Master Thesis

im Rahmen des
Universitätslehrganges „Geographical Information Science & Systems“
(UNIGIS MSc) am Interfakultären Fachbereich für GeoInformatik (Z_GIS)
der Paris Lodron-Universität Salzburg

zum Thema

Erkennung von städtischen Strukturen und Prozessen in sozialen Medien

Eine semantische und räumliche Analyse von Tweets

vorgelegt von

MSc Jan Grade
104261, UNIGIS MSc Jahrgang 2015

Zur Erlangung des Grades
„Master of Science (Geographical Information Science & Systems) – MSc(GIS)”

Bonn, 30.07.2018

Danksagung

Ich möchte mich bei denjenigen bedanken, die mich während des Studiums und während der Arbeit an dieser Master Thesis sowohl durch motivierende Worte als auch interessante Diskussionen und gute Ratschläge unterstützt haben: meine Familie, meine Freunde und meine ArbeitskollegInnen bei der empirica ag. Ein besonderes Dankeschön gilt Kathrin für die Aufmunterung und die Unterstützung bis zur Abgabe.

Weiterhin möchte ich mich darüber hinaus beim gesamten UNIGIS-Team bedanken. Mein Betreuer Ass.-Prof. Dr. Bernd Resch hat mich mit vielen guten Hinweisen und Anmerkungen für die Bearbeitung der Master Thesis sowie bei der Datenerhebung unterstützt. Das UNIGIS-Team stand darüber hinaus während des Studiums immer mit Rat und Tat zur Seite, hat bei den vielen wichtigen Kleinigkeiten geholfen und stets schnell und kompetent auf Fragen geantwortet.

Erklärung

Hiermit versichere ich, dass ich die vorliegende Master Thesis ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit wurde in gleicher oder ähnlicher Form noch bei keinem anderen Prüfungsamt eingereicht.

Bonn, 30.07.2018

Ort, Datum Jan Grade

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

In dieser Master Thesis geht es um die Auseinandersetzung mit der Frage, wie Themen aus georeferenzierten Tweets extrahiert werden können und ob diese Themen dazu geeignet sind, städtische Räume zu charakterisieren. Neue Daten aus Volunteered Geographic Informationen und sozialen Medien erfahren eine hohe Aufmerksamkeit im Bereich Stadtentwicklung und Stadtplanung. Die Kombination von Texten und Koordinaten in Beiträgen aus sozialen Medien ermöglicht die räumliche Analyse von Strukturen und Prozessen in der Stadt unabhängig von amtlichen Statistiken, die oftmals an gegebene Gebietseinheiten gebunden sind und nur anhand von Wohnadressen oder Arbeitsstellen verortet werden.

Mittels Latent Dirichlet Allocation (LDA) werden Themen zur Beschreibung städtischer Räume in geocodierten Tweets identifiziert. Dies erfolgt anhand von zwei Datensätzen für die Region Greater Manchester und die Stadt Birmingham. Die Analyse zeigt räumliche Zusammenhänge zwischen den identifizierten Clustern und verschiedenen Anziehungspunkten in den Städten. Das lässt sich sowohl anhand von Kenntnissen über reale Orte als auch über eine Verschneidung mit Standorten ausgewählter Versorgungsangebote auf Basis von OpenStreetMap-Daten nachweisen. Der letzte Teil der Analyse zeigt einen deutlichen Zusammenhang zwischen Fußballspielen und dem Aufkommen von semantisch passenden Tweets im Stadionumfeld an bestimmten Tagen. Am Beispiel von Bahnhaltestellen werden außerdem die täglichen Mobilitätsmuster in den Städten sichtbar. Die Analyse der Tweets ist also sowohl im Hinblick auf semantische als auch räumliche und zeitliche Muster möglich.

Die Ergebnisse zeigen die Potentiale für unterschiedliche Anwendungsfelder. Eine weitergehende Optimierung der Methoden, insbesondere bei der semantischen Klassifizierung mittels Algorithmen, ist aber noch erforderlich. Dennoch bietet die Kombination von Daten aus sozialen Medien mit weiteren Datengrundlagen, z.B. auf Basis von Volunteered Geographic Information und auch klassischen, amtlichen Datengrundlagen, bereits jetzt das Potential für ein tiefergehendes Verständnis von städtischen Strukturen und Prozessen.

1 Einleitung

In diesem Kapitel erfolgen die Beschreibung des Analyserahmens und die Ableitung der Fragestellung für die Master Thesis. Im Kapitel 1.1 geht es zunächst um Volunteered Geographic Information (VGI) und Daten aus sozialen Medien als neue Grundlagen für räumliche Analysen in der Stadt. Dabei steht die Frage nach den Unterschieden zu anderen Datengrundlagen im Vordergrund. In Kapitel 1.2 wird dann ein Blick auf die Literatur zur räumlichen Analyse von Daten aus sozialen Medien, insbesondere Twitter, als Datengrundlage geworfen. Darauf bauen die Fragestellung dieser Master Thesis und die Forschungsfragen auf. Die Herleitung erfolgt in Kapitel 1.3.

1.1 Volunteered Geographic Information, soziale Medien und die Stadt

Die quantitative, räumliche Datenanalyse im städtischen Kontext basiert bisher vorwiegend auf amtlichen Datengrundlagen oder baut unmittelbar darauf auf. Diese Daten werden von Städten bzw. Kommunen sowie übergeordneten, administrativen Ebenen und Organisationen oder durch Unternehmen bereitgestellt. Amtliche, aber oftmals auch kommerzielle Daten, beschränken sich zunächst auf abgegrenzte, teilweise durch administrative Zuständigkeit bestimmte, Gebiete. Dazu gehören z.B. Ortsteile, eigens für Monitoringzwecke gebildete statistische Bezirke oder Postleitzahlgebiete. Diese Bindung an festgelegte Gebietseinheiten kommt dadurch zustande, dass amtliche und auch kommerzielle Statistiken und Datensätze meistens nur in aggregierter Form angeboten werden. Jedoch können die jeweiligen Datenquellen nicht die unterschiedlichen Aufenthaltsorte von Personen im Tagesverlauf wiedergegeben, denn zumeist beschränken sich die Rohdaten auf die Wohn- oder Arbeitsadresse (Crooks et al., 2015). Beispielsweise erfassen Einwohnermeldedaten nur die Wohnadresse einer Person und die Beschäftigungsstatistik der Bundesagentur für Arbeit nur den Wohn- sowie den Arbeitsort. Untersuchungen, die dagegen die unterschiedlichen Nutzungsräume von Bewohnern oder Besuchern einer Stadt erfassen möchten, basieren dann oft auf Befragungen oder qualitativen Methoden, z.B. Interviews oder kognitiven Karten, um die alltäglichen Wege und Aufenthaltsorte der untersuchten Personen nachvollziehen zu können.

Laut Crooks et al. (2015) sind viele Studien auf Basis dieser klassischen Datengrundlagen limitiert, weil sie nur einen begrenzten Zeitraum und nur eine begrenzte Anzahl an kohärenten Daten verwenden. Ein typisches Beispiel für diese Limitierungen sind die administrativen Grenzen von Stadtteilen und Quartieren, die nicht unbedingt die Wahrnehmung von zusammenhängenden Räumen durch die Bewohner wiederspiegeln. Die sozialen Verbindungen und Aktivitäten der

Menschen enden nicht an Stadtteilgrenzen. Nur kann dies durch die amtlichen Daten nicht abgebildet werden.

In den unterschiedlichen Disziplinen, die sich mit städtischen Strukturen und den Nutzungsmustern in der Stadt beschäftigen, erfährt darum die Analyse von neuen Datengrundlagen aus unterschiedlichen Datenquellen ein steigendes Interesse. Heutzutage erzeugen viele Menschen mehr oder weniger freiwillig geographische Informationen und hinterlassen somit ihren persönlichen, digitalen Fußabdruck. Der Begriff der Volunteered Geographic Information (VGI) beschreibt dieses Phänomen (Goodchild, 2007). Unter den Begriff der VGI fallen sowohl von Nutzern aktiv generierte Daten als auch in einem weiteren Sinne Daten, die durch Aktivitäten von Nutzern erzeugt und dann zur Verfügung gestellt werden. Auch solche Daten enthalten oft geographische Informationen (See et al., 2016). Craglia et al. (2012) sprechen in diesem Zusammenhang von einer expliziten und einer impliziten Dimension. Explizite Daten werden mit dem Ziel der Erfassung geographischer Features bereitgestellt. Das bekannteste Beispiel für explizite VGI ist vermutlich OpenStreetMap als „freie Weltkarte“ (OpenStreetMap - Deutschland). Räumliche Daten aus sozialen Medien stellen dagegen implizite Daten dar, bei denen die räumliche Information zunächst kein integraler Bestandteil der Daten ist, aber dennoch eine Georeferenzierung vorliegen kann. Dazu gehören Bilder und Videos (z.B. Flickr, Instagram) sowie Microblogs (z.B. Twitter, Statusmeldungen auf Facebook). Auch wenn es sich dabei nicht um ein völlig neues Phänomen handelt, so hat die Entwicklung des Web 2.0 und die zunehmende Nutzung von Geräten mit GPS-Funktionalitäten und anderen Verortungsroutinen die Erzeugung dieser geographischen Daten vereinfacht (Elwood et al., 2012). Verschiedene soziale Medien sammeln dabei nicht nur Daten ihrer Nutzer, sondern stellen diese auch in unterschiedlicher Form als maschinenlesbare Datensätze für andere Zwecke zur Verfügung.

Mit sozialen Medien, aus dem Englischen von *Social Media* ins Deutsche übersetzt, werden Internetplattformen bezeichnet, die eine soziale Interaktion zwischen den Nutzern, in der Regel mittels Profilen, ermöglichen. Dabei werden nutzergenerierte Inhalte, wie Texte, Fotos oder Videos, ausgetauscht. Zu sozialen Medien gehören neben sozialen Netzwerken auch Weblogs, Microblogs, Wikis sowie Foto- und Videoplattformen (Springer Gabler Verlag). Liu et al. (2015) beschreiben die Analyse von Individualdaten aus sozialen, teilweise explizit geographischen Netzwerken als *social sensing* (übersetzt in etwa soziale Messfühler). Dabei kann jedes Individuum, das diese sozialen Medien nutzt, als Sensor betrachtet werden. Der Begriff des *social sensing* ist angelehnt an den Begriff der Fernerkundung (*remote sensing*), aber anstelle von Landschaften mit ihren Strukturen erfasst der *social sensor* sozioökonomische Daten und menschliche Aktivitätsmuster. Craglia et al. (2012) unterscheiden nochmals zwischen aktiven und passiven Sensoren, was sich wiederum mit der Definition der expliziten und impliziten, geographischen Daten deckt. An dieser Stelle sollte darum

auch bewusst getrennt werden zwischen „echten“ VGI, die explizit und auf freiwilliger Basis durch die Nutzer bereitgestellt werden, und passiven, impliziten geographischen Informationen aus sozialen Medien. Über die dargestellten Begriffe hinaus haben sich noch zahlreiche weitere Definitionen etabliert, die sich teilweise überschneiden oder nur sehr spezifische Anwendungsfelder beschreiben. Eine umfangreiche Übersicht zur Terminologie haben See et al. (2016) zusammengestellt.

Für die Forschung stellen diese Datengrundlagen eine neue Quelle von Informationen für Analysen dar. Ein Anwendungsbereich ist die Untersuchung von Strukturen und Prozessen in der Stadt (Crooks et al., 2015). In diesem Zusammenhang verwenden Crooks et al. (2015) die Form als Begriff für die physische Struktur einer Stadt und die Funktion als Begriff für die Aktivitäten oder Prozesse innerhalb der städtischen Formen. Dazu gehören insbesondere auch soziale Aktivitäten im städtischen Raum. Beide Begriffe sind stark miteinander verbunden, denn die Form entsteht aus den Interaktionen der verschiedenen Akteure einer Stadt wie den Menschen, Infrastrukturen, Gebäuden, Aktivitäten und Regeln. Im öffentlichen Raum treffen sich somit Formen und Funktionen. VGI und Daten aus sozialen Medien bilden dies wiederum explizit oder implizit ab. Sie ergänzen die amtlichen Daten und ermöglichen neue Methoden der Datenanalyse zur Analyse dieser Formen und Funktionen (Long and Liu, 2016). Aber auch neue Formen der Bürgerbeteiligung (Tenney and Sieber, 2016) sowie die Begleitung von Planungsprozessen stellen mögliche Anwendungsfelder dar (Resch et al., 2016, Brabham, 2009). Dabei geht es gar nicht um die Digitalisierung jedes einzelnen Prozessschrittes, sondern vor allem um die Veränderung einzelner Planungs- und Entscheidungsprozesse durch die Bereitstellung zusätzlicher Informationen. Long and Liu (2016) unterscheiden in diesem Zusammenhang zwischen drei Arten von neuen Datenquellen: (1) offizielle Datenportale mit Zugriff auf Datenquellen, die in der Vergangenheit nicht öffentlich verfügbar waren; (2) Big Data aus unterschiedlichen Datenquellen; (3) Volunteered Geographic Information (VGI) einschließlich georeferenzierten Daten aus sozialen Medien oder auch Geo Social Media (Kim et al., 2016). Diese neuen Daten können aus Perspektive der Nutzer sowohl freiwillig als auch unfreiwillig zur Verfügung gestellt werden, wodurch sich der dritte Punkt nochmals differenzieren lässt.

Georeferenzierte Beiträge aus sozialen Medien können sehr einfach mit beliebigen räumlichen Einheiten verschnitten werden, da es sich um Punktdaten handelt. Dazu gehören insbesondere auch Zellen oder vergleichbare Analyseraster, die wiederum unabhängig von vorgegebenen Raumeinheiten sind (Zhou and Zhang, 2016). Unter der Annahme, dass die Verortung der Beiträge in einem Zusammenhang mit der realen Welt steht, entstehen dadurch neue Möglichkeiten, räumliche Prozesse und Strukturen in der Stadt anhand der Nutzerbeiträge sichtbar zu machen. Zu den bisherigen Ursache-Wirkungs-Modellen auf Basis standardisierter Daten stößt dann die Auswertung dieser teilweise sehr heterogenen Datengrundlagen mittels Algorithmen (Schüller, 2017).

Goodchild (2007) schreibt, dass sich durch das Aufkommen der VGI und ähnlicher Datengrundlagen sogar die Rolle der klassischen Urheber geographischer Daten neu definieren wird. Das Aufkommen der neuen Datenquellen ermöglicht neue Forschungsansätze zu sozialen Aktivitäten und Interaktionen. Eine wichtige Grundannahme dafür ist, dass die große Menge an Tweets und anderen Datensätzen die soziale und physische Umwelt der realen Welt wiederspiegeln und diese Daten daher zur Beantwortung von Fragen zu diesen Räumen und ihren Nutzern verwendet werden können. Zur Untersuchung solcher Zusammenhänge nutzen viele Studien der vergangenen Jahre Daten aus sozialen Medien und untersuchen diese mit unterschiedlichen GIS-Methoden (Zhou and Zhang, 2016, Resch et al., 2016, Longley and Adnan, 2016, Steiger et al., 2015b, Crooks et al., 2015, Croitoru et al., 2015, Lee et al., 2013).

1.2 Semantische und räumliche Analyse von Beiträgen aus sozialen Medien

Webseiten mit sozialen Netzwerken sind laut Boyd and Ellison (2007) „web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other user with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system." Boyd and Ellison (2007) beschreiben die Verbindungen zwischen den Nutzern dieser Netzwerke als eine Reflektion ihrer erweiterten sozialen Netzwerke in der realen Welt. Nach dieser Definition haben soziale Netzwerke jedoch noch keine räumliche Dimension. See et al. (2016) betonen in diesem Zusammenhang, dass die meisten Inhalte dieser Netzwerke zunächst nicht räumlich sind. Aber Schnittstellen zur Verknüpfung mit geographischen Angaben aus GPS-Koordinaten oder Geotags stellen mittlerweile auf vielen Geräten wie Smartphones räumliche Informationen als Hintergrundfunktionen zur Verfügung. Diese sogenannten Location Based Services speichern über den gesamten Tag freiwillig oder unfreiwillig geographische Informationen über den Nutzer. Ein Twitter-Nutzer kann sich beispielsweise dazu entscheiden, seinen Standort mit einem Tweet zu speichern und somit anderen zur Verfügung zu stellen.

Twitter wurde als Mikroblog-Seite im Jahr 2006 gestartet. Die Mikroblog-Nutzer teilen hier ihre Interessen und täglichen Aktivitäten mit anderen Nutzern und suchen Informationen, die wiederum von Anderen geteilt wurden. Es handelt sich um eine schnellere Art der Kommunikation als bei klassischen Blogs. Anstelle von längeren Beiträgen können Nutzer kurze Nachrichten schneller schreiben und ihren Status somit schneller ändern (Java et al., 2007). Dabei sind die Nutzer nicht an einen Computer zu Hause gebunden, sondern können von überall, wo mobiles Internet zur Verfügung steht, Beiträge veröffentlichen. Laut Twitter waren etwa 328 Mio. Accounts im Sommer 2016 monatlich aktiv und etwa 82 % der Nutzer haben ein Mobiltelefon zum Zugriff benutzt (Twitter Inc.).

Tweets ermöglichen sowohl die Analyse der Texte selbst also auch der Uhrzeit und des Ort der Veröffentlichung, insofern dieser mit gespeichert wurde. Die Tweets liefern Texte mit einem spezifischem Inhalt (Lee et al., 2013). Diese Texte bestehen sowohl aus Wörtern als auch Links, Verweisen zu anderen Nutzern und Themen, die mit einem @ Symbol für Accounts oder einem Hashtag in der Form *#topic* indiziert werden. Lee et al. (2013) beschreiben zwei Vorteile der Analyse von georeferenzierten Tweets oder vergleichbaren Daten: die große Menge der Nutzer, die Daten veröffentlichen, sowie die Heterogenität der Daten, die von den Mitgliedern geteilt werden. Das ermöglicht die Analyse der unterschiedlichen Themen in den Beiträgen und auf Basis vorliegender Koordinaten den Vergleich mit den städtischen Räumen, mit denen diese Beiträge räumlich verbunden sind.

Die globale, regionale und lokale räumliche Verteilung von Tweets unterscheidet sich stark zwischen ländlichen und städtischen Räumen und folgt damit der Verteilung der Bevölkerungsdichte, wie dies z.B. anhand von Beiträgen zu Gesundheitsthemen in den USA durch Ghosh and Guha (2013) deutlich wird. Auch innerhalb städtischer Gebiete zeigt sich eine räumliche Differenzierung, die der Siedlungsstruktur folgt (Longley and Adnan, 2016). Kim et al. (2016) verwenden einen statistischen Ansatz um Punktmuster von Stichworten in Raum und Zeit zu kategorisieren. Lokale Themen in Form bestimmter Begriffen werden in der Studie auf Autokorrelation in Raum und Zeit untersucht. Die Untersuchung zeigt, dass bestimmte Begriffe in den Daten räumlich und semantisch stärker miteinander verbunden sind.

Zur weiteren Erläuterung dieses Zusammenhanges kann nochmals auf die Unterscheidung von Form und Funktion von Crooks et al. (2015) zurückgegriffen werden. Sie unterscheiden zwischen expliziten und impliziten Repräsentationen der beiden Begriffe. Die nutzergenerierten Daten von offenen Plattformen wie OpenStreetMap bilden demnach explizit städtische Formen ab. Implizite Formen stellen dagegen die räumlichen Daten von Geräten dar, die durch GPS und andere Mechaniken erstellt werden und die Formen räumlich wiederspiegeln. Ein Beispiel für explizite Funktionen sind die Apps der Location Based Social Networks, die auf Basis des Standortes Angebote an die Nutzer verschicken. Implizite Funktionen sind die Inhalte in den sozialen Netzwerken, die Funktionen der realen Welt wiederspiegeln. Die Beobachtung, Analyse und Modellierung der Formen (physische Strukturen und Geometrien) und insbesondere der Funktionen (Aktivitäten oder Prozesse) werden durch die Daten aus den sozialen Netzwerken ermöglicht.

Städtische Gebiete können somit mittels georeferenzierter Tweets charakterisiert werden. Das zeigen z.B. Lee et al. (2013) anhand der raum-zeitlichen Bewegung von Twitter-Nutzern in städtischen Räumen. Mit extrahierten Themen aus Twitter-Daten weisen Steiger et al. (2015b) nach, dass räumlich-zeitliche und semantische Muster von Aktivitäten am Arbeitsort und am Wohnort

identifiziert werden können. Dies wird mit einem räumlichen Vergleich mit Zensusdaten untermauert. Es zeigt sich, dass die Daten, die im digitalen Raum vom Menschen erzeugt werden, soziale und räumliche Muster im realen Raum wiederspiegeln. Anhand von Twitter-Daten können weiterhin auch Mobilitätsmuster identifiziert werden. Sowohl typische Pendlerzeiten als auch eine Differenzierung von Arbeitstagen und Wochenenden ist mit den Daten möglich (Steiger et al., 2016). Zhou and Zhang (2016) nutzen Twitter und Foursquare Daten um sechs unterschiedliche Typen von Aktivitäten in Raum und Zeit zu extrahieren. Die Analyse zeigt typische Muster einiger städtischer Funktionen. Eine Anwendung des Modells in einem Echtzeitsystem zeigt neben den verschiedenen Funktionen und funktionalen Grenzen einer Stadt auch, dass eine dynamische Analyse von realzeitlichen Daten anstelle eines statischen Datensatzes möglich ist.

