posting-Verhalten online (Sun et al. 2014), weswegen nicht davon ausgegangen werden kann, dass die über die produzierten Inhalte gewonnenen Ansichten auch die der Allgemeinheit auf der Plattform widerspiegeln. Die Exklusion von besonders aktiven Nutzenden, wie beispielsweise in Studie A1 angewendet, kann hier ein statistisches Mittel sein, um den Einfluss dieser Disparitäten zu reduzieren, löst aber nicht das grundsätzliche Bias-Problem.

Um die Ergebnisse einordnen zu können, wäre es hilfreich, Interview-basierte Umfragen in Migrationsstudien durch Fragen zum Online-Verhalten zu ergänzen, welche idealerweise eine Unterscheidung zwischen Lurking und aktivem Posten ermöglichen. Mit diesen Informationen könnte, sofern sie ausreichend repräsentativ sind, bei der Analyse von GSM-Daten zwischen Aspekten der Meinungsbildung, Sichtbarkeit und Wahrnehmung unterschieden werden. In Bezug auf die Nutzung von GSM würde dies theoretisch eine komplexere Auswertung der Ergebnisse erlauben. Beispielsweise könnte die Anzahl an Geotweets in Teilen der Stadt, welche in dieser Dissertation als Proxy für online-Sichtbarkeit eines Ortes interpretiert wird, so auch als Proxy für Gesehenwerden, Aufmerksamkeit und Bedeutung dieses Ortes interpretiert werden.

7.1.3. Verknüpfbarkeit von Fernerkundung und GSM

Die Dissertation zeigt, dass die gemeinsame Auswertung der grundsätzlich verschiedenen Datenquellen FED und GSM im Migrationskontext möglich ist. Die für die Verknüpfung genutzt Methodik muss jedoch ihre verschiedenen Eigenschaften berücksichtigen. Hierfür sind mehrere Ansätze hervorzuheben, welche sich gegenseitig nicht ausschließen.

Ein Ansatz ist die Nutzung von Embeddings, welche Textdaten in einem numerischen Raum überführen, und so mit quantitativen Methoden messbar machen können. Wie durch den für Studie A2 entwickelten Ansatz demonstriert wird, können Embeddings auch in Rasterformate konvertiert werden. Für diese neuartige Kombination von Methoden der NLP und Geostatistik steht prinzipiell eine Vielfalt an Interpolationsmethoden zur Verfügung. Die in Studie A2 genutzte Inverse Distanzwichtung ist nur eine von vielen möglichen Extrapolationsmethoden (Chang 2019). Sie wurde in der Annahme gewählt, dass die Relevanz eines Textinhaltes mit zunehmender Distanz quadratisch abnimmt. Ob diese Annahme der Realität entspricht, und inwiefern andere Extrapolationsmethoden für GSM-Textinhalte geeignet sind, könnte in künftigen Studien untersucht werden.

Eine Alternative zur Überführung von Embeddings ins Rasterformat kann darin bestehen, die Modellierung bereits auf Pixelebene durchzuführen, also Pixel-Embeddings zu generieren. Ein Beispiel ist Alpha Earth von Google, welches neben Satellitendaten auch Texte von Wikipedia miteinbezieht (Google, o.D.), um ein weltumspannendes Embedding auf Pixelebene zu generieren. Da die Fähigkeit multimodaler Modelle, unterschiedliche Inputs zu akzeptieren, sich perspektivisch weiter verbessern wird (Zhang et al. 2024), könnte dies in Zukunft auch auf weitere Datentypen ausgeweitet werden, sofern sie sich mit ausreichender

Genauigkeit georeferenzieren lassen. Damit können diverse räumliche Datenquellen verbunden werden, um regionale Embeddings auf Ebene von Pixeln oder vektorbasierten Raumeinheiten zu erstellen (Li et al. 2023), die dann als Inputs für geographische Modellierung oder explorative Anwendung (Yan et al. 2024) genutzt werden können.

Zwar ist zu vermuten, dass die Fähigkeit von Embeddings und multimodalen Modellen noch zunehmen wird. Dennoch steht zu erwarten, dass sie traditionelle Methoden nicht vollständig ersetzen werden, denn die Komplexität der Modellarchitekturen und die Größe des Datenvolumens stellt ein Hindernis für den Erkenntnisgewinn dar. Selbst in modularen Ansätzen, wie in dem in Studie A2 demonstrierten, welcher das Auslesen und Visualisieren von Zwischenergebnissen ermöglicht, erfolgt doch eine Generalisierung und Verzerrung dadurch, dass Inhalte aus der Form genommen werden, in welcher sie der Autor ursprünglich erschaffen hat. Das schmälert ein grundsätzlich attraktives Potenzial von Social-Media-Daten, welches darin besteht, dass sie von Menschen für Menschen geschrieben und daher interpretierbar sind. Wege, dieses Potenzial zu erhalten, bestehen in Mixed-Method Ansätzen, welche qualitative und quantitative Methoden verbinden (Andreotta et al. 2019; Biewer et al. 2026). Bei dem in Studie A2 demonstrierten Ansatz werden die Texte zwar auch über Embeddings quantifiziert, aber gleichzeitig mittels Topic-Modeling wieder kategorisch interpretierbar gemacht. Zusätzlich könnten die ursprünglichen Texte in diese Visualisierung integriert werden. Eine Möglichkeit, derartige Mixed-Methods in der räumlichen Dimension zu unterstützen, könnte in Dashboard-artigen interaktiven Visualisierungen¹¹¹ bestehen. Diese erlauben es Experten, Social-Media-Inhalte gezielt zu durchsuchen und in Bezug zu anderen Geodaten zu setzen. Fortschritte in Visualisierungs- und Suchmethoden erlauben es, auch große Datensätze interaktiv darzustellen und zu durchsuchen (Weckmüller et al. 2025).

Quantitative und qualitative Methoden schließen sich also nicht aus, sondern können parallel genutzt werden. Dies ist nicht bei allen Anwendungen notwendig und sinnvoll, aber in der geographischen Erforschung von Migration sicher angemessen.

7.2. Validität der Daten

Die Fähigkeit, valide Aussagen aus den Daten zu ziehen, erfordert eine gute Kenntnis der Repräsentativität und Generalisierbarkeit (Campbell 1957). Die demographische Repräsentativität wurde bereits oben aus der Perspektive der Abdeckung behandelt und ist nicht einfach zu erfassen, da verschiedene Bias-Faktoren einander überschneiden. Der Internetzugang von Menschen, ihre Neigung zu bestimmten Plattformen, zu bestimmten Themen und zur Nutzung von Geotagging-Funktionen variieren über Nutzertypen und über Länder hinweg (Gollin et al. 2021; Huang und Carley 2019; Pappalardo et al. 2023). Das macht es schwierig, einzuordnen, welche Bevölkerungsgruppen durch die Accounts repräsentiert werden.

Eine große Herausforderung stellen automatisierte Accounts (oft als „Bots“ bezeichnet) dar. Deren Identifikation ist Ziel zahlreicher Forschungsinitiativen (Subrahmanian et al. 2016; Efthimion et al. 2018; Orabi et al. 2020). Motiviert ist dieses Interesse größtenteils dadurch, dass Bots für schädliche oder kriminelle Zwecke eingesetzt werden können. Die Nutzung von Bots zur Marktmanipulation, der Verbreitung von Desinformation und politischer Einflussnahme wurde vielfach dokumentiert (Ferrara 2023).

Dennoch muss auch berücksichtigt werden, dass Automatisierung auf verschiedenen Ebenen

¹¹¹ Beispielsweise das LBSN-Dashboard für Bonn: <https://geo.rocks/dashboards/bonn/>

und für verschiedene Zwecke eingesetzt werden kann, welche nicht alle rechtmäßig und konstruktiv wird. So existieren beispielsweise Bots, welche Horoskope, Nachrichten, oder Bilder posten, und Software ermöglicht es auch menschlichen Nutzenden, das Veröffentlichen von Inhalten zu einem gewissen Grad zu automatisieren. Es gibt also zahlreiche Mischformen, von welchen nicht alle für jede Forschungsfrage problematisch sind (Orabi et al. 2020).

Da in der Nutzung im Migrationskontext die Informationsgewinnung über Menschen im Fokus steht, sind komplett automatisierte Accounts in der Regel nicht von Interesse. Das Verwechseln von Bots mit Menschen beeinträchtigt die Validität der Daten negativ. Maßnahmen zur Filterung automatisierter Accounts sind in diesem Forschungsfeld daher ein empfehlenswerter und weithin genutzter Vorverarbeitungsschritt vieler Studien (z.B. Hübl et al. 2017; Petutschnig et al. 2020; Lemoine-Rodríguez et al. 2024). Diese Maßnahmen sollten jedoch an die Forschungsfrage angepasst werden, um nicht auch nützliche Daten auszuschließen.

Studien B1 und B2 fokussierten sich beispielsweise auf die Charakterisierung von menschlicher Gruppen nach Mobilität, weswegen Mobilitätscharakteristika (wie Geschwindigkeit) einen essenziellen Bestandteil der Bot-Filterung ausmachten. Als Teil von Studie B1 wurde ein post-hoc Bot-Check durchgeführt, welcher ergab, dass unter den analysierten Accounts nur ein vernachlässigbar kleiner Bot-Verdacht detektiert wurde. Es steht zu vermuten, dass der recht komplexe Accountauswahlalgorithmus und insbesondere das segmentbasierte Migrationsidentifikationsverfahren eine effektive Barriere gegenüber nicht-bösartiger Automatisierung darstellt.

Akteure, welche zum Ziel haben, reale Migration zu simulieren, könnten auch diese Barriere selbstverständlich angreifen. Algorithmische Fortschritte, insbesondere im Bereich generativer KI, führen zu zunehmender Fähigkeit von Bots, komplexe Kommunikations- und Verhaltensmuster zu verstehen und zu imitieren, sodass es zunehmend schwerer wird, Chatbots von echten Menschen zu unterscheiden (Ferrara 2023; Radivojevic et al. 2024).

Zwischen Menschen, die Bots identifizieren wollen und den Entwicklern dieser Bots besteht daher ein regelrechtes Katz und Maus Spiel (Javed et al. 2024), welches auch zur Verbesserung von Bot-Erkennung führt. Aus Sicht der GSM-basierten Forschung könnte die Möglichkeit in Betracht gezogen werden, Studien nach einem gewissen Zeitraum nach Anwendung neuer Bot-Check Methoden zu re-evaluieren. Insbesondere bei Migration, welche in vieler Hinsicht ein politisches Thema ist, kann politische Einflussnahme möglich sein.

Angesichts der Möglichkeit der Manipulation sollte die Validität der Daten daher nur unter Vorbehalt angenommen und kritisch unter Berücksichtigung des gesellschaftlichen Umfelds der Anwendung evaluiert werden.

Aufgrund der eingangs beschriebenen Datenlücken konnten in in dieser Dissertation untersuchten Kontext afrikanischer Migration nur begrenzte Validierungen der Daten und der Ergebnisse vorgenommen werden, was als eine klare Limitation der Dissertation angesehen werden muss. Die implementierten Robustheits- und Plausibilitätschecks sind zwar ausführlich, können aber letztendlich eine tatsächliche Validierung nicht ersetzen. Die Entwicklung von annotierten Datensätzen zur Validierung ist herausfordernd und zeitintensiv, sollte im Zuge des Wachstums des Forschungsfeldes aber dennoch einen wichtigen Fokus darstellen, um die Validität, Kreditibilität und Transparenz der Forschung zu verbessern (siehe Olteanu et al. 2019).

7.3. Zugang zu Daten

Die Abdeckung der Daten ist eine wichtige Komponente der Praktikabilität und wurde bereits in Sektion 7.1.2 diskutiert. Es ist aber nicht nur wichtig, wo Daten vorhanden sind, sondern auch, ob sie für Forschende zugänglich sind.

Die jüngsten Entwicklungen auf Twitter, insbesondere nach der Übernahme durch Elon Musk (Bisbee und Munger 2025), haben gezeigt, dass sich die Verfügbarkeit von Datenquellen für wissenschaftliche Forschung schnell und unerwartet ändern kann. Die Studien dieser Dissertation wurden durch das zeitweise von Twitter angebotenen *Academic Research Track*-Programm ermöglicht. Dieses wurde im Januar 2021 geöffnet und im März 2023 geschlossen (Davidson et al. 2023). Während davor ein kostenloser, obschon weniger nutzerfreundlicher Stream verfügbar war, bestand danach nur noch ein eingeschränkter Datenzugang mit erheblichen Kosten, die für viele Wissenschaftler nicht tragbar sein dürften. Smith et al. (2025) finden entsprechend einen starken Abfall an Tweet-basierten wissenschaftlichen Publikationen im Zuge der Änderungen im API-Zugang.

Twitter ist nicht die einzige Social Media Plattform, welche ihren Datenzugang in den letzten Jahren eingeschränkt hat. Auch Reddit schloss 2023 den kostenlosen API-Zugang und im Jahr darauf beendete auch Meta das vergleichbare Programm *CrowdTangle* für Facebook und Instagram. Die Social-Media basierte Forschung befindet sich daher in einer Art Post-API Zeitalter (Mimizuka et al. 2025) mit einem merklichen Trend zur Kommerzialisierung des Datenzugangs.

Diese Kommerzialisierung ist wirtschaftlich nachvollziehbar, aber aus Sicht der Forschung problematisch, da sie finanzielle Disparitäten zwischen Forschungseinrichtungen verstärkt und die Reproduzierbarkeit von Studien beeinträchtigt.

Die Erschließung alternativer Datenquellen ist daher entscheidend für das Fortbestehen des Forschungsfeldes. Selbstverständlich besteht die Möglichkeit, Daten anderer Plattformen zu nutzen. Neben Twitter existieren beispielsweise dessen direkte Wettbewerber *Threads*, *Mastodon* und *Bluesky*, welche seit dem Rebranding Twitters große Zuwächse der Nutzeranzahl verzeichneten (Smith et al. 2025). Auch diese Plattformen sind für die Forschung nutzbar, auch wenn sie sich in der Struktur der Plattformen und Daten merklich unterscheiden. So bietet beispielsweise Bluesky die Möglichkeit des Erstellens von *feeds*, und es können aus den Daten Konversationsbäume rekonstruiert werden. Die Plattform bietet jedoch keine Geotag-Funktion. Als alternative Möglichkeit zur Verortung der Inhalte für räumliche Studien könnte hier Geoparsing genutzt werden. Wie Studie C2 zeigt, ist zwar nicht damit zu rechnen, dass dies zu vergleichbaren Inhalten führt, doch könnte das Volumen der nutzbaren Daten mit über 12% (siehe Kapitel 6.6.1) für viele Themen deutlich größer sein als die 0,5-2% der Tweets, die Geotags beinhalten (Kruspe et al. 2021; Pfeffer et al. 2023).

Die Vielfalt der Plattformen, ihrer Demographien und Formate könnte perspektivisch eine große Herausforderung für vergleichende Studien darstellen, denn das online-Verhalten ist zu einem Teil auch durch die Online Community geprägt (Sun et al. 2014). Andererseits könnte diese Vielfalt zu einer größeren Breite an Perspektiven in der Forschung führen, was insbesondere angesichts des großen Fokus auf Twitter in der Vergangenheit (Stock 2018), als positiv zu bewerten ist (Olteanu et al. 2019).

Für Plattformen mit unzureichendem oder gänzlich ohne API-Zugang könnten alternative Ansätze der Datensammlung genutzt werden. Die Datenschutz-Grundverordnung der

Europäischen Union ermöglicht Ansätze der Datenspende (engl. *data donation*) bei welcher Nutzende freiwillig (ggf. gegen finanzielle Anreize) ihre persönlichen Daten bei den Plattformbetreibern abfragen und Forschenden für Zwecke von Studien zur Verfügung stellen. Auch wenn hierbei nur relativ kleine Stichproben von Nutzenden erreicht werden können, zeichnet sich dieser Ansatz durch die Fülle an Metadaten aus, welche gewonnen werden können, und stellt durch die Freiwilligkeit und das explizite Einverständnis der Menschen eine ethische Verbesserung dar (Boeschoten et al. 2022).

Zusätzlich könnte das Gesetz über digitale Dienste (engl. Digital Services Act) der Europäischen Union zu einer grundsätzlichen Änderung der Beziehung zwischen Plattform und Datennutzenden führen, was von Mimizuka et al. (2025) als „Post-Post API Age“ betitelt wird. Als Teil dieses Gesetzes sollen Plattformbetreiber verpflichtet werden, Forschenden für bestimmte Zwecke (wie systemische Risiken) Zugang zu den Daten und Informationen über Algorithmen zu ermöglichen (Rudl 2025; Goanta et al. 2025). Dieses Framework ist jedoch praktisch noch nicht komplett umgesetzt (Mimizuka et al. 2025), zudem soll nach ersten Berichten eine komplizierte Antragstellung und das unklar definierte Konzept der abgedeckten Forschungsfragen den Zugang erschweren (Goanta et al. 2025). Eine Umfrage von Mimizuka et al. (2025) unter Forschenden zeigte deren Frustration mit dem Prozess der Antragstellung und Einschränkungen und Intransparenzen im Datenzugang. Web-Scraping, wie es in dieser Dissertation für Nairaland angewendet wurde, ist aus technischer Sicht für manche Plattformen umzusetzen, aus ethischer und rechtlicher Sicht für viele Plattformen jedoch problematisch, bei welchen algorithmische Abfragen nicht erwünscht sind oder gegen die Nutzungsbeschränkungen verstoßen (Mimizuka et al. 2025). Die Rechtslage ist also in vielen Bereichen unklar und nicht ausgearbeitet. Daher kann das Potenzial aktueller Entwicklungen hier nur vorsichtig angedeutet werden.

7.4. Ethische Überlegungen hinsichtlich der Datennutzung

Die Nutzung von FED und GSM-Daten bringt auch ethische Herausforderungen mit sich, welche bei der Forschung über personenbezogene Phänomene wie der Migration nicht vernachlässigt werden dürfen.

Social-Media-Daten sind Daten, welche von Menschen und über Menschen generiert wurden. Daher ergibt sich für diese Daten die besondere Notwendigkeit zu Berücksichtigung der Privatsphäre von Individuen, ihrer Anonymität, Freiwilligkeit (engl. *consent*), sowie Abschätzung von möglichen negativen Konsequenzen der Forschung (Townsend and Wallace 2016; Kochupillai et al. 2022; Kochupillai and Taubenböck 2023). Im Folgenden sollen ethische Fragestellungen nicht allgemein, sondern mit einem Fokus auf Herausforderungen der in dieser Dissertation genutzten Daten und Methoden diskutiert werden.

Herausforderungen bestehen insbesondere dann, wenn ein Zielkonflikt zwischen ethischen und wissenschaftlichen Anforderungen vorliegt. Die Notwendigkeit, die Forschung nach ethischen Standards durchzuführen kann in Konflikt mit dem wissenschaftlichen Bedarf nach Information stehen. So wird, wie bereits oben angesprochen, die korrekte Interpretation der Repräsentativität sehr durch möglichst detaillierte Informationen über die Autoren der Texte gefordert. Die Ableitung dieser steht jedoch im Zielkonflikt mit dem Schutz der Privatsphäre derselben Autoren. Im geographischen Kontext ist hier speziell die *Geoprivacy* schützenswert, welche als das Recht von Individuen definiert werden kann, die Preisgabe ihrer Aufenthaltsorte und Bewegungen zu verhindern (Kwan et al. 2004). Der Schutz der *Geoprivacy* muss gegen die Genauigkeit und Aussagekraft der Analyse abgewogen werden

(Kwan et al. 2004). Es ist unwahrscheinlich, dass hierbei allgemeingültige Richtlinien etabliert werden können, denn auch das geographische und sozioökonomische Umfeld der Datenanalyse muss berücksichtigt werden. Insbesondere Daten über Migrierte und Flüchtlinge können als sensibel betrachtet werden, da diese in vielen Situationen besonders vulnerable Gruppen darstellen (Sirbu et al. 2021) und besonderen Schutzes bedürfen. Dies gilt nicht nur in Bezug auf Geoprivacy, sondern auch für die Berücksichtigung möglicher negativer Konsequenzen.

Denn nur weil Daten (von beispielsweise Twitter) öffentlich sind, ist nicht jede Nutzung automatisch ethisch und rechtlich unbedenklich (Lomborg und Bechmann 2014). Für die in dieser Dissertation genutzten Methoden der Fernerkundung und der Sprachprozessierung gilt dies möglicherweise in besonderem Maße, da datentechnische Entwicklungen in hohem Tempo voranschreiten und Auswertungsmöglichkeiten erschließen, welche den Autoren zum Zeitpunkt der Veröffentlichung ihrer Beiträge nicht erwartet wurden.

Für Studien kleinerer und aktueller Datensätze besteht eine mögliche Lösung darin, die Zustimmung der Autoren zur Nutzung ihrer Beiträge zu erfragen. Für Studien großer Datensätze, wie in dieser Dissertation durchgeführt, ist dies aufgrund der großen Anzahl an Autoren nicht praktikabel.

Nichtsdestotrotz sind bei der Nutzung öffentlicher Plattformen, wie sie im Rahmen dieser Studie erfolgte, Aspekte von Privatsphärenschutz möglicherweise weniger bedenklich. Dies ist letztendlich auch vom kulturellen Hintergrund der Social-Media Nutzenden und vom Zweck und Methodik der Datenanalyse abhängig. Eine Studie von Hemphill et al. (2022) zeigte, dass Social Media Nutzende die Nutzung ihrer Daten durch Forschende¹¹² deutlich eher als akzeptabel bewerteten, als die Nutzung durch Journalisten oder Unternehmen. Die Bereitschaft der Bevölkerung, private Geoinformationen für Forschungszwecke bereitzustellen, variiert von Kultur zu Kultur. So finden Kim und Kwan (2021) höhere Zustimmungen zum Teilen der Geoinformationen für wissenschaftliche Zwecke in Südkorea als in den Vereinigten Staaten; zudem finden sie einen starken Einfluss der wahrgenommenen Sensibilität der Daten auf die Zustimmung. Bei wenig sensiblen Daten und Analysen auf einem hohen Grad der Aggregation¹¹³ ist der Schutz von Privacy weniger herausfordernd, da allein aufgrund der Methodik bereits ein hoher Anonymisierungsgrad erreicht wird (Lomborg und Bechmann 2014). Die Aggregation von Texten kann also eine technische Lösung darstellen, Anonymität und Geoprivacy zu schützen (Kounadi und Resch 2018) und damit die Nachteile fehlender Einwilligung zu verringern (Williams et al. 2017). In dieser Dissertation wurden Gruppen nur in einem hohen Aggregationsgrad präsentiert. Daher wurde in den Arbeiten dieser Dissertation die Anonymität oder Privatsphäre irgendeiner Person nicht gefährdet. Bei qualitativen Studien, welche auf dieser Methodik aufbauen, ist Aggregation keine sinnvolle Lösung, da Texte und persönliche Informationen essenzielle Komponenten der Analyse sind, und sich nicht ohne Informationsverlust aggregieren lassen. Daher sollte bei der Präsentation qualitativer Ergebnisse und Einschätzungen auf der Ebene von Individuen

¹¹² Eine frühere Studie von Williams et al. (2017) zeigte ferner eine deutlich höhere Akzeptanz von Forschung in universitären Umgebungen als durch Regierungen oder für kommerzielle Zwecke durchgeführte Forschung.

¹¹³ In dieser Theses wurden Ergebnisse nur in Aggregation auf Raumeinheiten und Account - Gruppen präsentiert und interpretiert. Wo dies nicht möglich war, wurden die Accounts von der Analyse ausgeschlossen.

ein größeres Augenmerk auf möglichen Konsequenzen der Veröffentlichung für die Textautoren liegen. Dies gilt sowohl für GSM-Daten als auch für hochaufgelöste FED und andere Geodaten wie beispielsweise aus kartenbasierten Bewertungsportalen¹¹⁴. In dieser Dissertation dargestellte Beispiele von Textinhalten wurden, wie von Williams et al. (2017) vorgeschlagen, rekonfiguriert¹¹⁵, um sowohl die Rechte der Autoren und die Twitter Nutzungsbedingungen zu respektieren. Dies ist ausreichend für die illustrative Nutzung innerhalb dieser geographischen Dissertation, für sprachwissenschaftliche Analysen sollten jedoch andere Lösungen gefunden werden. Auch sollte, falls technische Entwicklungen in Zukunft deutlich feinere räumliche Auflösungen erlauben, Geoprivacy einen integralen Teil des Forschungskonzeptes darstellen (siehe Kounadi und Resch 2018).

Ein weiterer Zielkonflikt besteht zwischen dem wissenschaftlichen Grundprinzip der Reproduzierbarkeit und dem Recht der Autoren, über die Sichtbarkeit ihrer Inhalte zu entscheiden, diese gegebenenfalls zu löschen und die Sichtbarkeit ihrer personenbezogenen Informationen auf Suchmaschinen zu verhindern. Während Letzteres in der EU als *right-to-be-forgotten* rechtlich implementiert ist (Bertram et al. 2019), ist auch Ersteres auf vielen Plattformen Teil der Funktionalität (Williams et al. 2017; Kaleel et al. 2025). Hinzu kommt auf moderierten Plattformen die Löschung von Inhalten, welche gegen Regeln oder Richtlinien der Plattform entsprechen. Die Nutzung dieser Funktionen führt dazu, dass Inhalte im Laufe der Zeit aus den Datenbanken verschwinden oder nicht mehr zugänglich sind. Dieser als *data decay* bezeichnete Verlust von Daten kann erheblich sein, so finden Pfeffer et al. (2023) in einer Analyse auf Twitter, dass nach einem Jahr nur noch zwei Drittel der Tweets verfügbar sind, und nach etwa fünf Jahren nur noch die Hälfte. Zubiaga (2018) identifiziert für eine Auswahl von 30 in wissenschaftlichen Studien zwischen 2012 und 2016 genutzten Datensätzen einen möglichen Verlust von mehr als 30% der Daten über vier Jahre. Die Studie von Zubiaga schätzt die verbleibenden Textinhalte als weiterhin repräsentativ für die ursprünglichen Datensätze ein, Metadaten wie Nutzerinformationen jedoch nicht. Diese Veränderlichkeit der Daten trägt zur Replikationskrise bei. Einerseits werden Forschende im Rahmen der Zugangsvereinbarung für die Datennutzung häufig von Plattformen verpflichtet, ihre Daten regelmäßig zu aktualisieren und gelöschte Inhalte aus den Analysedaten zu entfernen (Küpfer 2024). Hierfür stellen Plattformen teilweise auch spezielle Werkzeuge bereit¹¹⁶. Die Bereitstellung einer solchen Funktion ist zur Vereinfachung grundsätzlich begrüßenswert. Doch wird durch die Aktualisierung der Daten unweigerlich auch das Ergebnis der darauf durchgeführten Analysen beeinflusst, wodurch Reproduktionen der Studie selbst mit demselben Ansatz zu anderen Ergebnissen führen. Es ist damit zu rechnen, dass ein weitreichendes Replikationsproblem bei Studien mit Social Media besteht (Küpfer 2024). Bei längeren Analysen im Zeitverlauf, wie es Migrationsstudien wie die in dieser Dissertation oft sind, könnte die geringere Verfügbarkeit von älteren Tweets möglicherweise einen zusätzlichen Bias darstellen. Insbesondere bei kontroversen Themen (Küpfer 2024), welche besonders stark von Moderation und Redaktion betroffen sein könnten (Kaleel et al. 2025; Zubiaga 2018), kann der Verlust von Daten zu Verzerrungen führen. Küpfer (2024) warnt

¹¹⁴ Ein Beispiel wären Bewertungen auf Google Maps

¹¹⁵ Neuformulierung mit Nutzung von Synonymen um Aussage und Ton des Textes mit anderen Worten wiederzugeben (Williams et al., 2017)

¹¹⁶ So stellte Twitter eine Zeit lang einen *batch compliance API*-Zugang bereit (Twitter Inc. 2022a), welcher im Rahmen dieser Thesis genutzt wurde, bis er mit der Umstrukturierung der Plattform im Jahr 2023 nicht mehr verfügbar war.

davor, dass die Löschung von Inhalten in den meisten Fällen nicht durch Nutzende, sondern durch die (Betreiber der) Plattform initiiert ist, welche damit erheblichen Einfluss auf Ergebnisse ausüben kann.

Nachdem die bislang betrachteten ethischen Herausforderungen vorwiegend epistemischen Fragen betreffen, führt bei der Betrachtung ethischer Aspekte letztendlich der Bogen zurück zu den Potenzialen, deren Ergründung das Ziel dieser Dissertation darstellt. Alle diese Potenziale ermöglichen Forschenden neue Wege zur Informationsgewinnung über Migration und Migrierte, und deren Prioritäten, Motivationen und Handlungen. Diese Information befähigt Menschen in Machtpositionen, auf Migration Einfluss zu nehmen. Auf die ethische und moralische Umsetzung dieses Einflusses haben Forschende jedoch weitaus geringeren Einfluss. Umso wichtiger ist die Berücksichtigung von Ethik und Datenschutz während der Durchführung der wissenschaftlichen Datenanalysen.

Bezüglich dieser Dissertation ist meine Einschätzung, dass aus den Informationen, welche in dieser Veröffentlichung enthalten sind, keine Gefahr für Individuen hervorgeht. Selbstverständlich besteht die Möglichkeit, dass sich aus dem hier präsentierten Wissen negative Konsequenzen für Menschen ergeben könnten, welche von mir nicht gewollt und nicht erwartet werden. Doch ist es meine Überzeugung, dass es trotz dieser Gefahr nicht zu bevorzugen ist, sich in der Betrachtung von Migration auf Nicht-Wissen und Intuition zu beschränken. Denn Migration ist ein sehr facettenreiches Phänomen des menschlichen Daseins, zu welchem zwar viele Menschen einen persönlichen Bezug haben, welches tagtäglich politisch diskutiert wird, dessen volle Komplexität jedoch durch keine einzelne Person erfahren werden kann. Migration ist ein prägendes, aber schwer zu fassendes Element unserer Welt. Die in dieser Dissertation entwickelten Ansätze sollen durch empirische geographische Forschung zu einem akkuraterem Bild von Migration in unserer Welt beitragen, anhand dessen wir als menschliche Gesellschaft unser Verhalten diskutieren und planen können. Die Umsetzung geographischen Wissens in politische Handlungen ist Aufgabe aller Gesellschaftsmitglieder, und dabei nehme ich als Forscher keine privilegierte Stellung in Anspruch. Doch ich bin überzeugt, dass das aus Forschung resultierende Wissen die Grundlage für gute Entscheidungen bildet, und hoffe, dass diese Dissertation einen kleinen Teil hierzu beitragen kann.

8. Synthese

Mit dem höheren Ziel, geographisches Wissen für ein gesellschaftlich hochrelevantes Phänomen zu generieren, wurde durch diese Dissertation angestrebt, in der gemeinsamen Nutzung von Daten aus Fernerkundung und georeferenzierten sozialen Medien neue Möglichkeiten für die geographische Erforschung von Migration auszuloten.

Zu diesem Zweck wurden sechs Studien durchgeführt, von welchen sich jeweils zwei auf die Anwendung zum Verständnis städtischen Siedlungsraumes, der Erfassung von Mobilitätscharakteristika und der Untersuchung thematischer Verzerrungen in georeferenzierten Textdaten fokussierten. Über alle Studien hinweg konnten technische Potenziale identifiziert, die praktische Anwendung erprobt, die Datenverfügbarkeit- und Qualität getestet und Limitationen und Herausforderungen herausgearbeitet werden.

In diesem Kapitel sollen die Studien zunächst einzeln (8.1) und dann die Arbeit als Ganzes (8.2) zusammengefasst werden, um letztendlich eine Einschätzung und einen Ausblick zu ermöglichen (8.3).

8.1. Fazit der Studien

In der ersten Studie wurde folgende Leitfrage adressiert:

Leitfrage LA1: *Lässt sich durch die Verschneidung von FED und GSM eine Aussage über Muster digitaler Ungleichheiten in Afrikas expandierenden städtischen Räume treffen?*

In der Mehrheit der untersuchten 135 UTG waren ausreichend Geotweets vorhanden, um eine Verschneidung mit fernerkundlich kartiertem Bebauungsalter vorzunehmen. Dabei wurde in mehr als der Hälfte der Fälle beobachtet, dass die Dichte von Geotweets in einer negativen Beziehung zum Alter der bebauten Umgebung steht.

Dieser statistische Trend kann als eine statistische Ungleichheit betrachtet werden, welche nicht notwendigerweise einer sozialen Ungleichheit oder Ungerechtigkeit entspricht. Für eine tiefergehende Interpretation aus humanwissenschaftlichen oder ethischen Perspektiven muss noch analysiert werden, was durch die gemessene Geotweet-Dichte repräsentiert wird und welche Kausalitäten die Unterschiede bedingen. Im Rahmen der Studie durchgeführte Experimente mittels fernerkundlich kartierten Proxy-Variablen wie Intensität der Geplantheit und räumlichen Metriken lassen vermuten, dass sozioökonomische Beziehungen mit politischer Planung, Infrastrukturbereitstellung und Formalität des Siedlungsbaus vorliegen. Zur Etablierung kausaler Beziehungen sind weitere Untersuchungen mit sozioökonomischen Daten und Mixed-Method Ansätzen sinnvoll.

Die Verschneidung von FED und GSM im sich entwickelnden städtischen Raum Afrikas kann also zu Erkenntnissen über die Interaktion von physischer und digitaler Expansion führen. Großes Potenzial besteht in der Integration verschiedener Datentypen und von Methoden aus verschiedenen Wissenschaftsbereichen. Die folgende Studie befasste sich damit, durch die Unterteilung des urbanen Raumes nach Charakteristika aus FED und aus GSM-Texten eine mögliche Grundlage für die explorative Analyse und Mixed-Method Ansätze im Migrationskontext zu schaffen. Die Leitfrage dieser Studie war:

Leitfrage LA2: *Ermöglicht die Umwandlung von Textinhalten in ein Rasterformat und geographisch-textinhaltliches Clustering die Generierung von interpretierbaren urbanen Raumeinheiten für die Analyse mittels FED und GSM?*

In dieser Studie wurde ein bestehender Ansatz der räumlichen Clusterung auf die Nutzung mehrerer Daten erweitert. Die neuartige Umwandlung von Text-Embeddings in ein Rasterformat erlaubte es, Textinhalte und physische Informationen über die städtische Bebauung in einem gemeinsamen Datenformat zu vereinen und in einem Clustering-Ansatz zu nutzen. Die sich ergebenden Cluster konnten in der qualitativen Evaluation als interpretierbar und nachvollziehbar bewertet werden. Die Verteilungen von Text-Themen, Sprachen und Bebauungsalter waren zwischen den Clustern klar differenzierbar. Demzufolge ist der Ansatz, gerasterte Embeddings und Fernerkundungsdaten in räumlichem Clustern zu benutzen, eine Möglichkeit zur unüberwachten Strukturierung urbanen Raumes und kann eine sinnvolle Grundlage für Mixed-Method Studien sein. Der Ansatz eignet sich aufgrund seiner Flexibilität weniger für Analysen und Modellierungen, sondern vielmehr für explorative Zwecke, Mixed-Method Ansätze, Integration von Stakeholdern, Kommunikation und Visualisierung. Im Rahmen von Migrationsstudien kann somit der wachsende urbane Raum, welcher als Ziel und Umwelt von der meisten Migrierten eine hohe Relevanz hat, besser verstanden werden. Hierbei können Textinhalte eine große Rolle spielen.

Jedoch ist umgekehrt die statische Verortung von Textinhalten in (neuen) urbanen Räumen

nur bedingt und mittelbar aussagekräftig über die Mobilität der dort lebenden Menschen. Eine direkte Zuordnung von Textinhalten zur Mobilität von Migrierten kann hier möglicherweise unmittelbare Einblicke ermöglichen. Vor diesem Hintergrund beschäftigte sich Studie B1 mit der Frage, inwiefern Analysen von Textinhalten und von Mobilität auf Ebene von Nutzenden zusammengeführt werden können:

Leitfrage LB1: *Können thematische Präferenzen von Twitter-Nutzenden aus deren Texten abgeleitet und in Verbindung zu Mobilitätscharakteristika der Nutzenden gesetzt werden?*

Sowohl zur Extraktion von Informationen aus Texten als auch zur Ableitung von Mobilitätsinformationen aus digitalen Spuren existierten bereits verschiedene etablierte Methoden. Diese wurden in Studie B1 auf bislang neuartige Weise zusammengeführt. Hierdurch konnten die thematischen Schwerpunkte von Nutzenden auf Twitter nach der Mobilität der Nutzenden aufgelöst werden.

Die Kombination erwies sich als erfolgreich. Mehrere tausend Accounts konnten einem Mobilitätstyp zugeordnet werden. Auch die Mobilitätsinformationen, welche hierbei als Zwischenprodukt anfallen, können als Migrationsströme einen Mehrwert für Geographie und Migrationsforschung bieten. Limitierend wirkt primär die begrenzte Demographie der einzelnen Plattform Twitter. Der Ansatz ist hauptsächlich für sehr aktive Nutzende der Geotagging-Funktion erfolgreich, und eignet sich weniger für die Erforschung von Flucht und ländlicher Migration als von urbaner und internationaler Migration.

Mittels Domänenadaptionstechniken konnte eine Themenstruktur aus dem nigerianischen Webforum Nairaland abgeleitet und auf Tweets übertragen werden. Trotz geringer Genauigkeit des Klassifizierungsmodells kann der Transfer für viele Themen als ausreichend erfolgreich bewertet werden, es ergeben sich messbare Unterschiede nach Mobilitätstypen. Themen, welche bei international Migrierten einen höheren Anteil der Tweets ausmachten, waren Unterhaltung, Sport und Inhalte mit Verbindungen zur islamischen Religion und arabischen Sprache. Bemerkenswert war, dass für Texte über Reisen und Mobilität keine vergleichbaren Unterschiede der Mobilität erfasst wurden. Die qualitative Analyse einer Stichprobe legte nahe, dass dies auf die erhebliche Komplexität und Vielfalt dieser (und anderer) Themen zurückgeführt werden könnte.

Angesichts dieser Komplexität können kategorische Einteilungen von Mobilität als hilfreicher Ansatzpunkt für Forschung genutzt werden, ihre Inhalte können und sollten aber mit qualitativen und sprachwissenschaftlichen Methoden herausgearbeitet werden, um tiefergehende Interpretationen der beobachteten Muster zu unterstützen.

Neben dem objektiven thematischen Inhalt des Geschriebenen können auch andere Aspekte der Kommunikation von Interesse für die Migrationsforschung sein, beispielsweise die Sprache, in welcher der Text verfasst wurde. Eine Analyse der Texte in Bezug auf Sprache ist nicht nur zur Einordnung der thematischen Inhalte wichtig, sondern kann zu einem gewissen Grad selbst wichtige kontextuelle Informationen über Migration generieren. Aufgrund der wechselseitigen Beziehung von Sprache und Migration testet Studie B2 die Verbindung der algorithmisch aus GSM abgeleiteten Mobilitätsinformationen mit algorithmisch abgeleiteten Informationen über die Nutzung lokaler Sprachmerkmale:

Leitfrage LB2: *Kann die Verwendung von charakteristischen sprachlichen Merkmalen aus Twitter-Texten abgeleitet und in Verbindung zu Mobilität und Ort der Nutzenden gesetzt werden?*

Ein lexikalischer Ansatz, welcher lokale Begriffe aus Web-Wörterbüchern ableitete und sie in einem kuratierten aber letztendlich algorithmischen Ansatz in Tweets detektierte, erlaubte die Erfassung von Merkmalen einer lokalen Form des Englischen in Nigeria innerhalb von Tweets westafrikanischer Twitter-Nutzender. Über die in Studie B1 entwickelte Mobilitätsanalyse wurde für diese Nutzenden auch Mobilitätsinformationen abgeleitet. Die Vereinigung von räumlichen Informationen und Informationen über die Sprachverwendung erlaubte es, die Verbreitung lokaler Sprachmerkmale im Raum und ihre Beziehung zur Mobilität der Nutzenden messbar zu machen.

Der lexikalische Ansatz war erfolgreich zur Schätzung der Nutzung lokaler Sprachmerkmale. Auch wenn diese nur vergleichsweise selten auftraten, zeigte ihre Verbreitung messbare und nachvollziehbare Unterschiede nach Lokalitäten und Mobilitäten. Dabei zeigte eine post-hoc durchgeführte Analyse, dass lexikalische Merkmale nicht notwendigerweise mit pragmatischen und morphologischen Sprachmerkmalen korreliert sind. Da die algorithmische Erkennung der letzteren eine größere Herausforderung darstellt, kann davon ausgegangen werden, dass manuelle Kodierung weiterhin ein wichtiges Werkzeug in diesem wissenschaftlichen Feld bleibt.

Schlussfolgernd kann also gesagt werden, dass nicht nur die thematischen Inhalte von GSM-Texten algorithmisch herausgearbeitet werden können, sondern auch, **wie** Inhalte kommuniziert werden, was für die Analyse von Texten einen großen Mehrwert darstellen kann. Die Identifikation von Diskursaspekten¹¹⁷ kann mittels sprachwissenschaftlicher Methoden erfolgen und unterstützt sowohl die korrekte Interpretation der Textinhalte als auch die Erforschung von Sprache und Mobilität. Die Studie demonstriert, dass verschiedene Auswertemöglichkeiten von Textinhalten sowohl miteinander als auch mit räumlichen Charakteristika wie Ort und Mobilität verknüpft werden können.

Bei der räumlichen Nutzung von georeferenzierten Textinhalten ist zu berücksichtigen, dass verschiedene Themen unterschiedlich häufig räumliche Informationen beinhalten könnten. In Studie C1 wurde dieses Phänomen zum ersten Mal systematisch empirisch untersucht.

***Leitfrage Lc1:** Hat das Thema eines Web-Textes einen Einfluss auf dessen Wahrscheinlichkeit, eine geographisch verortbare Ortsreferenz zu enthalten, und wie unterscheidet sich dies nach Textsorte?*

Für mehrere Millionen Texte aus verschiedenen Webdatensätzen wurden sowohl die Präsenz von verortbaren Ortserwähnungen als auch deren Themen algorithmisch detektiert.

Es zeigt sich, dass die Präsenz von verortbaren Ortserwähnungen tatsächlich sehr stark mit den Themen der Texte variiert. Texte zu Weltpolitik und Reisen enthalten deutlich öfter mindestens eine Ortserwähnung, für welche sich mindestens ein Ort auf der Erde als Referenz identifizieren lässt.

In einem GLMM-Modell konnte dieser Effekt durch die Kontrolle von Effekten der Verfasser, Zeit und Textlänge auf die Themen selbst zurückgeführt werden. Der dabei quantifizierte Logit kann als Stärke einer themenspezifischen Geospatiality interpretiert werden. Inwiefern diese thematische Geospatiality eine universelle Eigenschaft von Themen darstellt, kann an dieser Stelle nur vermutet werden, denn die Themen sind nur begrenzt zwischen Datenquellen vergleichbar, und die analysierten Daten stellen keinen repräsentativen Auszug der Webtextdaten allgemein dar. Innerhalb dieser Daten, welche mehrere verschiedene

¹¹⁷ Wie Sentiment, Absicht oder Sprachregister

Textsorten abdecken, zeigte sich jedoch eine bemerkenswerte und signifikante, wenn auch moderat ausgeprägte, Übereinstimmung im Ranking der Themen.

Somit wurde zum ersten Mal empirisch ein Effekt nachgewiesen, welcher einerseits als möglicher Verzerrungsfaktor bei der räumlichen Nutzung von Textdaten berücksichtigt werden muss, andererseits ein Fokuskonzept für künftige Forschung zur Konzeption von Raum in Textdaten darstellen kann. Zudem kann geschlussfolgert werden, dass nicht alle Themen gleichsam in GSM verfügbar sind. Forschende, welche die Nutzung von GSM erwägen, sollten die Verfügbarkeit der für die Forschungsfrage relevanten Themen in dem jeweiligen Kontext prüfen.

Diese Studie nutzte mit Geoparsing einen einzelnen Ansatz zur geographischen Verortung von Texten. Es ist damit nicht geklärt, ob bei der Nutzung von anderen Arten der Georeferenzen vergleichbare Effekte zu erwarten sind. Um Unterschiede zwischen unterschiedlichen Ansätzen greifbar zu machen, wurde in Studie C2 ein Vergleich von Geoparsing- und Geotagging-basierten Ansätzen vorgenommen.

***Leitfrage Lc2:** Unterscheiden sich Geoparsing- und Geotagging-basierte Ansätze zur Verortung von Texten hinsichtlich der themenspezifischen Einflüsse auf die Verortbarkeit?*

Dieser Vergleich konnte mit aus Studie C1 übernommenen Twitter-Daten und Methoden in Studie C2 durchgeführt werden. Für einen Datensatz von Tweets nigerianischer Twitter-Nutzenden zeigte sich, dass die Geoparsing- und Geotag-basierte Ansätze zu unterschiedlichen Themenverteilungen führen.

Zudem wurde der Effekt von Themen bei der Nutzung von Geoparsing als deutlich stärker gemessen. Es kann daher nicht davon ausgegangen werden, dass diese beiden Ansätze zur Georeferenz von Texten einander substituieren. Ferner kann vermutet werden, dass Studien, welche Geotweets nutzten, nicht mittels Geoparsing repliziert werden können. Stattdessen bilden die Methoden vermutlich ähnliche, aber nicht identische Perspektiven von Raum ab, was sich in den hier gemessenen Unterschieden in Verteilung und Verfügbarkeit von Themen äußert.

Alle diese Studien wurden für einen Ausschnitt von Methoden und Daten durchgeführt, welche im praktischen State-of-the-Art verortet werden können, jedoch nur einen kleinen und nicht repräsentativen Ausschnitt der Vielfalt an Methoden und Datenquellen bilden. Sowohl im Bereich der Fernerkundungsdaten als auch in Webtextdaten existieren zahlreiche weitere Ansätze und es ist zu erwarten, dass perspektivisch weiterhin Ansätze entwickelt und verbessert werden.

Die Studien dieser Dissertation zielten darauf ab, eine Auslotung von Potenzialen in einem bestimmten Anwendungsgebiet der geographischen Forschung zu ermöglichen.

8.2. Schlussfolgerung

Insgesamt zeigen die Ergebnisse der Dissertation, dass in der Verbindung von GSM und FED eine erhebliche Vielfalt an modular kombinierbaren Methoden und damit ein bemerkenswertes Potenzial für die geographische Erforschung von Migration besteht. Die Fernerkundung kann die physische und objektive Sicht auf die Umwelt erfassen, während über GSM soziale Faktoren und subjektive Wahrnehmungen der Umwelt beobachtet und messbar gemacht werden können.

In der hochgradig komplementären Natur von GSM und FED bestehen dabei sowohl große

potenzielle Mehrwerte als auch Herausforderungen. So sind für FED typische Auswertemethoden nicht direkt auf Textdaten anwendbar und umgekehrt. Die Verbindung der Daten ist aber durch diverse Ansätze technisch möglich. Im Rahmen dieser Dissertation wurde dies durch Embeddings, Aggregation oder Visualisierung erreicht. Da mit diesen Methoden ein gewisser Informationsverlust einhergeht, sollten bei der Analyse von Textdaten zumindest ein Teil der Texte zusätzlich mit sprachwissenschaftlichen Methoden untersucht werden, welche für diesen Datentyp entwickelt wurden. So können eine korrekte Interpretation ermöglicht und Potenziale erkannt werden.

Auch, wenn die Methoden zur Auswertung von Texten zahlreich sind und sich perspektivisch verbessern werden, ist die Interpretation der von Menschen geschriebenen Textdaten herausfordernd. Daten sozialer Medien unterliegen zahlreichen sozioökonomischen Verzerrungen. Speziell für ihre geographische Auswertung als *georeferenzierte* Soziale Medien wurden im Rahmen dieser Dissertation weitere Verzerrungen in Bezug auf Themen und Bebauungsalter identifiziert.

Aus diesem Grund ist die Anwendung von Mixed-Methods und multidisziplinären Auswertungen in diesem Anwendungsbereich sinnvoll, da dabei Verzerrungen erkannt und die Daten richtig eingeordnet werden können. Dabei sind auch andere traditioneller Datenquellen, wie Umfragedaten bedeutungsvoll, welche als Ground-Truth dienen können. Dass deren Bedeutung in Zeiten von Big-Data Analysen abnehmen wird, ist daher nicht zu erwarten. Außerdem sind weiterhin die Fachkenntnisse und praktischen Erfahrungen von Fachleuten und Stakeholdern essenziell zur Einordnung der Ergebnisse.

Der Erforschung von Migration und von Migrierten, welche eine potenziell besonders vulnerable Gruppe darstellen, geht mit besonderen ethischen Herausforderungen einher. Dies gilt insbesondere für GSM, bei welchen es sich um von Personen erstellte Daten handelt, die oftmals auch personenbezogene Daten sind.

Wenn ethische Anforderungen und die Repräsentativität der Daten angemessen berücksichtigt werden, und die Anwendung in einem multidisziplinären Kontext erfolgt, ist die Verbindung von Big-GSM und Big-FED vielversprechend. Im Rahmen dieser Dissertation zeichnet sich hierbei klar eine besondere Eignung für die Erforschung städtischer Nutzenden und Räume ab. Auch wenn hiermit nur ein Teil der afrikanischen Migration abgedeckt werden kann, ist dieser doch, um auf die Einleitung dieser Arbeit zurückzukommen: „*Migration is essentially an urban affair*“ (IOM 2015), sehr relevant. Auch wenn die geographische Perspektive auf Migration auch die ländlichen Räume berücksichtigt, ist das Verständnis von Migration in und zu urbanen Räumen eine wichtige Komponente, bei welcher die räumliche Auswertung von GSM die Fernerkundung sinnvoll ergänzen kann.

Für diese urbanen Räume erlauben GSM die großflächige Kartierung (und damit die empirische Erforschung) von mehreren Arten soziokultureller und individueller Migrationsfaktoren, die für die Fernerkundung nicht erfassbar sind, und sowohl Ursachen als auch Folgen von Migration sein können. Dabei ist nicht sicher, ob durch GSM und FED tatsächlich alle migrationsrelevanten Faktoren ausreichend abgebildet werden können, zu denen neben Umweltfaktoren, individuellen, demographischen und ökonomischen Faktoren auch politisch-institutionelle Faktoren, Sicherheitsfaktoren und supranationale Faktoren gehören (Czaika und Reinprecht 2022). Es ist allerdings vorstellbar, dass auch Letztere grundsätzlich aus Texten und anderen Web-Datenquellen abgeleitet werden können.

Dies zu belegen war jedoch nicht Ziel dieser Dissertation, welche nicht alle Möglichkeiten ausloten soll und kann. Anhand einer Auswahl von Plattformen identifiziert sie aktuelle Potenziale und Herausforderungen der für die geographische Forschung relevantesten Faktoren, welche bei der Konzipierung weiterer geographischer Studien in diesem Kontext berücksichtigt werden können.

8.3. Einordnung und Ausblick

Ein abschließendes Urteil ist –auch aufgrund der Dynamik des Forschungsfeldes– nicht möglich. Sowohl Migrationsmuster als auch die Internet-Landschaft sind in ständigem Wandel, und die technischen Möglichkeiten, um sie abzubilden, werden stetig weiterentwickelt. Forschende, die an der Schnittstelle zwischen Fernerkundung und Textdaten arbeiten, werden sich daher kontinuierlich an die sich ändernden Möglichkeiten und Limitationen anpassen müssen. Hierbei dürften nicht nur methodische, sondern auch rechtliche und wirtschaftliche Aspekte eine Rolle spielen. Die aktuelle Arbeit kann daher nur eine Momentaufnahme darstellen.

Transferierbar können grundlegende Ansätze und Erkenntnisse, wie die Algorithmen zur Ableitung von Wohnorten aus digitalen Spuren, empirisch gewonnene Kenntnisse von Wechselbeziehungen zwischen Migrationsfaktoren und Mobilitätscharakteristika, und nicht zuletzt auch die Transfermethoden selbst sein.

Angesichts der Dynamik des Forschungsfeldes und des Forschungsgegenstandes gestaltet sich dieser Transfer von Methoden in der gemeinsamen Nutzung von GSM und FED als herausfordernd. Diese gemeinsame Nutzung ist aufgrund des erheblichen Potenzials dennoch attraktiv. Es sind die inhärenten Eigenschaften von Satellitenbilddaten und GSM, die sie zu ungleichen, aber hochgradig komplementären Datenquellen für die Erforschung von Migration machen. Die Vereinigung ihrer Perspektiven ermöglicht ein umfassenderes und vollständigeres Bild eines facettenreichen Aspektes des globalen Wandels. Ein Aspekt, dessen Bedeutung für die menschliche Gesellschaft auch in Zukunft nicht abnehmen dürfte.

9. Literatur

- Adelani, David Ifeoluwa, Jade Abbott, Graham Neubig, et al. 2021. 'MasakhaNER: Named Entity Recognition for African Languages'. *Transactions of the Association for Computational Linguistics* 9: 1116–31. https://doi.org/10.1162/tacl_a_00416.
- Adeleke, Richard. 2021. 'Digital Divide in Nigeria: The Role of Regional Differentials'. *African Journal of Science, Technology, Innovation and Development* 13 (3): 333–46. <https://doi.org/10.1080/20421338.2020.1748335>.
- Adika, Newton. 2023. 'Social Media Usage Trends in Africa: GeoPoll Report'. *GeoPoll*, September 6. <https://www.geopoll.com/blog/social-media-usage-trends-in-africa-geopoll-report/>.
- Afolabi, Oluwaseyi Joseph, Olalekan Adedamola Oluwaji, and Olawunmi K. Fashola. 2017. 'Socio-Economic Impact of Road Traffic Congestion on Urban Mobility: A Case Study of Ikeja Local Government Area of Lagos, State, Nigeria'. *Pacific Journal of Science and Technology* 18 (246–255). https://www.akamai.university/files/theme/AkamaiJournal/PJST18_2_246.pdf.
- Aggarwal, Charu C. 2018. *Neural Networks and Deep Learning: A Textbook*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-94463-0>.
- Aggour, Sara. 2014. IFC, EBRD Invest \$60m in Mall of Arabia - *Dailynewsegypt*. August 4. <https://www.dailynewsegypt.com/2014/08/04/ifc-ebd-invest-60m-mall-arabia/>.
- Akanle, Olayinka, OA Fayehun, and S Oyelakin. 2021. 'The Information Communication Technology, Social Media, International Migration and Migrants' Relations with Kin in Nigeria'. *Journal of Asian and African Studies* 56 (6): 1212–25. <https://doi.org/10.1177/0021909620960148>.
- Akoglu, Haldun. 2018. 'User's Guide to Correlation Coefficients'. *Turkish Journal of Emergency Medicine* 18 (3): 91–93.
- Albertz, Jörg. 2001. *Einführung in die Fernerkundung: Grundlagen der Interpretation von Luft- und Satellitenbildern*. Wiss. Buchges., [Abt. Verlag].
- Altieri, L., D. Cocchi, and G. Roli. 2019. 'Measuring Heterogeneity in Urban Expansion via Spatial Entropy: Measuring Heterogeneity in Urban Expansion via Spatial Entropy'. *Environmetrics* 30 (2): e2548. <https://doi.org/10.1002/env.2548>.
- Altieri, Linda, and Daniela Cocchi. 2022. 'Understanding the Expansion of Italian Metropolitan Areas: A Study Based on Entropy Measures'. *Environment and Planning B: Urban Analytics and City Science* 49 (2): 447–63. <https://doi.org/10.1177/23998083211012699>.
- Altieri, Linda, Daniela Cocchi, and Giulia Roli. 2021. 'Spatial Entropy for Biodiversity and Environmental Data: The R-Package SpatEntropy'. *Environmental Modelling & Software* 144 (October): 105149. <https://doi.org/10.1016/j.envsoft.2021.105149>.
- Andreotta, Matthew, Robertus Nugroho, Mark J. Hurlstone, et al. 2019. 'Analyzing Social Media Data: A Mixed-Methods Framework Combining Computational and Qualitative Text Analysis'. *Behavior Research Methods* 51 (4): 1766–81. <https://doi.org/10.3758/s13428-019-01202-8>.
- Anguera, M. Teresa, Angel Blanco-Villaseñor, José Luis Losada, Pedro Sánchez-Algarra, and Anthony J. Onwuegbuzie. 2018. 'Revisiting the Difference between Mixed Methods and Multimethods: Is It All in the Name?' *Quality & Quantity* 52 (6): 2757–70. <https://doi.org/10.1007/s11135-018-0700-2>.
- Aniche, Ernest Tooohi. 2022. 'Borders, Migration and Xenophobic Policies in West Africa: Contraventions of the ECOWAS Free Movement Protocol, and the Ghana-Nigeria Conundrum'. *Africa Review* 14 (1): 24–47. <https://doi.org/10.1163/09744061-20220121>.
- Anthony, David W. 2023. 'Ancient DNA and Migrations: New Understandings and Misunderstandings'. *Journal of Anthropological Archaeology* 70: 101508.
- Antypas, Dimosthenis, Asahi Ushio, Jose Camacho-Collados, Leonardo Neves, Vítor Silva, and Francesco Barbieri. 2022. 'Twitter Topic Classification'. *arXiv:2209.09824*. Preprint, arXiv, September 20. <https://doi.org/10.48550/arXiv.2209.09824>.
- Arcila-Calderón, Carlos, Patricia Sánchez-Holgado, Cristina Quintana-Moreno, Javier-J. Amores, and David Blanco-Herrero. 2022. 'Hate Speech and Social Acceptance of Migrants in Europe: Analysis of Tweets with Geolocation'. *Comunicar: Media Education Research Journal* 30 (71): 21–34.
- Armstrong, Caitrin, Ate Poorthuis, Matthew Zook, Derek Ruths, and Thomas Soehl. 2021. 'Challenges When Identifying Migration from Geo-Located Twitter Data'. *EPJ Data Science* 10 (1): 1. <https://doi.org/10.1140/epjds/s13688-020-00254-7>.