Longley and Adnan (2016) leiten aus Tweets Variablen zu demografischen Merkmalen der Nutzer ab. Zu den Merkmalen gehören neben dem Alter und der Ethnie auch Informationen zum jeweiligen Umfeld, in dem ein Tweet abgeschickt wurde, zur Tageszeit des Tweets sowie zum Wohnort des Nutzers und der besuchten Staaten während des Untersuchungszeitraums. Vor diesem Hintergrund stellt sich jedoch die Frage, wie repräsentativ die Daten in Bezug auf die Bevölkerung in einem Raum sind. Gemeinsam ist allen geographischen Daten aus sozialen Medien ein starkes Rauschen im Sinne von Ungenauigkeiten in den Inhalten und den vorliegenden Koordinaten. Die Daten können wenig präzise und mit Unsicherheiten verbunden sein (Kim et al., 2016).

Neben den räumlichen und zeitlichen Mustern sollte die sozioökonomische Dimension laut Li et al. (2013) nicht unterschätzt werden. Sie haben einen genauen Blick auf die Verbindungen zwischen der räumlichen Konzentration von Daten aus sozialen Medien (Twitter, Flickr) und den sozioökonomischen Charakteristika der lokalen Bevölkerung in Kalifornien geworfen. Sie schließen aus den Daten, dass Einwohner mit höherer Bildung mit einer höheren Wahrscheinlichkeit Beiträge mit Raumbezug in sozialen Medien produzieren. Shelton et al. (2015) zeigen am Beispiel der Stadt Louisville (USA), wie sich mittels geocodierter Tweets Segregationsmuster der Bevölkerung in der Stadt abbilden lassen, indem sie die räumlichen Bewegungsmuster von besonders aktiven Nutzern untersucht haben. Die unterschiedlichen Aktionsräume segregierter Bevölkerungsgruppen in der Stadt zeigen sich auch in den räumlichen Mustern der Twitter-Daten.

Neben Twitter werden auch die Daten weiterer sozialer Medien für die Analyse von räumlichen Mustern verwendet. Einige Studien greifen auf Check-Ins von Foursquare zurück (Zhou and Zhang, 2016), nutzen georeferenzierte Fotos von Flickr (Garcia-Palomares et al., 2015, Salesses et al., 2013, Li et al., 2013) oder Beiträge von Instagram (Boy and Uitermark, 2016). Gemeinsam ist diesen Studien, dass die räumliche und teilweise zeitliche Verteilung von sozialen Aktivitäten in Form von Beiträgen in den jeweiligen Medien gemessen werden, um Rückschlüsse auf räumliche Cluster

menschlicher Aktivitäten im realen Raum – insbesondere mit dem Fokus auf Städte – zu treffen. Die dargestellten Methoden finden darüber hinaus aber auch Anwendung in anderen Bereichen, z.B. der Verkehrsforschung, der Gesundheitsforschung oder im Katastrophenmanagement (Steiger et al., 2015a).

1.3 Zielsetzung der Master Thesis

Unter der Annahme, dass Menschen, während sie sich im städtischen Raum bewegen, auch soziale Medien nutzen, müsste eine stärkere Nutzung von öffentlichen Plätzen, Infrastrukturen, Geschäften und anderen Angeboten sowie eine verstärkte Interaktion der Nutzer miteinander und mit ihrer Umgebung auch zu einer höheren Dichte an Beiträgen in sozialen Netzwerken an diesen Orten führen. Das gilt insbesondere dann, wenn Nutzer ihre Erfahrung im öffentlichen Raum in den digitalen Netzwerken mit anderen teilen möchten. Für die Master Thesis stellt sich in diesem Zusammenhang die Frage, wie Tweets für die Analyse von Forschungsfragen bezogen auf bestimmte Räume einer Stadt genutzt werden können. Die benötigten räumlichen und zeitlichen Informationen können, wie bereits dargestellt, aus klassischen Datenquellen, wie Zensusbefragungen, nicht oder nur eingeschränkt abgeleitet werden (Longley and Adnan, 2016, Shelton et al., 2015).

Dieser Master Thesis liegt somit die Annahme zugrunde, dass eine höhere Nutzungsintensität von städtischen Räumen dazu führt, dass mehr Menschen in diesen Räumen soziale Medien nutzen. Je mehr Menschen den städtischen Raum nutzen, desto stärker und diversifizierter sind ihre Aktivitäten in den sozialen Netzwerken. Ein stark frequentierter Raum mit einer hohen Aufenthaltsqualität weist dann mitunter eine höhere Diversifizierung an Themen in sozialen Medien auf, als ein stark frequentierter Raum mit geringer Aufenthaltsqualität. Die Anziehungskraft dieser Räume beeinflusst somit die Menge und Art der Beiträge in sozialen Medien unter der Bedingung, dass der reale Raum einen Einfluss auf den virtuellen Raum hat und beide miteinander verbunden sind. Ein stark genutzter städtischer Raum sollte demnach anhand eines höheren Aufkommens an Beiträgen in sozialen Medien zu unterschiedlichen Themen zu identifizieren sein. Es mag aber auch Räume geben, die zwar überdurchschnittlich stark frequentiert, jedoch wenig attraktiv in einem qualitativen Sinne sind. Hierbei handelt es sich z.B. um Räume, die auf Grund stark frequentierter Nutzungen im Umfeld (z.B. Shopping Center, Bahnhof) eine hohe Nutzungsdichte aufweisen, ohne dass hiermit eine hohe Aufenthaltsqualität einhergeht.

Die Frage ist, ob die unterschiedlichen Räume einer Stadt durch Themen charakterisiert werden können, die durch die Nutzer der sozialen Medien mit den jeweiligen Räumen in Form von georeferenzierten Beiträgen verbunden sind. Wenn die Erkennung von städtischen Strukturen und Prozessen möglich ist, dann kann eine Form der kollektiven Wahrnehmung eines Raumes daraus

abgeleitet werden (Jenkins et al., 2016). Das gilt es mittels quantitativer Analysemethoden sichtbar zu machen. Daran schließt sich dann auch die Frage an, ob die Ergebnisse solcher räumlichen Analysen im Weiteren dazu geeignet sind, Planungsprozesse in der Stadt zu unterstützten.

Das Ziel dieser Master Thesis vor diesem Hintergrund ist eine semantische und räumliche Analyse von Beiträgen aus sozialen Medien im Hinblick auf die Nutzung städtischer Räume. Die zentrale These ist, dass mittels Extrahierung von Themen aus georeferenzierten Tweets städtische Räume erkannt und charakterisiert werden können. Die Analyse der empirischen Daten unterteilt sich dazu in mehrere Teilschritte, die den folgenden Forschungsfragen untergeordnet sind:

1) Wie können geeignete Themen zur Beschreibung städtischer Räume in geocodierten Tweets identifiziert werden?
2) Wo gibt es räumliche Cluster, die sich spezifischen städtischen Strukturen oder Prozessen zuordnen lassen?
3) Welchen Zusammenhang gibt es zwischen raum-zeitlichen Mustern ausgewählter Themen und lokalen Ereignissen mit einem hohen Aufkommen an Menschen?

Zur Beantwortung dieser Fragen wird ein Korpus mit georeferenzierten Tweets aufgebaut, bereinigt und semantisch klassifiziert. Als Datensatz liegen Tweets mit Koordinaten für die Region Greater Manchester aus dem Jahr 2017 sowie für die Stadt Birmingham aus dem Jahr 2015 vor. Die Zuordnung der Tweets zu Themen erfolgt mittels Latent Dirichlet Allocation (LDA). Auf Basis der dann vorliegenden Themen erfolgt die räumliche Analyse der Daten. Ausgewählte Themen aus dem Corpus werden hinsichtlich ihrer räumlichen Verteilung in der Stadt analysiert. Dabei steht die Frage im Vordergrund, ob die Themen des ersten Teilschritts sich räumlich konzentrieren und ob sich ein räumlicher Zusammenhang zu bekannten städtischen Strukturen herstellen lässt. Die Beschreibung der einzelnen Methoden der Datenerhebung und –aufbereitung, die Anwendung der LDA und die daraus resultierenden Themen sowie die Vorstellung der verwendeten GIS-Methoden erfolgt in Kapitel 2. Die Vorstellung der räumlichen Analyse der Themen in Manchester steht in Kapitel 3.1 und für Birmingham in Kapitel 3.2.

Im nächsten Schritt werden dann die identifizierten räumlichen Cluster ausgewählter Themen mit weiteren Datengrundlagen verglichen. Dafür stehen Freizeitangebote und Geschäfte aus OpenStreetMap (OSM) als Punktdaten zur Verfügung. Die räumliche Lage dieser Daten wird mit den räumlichen Clustern der Themen verglichen. Neben der Frage, ob OSM-Daten zur Validierung der Themen geeignet sind, geht es auch darum, inwiefern sich beide Datensätze bei der Charakterisierung städtischer Räume ergänzen. Die Ergebnisse dieses Abgleichs werden in Kapitel 3.3 dargestellt.

Schließlich erfolgt im letzten Schritt der Analyse eine raum-zeitliche Betrachtung der Tweets anhand von zwei ausgewählten Themenkomplexen in Birmingham. Diese werden beispielhaft herausgegriffen, weil die semantische Analyse hierfür Themen mit einer hohen Signifikanz liefert: Verkehr und Arbeit sowie Fußball. Die Tweet-Häufigkeiten dieser Themen werden dabei nicht nur im räumlichen Zusammenhang von Bahnhaltestellen bzw. einem Stadion untersucht. Auch das Auftreten von unterschiedlichen Häufigkeiten im Tages- bzw. Jahresverlauf fließt in die Betrachtung mit ein. Somit soll anhand dieser Beispiele gezeigt werden, wie neben der rein räumlichen Perspektive auch die Betrachtung im zeitlichen Kontext mittels Tweets und ausgewählter Themen im Hinblick auf städtische Prozesse möglich ist. Es kann somit gezeigt werden, wie sich die räumlich-semantischen Muster auch in der zeitlichen Dimension im Zusammenhang mit bestimmten Events niederschlagen. Die Ergebnisse der Datenanalyse werden in Kapitel 3.4 dargestellt.

In der Diskussion der Ergebnisse in Kapitel 4 steht dann die kritische Betrachtung der empirischen Ergebnisse aus den einzelnen Analyseschritten im Vordergrund. Abschließend wird aber auch auf Anknüpfungspunkte für die Stadtentwicklung und Stadtplanung eingegangen. Die Grenzen und Einschränkungen, die mit den verwendeten Daten verbunden sind, stehen hierbei besonders im Fokus.

2 Methoden und Analyse

Die zentrale Herausforderung des Forschungsvorhabens ist es, aus dem „digitalen Rauschen“ der Tweets die wichtigsten Themen zu identifizieren und diese dann mit der räumlichen Betrachtungsweise zu verbinden. Dieses Kapitel beschreibt die Methoden, die zur Lösung dieser Herausforderung angewendet wurden. Das Kapitel ist daher in verschiedene Abschnitte untergliedert, die sich jeweils mit einem methodischen Teilschritt befassen. Der methodische Ansatz wird zunächst schematisch beschrieben (Kapitel 2.1). Die zwei folgenden Kapitel stellen den Untersuchungsraum und die Methoden der Datenerhebung mittels eines Crawlers vor (Kapitel 2.2) und beschreiben die Aufbereitung des Korpus mit Methoden der Textanalyse (Kapitel 2.3). Die Methodik des Topic Modelling mit dem Latent Dirichlet Allocation Algorithmus wird in einem eigenen Kapitel dargestellt (Kapitel 2.4). Die Identifikation räumlicher Cluster folgt mittel GIS Methoden, die im letzten Methodenkapitel erklärt werden (Kapitel 2.5).

2.1 Forschungsansatz

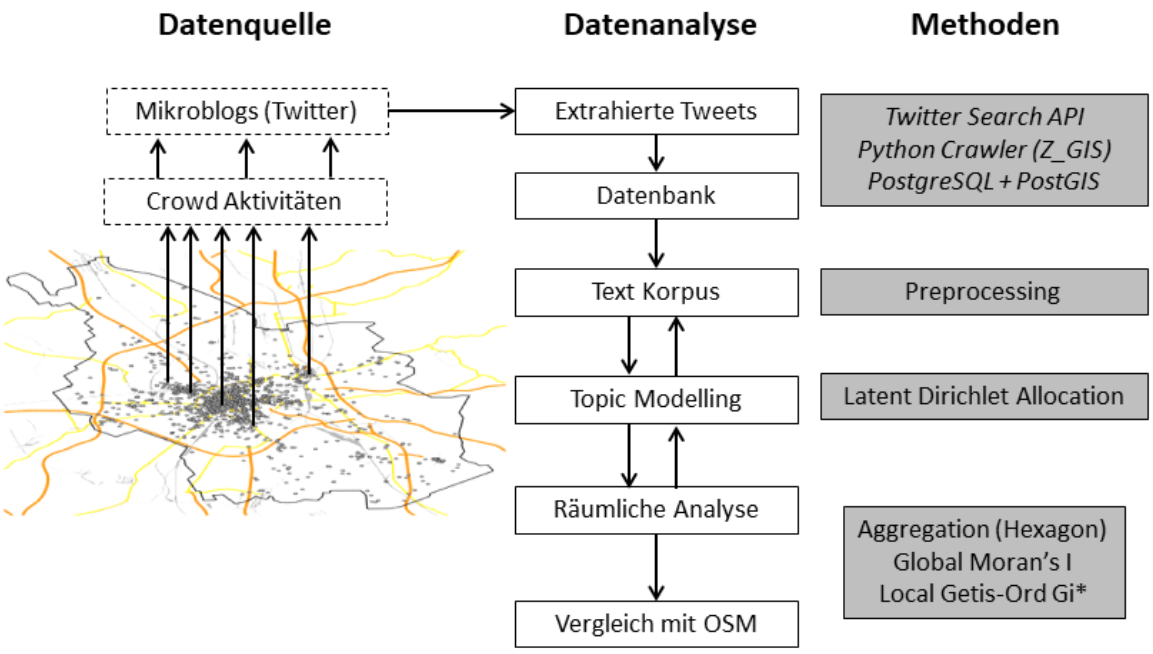


Abbildung 1: Forschungsmodell

Abbildung 1 visualisiert den Forschungsansatz. Die Agenda lehnt sich am Information Mining Ansatz an, der von Tsou and Leitner (2013) beschrieben wird. Anhand des Literaturstudiums wurden zunächst geeignete Methoden identifiziert und erprobt. Der Forschungsansatz nutzt eine top-down Perspektive, bei der Daten zunächst gesammelt und dann mit Text Mining- und GIS-Methoden

analysiert werden. Von den Ergebnissen der Datenanalyse aus erfolgen dann Rückschlüsse auf die realen Räume in der Stadt.

Der verwendete Python Crawler zur Speicherung der Daten aus der Twitter API stellt alle Methoden zur Datenextraktion bereit. Die Tweets wurden durch den Crawler extrahiert und die Rohdaten in einer PostGIS-Datenbank gespeichert. Der Crawler basiert auf Python und wird durch den Interfakultären Fachbereich Geoinformation – Z_GIS an der Universität Salzburg bereitgestellt.

Die Auswahl der weiteren Methoden für das Forschungsprojekt erfolgte in der beschriebenen Form, weil die einzelnen Methoden in der Literatur gut beschrieben sind und sich für die gewählten Forschungsfragen eignen. Wie in Kapitel 1.2 dargestellt, führen Topic Modelling Methoden zu validen Resultaten für räumliche Analysen. Eine Herausforderung war es jedoch, die Methoden in der Statistiksprache R mittels vorhandener Pakete umzusetzen und die einzelnen Analyseschritte miteinander zu verknüpfen. Die gesamte Datenaufbereitung und Textanalyse erfolgte in R. Die Basismethoden in R und die enorme Anzahl an zusätzlichen Paketen bieten eine Vielzahl an Algorithmen für unterschiedliche Anwendungsfälle. Für die Aufbereitung und Analyse von Textdokumenten einschließlich des LDA Algorithmus von Blei et al. (2003) stehen mehrere Pakete zur Verfügung. In jedem der folgenden Kapitel werden die verwendeten Pakete mit den spezifischen Methoden aufgeführt und die jeweiligen Autoren genannt. Für die Analysen mittels GIS-Methoden wurde auf ArcGIS Pro zurückgegriffen, das durch seine graphische Oberfläche eine einfachere Anwendung von GIS-Methoden und eine leichtere Visualisierung der Ergebnisse erlaubt. Prinzipiell wäre aber auch eine vollständige Umsetzung der Analyse einschließlich der Kartendarstellungen in R möglich.

Die gesammelten Daten wurden aufbereitet und bereinigt. Dieser Schritt von den Rohdaten zu einem auswertbaren Korpus stellt die Grundlage für den Topic Modelling Ansatz dar, auf dem wiederum die räumliche Analyse der identifizierten Themen aufbaut. Die Auswertung der Daten erfordert eine Analyse in Schleifen, wodurch die Ergebnisse einzelner Analyseschritte geprüft und ggf. verworfen werden. Daraus folgen dann wiederum Anpassungen in den vorherigen Analyseschritten, die zu angepassten Ergebnissen und zu einer erneuten Prüfung führen. Die einzelnen Analyseschritte sind wie folgt:

- Aufbereitung und Bereinigung der Rohdaten (Filter, Stoppwörter, Fehlerprüfung, manuelle Bereinigungsschritte, Stemming)
- Erstellung einer Dokument-Begriff-Matrix als Datenbasis für das Topic Modelling
- Durchführung einer semantischen Analyse mittels des Latent Dirichlet Allocation Algorithmus; Prüfung der Ergebnisse und ggf. angepasste Bereinigung der Rohdaten und erneute Durchführung

- Erprobung und schließlich Auswahl verschiedener räumlicher und statistischer Analysemethoden zur Untersuchung der identifizierten Themen; Prüfung der Ergebnisse und ggf. angepasste Bereinigung der Rohdaten sowie erneute semantische Analyse mit angepassten Parametern
- Transformation der Daten in analytische Formate (Karten, Grafiken, Tabellen, etc.)
- Interpretation der Resultate im Abgleich mit identifizierten städtischen Räumen sowie weiterer Datengrundlagen aus OpenStreetMap; Überprüfung von raum-zeitlichen Mustern
- Kritische Diskussion der Ergebnisse

2.2 Datensammlung

Das Twitter Application Programming Interface (API) liefert eine Stichprobe von 1 % aller Tweets. Laut Jenkins et al. (2016) und Longley et al. (2015) haben davon wiederum 1 bis 2 % der Tweets Koordinaten. Von der Twitter API wurden für dieses Forschungsprojekt nur diejenigen Tweets ausgelesen, die eine Koordinate beinhalten. Die API stellt weitere Tweets mit Ortsangaben zur Verfügung. Diese Ortsnamen sind jedoch unzuverlässig für eine Datenanalyse, die sich auf einen großen Maßstab mit kleinen Teilgebieten fokussiert (Kim et al., 2016). Die Genauigkeit dieser Angaben ist sehr heterogen und basiert zum Großteil auf größeren räumlichen Einheiten wie Stadtteilen, Städten oder sogar Regionen und Staaten (Cheng et al., 2010). Dadurch kann jedoch ein Teil der von der API zur Verfügung gestellten Tweets mit räumlichen Informationen gar nicht genutzt werden.

Ein Tweet bestand bis zum Oktober 2017 aus maximal 140 Zeichen. Nach einer Testphase mit ausgewählten Nutzern ab September 2017 hat Twitter im November 2017 die Zahl der Zeichen auf 280 erhöht. Mehr Zeichen bedeuten, dass natürlich auch mehr Wörter als vorher in einem Tweet untergebracht werden können. Für die vorliegende Master Thesis war dies jedoch noch nicht relevant, da die Änderung der Begrenzung nach Abschluss der Datenerhebung erfolgt ist.

Für die Untersuchung wurde die Metropolregion Greater Manchester als eine Beobachtungsregion ausgewählt. Die Region untergliedert sich in 10 Metropolitan Boroughs, wovon einer die Stadt Manchester selbst ist. Parallel zur Datenerhebung in der Region wurden zunächst auch Tweets im Gebiet der Stadt Köln erhoben. Der Datensatz wurde jedoch nach mehrfacher Evaluation verworfen. Die wichtigsten Gründe dafür sind die deutlich geringere Anzahl an Tweets im Gegensatz zu Manchester und somit eine geringere Dichte und räumliche Abdeckung sowie eine hohe Anzahl an vermeintlichen Bots und eine Konzentration von Beiträgen in unterschiedlichen Sprachen im Zentrum. Die Zahl der auszuwertenden Tweets hätte sich nach einer Bereinigung voraussichtlich stärker verringert als bei den englischsprachigen Datensätzen. Die angestrebte Kombination aus

semantischen und räumlichen Analysen wäre mit dem Teildatensatz für Köln nur schwierig oder gar nicht umzusetzen gewesen.