- Babawilly. 2023. 'Babawilly's Dictionary of Pidgin English Words and Phrases. NigeriaExchange.' <https://www.ngex.com/personalities/babawilly/dictionary/>.
- Baddeley, Adrian, and Rolf Turner. 2005. 'Spatstat: An R Package for Analyzing Spatial Point Patterns'. *Journal of Statistical Software* 12 (6): 1–42. <https://doi.org/10.18637/jss.v012.i06>.
- Badreldin Group. 2024. 'About the Developer – Arkan Palm Behind 205 Towers'. About Arkan Palm, October 1. <https://arkanpalm.com/about-the-developer/>.
- Barrie, Christopher, and Justin Chun-ting Ho. 2021. 'academicwitter: An R Package to Access the Twitter Academic Research Product Track v2 API Endpoint'. *Journal of Open Source Software* 6 (62): 3272. <https://doi.org/10.21105/joss.03272>.
- Barwińska-Małajowicz, Anna, and Alexandra David. 2014. Motive der Migration: Suche nach besseren Berufschancen oder sozioökonomische Notwendigkeit? Research Report No. 05/2014. *Forschung Aktuell*. <https://www.econstor.eu/handle/10419/97168>.
- Basmus, Rasheed, Simon Ayodele, and Lambe Kayode. 2019. Agora for the Diaspora: Exploring the Use of Nairaland Online Forum for Political Deliberations among Nigerian Emigrants. 16 (1): 20.
- Bean, Frank D., and Susan K. Brown. 2014. 'Demographic Analyses of Immigration'. In *Migration Theory*. Routledge.
- Becker, Charles M. 2008. 'Urbanization and Rural–Urban Migration'. *International Handbook of Development Economics, Volumes 1 & 2*. <https://www.elgaronline.com/downloadpdf/edcollchap/9781845423278.00044.pdf>.
- Becker, Martin. 2018. 'Understanding Human Navigation Using Bayesian Hypothesis Comparison'. Doctoralthesis, Universität Würzburg.
- Bernt, Matthias. 2019. 'Migration and Strategic Urban Planning: The Case of Leipzig'. *disP - The Planning Review* 55 (3): 56–66. <https://doi.org/10.1080/02513625.2019.1671002>.
- Berragan, Cillian, Alex Singleton, Alessia Calafiore, and Jeremy Morley. 2022. Geoparsing Comments from Reddit to Extract Mental Place Connectivity within the United Kingdom. <https://escholarship.org/uc/item/4wg37921>.
- Bertram, Theo, Elie Bursztein, Stephanie Caro, et al. 2019. 'Five Years of the Right to Be Forgotten'. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (New York, NY, USA), CCS '19, November 6, 959–72*. <https://doi.org/10.1145/3319535.3354208>.
- Bieswanger, Markus, and Annette Becker. 2021. *Introduction to English Linguistics*. 5th edn. UTB GmbH.
- Biewer, Carolin, Lisa Lehnen, Ninja Schulz, and Hannes Taubenböck. forthcoming. *Geolinguistic Studies on Urban Space*. Edinburgh University Press.
- Biljecki, Filip, and Koichi Ito. 2021. 'Street View Imagery in Urban Analytics and GIS: A Review'. *Landscape and Urban Planning* 215 (November): 104217. <https://doi.org/10.1016/j.landurbplan.2021.104217>.
- Bircan, Tuba. 2023. *Remote Sensing Data for Migration Research*. <https://doi.org/10.5871/bacad/9780197267103.003.0006>.
- Bisbee, James, and Kevin Munger. 2025. 'The Vibes Are Off: Did Elon Musk Push Academics Off Twitter?' *PS: Political Science & Politics* 58 (1): 139–46. <https://doi.org/10.1017/S1049096524000416>.
- Black, Richard, Stephen R. G. Bennett, Sandy M. Thomas, and John R. Beddington. 2011. 'Migration as Adaptation'. *Nature* 478 (7370): 447–49. <https://doi.org/10.1038/478477a>.
- Blanchard, Paul, Douglas Gollin, and Martina Kirchberger. 2020. *Perpetual Motion: Human Mobility and Spatial Frictions in Three African Countries*. 63.
- Blanqué, Pascal, Mohamed Ben Slimane, Amina Cherief, Théo Le Guenedal, Takaya Sekine, and Lauren Stagnol. 2022. 'Monitoring Narratives: An Application to the Equity Market'. *SSRN Electronic Journal*, ahead of print. <https://doi.org/10.2139/ssrn.4081958>.
- Blaschke, Th. 2000. 'Landscape Metrics: Konzepte Eines Jungen Ansatzes Der Landschaftsökologie Und Anwendungen in Naturschutz Und Landschaftsforschung'. *Archives of Nature Conservation and Landscape Research* 39 (4): 267–99.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. 'Latent Dirichlet Allocation'. *Journal of Machine Learning Research* 3 (Jan): 993–1022.
- Blench, Roger, and Mallam Dendo. 2005. 'A Dictionary of Nigerian English'. Cambridge, United Kingdom, CB1 2AL 26.

- Bloch, Robin, Sean Fox, Jose Monroy, and Adegbola Ojo. 2015. Urbanisation and Urban Expansion in Nigeria. *Urbanisation Research Nigeria (URN)*, 73.
- Bodas-Sagi, Diego, and José Labeaga. 2016. Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy. <https://reunir.unir.net/handle/123456789/10962>.
- Boeschoten, Laura, Jef Ausloos, Judith E. Möller, Theo Araujo, and Daniel L. Oberski. 2022. 'A Framework for Privacy Preserving Digital Trace Data Collection through Data Donation'. *Computational Communication Research* 4 (2): 388–423. <https://doi.org/10.5117/CCR2022.2.002.BOES>.
- Borlongan, Ariane Macalinga. 2023. 'Migration Linguistics: A Synopsis'. *AILA Review* 36 (1): 38–63. <https://doi.org/10.1075/aila.22014.bor>.
- Boy, John D, and Justus Uitermark. 2017. 'Reassembling the City through Instagram'. *Transactions of the Institute of British Geographers* 42 (4): 612–24. <https://doi.org/10.1111/tran.12185>.
- Brettell, Caroline, and James Frank Hollifield, eds. 2015. *Migration Theory: Talking across Disciplines*. Third edition. Routledge.
- Brooks, Mollie E., Kasper Kristensen, Koen J. van Benthem, et al. 2017. 'glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling'. *The R Journal* 9 (2): 378–400. <https://doi.org/10.32614/RJ-2017-066>.
- Brown, Christopher F., Michal R. Kazmierski, Valerie J. Pasquarella, et al. 2025. 'AlphaEarth Foundations: An Embedding Field Model for Accurate and Efficient Global Mapping from Sparse Label Data'. arXiv:2507.22291. Preprint, arXiv, July 29. <https://doi.org/10.48550/arXiv.2507.22291>.
- Cai, Jixuan, Bo Huang, and Yimeng Song. 2017. 'Using Multi-Source Geospatial Big Data to Identify the Structure of Polycentric Cities'. *Remote Sensing of Environment, Big Remotely Sensed Data: tools, applications and experiences*, vol. 202 (December): 210–21. <https://doi.org/10.1016/j.rse.2017.06.039>.
- Campbell, Donald T. 1957. 'Factors Relevant to the Validity of Experiments in Social Settings'. *Psychological Bulletin (US)* 54 (4): 297–312. <https://doi.org/10.1037/h0040950>.
- Canagarajah, Suresh. 2017. 'Introduction: The Nexus of Migration and Language: The Emergence of a Disciplinary Space'. In *The Routledge Handbook of Migration and Language*. Routledge.
- Carling, Jørgen, Mathias Czaika, and Marta Bivand Erdal. 2020. 'Translating Migration Theory into Empirical Propositions'. *Changes* 1: 27.
- Castelvecchi, Davide. 2016. 'Can We Open the Black Box of AI?' *Nature News* 538 (7623): 20.
- Cecaj, Alket, Marco Mamei, and Nicola Biccocchi. 2014. 'Re-Identification of Anonymized CDR Datasets Using Social Network Data'. 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS), March, 237–42. <https://doi.org/10.1109/PerComW.2014.6815210>.
- Chancellor, Stevie, and Munmun De Choudhury. 2020. 'Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review'. *Npj Digital Medicine* 3 (1): 1–11. <https://doi.org/10.1038/s41746-020-0233-7>.
- Chang, Kang-Tsung. 2019. *Introduction to Geographic Information Systems*. McGraw-Hill Education,. <https://dlib.scu.ac.ir/bitstream/Hannan/682138/1/9781259929649.pdf>.
- Charmaz, Kathy. 2006. *Constructing Grounded Theory*. Sage Publications.
- Chavent, Marie, Vanessa Kuentz-Simonet, Amaury Labenne, and Jérôme Saracco. 2018. 'ClustGeo: An R Package for Hierarchical Clustering with Spatial Constraints'. *Computational Statistics* 33 (4): 1799–822. <https://doi.org/10.1007/s00180-018-0791-1>.
- Chazalnoël, Mariam Traore, and Alex Randall. 2022. '9 Migration and the Slow-onset Impacts of Climate Change: Taking Stock and Taking Action'. *World Migration Report 2022* (1): e00030. <https://doi.org/10.1002/wom3.30>.
- Chen, Yile, Weiming Huang, Kaiqi Zhao, Yue Jiang, and Gao Cong. 2024. 'Self-Supervised Learning for Geospatial AI: A Survey'. arXiv:2408.12133. Preprint, arXiv, August 22. <http://arxiv.org/abs/2408.12133>.
- Chi, Guanghua. 2020. 'Migration and Social Networks: New Insights from Novel Data'. UC Berkeley. <https://escholarship.org/uc/item/6rs5n7jq>.
- Chi, Guanghua, Guy J. Abel, Drew Johnston, Eugenia Giraudy, and Mike Bailey. 2025. 'Measuring Global Migration Flows Using Online Data'. arXiv:2504.11691. Preprint, arXiv, April 16. <https://doi.org/10.48550/arXiv.2504.11691>.

- Chi, Guanghua, Fengyang Lin, Guangqing Chi, and Joshua Blumenstock. 2020. 'A General Approach to Detecting Migration Events in Digital Trace Data'. *PLOS ONE* 15 (10): e0239408. <https://doi.org/10.1371/journal.pone.0239408>.
- Chiluwa, Innocent, and Akin Odebunmi. 2016. 'On Terrorist Attacks in Nigeria: Stance and Engagement in Conversations on Nairaland'. *Communication and the Public* 1 (1): 91–109. <https://doi.org/10.1177/2057047315624926>.
- CIA. 2025a. 'Benin'. In *The World Factbook*. Central Intelligence Agency, August 20. <https://www.cia.gov/the-world-factbook/countries/benin/>.
- CIA. 2025b. 'Burkina Faso'. In *The World Factbook*. Central Intelligence Agency, August 20. <https://www.cia.gov/the-world-factbook/countries/burkina-faso/>.
- CIA. 2025c. 'Nigeria'. In *The World Factbook*. Central Intelligence Agency, August 20. <https://www.cia.gov/the-world-factbook/countries/nigeria/>.
- CIA. 2025d. 'Togo'. In *The World Factbook*. Central Intelligence Agency, August 20. <https://www.cia.gov/the-world-factbook/countries/togo/>.
- Cleveland, William S., and Clive Loader. 1996. 'Smoothing by Local Regression: Principles and Methods'. Edited by Wolfgang Härdle and Michael G. Schimek. *Contributions to Statistics*, edited by Werner A. Müller and Peter Schuster. Physica-Verlag HD. https://doi.org/10.1007/978-3-642-48425-4_2.
- Cohen, Jacob. 1960. 'A Coefficient of Agreement for Nominal Scales'. *Educational and Psychological Measurement* 20 (1): 37–46. <https://doi.org/10.1177/001316446002000104>.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Routledge. <https://doi.org/10.4324/9780203771587>.
- Common Crawl. 2024. 'Distribution of Languages'. <https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html>.
- Connell, Bruce, and David Zeitlyn. 2023. 'Sociolinguistic Studies of West And Central Africa'. In *The Routledge Handbook of Sociolinguistics Around the World*, 2nd edn. Routledge.
- Corum, Micah. 2015. 'Chapter 4. Meanings and Functions of for in Nigerian Pidgin and Ghanaian Pidgin English'. In *Substrate and Adstrate: The Origins of Spatial Semantics in West African Pidgincreoles*. De Gruyter Mouton. <https://www.degruyterbrill.com/document/doi/10.1515/9781614514626-009/html>.
- Curry, A, I Bashir, and P Raven. 2024. *Reimagining Migration and Mobility: New Ideas for an Age-Old Human Phenomenon*. International Organization for Migration (IOM) and School of International Futures (SOIF), Geneva and London. <https://publications.iom.int/books/reimagining-migration-and-mobility-new-ideas-age-old-human-phenomenon>.
- Czaika, Mathias, and Sultan Orazbayev. 2018. 'The Globalisation of Scientific Mobility, 1970–2014'. *Applied Geography* 96 (July): 1–10. <https://doi.org/10.1016/j.apgeog.2018.04.017>.
- Czaika, Mathias, and Constantin Reinprecht. 2022. 'Migration Drivers: Why Do People Migrate'. *Introduction to Migration Studies: An Interactive Guide to the Literatures on Migration and Diversity*, 49–82.
- Dangschat, Jens S. 2019. 'Räumliche Daten'. In *Handbuch Methoden der empirischen Sozialforschung*, edited by Nina Baur and Jörg Blasius. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-18939-0_76.
- Davidson, Brittany I, Darja Wischerath, Daniel Racek, et al. 2023. *Social Media APIs: A Quiet Threat to the Advancement of Science*. Preprint. PsyArXiv. <https://doi.org/10.31234/osf.io/ps32z>.
- Debray, Henri, Nicolas J. Kraff, Xiao Xiang Zhu, and Hannes Taubenböck. 2023. 'Planned, Unplanned, or in-between? A Concept of the Intensity of Plannedness and Its Empirical Relation to the Built Urban Landscape across the Globe'. *Landscape and Urban Planning* 233 (May): 104711. <https://doi.org/10.1016/j.landurbplan.2023.104711>.
- Dekker, Rianne, and Godfried Engbersen. 2014. 'How Social Media Transform Migrant Networks and Facilitate Migration'. *Global Networks* 14 (4): 401–18. <https://doi.org/10.1111/glob.12040>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. *Proceedings of NAACL-HLT 2019*, 4171–86. <https://doi.org/10.48550/arXiv.1810.04805>.
- D'Ignazio, C., Bhargava, R., Zuckerman, E., & Beck, L. (2014). *Cliff-clavin: Determining geographic focus for news articles*. <https://hdl.handle.net/1721.1/123451>.

- Dinbabo, M. F. 2018. 'Migration Research and Analysis: Growth, Reach and Recent Contributions'. International Organization for Migration. World Migration Report, 318–20.
- Distefano, Veronica, Valentina Mameli, and Irene Poli. 2020. 'Identifying Spatial Patterns with the Bootstrap ClustGeo Technique'. *Spatial Statistics* 38 (August): 100441. <https://doi.org/10.1016/j.spasta.2020.100441>.
- Doane, David P., and Lori E. Seward. 2011. 'Measuring Skewness: A Forgotten Statistic?' *Journal of Statistics Education* 19 (2). <https://doi.org/10.1080/10691898.2011.11889611>.
- Dodman, D., B. Hayward, M. Pelling, et al. 2022. Cities, Settlements and Key Infrastructure. In: *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. <https://research.fs.usda.gov/treesearch/64737>.
- Duncan, Howard, and Ioana Popp. 2018. 'Chapter 10 – Migrants and Cities: Stepping beyond World Migration Report 2015'. *World Migration Report 2018* (1): e00010. <https://doi.org/10.1002/wom3.10>.
- Economic Intelligence Unit, Ministry of Economic Planning & Budget. 2013. 'The Socio-Economic Costs of Traffic Congestion in Lagos'. Working Paper Series 2.
- Efthimion, Phillip George, Scott Payne, and Nicholas Proferes. 2018. 'Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots'. *SMU Data Science Review* 1 (2): 71.
- Egger, Roman, and Joanne Yu. 2022. 'A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts'. *Frontiers in Sociology* 7 (May): 886498. <https://doi.org/10.3389/fsoc.2022.886498>.
- Egghe, Leo, and Loet Leydesdorff. 2009. 'The Relation between Pearson's Correlation Coefficient r and Salton's Cosine Measure'. *Journal of the American Society for Information Science and Technology* 60 (5): 1027–36.
- El Ouadi, Ameir, and David Beskow. 2024. 'Comparison of Common Crawl News & GDELT'. 2024 IEEE International Systems Conference (SysCon), 1–3. https://ieeexplore.ieee.org/abstract/document/10553540/?casa_token=SelHzCNBRuMAAAA:XMdJhmztQBabIAJEwwdIR01OekKcvJySFKFTjG3_ftR-9DJ-NO-pt3P3BFa-mikAU13D2AjtPOk.
- Esch, Thomas, Wieke Heldens, Andreas Hirner, et al. 2017. 'Breaking New Ground in Mapping Human Settlements from Space – The Global Urban Footprint'. *ISPRS Journal of Photogrammetry and Remote Sensing* 134 (December): 30–42. <https://doi.org/10.1016/j.isprsjprs.2017.10.012>.
- Esmaail, Naji, Nazlia Omar, Masnizah Mohd, Fariza Fauzi, and Zainab Mansur. 2024. 'Named Entity Recognition in User-Generated Text: A Systematic Literature Review'. *IEEE Access* 12: 136330–53. <https://doi.org/10.1109/ACCESS.2024.3427714>.
- Esser, Hartmut. 2006. *Migration, Sprache Und Integration*. Vol. 4. AKI-Forschungsbilanz. Wissenschaftszentrum Berlin für Sozialforschung gGmbH FSP Zivilgesellschaft, Konflikte und Demokratie Arbeitsstelle Interkulturelle Konflikte und gesellschaftliche Integration -AKI-.
- Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'. *Kdd* 96 (34): 226–31. https://cdn.aaai.org/KDD/1996/KDD96-037.pdf?source=post_page-----.
- European Commission. 2022. 'Data Innovation in Demography, Migration and Human Mobility'. Publications Office of the European Union. <https://publications.jrc.ec.europa.eu/repository/handle/JRC127369>.
- Ezli, Özkan. 2022. *Narrative der Migration : Eine andere deutsche Kulturgeschichte*. De Gruyter. <https://kops.uni-konstanz.de/handle/123456789/55657>.
- Farahani, Abolfazl, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. 2021. 'A Brief Review of Domain Adaptation'. In *Advances in Data Science and Information Engineering*, edited by Robert Stahlbock, Gary M. Weiss, Mahmoud Abou-Nasr, Cheng-Ying Yang, Hamid R. Arabnia, and Leonidas Deligiannidis. Springer International Publishing. https://doi.org/10.1007/978-3-030-71704-9_65.
- Fargues, Philippe. 2011. 'International Migration and the Demographic Transition: A Two-Way Interaction'. *International Migration Review* 45 (3): 588–614.
- Farzana, Sheikh Mastura, and Tobias Hecking. 2024. 'Towards a Scalable Geoparsing Approach for the Web'. 46th European Conference on Information Retrieval, ECIR 2024.
- Fayissa, Bichaka, and Christian Nsiah. 2010. 'The Impact of Remittances on Economic Growth and Development in Africa'. *The American Economist* 55 (2): 92–103. <https://doi.org/10.1177/056943451005500210>.

- Feng, Zhengpeng, Clement Atzberger, Sadiq Jaffer, et al. 2025. 'TESSERA: Precomputed FAIR Global Pixel Embeddings for Earth Representation and Analysis'. arXiv:2506.20380. Preprint, arXiv, September 19. <https://doi.org/10.48550/arXiv.2506.20380>.
- Ferrara, Emilio. 2023. 'Social Bot Detection in the Age of ChatGPT: Challenges and Opportunities'. First Monday, ahead of print, June 15. <https://doi.org/10.5210/fm.v28i6.13185>.
- Fiallos, Angel, and Karina Jimenes. 2019. 'Using Reddit Data for Multi-Label Text Classification of Twitter Users Interests'. 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG), 324–27. <https://doi.org/10.1109/ICEDEG.2019.8734365>.
- Findlay, Allan M., and Jackline Wahba. 2013. 'Migration and Demographic Change'. Population, Space and Place 19 (6): 651–56. <https://doi.org/10.1002/psp.1786>.
- Firouzjaei, Hassan Abedi. 2024. 'A Deep Learning-Based Approach for Identifying Unresolved Questions on Stack Exchange Q & A Communities through Graph-Based Communication Modelling'. International Journal of Data Science and Analytics 18 (2): 205–18. <https://doi.org/10.1007/s41060-023-00454-0>.
- Flahaux, Marie-Laurence, and Hein De Haas. 2016. 'African Migration: Trends, Patterns, Drivers'. Comparative Migration Studies 4 (1): 1. <https://doi.org/10.1186/s40878-015-0015-6>.
- Foster, Kathryn, and Matthew Hall. 2025. 'The Demographic Future of Migration Is African'. International Migration 63 (1): e13377. <https://doi.org/10.1111/imig.13377>.
- Fox, Nathan, Laura J. Graham, Felix Eigenbrod, James M. Bullock, and Katherine E. Parks. 2021. 'Reddit: A Novel Data Source for Cultural Ecosystem Service Studies'. Ecosystem Services 50: 101331.
- Freedman, David A. 2009. Statistical Models: Theory and Practice. Cambridge University Press.
- Gan, Qingqing, Josh Attenberg, Alexander Markowetz, and Torsten Suel. 2008. 'Analysis of Geographic Queries in a Search Engine Log'. Proceedings of the First International Workshop on Location and the Web, April 22, 49–56. <https://doi.org/10.1145/1367798.1367806>.
- Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, et al. 2016. 'Domain-Adversarial Training of Neural Networks'. Journal of Machine Learning Research 17 (59): 1–35.
- Gardazi, Nadia Mushtaq, Ali Daud, Muhammad Kamran Malik, Amal Bukhari, Tariq Alsahfi, and Bader Alshemaimri. 2025. 'BERT Applications in Natural Language Processing: A Review'. Artificial Intelligence Review 58 (6): 166. <https://doi.org/10.1007/s10462-025-11162-5>.
- Gauss, Carl-Friedrich. 1823. Theoria combinationis observationum erroribus minimis obnoxiae. Henricus Dieterich.
- GDELT Project. 2015. 'THE GDELT GLOBAL KNOWLEDGE GRAPH (GKG) DATA FORMAT CODEBOOK V2.1'. http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook-V2.1.pdf.
- GDELT Project. o.D. The World As Seen Through Books: Comparing the Internet Archive, HathiTrust, and Google Books Ngrams Collections – The GDELT Project. Accessed 9 September 2025. <https://blog.gdeltproject.org/the-world-as-seen-through-books-comparing-the-internet-archive-hathitrust-and-google-books-ngrams-collections/>.
- Geis-Thöne, Wido. 2020. 'Der Beitrag Der Zuwanderung Zur Stabilisierung Der Demografischen Entwicklung'. IW-Trends-Vierteljahresschrift Zur Empirischen Wirtschaftsforschung 47 (2): 129–48.
- GeoNames. o.D. Zugriff am 9 November 2022. <https://www.geonames.org/>.
- Ghana Statistical Service. 2023. 'Ghana 2021 Population and Housing Census Thematic Report: Migration'. Ghana Statistical Service Accra, Ghana.
- Glaeser, Edward. 2011. Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier. Penguin Publishing Group.
- Goanta, Catalina, Savvas Zannettou, Rishabh Kaushal, et al. 2025. 'The Great Data Standoff: Researchers vs. Platforms Under the Digital Services Act'. arXiv:2505.01122. Preprint, arXiv, May 2. <https://doi.org/10.48550/arXiv.2505.01122>.
- Gollin, Douglas, Paul Blanchard, and Martina Kirchberger. 2021. 'Perpetual Motion: Human Mobility and Spatial Frictions in Three African Countries'. SSRN Scholarly Paper No. 3960245. Rochester, NY, October 1. <https://papers.ssrn.com/abstract=3960245>.

- Gomes, Vitor C. F., Gilberto R. Queiroz, and Karine R. Ferreira. 2020. 'An Overview of Platforms for Big Earth Observation Data Management and Analysis'. *Remote Sensing* 12 (8): 8. <https://doi.org/10.3390/rs12081253>.
- Gonzales, Jack Joseph. 2023. 'Building-Level Comparison of Microsoft and Google Open Building Footprints Datasets (Short Paper)'. *LIPICs*, Volume 277, *GIScience* 2023 277: 35:1-35:6. <https://doi.org/10.4230/LIPICS.GISCIENCE.2023.35>.
- Goodfellow, Tom. 2022. *Politics and the Urban Frontier: Transformation and Divergence in Late Urbanizing East Africa*. Oxford University Press. <https://doi.org/10.1093/oso/9780198853107.001.0001>.
- Google. (2016) 2016. Google/Cld3. C++. June 2; Google, released. <https://github.com/google/cld3>.
- Google. 2024. 'Google Places'. Google for Developers. <https://developers.google.com/maps/documentation/places/web-service/overview>.
- Google. o.D. 'Satellite Embedding V1'. Earth Engine Data Catalog. Zugriff am 3 August 2025. https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_SATELLITE_EMBEDDING_V1_ANNUAL.
- Graham, Mark, and Stefano De Sabbata. 2020. 'The Geographic Contours of Openness'. *Making Open Development Inclusive: Lessons from IDRC Research*, 119.
- Gregory, Ian, David Cooper, Andrew Hardie, and Paul Rayson. 2015. 'Spatializing and Analyzing Digital Texts: Corpora, GIS and Places'. *Deep Maps and Spatial Narratives*, 150–78.
- Gries, Stefan. 2015. 'The Most Under-Used Statistical Method in Corpus Linguistics: Multi-Level (and Mixed-Effects) Models'. *Corpora* 10 (1): 95–125. <https://doi.org/10.3366/cor.2015.0068>.
- Gritta, Milan, Mohammad Taher Pilehvar, and Nigel Collier. 2018. 'Which Melbourne? Augmenting Geocoding with Maps'. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1285–96. <https://doi.org/10.18653/v1/P18-1119>.
- Grootendorst, Maarten. 2022. 'BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure'. *arXiv Preprint arXiv:2203.05794*.
- Gumuchian, Marie-Louise. 2012. 'Pro-Government Forces Seize Old Gaddafi Bastion Bani Walid'. *China*. Reuters, October 24. <https://www.reuters.com/article/world/china/pro-government-forces-seize-old-gaddafi-bastion-bani-walid-idUSBRE89N0WQ/>.
- Guyer, Jane I., and Eric F. Lambin. 1993. 'Land Use in an Urban Hinterland: Ethnography and Remote Sensing in the Study of African Intensification'. *American Anthropologist* 95 (4): 839–59. <https://doi.org/10.1525/aa.1993.95.4.02a00040>.
- Häberle, Matthias, Eike Jens Hoffmann, and Xiao Xiang Zhu. 2022. 'Can Linguistic Features Extracted from Geo-Referenced Tweets Help Building Function Classification in Remote Sensing?' *ISPRS Journal of Photogrammetry and Remote Sensing* 188 (June): 255–68. <https://doi.org/10.1016/j.isprsjprs.2022.04.006>.
- Haraguchi, Masahiko, Akihiko Nishino, Akira Kodaka, et al. 2022. 'Human Mobility Data and Analysis for Urban Resilience: A Systematic Review'. *Environment and Planning B: Urban Analytics and City Science* 49 (5): 1507–35. <https://doi.org/10.1177/23998083221075634>.
- Havas, Clemens, Lorenz Wendlinger, Julian Stier, et al. 2021. 'Spatio-Temporal Machine Learning Analysis of Social Media Data and Refugee Movement Statistics'. *ISPRS International Journal of Geo-Information* 10 (8): 498. <https://doi.org/10.3390/ijgi10080498>.
- Hawelka, Bartosz, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. 2014. 'Geo-Located Twitter as Proxy for Global Mobility Patterns'. *Cartography and Geographic Information Science* 41 (3): 260–71. <https://doi.org/10.1080/15230406.2014.890072>.
- Hecht, Brent, and Monica Stephens. 2014. 'A Tale of Cities: Urban Biases in Volunteered Geographic Information'. *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1): 197–205. <https://ojs.aaai.org/index.php/ICWSM/article/view/14554>.
- Heidenreich, Tobias, Jakob-Moritz Eberl, Fabienne Lind, and Hajo Boomgaarden. 2020. 'Political Migration Discourses on Social Media: A Comparative Perspective on Visibility and Sentiment across Political Facebook Accounts in Europe'. *Journal of Ethnic and Migration Studies* 46 (7): 1261–80. <https://doi.org/10.1080/1369183X.2019.1665990>.
- Hemphill, Libby, Angela Schöpke-Gonzalez, and Anmol Panda. 2022. 'Comparative Sensitivity of Social Media Data and Their Acceptable Use in Research'. *Scientific Data* 9 (1): 643. <https://doi.org/10.1038/s41597-022-01773-w>.

- Hermit, Dave. (2016) 2023. FrequencyWords. C#. August 16, released August 30. <https://github.com/hermitdave/FrequencyWords>.
- Hesselbarth, Maximilian H. K., Marco Sciaini, Kimberly A. With, Kerstin Wiegand, and Jakub Nowosad. 2019. 'LandscapeMetrics: An Open-Source R Tool to Calculate Landscape Metrics'. *Ecography* 42: 1648–57.
- Heyd, Theresa. 2014. 'Doing Race and Ethnicity in a Digital Community: Lexical Labels and Narratives of Belonging in a Nigerian Web Forum'. *Discourse, Context & Media* 4–5: 38–47. <https://doi.org/10.1016/j.dcm.2013.11.002>.
- Hickey, Daniel, Matheus Schmitz, Daniel Fessler, Paul E. Smaldino, Goran Muric, and Keith Burghardt. 2023. 'Auditing Elon Musk's Impact on Hate Speech and Bots'. *Proceedings of the International AAAI Conference on Web and Social Media* 17 (June): 1133–37. <https://doi.org/10.1609/icwsm.v17i1.22222>.
- Hijmans, Robert J. 2025. Terra: Spatial Data Analysis. Released. <https://github.com/rspatial/terra>.
- Hobbs, Jerry R. 1990. 'Topic Drift'. *Conversational Organization and Its Development* 38: 3–22.
- Hoffmann, Eike Jens. 2022. 'Predicting Building Functions by Fusing Social Media and Remote Sensing Data'. In *mediaTUM*. Thesis, Technical University of Munich. <https://mediatum.ub.tum.de/?id=1655456>.
- Hollander, Myles, Douglas A. Wolfe, and Eric Chicken. 1973. *Nonparametric Statistical Methods*. John Wiley & Sons.
- Holmes, Janet, and Nick Wilson. 2022. *An Introduction to Sociolinguistics*. 6th edn. Routledge. <https://doi.org/10.4324/9780367821852>.
- Honkanen, Mirka. 2020. *World Englishes on the Web: The Nigerian Diaspora in the USA*. Vol. G63. *Varieties of English Around the World*. John Benjamins Publishing Company. <https://doi.org/10.1075/veaw.g63>.
- Honkanen, Mirka, and Julia Müller. 2021. 'Interjections and Emojis in Nigerian Online Communication'. *World Englishes*. https://www.academia.edu/66834489/Interjections_and_emojis_in_Nigerian_online_communication.
- Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-Strength Natural Language Processing in Python*. <https://doi.org/10.5281/zenodo.1212303>.
- Hu, Xuke, Tobias Elßner, Shiyu Zheng. 2024. 'DLRGeoTweet: A Comprehensive Social Media Geocoding Corpus Featuring Fine-Grained Places'. *Information Processing & Management* 61 (4): 103742. <https://doi.org/10.1016/j.ipm.2024.103742>.
- Hu, Xuke, Jens Kersten, Friederike Klan, and Sheikh Mastura Farzana. 2024. 'Toponym Resolution Leveraging Lightweight and Open-Source Large Language Models and Geo-Knowledge'. *International Journal of Geographical Information Science*, September 24, 1–28. <https://doi.org/10.1080/13658816.2024.2405182>.
- Hu, Xuke, Yeran Sun, Jens Kersten, Zhiyong Zhou, Friederike Klan, and Hongchao Fan. 2023. 'How Can Voting Mechanisms Improve the Robustness and Generalizability of Toponym Disambiguation?' *International Journal of Applied Earth Observation and Geoinformation* 117: 103191.
- Hu, Xuke, Zhiyong Zhou, Hao Li, et al. 2023. 'Location Reference Recognition from Texts: A Survey and Comparison'. *ACM Computing Surveys* 56 (5): 1–37. <https://doi.org/10.1145/3625819>.
- Hua, Wen, Dat T Huynh, Saeid Hosseini, Jiaheng Lu, and Xiaofang Zhou. 2012. 'Information Extraction From Microblogs: A Survey'. *Int. J. Softw. Informatics* 6 (4): 495–522.
- Huang, Binxuan, and Kathleen M. Carley. 2019. 'A Large-Scale Empirical Study of Geotagging Behavior on Twitter'. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (New York, NY, USA), ASONAM '19*, 365–73. <https://doi.org/10.1145/3341161.3342870>.
- Hübl, Franziska, Sreten Cvetojevic, Hartwig Hochmair, and Gernot Paulus. 2017. 'Analyzing Refugee Migration Patterns Using Geo-Tagged Tweets'. *ISPRS International Journal of Geo-Information* 6 (10): 302. <https://doi.org/10.3390/ijgi6100302>.
- Hughes, Christina, Emilio Zagheni, Guy J. Abel, et al. 2016. *Inferring Migrations: Traditional Methods and New Approaches Based on Mobile Phone, Social Media, and Other Big Data: Feasibility Study on Inferring (Labour) Mobility and Migration in the European Union from Big Data and Social Media Data*.
- Hundt, Marianne, Nadja Nesselhauf, and Carolin Biewer. 2007. 'Corpus Linguistics and the Web'. In *Corpus Linguistics and the Web*. Brill. https://doi.org/10.1163/9789401203791_002.
- IDMC. 2025. 'About Us'. About Us - Internal Displacement Monitoring Centre (IDMC). <https://www.internal-displacement.org/about-us>.

- Iheke, O R. 2012. The Effect of Remittances on the Nigerian Economy.
- Ilieva, Rositsa T., and Timon McPhearson. 2018. 'Social-Media Data for Urban Sustainability'. *Nature Sustainability* 1 (10): 553–65. <https://doi.org/10.1038/s41893-018-0153-6>.
- Intergovernmental Panel On Climate Change (Ipcc). 2022. Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. 1st edn. Cambridge University Press. <https://doi.org/10.1017/9781009325844>.
- International Organization for Migration. 2015. World Migration Report 2015: Migrants and Cities-New Partnerships to Manage Mobility. UN.
- International Organization for Migration. 2022. Displacement Tracking Matrix (DTM) of the International Organization for Migration. DTM. <https://dtm.iom.int/reports/nigeria-%E2%80%94-emergency-tracking-tool-report-280-13-19-june-2022>.
- ITU. 2021. Measuring Digital Development - Facts and Figures 2021. 31.
- ITU. 2024. Measuring Digital Development - Facts and Figures 2024. <https://digitallibrary.un.org/record/4074377>.
- Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. 'Why We Twitter: Understanding Microblogging Usage and Communities'. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (New York, NY, USA), WebKDD/SNA-KDD '07, August 12, 56–65. <https://doi.org/10.1145/1348549.1348556>.
- Javed, Danish, NZ Jhanjhi, Navid Ali Khan, et al. 2024. 'Towards the Future of Bot Detection: A Comprehensive Taxonomical Review and Challenges on Twitter/X'. *Computer Networks* 254 (December): 110808. <https://doi.org/10.1016/j.comnet.2024.110808>.
- Jeremy, Gelb. 2023. 'Geomeans: An R Package for Spatial Fuzzyc-Means'. *Journal of Open Source Software* 8 (89): 5259. <https://doi.org/10.21105/joss.05259>.
- Jheng, Decyuan, Richard Lemoine-Rodríguez, Carolin Biewer, Karen C. Seto, and Hannes Taubenböck. 2025. 'The Relationship Between Urban Morphology and the Citizens' Social Media Footprints: A Case Study on Tokyo'. 2025 Joint Urban Remote Sensing Event (JURSE) CFP25RSD-ART (May): 1–4. <https://doi.org/10.1109/JURSE60372.2025.11076032>.
- Jiang, Hong. 2003. 'Stories Remote Sensing Images Can Tell: Integrating Remote Sensing Analysis with Ethnographic Research in the Study of Cultural Landscapes'. *Human Ecology* 31 (2): 215–32. <https://doi.org/10.1023/A:1023980713048>.
- Jiang, Bin, Ding Ma, Junjun Yin, and Mats Sandberg. 2016. 'Spatial Distribution of City Tweets and Their Densities: City Tweets and Their Densities'. *Geographical Analysis* 48 (3): 337–51. <https://doi.org/10.1111/gean.12096>.
- Jiang, Julie, Jesse Thomason, Francesco Barbieri, and Emilio Ferrara. 2023. 'Geolocated Social Media Posts Are Happier: Understanding the Characteristics of Check-in Posts on Twitter'. Proceedings of the 15th ACM Web Science Conference 2023, April 30, 136–46. <https://doi.org/10.1145/3578503.3583596>.
- Jiang, Yuqin, Zhenlong Li, and Xinyue Ye. 2019. 'Understanding Demographic and Socioeconomic Biases of Geotagged Twitter Users at the County Level'. *Cartography and Geographic Information Science* 46 (3): 228–42. <https://doi.org/10.1080/15230406.2018.1434834>.
- Johnson, Isaac L., Subhasree Sengupta, Johannes Schöning, and Brent Hecht. 2016. 'The Geography and Importance of Localness in Geotagged Social Media'. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, May 7, 515–26. <https://doi.org/10.1145/2858036.2858122>.
- Jurdak, Raja, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. 2015. 'Understanding Human Mobility from Twitter'. *PLOS ONE* 10 (7): e0131469. <https://doi.org/10.1371/journal.pone.0131469>.
- Kaleel, Akhmed, Abdulkrim Ziani, Salma Abdullah Mohamed saeed Alkaabi, and Beena Khalid Alhajri. 2025. 'Platform-Specific Data Decay Patterns: A Comparative Study of Twitter, Reddit, and TikTok'. *MEDAAD 2025* (February): 8–26. <https://doi.org/10.70470/MEDAAD/2025/002>.
- Kamana, Alanda A., Hassan Radoine, and Chimango Nyasulu. 2024. 'Urban Challenges and Strategies in African Cities – A Systematic Literature Review'. *City and Environment Interactions* 21 (January): 100132. <https://doi.org/10.1016/j.cacint.2023.100132>.
- Kang, Chaogui, Xiujun Ma, Daoqin Tong, and Yu Liu. 2012. 'Intra-Urban Human Mobility Patterns: An Urban Morphology Perspective'. *Physica A: Statistical Mechanics and Its Applications* 391 (4): 1702–17.

- Karami, Amir, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, Harshini Bheemreddy, and Parisa Bozorgi. 2021. 'Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population?' *ISPRS International Journal of Geo-Information* 10 (6): 373. <https://doi.org/10.3390/ijgi10060373>.
- KC, Hari. 2025. 'Leveraging Transnational Remote Work for Inclusive Sustainable Development'. *Migration Policy Practice*, no. Vol. XIV(3). <https://doi.org/doi>:
- Kelle, Udo. 2014. 'Mixed Methods'. In *Handbuch Methoden der empirischen Sozialforschung*, edited by Nina Baur and Jörg Blasius. Springer Fachmedien. https://doi.org/10.1007/978-3-531-18939-0_8.
- Kemp, Simon. 2015. 'Digital 2015 Nigeria'. Datareportal. <https://datareportal.com/reports/digital-2015-nigeria>.
- Kemp, Simon. 2019a. 'Digital 2019: Ghana'. DataReportal – Global Digital Insights, January 31. <https://datareportal.com/reports/digital-2019-ghana>.
- Kemp, Simon. 2019b. 'Digital 2019 Nigeria'. Datareportal. <https://datareportal.com/reports/digital-2019-nigeria>.
- Kemp, Simon. 2025. 'Digital 2025: Global Overview Report'. DataReportal – Global Digital Insights, February 5. <https://datareportal.com/reports/digital-2025-global-overview-report>.
- Kersten, Jens, and Friederike Klan. 2020. 'What Happens Where during Disasters? A Workflow for the Multifaceted Characterization of Crisis Events Based on Twitter Data'. *Journal of Contingencies and Crisis Management* 28 (3): 262–80. <https://doi.org/10.1111/1468-5973.12321>.
- Khatua, Aparup, and Wolfgang Nejdl. 2021. 'Struggle to Settle down! Examining the Voices of Migrants and Refugees on Twitter Platform'. Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing, October 23, 95–98. <https://doi.org/10.1145/3462204.3481773>.
- Kim, Jisu, Alina Sirbu, Fosca Giannotti, and Lorenzo Gabrielli. 2020. 'Digital Footprints of International Migration on Twitter'. In *Advances in Intelligent Data Analysis XVIII*, edited by Michael R. Berthold, Ad Feelders, and Georg Krempel.
- Kim, Jisu, Alina Sirbu, Fosca Giannotti, Giulio Rossetti, and Hillel Rapoport. 2022. 'Origin and Destination Attachment: Study of Cultural Integration on Twitter'. *EPJ Data Science* 11 (1): 55.
- Kim, Jisu, Alina Sirbu, Giulio Rossetti, and Fosca Giannotti. 2022. 'Characterising Different Communities of Twitter Users: Migrants and Natives'. *Complex Networks & Their Applications X: Volume 1, Proceedings of the Tenth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021* 10, 130–41.
- Kim, Junghwan, and Mei-Po Kwan. 2021. 'An Examination of People's Privacy Concerns, Perceptions of Social Benefits, and Acceptance of COVID-19 Mitigation Measures That Harness Location Information: A Comparative Study of the U.S. and South Korea'. *ISPRS International Journal of Geo-Information* 10 (1): 25. <https://doi.org/10.3390/ijgi10010025>.
- King, Russell. 2012. 'Geography and Migration Studies: Retrospect and Prospect'. *Population, Space and Place* 18 (2): 134–53. <https://doi.org/10.1002/psp.685>.
- Kirwin, Matthew, and Anderson. 2018. Identifying the Factors Driving West African Migration. *West African Papers* No. 17. Vol. 17. West African Papers. OECD. <https://doi.org/10.1787/eb3b2806-en>.
- Kochupillai, Mrinalini, M Kahl, M Schmitt, H Taubenböck, and Xiao Xiang Zhu. 2022. 'Artificial Intelligence for Earth Observations: An Overview of Emerging Ethical Issues and Opportunities'. Under Review at *IEEE Geoscience and Remote Sensing Magazine*.
- Kochupillai, Mrinalini, and Hannes Taubenböck. 2023. 'Conducting Ethically Mindful Earth Observation Research: The Case of Slum Mapping'. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, July, 1937–40. <https://doi.org/10.1109/IGARSS52108.2023.10281725>.
- Kounadi, Ourania, and Bernd Resch. 2018. 'A Geoprivacy by Design Guideline for Research Campaigns That Use Participatory Sensing Data'. *Journal of Empirical Research on Human Research Ethics* 13 (3): 203–22. <https://doi.org/10.1177/1556264618759877>.
- Kraler, Albert, and David Reichel. 2022. 'Migration Statistics'. *Introduction to Migration Studies*, 439–62.
- Krieger, Gerhard, Alberto Moreira, Hauke Fiedler, et al. 2007. 'TanDEM-X: A Satellite Formation for High-Resolution SAR Interferometry'. *IEEE Transactions on Geoscience and Remote Sensing* 45 (11): 3317–41.
- Krotov, Vlad, Murray State University, Leiser Silva, and University of Houston. 2020. 'Legality and Ethics of Web Scraping'. *Communications of the Association for Information Systems* 47: 539–63. <https://doi.org/10.17705/1CAIS.04724>.

- Kruskal, William H., and W. Allen Wallis. 1952. 'Use of Ranks in One-Criterion Variance Analysis'. *Journal of the American Statistical Association* 47 (260): 583–621. <https://doi.org/10.1080/01621459.1952.10483441>.
- Kruspe, Anna, Matthias Häberle, Eike J Hoffmann, Samyo Rode-Hasinger, Karam Abdulahhad, and Xiao Xiang Zhu. 2021. 'Changes in Twitter Geolocations: Insights and Suggestions for Future Usage'. *Proceedings of the 2021 EMNLP Workshop W-NUT*, 212–21. <https://doi.org/10.48550/arXiv.2108.12251>.
- Kuhn, Max, and Kjell Johnson. 2019. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781315108230>.
- Kühnl, Marlene, Marta Sapena, Michael Wurm, Christian Geiß, and Hannes Taubenböck. 2023. 'Multitemporal Landslide Exposure and Vulnerability Assessment in Medellín, Colombia'. *Natural Hazards* 119 (2): 883–906. <https://doi.org/10.1007/s11069-022-05679-z>.
- Küpfer, Andreas. 2024. 'Nonrandom Tweet Mortality and Data Access Restrictions: Compromising the Replication of Sensitive Twitter Studies'. *Political Analysis* 32 (4): 493–506. <https://doi.org/10.1017/pan.2024.7>.
- Kuzu, Ridvan Salih, Thomas Brunschwiler, Gabriele Cavallaro, et al. 2025. 'FAST-EO: Multi-Modal Foundation Models for Scalable Earth Observation and Earth Sciences'. Paper presented at ESA-NASA International Workshop on AI Foundation Model for EO. May 5. <https://elib.dlr.de/212163/>.
- Kwan, Mei-Po, Irene Casas, and Ben Schmitz. 2004. 'Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks?' *Cartographica* 39 (2): 15–28. <https://doi.org/10.3138/X204-4223-57MK-8273>.
- Lakes, Tobia. 2019. 'Geodaten'. In *Handbuch Methoden der empirischen Sozialforschung*, edited by Nina Baur and Jörg Blasius. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21308-4_99.
- Lamanna, Fabio, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, and José J. Ramasco. 2018. 'Immigrant Community Integration in World Cities'. *PLOS ONE* 13 (3): e0191612. <https://doi.org/10.1371/journal.pone.0191612>.
- Lamidi, Idayat Modupe. 2016. Humour Markers and Their Interpretations in the Nairaland Virtual Community.
- Lang, Wei, Hao Lang, Eddie C. M. Hui, Tingting Chen, Jiemin Wu, and Marianne Jahre. 2022. 'Measuring Urban Vibrancy of Neighborhood Performance Using Social Media Data in Oslo, Norway'. *Cities* 131 (December): 103908. <https://doi.org/10.1016/j.cities.2022.103908>.
- Lansley, Guy, and Paul A. Longley. 2016. 'The Geography of Twitter Topics in London'. *Computers, Environment and Urban Systems* 58 (July): 85–96. <https://doi.org/10.1016/j.compenvurbsys.2016.04.002>.
- Lee, Chris van der, and Antal van den Bosch. 2017. 'Exploring Lexical and Syntactic Features for Language Variety Identification'. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, edited by Preslav Nakov, Marcos Zampieri, Nikola Ljubešić, Jörg Tiedemann, Shevin Malmasi, and Ahmed Ali. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-1224>.
- Lee, Everett S. 1966. 'A Theory of Migration'. *Demography* 3 (1): 47–57. <https://doi.org/10.2307/2060063>.
- Lee, June J. H. 2015. *World Migration Report 2015 - Chapter 4: Urban Migration and Economic Development*. International Organization for Migration, Geneva. <https://publications.iom.int/books/world-migration-report-2015-migrants-and-cities-new-partnerships-manage-mobility>.
- Lee, Ju-Sung, and Adina Nerghe. 2018. 'Refugee or Migrant Crisis? Labels, Perceived Agency, and Sentiment Polarity in Online Discussions'. *Social Media+ Society* 4 (3).
- Leetaru, Kalev Hannes. 2015. 'Mining Libraries: Lessons Learned from 20 Years of Massive Computing on the World's Information'. *Information Services & Use* 35 (1–2): 31–50. <https://doi.org/10.3233/ISU-150767>.
- Leetaru, Kalev, and Philip A. Schrodt. 2013. 'Gdelt: Global Data on Events, Location, and Tone, 1979–2012'. *ISA Annual Convention* 2 (4): 1–49.
- Leichter, A., D. Wittich, F. Rottensteiner, M. Werner, and M. Sester. 2018. 'Improved Classification of Satellite Imagery Using Spatial Feature Maps Extracted from Social Media'. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-4* (September): 335–42. <https://doi.org/10.5194/isprs-archives-XLII-4-335-2018>.
- Leichter, Artem, Dennis Wittich, Franz Rottensteiner, Martin Werner, and Monika Sester. 2018. Improved Classification of Satellite Imagery Using Spatial Feature Maps Extracted from Social Media. <https://repo.uni-hannover.de/handle/123456789/4105>.

- Lemoine-Rodríguez, Richard, Johannes Mast, Martin Mühlbauer, Nico Mandery, Carolin Biewer, and Hannes Taubenböck. 2024. 'The Voices of the Displaced: Mobility and Twitter Conversations of Migrants of Ukraine in 2022'. *Information Processing & Management* 61 (January): 103670. <https://doi.org/10.1016/j.ipm.2024.103670>.
- Levene, Howard. 1960. 'Robust Tests for Equality of Variances'. *Contributions to Probability and Statistics*, 278–92.
- Levon, Erez. 2015. 'Integrating Intersectionality in Language, Gender, and Sexuality Research'. *Language and Linguistics Compass* 9 (7): 295–308. <https://doi.org/10.1111/lnc3.12147>.
- Lewis, Benjamin, and Devika Kakkar. 2025. 'Harvard CGA Geotweet Archive v2.0'. Version 2.23. Harvard Dataverse, March 11. <https://doi.org/10.7910/DVN/3NCMB6>.
- Leydesdorff, Loet. 2005. 'Similarity Measures, Author Cocitation Analysis, and Information Theory'. *Journal of the American Society for Information Science and Technology* 56 (7): 769–72. <https://doi.org/10.1002/asi.20130>.
- Li, Fei, Tan Yigitcanlar, Madhav Nepal, Kien Nguyen, and Fatih Dur. 2023. 'Machine Learning and Remote Sensing Integration for Leveraging Urban Sustainability: A Review and Framework'. *Sustainable Cities and Society* 96 (September): 104653. <https://doi.org/10.1016/j.scs.2023.104653>.
- Li, Xiang, Congcong Wen, Yuan Hu, Zhenghang Yuan, and Xiao Xiang Zhu. 2024. 'Vision-Language Models in Remote Sensing: Current Progress and Future Trends'. *IEEE Geoscience and Remote Sensing Magazine* 12 (2): 32–66. <https://doi.org/10.1109/MGRS.2024.3383473>.
- Li, Yi, Weiming Huang, Gao Cong, Hao Wang, and Zheng Wang. 2023. 'Urban Region Representation Learning with OpenStreetMap Building Footprints'. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (New York, NY, USA), KDD '23, August 4, 1363–73*. <https://doi.org/10.1145/3580305.3599538>.
- Lin, Jianhua. 1991. 'Divergence Measures Based on the Shannon Entropy'. *IEEE Transactions on Information Theory* 37 (1): 145–51.
- Lloyd, Stuart. 1982. 'Least Squares Quantization in PCM'. *IEEE Transactions on Information Theory* 28 (2): 129–37.
- Lomborg, Stine, and Anja Bechmann. 2014. 'Using APIs for Data Collection on Social Media'. *The Information Society* 30 (4): 256–65. <https://doi.org/10.1080/01972243.2014.915276>.
- Longley, Paul A, Muhammad Adnan, and Guy Lansley. 2015. 'The Geotemporal Demographics of Twitter Usage'. *Environment and Planning A: Economy and Space* 47 (2): 465–84. <https://doi.org/10.1068/a130122p>.
- Luber, Stefan. 2021. 'Was ist GDELT?'. *BigData-Insider*, November 30. <https://www.bigdata-insider.de/was-ist-gdelt-a-1078771/>.
- Lüdeling, Anke, Stefan Evert, and Marco Baroni. 2007. 'Using Web Data for Linguistic Purposes'. In *Corpus Linguistics and the Web*. Brill. https://doi.org/10.1163/9789401203791_003.
- Ma, Yuchi, Shuo Chen, Stefano Ermon, and David B. Lobell. 2024. 'Transfer Learning in Environmental Remote Sensing'. *Remote Sensing of Environment* 301 (February): 113924. <https://doi.org/10.1016/j.rse.2023.113924>.
- Malik, Momin, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. 'Population Bias in Geotagged Tweets'. *Proceedings of the International AAAI Conference on Web and Social Media* 9 (4): 18–27. <https://ojs.aaai.org/index.php/ICWSM/article/view/14688>.
- Mamei, Marco, Nicola Bicocchi, Marco Lippi, Stefano Mariani, and Franco Zambonelli. 2019. 'Evaluating Origin–Destination Matrices Obtained from CDR Data'. *Sensors* 19 (20): 4470. <https://doi.org/10.3390/s19204470>.
- Manfredi, Stefano, and Slavomír Čéplö. 2023. 'A Corpus-Driven Description of o in Naijá (Nigerian Pidgin)'. *Journal of Pidgin and Creole Languages*, ahead of print, May. <https://doi.org/10.1075/jpcl.22016.man>.
- Marconcini, Mattia, Annekatrin Metz-Marconcini, Thomas Esch, and Noel Gorelick. 2021. 'Understanding Current Trends in Global Urbanisation - The World Settlement Footprint Suite'. *GI_Forum* 1: 33–38. https://doi.org/10.1553/giscience2021_01_s33.
- Marconcini, Mattia, Annekatrin Metz-Marconcini, Soner Üreyen, et al. 2020. 'Outlining Where Humans Live, the World Settlement Footprint 2015'. *Scientific Data* 7 (1): 1. <https://doi.org/10.1038/s41597-020-00580-5>.