Für die Stadt Birmingham liegt ein weiterer Datensatz aus den Jahren 2012 bis 2015 vor, der vom Interfakultären Fachbereich Geoinformation – Z_GIS an der Universität Salzburg bereitgestellt wurde. Aus diesem Datensatz wurden 826.282 Tweets aus dem Jahr 2015 extrahiert. Der Rohdatensatz für die Metropolregion Manchester besteht dagegen nur aus 99.884 Tweets aus dem Zeitraum April bis Oktober 2017. Andere Studien haben in vergleichbaren oder sogar kürzeren Zeiträumen eine weitaus höhere Anzahl an Tweets sammeln können, als dies in der Region Manchester der Fall war. Die genauen Gründe für die unterschiedlichen Stichprobengrößen konnten nicht ermittelt werden. Möglicherweise ist das Aufkommen an Tweets in einigen Regionen deutlich höher als in anderen. Viele Studien stützen sich auf Erhebungen aus stark frequentierten Städten mit einem sehr hohen Besucheraufkommen wie New York, Boston, Chicago, Los Angeles, London oder Singapur (Zhou and Zhang, 2016, Longley and Adnan, 2016, Jiang et al., 2016, Jenkins et al., 2016, Steiger et al., 2015b). Bis Ende 2016 lieferte die API bei räumlichen Abfragen zudem nur Tweets mit Koordinaten. Seit Anfang 2017 werden jedoch, wie geschildert, auch Tweets mit Ortsangaben geliefert. Da diese Tweets bei der Erhebung herausgefiltert wurden, hat sich dadurch womöglich die Stichprobe weiter verringert.

Abbildung 2 zeigt die Dichte der erhobenen Tweets im Untersuchungsraum um die Stadt Manchester. Abbildung 3 stellt die Dichte mit einer anderen Klassierung für die Stadt Birmingham dar. Die Karten wurden mittels Kernel Density Estimation (KDE) erzeugt und basieren auf den vollständig bereinigten Datensätzen (siehe Kapitel 2.3). Die Dichtewerte beider Karten sind nicht miteinander zu vergleichen, da die zugrunde liegende Stichprobe für Birmingham etwa 7,5-mal so groß ist, als die Stichprobe für Manchester. Somit liegen im Ergebnis andere Dichtewerte vor. Die Dichten spiegeln die Siedlungsstruktur der Region Greater Manchester sowie der Stadt Birmingham sehr gut wieder. Die höchste Dichte in Manchester befindet sich im Zentrum und hohe Dichten können weiterhin in den kleineren Städten der Region, wie z.B. Salford und Trafford im Westen oder Stockport im Süden identifiziert werden. In Birmingham werden für die Analyse nur die Stadt und ihr unmittelbares Umfeld berücksichtigt. Gut nachvollziehbar ist darum im Vergleich, dass in den peripheren Bereichen der Region Greater Manchester noch Lücken ohne Tweets vorhanden sind. Das gibt es in der Kernregion Birmingham hingegen nicht. Hier wird aber zumindest das Zentrum-Peripherie-Gefälle anhand der Tweetdichte deutlich. Insbesondere in den Stadtkernen gibt es Dichtewerte, die alle anderen Werte in der Region deutlich übersteigen. In Manchester im Northern Quarter, einem beliebten Viertel, liegt das Maximum bei einer Dichte von knapp unter 10.000 Tweets je km². In Birmingham liegt hingegen das Maximum im Gebiet der Birmingham New Street Station

und dem Grand Central Shopping Centre bei einer Dichte von knapp 75.000 Tweets je km². Hier befinden sich ein wichtiger Bahnhof und ein großes Einkaufszentrum unmittelbar übereinander.

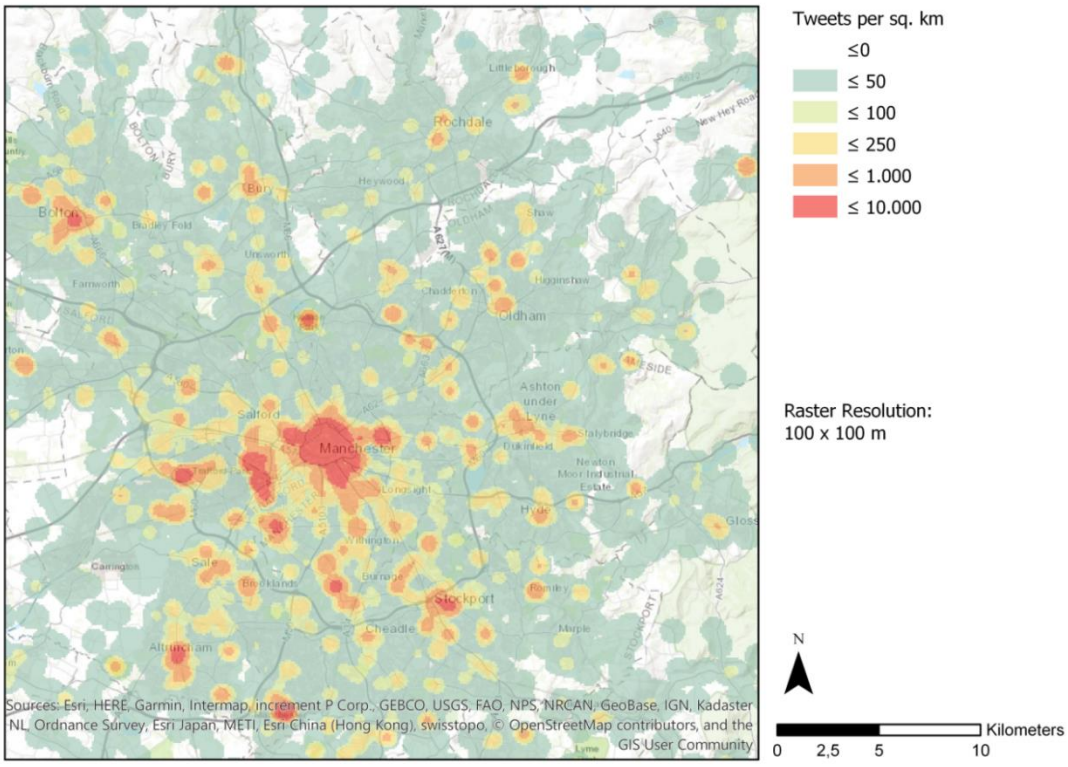


Abbildung 2: Erhebungsraum Manchester und Dichte der Tweets (bereinigter Datensatz)

Abbildung 3: Erhebungsraum Birmingham und Dichte der Tweets (bereinigter Datensatz)

In der Manchester-Stichprobe kommen im Durchschnitt 4,9 Beiträge auf einen Nutzer. Der Birmingham-Datensatz weist dagegen einen Durchschnitt von 16,8 Beiträgen je Nutzer bei einem längeren Erhebungszeitraum auf. Die Beiträge verteilen sich jedoch sehr heterogen auf die einzelnen Accounts. Bots und sogenannte Power User teilen auf Twitter eine große Menge an Beiträgen. Dies kann die Analyse erschweren, da diesen Nutzern und den teilweise sehr einheitlich geschriebenen Tweets mit gleichen oder nahezu gleichen Wörtern ein großes Gewicht im Datensatz zukommt.

Im Manchester-Datensatz wurden sechs der acht Accounts mit der höchsten Anzahl an Beiträgen gelöscht. Im Birmingham-Datensatz konnten ebenfalls mehrere der besonders aktiven Nutzer eindeutig als Bots identifiziert werden. Auf Grund der Vielzahl an Beiträgen handelt es sich bei solchen Accounts sehr wahrscheinlich um Bots, auch wenn das nicht immer zweifelsfrei bestimmt werden kann. Das hat vor allem mit der Art der geschriebenen Beiträge zu tun, die teilweise versuchen, unterschiedliche Schreibweisen menschlicher Nutzer zu imitieren. Zwei der gelöschten Bots aus den Manchester-Daten informieren über Angebote einer Sprachschule. Zwei weitere Bots veröffentlichen die gespielten Musikstücke in einem Fitnessstudio. Ein ähnlicher Bot wurde im Birmingham-Datensatz gefunden. Ein weiterer Nutzer in Manchester umfasst ausschließlich Beiträge in arabischen Schriftzeichen, die auch auf Grund der Sprache nicht ausgewertet werden können und durch die hohe Anzahl wahrscheinlich ebenfalls von einem Bot stammen. Auffällig sind weitere Accounts insbesondere dann, wenn die Beiträge nur von einem Standort veröffentlicht werden, wie z.B. Kurzmeldungen zu Fußballspielen mit drei fixen Koordinaten in Manchester. Auf die letztendlich gelöschten Nutzer entfallen im Manchester-Datensatz 6.307 Tweets. Das sind etwa 6,3 % der Tweets im Rohdatensatz. Im Birmingham-Datensatz wurden 49.386 Tweets (5,9 %) gelöscht.

Für die Analyse stehen ohne diese entfernten Beiträge 93.577 Tweets aus der Metropolregion Manchester und 776.896 Beiträge aus Birmingham zur Verfügung. Abbildung 4 und Abbildung 5 zeigen die Verteilung der Tweets auf die Nutzer als Lorenz-Kurve für beide Datensätze. Dabei ist die Zahl der Beiträge gegen die Nutzer aufgetragen. Aus den Lorenz-Kurven kann abgelesen werden, dass rund 20 % der Tweets von etwas weniger als 5 % der Accounts veröffentlicht werden. Die Hälfte der Tweets stammt von nur rund einem Viertel der Nutzer. Der Gini-Koeffizient für die Verteilung in Manchester beträgt 0,32 und in Birmingham 0,33. Das bedeutet: Es gibt im Datensatz viele Nutzer mit nur sehr wenigen Beiträgen sowie einige Nutzer mit sehr vielen Beiträgen. Beide Daten unterscheiden sich zudem im relativen Verhältnis von Nutzern zu Beiträgen nicht, auch wenn die absoluten Fallzahlen auf Grund der beschriebenen Stichprobenunterschiede nicht direkt vergleichbar sind.

Durch Herausfiltern der aktivsten Bots und ausgewählter Accounts konnte jedoch ein guter Kompromiss gefunden werden. Weitere Bots wurden nicht identifiziert. Dies würde eine

automatisierte Erkennung solcher Accounts und Beiträge erfordern. Das gestaltet sich jedoch insbesondere als schwierig, weil Bots u.a. darauf getrimmt werden, menschliches Verhalten zu imitieren (Haustein et al., 2016).

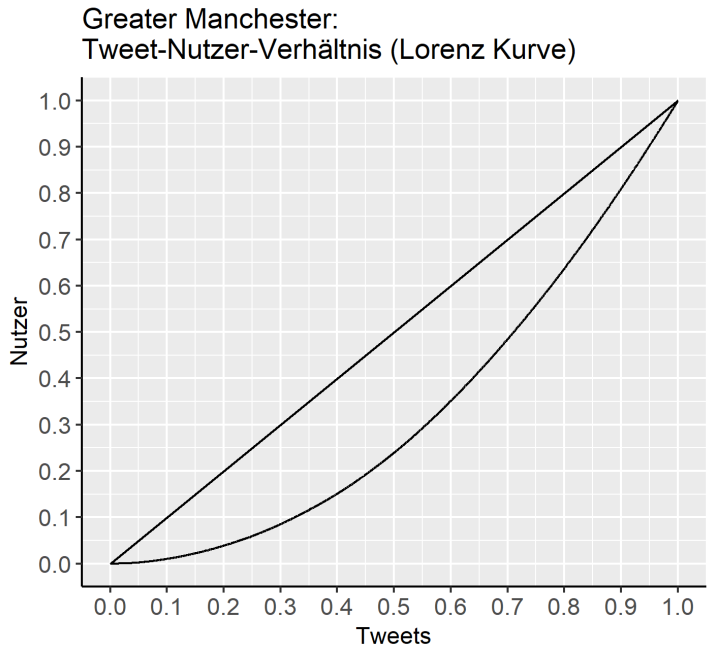


Abbildung 4: Lorenz Kurve des Tweet-Nutzer-Verhältnisses in Manchester

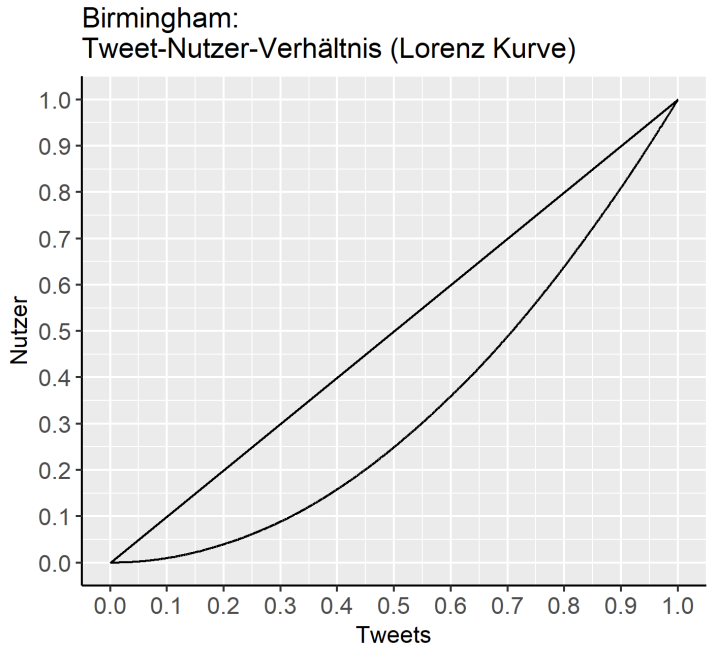


Abbildung 5: Lorenz Kurve des Tweet-Nutzer-Verhältnisses in Birmingham

2.3 Aufbau des Korpus

Die einzelnen Tweets sind als Rohdatensatz nur schwer zu analysieren. Die Texte bestehen aus unterschiedlichen Inhalten, die neben Wörtern auch URLs, Abkürzungen, Smileys und Zeichen aus anderen Schriftarten enthalten. Hier ist zunächst eine Bereinigung der Texte erforderlich. Das Text Mining Package für R (tm package) von Feinerer and Hornik (2017) stellt die benötigten Methoden zum Bereinigen bereit. Das tm Paket stellt eine Middleware zwischen R und dem später in der Analyse verwendeten topicmodels Paket dar (Ponweiser, 2012), womit dann die LDA Analyse durchgeführt wird (siehe Kapitel 2.4). Im Folgenden werden die einzelnen Bereinigungsschritte erläutert. Vergleichbare Methoden zur Bereinigung der Daten werden auch in anderen Studien beschrieben und sind etabliert (Resch et al., 2016, Steiger et al., 2015b, Longley et al., 2015).

Zunächst werden alle Großbuchstaben zu kleinen Buchstaben umgewandelt, um die unterschiedlichen Eingaben der Nutzer zu vereinheitlichen (insbesondere Wörter am Satzanfang und Tippfehler). Dies kann in Einzelfällen zu falschen Assoziationen führen. Die Unterscheidung von Eigennahmen mit Großbuchstaben am Wortbeginn im Vergleich zu Nomen mit kleinen Buchstaben in der englischen Sprache wird dadurch aufgehoben. Generell ist dies aber in der englischen Sprache weniger problematisch als im Deutschen, weil hier generell alle Nomen mit einem Großbuchstaben beginnen und nach der Bereinigung nicht mehr von anderen Wörtern unterschieden werden können (z.B. ich spiele und die Spiele, es ist grün und das Grün). Die spätere semantische Analyse kann hier nicht mehr zwischen den unterschiedlichen Wortformen unterscheiden.

Alle URLs werden aus den Daten entfernt, da diese nicht mit den gewählten Methoden analysiert werden können. Des Weiteren werden Zahlen aus den Tweets entfernt. Zahlen können nicht in Bezug zu anderen Wörtern gesetzt werden, weil die Reihenfolge der Begriffe in der Analyse keine Rolle spielt. Die Zahlen stellen für sich keine untersuchbare Einheit in dieser Analyse dar. Einige Tweets enthalten Sonderzeichen, teilweise mit spezifischen Funktionen (z.B. # für Themen oder @ für Nutzernamen), oder Buchstaben aus anderen Sprachen. Alle Sonderzeichen und nicht englische Buchstaben werden aus den Tweets entfernt. Doppelte Leerzeichen werden darüber hinaus durch einfache Leerzeichen ersetzt.

Es werden alle Wörter mit einem oder zwei Buchstaben entfernt. Weiterhin werden sogenannte Stoppwörter aus dem Datensatz entfernt. Meyer et al. (2008) beschreiben Stoppwörter als “words that are so common in a language that their information value is almost zero, in other words their entropy is very low.” Im Deutschen fallen hierunter auch sogenannte Füllwörter. Jedoch nicht nur Füllwörter sind als Stoppwörter zu betrachten. Grammatikalisch betrachtet handelt es sich vor allem um Präpositionen, Adverbien und Partikel. Diese Wörter sind zum Teil wichtig zum Verständnis eines Satzes. Jedoch werden in der Analyse keine Sätze und Satzstrukturen untersucht. Zunächst wird für

die Bereinigung die vom tm Paket bereitgestellte Stoppwortliste für die englische Sprache verwendet. Bei den ersten Testdurchläufen der Analyse wurden dann weitere Begriffe identifiziert, die später aus dem Datensatz entfernt wurden. Die Liste der zusätzlich entfernten Begriffe ist im Anhang zu finden (Tabelle 9).

Der letzte Bereinigungsschritt ist das Stemming. Der Begriff leitet sich vom Wortstamm ab, denn es geht um die Reduktion auf die jeweiligen Wortstämme (Feinerer, 2010). Im tm Paket ist der Stemming Algorithmus von Porter für verschiedene Sprachen integriert (Meyer et al., 2008). Es handelt sich um eine wichtige Methode für alle Text Mining Analysen. Die Deklination von Wörtern führt zu unterschiedlichen Wortformen desselben Wortstammes, die wiederum bei den hier verwendeten Methoden nicht als der gleiche Wortstamm erkannt werden (z.B. im Englischen: *play, played, plays, playing*). Das Stemming hat jedoch auch einige Nachteile. Stemmingalgorithmen können zu einer zu starken oder zu geringen Reduzierung des Wortes führen (overstemming bzw. understemming). Außerdem ist es möglich, dass zwei in der Bedeutung unterschiedliche Wörter auf den gleichen Wortstamm reduziert werden und damit dem Algorithmus einen einheitlichen Begriff vortäuschen (z.B. im Englischen: *I liked* und *something like* oder *I play* and *the play*). Dies geschieht in einzelnen Fällen, wie dem englischen Wort *like*, aber auch schon ohne Stemming.

Nach dem Stemming kann in einem zweiten Schritt eine Vervollständigung des Wortstammes zu einem deklinierten und damit besser lesbaren Wort erfolgen, z.B. die häufigste Deklination im Datensatz oder eine zufällige Deklination aus dem Datensatz. Der Algorithmus im tm Paket führte jedoch in verschiedenen Testdurchläufen zu Fehlern. Somit werden nur die Wortstämme in dieser Arbeit ausgewiesen, aber bei Bedarf zur besseren Lesbarkeit vom Autor vervollständigt (siehe Kapitel 3).

Sobald die einzelnen Wörter im Korpus bereinigt und auf ihren Wortstamm reduziert sind, erfolgt die Tokenisierung. Das bedeutet, die Texte (Tweets) werden in die kleinste Einheit zerlegt, die für die Analyse benötigt wird: das Wort. Das Vorhandensein oder Nichtvorhandensein von Wörtern in einem Dokument bzw. Tweet wird wiederum in einer Matrix gespeichert, die nur noch Zahlen enthält. Das tm Paket liefert die dafür benötigte Funktion, die aus dem Korpus eine Dokument-Wort-Matrix erstellt. In dieser Matrix beschreibt jede Zeile genau ein Dokument und jede Spalte genau einen Begriff. Die Matrix hat daher so viele Zeilen wie es im Korpus Dokumente gibt und so viele Spalten wie es im Korpus unterschiedliche Begriffe gibt. Die Matrix umfasst alle Wörter, die im Korpus vorhanden sind. Jedes Dokument bildet in der Matrix einen Vektor bzw. eine Anordnung von Wörtern mit spezifischen Gewichten. Jede Zelle dieses Vektors in der Matrix enthält die absolute Häufigkeit des Begriffes im Dokument als Zahlenwert. Man spricht in diesem Zusammenhang auch von einem *bag-of-words*-Modell. Ein Text wird als ein Set aus Wörtern repräsentiert. Die Grammatik

und die Reihenfolge der Wörter spielt dabei keine Rolle. Für die Klassifizierung der Dokumente ist nur das Vorhandensein von Wörtern bzw. deren Häufigkeit entscheidend (siehe Kapitel 2.4).

Aus der Dokument-Wort-Matrix werden alle Wörter entfernt, die insgesamt nur einmal im Datensatz vorkommen. Anhand der Matrix können im letzten Schritt dann noch diejenigen Dokumente identifiziert werden, die keine Begriffe mehr enthalten. Dies können z.B. Tweets sein, die nur aus Smileys und einer URL bestanden haben und durch die Bereinigung keine Elemente mehr enthalten. Weiterhin werden auch diejenigen Tweets aus dem Korpus entfernt, die nach der Bereinigung nur noch aus einem Wort bestehen.