- Martí, Pablo, Leticia Serrano-Estrada, and Almudena Nolasco-Cirugeda. 2019. 'Social Media Data: Challenges, Opportunities and Limitations in Urban Studies'. *Computers, Environment and Urban Systems* 74 (March): 161–74. <https://doi.org/10.1016/j.compenvurbsys.2018.11.001>.
- Martin, Philip. 2014. 'Economic Aspects of Migration'. In *Migration Theory*. Routledge. <https://api.taylorfrancis.com/content/chapters/edit/download?identifierName=doi&identifierValue=10.4324/9781315814933-4&type=chapterpdf>.
- Martiniello, Marco. 2022. 'Researching Arts, Culture, Migration and Change: A Multi (Trans)Disciplinary Challenge for International Migration Studies'. *Comparative Migration Studies* 10 (1): 7. <https://doi.org/10.1186/s40878-022-00281-5>.
- Mast, Johannes, Marta Sapena, Henri Debray, et al. 2025. 'The Digital Urban Frontier: Disparities in Social Media Activity between Consolidated and Newly Urbanized Areas in Africa'. *Applied Geography* 181 (August): 103687. <https://doi.org/10.1016/j.apgeog.2025.103687>.
- Mau, Steffen, Fabian Gülzau, Lena Laube, and Natascha Zaun. 2015. 'The Global Mobility Divide: How Visa Policies Have Evolved over Time'. *Journal of Ethnic and Migration Studies* 41 (8): 1192–213. <https://doi.org/10.1080/1369183X.2015.1005007>.
- Mazzoli, Mattia, Boris Diechtiareff, Antonia Tugores, et al. 2020. 'Migrant Mobility Flows Characterized with Digital Data'. *PLOS ONE* 15 (3): e0230264. <https://doi.org/10.1371/journal.pone.0230264>.
- McAuliffe, Marie. 2021. *International Migration and Digital Technology: An Overview*. Edward Elgar Publishing. <https://www.elgaronline.com/edcollchap/book/9781839100611/book-part-9781839100611-8.xml>.
- McAuliffe, Marie. 2024. *World Migration Report 2024*. Edited by International Organization for Migration. United Nations.
- McAuliffe, Marie, Luisa Feline Freier, Ronald Skeldon, and Jenna Blower. 2022. '5 The Great Disrupter: COVID-19's Impact on Migration, Mobility and Migrants Globally'. *World Migration Report 2022* (1): e00026.
- McAuliffe, Marie, Adrian Kitimbo, Alexandra M. Goossens, and Akm Ahsan Ullah. 2018. 'World Migration Report 2018 - Chapter 7 –Understanding Migration Journeys from Migrants' Perspectives'. *World Migration Report 2018* (1).
- McCreadie, Richard, Cody Buntain, and Ian Soboroff. 2019. *TREC Incident Streams: Finding Actionable Information on Social Media*.
- McInnes, Leland, John Healy, and Steve Astels. 2017. 'Hdbscan: Hierarchical Density Based Clustering.' *J. Open Source Softw.* 2 (11): 205.
- McInnes, Leland, John Healy, and James Melville. 2018. 'Umap: Uniform Manifold Approximation and Projection for Dimension Reduction'. *arXiv Preprint arXiv:1802.03426*.
- 'Measuring Digital Development: Facts and Figures 2024'. 2024. (Geneva), Measuring Digital Development. <https://digitallibrary.un.org/record/4074377>.
- Mekaoui, Salma, Ilham Chaker, Aarsalane Zarghili, and Nikola S. Nikolov. 2025. 'Systematic Literature Review of Topic Labeling'. *IEEE Access*. <https://ieeexplore.ieee.org/abstract/document/11015466/>.
- Merry, Krista, and Pete Bettinger. 2019. 'Smartphone GPS Accuracy Study in an Urban Environment'. *PLOS ONE* 14 (7): e0219890. <https://doi.org/10.1371/journal.pone.0219890>.
- Migration Data Portal. 2021. 'Bestand an internationalen Migrantinnen und Migranten'. February 5. <https://www.migrationdataportal.org/de/themes/bestand-internationalen-migrantinnen-und-migranten>.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. 'Distributed Representations of Words and Phrases and Their Compositionality'. *Advances in Neural Information Processing Systems* 26. <https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html>.
- Mimizuka, Kayo, Megan A. Brown, Kai-Cheng Yang, and Josephine Lukito. 2025. 'Post-Post-API Age: Studying Digital Platforms in Scant Data Access Times'. *arXiv:2505.09877*. Preprint, arXiv, May 15. <https://doi.org/10.48550/arXiv.2505.09877>.
- Moed, Henk F., and Gali Halevi. 2014. 'A Bibliometric Approach to Tracking International Scientific Migration'. *Scientometrics* 101 (3): 1987–2001. <https://doi.org/10.1007/s11192-014-1307-6>.
- Mukherjee, Prof Dr Joybrato. 2008. *Anglistische Korpuslinguistik: Eine Einführung - Grundlagen der Anglistik und Amerikanistik-33*. Erich Schmidt Verlag.
- Naija Lingo. 2023. 'Naijalingo: The Nigerian Pidgin English Dictionary Created by You for You!' <http://naijalingo.com/>.

- 'Nairaland Forum'. o.D. Zugriff am 5 May 2023. <https://www.nairaland.com/>.
- Nchake, Mamello A., and Mohammed Shuaibu. 2022. 'Investment in ICT Infrastructure and Inclusive Growth in Africa'. *Scientific African* 17 (September): e01293. <https://doi.org/10.1016/j.sciaf.2022.e01293>.
- Nelson, Okorie, Grace Loto, and Oladokun Omojola. 2018. 'Blogging, Civic Engagement, and Coverage of Political Conflict in Nigeria: A Study of Nairaland.Com'. *Kasetsart Journal of Social Sciences* 39 (2): 291–98. <https://doi.org/10.1016/j.kjss.2018.05.007>.
- Nesselhauf, Jonas. 2018. 'Migration Und Europa in Kulturwissenschaftlicher Perspektive - Zweite Jahrestagung Der Kulturwissenschaftlichen Gesellschaft (KWG), Universität Vechta, 6.–8.10.2016'. *Kulturwissenschaftliche Zeitschrift* 2017 (1): 94–96. <https://doi.org/10.2478/kwg-2017-0006>.
- Nguyen, Dat Quoc, Thanh Vu, and Anh Tuan Nguyen. 2020. 'BERTweet: A Pre-Trained Language Model for English Tweets'. *arXiv:2005.10200 [Cs]*, October 5. <http://arxiv.org/abs/2005.10200>.
- NOI Polls Ltd. 2015. US And UK Top List Of Countries Most Nigerians Abroad Reside In. <https://noi-polls.com/us-and-uk-top-list-of-countries-most-nigerians-abroad-reside-in-key-reason-for-migration-is-for-economic-opportunities/>.
- Nutakki, Brahmani, Rosa M. Navarrete, Giuseppe Carteny, and Ingmar Weber. 2025. 'Is There Anything Left?: A Global Analysis on Changes in Engagement with Political Content on Twitter in the Musk Era'. *Journal of Quantitative Description: Digital Media* 5 (Januar). <https://doi.org/10.51685/jqd.2025.004>.
- OECD, African Development Bank, Cities Alliance, and United Cities And Local Governments Of Africa. 2025. *Africa's Urbanisation Dynamics 2025: Planning for Urban Expansion*. West African Studies. OECD Publishing. <https://doi.org/10.1787/2a47845c-en>.
- Ofulue, C. I. 2012. 'Nigerian Pidgin and West African Pidgins: A Sociolinguistic Perspective'. *Legon Journal of the Humanities* 1: 1–42. <https://doi.org/10.4314/ljh.v1i0>.
- Ojedokun, Usman A, Yetunde O Ogunleye, and Adeyinka A Aderinto. 2021. 'Mass Mobilization for Police Accountability: The Case of Nigeria's #EndSARS Protest'. *Policing: A Journal of Policy and Practice* 15 (3): 1894–903. <https://doi.org/10.1093/policing/paab001>.
- Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. 'Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries'. *Frontiers in Big Data* 2 (July): 13. <https://doi.org/10.3389/fdata.2019.00013>.
- Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. 'CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises'. *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1): 376–85. <https://doi.org/10.1609/icwsm.v8i1.14538>.
- Omotuyi, Sunday. 2022. 'Franco-Nigerian Détente? Nigeria, France and the Francophone States of the Lake Chad Region in the Era of the Boko Haram Terrorism'. *African Identities* 20 (1): 26–41. <https://doi.org/10.1080/14725843.2020.1805296>.
- openAI. 2022. 'Introducing ChatGPT'. <https://openai.com/index/chatgpt/>.
- Openshaw, Stan. 1984. 'The Modifiable Areal Unit Problem'. *Concepts and Techniques in Modern Geography*. <https://cir.nii.ac.jp/crid/1570291225725496704>.
- Orabi, Mariam, Djedjiga Mouheb, Zaher Al Aghbari, and Ibrahim Kamel. 2020. 'Detection of Bots in Social Media: A Systematic Review'. *Information Processing & Management* 57 (4): 102250. <https://doi.org/10.1016/j.ipm.2020.102250>.
- Osborne, Larry N. 1998. 'HUMAN COMMUNICATIONS ISSUES Topic Development in USENET Newsgroups'. *Journal of the American Society for Information Science* (1986-1998) (New York, United States) 49 (11): 1010–16.
- Owuor, Innocensia, Hartwig H. Hochmair, and Sreten Cvetojevic. 2020. 'Tracking Hurricane Dorian in GDELT and Twitter'. *AGILE: GIScience Series* 1: 19.
- Oyebode, Oladapo, and Rita Orji. 2019a. 'Detecting Factors Responsible for Diabetes Prevalence in Nigeria Using Social Media and Machine Learning'. 2019 15th International Conference on Network and Service Management (CNSM), 1–4. <https://doi.org/10.23919/CNSM46954.2019.9012679>.
- Oyebode, Oladapo, and Rita Orji. 2019b. 'Social Media and Sentiment Analysis: The Nigeria Presidential Election 2019'. 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 0140–46. <https://doi.org/10.1109/IEMCON.2019.8936139>.

- Pappalardo, Luca, Leo Ferres, Manuel Sacasa, Ciro Cattuto, and Loreto Bravo. 2020. 'An Individual-Level Ground Truth Dataset for Home Location Detection'. arXiv:2010.08814. Preprint, arXiv, October 17. <https://doi.org/10.48550/arXiv.2010.08814>.
- Pappalardo, Luca, Ed Manley, Vedran Sekara, and Laura Alessandretti. 2023. 'Future Directions in Human Mobility Science'. *Nature Computational Science* 3 (7): 588–600. <https://doi.org/10.1038/s43588-023-00469-4>.
- Pappalardo, Luca, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. 2015. 'Returners and Explorers Dichotomy in Human Mobility'. *Nature Communications* 6 (1): 1. <https://doi.org/10.1038/ncomms9166>.
- Park, Hyunji Hayley, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. 'Morphology Matters: A Multilingual Language Modeling Analysis'. *Transactions of the Association for Computational Linguistics* 9 (March): 261–76. https://doi.org/10.1162/tacl_a_00365.
- Partridge, Virginia, Jasmine Mangat, Rebecca Curran, Ryan Mcgrady, and Ethan Zuckerman. 2024. 'Here Be Livestreams: Trade-Offs in Creating Temporal Maps of Reddit'. *ACM Web Science Conference*, May 21, 81–91. <https://doi.org/10.1145/3614419.3643999>.
- Patil, Rajvardhan, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. 'A Survey of Text Representation and Embedding Techniques in NLP'. *IEEE Access* 11: 36120–46. <https://doi.org/10.1109/ACCESS.2023.3266377>.
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. 'Glove: Global Vectors for Word Representation'. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43. <https://aclanthology.org/D14-1162.pdf>.
- Pesaresi, M., and P. Politis. 2023. 'GHS Built-up Surface Grid, Derived from Sentinel2 Composite and Landsat, Multitemporal (1975–2030). European Commission, Joint Research Centre (JRC). 10.2905'. D07D81B4-7680-4D28-B896-583745C27085. <https://doi.org/10.2905/9F06F36F-4B11-47EC-ABB0-4F8B7B1D72EA>.
- Pesarin, Fortunato, and Luigi Salmaso. 2010. 'The Permutation Testing Approach: A Review'. *Statistica* 70 (4): 481–509. <https://doi.org/10.6092/issn.1973-2201/3599>.
- Petutschnig, Andreas, Clemens Rudolf Havas, Bernd Resch, Veronika Krieger, and Cornelia Ferner. 2020. 'Exploratory Spatiotemporal Language Analysis of Geo-Social Network Data for Identifying Movements of Refugees'. *GI_Forum* 1: 137–52. https://doi.org/10.1553/giscience2020_01_s137.
- Pfeffer, Juergen, Daniel Matter, Kokil Jaidka, et al. 2023. 'Just Another Day on Twitter: A Complete 24 Hours of Twitter Data'. arXiv:2301.11429. Preprint, arXiv, April 11. <http://arxiv.org/abs/2301.11429>.
- Pfeffer, Jürgen, Angelina Mooseder, Jana Lasser, Luca Hammer, Oliver Stritzel, and David Garcia. 2023. 'This Sample Seems to Be Good Enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API'. *Proceedings of the International AAAI Conference on Web and Social Media* 17 (June): 720–29. <https://doi.org/10.1609/icwsm.v17i1.22182>.
- Potere, David, Annemarie Schneider, Shlomo Angel, and Daniel L. Civco. 2009. 'Mapping Urban Areas on a Global Scale: Which of the Eight Maps Now Available Is More Accurate?' *International Journal of Remote Sensing* 30 (24): 6531–58. <https://doi.org/10.1080/01431160903121134>.
- Princewill, Nimi, Stephanie Busari, and CNN. 2021. 'Nigeria Bans Twitter after Company Deletes President Buhari's Tweet'. CNN, June. <https://www.cnn.com/2021/06/04/africa/nigeria-suspends-twitter-operations-intl/index.html>.
- Qiao, Fengcai, Pei Li, Xin Zhang, Zhaoyun Ding, Jiajun Cheng, and Hui Wang. 2017. 'Predicting Social Unrest Events with Hidden Markov Models Using GDELT'. *Discrete Dynamics in Nature and Society* 2017: 1–13. <https://doi.org/10.1155/2017/8180272>.
- Quercia, Daniele, Harry Askham, and Jon Crowcroft. 2012. 'TweetLDA: Supervised Topic Classification and Link Prediction in Twitter'. *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12*, 247–50. <https://doi.org/10.1145/2380718.2380750>.
- R Core Team. 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Radivojevic, Kristina, Nicholas Clark, and Paul Brenner. 2024. 'Llms among Us: Generative Ai Participating in Digital Discourse'. *Proceedings of the AAAI Symposium Series* 3 (1): 209–18. <https://ojs.aaai.org/index.php/AAAI-SS/article/view/31202>.

- Rampazzo, Francesco, Jakub Bijak, Agnese Vitali, Ingmar Weber, and Emilio Zagheni. 2021. 'A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom'. *Demography* 58 (6): 2193–218. <https://doi.org/10.1215/00703370-9578562>.
- Rapoport, Hillel, Sulin Sardoschau, and Arthur Silve. 2021. *Migration and Cultural Change*.
- Ravenstein, Ernst Georg. 1876. *The Birthplaces of the People and the Laws of Migration*. Trübner. <https://books.google.de/books?hl=de&lr=&id=i0e7V1WeNaMC&oi=fnd&pg=PP9&dq=Ernst+Georg+Ravenstein&ots=4vBOeNbBbx&sig=FSxIP415e15aQzDhcg7IxzG2ffl>.
- Reimers, Nils, and Iryna Gurevych. 2019a. 'Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks'. arXiv Preprint arXiv:1908.10084.
- Reimers, Nils, and Iryna Gurevych. 2019b. 'Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks'. arXiv:1908.10084. Preprint, arXiv, August 27. <https://doi.org/10.48550/arXiv.1908.10084>.
- Reips, Ulf-Dietrich, and Laura E. Buffardi. 2012. 'Studying Migrants with the Help of the Internet: Methods from Psychology'. *Journal of Ethnic and Migration Studies* 38 (9): 1405–24. <https://doi.org/10.1080/1369183X.2012.698208>.
- Rentschler, Jun, Paolo Avner, Mattia Marconcini, et al. 2023. 'Global Evidence of Rapid Urban Growth in Flood Zones since 1985'. *Nature* 622 (7981): 7981. <https://doi.org/10.1038/s41586-023-06468-9>.
- Reuter, Christian, und Marc-André Kaufhold. 2018. 'Fifteen Years of Social Media in Emergencies: A Retrospective Review and Future Directions for Crisis Informatics'. *Journal of Contingencies and Crisis Management* 26 (1): 41–57. <https://doi.org/10.1111/1468-5973.12196>.
- Robertson, Colin, und Robert Feick. 2016. 'Bumps and Bruises in the Digital Skins of Cities: Unevenly Distributed User-Generated Content across US Urban Areas'. *Cartography and Geographic Information Science* 43 (4): 283–300. <https://doi.org/10.1080/15230406.2015.1088801>.
- Rowe, Francisco, Michael Mahony, Eduardo Graells-Garrido, Marzia Rango, and Niklas Sievers. 2021. 'Using Twitter to Track Immigration Sentiment during Early Stages of the COVID-19 Pandemic'. *Data & Policy* 3. <https://doi.org/10.1017/dap.2021.38>.
- Royston, J. P. 1982. 'An Extension of Shapiro and Wilk's W Test for Normality to Large Samples'. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31 (2): 115–24. <https://doi.org/10.2307/2347973>.
- Rudl, Tomas. 2025. 'Zugang für Forschung: So müssen Online-Dienste ihre Datensilos öffnen'. [netzpolitik.org](https://netzpolitik.org/2025/zugang-fuer-forschung-so-muessen-online-dienste-ihre-datensilos-oeffnen/), July 3. <https://netzpolitik.org/2025/zugang-fuer-forschung-so-muessen-online-dienste-ihre-datensilos-oeffnen/>.
- Russell, Stuart, und Peter Norvig. 2023. *Künstliche Intelligenz*. Pearson Deutschland. <https://elibrary.pearson.de/book/99.150005/9783863263263>.
- Ruths, Derek, und Caitrin Armstrong. 2020. 'The Residence History Inference Problem'. arXiv:2003.04155 [Cs], March 9. <http://arxiv.org/abs/2003.04155>.
- Sapena, Marta, Johannes Mast, Elisabeth Schoepfer, und Hannes Taubenböck. 2024. 'Do-It-Yourself Built-Up Mapping Tool: A Practical Cloud-Based Solution Using Sentinel Imagery for Mapping Urban Expansion in Africa'. SSRN Scholarly Paper No. 4762416. Rochester, NY, March 17. <https://doi.org/10.2139/ssrn.4762416>.
- Sapena, Marta, Luis Ruiz, und Hannes Taubenböck. 2020. 'Analyzing Links between Spatio-Temporal Metrics of Built-Up Areas and Socio-Economic Indicators on a Semi-Global Scale'. *ISPRS International Journal of Geo-Information* 9 (7): 436. <https://doi.org/10.3390/ijgi9070436>.
- Sapena, Marta, Michael Wurm, Hannes Taubenböck, Devis Tuia, und Luis A. Ruiz. 2021. 'Estimating Quality of Life Dimensions from Urban Spatial Pattern Metrics'. *Computers, Environment and Urban Systems* 85 (January): 101549. <https://doi.org/10.1016/j.compenvurbsys.2020.101549>.
- Saunders, Doug. 2010. *Arrival City: How the Largest Migration in History Is Reshaping Our World*. Random House.
- Schewel, Kerilyn. 2020. 'Understanding Immobility: Moving Beyond the Mobility Bias in Migration Studies'. *International Migration Review* 54 (2): 328–55. <https://doi.org/10.1177/0197918319831952>.
- Schmidt, Volker. 2013. *Räumliche Statistik*. https://www.mathematik.uni-ulm.de/stochastik/lehre/ss07/stochgeom/skript_alt.pdf.
- Schofield, Alexandra, Siqi Wu, Theo Bayard De Volo, Tatsuki Kuze, Alfredo Gomez, und Sharifa Sultana. 2025. "'My Very Subjective Human Interpretation": Domain Expert Perspectives on Navigating the Text

- Analysis Loop for Topic Models'. *Proceedings of the ACM on Human-Computer Interaction* 9 (1): 1–30. <https://doi.org/10.1145/3701201>.
- Scholten, Peter, Asya Pisarevskaya, und Nathan Levy. 2022. 'An Introduction to Migration Studies: The Rise and Coming of Age of a Research Field'. In *Introduction to Migration Studies*, edited by Peter Scholten. IMISCOE Research Series. Springer International Publishing. https://doi.org/10.1007/978-3-030-92377-8_1.
- Schumann, Guy J.-P., G. Robert Brakenridge, Albert J. Kettner, Rashid Kashif, und Emily Niebuhr. 2018. 'Assisting Flood Disaster Response with Earth Observation Data and Products: A Critical Assessment'. *Remote Sensing* 10 (8): 8. <https://doi.org/10.3390/rs10081230>.
- Schürmann, Alina, Janina Kleemann, Mike Teucher, und Christopher Conrad. 2024. 'Mapping Socio-Environmental Vulnerability to Assess Rural Migration in Ghana'. *Applied Geography* 167: 103283. <https://doi.org/10.1016/j.apgeog.2024.103283>.
- Schürmann, Alina, Janina Kleemann, Mike Teucher, Christine Fürst, und Christopher Conrad. 2022. 'Migration in West Africa: A Visual Analysis of Motivation, Causes, and Routes'. *Ecology and Society* 27 (3). <https://doi.org/10.5751/ES-13489-270316>.
- Schweter, Stefan, und Alan Akbik. 2021. 'FLERT: Document-Level Features for Named Entity Recognition'. arXiv:2011.06993. Preprint, arXiv, May 14. <http://arxiv.org/abs/2011.06993>.
- Şen, İpek. 2024. *A Decolonial Analysis of Urban and Regional Planning Education in the Global South*. November 4. <http://hdl.handle.net/11527/26452>.
- Senaratne, Hansi, Arne Bröring, Tobias Schreck, und Dominic Lehle. 2014. 'Moving on Twitter: Using Episodic Hotspot and Drift Analysis to Detect and Characterise Spatial Trajectories'. *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks - LBSN '14*, 23–30. <https://doi.org/10.1145/2755492.2755497>.
- Senaratne, Hansi, Martin Mühlbauer, Ralph Kiefl, et al. 2023. 'The Unseen—An Investigative Analysis of Thematic and Spatial Coverage of News on the Ongoing Refugee Crisis in West Africa'. *ISPRS International Journal of Geo-Information* 12 (4): 4. <https://doi.org/10.3390/ijgi12040175>.
- Serere, Helen Ngonidzashe, und Bernd Resch. 2024. 'Understanding the Impact of Geotagging on Location Inference Models for Accurate Generalization to Non-Geotagged Datasets'. *Geomatica* 76 (1): 100004. <https://doi.org/10.1016/j.geomat.2024.100004>.
- Seto, Karen C., Burak Güneralp, und Lucy R. Hutya. 2012. 'Global Forecasts of Urban Expansion to 2030 and Direct Impacts on Biodiversity and Carbon Pools'. *Proceedings of the National Academy of Sciences* 109 (40): 16083–88. <https://doi.org/10.1073/pnas.1211658109>.
- Shannon, Claude E. 1948. 'A Mathematical Theory of Communication'. *The Bell System Technical Journal* 27 (3): 379–423.
- Sîrbu, Alina, Gennady Andrienko, Natalia Andrienko, et al. 2021. 'Human Migration: The Big Data Perspective'. *International Journal of Data Science and Analytics* 11 (4): 341–60. <https://doi.org/10.1007/s41060-020-00213-5>.
- Smart Village. 2022. *About - Smart Village - About Us*. <https://www.smart-villages.com/about-smart-village/>.
- Smith, Paddy, Myles Gould, und Ed Manley. 2025. 'The Rise and Fall of Twitter Data in Urban Analytics'. Findings, ahead of print, March 26. <https://doi.org/10.32866/001c.133245>.
- Sparck Jones, Karen. 1972. 'A Statistical Interpretation of Term Specificity and Its Application in Retrieval'. *Journal of Documentation* 28 (1): 11–21.
- Spearman, C. 1904. 'The Proof and Measurement of Association between Two Things'. *The American Journal of Psychology* 15 (1): 72–101.
- Spyratos, Spyridon, Michele Vespe, Fabrizio Natale, Ingmar Weber, Emilio Zagheni, und Marzia Rango. 2018. *Migration Data Using Social Media: A European Perspective*. no. Luxembourg: Publications Office of the European Union. <https://hdl.handle.net/21.11116/0000-0004-7C55-2>.
- Spyratos, Spyridon, Michele Vespe, Fabrizio Natale, Ingmar Weber, Emilio Zagheni, und Marzia Rango. 2019. 'Quantifying International Human Mobility Patterns Using Facebook Network Data'. *PLOS ONE* 14 (10): e0224134. <https://doi.org/10.1371/journal.pone.0224134>.
- Stack Exchange Inc. 2024. 'Stack Exchange'. <https://stackoverflow.com/>.

- Stack Exchange Inc. 2024. 'Stack Overflow and OpenAI Partner to Strengthen the World's Most Popular Large Language Models - Press Release - Stack Overflow'.
<https://stackoverflow.co/company/press/archive/openai-partnership/>.
- State, Bogdan, Mario Rodriguez, Dirk Helbing, und Emilio Zagheni. 2014. 'Migration of Professionals to the U.S.' In *Social Informatics*, edited by Luca Maria Aiello and Daniel McFarland, vol. 8851. Lecture Notes in Computer Science. Springer International Publishing. https://doi.org/10.1007/978-3-319-13734-6_37.
- Steiger, Enrico, René Westerholt, Bernd Resch, und Alexander Zipf. 2015. 'Twitter as an Indicator for Whereabouts of People? Correlating Twitter with UK Census Data'. *Computers, Environment and Urban Systems* 54 (November): 255–65. <https://doi.org/10.1016/j.compenvurbsys.2015.09.007>.
- Stock, Kristin. 2018. 'Mining Location from Social Media: A Systematic Review'. *Computers, Environment and Urban Systems* 71 (September): 209–40. <https://doi.org/10.1016/j.compenvurbsys.2018.05.007>.
- Strübing, Jörg. 2019. 'Grounded Theory und Theoretical Sampling'. In *Handbuch Methoden der empirischen Sozialforschung*, edited by Nina Baur and Jörg Blasius. Springer Fachmedien Wiesbaden.
https://doi.org/10.1007/978-3-658-21308-4_36.
- Subrahmanian, V.S., Amos Azaria, Skylar Durst, et al. 2016. 'The DARPA Twitter Bot Challenge'. *Computer* 49 (6): 38–46. <https://doi.org/10.1109/MC.2016.183>.
- Sun, Na, Patrick Pei-Luen Rau, und Liang Ma. 2014. 'Understanding Lurkers in Online Communities: A Literature Review'. *Computers in Human Behavior* 38 (September): 110–17.
<https://doi.org/10.1016/j.chb.2014.05.022>.
- Tacoli, Cecilia, Gordon McGranahan, und David Satterthwaite. 2015. *Urbanisation, Rural-Urban Migration and Urban Poverty*. With International Institute for Environment and Development. IIED Working Paper, march 2015. IIED.
- Tally, Robert. 2012. *Spatiality*. Routledge.
<https://www.taylorfrancis.com/books/mono/10.4324/9780203082881/spatiality-robert-tally-jr-john-drakakis>.
- Tao, Yuan, Wanzeng Liu, Jun Chen, et al. 2025. 'A Graph-Based Multimodal Data Fusion Framework for Identifying Urban Functional Zone'. *International Journal of Applied Earth Observation and Geoinformation* 136 (February): 104353. <https://doi.org/10.1016/j.jag.2024.104353>.
- Taubenböck, H., J. Mast, C. Geiß, M. Wurm, T. Esch, und K. C. Seto. 2024. 'Global Differences in Urbanization Dynamics from 1985 to 2015 and Outlook Considering IPCC Climate Scenarios'. *Cities* 151 (August): 105117. <https://doi.org/10.1016/j.cities.2024.105117>.
- Taubenböck, H., J. Mast, R. Lemoine Rodríguez, H. Debray, M. Wurm, und C. Geiß. 2025. 'Was Global Urbanization from 1985 to 2015 Efficient in Terms of Land Consumption?' *Habitat International* 160 (June): 103397. <https://doi.org/10.1016/j.habitatint.2025.103397>.
- Taubenböck, H., C. Otto, F. Gülzau, und S. Mau. 2023. 'Border Regions across the Globe: Analyzing Border Typologies, Economic and Political Disparities, and Development Dynamics'. *Applied Geography* 151 (February): 102866. <https://doi.org/10.1016/j.apgeog.2022.102866>.
- Taubenböck, H., M. Weigand, T. Esch, et al. 2019. 'A New Ranking of the World's Largest Cities —Do Administrative Units Obscure Morphological Realities?' *Remote Sensing of Environment* 232 (October): 111353. <https://doi.org/10.1016/j.rse.2019.111353>.
- Taubenböck, Hannes, Henri Debray, Chunping Qiu, Michael Schmitt, Yuanyuan Wang, und Xiao Xiang Zhu. 2020. 'Seven City Types Representing Morphologic Configurations of Cities across the Globe'. *Cities* 105: 102814.
- Taubenböck, Hannes, Jeroen Staab, Xiao Zhu, Christian Geiß, Stefan Dech, und Michael Wurm. 2018. 'Are the Poor Digitally Left Behind? Indications of Urban Divides Based on Remote Sensing and Twitter Data'. *ISPRS International Journal of Geo-Information* 7 (8): 304. <https://doi.org/10.3390/ijgi7080304>.
- Taylor, Zack. 2013. 'Rethinking Planning Culture: A New Institutional Approach'. *Town Planning Review* 84 (6): 683–702. <https://doi.org/10.3828/tpv.2013.36>.
- Tear, Adrian. 2020. 'Geotagging Matters?: Second International Symposium on Platial Information Science'. *Proceedings of the 2nd International Symposium on Platial Information Science*, Januar 27, 61–72.
<https://doi.org/10.5281/zenodo.3628879>.
- Tjaden, Jasper. 2021. 'Measuring Migration 2.0: A Review of Digital Data Sources'. *Comparative Migration Studies* 9 (1): 59. <https://doi.org/10.1186/s40878-021-00273-x>.

- Tobaili, Taha. 2016. 'Arabizi Identification in Twitter Data'. Proceedings of the Acl 2016 Student Research Workshop, 51–57. <https://aclanthology.org/P16-3008.pdf>.
- Tobler, W. R. 1970. 'A Computer Movie Simulating Urban Growth in the Detroit Region'. *Economic Geography* 46: 234–40. <https://doi.org/10.2307/143141>.
- Tobler, Waldo. 1995. 'Migration: Ravenstein, Thorntwaite, and Beyond'. *Urban Geography - URBAN GEOGR* 16 (May): 327–43. <https://doi.org/10.2747/0272-3638.16.4.327>.
- Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. <https://www.torrossa.com/gs/resourceProxy?an=5002198&publisher=FZ4850>.
- Torch Contributors. 2019. 'CrossEntropyLoss — PyTorch 1.11.0 Documentation'. <https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html>.
- Törnberg, Petter. 2018. 'Echo Chambers and Viral Misinformation: Modeling Fake News as Complex Contagion'. *PLOS ONE* 13 (9): e0203958. <https://doi.org/10.1371/journal.pone.0203958>.
- Törnberg, Petter. 2023. 'The Urban Attention Economy: How Our Cities Are Rebuilt for Social Media'. *Mediapolis: A Journal of Cities and Culture* 8 (1). <https://www.mediapolisjournal.com/2023/03/urban-attention-economy/>.
- Touvron, Hugo, Louis Martin, Kevin Stone, et al. 2023. 'Llama 2: Open Foundation and Fine-Tuned Chat Models'. arXiv:2307.09288. Preprint, arXiv, Juli 19. <https://doi.org/10.48550/arXiv.2307.09288>.
- Townsend, Leanne, und Claire Wallace. 2016. 'Social Media Research: A Guide to Ethics'. University of Aberdeen 1: 16.
- Trask, Bahira Sherif. 2022. 'Migration, Urbanization, and the Family Dimension'. Division for Inclusive Social Development, Focal Point on the Family.
- Twitter Inc. 2022. 'Twitter API for Academic Research | Products'. <https://developer.twitter.com/en/products/twitter-api/academic-research>.
- UN DESA. 2019. *World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420)*. United Nations.
- UN DESA. 2024. *World Population Prospects 2024: Summary of Results (UN DESA/POP/2024/TR/NO. 9)*.
- United Nations Commission on Human Rights. 1998. *Guiding Principles on Internal Displacement*. United Nations Commission on Human Rights.
- United Nations Department of Economic and Social Affairs. 1998. *Recommendations on Statistics of International Migration*. no. Revision 1.
- United Nations Statistical Office. 1982. *Standard Country or Area Codes for Statistical Use*. UN.
- Unuabonah, Foluke O., und Rotimi O. Oladipupo. 2021. 'Bilingual Pragmatic Markers in Nigerian English'. *World Englishes* 40 (3): 390–406. <https://doi.org/10.1111/weng.12453>.
- Urry, John. 2007. *Mobilities*. Polity.
- Varghese, Bindiya M., Avittathur Unnikrishnan, und K. Jacob. 2013. 'Spatial Clustering Algorithms-an Overview'. *Asian Journal of Computer Science and Information Technology* 3 (1): 1–8.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. 'Attention Is All You Need'. In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, et al., vol. 30. Curran Associates, Inc.
- Völske, Michael, Martin Potthast, Shahbaz Syed, und Benno Stein. 2017. 'Tl; Dr: Mining Reddit to Learn Automatic Summarization'. Proceedings of the Workshop on New Frontiers in Summarization, 59–63.
- Wang, Jimin, und Yingjie Hu. 2020. *Enhancing Spatial and Textual Analysis with EUPEG: An Extensible and Unified Platform for Evaluating Geoparsers*. <https://doi.org/10.48550/arXiv.2007.04524>.
- Wang, Ruikun, Lei Ma, Guangjun He, et al. 2024. 'Transformers for Remote Sensing: A Systematic Review and Analysis'. *Sensors* 24 (11): 3495. <https://doi.org/10.3390/s24113495>.
- Wang, Zijian, Scott A. Hale, David Adelani, et al. 2019. 'Demographic Inference and Representative Population Estimates from Multilingual Social Media Data'. *The World Wide Web Conference*, Mai 13, 2056–67. <https://doi.org/10.1145/3308558.3313684>.
- Weckmüller, Dominik, Alexander Dunkel, und Dirk Burghardt. 2025. 'Embedding-Based Multilingual Semantic Search for Geo-Textual Data in Urban Studies'. *Journal of Geovisualization and Spatial Analysis* 9 (2): 31. <https://doi.org/10.1007/s41651-025-00232-5>.

- Wentrup, Robert, Patrik Ström, and H. Richard Nakamura. 2016. 'Digital Oases and Digital Deserts in Sub-Saharan Africa'. *Journal of Science & Technology Policy Management* 7 (1): 77–100. <https://doi.org/10.1108/JSTPM-03-2015-0013>.
- Weigand, Matthias, Simon Worbis, Marta Sapena, und Hannes Taubenböck. 2023. 'A Structural Catalogue of the Settlement Morphology in Refugee and IDP Camps'. *International Journal of Geographical Information Science* 37 (6): 1338–64. <https://doi.org/10.1080/13658816.2023.2189724>.
- Weigand, Matthias, Michael Wurm, Stefan Dech, und Hannes Taubenböck. 2019. 'Remote Sensing in Environmental Justice Research—A Review'. *ISPRS International Journal of Geo-Information* 8 (1): 1. <https://doi.org/10.3390/ijgi8010020>.
- Wieland, Marc, Sebastian Schmidt, Bernd Resch, Andreas Abecker, und Sandro Martinis. 2025. 'Fusion of Geospatial Information from Remote Sensing and Social Media to Prioritise Rapid Response Actions in Case of Floods'. *Natural Hazards*, ahead of print, Januar 22. <https://doi.org/10.1007/s11069-025-07120-7>.
- Willekens, Frans. 2019. 'Evidence-Based Monitoring of International Migration Flows in Europe'. *Journal of Official Statistics* 35 (1): 231–77. <https://doi.org/10.2478/jos-2019-0011>.
- Williams, Matthew L., Pete Burnap, und Luke Sloan. 2017. 'Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation'. *Sociology* 51 (6): 1149–68.
- Williams, Susan. 2020. Global Database of Events, Language und Tone (GDELT) Appendix. Office for National Statistics.
- Wilson, Robin, Elisabeth zu Erbach-Schoenberg, Maximilian Albert, et al. 2016. 'Rapid and Near Real-Time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake'. *PLoS Currents* 8 (February): ecurrents.dis.d073fbeece328e4c39087bc086d694b5c. <https://doi.org/10.1371/currents.dis.d073fbeece328e4c39087bc086d694b5c>.
- Wojcik, Stefan, und Adam Hughes. 2019. 'Sizing up Twitter Users'. *PEW Research Center* 24: 1–23.
- World Bank. 2022. 'Remittances Brave Global Headwinds. Special Focus: Climate Migration'. Knomad. https://www.knomad.org/sites/default/files/publication-doc/migration_and_development_brief_37_nov_2022.pdf.
- World Bank. 2015. 'WBG Topical Taxonomy'. Text. <https://Vocabularyserver.Com/Worldbank/Taxonomy/>, TemaTres, Januar 23. en. <https://vocabularyserver.com/worldbank/taxonomy/>.
- Wurm, Michael, Thomas Stark, Xiao Xiang Zhu, Matthias Weigand, und Hannes Taubenböck. 2019. 'Semantic Segmentation of Slums in Satellite Images Using Transfer Learning on Fully Convolutional Neural Networks'. *ISPRS Journal of Photogrammetry and Remote Sensing* 150: 59–69.
- Xing, Xiaoyue, Bailang Yu, Chaogui Kang, Bo Huang, Jianya Gong, und Yu Liu. 2024. 'The Synergy Between Remote Sensing and Social Sensing in Urban Studies: Review and Perspectives'. *IEEE Geoscience and Remote Sensing Magazine* 12 (1): 108–37. <https://doi.org/10.1109/MGRS.2023.3343968>.
- Xu, Dan, Peng Cui, Wenwu Zhu, und Shiqiang Yang. 2014. 'Graph-Based Residence Location Inference for Social Media Users'. *IEEE MultiMedia* 21 (4): 76–83. <https://doi.org/10.1109/MMUL.2014.62>.
- Yan, Yibo, Haomin Wen, Siru Zhong, et al. 2024. 'UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web'. *Proceedings of the ACM Web Conference 2024* (New York, NY, USA), WWW '24, Mai 13, 4006–17. <https://doi.org/10.1145/3589334.3645378>.
- Yang, Jun, Yu-Gang Jiang, Alexander G. Hauptmann, und Chong-Wah Ngo. 2007. 'Evaluating Bag-of-Visual-Words Representations in Scene Classification'. *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval* (New York, NY, USA), MIR '07, September 24, 197–206. <https://doi.org/10.1145/1290082.1290111>.
- Yin, Wenpeng, Katharina Kann, Mo Yu, und Hinrich Schütze. 2017. 'Comparative Study of CNN and RNN for Natural Language Processing'. *arXiv:1702.01923*. Preprint, arXiv, Februar 7. <https://doi.org/10.48550/arXiv.1702.01923>.
- Zagheni, Emilio, Venkata Rama Kiran Garimella, Ingmar Weber, und Bogdan State. 2014. 'Inferring International and Internal Migration Patterns from Twitter Data'. *Proceedings of the 23rd International Conference on World Wide Web*, April 7, 439–44. <https://doi.org/10.1145/2567948.2576930>.
- Zanaga, Daniele, Ruben Van De Kerchove, Dirk Daems, et al. 2022. *ESA WorldCover 10 m 2021 V200*. <https://pure.iiasa.ac.at/18478>.

- Zhang, Duzhen, Yahan Yu, Jiahua Dong, et al. 2024. 'MM-LLMs: Recent Advances in MultiModal Large Language Models'. In Findings of the Association for Computational Linguistics: ACL 2024, edited by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.738>.
- Zhang, Xinyang, Yury Malkov, Omar Florez, et al. 2023. 'TwHIN-BERT: A Socially-Enriched Pre-Trained Language Model for Multilingual Tweet Representations at Twitter'. arXiv:2209.07562. Preprint, arXiv, August 26. <https://doi.org/10.48550/arXiv.2209.07562>.
- Zhao, Fei, Chengcui Zhang, und Baocheng Geng. 2024. 'Deep Multimodal Data Fusion'. ACM Comput. Surv. 56 (9): 216:1-216:36. <https://doi.org/10.1145/3649447>.
- Zheng, Chenyu. 2020. 'Comparisons of the City Brand Influence of Global Cities: Word-Embedding Based Semantic Mining and Clustering Analysis on the Big Data of GDELT Global News Knowledge Graph'. Sustainability 12 (16): 6294.
- Zhu, Linhong, und Kristina Lerman. 2016. 'Attention Inequality in Social Media'. arXiv:1601.07200. Preprint, arXiv, Januar 26. <https://doi.org/10.48550/arXiv.1601.07200>.
- Zhu, Xiao Xiang, Yuanyuan Wang, Mrinalini Kochupillai, et al. 2022. 'Geoinformation Harvesting From Social Media Data: A Community Remote Sensing Approach'. IEEE Geoscience and Remote Sensing Magazine 10 (4): 150–80. <https://doi.org/10.1109/MGRS.2022.3219584>.
- Zubiaga, Arkaitz. 2018. 'A Longitudinal Assessment of the Persistence of Twitter Datasets'. Journal of the Association for Information Science and Technology 69 (8): 974–84. <https://doi.org/10.1002/asi.24026>.

Anhang

A1. Beschreibung des Nairaland-Datensatzes

Bezeichnung des Themas		Anzahl Kommentare vor & nach Filterung		Anteil der nicht- gefilterten Kommentare [%]	Beschreibung des Themas durch Nairaland
Deutsche Übersetzung	Original	Vor	Nach		
<i>Werbung</i>	<i>adverts</i>	86 757	14 567	16,79	Announcements, commercial posts, public notices.
<i>Landwirtschaft</i>	<i>agriculture</i>	86 792	49 148	56,63	The business and science of crop and livestock production, Agric Science, Agric Economics, etc
<i>Autos</i>	<i>autos</i>	121 529	15 660	12,89	Cars, motorbikes, airplanes, et cetera.
<i>B2B</i>	<i>b2b</i>	91 038	18 696	20,54	A Market For Corporate Suppliers And Buyers
<i>Geschäftliches</i>	<i>business</i>	78 761	48 155	61,14	Entrepreneurship, Startups, Economics, etc
<i>Karriere</i>	<i>career</i>	75 644	49 053	64,85	Talk about workplace experiences and incidents, professional training and certification, career management, etc. Employed, self-employed, job seekers
<i>Auto-Talk</i>	<i>cartalk</i>	134 358	96 445	71,78	Let's talk about cars here
<i>Prominente</i>	<i>celebs</i>	110 280	83 725	75,92	Celebrity obsession, Nollywood gossip, etc.
<i>Computer</i>	<i>computers</i>	73 149	29 132	39,83	Personal Computing, etc.
<i>Verbrechen</i>	<i>crime</i>	139 450	104 777	75,14	Crime
<i>Dating</i>	<i>dating</i>	128 087	96 460	75,31	A place to hook up with fellow Nairaland members and find love!
<i>Bildung</i>	<i>education</i>	83 397	59 010	70,76	Nairaland Students Forum: Primary and secondary school, universities, polytechnics, et cetera.
<i>Ereignisse</i>	<i>events</i>	85 429	57 521	67,33	Birthdays, Weddings, other Occasions. Planning, Announcement, Gifts.
<i>Familie</i>	<i>family</i>	172 946	117 388	67,88	Marriage/Family issues including husband to wife, parent to child, sibling and extended family relationships
<i>Mode</i>	<i>fashion</i>	73 636	41 745	56,69	Clothes, dresses, make-up routines, and modeling.
<i>Essen</i>	<i>food</i>	104 558	72 601	69,44	Delicious foods and how to prepare them. Anything about food or drink goes here. This is the Nairaland kitchen!
<i>Außenpolitik</i>	<i>foreign-affairs</i>	127 776	90 622	70,92	International Politics & other issues
<i>Gaming</i>	<i>gaming</i>	87 334	66 214	75,82	Welcome to the world of computer, Internet, video, and board games. Xbox, PS2, Chess, whatever!
<i>Graphik & Video</i>	<i>graphics-video</i>	82 509	49 507	60,00	Digital Video and Film, Computer Graphics and Animation. Tips, Tricks & Tools.

<i>Gesundheit</i>	<i>health</i>	84 347	55 068	65,29	Health related topics.
<i>Investment</i>	<i>investment</i>	75 155	42 955	57,16	For Investors. Nigerian Stock Exchange (NSE) Stocks, Bonds, T-bills, Real-Estate, etc.
<i>Islam</i>	<i>islam</i>	148 677	88 279	59,38	A forum for Muslims to share their faith. Muslims only please, even the moderators are Muslims.
<i>Jobs</i>	<i>jobs</i>	73 007	44 073	60,37	Job/Employment Opportunities; Vacancies In Nigeria!
<i>Witze</i>	<i>jokes</i>	60 065	41 356	68,85	Riddles and jokes that fellow Nigerians can appreciate.
<i>Musik & Radio</i>	<i>music-radio</i>	32 028	20 771	64,85	Talk about your favorite songs, music albums, artists and bands.
<i>NYSC</i>	<i>nysc</i>	109 915	82 392	74,96	National Youth Service Corps (NYSC) Discussions. Corpers, etc.
<i>Haustiere</i>	<i>pets</i>	100 667	69 952	69,49	Discuss dogs/puppies, cats/kittens, and other pets. Buy and sell them
<i>Handy-Anzeigen</i>	<i>phone-ads</i>	104 508	45 427	43,47	Phones & Accessories For Sale / Adverts
<i>Mobiltelefone</i>	<i>phones</i>	111 384	87 173	78,26	Nigerian GSM networks, telephone companies, et cetera. ISPs, Modems, Websites, etc
<i>Politik</i>	<i>politics</i>	153 258	112 175	73,19	Our country Nigeria is the giant of Africa!
<i>Programmieren</i>	<i>programming</i>	68 543	45 536	66,43	Software programming, development of applications..
<i>Immobilien</i>	<i>properties</i>	97 548	16 299	16,71	Real Estate: Land, Houses, Flats, Etc.
<i>Religion</i>	<i>religion</i>	131 831	83 338	63,22	Share your faith and belief in God or higher powers here.
<i>Romantik</i>	<i>romance</i>	174 977	134 706	76,98	Discuss dating, courtship, and romance in marriage.
<i>Wissenschaft</i>	<i>science</i>	67 563	42 946	63,56	Science/Technology
<i>Sport</i>	<i>sports</i>	84 581	61 749	73,01	We are Nigerians; and we love soccer and many other sports!
<i>Technologiemarkt</i>	<i>techmarket</i>	108 384	28 879	26,65	Buy and sell Phones, Computers and PC accessories here.
<i>Reisen</i>	<i>travel</i>	166 959	123 976	74,26	Tourism, travel. Interesting destinations within Nigeria and abroad. And motoring!
<i>TV & Filme</i>	<i>tv-movies</i>	94 081	69 536	73,91	Local television programmes, local and foreign movies available in Nigeria.
<i>Webmasters</i>	<i>webmasters</i>	55 470	33 330	60,09	Website design and development, management of forums, blogs, wikis, and all sorts of websites.

Tabelle A1: Beschreibung der Themen des Nairaland-Datensatzes

A2. Vergleich von Divergenzmetriken

Verglichen mit Divergenzmetriken wie der Kullback-Leibler Divergenz oder Jensen-Shannon Divergenz (Lin, 1991, grün in untenliegender Abbildung), erfasst der DDS (blau in Abbildung unten) nicht nur die Stärke der Disparität, sondern auch deren Richtung, also ob die Dichte in älteren oder neueren Stadtgebieten höher ist. Verglichen mit einer linearen Regression zwischen Bebauungsalter und Dichte (rot in der Abbildung unten) bietet der DDS den Vorteil

von Invarianz gegenüber der Länge der Inputsequenz, ist aber ansonsten dem Korrelationskoeffizienten einer linearen Regression stark ähnlich und für Sequenzen gleicher Länge tatsächlich perfekt korreliert mit diesem.

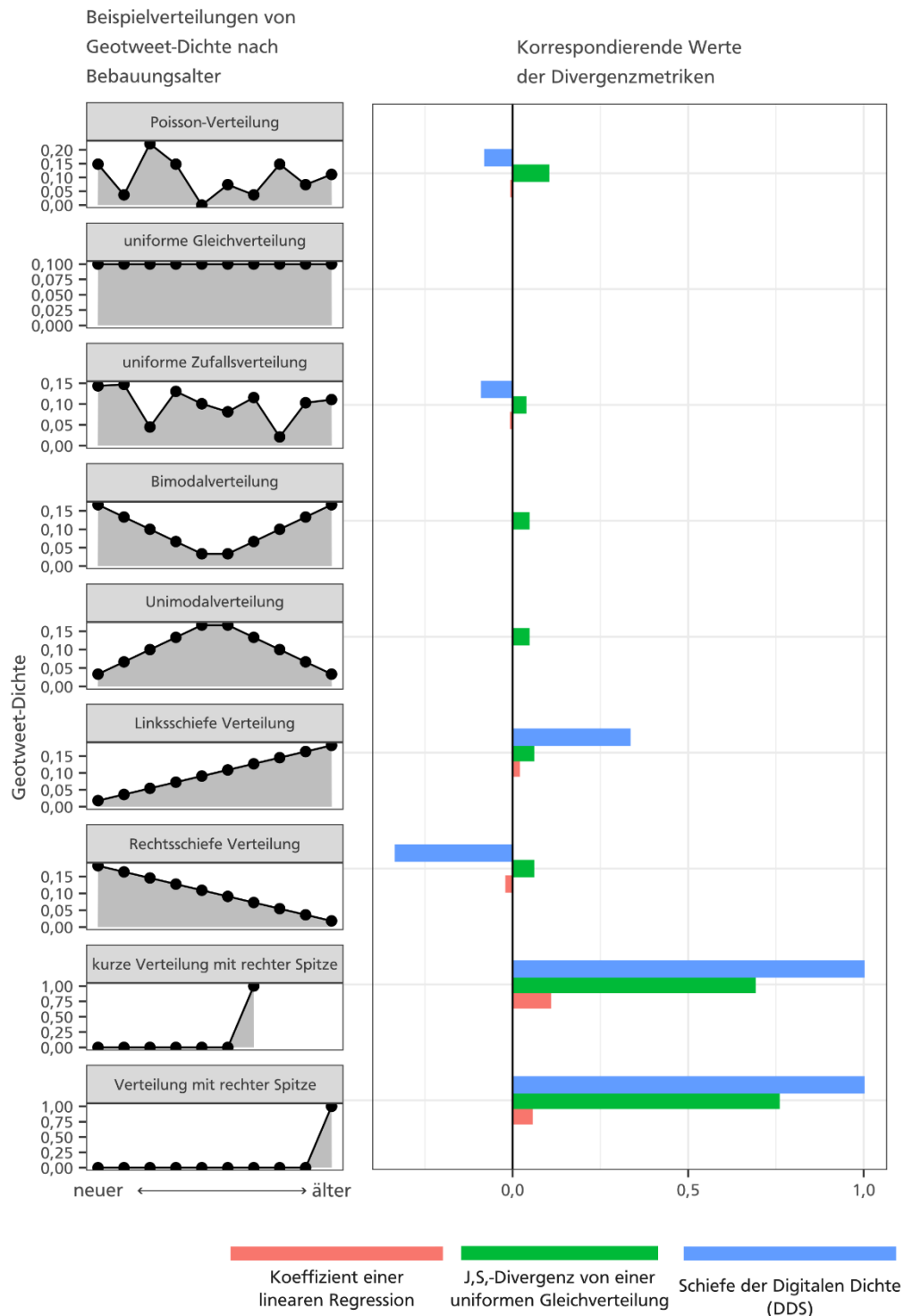


Abbildung A2: Vergleich von DDS mit anderen Metriken, welche in Betracht gezogen wurden. Links: Hypothetische Verteilungen von Geotweet-Dichte über Bebauungsalter. Rechts: Die korrespondierenden Werte dreier verschiedener Metriken für diese Verteilungen. Diese Metriken sind A: Koeffizienten eines linearen Regressionsmodells; B: Divergenzmaß (Jenssen-Shannon-Divergenz von einer Gleichverteilung); C: Die in dieser

A3. Robustheitstests in Studie A1

a) Robustheit gegenüber GPS-Ungenauigkeiten

Ungenauigkeiten von GPS-Systemen auf den Geräten der Nutzenden könnten zu den erfassten Disparitäten beitragen, da neuere Gebiete typischerweise am Rand der Städte liegen. Randgebiete, welche an Flächen ohne Geotweets grenzen, tendieren dazu, durch Ungenauigkeiten Geotweets an diese Flächen zu verlieren. Kernzonen verlieren auch Geotweets, tendieren jedoch dazu, in gleichem Maße auch Geotweets zu gewinnen, wie untenliegende Abbildung veranschaulicht.

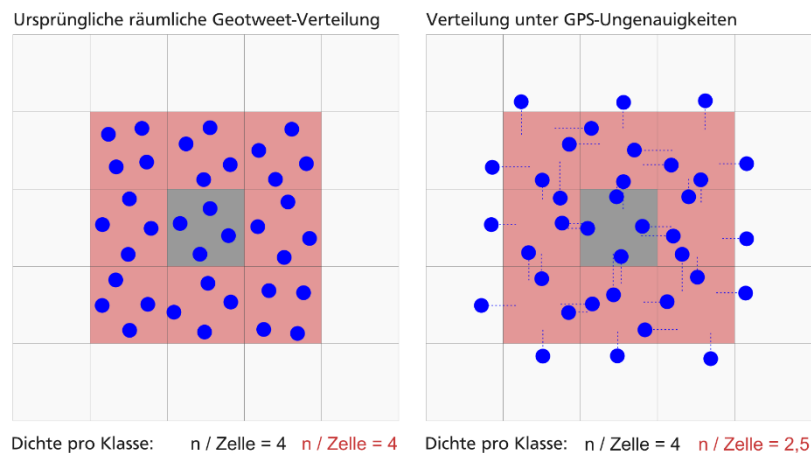


Abbildung A3a: Illustration des hypothetischen Randeffektes, bei welchem Ungenauigkeiten zu Dichteunterschieden führen könnten.

Obwohl die wahre Dichte in Kernzonen (grau) und Randzonen (rot) mit 4 Geotweets pro Zelle gleich hoch ist, ergibt sich durch die Verschiebung von Geotweets in angrenzende Zellen eine Disparität in der Dichte (4 Geotweets pro Zelle in Kernzonen, 2,5 Geotweets pro Zelle in Randzonen). Es soll daher getestet werden, ob die in Studie A1 gemessene Disparität durch GPS-Ungenauigkeiten verursacht sein könnte.

Hierfür wird die Disparität (DDS) unter verschiedenen simulierten Ungenauigkeiten neu berechnet. Zur Simulation der Ungenauigkeiten werden die gemessenen Geotweet-Positionen zufällig unter verschiedenen Ungenauigkeitsstärken verschoben, welche durch einen mittleren Fehler von 10, 30, 100, 200, und 655 (der höchsten in der Fachliteratur gemessenen Ungenauigkeit von GPS-Geräten, siehe Merry & Bettinger, 2019). Die unter diesen zusätzlichen Ungenauigkeiten neu berechneten Disparitäten für alle UTG werden dann in Bezug zu den ursprünglich berechneten Ungenauigkeiten gesetzt (siehe Abbildung unten).