Nach der Erstellung des Korpus verbleiben 91.952 Dokumente bzw. Tweets aus Manchester im Datensatz. Es wurden 1.625 Tweets aus dem Datensatz entfernt, die letztendlich nach den beschriebenen Bereinigungsschritten keine Wörter oder nur ein Wort enthielten. Die Matrix enthält als Ausgangsdatensatz für das Topic Modelling 6.712 unterschiedliche Wörter bzw. Wortstämme. Der Birmingham-Datensatz wurde um 74.328 Tweets bereinigt. Im Ergebnis liegen 702.568 Dokumente mit 4.911 unterschiedlichen Wörtern bzw. Wortstämmen als Ausgangsdatensatz für das Topic Modelling vor.

2.4 Latent Dirichlet Allocation (LDA)

Das Ziel der Anwendung von Topic Modelling Methoden und damit auch des Latent Dirichlet Allocation Algorithmus ist die Analyse von zunächst nicht sichtbaren, semantischen Mustern in einem Korpus aus Dokumenten. Jedes Dokument soll automatisiert mit dem jeweiligen Thema oder mehreren Themen annotiert werden. Die jeweiligen Wahrscheinlichkeiten der Themenzuordnungen stellen eine explizite Repräsentation eines Dokumentes dar. Die LDA ist somit eine dimensionsreduzierende Methode, die selbstständig Themen in einer Sammlung von Dokumenten identifiziert. Sie kann daher zu den Methoden des Machine Learning gezählt werden. Die Methode wurde von Blei et al. (2003) eingeführt. Viele Studien haben bereits die LDA als Methode verwendet und das insbesondere auch bei der Klassifizierung von Tweets (siehe Kapitel 1.2). Die Methode kann deshalb als etabliert gelten.

Blei et al. (2003) beschreiben drei zentrale Begriffe, die der LDA zugrunde liegen: Wort, Dokument und Korpus. Ein Wort ist die Basiseinheit eines diskreten Datensatzes und jedes Wort im Vokabular kann im Datensatz als *{1, …, V}* indiziert werden. Jedes Dokument wird als eine Sequenz von *N* Wörtern oder als w = (w_1, w_2, …, w_N) beschrieben. Ein Korpus ist eine Auswahl von *M* Dokumenten in der Form *D* = (w_1, w_2, …, w_M). Ein Dokument wird zudem durch eine zufällige Mischung verborgener Themen repräsentiert. Ein Thema kann wiederum durch ein Set an Wörtern beschrieben werden. Ein

Wort kann daher in unterschiedlichen Themen vorkommen, dann aber mit einem unterschiedlichen Set an weiteren Wörtern.

Die graphische Repräsentation dieses Modells ist in Abbildung 6 dargestellt. Die äußere Box repräsentiert die *M* Dokumente und die innere Box die Themen und Wörter in einem Dokument. Blei et al. (2003) spricht von drei Stufen der LDA Repräsentation:

- α ist ein Parameter für die Dirichlet A-priori-Wahrscheinlichkeit der Themenverteilung in den Dokumenten. β ist ein Parameter für die Dirichlet A-priori-Wahrscheinlichkeit der Wortverteilung in den Themen. Diese Parameter werden auf der Stufe des Korpus erzeugt.
- Die Variable ϑ liegt auf der Stufe des Dokuments vor. ϑ_m ist die Themenverteilung für ein Dokument *m*.
- Die Variablen *z* und *w* bestehen auf der Stufe des Wortes und liegen einmal für jedes Wort in jedem Dokument vor. z_{mn} ist das Thema des n-ten Wortes im Dokument m und w_{mn} ist ein spezifisches Wort.

Nur die Wörter sind beobachtbar und die anderen Variablen sind latent. Da eine Dirichlet A-priori-Wahrscheinlichkeit für die Verteilung der Themen in den Dokumenten und der Wörter in den Themen unterstellt wird, spricht man von Latent Dirichlet Allocation (LDA).

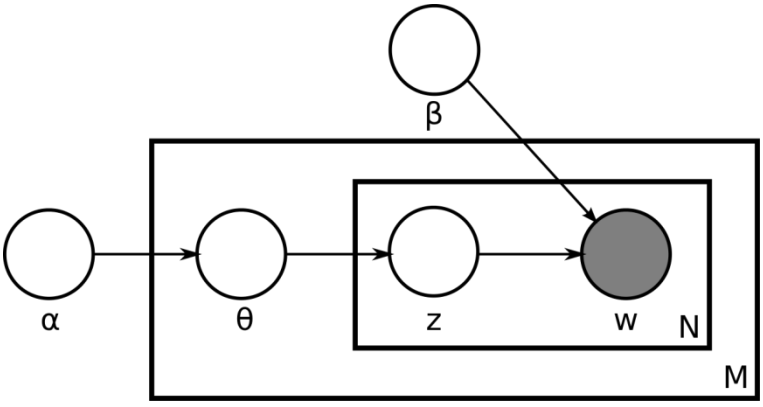


Abbildung 6: Graphische Repräsentation der LDA; Darstellung nach Blei et al. (2003)

Topic Modelling Methoden klassifizieren automatisch große Sammlungen von Dokumenten, wie z.B. Tweets, zu Themen. Die Grundannahme ist, dass die Themen bereits vor den Dokumenten vorhanden waren (Blei, 2012). Die LDA Methode identifiziert diese Themen, indem sie iterativ die relative Bedeutung von Themen in einem Dokument und Wörtern in Themen abgleicht. Es handelt sich um ein generatives Wahrscheinlichkeitsmodell für einen Korpus (Blei et al., 2003).

Das topicmodels Paket stellt den LDA Algorithmus von Blei et al. (2003) in R zur Verfügung (Hornik and Grün, 2011). Das Paket kann Korpora analysieren, die mit dem tm Paket aufbereitet wurden (siehe Kapitel 2.3). Der verwendete Algorithmus nutzt Gibbs-Sampling, um die Verteilung der

Themen über die Dokumente und die Verteilung der Wörter über die Themen zu ermitteln. Mit einem gegebenen Wert von *k* für die Anzahl der Themen wurden in der Analyse fünf unabhängige Durchläufe mit unterschiedlichen Startpunkten für das Gibbs-Sampling berechnet und jeder Durchlauf umfasst dabei 4.000 Iterationen als Burn-In-Phase sowie weitere 2.000 Iterationen, bis das Modell stabil ist. Vergleichbare Berechnungen mit dem Birmingham-Datensatz sind mit einer so großen Anzahl an Iterationen und der verfügbaren Hardware in R nicht möglich, so dass hier die Parameter auf 1.000 Iterationen als Burn-In Phase und weitere 1.000 Iterationen zur Stabilisierung des Modells reduziert werden mussten. Die R-Umgebung kommt hier auch bei einem leistungsfähigen Desktop-Rechner an ihre Grenzen.

Weitere Eingangsvariablen müssen a priori festgelegt werden. Nach Griffiths and Steyvers (2004) sollte $\alpha = 50/k$ und $\delta = 0{,}1$ sein, um einen hochgradig granulierten Korpus in Themen einzuteilen. Das topicmodel Paket nutzt diese Vorgaben als Standardwerte (Hornik and Grün, 2011). Es werden je Modell (mit festgelegten Parametern k, α, δ) eine fixe Anzahl von Durchläufen mit zufälliger Initialisierung durchgeführt. Im Ergebnis wird jeweils das Modell mit dem besten Ergebnis der log-likelihood Werte gespeichert und die übrigen Modelle verworfen.

Das prinzipielle Vorgehen des Algorithmus ist in die folgenden Schritte unterteilt: Zunächst wird in jedem Dokument jedes Wort einem zufälligen Thema zugewiesen. Diese zufällige Setzung ist mit einer hohen Wahrscheinlichkeit nicht korrekt. Der Algorithmus nimmt daher in jedem Dokument für jedes Wort und jedes Thema zwei Berechnungen vor: das Verhältnis der Wörter im Dokument, die derzeit zum Thema zugeordnet sind sowie das Verhältnis der Zuweisungen zum Thema über alle Dokumente, die von diesem Wort stammen. Alle anderen Zuweisungen sind in diesem Moment fixiert und anhand der statistischen Verteilung der Themen wird dem Wort ein neues Thema zugewiesen, dessen Wahrscheinlichkeit sich mit der derzeitigen Verteilung im Modell deckt.

Dieser Vorgang wird durch den Algorithmus mehrfach wiederholt, bis die Zuweisungen der Themen ein stabiles Niveau erreichen. Im Ergebnis weist ein Dokument mehrere Themen auf und jedes Dokument besteht aus einem unterschiedlichen Verhältnis von Themen. Jedes Thema kann in der gesamten Sammlung von Dokumenten, aber nicht in jedem Dokument, gefunden werden. Jedes Thema wird dabei durch bestimmte Wörter identifiziert. Die LDA nimmt somit an, dass die Wörter von unterschiedlichen Themen mit festgelegten Verteilungen stammen.

Die Zahl der Themen *k* muss vor der Berechnung festgelegt werden. Bei anderen Fragestellungen und Korpora ergibt sich mitunter die Zahl der Themen aus dem Zusammenhang, z.B. bei der Analyse von Zeitschriftenartikeln zu einem bestimmten Themenbereich. Jedoch ist bei einem Twitter-Datensatz die Zahl der Themen für den Anwender nicht vorab bekannt. Die Anzahl der Themen soll aber auch

nicht willkürlich festgelegt werden. Darum sollten zunächst verschiedene Modelle mit unterschiedlichen Werten für *k* gerechnet werden (Hornik and Grün, 2011).

Ein Ansatz zur Ermittlung der optimalen Zahl für *k* stellt die harmonic mean Methode dar (Griffiths and Steyvers, 2004). Die Implementierung dieser Methode in R wird von Ponweiser (2012) beschrieben. Hierbei wird das Maximum des harmonischen Mittelwertes (harmonic mean) der log Wahrscheinlichkeiten der Daten für verschiedene Werte von *k* berechnet. Das Ziel der Methode ist es, den Wert für *k* solange zu steigern, bis der resultierende Wert wieder absinkt. In verschiedenen Modelldurchläufen können so unterschiedliche Werte von *k* getestet werden.

Abbildung 7 zeigt das Ergebnis der Methode beispielhaft für unterschiedliche Werte von *k* im Manchester-Datensatz. Der optimale Wert von *k* entspricht nach der Methode dem Maximum der Kurve. Die Kurve impliziert eine relativ hohe Anzahl an Themen im Datensatz bei *k = 60*. Dies widerspricht sich aber deutlich mit anderen Studien. Steiger et al. (2015b) kamen beispielsweise bei dieser Methode mit einem vergleichbaren Twitter-Datensatz auf nur 11 Themen.

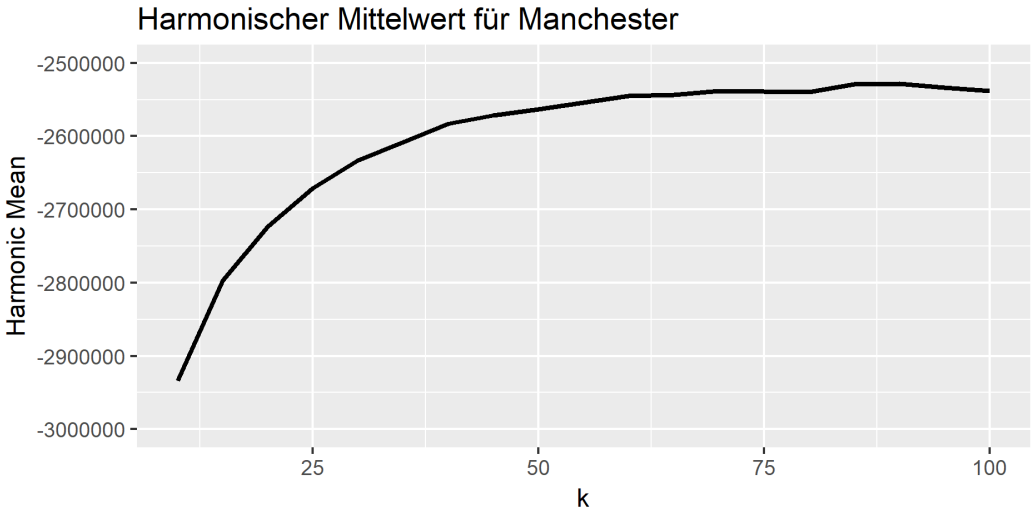


Abbildung 7: Ergebnis der harmonic mean Methode für unterschiedliche Werte von *k* in der LDA für Manchester

Eine nähere Betrachtung der einzelnen Themen sowohl im Birmingham- als auch im Manchester-Datensatz zeigt zudem, dass bei mehr als 30 Themen eine Aufteilung von vorher gut interpretierbaren Themen in teilweise nicht mehr interpretierbare Themen erfolgt. Bei weniger als 20 Themen werden dagegen Tweets unter ein Thema kategorisiert, die inhaltlich auf mindestens zwei Themen aufzuteilen wären.

Chang and Blei (2009) plädieren dafür, keine übermäßige Optimierung der likelihood-basierten Werte durchzuführen, sondern sich mehr auf die realweltlichen Aufgaben zu fokussieren. Die jeweiligen latenten Strukturen der Wörter und Themen sollten jeweils vom Forscher anhand der

Ergebnisse der Modelle geprüft und evaluiert werden. Mit einem Wert von *k = 60* konnte im Manchester-Datensatz kein Modell gefunden werden, dass plausible Themen und Kategorisierungen der Tweets ermöglicht und gleichzeitig im Rahmen der hier betrachteten Fragestellung handhabbar ist. Bei einer höheren Anzahl an Themen nahm die Zahl der Themen, die nicht mehr eindeutig identifiziert werden können, deutlich zu. Aber auch schon bei 60 Themen waren viele Themen anhand der wahrscheinlichsten Wörter nicht mehr interpretierbar. Als Alternative wurden die vorliegenden Ergebnisse zunächst anhand der „Lesbarkeit" der Themen mit den fünf bis zehn Wörtern mit der höchsten Wahrscheinlichkeit überprüft. Dies führt zu einer besseren Verwertbarkeit der Daten bei rund 30 Themen. Weitere Durchläufe mit Werten für *k* zwischen 20 und 40 lieferten bei *k = 32* ein ausgewogenes Ergebnis für Manchester. Im Birmingham-Datensatz konnte bei *k = 30* eine ausgewogene Zusammensetzung der Themen erreicht werden. Ausgewogen bedeutet, dass inhaltlich gut zu interpretierende Themen für die Fragestellung erreicht wurden, die bei höheren oder niedrigeren Werten von *k* entweder mit anderen Themen zusammengelegt oder noch weiter aufgefächert werden, ohne dass dies inhaltlich sinnvoll erscheint. Jeder Tweet im Datensatz wurde schließlich mit dem Thema mit der höchsten Wahrscheinlichkeit aus dem jeweils letzten Modelldurchlauf klassifiziert. Tabelle 1 und Tabelle 2 zeigen für alle Themen aus beiden Datensätzen die jeweils zehn Wörter mit der höchsten Wahrscheinlichkeit.

Tabelle 1: Verteilung der Wörter über Themen in Manchester (*k = 32*)

Thema	Verteilung der Wörter über Themen
1	0,1596*night + 0,1282*great + 0,0684*amaz + 0,0315*thank + 0,0235*anoth + 0,0212*day + 0,0211*absolut + 0,0198*meet + 0,0183*saturday + 0,0182*awesom
2	0,0364*food + 0,0284*lunch + 0,0244*cake + 0,0219*wetherspoon + 0,0216*breakfast + 0,0205*tea + 0,0172*chocol + 0,0156*made + 0,0142*chicken + 0,014*fresh
3	0,1029*back + 0,0669*home + 0,0354*go + 0,0317*come + 0,0278*head + 0,0241*tomorrow + 0,0204*tou + 0,0201*excit + 0,0193*servic + 0,0191*miss
4	0,1246*good + 0,0694*look + 0,069*morn + 0,0455*readi + 0,0426*get + 0,0386*feel + 0,0385*friday + 0,0382*andu + 0,0304*realli + 0,0162*better
5	0,109*love + 0,0396*wed + 0,0382*beauti + 0,0275*walk + 0,0249*dog + 0,0219*afternoon + 0,021*pic + 0,0189*lot + 0,0188*yesterday + 0,0165*hill
6	0,0539*run + 0,0515*finish + 0,0348*just + 0,0285*endomondo + 0,028*endorphin + 0,0239*walk + 0,0225*mile + 0,0203*stop + 0,0197*around + 0,0188*end
7	0,0408*art + 0,0242*design + 0,023*pictur + 0,0213*green + 0,0164*photographi + 0,0163*hand + 0,0154*made + 0,0144*inspir + 0,0141*visit + 0,0131*galleri
8	0,0717*best + 0,0469*hotel + 0,0287*red + 0,0265*black + 0,0249*world + 0,0243*white + 0,0225*room + 0,0181*lancashir + 0,0165*arm + 0,0154*pub
9	0,1444*time + 0,0801*first + 0,0702*year + 0,05*stockport + 0,0338*hall + 0,018*town + 0,0155*albert + 0,0149*ago + 0,0142*cheadl + 0,0135*trip
10	0,0742*weekend + 0,0559*sunday + 0,0442*wait + 0,0438*cant + 0,0379*cheshir + 0,0349*holiday + 0,0276*wilmslow + 0,0263*saturday + 0,0247*bank + 0,0216*northern
11	0,1867*new + 0,0388*check + 0,0208*break + 0,019*car + 0,0169*arriv + 0,0162*fabul + 0,0153*castlefield + 0,0153*stock + 0,0142*latest + 0,0134*vehicl

Thema	Verteilung der Wörter über Themen
12	0,1481*love + 0,0643*club + 0,0469*littl + 0,0441*friend + 0,0375*famili + 0,0325*fun + 0,0276*golf + 0,0231*babi + 0,0216*ladi + 0,021*selfi
13	0,0551*thank + 0,0504*team + 0,0457*well + 0,0382*done + 0,0359*big + 0,0249*win + 0,0237*everyon + 0,0221*pleas + 0,0203*proud + 0,0186*help
14	0,0592*beauti + 0,0457*hair + 0,0368*book + 0,031*colour + 0,0286*use + 0,0283*gorgeous + 0,0213*avail + 0,0209*full + 0,0194*makeup + 0,0155*offer
15	0,4799*manchest + 0,1429*greater + 0,0582*airport + 0,043*railway + 0,0423*station + 0,0361*man + 0,0278*intern + 0,0261*piccadilli + 0,0082*arndal + 0,0061*grill
16	0,2369*just + 0,1931*photo + 0,1163*post + 0,0241*floor + 0,0196*carpet + 0,0185*planet + 0,0184*video + 0,0095*squar + 0,0079*clean + 0,0069*huddersfield
17	0,0801*tonight + 0,052*stadium + 0,0442*play + 0,0415*live + 0,0368*etihad + 0,0336*music + 0,0322*show + 0,0189*band + 0,0165*watch + 0,0151*perform
18	0,0915*happi + 0,0652*birthday + 0,0549*bolton + 0,04*summer + 0,0292*boy + 0,0284*celebr + 0,0272*sun + 0,0261*parti + 0,0256*girl + 0,0244*blue
19	0,1283*old + 0,12*trafford + 0,0665*centr + 0,0304*footbal + 0,0294*game + 0,0263*cricket + 0,0233*final + 0,019*dream + 0,018*season + 0,0177*ground
20	0,4641*manchest + 0,1178*unit + 0,0804*kingdom + 0,0222*oldham + 0,0121*gtr + 0,0084*wythenshaw + 0,0063*unitedu + 0,0057*mot + 0,0055*termin + 0,0049*cathedr
21	0,0617*job + 0,0591*england + 0,0451*work + 0,0438*open + 0,0357*hire + 0,0322*free + 0,0292*month + 0,0233*join + 0,0211*sale + 0,0202*latest
22	0,2312*drink + 0,0649*nice + 0,0546*beer + 0,0282*photo + 0,027*ale + 0,0228*ipa + 0,0218*pale + 0,0163*bitter + 0,0158*flavour + 0,0151*tast
23	0,2065*day + 0,0924*week + 0,0569*start + 0,0425*next + 0,0356*anoth + 0,0302*busi + 0,0214*everi + 0,02*enjoy + 0,0157*cours + 0,0135*coupl
24	0,0422*coffe + 0,033*place + 0,0323*altrincham + 0,0297*favourit + 0,0295*light + 0,0295*street + 0,0253*shop + 0,0226*market + 0,0213*project + 0,0189*victoria
25	0,0409*know + 0,0361*think + 0,0354*tri + 0,032*life + 0,0315*alway + 0,0315*peopl + 0,0177*find + 0,0166*someth + 0,0149*thing + 0,0145*mani
26	0,0641*work + 0,0519*train + 0,0422*session + 0,0309*fit + 0,0294*gym + 0,0227*morn + 0,0206*class + 0,02*hard + 0,02*box + 0,0147*workout
27	0,1189*park + 0,0388*make + 0,0297*garden + 0,0293*heaton + 0,0212*chorlton + 0,019*call + 0,0166*church + 0,0148*urban + 0,0144*water + 0,0128*tree
28	0,2749*manchest + 0,0506*festiv + 0,0355*parklif + 0,0298*school + 0,0269*arena + 0,0263*museum + 0,0237*academi + 0,02*apollo + 0,0149*even + 0,0137*north
29	0,0594*today + 0,0372*top + 0,0213*avail + 0,0191*left + 0,0171*super + 0,0171*onlin + 0,0152*dress + 0,0143*ticket + 0,0137*inu + 0,0133*boutiqu
30	0,1791*today + 0,0397*studio + 0,0332*littl + 0,0261*rain + 0,0232*tattoo + 0,0207*nail + 0,017*client + 0,0164*shot + 0,0163*went + 0,0151*pretti
31	0,1184*salford + 0,0922*citi + 0,0272*univers + 0,0248*media + 0,0245*quay + 0,0223*ofu + 0,0214*build + 0,0203*sign + 0,0146*lad + 0,014*mediacityuk
32	0,0581*bar + 0,0576*hous + 0,0469*let + 0,0398*didsburi + 0,0311*restaur + 0,029*buri + 0,0227*cocktail + 0,0214*apart + 0,0208*road + 0,02*cafe