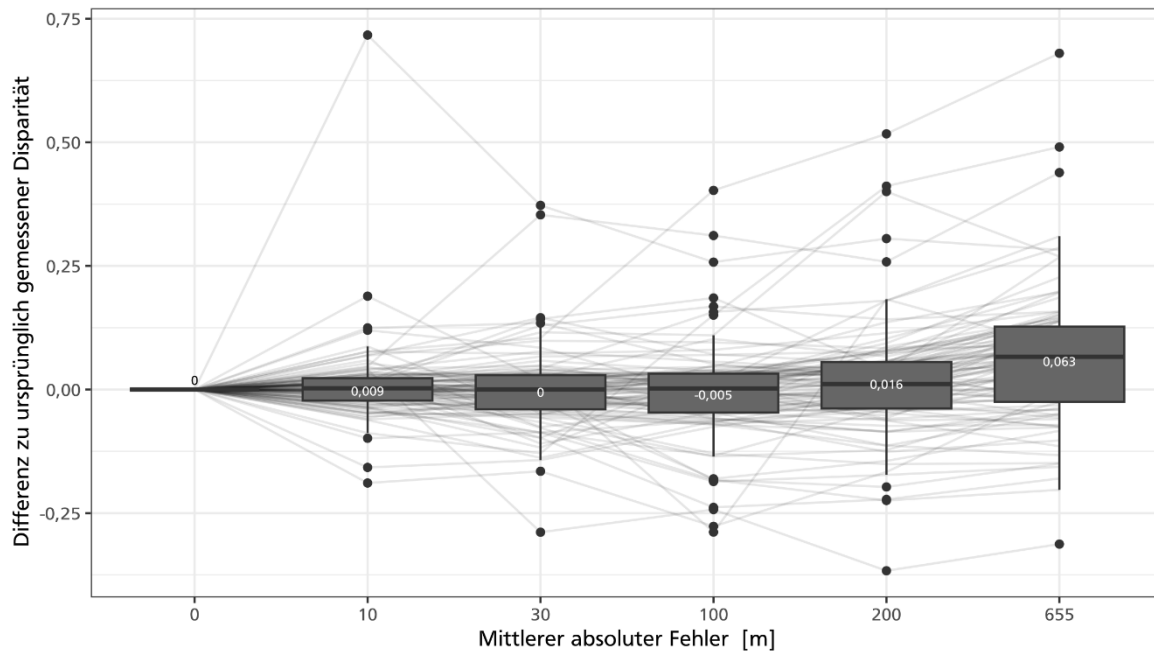


Abbildung A3b: Auswirkung von simulierten GPS-ungenauigkeiten auf gemessene Disparität (DDS_{UTG}).

Wie erwartet erhöhen sich die berechneten Disparitäten, wenn die Ungenauigkeiten ein sehr hohes Niveau erreichen. Ein mittlerer GPS-ungenauigkeit von 10m führt zu sehr kleinen Unterschieden, ein Ungenauigkeit von 100m führt zu einer Senkung des DDS um durchschnittlich -0,005. Ungenauigkeiten von 655m sind sehr einflussreich und erhöhen die gemessene Disparität im Durchschnitt um 0,063. Im Vergleich mit der in der Studie gemessenen DDS von 0,206 ist eine mittlere Überschätzung von 0,063 (31% von 0,206) ein bemerkenswert hoher Fehler, der die Analyse maßgeblich beeinträchtigen könnte, auch wenn die übergreifenden Schlussfolgerungen selbst durch diese extremen Ungenauigkeiten nicht grundsätzlich in Frage gezogen werden.

Solch hohe Ungenauigkeiten würden jedoch, wenn überhaupt, nur zu Beginn des Analysezeitraums auftreten, als die GPS-Genauigkeit der Geräte am niedrigsten war, und stellt selbst dann nur ein worst-case Szenario dar. Eine realistischere Ungenauigkeit von im Mittel 30 Metern würde den DDS nicht merklich ändern.

Zudem wurden im Rahmen von Studie A1 nur wenige Geotweets außerhalb der bebauten Flächen beobachtet (welche möglicherweise durch die Randgebiete verloren wurden), weswegen es wahrscheinlich ist, dass der Beitrag der GPS-ungenauigkeiten zur gemessenen Disparität geringer ist als das oben beschriebene Maximum.

Zusammenfassend bestätigen die Tests einen Einfluss der GPS-ungenauigkeiten auf die gemessene Disparität, doch selbst unter Annahme extremer Ungenauigkeiten würden diese nur etwa 31% der gemessenen Ungenauigkeit erklären. Aus diesem Grund muss eingeräumt werden, dass die Disparität zu einem ungewissen, aber vermutlich sehr geringen, Grad überschätzt wird, jedoch nicht in einem Ausmaß, welches die Schlussfolgerungen der Studie in Frage stellt.

Robustheit gegenüber geringer Datenverfügbarkeit

Eine geringe Anzahl von Geotweets für ein UTG und Jahr führt vermutlich zu weniger robusten Schätzungen der Disparität. Um den Effekt Geringer Datenverfügbarkeit auf den DDS zu testen, wird ein bootstrapping-Ansatz genutzt. Für Anteile von 1%, 2%, 3%, ... bis zu 99% werden 1000 unterschiedliche zufällige Stichproben aus den gesammelten Geotweets genommen und auf Basis dieser Stichproben jeweils der DDS berechnet. Die Verteilung des DDS über diese Stichproben ermöglicht einen Einblick in die Robustheit der Disparität gegenüber unterschiedlich geringeren Datenverfügbarkeiten.

Die untenliegende Abbildung zeigt diesen Effekt für einige ausgewählte UTG verschiedener Größenordnungen. Die Boxplots zeigen die Verteilung der Disparität über die 1000 Permutationen. Sichtbar ist, wie sich die gemessene Disparität stabilisiert, wenn der Anteil sich den 100% einer vollen Inklusion aller Daten annähert.

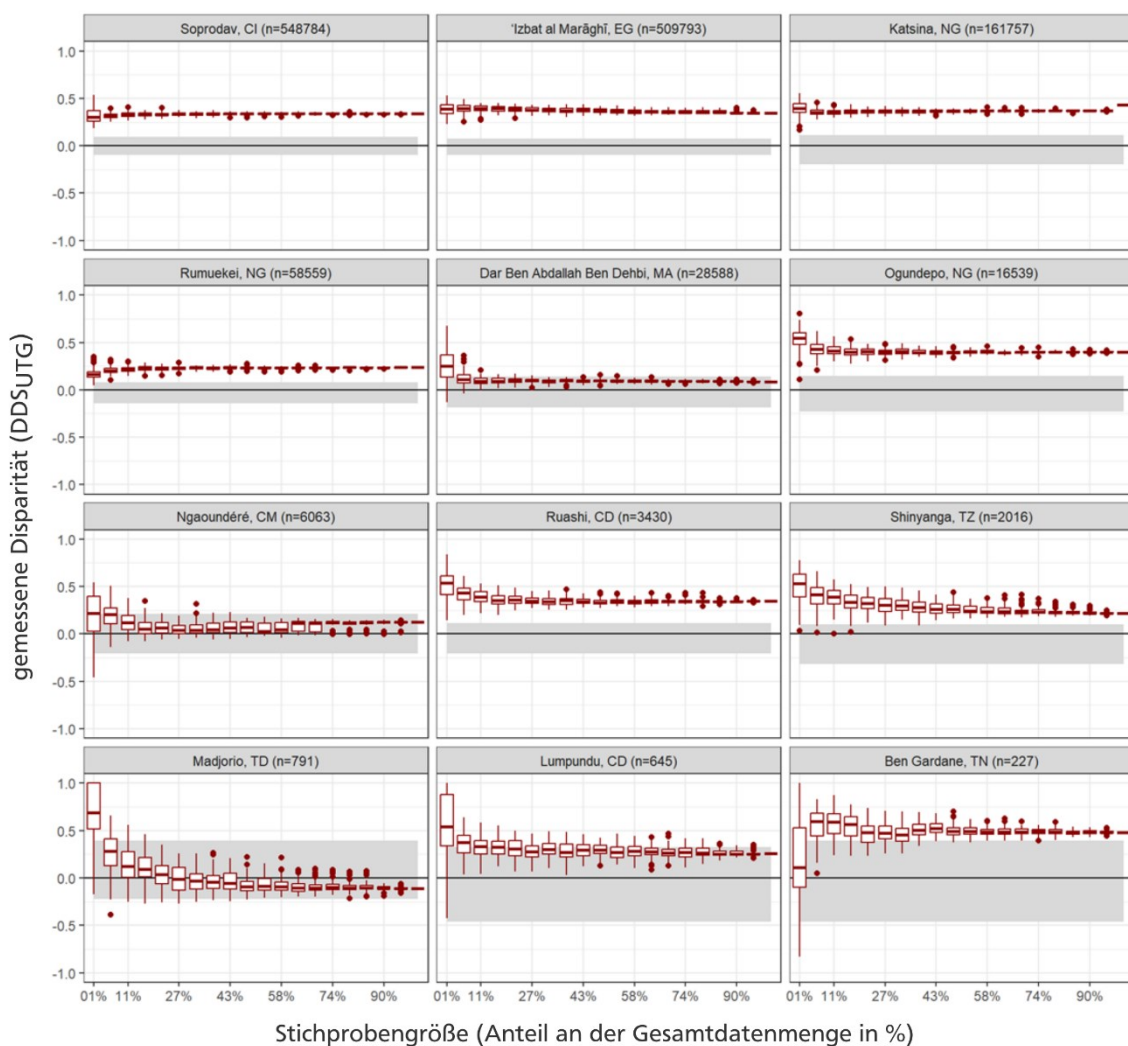


Abbildung A3c: Auswirkung von Stichprobengröße auf geschätzte Disparität (DDS_{UTG}), In grau: 95% Signifikanzbereich des in der Studie A1 durchgeführten Permutationstests.

Die Ergebnisse zeigen, dass für UTG mit einer hohen Anzahl von Tweets bereits eine kleine Stichprobe bereits zu einem stabilen Ergebnis führt (obere Reihe). Für UTG mit wenigen Tweets (untere Reihe) konvergieren die Ergebnisse deutlich langsamer. In wenigen Fällen finden sich sogar Fälle, bei welchem das Vorzeichen der Disparität sich ändert. Doch in Bezug

auf das Signifikanzniveau wird ein stabiles Ergebnis bei bereits 30% der einbezogenen Daten erreicht, und dies gilt auch für UTG mit sehr wenigen Geotweets. Dies belegt, dass selbst mit kleineren Geotweet-Mengen belastbare Ergebnisse erzielt werden können. Selbstverständlich stellen sich Fragen von Repräsentativität, und es ist zu bevorzugen, Ergebnisse für UTG mit wenigen Geotweets nur sehr vorsichtig zu interpretieren.

Robustheit gegenüber sehr aktiven Accounts

Nicht alle Accounts tragen in gleichem Maße zu den in Studie A1 genutzten Datensätzen bei. Im Gegenteil: The Der Gini-Koeffizient der Geotweet-Beiträge liegt mit 0,819 bemerkenswert hoch. Dies legt nahe, dass die Mehrheit von Tweets von einer kleinen Gruppe sehr aktiver Accounts stammt.

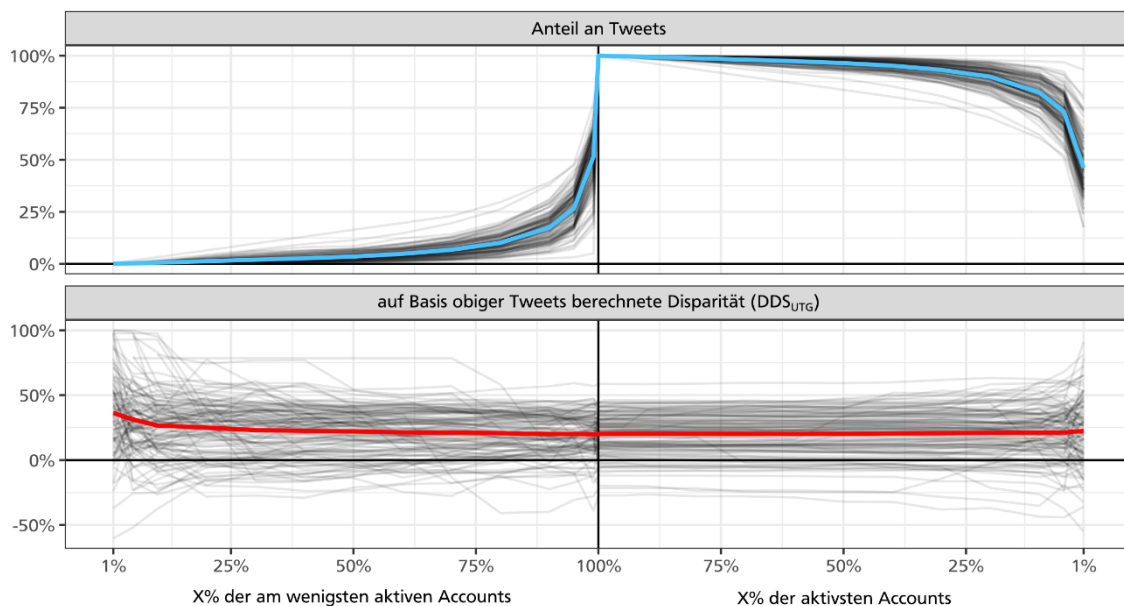


Abbildung A3d: Auswirkung der Inklusion/Exklusion von besonders aktiven Twitter-Accounts auf die Vollständigkeit des Datensatzes (oben) und die daraus berechnete Disparität (unten). Jede schwarze Linie entspricht einem UTG. In blau bzw. rot das arithmetische Mittel aller UTGs.

Die obenliegende Abbildung zeigt den Effekt besonders aktiver Accounts auf die gemessene Disparität (DDS). Oben: P-P Plot der kumulativen Geotweet-Anzahl pro Anteil der Accounts. Unten: Korrespondierende Disparität (DDS) auf Basis der Tweets dieser Accounts.

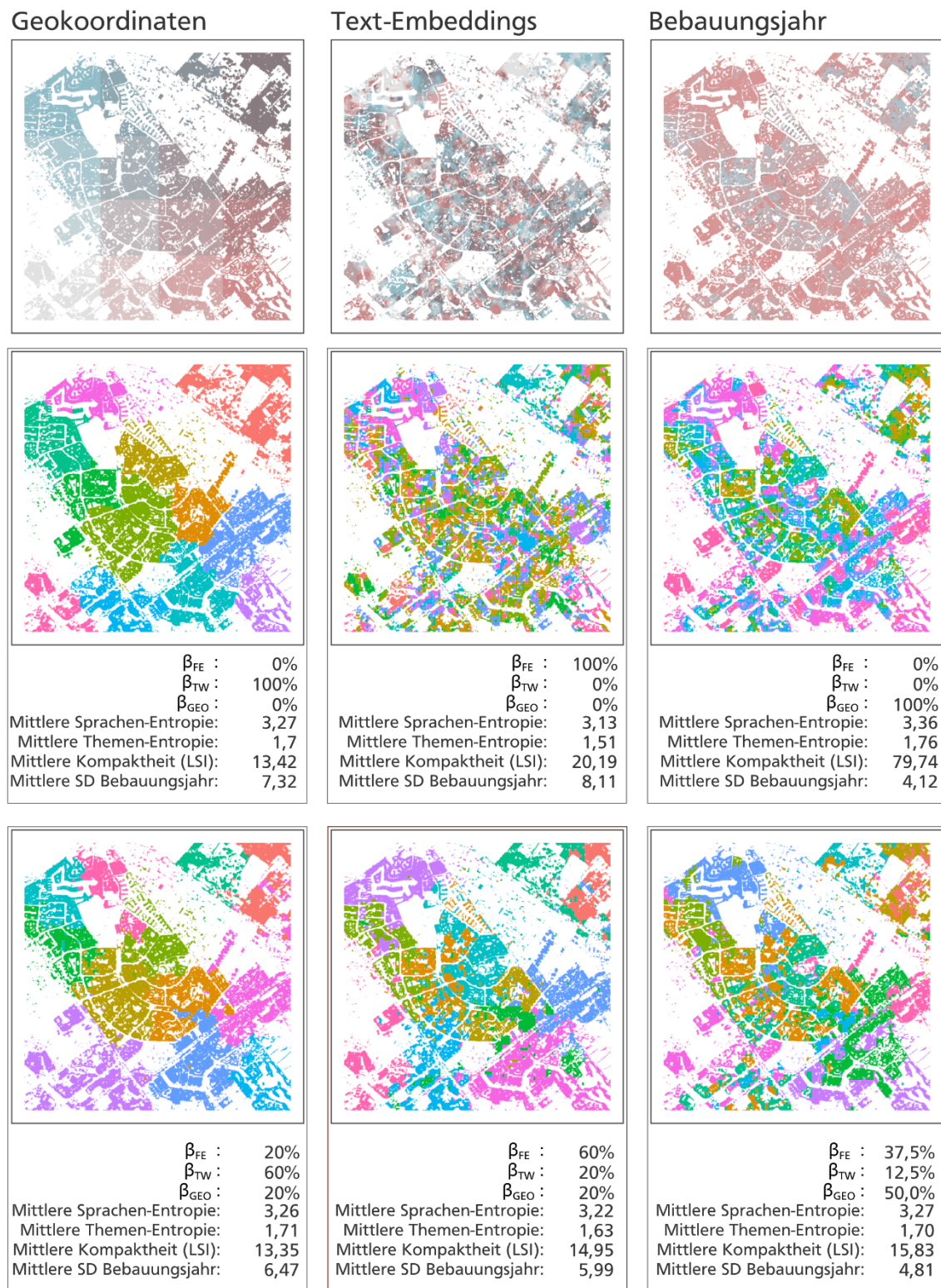
Klar zeigt sich, dass der Ausschluss der hochaktiven Accounts die Anzahl an Tweets stark beeinflusst. Jedoch ändert sich die gemessene Disparität nur sehr leicht. Fokussiert man die aktivsten 1% oder 10% ist die Disparität mit 0,225 beziehungsweise 0,211 leicht höher als die 0,201, welche sich bei Einschluss aller Accounts ergeben. Bei Ausschluss der 10% und 1% hochaktiven Accounts ergibt sich eine leicht geringere mittlere Disparität von 0,200 und 0,199. Dies legt nahe, dass die hochaktiven Accounts in ihrem UTG-internen Posting-verhalten nur eine leicht stärkere Neigung dazu haben, disproportional mehr Geotweets von älteren Gebieten zu posten. Interessanterweise ist die Disparität mit 0,363 und 0,266 für die geringstaktiven 1% und 10% merkbar höher. Aufgrund der sehr geringen Anzahl von Tweets dieser Accounts ist dies jedoch keine belastbare Messung.

d) Zusammenfassung der Robustheitstests

Die Simulationen der GPS-Ungenauigkeiten zeigten eine positive Beziehung mit der gemessenen Disparität, welche nur bei extremen Ungenauigkeiten (>30m mittlere Ungenauigkeit) merklich ist. Die Simulation eines Fehlers von 100 Metern erhöhte die Disparität um etwa 8% DDS. Im Umkehrschluss legt dies nahe, dass etwa 8% der gemessenen Disparität auf diese Art von Ungenauigkeit zurückgeführt werden kann, was die Ergebnisse von Studie A1 nicht entscheidend beeinflusst. Schlussfolgernd sind die in Studie A1 berichteten Ergebnisse robust gegenüber realistischen GPS-Ungenauigkeiten.

Bezüglich des Einflusses besonders aktiver Accounts zeigt sich, dass diese einen großen Teil der Daten ausmachen, aber dies die gemessene Disparität nicht entscheidend beeinflusst. Ein Ausschluss der 10% oder 1% aktivsten Accounts ändert den gemessenen mittleren DDS über alle UTG hinweg von 0,201 auf 0,200 beziehungsweise 0,199. Es lässt sich schlussfolgern, dass diese hochaktiven Accounts sich in Bezug auf die Studie A1 nicht entscheidend von den weniger aktiven Accounts unterscheiden.

A4. Variationen des in Studie 2A demonstrierten Clusterings



In der Studie A2 genutzte Gewichtung

Abbildung A4: Raumunterteilungen eines UTG, welche sich bei unterschiedlichen Gewichtungen im Clusteringverfahren ergeben (siehe Studie A2). Obere Reihe: Die Input Daten Geokoordinaten (links), dimensionsreduzierte Text-Embeddings (Mitte) und Bebauungsjahr (rechts). Mittlere Reihe: Raumunterteilung basierend auf ausschließlich einem dieser Inputs. Untere Reihe: Gemischte Raumunterteilungen.

A5. Ableitung von Raumeinheiten für Studie B1

Da die kleinräumige Mobilität in der Studie nicht von Interesse ist, werden die Tweets von S_{GEO} für die Mobilitätsanalyse zu räumlichen Einheiten (RE) aggregiert, die auf dem Schwerpunkt ihres getaggtten Ortes basieren. Dies dient auch der weiteren Anonymisierung der Daten. Da die Mobilität innerhalb anderer Länder für die Analyse der nigerianischen Migrationsmuster von geringerem Interesse für die Studie ist, werden die Tweets von außerhalb Nigerias auf Länderebene aggregiert. Innerhalb Nigerias hingegen werden die Tweets in einem sechseckigen Raster der Zellengröße z aggregiert. Ein Raster hat den Vorteil der Regelmäßigkeit und der reduzierten Normativität, aber es respektiert nicht die Integrität funktional verbundener städtischer Gebiete, durch die Grenzen von Rasterzellen aufgeteilt sein könnten. Um dies zu verhindern, werden die SU um städtische Gebiete wie folgt abgegrenzt:

Nach der Identifizierung signifikanter Zentren mit Hilfe eines mehrstufigen Ansatzes, der (1) die Bekanntheit, (2) die Aktivität im Twitter-Datensatz, (3) die politische Bedeutung und (4) die physische Größe berücksichtigt, wird jedem dieser Zentren eine eigene RE zugewiesen. Da a priori keine Informationen zur Berechnung der typischen Reisezeit-Isochronen auf der Grundlage der Pendlerentfernungen vorlagen, werden kreisförmige RE mit dem Radius r um die Zentren abgegrenzt und die Sensitivität der Mobilitätsanalyse in Bezug auf r und z getestet (Abbildung unten).

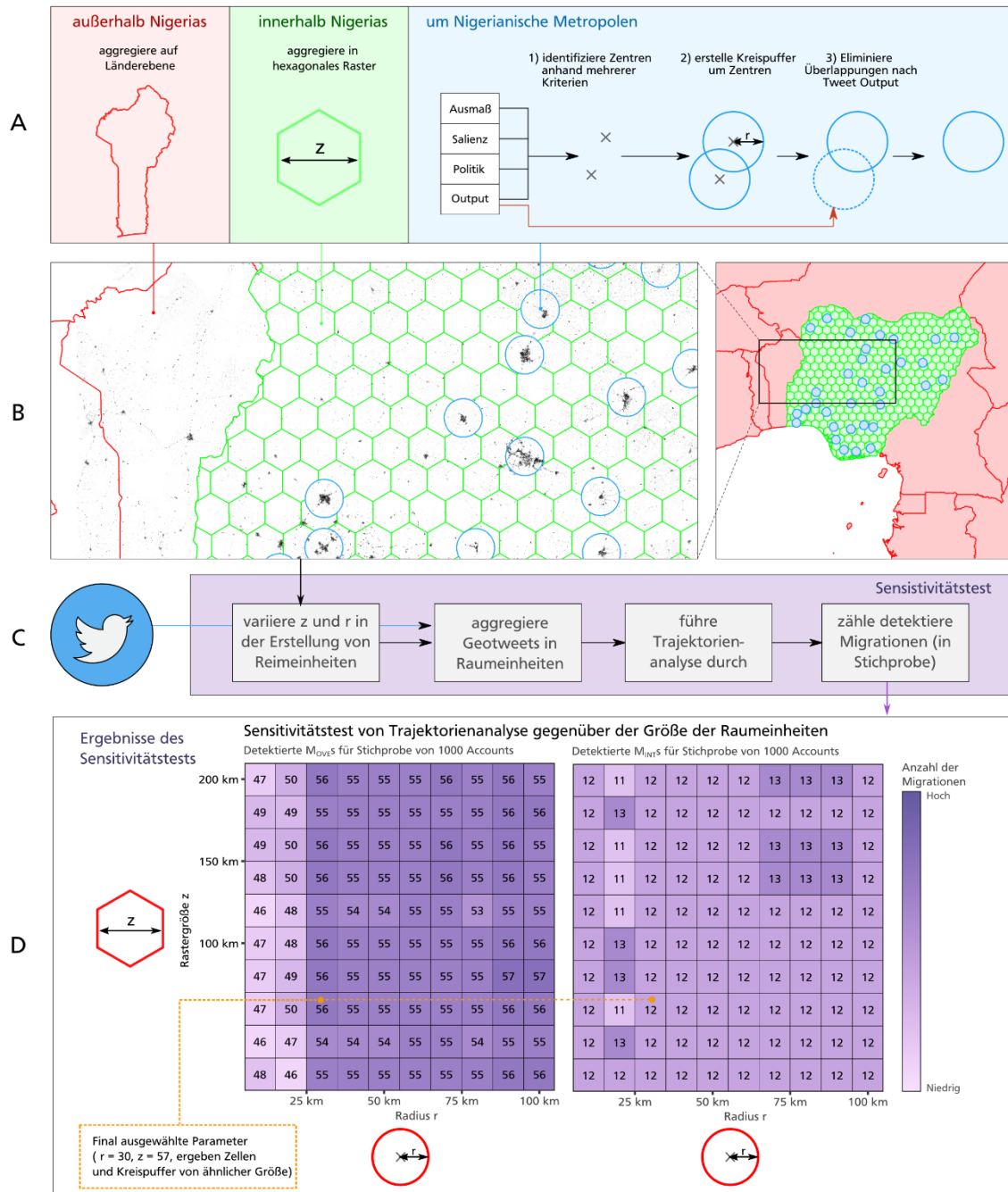


Abbildung A5: Ableitung der räumlichen Einheiten (RE) und Sensitivitätstest. A: Konzept der drei komplementären Typen von RE; B: räumliche Ansicht der in der Studie verwendeten RE; C: Konzept des Sensitivitätstests; D: Ergebnis des Sensitivitätstests.

Es zeigt sich eine hohe Robustheit der Ergebnisse, außer bei extremen Werten von r und z . Um Rasterzellen mit ähnlicher Fläche wie die städtischen RE zu erzeugen, wird ein Radius von 30 km und eine Zellengröße von 57 km gewählt. Wo solche Überschneidungen auftreten, wird das städtische RE mit der geringeren Anzahl von Tweets eliminiert. Außerdem werden als Anonymisierungsmaßnahme Informationen aus RE mit einer Einwohnerzahl von weniger als fünf ausgeschlossen.

A6. Ergebnisse der Analyse von Themen nach Mobilitätsgrad

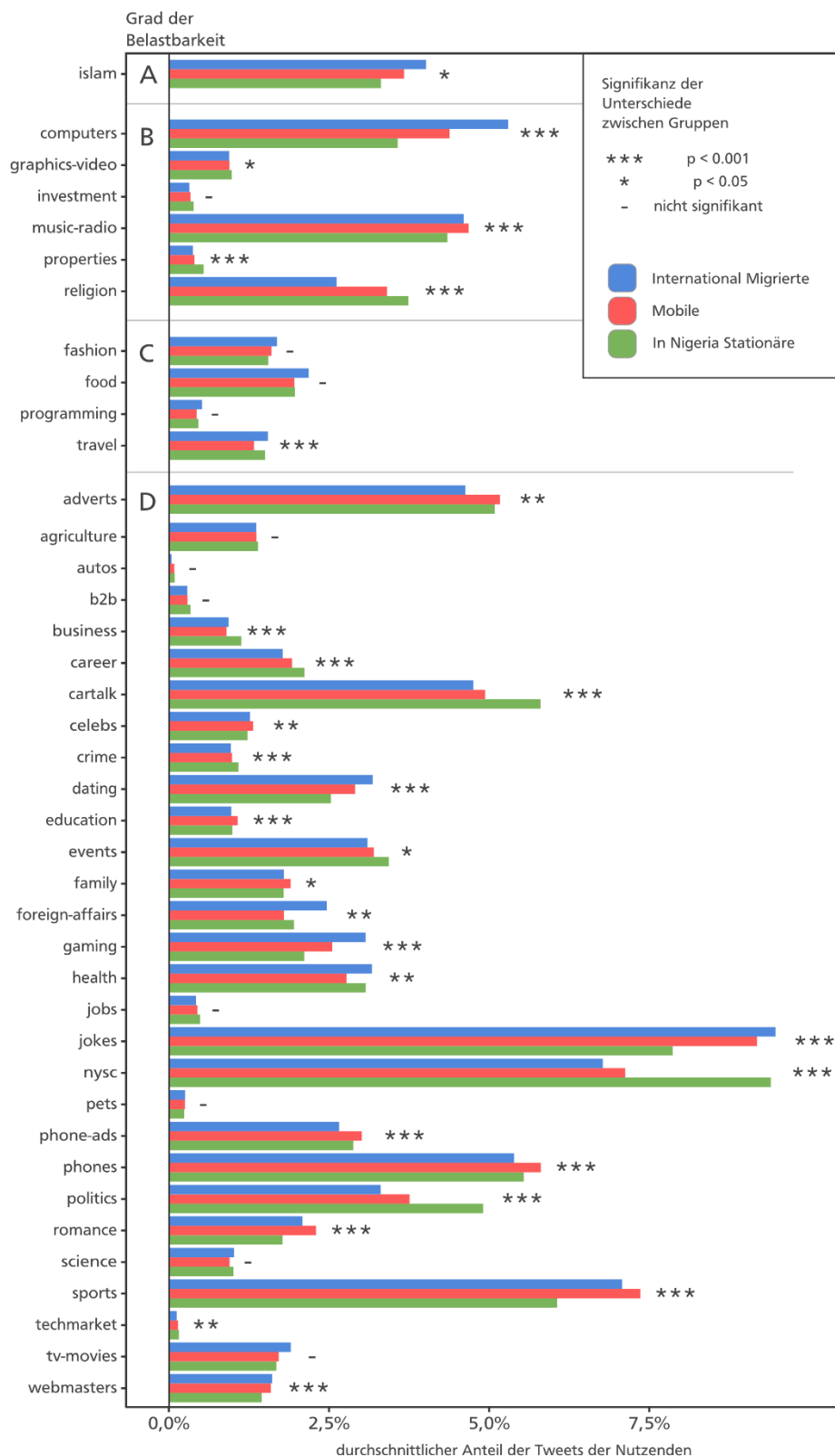


Abbildung A6: Gemittelte Themenanteile der Mobilitätstypen. Themen sind nach Belastbarkeit in Grade (a), (b) und (c) eingeteilt. Ein Test auf signifikante Unterschiede zwischen Mobilitätstypen erfolgte mittels der Kruskal-Wallis Methode.