Tabelle 2: Verteilung der Wörter über Themen in Birmingham (*k = 30*)

Thema	Verteilung der Wörter über Themen
1	0,1308*back + 0,0725*home + 0,048*train + 0,0333*come + 0,0273*london + 0,0238*bring + 0,0193*drive + 0,019*bus + 0,0171*road + 0,0164*brum
2	0,1106*watch + 0,0595*show + 0,0515*live + 0,0488*tonight + 0,0287*walsal + 0,0273*music + 0,0263*christma + 0,0234*film + 0,02*amaz + 0,0173*arena
3	0,1218*happi + 0,0994*drink + 0,0841*birthday + 0,0735*photo + 0,0242*black + 0,0221*post + 0,0203*blue + 0,0191*beer + 0,0164*celebr + 0,0152*bro
4	0,1422*new + 0,0307*phone + 0,0302*tweet + 0,0278*twitter + 0,0262*video + 0,0228*check + 0,0175*world + 0,015*pictur + 0,0138*number + 0,0133*snapchat
5	0,119*love + 0,0682*girl + 0,0505*littl + 0,038*beauti + 0,036*boy + 0,0249*listen + 0,0245*babi + 0,0242*song + 0,0235*famili + 0,0176*perfect
6	0,0999*week + 0,0758*next + 0,0693*wait + 0,0515*start + 0,0434*excit + 0,0376*weekend + 0,0321*book + 0,0305*month + 0,0285*tomorrow + 0,0243*alreadi
7	0,1081*feel + 0,0825*life + 0,0546*make + 0,0421*made + 0,0307*better + 0,0275*real + 0,0219*sick + 0,021*without + 0,0168*sad + 0,0167*whole
8	0,064*tri + 0,0516*just + 0,0436*keep + 0,0405*stop + 0,0369*find + 0,0315*face + 0,0214*eye + 0,0213*someon + 0,021*head + 0,02*hard
9	0,0954*thank + 0,0667*your + 0,0666*miss + 0,0549*ill + 0,0517*friend + 0,0514*best + 0,0347*amaz + 0,0319*everyon + 0,0305*soon + 0,0193*cheer
10	0,0338*hous + 0,0293*car + 0,0271*hair + 0,0259*top + 0,0212*room + 0,0211*wear + 0,0203*full + 0,0165*cold + 0,0141*high + 0,013*sit
11	0,1739*love + 0,1256*pleas + 0,0777*follow + 0,0672*day + 0,063*thank + 0,0388*omg + 0,0335*best + 0,0301*support + 0,0276*mean + 0,0231*hand
12	0,2044*time + 0,1264*night + 0,0734*first + 0,0601*guy + 0,0491*long + 0,0423*everi + 0,0237*meet + 0,0201*second + 0,0197*wait + 0,0184*saturday
13	0,1142*year + 0,0485*old + 0,0432*call + 0,0295*school + 0,0284*kill + 0,0271*believ + 0,0244*kid + 0,0222*mom + 0,0219*name + 0,0218*dad
14	0,0314*food + 0,0278*bar + 0,0259*eat + 0,024*solihul + 0,0184*coffe + 0,0173*hot + 0,0154*lunch + 0,0148*tea + 0,0141*dinner + 0,014*breakfast
15	0,0351*club + 0,035*shop + 0,0332*ticket + 0,0258*buy + 0,0216*free + 0,0159*hotel + 0,0153*snow + 0,0146*park + 0,0133*order + 0,0132*fashion
16	0,3196*just + 0,0396*finish + 0,0392*walk + 0,0376*run + 0,0339*wanna + 0,028*turn + 0,0267*around + 0,0236*use + 0,0186*round + 0,0184*saw
17	0,1481*work + 0,0545*hour + 0,0502*sleep + 0,0428*bed + 0,0367*tomorrow + 0,0341*get + 0,0314*readi + 0,0313*go + 0,0289*leav + 0,028*half
18	0,3739*birmingham + 0,1275*west + 0,1035*midland + 0,0309*street + 0,0295*citi + 0,0257*centr + 0,0252*station + 0,0237*unit + 0,0177*airport + 0,0176*univers
19	0,0552*big + 0,0541*play + 0,0283*man + 0,0164*fair + 0,0156*ball + 0,0145*kick + 0,0144*fight + 0,0128*super + 0,0124*whos + 0,0109*rip
20	0,0671*game + 0,0545*villa + 0,0346*avfc + 0,0344*win + 0,0288*season + 0,028*goal + 0,0279*fan + 0,0255*park + 0,0248*team + 0,024*player
21	0,1096*look + 0,1043*realli + 0,0603*actual + 0,0429*nice + 0,0421*tonight + 0,0334*go + 0,0322*forward + 0,0266*good + 0,0242*sound + 0,0238*see
22	0,0242*word + 0,024*yet + 0,023*need + 0,0209*pay + 0,0207*sign + 0,0183*lost + 0,0182*money + 0,018*bet + 0,0173*read + 0,0168*definit
23	0,1206*well + 0,086*today + 0,0687*done + 0,0681*great + 0,0251*gym + 0,0197*lad + 0,0194*proud + 0,0181*fit + 0,0178*class + 0,0149*win
24	0,2012*day + 0,1749*good + 0,0869*today + 0,071*hope + 0,0524*morn + 0,0516*great + 0,0305*enjoy + 0,0195*even + 0,0188*fun + 0,0187*luck

Thema	Verteilung der Wörter über Themen
25	0,1257*peopl + 0,041*mani + 0,0363*talk + 0,0352*hate + 0,0339*think + 0,0295*someon + 0,0284*person + 0,0267*doesnt + 0,024*care + 0,0197*theyr
26	0,0203*parti + 0,0184*news + 0,0169*vote + 0,0121*women + 0,0102*must + 0,01*men + 0,0099*social + 0,0095*labour + 0,0091*issu + 0,0089*media
27	0,1365*fuck + 0,0691*shit + 0,0564*man + 0,041*give + 0,0371*god + 0,0209*hell + 0,0209*piss + 0,0195*absolut + 0,0181*bitch + 0,0175*joke
28	0,1535*know + 0,0975*lol + 0,0721*haha + 0,0635*let + 0,0516*that + 0,0447*yeah + 0,0435*mate + 0,0313*think + 0,0209*true + 0,0202*gonna
29	0,0651*alway + 0,0583*think + 0,0491*bad + 0,0384*get + 0,0365*happen + 0,0349*thought + 0,0312*noth + 0,0304*there + 0,0298*what + 0,0296*place
30	0,0386*job + 0,0259*busi + 0,0245*help + 0,0244*open + 0,0216*manag + 0,0187*servic + 0,0152*cours + 0,015*event + 0,0149*interest + 0,0135*join

2.5 Methoden der räumlichen Analyse

Die räumliche Analyse erfolgt zunächst jeweils für die gesamte Region und auf Basis eines 250 m Hexagon-Rasters. Dabei bleibt die Zuweisung der Themen zu den Tweets stets gleich auf Basis des im vorherigen Kapitel dargestellten Modells. Die Tweets werden zunächst mittels eines Gitternetzes aggregiert (Kapitel 2.5.1) und dann ausgewählte Themen hinsichtlich räumlicher Autokorrelation überprüft (Kapitel 2.5.2). Für diese Themen erfolgt schließlich die Identifikation von Hot Spots in der Region (Kapitel 2.5.3).

2.5.1 Aggregation in ein Hexagon-Gitter

Die Punktdaten werden für die räumliche Analyse in einem Gitter aggregiert. Gemessen wird jeweils die Zahl der Ereignisse (Punkte) je Gitterzelle und das differenziert nach den Themen. Somit können die Unterschiede zwischen den beobachteten Ereignissen und den erwarteten Ereignissen hinsichtlich der räumlichen Zusammenhänge untersucht werden. Ein Vorteil dieser Aggregation ist eine Minimierung der notwendigen Rechenleistung, da mehrere Ereignisse (Punkte) zu einer Zelle (Polygon) zusammengefasst werden. Die Analyse von räumlichen Clustern erfolgt somit über eine begrenzte Anzahl von Polygonen anstelle einer teilweise erheblichen Anzahl von Einzelereignissen. Rasterzellen sind zudem vorteilhaft gegenüber administrativen Einheiten, wie Stadtteilen, weil sie einheitliche Größen haben und ein regelmäßig angeordnetes Muster aufweisen.

Die Zellen sollen für diese Art der Analyse gleichmäßig geformte Polygone sein. Dafür werden Sechsecke mit gleicher Kantenlänge verwendet. Jedes Hexagon hat einen Durchmesser gemessen an zwei gegenüberliegenden Kanten von 250 m, einen Umfang von 1.000 m und eine Fläche von rund 72.168 m². Der Vorteil der Sechsecke gegenüber den oftmals verwendeten Quadraten ist, dass diese in ihrer Form näher an einem Kreis sind und damit Artefakte in den Ecken minimieren. Zugleich hat ein Sechseck gemessen an den Kanten zwei Nachbarn mehr als ein Viereck, insofern Nachbarschaften

beim Viereck nur über gemeinsame Kanten abgeleitet werden. Rasterzellenstrukturen werden auch in vergleichbaren Studien mit georeferenzierten Daten aus sozialen Medien verwendet (Zhou and Zhang, 2016, Garcia-Palomares et al., 2015).

Abbildung 8 zeigt die Zahl der Tweets je Rasterzelle für die Region Greater Manchester. Wie schon in Abbildung 2 werden die räumlichen Schwerpunkte mit dem Zentrum der Stadt Manchester sowie den Städten Salford, Trafford, Stockport und Bolton deutlich sichtbar. Die Darstellung von Birmingham zeigt wiederum in fast allen Zellen Tweets, so dass nur sehr wenige Lücken um die Stadt herum vorhanden sind. Hierbei kommt die höhere Zahl der erhobenen Tweets pro Tag genauso zu tragen wie der längere Zeitraum der Erhebung und der Fokus auf die Kernstadt. Beide Karten sind, auf Grund der unterschiedlich großen Stichproben und verschiedenen Größen der Gebiete, nicht direkt miteinander zu vergleichen.

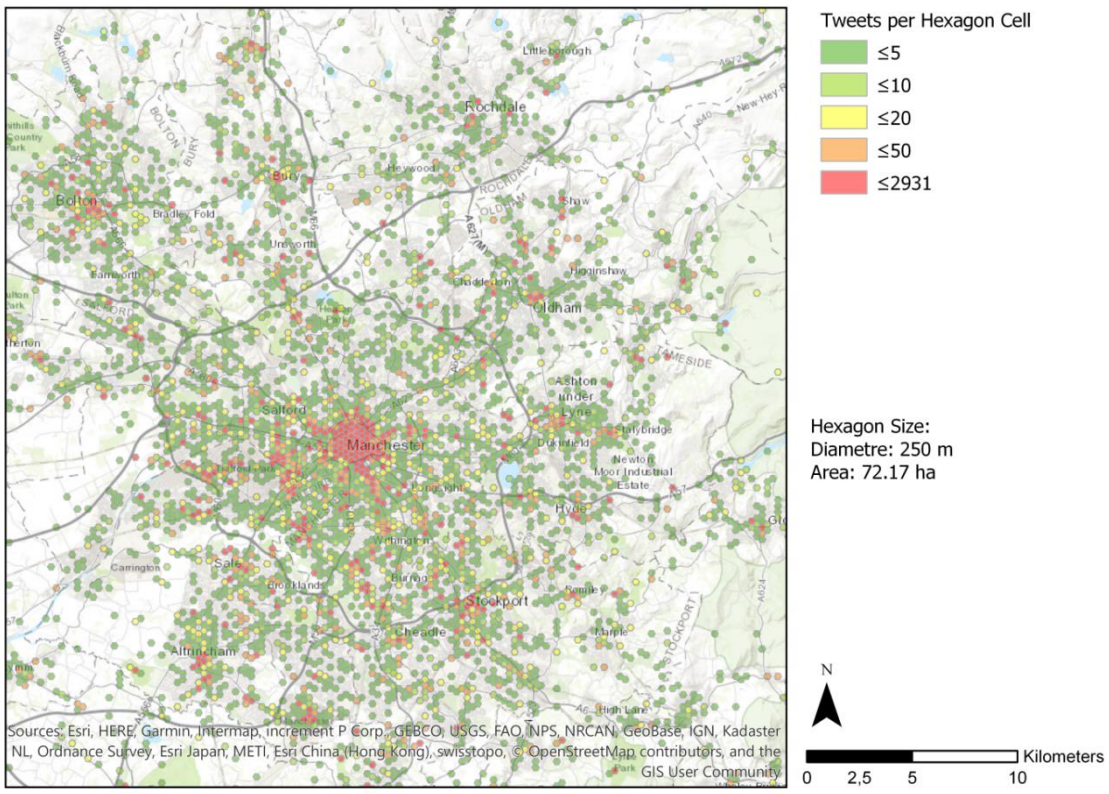


Abbildung 8: Tweets je Zelle in der Region Greater Manchester (bereinigter Datensatz)

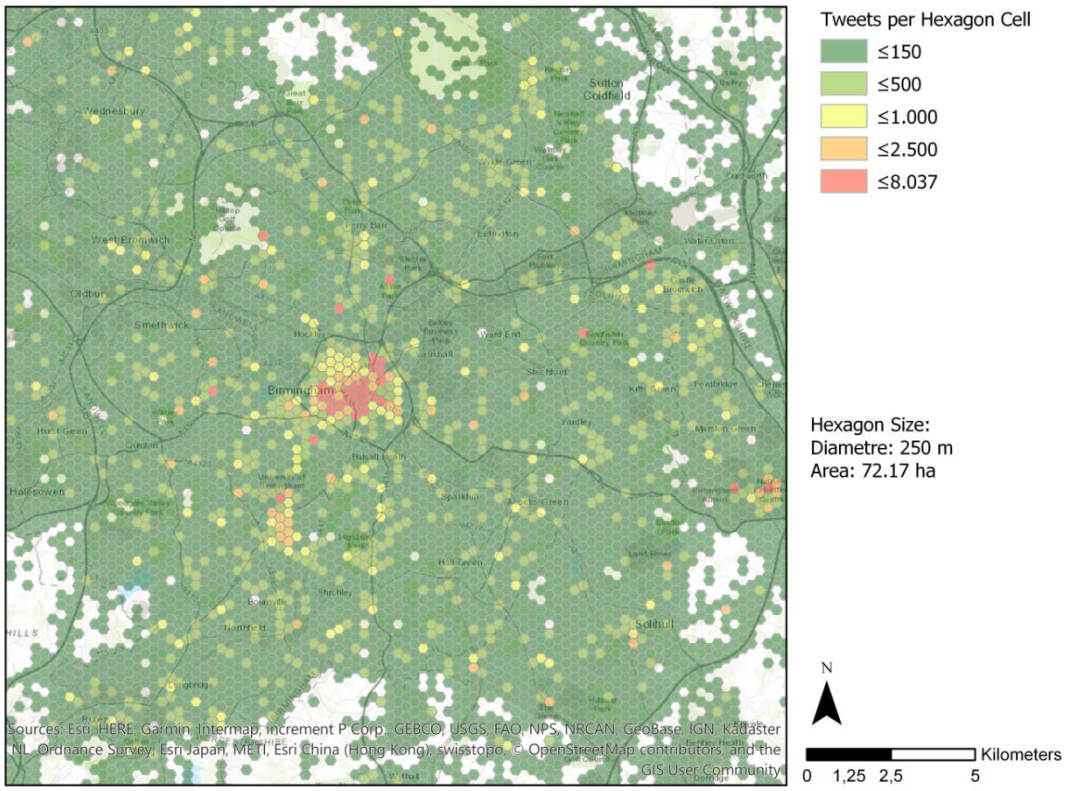


Abbildung 9: Tweets je Zelle in der Region Birmingham (bereinigter Datensatz)

2.5.2 Global Moran's I

Jeder Punkt in einem räumlichen Datensatz stellt ein einzelnes Ereignis mit einer x,y-Koordinate dar. Die Punkte können zufällig, gebündelt oder gleichmäßig verteilt sein:

- Zufällige Verteilung: Jeder Punkt existiert völlig unabhängig von anderen Punkten in einem Untersuchungsgebiet.
- Gebündelte Verteilung: Zwischen Punkten derselben Gruppe bestehen kurze Distanzen und zwischen Punkten unterschiedlicher Gruppen große Distanzen.
- Gleichmäßige Verteilung: Die Abstände zwischen den Punkten sind sehr ähnlich oder im Extremfall gleich.

Punktmuster (point patterns) resultieren aus zugrundeliegenden räumlichen Prozessen und damit aus einem stochastischen Modell. Das geht z.B. aus Toblers erstem Gesetz der Geographie hervor (1970), wonach benachbarte Objekte einander ähnlicher sind als weit entfernte Objekte. Zusammengefasst wird dies unter dem Begriff der räumlichen Autokorrelation. Diese gilt es zunächst für ausgewählte Themen nachzuweisen.

Moran (1950) entwickelte ein globales Maß zum Nachweis räumlicher Autokorrelation. Diese Methode wird heute noch häufig angewendet und ist beispielsweise in ArcGIS Pro implementiert. Im Ergebnis können Aussagen darüber getroffen werden, ob die räumliche Verteilung von Ereignissen

gruppiert (gebündelt), gleichmäßig oder zufällig ist. Neben dem Moran's I Indexwert, der Werte zwischen -1 und +1 annehmen kann und bei dem negative Werte eine negative Autokorrelation und positive Werte eine positive Autokorrelation indizieren, wird ein p- sowie ein z-Wert berechnet. Geht man von der Annahme einer zufälligen räumlichen Verteilung als Nullhypothese aus, dann ist diese Hypothese für eine gegebene räumliche Verteilung abzulehnen, wenn sehr hohe oder sehr niedrige (negative) z-Werte mit einem kleinen p-Wert vorliegen. Sehr hohe z-Werte weisen dann auf statistisch signifikante Hot-Spots oder Cold-Spots hin.

Zunächst wird für die Metropolregion Manchester sowie für das Zentrum die Distanz bestimmt, innerhalb der die gemessenen Ereignisse (Tweets) Autokorrelation aufweisen. Dafür wird Global Moran's I als Maß verwendet. Moran's Index weist auf eine starke Autokorrelation mit räumlichen Clustern hin (p-Wert bei fast allen Analysen bei etwa 0,00000 oder knapp darüber). Die Autokorrelation wird mit unterschiedlichen Distanzen, angefangen bei 300 m und in 100 m Schritten bis 3.000 m, gemessen. Die maximalen z-Werte werden für die einzelnen Themen bei unterschiedlichen Distanzen erreicht. Diese reichen von 800 m bis 2.100 m in Manchester und von 800 m bis 2.000 m in Birmingham. Die maximalen z-Werte mit Angabe der Distanz für alle Themen sind im Anhang in Tabelle 10 und Tabelle 11 aufgelistet. Ein vergleichbares Vorgehen wurde auch von Jenkins et al. (2016) verwendet. Sie haben für die Städte Los Angeles, New York und London Distanzen zwischen rund 620 und 730 m berechnet. Für Singapur kommen sie auf eine Distanz von etwa 1.335 m.

2.5.3 Local Getis-Ord Gi*

Getis and Ord (1992) haben Methoden eingeführt, um räumliche Korrelation zwischen Daten bzw. Ereignisse zu messen. G-Statistiken machen es möglich, Verbindungen in einer räumlich verteilten Variable messbar zu machen. Global Moran's I sagt nur aus, ob Autokorrelation besteht, aber zeigt nicht, wo diese im Untersuchungsgebiet signifikant ist und kann somit keine lokalen Cluster identifizieren. Gi* Statistiken können dagegen lokale räumliche Muster sichtbar machen, die durch die Anwendung globaler Statistiken nicht sichtbar sind. Local Getis-Ord Gi* ermöglicht die Identifizierung und Visualisierung solcher lokaler Cluster (Ord and Getis, 1995). Der Test prüft, ob die lokale Summe von Werten in einer Nachbarschaft eines gegebenen Features signifikant von der globalen Summe im Untersuchungsgebiet abweicht. Eine Region mit signifikant hohen Werten wird als Hot Spot und eine Region mit signifikant niedrigen Werten als Cold Spot bezeichnet.

Eine hohe Bedeutung bei der Anwendung von Gi* hat die Definition der Nachbarschaft. Die Distanz, innerhalb der alle Features um ein untersuchtes Feature als Nachbarschaft definiert werden, wird aus den durchgeführten Moran's I Tests bestimmt. Die Distanz mit dem Maximum des z-Wertes wird als Distanz für Gi* vorgegeben. Die p-Werte für die gewählten Distanzen zeigen in allen Fällen eine

hohe statistische Signifikanz (p-Werte von 0,00000 oder knapp darüber). Für die Analyse der auf Ebene der Hexagon-Zellen aggregierten Tweets wird des Weiteren unterstellt, dass der Einfluss untereinander umso höher ist, je näher sich die Punkte sind. Deshalb erfolgt die Modellierung der räumlichen Beziehung für die Analyse nach Getis und Ord mittels inverser Entfernung.

Die Methoden zur Berechnung der lokalen Cluster können zu Problemen bei Mehrfachtests und Abhängigkeiten führen. Ein Konfidenzniveau von 95 % entspricht einem möglichen Irrtum in 5 von 100 Fällen, die tatsächlich nicht geclustert, sondern räumlich zufällig verteilt sind. Das führt bei Datensets mit mehreren hundert oder tausend Features zu einer relativ hohen Anzahl an falsch erwarteten Ereignissen. Zudem werden Features immer im Kontext ihrer Nachbarfeatures untersucht und diese weisen wiederum viele gemeinsame Nachbarn auf, wodurch die räumlichen Abhängigkeiten womöglich überschätzt werden. In ArcGIS Pro ist für diesen Zweck die FDR-Korrektur (False Discovery Rate) implementiert. Dabei wird für ein gegebenes Konfidenzniveau die Anzahl der falschen positiven Ereignisse geschätzt. Eine Sortierung der statistisch signifikanten p-Werte und der darauf basierende Ausschluss der p-Werte mit den größten (schwächsten) Werten ermöglicht bessere Ergebnisse (Caldas de Castro and Singer, 2006). Diese Methode wird in vergleichbaren Studien ebenfalls eingesetzt (Steiger et al., 2015b).

3 Ergebnisse

Die Ergebnisse der empirischen Analysen werden in den folgenden Kapiteln beschrieben. Kapitel 3.1 beschreibt die räumlichen Muster und Cluster ausgewählter Themen in der Region Manchester und in Kapitel 3.2 erfolgt dies für die Stadt Birmingham. Für die Untersuchung werden diejenigen Themen herausgegriffen, für die inhaltliche Zusammenhänge im Hinblick auf städtische Räume angenommen werden können. Die Themen sind im folgenden Text mit eckigen Klammern nummeriert (z.B. [1]). Die assoziierten Wörter werden *kursiv* dargestellt. Es handelt sich jeweils um die Wortstämme, so dass einzelne Begriffe nicht vollständig sind (z.B. *morn* statt *morning*, *amaz* statt *amazing* oder *amazed*). Zur besseren Lesbarkeit wurden die Wortstämme zum Teil ergänzt und diese Ergänzungen kenntlich gemacht (z.B. *wed*[ding]). Nach dem Stemming ist jedoch nicht eindeutig die ursprüngliche, konjungierte Form des Wortes im jeweiligen Kontext bekannt. Dieser Vorgang wurde im Kapitel 2.3 näher beschrieben.

Die räumlichen Cluster werden im Hinblick auf bekannte Orte (Strukturen) sowie zu erwartende oder bekannte Mobilitätsmuster und Events (Prozesse) in den Regionen analysiert. Kapitel 3.3 beschreibt den quantitativen Abgleich der Cluster mit Punktdaten aus OpenStreetMap und zeigt räumliche sowie semantische Zusammenhänge zwischen den unterschiedlichen Datensätzen auf. In Kapitel 3.4 erfolgt schließlich eine Analyse von raum-zeitlichen Mustern in Birmingham. Dabei geht es zum einen um den Tagesverlauf von Tweethäufigkeiten im Umfeld von Bahnhaltestellen und zum anderen um den Abgleich von Fußballspielen des Aston Villa Football Club mit Tweethäufigkeiten im Stadionumfeld.

3.1 Identifizierte Themen und räumliche Muster in Manchester

Wie bereits in Kapitel 2.4 beschrieben wurden für Manchester 32 Themen als eine plausible, semantische Klassifikation bestimmt. Die 32 Themen unterscheiden sich jedoch in dem Grad, in dem sie interpretierbar sind. Die überwiegende Anzahl der identifizierten Themen spiegelt Freizeitaktivitäten wieder. Herausgegriffen werden Themen zum Nachtleben [1, 10, 22, 32], zu Essen [2, 24], Einkaufen und Lifestyle [14, 29, 30], zu touristischen Anlaufpunkten [7, 8], (Fußball)Stadien [13, 17, 19] sowie zu Aktivitäten im Freien, Sport und Training [6, 26, 27, 28]. Nicht alle identifizierten Themen lassen sich einem der genannten Themenbereiche zuordnen oder sie zeigen keine interpretierbaren räumlichen Muster. Bei einzelnen Themen ergeben sich zudem Überschneidungen, die in den folgenden Kapiteln auch beschrieben werden. Die im Folgenden dargestellten Karten zeigen die identifizierten Hot Spots für ein Gebiet im Umkreis von 10 km um das Zentrum von Manchester auf Basis des beschriebenen 250 m Hexagonrasters.

Hinweise zu den in der Analyse nicht weiter behandelten Themen

Einige Themen sind für die vorliegende Analyse nicht weiter relevant. Insofern sie inhaltlich interpretiert werden können, werden sie hier kurz beschrieben. Weiterhin liefert die Auswertung Themen, die gar nicht interpretiert werden können. Auf die Themen [3, 4, 9, 11, 12, 18, 20, 21, 25] wird darum nicht weiter eingegangen.

Thema [15] verweist mit den Wörtern *airport*, *railway*, *station* und *piccadilli* auf den Flughafen und die Bahnhöfe in Manchester. Die Wörter *manchest*[er] und *greater* haben in diesem Thema jedoch mit Abstand die größte Bedeutung und somit werden auch viele Tweets diesem Thema zugeordnet, die keinen Bezug zu Bahnhöfen oder zum Flughafen haben. Das Thema eignet sich daher nicht zur weiteren Analyse. Thema [23] umfasst verschiedene Zeitangaben, wie z.B. *day*, *week*, *start*, *next*, *anoth*[er] und *everi*, sowie mit sehr geringen Wahrscheinlichkeiten alle Werktage bis auf Freitag. Unter dieses Thema fallen sehr unterschiedliche Tweets, die nicht einem spezifischen Thema über den Zeitbezug hinaus zuzuordnen sind. Vielmehr fallen hier viele nicht hinreichend differenzierte Themen zusammen. Thema [31] ist ein Indikator für Universitätsstandorte. Das Stichwort *univers[ity]* ist dafür ein guter Hinweis. Die Tweets konzentrieren sich aber vor allem im Zentrum von Salford und an der MediaCity UK.

3.1.1 Nachtleben und Freizeit

Anhand der identifizierten Themen können viele Tweets dem Nachtleben in einer Stadt zugeordnet werden. Zu den häufigsten Themen im Datensatz gehören Pub- und Barbesuche [22, 32] und abendliche Touren am Wochenende [1, 10]. Thema [1] konzentriert sich insbesondere im Zentrum von Manchester. Weitere Hot Spots befinden sich am Etihad Stadion von Manchester City, am Old Trafford Stadion von Manchester United, an und in der Nähe der University of Manchester, in der MediaCity UK, am Apollo Business Park und am intu Trafford Centre. Die MediaCity ist ein Gewerbe-, Unterhaltungs- und Wohnkomplex an den Salford Quays in Salford. Am Apollo Business Park steht das O2 Apollo Manchester. Dabei handelt es sich um eine Konzerthalle. Das intu Trafford Centre ist eine große Shopping Mall. Weitere Hot Spots sind nicht eindeutig einzelnen Einrichtungen oder Ortsteilzentren zuzuordnen. Thema [1] wird insbesondere durch die Wörter *night*, *great* und *amaz*[ing] geprägt und umfasst auch viele eher allgemeine, persönliche Statusupdates.

Thema [10] hat ebenfalls Hot Spots an den beschriebenen Punkten (Ausnahme Etihad Stadion). Das Zentrum von Manchester wird jedoch nicht flächendeckend als Hot Spot ausgewiesen, aber unter anderem das Northern Quarter. Das Thema verweist mit den Wörtern *weekend*, *sunday*, *holiday*, *saturday* oder *bank* [holiday] auf diverse Wochenendaktivitäten. Das Thema [22] konzentriert sich im Zentrum auf die Bereiche des Northern Quarter im Nordosten und einige Bereiche im Südwesten des Zentrums, die wiederum nördlich der Universität liegen. Ein weiterer Hot Spot liegt im Gebiet des

Piccadilly Bahnhofs sowie entlang der Gleise nach Südosten. Die südlichste dieser Zellen überschneidet sich mit einer Brauerei und erklärt hier die hohe Korrelation mit den Tweets mit Bierbezug. Die wichtigsten Wörter sind *drink*, *nice* und *beer*.

Thema [32] konzentriert sich neben dem Zentrum von Manchester an der MediaCity UK bzw. den dort gelegenen Salford Quays und dem südlich gelegenen Capital Quay. Im Süden gibt es einen Hot Spot im Stadtteil Didsburry, der zu Manchester gehört. Im Osten taucht die Stadt Stalybridge als Hot Spot auf. Die beiden letztgenannten Punkte waren in den vorherigen Themen nicht als Hot Spot gekennzeichnet. Das Thema [32] umfasst u.a. die Wörter *bar*, *restaurant* oder *cocktail*.

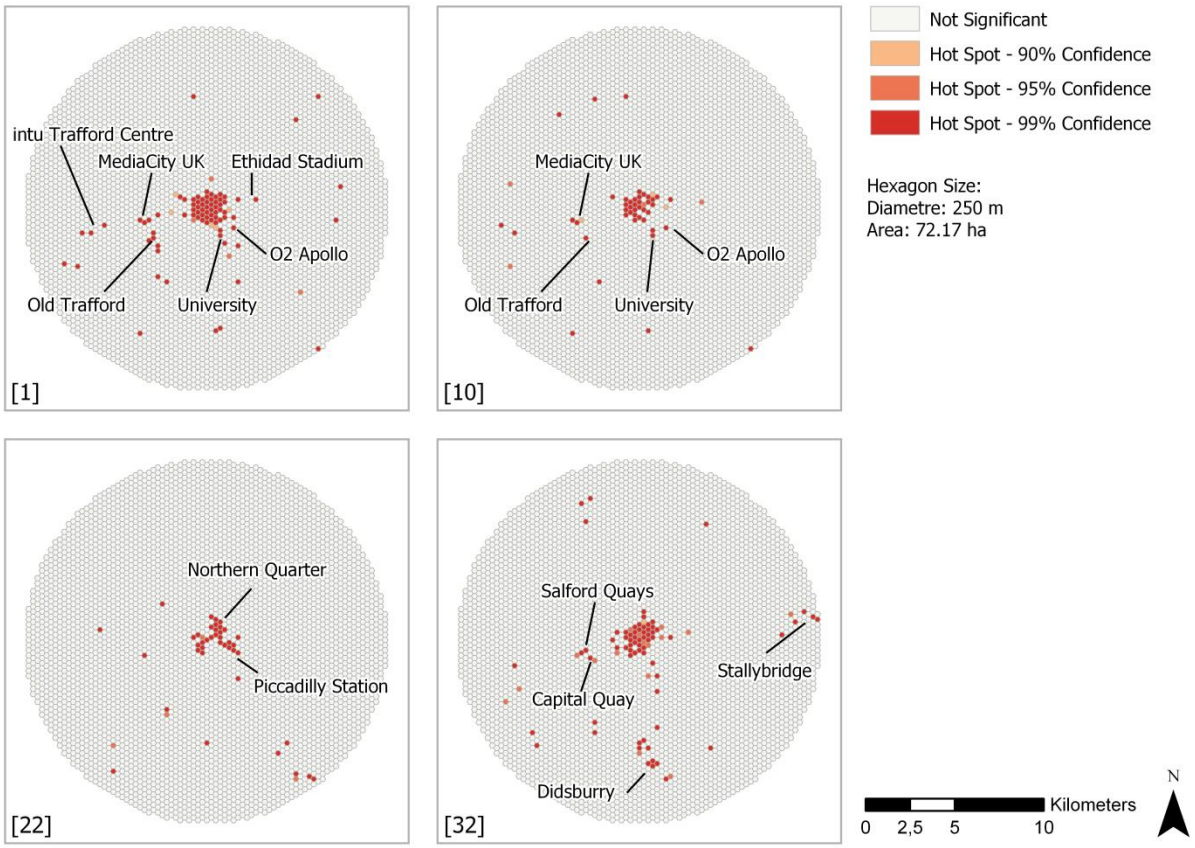


Abbildung 10: Local Getis-Ord Gi* Statistiken für Themen im Bereich Nachtleben in Manchester

3.1.2 Essen, Einkaufen und Lifestyle

Tweets mit Bezug zum Essen können vor allem mit dem Thema [2] identifiziert werden. Die Begriffe *food*, *lunch*, *cake*, *breakfast*, *tea*, *chocol*[ate] und *chicken* verweisen auf diverse Mahlzeiten und Essensangebote. Die räumlichen Muster der mit [2] klassifizierten Tweets zeigen Hotspots im Zentrum von Manchester, aber auch in den Stadtzentren von Didsburry und Chorlton-cum-Hardy. Die Shopping Mall Trafford Centre taucht ebenfalls als Hot Spot auf. Das Thema [24] enthält neben dem Wort *coffee* auch Verweise zu konkreten Orten und Einkaufsgelegenheiten (*altrincham*, *street*, *shop*, *market*, *victoria*).

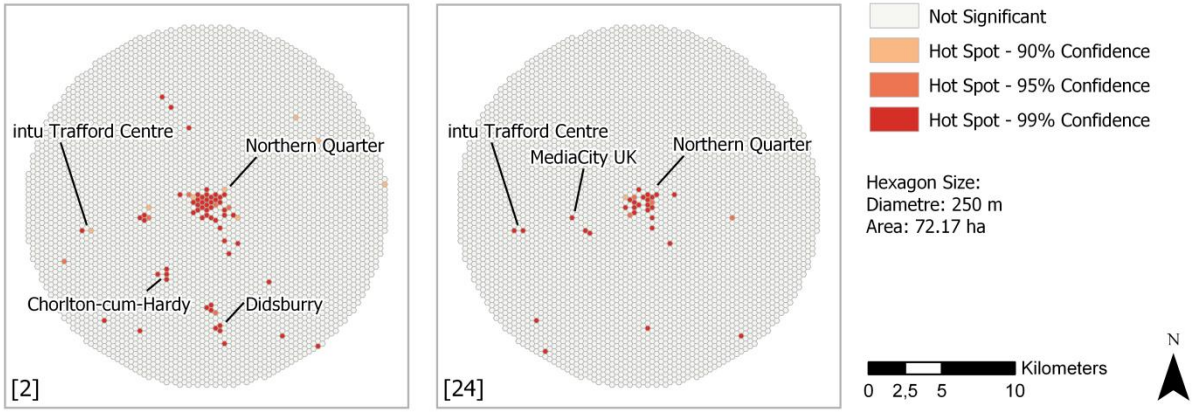


Abbildung 11: Local Getis-Ord Gi* Statistiken für Themen im Bereich Essen und Einkaufen in Manchester

Die Themen [14], [29] und [30] haben gemeinsam, dass sie im weiteren Sinne verschieden Begriffe im Bereich Mode und zur Beschreibung von Mode und Accessoires aufgreifen. Thema [14] vereint eine Vielzahl an Wörtern aus diesem Zusammenhang, wie z.B. *beauti*[ful], *hair*, *color*, *gorgeous*, *makeup*, *offer*, *salon* und u.v.a.. Jedoch haben alle diese Wörter im Thema eine relativ geringe Bedeutung. Thema [29] verweist auf Angebote in Geschäften und Onlineshops, wie einige der wichtigsten Wörter zeigen: *today*, *top*, *avail*[able], *super*, *onlin*[e], *dress*, *ticket*, *boutiqu*[e]. Aber auch dieses Thema ist durch geringe Wahrscheinlichkeiten der einzelnen Wörter geprägt. Dies gilt auch für Thema [30], dass Verweise auf Tattoo- und Nagelstudios beinhaltet: *today*, *studio*, *tattoo*, *nail*, *client*.

Das Northern Quarter taucht auch bei den Themen [29] und [30] als Hot Spot auf. Die Hot Spots von Thema [14] konzentrieren sich unter anderem an der MediaCity UK und im Northern Quarter des Stadtzentrums. Diese beiden Standorte sind insofern stimmig, als dass das Northern Quarter als „In“-Stadtteil bekannt ist und die MediaCity UK als Wohn- und Unterhaltungskomplex am Wasser ebenfalls Lifestyle orientierte Themen in sozialen Medien begünstigen kann.

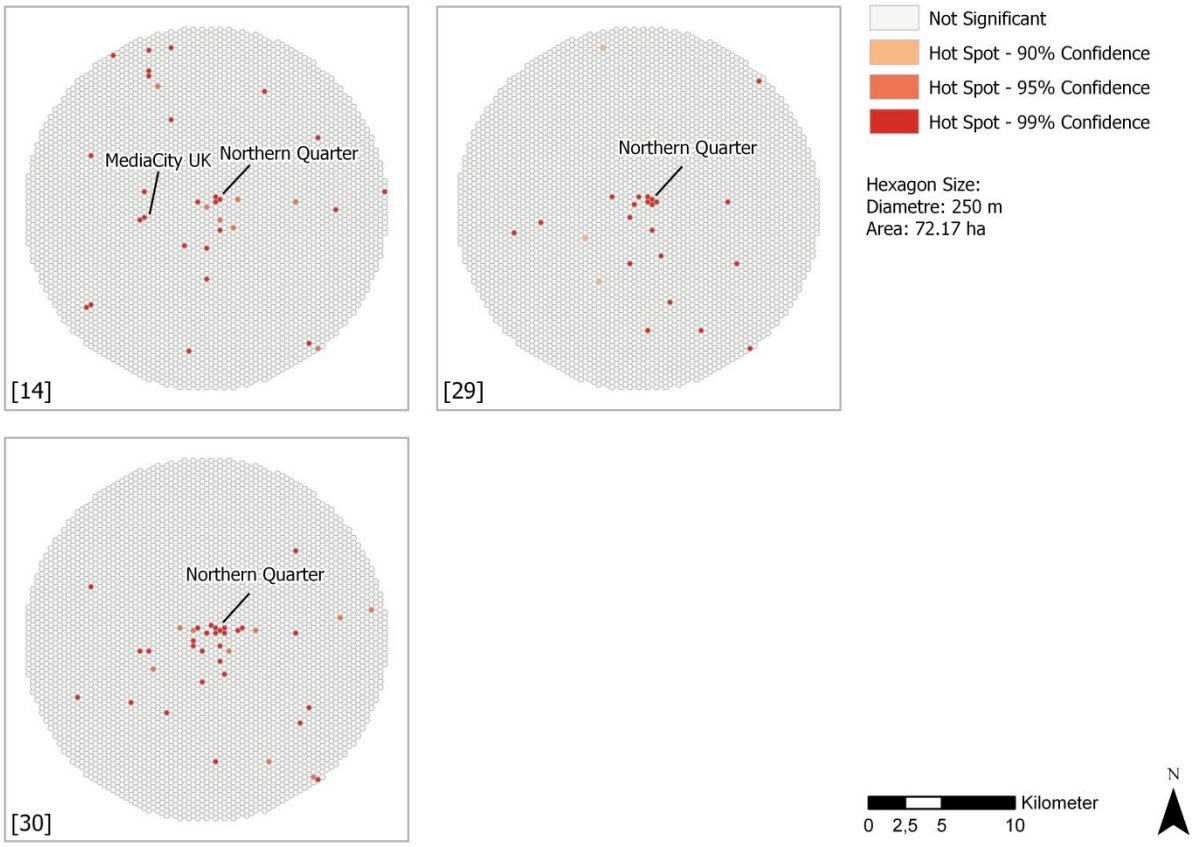


Abbildung 12: Local Getis-Ord Gi* Statistiken für Themen im Bereich Mode und Lifestyle in Manchester

3.1.3 Tourismus

Thema [7] verweist inhaltlich auf Kunst und Ausstellungen. Die Wörter *art*, *design*, *pictur*[e], *photographi*, *hand*, *made*, *inspir*[ed], *visit* und *galleri* verweisen auf Ausstellungen und Kunstwerke. Die Tweets dieses Thema konzentrieren sich insbesondere im Zentrum von Manchester, wo es viele Ausstellungen und Galerien gibt, südlich der Innenstadt beim Manchester Museum und der Whitworth Art Gallery sowie am O2 Apollo Theatre. Weitere Hot Spots sind am Salford Museum and Art Gallery, an der MediaCity UK, am Etihad Stadion und dem Old Trafford zu finden. Thema [7] kann somit als ein guter Indikator für touristische Hot Spots gesehen werden.