A7. Plausibilitätschecks in Studie B1

Bezeichnung des Themas	Präzision	Sensitivität	F1	CV	COR	COS	Note
<i>islam</i>	0,628	0,629	0,629	0,618	0,709	0,731	A
<i>gaming</i>	0,776	0,634	0,698	0,032	0,110	0,463	B
<i>investment</i>	0,418	0,410	0,414	0,051	0,527	0,585	B
<i>music-radio</i>	0,359	0,603	0,450	0,148	0,509	0,678	B
<i>programming</i>	0,501	0,498	0,499	0,033	0,356	0,491	B
<i>properties</i>	0,341	0,598	0,434	0,094	0,753	0,434	B
<i>religion</i>	0,453	0,478	0,465	0,611	0,968	0,472	B
<i>fashion</i>	0,284	0,414	0,337	0,085	0,596	0,482	C
<i>food</i>	0,441	0,506	0,471	0,053	0,351	0,368	C
<i>travel</i>	0,526	0,326	0,402	0,085	0,663	0,312	C
<i>adverts</i>	0,235	0,356	0,283	0,115	0,159	0,302	D
<i>agriculture</i>	0,578	0,524	0,550	0,057	0,250	0,304	D
<i>autos</i>	0,416	0,597	0,490	0,041	0,000	0,055	D
<i>b2b</i>	0,234	0,348	0,280	0,105	0,761	0,057	D
<i>business</i>	0,228	0,207	0,217	0,084	0,765	0,393	D
<i>career</i>	0,251	0,283	0,266	0,138	0,660	0,318	D
<i>cartalk</i>	0,659	0,534	0,590	0,069	0,163	0,421	D
<i>celebs</i>	0,311	0,265	0,286	0,116	-0,132	0,273	D
<i>computers</i>	0,495	0,491	0,493	0,041	0,367	0,062	D
<i>crime</i>	0,371	0,373	0,372	0,127	-0,075	0,294	D
<i>dating</i>	0,352	0,480	0,406	0,152	-0,410	0,089	D
<i>education</i>	0,442	0,388	0,413	0,164	0,624	0,269	D
<i>events</i>	0,301	0,299	0,300	0,140	0,539	0,297	D
<i>family</i>	0,402	0,312	0,351	0,067	-0,352	0,328	D
<i>foreign-affairs</i>	0,448	0,392	0,418	0,094	0,018	0,513	D
<i>graphics-video</i>	0,467	0,528	0,495	0,041	0,754	0,201	D
<i>health</i>	0,278	0,391	0,325	0,088	0,026	0,509	D
<i>jobs</i>	0,509	0,496	0,502	0,212	0,894	0,224	D
<i>jokes</i>	0,104	0,334	0,158	0,095	-0,054	0,073	D
<i>nysc</i>	0,664	0,479	0,557	0,195	0,400	0,275	D
<i>pets</i>	0,728	0,581	0,646	0,039	-0,253	0,250	D
<i>phone-ads</i>	0,407	0,409	0,408	0,037	0,241	0,141	D
<i>phones</i>	0,592	0,516	0,552	0,072	-0,533	0,545	D
<i>politics</i>	0,467	0,475	0,471	0,104	-0,095	0,342	D
<i>romance</i>	0,347	0,232	0,278	0,066	-0,329	0,242	D
<i>science</i>	0,255	0,349	0,295	0,068	0,586	0,371	D
<i>sports</i>	0,601	0,573	0,587	0,094	0,156	0,732	D
<i>techmarket</i>	0,285	0,316	0,300	0,078	0,670	0,168	D
<i>tv-movies</i>	0,611	0,485	0,541	0,197	0,173	0,505	D
<i>webmasters</i>	0,448	0,482	0,464	0,049	-0,332	0,175	D

Tabelle A7: Ergebnisse des aus drei Komponenten bestehenden Plausibilitätschecks.

A8. Codes der Qualitativen Analyse in Studie B1

<i>religion</i>	
Code	Beschreibung
Ausruf oder Gebet	Bitten, Gebete, Wunschausrufe oder direkte Gespräche mit einer Gottheit.
Kirche und Klerus	Allgemeiner Diskurs über Kirchenangelegenheiten und die Handlungen der Kirchenleitung.
Gespräch über Religion	Diskurs über das Verhältnis der Religionen, die Rolle der Religion oder Äußerungen über den Glauben des Nutzenden.
Spendensammlung / Anzeige	Bitten um Spenden oder Werbung für Andachtsbücher oder religiöses Selbsthilfematerial.
Andere oder unidentifizierte Themen	Nicht-religiöse Themen und Gesprächsfragmente mit zu wenig Kontext, um das Thema zu identifizieren.
Zitat von Person	Zitat, das eindeutig einer öffentlichen Person zugeschrieben wird.
Zitat aus religiösem Text	Ausdrücklicher Verweis oder direktes Zitat aus einem religiösen Text (fast immer Bibelzitate), häufig von Nutzenden personalisiert. Manchmal verknüpft mit täglichen Andachten.
Religion und Politik	Politische Themen mit religiösem Hintergrund oder Diskurse über die Überschneidung von Religion und Politik.
Religiöse Veranstaltungen oder Ereignisse	Ankündigung von oder Kommentar zu religiösen Versammlungen und Kongressen.
Religiöse Musik	Religiöse Musikdiskussionen oder Liedtexte.
Religiöse oder Philosophische Aussage oder Rat	Ratschläge zu religiösem Verhalten, Lebensstil und Denkweise, vorwiegend religiöser Natur.

Tabelle A8a: Im Thema *Religion* identifizierte Codes.

<i>gaming</i>	
Code	Beschreibung
Öffentliches Leben	Kommentare zu Prominenten, Partys, aber auch Diskussionen über Stil und Mode, sowie Darstellungen von Mode und Kunst.
Film und Fernsehen	Besprechungen von Filmen und Serien.
Fußball	Eine Vielzahl von fußballbezogenen Inhalten. Vor allem Kommentare und Diskussionen über den professionellen Verbandsfußball, insbesondere den europäischen Männerfußball und die nigerianische Fußballnationalmannschaft. Retweets von Fußballspielern, Fußballmarken und Nachrichtenagenturen. Enthält auch Kommentare zur Leistung der Fantasy-Football-Teams der Nutzenden.
Diskussion über Musik	Kommentare und Meinungen zu Liedern und musikalischen Darbietungen von Profis und Amateursängern.
Werbung für Unterhaltung	Werbung für Musik, Veranstaltungen oder andere Unterhaltungsinhalte. Überwiegend Retweets.
Smalltalk	Persönliche Interaktion, Gesprächsfragmente, Witze und Memes.
Alltag	Verschiedene Aspekte des Privatlebens. Kommentare zum Wetter, zur Arbeit, zum Essen, aber auch zu persönlichen Beziehungen und zur Familie.
Politik	Nachrichten über politische Ereignisse und Kommentare zur Politik.
Inspirierendes	Gespräche und Äußerungen spiritueller Natur oder allgemeine Lebensratschläge.
Sport und Spiele	Sportarten, die kein Fußball sind, körperliche Aktivitäten und Bewegung, Brettspiele.
Reisen und Transport	Diskussionen und Nachrichten über Politik und Technologie im Zusammenhang mit Reisen sowie persönliche Reiseerfahrungen.
Videospiele	Nachrichten und persönliche Erfahrungen und Meinungen über Videospiele.

Tabelle A8b: Im Thema *Gaming* identifizierte Codes.

<i>travel</i>	
Code	Beschreibung
Anzeige	Werbung oder Anzeigen für Unterkünfte, Gastronomie oder Transport.
Diskurs mit Bezug auf Nigeria	Ernsthafte Diskussionen über Nigeria oder Nigerianer, oft zu politischen und wirtschaftlichen Themen. Gelegentlich humorvolle Kommentare über Nigerianer.
Mobilitätsangelegenheiten und Probleme	Beschwerden über den Zustand der Infrastruktur, den Verkehr, die Leistungen der Mobilitätsanbieter und kritische Kommentare zur Verkehrspolitik.
Ortsbezogene Ereignisse oder Veranstaltungen	Kommentieren Sie aktuelle Ereignisse und Tragödien mit einem eindeutigen Verweis auf den Ort. Oft in Form von Beileidsbekundungen und Gebeten.
Nachrichten und Politik	Nachrichten (meist in Form von Retweets) oder Kommentare zu bestimmten politischen Themen oder zur Politik im Allgemeinen.
Andere oder unidentifizierte Themen	Nicht reisebezogene Themen und Gesprächsfragmente mit zu wenig Kontext, um das Thema zu identifizieren.
Persönliche Reiseberichte	Bilder, Eindrücke und Berichte von aktuellen oder vergangenen Reisen.
Reisepläne oder Migrationspläne	Ausdruck der Absicht, zu reisen oder auszuwandern, sei es an einen bestimmten Ort oder allgemein.
Inspirierendes	Gespräche und Äußerungen spiritueller Natur oder allgemeine Lebensberatung.
Bestimmte Orte	Tweets, die Informationen, Nachrichten oder Meinungen über einen bestimmten Ort enthalten und die nicht zu anderen Codes passen.
Information über Reisen / Migration	Nachrichten zum Thema Verkehr und Transport aus dem lokalen und internationalen Bereich. Tipps von Benutzern zum Reisen und zu bestimmten Orten.
Reisen als abstraktes Konzept	Tweets, die sich über die Vorteile und Schwierigkeiten des Reisens äußern, oder Tweets, die reisebezogene Begriffe (Reise, Weg, Ziel) in einem metaphorischen Sinne verwenden.

Tabelle A8c: Im Thema *Travel* identifizierte Codes.

A9. Vereinigung von Themen in Experiment in Studie B1

Ursprüngliches Thema	Vereinigtes Thema
<i>family</i> <i>romance</i> <i>dating</i>	<i>Relationships</i>
<i>religion</i> <i>islam</i>	<i>Religion</i>
<i>phones</i> <i>phone-ads</i> <i>techmarket</i>	<i>Phones</i>
<i>politics</i> <i>crime</i>	<i>Politics</i>
<i>foreign-affairs</i> <i>education</i> <i>nysc</i>	<i>Education</i>
<i>programming</i> <i>webmasters</i>	<i>Programming</i>
<i>cartalk</i> <i>autos</i>	<i>Cars</i>
<i>adverts</i> <i>business</i> <i>b2b</i> <i>investment</i> <i>career</i> <i>jobs</i>	<i>Commerce</i>
<i>jokes</i>	- (Thema wurde gelöscht)

Tabelle A9: In einem der Studie C1 nachgestellten Experiment wurden Themen vereinigt. Hierdurch, und die Exklusion von sehr kurzen Texten, erhöhte sich die Klassifikationsgenauigkeit von 42% auf 60%.

A10. Themen der in Studien C1 und C2 genutzten Taxonomie

Themenbezeichnung (Englisch)	Themenbezeichnung (Englisch)	Kommentare
Health	Gesundheit	Körperliche und geistige Gesundheit (mit Ausnahme von körperlicher Betätigung zu Leistungszwecken, die eher der Kategorie Sport & Spiele zugeordnet werden sollte)
History & Culture	Geschichte & Kultur	Geisteswissenschaften, darstellende und kreative Künste (im Allgemeinen ohne Kunsthandwerk wie Holzbearbeitung), Geschichte (mit Ausnahme der jüngsten Geschichte), Kulturerbe
Science, Education & Mathematics	Wissenschaft, Bildung & Mathematik	Naturwissenschaften, Mathematik, Statistik, Wissenschaft, Schulen, Bildung, Forschung
Economy, Business & Finance	Wirtschaft, Geschäftliches & Finanzen	Persönliche Finanzen und Wirtschaft
Sports & Games	Sport & Spiele	Videospiele, Karten- und Brettspiele, Profi- und Freizeitsport, Fitness, andere Wettbewerbe mit körperlicher Betätigung.
Technology	Technologie	Einschließlich Elektronik, Softwaretechnik, Entwicklung von Software (auch Spiele). Praktisch angewandte Datenwissenschaft und maschinelles Lernen, aber ohne theoretische Statistik und Datenwissenschaft (diese sollten eher zu Wissenschaft, Bildung & Mathematik gehören)
Adverts	Anzeigen	Eigenwerbung, typischerweise für kommerzielle Inhalte, oder Jobangebote
Private Life, Family & Relationships	Privatleben, Familie & Beziehungen	Diverse Aspekte des Privatlebens und private Beziehungen zwischen Menschen, auch Partnersuche, Haustiere & Hausarbeit
Religion	Religion	Religion, Spiritualität und Philosophie
Politics & Government	Politik & Regierung	Innenpolitik (jedes Land). Parteipolitik, Wahlen.
Travel, Tourism & Migration	Reisen, Tourismus & Migration	Einschließlich der Politik der Migration. Diskussion über Reise- und Transportmittel, Verkehrsinfrastruktur und Reisetipps. Einschließlich Themen über Flüchtlinge und Migrierte.
Other	Sonstiges	Eine gemischte Kategorie. Dinge, die von Natur aus ein breites Spektrum verschiedener Themen betreffen, wie z.B.: Witze, Literatur, Ratschläge.
International Politics	Internationale Politik	Beziehungen zwischen Ländern und internationalen Organisationen, Geopolitik, Diplomatie, grenzüberschreitende Militäroperationen.
Celebrities, Entertainment & Music	Prominente, Unterhaltung & Musik	Einschließlich Filme, TV-Serien, Anime, Cartoons und Comics. Berühmte Personen, wenn sie aus der Unterhaltungsbranche kommen. Berühmte Personen, die in einem anderen Bereich tätig sind (z.B. religiöse Führer oder Politiker im Amt), sollten nach Möglichkeit lieber einer Kategorie zugeordnet werden, die zu ihrem Bereich passt, es sei denn, der Text behandelt nur ihr Privatleben. Zeitgenössische und klassische Musik und Musiker, auch einschließlich Musiktheorie und Instrumente.
Food & Agriculture	Ernährung & Landwirtschaft	Kochen, professionelle und Hobby-Landwirtschaft (außer Drogenproduktion), Gartenarbeit, Vegetarismus/Veganismus
Events	Ereignisse & Veranstaltungen	Geburtstage, Hochzeiten, ähnliche Feierlichkeiten, Klassentreffen und andere organisierte Zusammenkünfte, die nicht eindeutig in eine andere Kategorie passen
Architecture, Construction & Real Estate	Architektur, Bau & Immobilien	Tiefbau und Hochbau, sowohl Großprojekte als auch Ausbau des Eigenheims. Inklusive Stadtplanung und Gartenbau.
Crime	Verbrechen	Kriminelle Aktivitäten, Korruption, Verstöße gegen nationales Recht, fahrlässiges Zuschadenkommenlassen von Menschen, ausgenommen sind Menschen verursachte Katastrophen größeren Ausmaßes
Natural Disasters and Hazards	Naturkatastrophen & Naturgefahren	Naturkatastrophen verschiedener Art, einschließlich kleinerer Katastrophen und Unfälle, die nur wenige Menschen betreffen. Ohne von Menschen verursachte Katastrophen wie Umweltverschmutzung, Ölverschmutzung usw.

Tabelle A10: Themen der in Studien C1 und C2 genutzten Taxonomie und ihre Beschreibung. Die hier angegebenen Beschreibungen setzen sich aus Kommentaren zusammen, welche von den Analysten in der Studie zur Abgrenzung der Themen genutzt wurden, und sollen keine vollständige Definitionen der Themen darstellen.

A11. Konfusionsmatrizen der Validierung der Themenklassifizierung in Studie C1



Abbildung A11. Konfusionsmatrizen des Vergleichs der Themenannotation von Studie C1 mit Annotation durch menschlichen Analysten. Abkürzungen sind aus den englischen Themenbezeichnungen abgeleitet (siehe A10).

A12. Modelloutputs der Studie C1

Term	Koeffizient	Standardfehler	p-Wert
(Konstante)	-0,4559	0,0138	0,0000
Thema: Architektur, Bau & Immobilien	0,0876	0,0028	0,0000
Thema: Prominente, Unterhaltung & Musik	-0,0234	0,0019	0,0000
Thema: Verbrechen	0,1715	0,0018	0,0000
Thema: Wirtschaft, Geschäftliches & Finanzen	-0,1027	0,0017	0,0000
Thema: Ernährung & Landwirtschaft	0,0068	0,0020	0,0007
Thema: Gesundheit	-0,064	0,0018	0,0000
Thema: Geschichte & Kultur	0,3635	0,0079	0,0000
Thema: Internationale Politik	0,4293	0,0020	0,0000
Thema: Naturkatastrophen & Naturgefahren	0,1943	0,0024	0,0000
Thema: Politik & Regierung	0,1134	0,0017	0,0000
Thema: Religion	0,1686	0,0026	0,0000
Thema: Wissenschaft, Bildung & Mathematik	-0,0044	0,0018	0,0124
Thema: Sport & Spiele	-0,0544	0,0103	0,0000
Thema: Technologie	-0,0688	0,0019	0,0000
Thema: Reisen, Tourismus & Migration	0,3042	0,0020	0,0000
Thema: Privatleben, Familie & Beziehungen	-0,3798	0,0100	0,0000
Textlänge	0,0003	0,0000	0,0000
zufälliger Effekt (σ Konstante) für Autor	1,0563		
zufälliger Effekt (Σ Konstante) für Zeitstempel	0,0504		

Tabelle A12a. Geschätzte Koeffizienten des Modells auf Datensatz: GDELТ-GKG

Term	Koeffizient	Standardfehler	p-Wert
(Konstante)	-2,1197	0,0463	0,0000
Thema: Architektur, Bau & Immobilien	-0,4787	0,0352	0,0000
Thema: Prominente, Unterhaltung & Musik	-0,5457	0,039	0,0000
Thema: Verbrechen	-0,332	0,0292	0,0000
Thema: Wirtschaft, Geschäftliches & Finanzen	-0,4123	0,0263	0,0000
Thema: Ernährung & Landwirtschaft	0,0071	0,0266	0,7902
Thema: Gesundheit	-0,1762	0,0258	0,0000
Thema: Geschichte & Kultur	-0,0343	0,0864	0,6915
Thema: Internationale Politik	-0,0025	0,0329	0,9404
Thema: Naturkatastrophen & Naturgefahren	-0,548	0,0364	0,0000
Thema: Politik & Regierung	-0,0362	0,0261	0,1648
Thema: Privatleben, Familie & Beziehungen	-0,5622	0,2475	0,0231
Thema: Religion	0,0737	0,0277	0,0078
Thema: Wissenschaft, Bildung & Mathematik	1,3805	0,0254	0,0000
Thema: Sport & Spiele	-0,3269	0,2043	0,1096
Thema: Technologie	-0,7739	0,0389	0,0000
Thema: Reisen, Tourismus & Migration	-0,3334	0,0308	0,0000
zufälliger Effekt (Σ Konstante) für Autor	1,499		
zufälliger Effekt (Σ Konstante) für Zeitstempel	0,0757		

Tabelle A12b. Geschätzte Koeffizienten des Modells auf Datensatz: GDELT-IA-AMERICANA

Term	Koeffizient	Standardfehler	p-Wert
(Konstante)	-2,4976	0,0131	0,0000
Thema: Anzeigen	0,0942	0,0186	0,0000
Thema: Architektur, Bau & Immobilien	0,7574	0,0273	0,0000
Thema: Prominente, Unterhaltung & Musik	-0,2612	0,0142	0,0000
Thema: Verbrechen	0,3776	0,0140	0,0000
Thema: Wirtschaft, Geschäftliches & Finanzen	0,1424	0,0131	0,0000
Thema: Ereignisse & Veranstaltungen	-0,0437	0,0187	0,0192
Thema: Ernährung & Landwirtschaft	0,2302	0,0147	0,0000
Thema: Gesundheit	0,2942	0,0170	0,0000
Thema: Geschichte & Kultur	0,7703	0,0202	0,0000
Thema: Internationale Politik	1,2668	0,0133	0,0000
Thema: Politik & Regierung	0,7385	0,0132	0,0000
Thema: Privatleben, Familie & Beziehungen	-0,5774	0,0131	0,0000
Thema: Religion	-0,2365	0,0146	0,0000
Thema: Wissenschaft, Bildung & Mathematik	0,1949	0,0147	0,0000
Thema: Sport & Spiele	0,3051	0,0146	0,0000
Thema: Technologie	-0,4059	0,0139	0,0000
Thema: Reisen, Tourismus & Migration	1,2313	0,0127	0,0000
Textlänge	0,0014	0,0000	0,0000
zufälliger Effekt (Σ Konstante) für Autor	0,7743		
zufälliger Effekt (Σ Konstante) für Zeitstempel	0,0638		

Tabelle A12c. Geschätzte Koeffizienten des Modells auf Datensatz: Nairaland

Term	Koeffizient	Standardfehler	p-Wert
(Konstante)	-3,1566	0,0535	0,0000
Thema: Anzeigen	1,0645	0,0539	0,0000
Thema: Architektur, Bau & Immobilien	0,9177	0,0601	0,0000
Thema: Prominente, Unterhaltung & Musik	0,0202	0,0506	0,6905
Thema: Verbrechen	0,5449	0,0578	0,0000
Thema: Wirtschaft, Geschäftliches & Finanzen	-0,0188	0,0548	0,7322
Thema: Ereignisse & Veranstaltungen	0,1909	0,0488	0,0001
Thema: Ernährung & Landwirtschaft	0,0442	0,0471	0,3480
Thema: Gesundheit	0,2439	0,0624	0,0001
Thema: Geschichte & Kultur	0,3451	0,0559	0,0000
Thema: Internationale Politik	1,5474	0,0527	0,0000
Thema: Politik & Regierung	0,7973	0,0483	0,0000
Thema: Privatleben, Familie & Beziehungen	-0,8712	0,0592	0,0000
Thema: Religion	-0,4410	0,0610	0,0000
Thema: Wissenschaft, Bildung & Mathematik	0,2735	0,0644	0,0000
Thema: Sport & Spiele	0,6450	0,0505	0,0000
Thema: Technologie	-0,0744	0,0670	0,2673
Thema: Reisen, Tourismus & Migration	0,7730	0,0478	0,0000
Textlänge	0,0064	0,0001	0,0000
zufälliger Effekt (Σ Konstante) für Autor	0,9083		
zufälliger Effekt (Σ Konstante) für Zeitstempel	0,1046		

Tabelle A12d. Geschätzte Koeffizienten des Modells auf Datensatz: Twitter (Geotweets)

Term	Koeffizient	Standardfehler	p-Wert
(Konstante)	-3,5309	0,1077	0,0000
Thema: Anzeigen	-0,0852	0,1340	0,5246
Thema: Architektur, Bau & Immobilien	0,7731	0,1407	0,0000
Thema: Prominente, Unterhaltung & Musik	-0,3402	0,1169	0,0036
Thema: Verbrechen	0,5572	0,1220	0,0000
Thema: Wirtschaft, Geschäftliches & Finanzen	-0,2825	0,1243	0,0231
Thema: Events	0,0873	0,1073	0,4162
Thema: Ernährung & Landwirtschaft	-0,1297	0,1050	0,2169
Thema: Gesundheit	0,0389	0,1358	0,7747
Thema: Geschichte & Kultur	0,0582	0,1275	0,6482
Thema: Internationale Politik	1,5314	0,1148	0,0000
Thema: Politik & Regierung	0,5413	0,1089	0,0000
Thema: Privatleben, Familie & Beziehungen	-1,5965	0,1505	0,0000
Thema: Religion	-0,9646	0,1429	0,0000
Thema: Wissenschaft, Bildung & Mathematik	0,3596	0,1391	0,0097
Thema: Sport & Spiele	0,5328	0,1124	0,0000
Thema: Technologie	-0,2203	0,1413	0,1188
Thema: Reisen, Tourismus & Migration	1,4351	0,1127	0,0000
Textlänge	0,0084	0,0003	0,0000
zufälliger Effekt (Σ Konstante) für Autor	0,8015		
zufälliger Effekt (Σ Konstante) für Zeitstempel	0,0287		

Tabelle A12e. Geschätzte Koeffizienten des Modells auf Datensatz: Twitter (ohne geotags)

Term	Koeffizient	Standardfehler	p-Wert
(Konstante)	-2,0019	0,0112	0,0000
Thema: Prominente, Unterhaltung & Musik	0,0185	0,0168	0,2708
Thema: Wirtschaft, Geschäftliches & Finanzen	0,0299	0,0140	0,0324
Thema: Ernährung & Landwirtschaft	-0,5047	0,0171	0,0000
Thema: Geschichte & Kultur	1,6023	0,0168	0,0000
Thema: Politik & Regierung	1,9984	0,0164	0,0000
Thema: Privatleben, Familie & Beziehungen	-1,2213	0,0204	0,0000
Thema: Religion	-0,2665	0,0144	0,0000
Thema: Wissenschaft, Bildung & Mathematik	-1,9674	0,0116	0,0000
Thema: Sport & Spiele	-1,1027	0,0128	0,0000
Thema: Technologie	-2,4744	0,0121	0,0000
Thema: Reisen, Tourismus & Migration	2,5225	0,0155	0,0000
Textlänge	0,0003	0,0000	0,0000
zufälliger Effekt (Σ Konstante) für Autor	0,8669		
zufälliger Effekt (Σ Konstante) für Zeitstempel	0,0162		

Tabelle A12f Geschätzte Koeffizienten des Modells auf Datensatz: Stackexchange

Term	Koeffizient	Standardfehler	p-Wert
(Konstante)	-1,8796	0,0112	0,0000
Thema: Architektur, Bau & Immobilien	0,5279	0,1369	0,0001
Thema: Prominente, Unterhaltung & Musik	-0,2166	0,0175	0,0000
Thema: Wirtschaft, Geschäftliches & Finanzen	-0,3224	0,0254	0,0000
Thema: Ernährung & Landwirtschaft	0,4294	0,0473	0,0000
Thema: Gesundheit	-1,278	0,0337	0,0000
Thema: Geschichte & Kultur	0,4900	0,0519	0,0000
Thema: Internationale Politik	1,5405	0,0199	0,0000
Thema: Politik & Regierung	0,5612	0,0158	0,0000
Thema: Privatleben, Familie & Beziehungen	-1,1533	0,0181	0,0000
Thema: Religion	-0,2912	0,0190	0,0000
Thema: Wissenschaft, Bildung & Mathematik	-0,3394	0,0198	0,0000
Thema: Sport & Spiele	-0,5890	0,0124	0,0000
Thema: Technologie	-1,1282	0,0181	0,0000
Thema: Reisen, Tourismus & Migration	2,2714	0,0620	0,0000
Textlänge	0,0004	0,0000	0,0000
zufälliger Effekt (Σ Konstante) für Autor	0,7338		

Tabelle A12g. Geschätzte Koeffizienten des Modells auf Datensatz: WEBIS Reddit