Thema [16] kann an drei Wörtern festgemacht werden: *just*, *photo*, *post*. Diese Wörter verweisen wiederum auf typische Beiträge in sozialen Medien, bei denen die Nutzer Fotos veröffentlichen. Dies kann über Apps parallel in unterschiedlichen Netzwerken erfolgen. Ausgewählte Hot Spots verteilen sich über die Innenstadt, die Stadien von Manchester City und Manchester United, die MediaCity UK, die Universität in Salford sowie das Salford Museum and Art Gallery, das O2 Apollo Theatre und die Universität von Manchester. Das Thema kann inhaltlich nicht weiter eingeordnet werden, zeigt aber mögliche Besuchsziele, von denen besonders häufig Fotos veröffentlicht werden. Da diese Orte auch mit möglichen Anlaufpunkten für Touristen korrespondieren, wird das Thema hier mit aufgeführt.

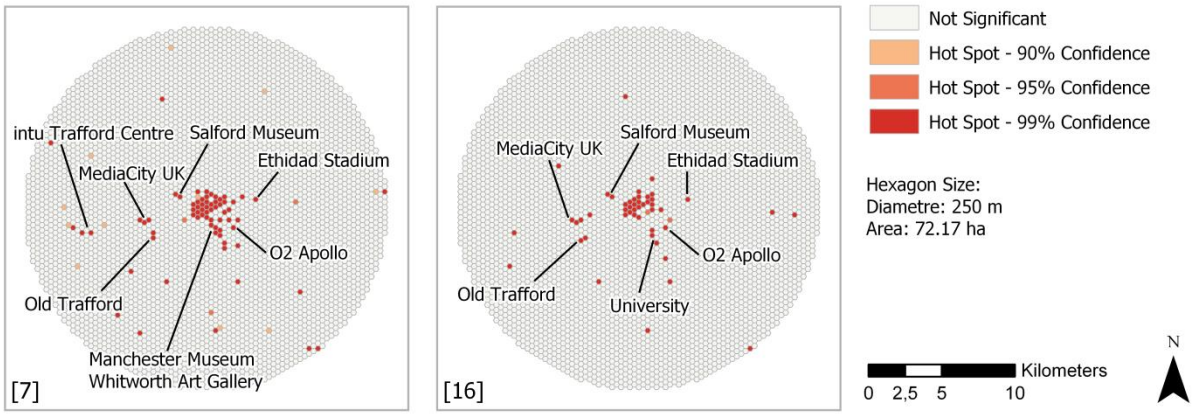


Abbildung 13: Local Getis-Ord Gi* Statistiken für Themen im Bereich Tourismus in Manchester

Thema [8] deutet mit dem Wort *hotel* zunächst auch auf ein Thema für den Bereich Tourismus hin. Das Thema ist jedoch darüber hinaus inhaltlich anhand der weiteren häufigsten Wörter nicht gut interpretierbar und zeigt auch keine eindeutigen Hot Spots außerhalb des Zentrums von Manchester.

3.1.4 Stadien

Mehrere Themen in Manchester können dem Themenfeld Fußball und konkreten damit verbundenen Orten zugeordnet werden. Eindeutig abgegrenzt werden können die beiden Fußballvereine Manchester City [17] und Manchester United [19]. Die Tweets bezogen auf Manchester City gruppieren sich vor allem Anhand des Begriffs *etihad*, was sich auf den Namen des Stadions bezieht. Ein weiterer Hot Spot des Themas [17] liegt beim Salford Stadion im Westen der Region. Es gibt aber auch einzelne Hot Spots im Zentrum von Manchester. Die wichtigsten Wörter sind neben *etihad* noch *tonight*, *stadium*, *play*, *live*, *music*, *show*, *band*, *watch* und *perform*. Die Überschneidung zwischen dem Stadion und Begriffen im Zusammenhang mit Livekonzerten erklärt sich dadurch, dass im Etihad Stadion auch große Live-Konzerte stattfinden. Viele Tweets des Themas [17] verweisen im Zusammenhang mit dem Stadion nicht auf Fußball, sondern auf diese Live-Konzerte. Dadurch ist das Thema [17] sowohl mit Etihad, Fußball und Manchester City als auch mit Etihad, Musik und Konzerten verbunden. Die Hot Spots im Zentrum sind wiederum Gebiete, in denen es ebenfalls zahlreiche Musikveranstaltungen gibt, z.B. im Northern Quarter.

Das Thema [19] bezieht sich vordergründig vor allem auf Old Trafford, dem Stadion von Manchester United. Wichtige Wörter in diesem Thema sind *old*, *trafford*, *centr*, *footbal*, *game*, *cricket*, *final*, *dream*, *season* und *ground*. Mit Thema ist unter anderem das Wort *cricket* verbunden, weil es in der Nähe des Fußballstadions auch den Old Trafford Cricket Ground gibt. Die beiden Stadien sind in verschiedenen Durchläufen der LDA Methode oftmals nicht getrennt worden und somit liegt in der Regel nur ein Thema für beide vor. Cricket ist in England sehr populär und es gibt auch sehr viele Tweets zu dieser Sportart. Erst bei einer deutlichen Erhöhung von *k* findet eine Differenzierung statt.

Weiterhin taucht in der räumlichen Analyse die Shopping Mall intu Trafford Centre auf, die auch bei anderen Themen als Hot Spot identifiziert wird. Diese drei Orte liegen alle im gleichnamigen Metropolitan Borough Trafford. Das Wort *trafford* ist entsprechend ein häufig genutzter Begriff in dieser Region des Untersuchungsgebietes und Namensgeber für viele Orte.

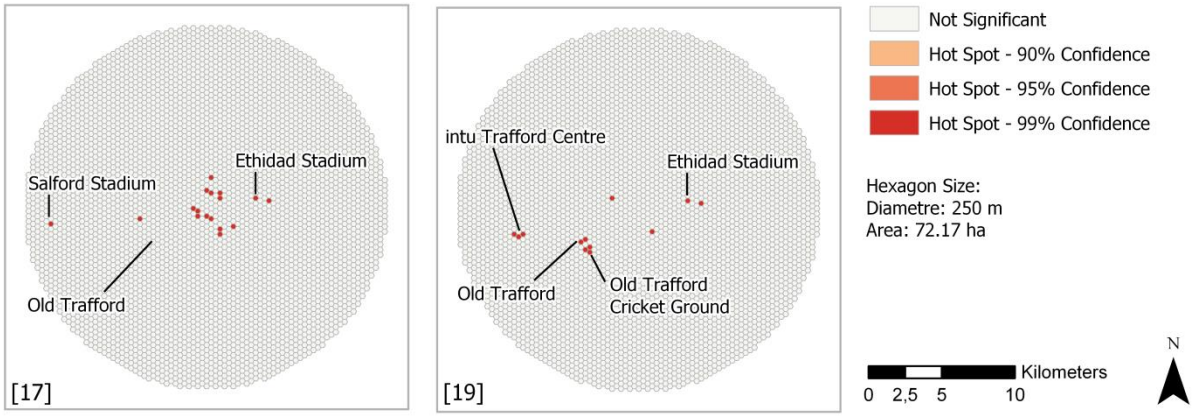


Abbildung 14: Local Getis-Ord Gi* Statistiken für Themen im Bereich Stadien in Manchester

Das Thema [13] umfasst viele Tweets, die sich ebenfalls mit Fußball und anderen Sportarten beschäftigen. Die Tweets sind jedoch allgemeiner gefasst bzw. die verwendeten Wörter beziehen sich nicht auf konkrete Orte, wie die dargestellten Stadien. Die wichtigsten Wörter umfassen unter anderem *thank*, *team, well*, *done*, *big*, *win*. Die geringen Wahrscheinlichkeiten und ein fehlendes räumliches Muster, das über ein Zentrum-Peripherie-Gefälle hinausgeht, erschweren jedoch die Interpretierbarkeit des Themas.

3.1.5 Aktivitäten im Freien und Sport

Das Thema [6] umfasst verschiedene Wörter, die insbesondere auf sportliche Aktivitäten der Nutzer verweisen. Neben den Wörtern *run*, *finish*, *just*, *endorphin*, *walk* und *mile* ist der Begriff *endomondo* ein wichtiger Indikator dafür. Dabei handelt es sich um eine App zum Aufzeichnen von Routen beim Laufen oder Fahrradfahren. Unter den Tweets befinden sich entsprechend einige Statusupdates von Nutzern, die ihre gelaufenen Routen und Distanzen veröffentlichen. Die Hot Spots sind vor allem im Zentrum von Manchester, um die Stadien von Manchester City und Manchester United sowie an der MediaCity UK zu finden. Weitere Hot Spots können keinem eindeutigen Ort zugeordnet werden, sondern verteilen sich eher zufällig auf verschiedene Wohnviertel. Hier wäre aber z.B. eine räumliche Korrelation mit Parkanlagen zu erwarten gewesen, die aber nicht vorliegt. Das Thema ist anhand der vorliegenden räumlichen Konzentration nur schwer zu interpretieren.

Ähnlich verhält es sich mit dem Thema [26]. Die Hot Spots zeigen keinen direkten Zusammenhang zu Parks oder Sportstätten, sondern finden sich in verschiedenen Wohnvierteln sowie dem Zentrum von

Manchester wieder. Die wichtigsten Begriffe in diesem Thema sind *work*, *train*, *session*, *fit*, *gym*, *morn*, *class*, *hard*, *box* und *workout*. Ein Abgleich mittels Internetsuchmaschinen zeigt, dass einige Hot Spots mit den Standorten von Fitnessstudios oder Haltestellen des Öffentlichen Personennahverkehrs in der Nähe dieser Fitnessstudios korrespondieren. Die Wörter mit der höchsten Wahrscheinlichkeit passen in diesem Kontext sehr gut. Zur Charakterisierung städtischer Räume ist dieses Thema dagegen weniger geeignet.

Thema [27] zeigt dagegen Überschneidungen mit verschiedenen Parkanlagen in der Region. Dies korreliert zudem mit einigen der wichtigsten Wörter: *park*, *garden*, *heaton*, *chorlton*, *water* und *tree*. Heaton Park und Chorlton Water Park sind zwei Parkanlagen in Manchester, die auch als Hot Spot identifiziert werden können. Diese und einige weitere Parkanlagen sind in Abbildung 15 verortet. Es gibt aber auch zwei Hot Spots am Manchester Science Park sowie am Towers Business Park, die zwar mit dem Wort *park* korrelieren, aber nicht zum angenommenen Thema passen. Bis auf wenige Ausnahmen ist das Thema [27] aber ein guter Indikator für Parkanlagen im Untersuchungsgebiet. Es werden jedoch nicht alle Parkanlagen auch als Hot Spots ausgewiesen.

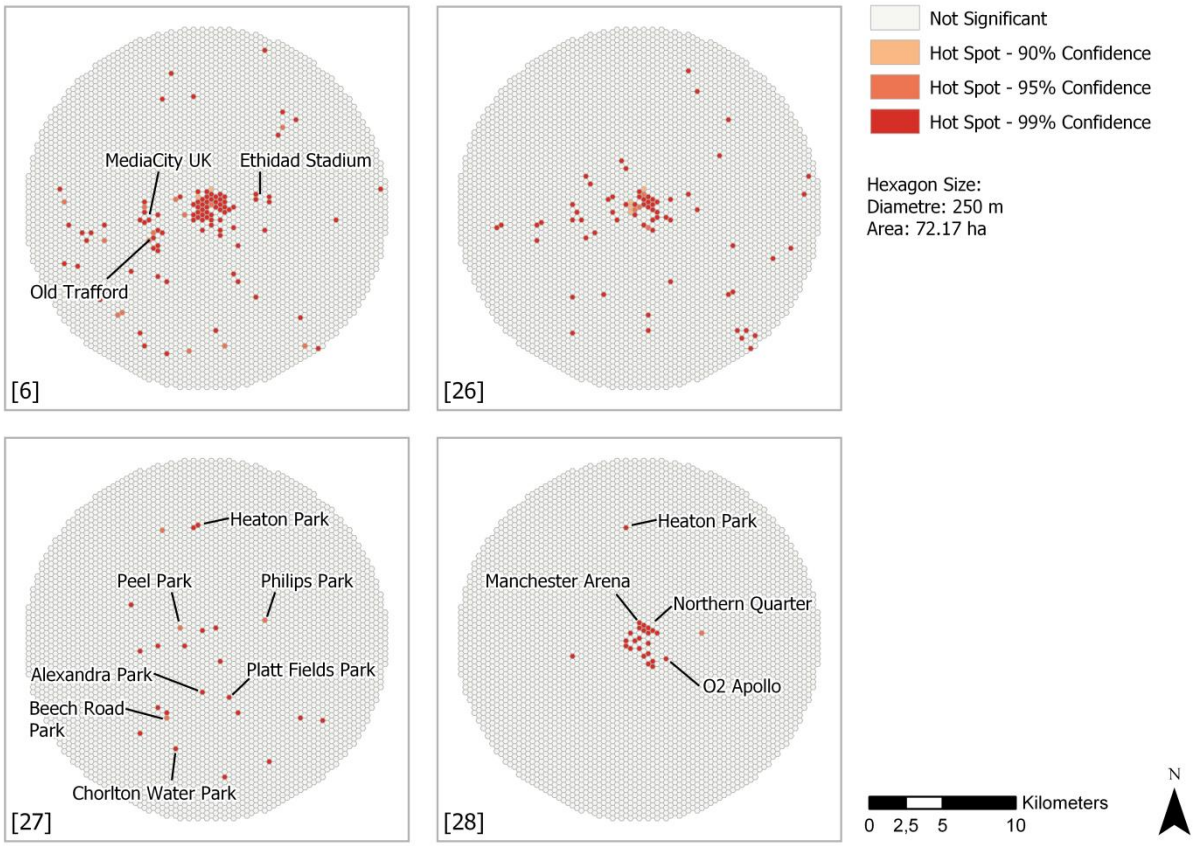


Abbildung 15: Local Getis-Ord Gi* Statistiken für Themen im Bereich Aktivitäten im Freien, Sport und Training in Manchester

Thema [28] beinhaltet unter anderem die Wörter *manchest*[er], *festiv*[al] und *parklif*[e] und bezieht sich auf das Parklife Festival im Heaton Park. Der Ort des Festivals erscheint in der räumlichen Analyse auch als Hot Spot. Die übrigen Wörter *school*, *arena*, *museum*, *academi*, *apollo*, *even*, *north* haben eine niedrige Wahrscheinlichkeit und sind nicht eindeutig zu interpretieren. Die Hot Spots im Zentrum liegen aber unter anderem im Northern Quarter, am Apollo Theatre und an der Manchester Arena. In dem Thema überschneiden sich somit Tweets zum Festival sowie zu unterschiedlichen Konzerten (siehe auch Kapitel 3.1.1).

Thema [5] ist inhaltlich nicht zu interpretieren. Zum einen weisen die Wörter *love* und *wed*[ding] auf Tweets in Zusammenhang mit Hochzeiten hin. Die Wörter *walk* und *dog* zeigen inhaltliche Zusammenhänge mit Spaziergängen. Das Wort *pic* (kurz für *picture*) weist auf das Veröffentlichen von Fotos hin, z.B. im Zusammenhang mit Instagram oder anderen Fotoapps auf Smartphones. Das Wort *beauti*[ful] wird in diesem Zusammenhang oft verwendet und dabei sehr häufig mit einem Hashtag benutzt (#beautiful). Dieses Thema ist nicht nur inhaltlich nicht zu interpretieren, sondern zeigt auch in der räumlichen Verteilung keine Muster, die mit den Inhalten korrespondieren.

3.2 Identifizierte Themen und räumliche Muster in Birmingham

Bei der Analyse von Themen für Birmingham zeigt sich, dass zwar deutlich mehr Tweets als Stichprobe vorhanden sind als für Manchester, aber bei einer vergleichbaren Anzahl an Themen weniger aussagekräftige Themen identifiziert werden können. In den folgenden drei Kapiteln wird auf drei Themenkomplexe näher eingegangen, für die mehrere Themen zuordbar sind. Aus den weiteren Themen sind dagegen keine Rückschlüsse auf räumlich-semantische Zusammenhänge möglich. Auf Grund der deutlich größeren Stichprobe und somit einer höheren Anzahl Tweets je Raumeinheit (Zelle oder Umkreis) sowie Zeiteinheit (Tag, Stunde) ermöglicht der Birmingham-Datensatz aber weitere Detailanalysen. Diese werden in einem gesonderten Kapitel 3.4 für ausgewählte Themen dargestellt.

Wie in Manchester können auch in Birmingham insbesondere Themen mit Freizeitbezug identifiziert werden: Die Bereiche Nachtleben [2, 3, 12], Essen und Trinken [14] sind genauso in den Daten wiederzufinden wie sportliche Aktivtäten [23] und Events im Allgemeinen [19] sowie Fußball im Speziellen [20]. Im Birmingham-Datensatz treten zusätzlich Themen mit Bezug zu Arbeit [17, 30] und verkehrliche Infrastrukturen [1, 18] deutlicher hervor. Themen zu Tourismus und Einkaufen tauchen dagegen nur rudimentär auf Basis einzelner Stichwörter auf, ohne dass die übrigen Begriffe der Themen einen klaren Zusammenhang zulassen.

Hinweise zu den in der Analyse nicht weiter behandelten Themen

Bei der gegenüber dem Manchester-Datensatz größeren Gruppe der Themen, die für die weitere Analyse nicht weiter verwendet werden, gibt es einzelne Themen, die keinen unmittelbaren räumlichen Kontext erkennen lassen sowie Themen, die anhand der häufigsten Wörter nicht oder nur teilweise interpretiert werden können. Die Mehrheit dieser Themen konzentriert sich im Zentrum von Birmingham. Einige Themen finden sich auch im Umfeld der University of Birmingham sowie im südlich gelegenen Stadtteil Bournbrook.

Ein Thema dreht sich um Politik und Wahlen. Ein vergleichbares Thema konnte im Manchester-Datensatz nicht identifiziert werden. Im Thema [26] erscheinen die Wörter *parti*, *news*, *vote*, *women*, *must*, *men*, *social* und *labour*. Die resultierenden Cluster konzentrieren sich auf die University of Birmingham, die Aston University und Teile der Innenstadt. Die Tweets stehen hier vor allem im Zusammenhang mit den britischen Unterhauswahlen am 7. Mai 2015.

Mehrere der unklaren Themen sind durch Adjektive, Verben und Zeitwörter mit relativ hohen Wahrscheinlichkeiten geprägt, die zugleich nicht eindeutig bestimmten Themen zuzuordnen sind. Besonders prägnante Beispiele sind dafür die Themen [6] (*week, next, wait, start, excit, weekend*), [7] (*feel, life, make, made, beetter, real, sick*), [8] (*tri, keep, stop, find*), [21] (*look, really, actual, nice, tonight, go*) und [24] (*day, good, today, hope, morn, great*). Die Themen umfassen unterschiedliche Bewertungen der Nutzer, die aber nicht in einen konkreten Kontext gesetzt werden können.

3.2.1 Nachtleben, Wochenende, Essen und Trinken

Wie bereits anhand des Manchester-Datensatzes gezeigt, enthalten viele Tweets Äußerungen, die mit dem Nachtleben in einer Stadt in Zusammenhang gebracht werden können. In Birmingham können vergleichbare Tweets zu Feiern [2, 3], zum Wochenende [12] und zu Essen und Bars [14] identifiziert werden. Die Tweets dieser Themen konzentrieren sich insbesondere im Zentrum und in der Nähe der University of Birmingham im Südwesten der Stadt sowie der Aston University nordöstlich des Zentrums.

Im Thema [2] verweisen die häufigsten Wörter *watch*, *show*, *live* und *tonight* auf Veranstaltungen, die durch die Begriffe *music* oder *film* vor allem auf Livekonzert oder Filmvorstellungen bezogen sind. *christma[s]* verweist auf Tweets im Zusammenhang mit Weihnachten. Im Datensatz sind tatsächlich viele Tweets mit Weihnachtsbezug aus dem Dezember zu finden, die zum Teil unter dieses Thema fallen. Mit *walsal* mischt sich auch eine Ortsbezeichnung für die Stadt Walsall im Nordwesten von Birmingham unter die häufigsten Begriffe. Hierbei handelt es sich allerdings um eine Verzerrung durch einen Bot, der bei der Bereinigung nicht identifiziert wurde und auf ein Online-Spiel hinweist. Die meisten Tweets mit diesem Wort im Thema [2] sind darum nicht interpretierbar, stellen aber nur

einen kleinen Anteil der Beiträge des Themas dar. Die übrigen Tweets im Thema [2] konzentrieren sich vor allem räumlich sehr stark auf das Zentrum von Birmingham.

Das Thema [3] umfasst vor allem die Wörter *happi*, *drink*, *birthday* und *photo* und mit geringerer Wahrscheinlichkeit *black*, *post*, *blue*, *beer*, *celebr[atio]* und *bro*. Somit liegt auch hier eine Durchmischung vor, bei der sowohl Tweets mit konkretem Bezug zu Geburtstagsfeiern auf der einen Seite als auch sonstige Anlässe und eher allgemeine Tweets zum Thema Trinken und Ausgehen auf der anderen Seite unter das Thema [3] fallen. Diese Tweets haben räumliche Schwerpunkte im Zentrum in der Nähe der Custard Factory, einem Einkaufs- und Wirtschaftsstandort, im Umfeld der Aston University sowie im Zentrum von Birmingham mit mehreren größeren Einkaufszentren (u.a. Bullring, Grand Central). An der University of Birmingham gibt es dagegen nur einzelne Zellen mit besonders hohen Häufigkeiten.

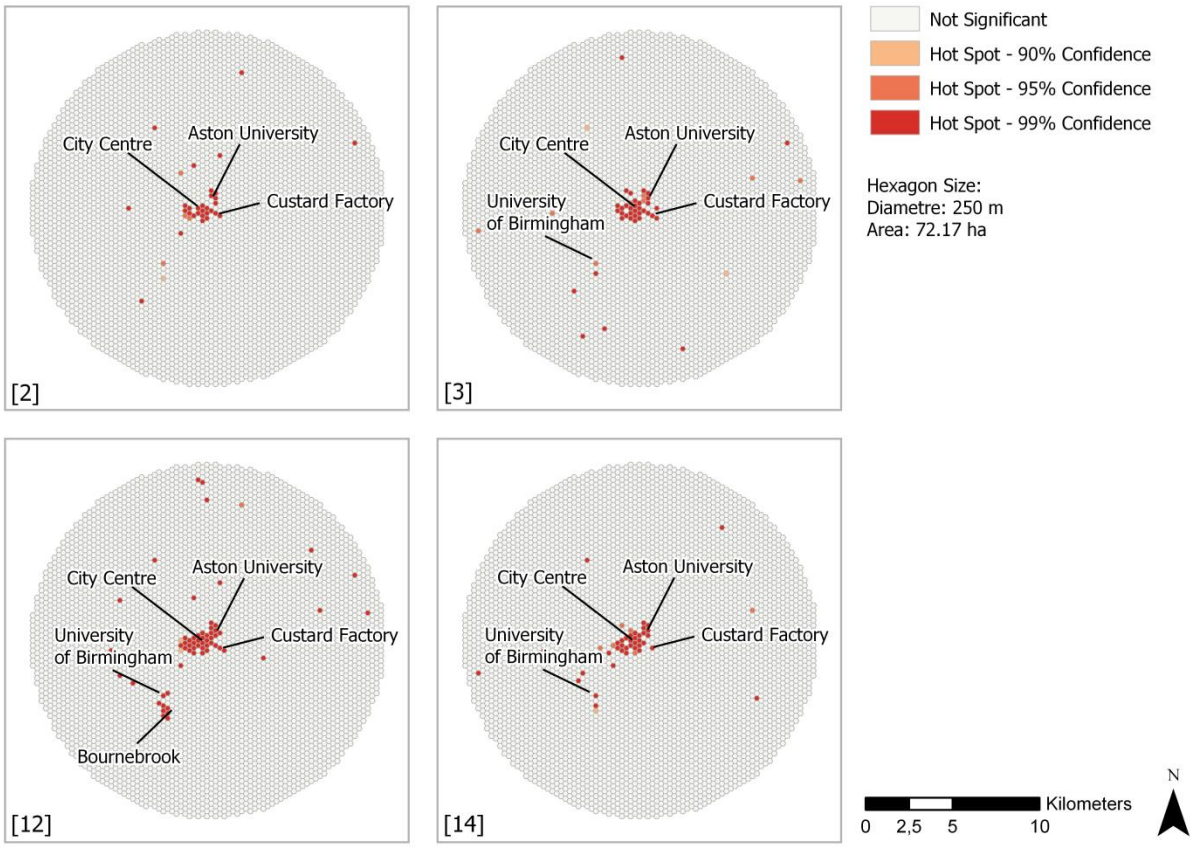


Abbildung 16: Local Getis-Ord Gi* Statistiken für Themen im Bereich Nachtleben in Birmingham

Der Stadtteil Bournbrook südlich der Universität ist unter anderem als Wohnstandort für Studierende beliebt. Hier gibt es einen räumlichen Schwerpunkt des Themas [12], das darüber hinaus ebenfalls vorwiegend im Zentrum von Birmingham anzutreffen ist. Das Thema [12] umfasst unterschiedliche Begriffe mit zeitlichen Bezügen (*time*, *night, long*, *everi*, *saturday*) sowie Wörter, die auf gemeinsame Treffen von Nutzern hinweisen können (*meet*, *wait*). Die übrigen Begriffe sind dagegen schwieriger in

den semantischen Kontext zu setzen. Aus dem räumlichen Kontext heraus kann hier aber zumindest teilweise auf einen Zusammenhang mit abendlichen Aktivitäten geschlossen werden.

Das Thema [14] fasst die häufigsten Begriffe aus dem Bereich Essen zusammen, wie z.B. *food*, *bar*, *eat*, *coffe*, *hot*, *lunch*, *tea*, *dinner*, *breakfast*, *chocol*, *cake*, *sweet*, *chicken*, *water*, *cocktail*, *wine*, *meal*, *pizza*, usw. Auch wenn die Wahrscheinlichkeiten für die einzelnen Wörter relativ gering sind (siehe Tabelle 2; dort sind allerdings nur die 10 häufigsten Wörter aufgeführt), zeigt die Fülle der unterschiedlichen Begriffe in Bezug auf Nahrung den deutlichen Zusammenhang mit dem Thema Essen. Räumliche Cluster sind auch hier insbesondere im Zentrum von Birmingham und vereinzelt im Bereich der Universität zu finden.

3.2.2 Arbeit und Verkehr

Wie bereits eingangs geschildert, treten im Birmingham-Datensatz Themen mit Bezug zu Verkehr und Arbeit deutlicher hervor als im Manchester-Datensatz. Die räumlichen Muster der im vorherigen Kapitel geschilderten Cluster im Zentrum und in der Nähe der University of Birmingham bzw. im südlich gelegenen Stadtteil Bournebrook finden sich auch bei diesen Themen wieder.

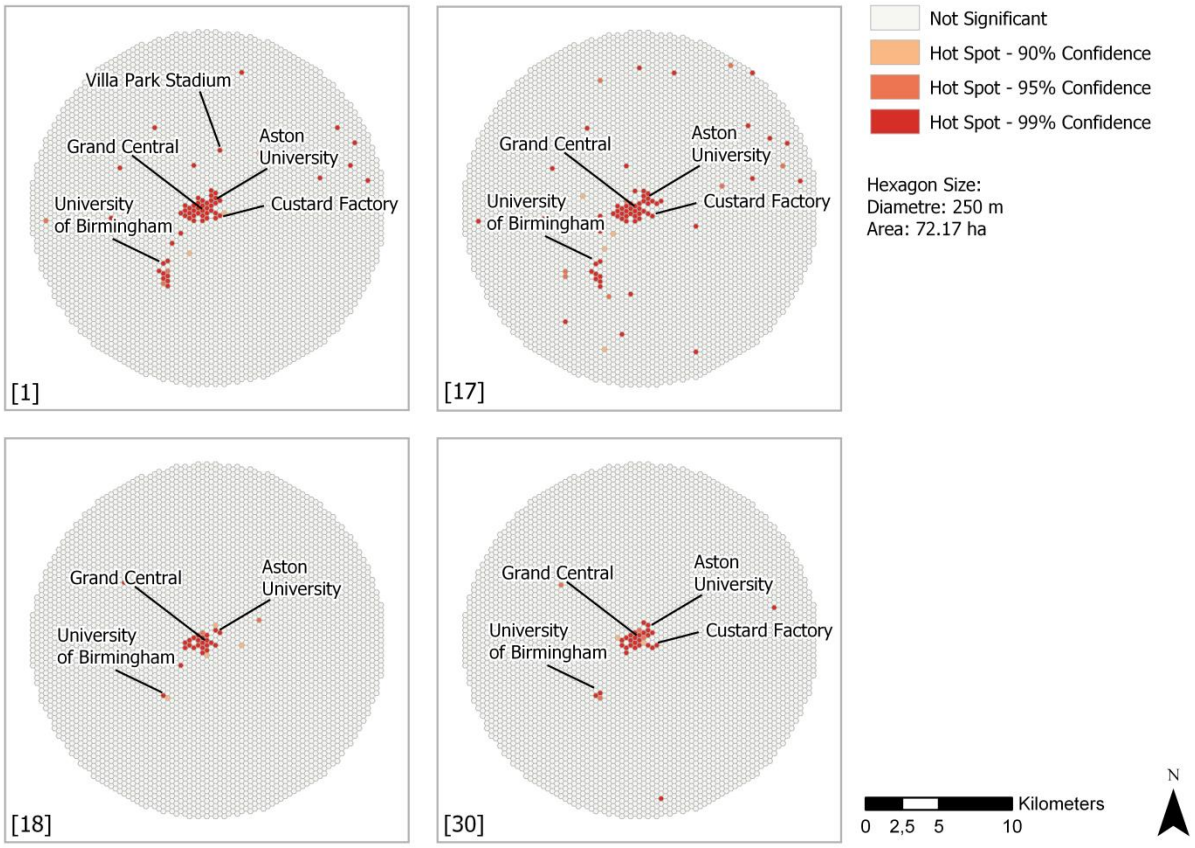


Abbildung 17: Local Getis-Ord Gi* Statistiken für Themen im Bereich Arbeit und Reisen in Birmingham

Thema [1] hat einen Bezug zu Verkehrsthemen sowie zu Schilderungen von Reisezielen bzw. Ortsreferenzen. Die Wörter *back*, *home* und *london* verweisen auf Orte. *come*, *bring*, und *go* verweisen auf Bewegungen der Nutzer. Die Begriffe *train*, *bus* und *road* beziehen sich auf unterschiedliche Verkehrsmittel. Das Thema [18] ist insofern vergleichbar, als dass auch hier Ortsangaben überwiegen, z.B. *birmingham*, *west*, *midland*, *street*, *citi*, *centr[al]*, *station* und *airport*. Die letzten beiden Begriffe verweisen sowohl auf Ortsbezeichnungen als auch auf Verkehrsinfrastrukturen. Im Thema [17] überwiegen zeitliche Bezüge (*hour*, *tomorrow*) und Wörter im Zusammenhang mit Wohnung (*sleep, bed*) und Arbeit (*work*). *get*, *readi*, *go* und *leav* verweisen u.a. auf Bewegungen. Im Thema [30] tauchen weitere Begriffe auf, die im Zusammenhang mit Arbeit stehen können, z.B. *job*, *busi*, *help*, *manag[e]* und *servic[e]*.

Alle vier Themen zeigen ähnliche räumliche Muster, wobei sich die Themen [18] und [30] noch etwas stärker auf nur wenige Zellen konzentrieren. Das Zentrum von Birmingham und die University of Birmingham treten hier jeweils deutlich hervor. In den Themen [1] und [17] gibt es mehrere einzelne Zellen mit signifikanten Clustern, die aber nicht im besonderen Bezug zu Verkehrsinfrastrukturen oder Arbeitsmarktzentren stehen.

3.2.3 Stadien und Sport

Sportthemen treten in drei Themen im Birmingham-Datensatz hervor. Die wichtigsten Begriffe im Thema [19] sind *big*, *play*, *man*, *fair*, *ball*, *kick* und *fight*. Neben den wiederum deutlichen Clustern im Zentrum, an der University of Birmingham und im Stadtteil Bournebrook erscheinen auch verschiedene Stadien als Hot Spot. Dazu gehören das Aston Park Stadium (Aston Villa Football Club), das St. Andrew's Stadium (Birmingham City Football Club) und der Edgbaston Cricket Ground (Warwickshire County Cricket Club). Das Thema [20] verweist explizit auf den Aston Villa Football Club. Die wichtigsten Wörter *game*, *villa*, *avfc*, *win*, *season*, *goal*, *fan*, *park*, *team* und *player* stellen einen deutlichen Zusammenhang zu diesem Verein her. In diesem Thema erscheint zudem nur ein kleiner Hot Spot im Zentrum. Neben dem Stadion des Vereins tauchen aber auch hier weitere Stadien als Hot Spots auf.

Das Thema [23] ist auch mit Sport assoziiert. Das wird anhand der Begriffe *gym*, *fit*, *class* und *win* deutlich. Die weiteren Begriffe *well*, *today*, *done*, *great* und *proud* zeigen aber auch, dass hier vor allem die sportlichen Aktivitäten der Nutzer selbst im Vordergrund stehen. Es gibt einen großen Hot Spot im Zentrum sowie einzelne Hotspots im Umland, z.B. um die Universitäten und den Edgbaston Cricket Ground sowie dem benachbarten Cannon Hill Park.

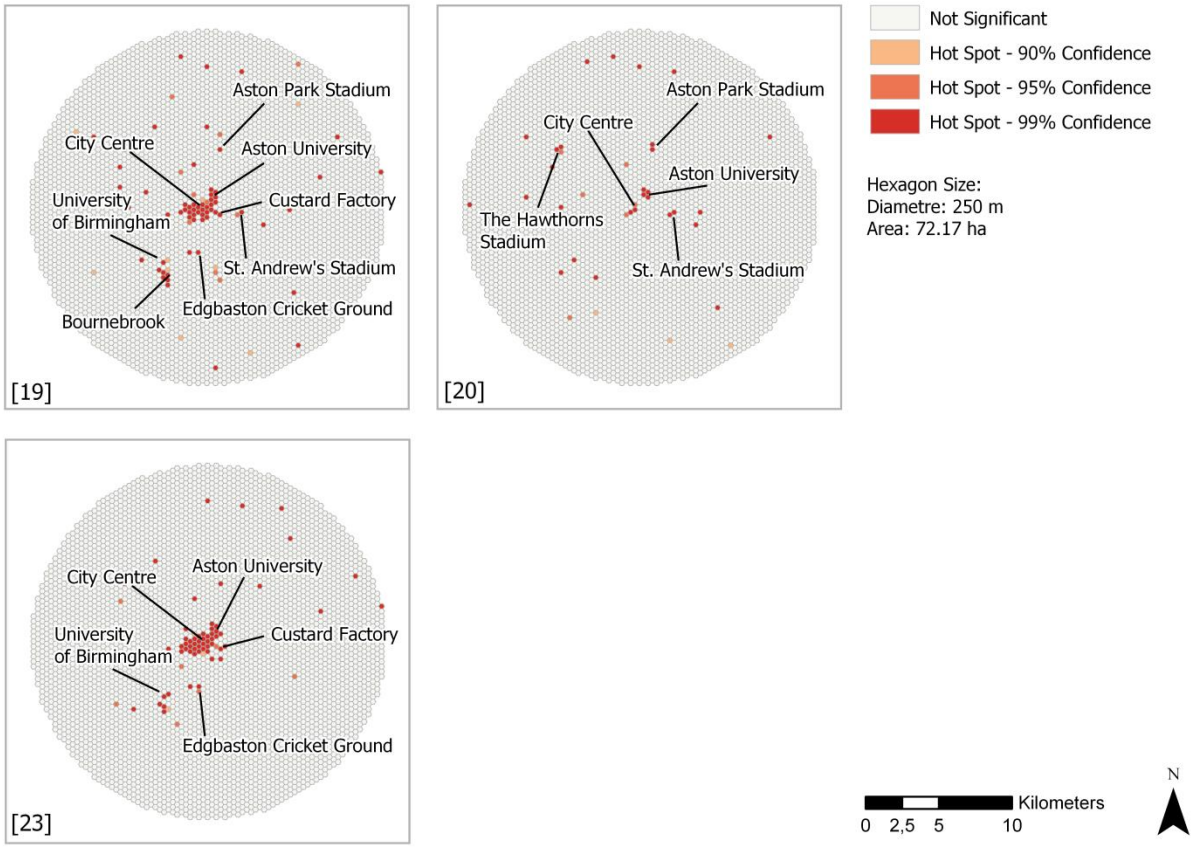


Abbildung 18: Local Getis-Ord Gi* Statistiken für Themen im Bereich Sport in Birmingham

3.3 Abgleich mit OpenStreetMap

Daten aus digitalen Netzwerken sollten nicht isoliert betrachtet werden, sondern als eine zusätzliche Datenschicht in einem erweiterten Analyseraster bzw. im Abgleich mit weiteren Datengrundlagen. Wichtig ist daher der Vergleich mit bestehenden Daten, so dass sich die Resultate gegenseitig ergänzen können (Li et al., 2013). Dazu wird die räumliche Verteilung der ausgewerteten Tweets mit ausgewählten Punktdaten aus OpenStreetMap (OSM) verglichen. Dieser Vergleich erfolgt für ausgewählte Themen, für die ein Referenzdatensatz aus OSM-Daten mit zu den Themen passenden Angeboten (Points of Interest) erstellt werden kann. Da es sich bei OSM um ein Projekt im Sinne von Volunteered Geographic Information handelt, also die Daten von diversen Akteuren aus unterschiedlichen Datenquellen und eigenen Erhebungen zusammengetragen werden, muss zunächst davon ausgegangen werden, dass die vorliegenden Standorte keine vollständige Erhebung darstellen und einzelne Objekte nicht aktuell oder sogar falsch sein können (Goodchild and Li, 2012). Für den vorliegenden Vergleich der beiden Datenquellen, Twitter und OSM, wird zunächst angenommen, dass sie vergleichbare Freizeitnutzungen in der Stadt abbilden können. Während die Daten aus OSM die Dimension der Angebote als Punkte repräsentieren, z.B. Bars, Restaurants, Geschäfte, stellen die georeferenzierten Tweets die Nutzerdimension in Form von Ereignissen

(Beiträgen) dar. Beide Perspektiven, die Angebote und ihre Nutzer, sollten räumliche Überschneidungen innerhalb einer Stadt haben.

Es liegen Referenzdatensätze aus OSM vor mit den Standorten von Einrichtungen, die zu den folgenden Themen zusammengefasst werden:

- Nachtleben: Bar, Pub, Nachtclub, Biergarten
- Essen: Restaurant, Fast Food, Café
- Einkaufen: Kaufhaus, Kleidungsgeschäft, Schuhgeschäft, Sportgeschäft, Spielwarengeschäft, Juwelier, Kosmetikgeschäft, Buchgeschäft, Florist, Friseur, Optiker, Souvenirladen, Bank
- Tourismus: Museen, Attraktionen, Monumente

Mittels Verschneidung wird die Anzahl der Standorte je Thema und je Hexagon bestimmt. Für diese aggregierten Daten erfolgt die Identifikation von Clustern mit den beschriebenen Methoden zu Global Moran's I und Local Getis-Ord Gi* (Kapitel 2.5.2 und Kapitel 2.5.3). Eine Übersicht der Ergebnisse von Global Moran's I auf Basis der 250-Hexagonzellen ist im Anhang in Tabelle 12 für Manchester und Tabelle 13 für Birmingham zu finden. Für eine Analyse des Tweetaufkommens im Umfeld von Bahnhöfen (Kapitel 3.4.1) werden zudem alle Bahnhaltestellen aus dem OSM-Datensatz gespeichert.

3.3.1 Manchester

Die räumlichen Muster der identifizierten Cluster spiegeln die Siedlungsstruktur der Region wieder. Zwischen den vier Datensätzen zeigen sich aber auch deutliche Unterschiede, insbesondere bedingt durch die unterschiedliche Anzahl an Objekten. Die höchsten z-Werte werden für Bars, Pubs und Nachtclubs bei 1.100 m sowie für Restaurants und Fast Food-Angebote bei 1.000 m erreicht. Beide Datensätze zeigen eine Vielzahl von Hot Spots im Untersuchungsgebiet. Einkaufsmöglichkeiten sind dagegen stärker konzentriert und es werden deutlich weniger Hot Spots sichtbar. Für touristische Anlaufpunkte gibt es nur eine geringere Anzahl an Punkten im Untersuchungsgebiet, so dass es nur wenige signifikante Hot Spots gibt. Insbesondere für die Sehenswürdigkeiten ist also davon auszugehen, dass sie sich nur in geringem Maße als Referenzdatensatz eignen. Das hängt unter anderen aber auch damit zusammen, dass es bei OSM zwar Vorgaben für die Benennung und Klassifizierung von Objekten gibt, diese aber nicht immer einheitlich genutzt werden.

Die vergleichsweise hohe Anzahl an Hot Spots mit Bezug auf Bars, Pubs und Nachtclubs erklärt sich durch die hohe Anzahl an Punkten im gesamten Gebiet. Im Zentrum von Manchester genauso wie in vielen Ortsteilzentren sowie Stadtzentren der kleineren Umlandstädte in der Region sind vergleichsweise viele Punkte kartiert. Dies gilt gleichermaßen, wenn auch nicht in so hoher Dichte, für die Angebote an Restaurants und Fast Food Angeboten.

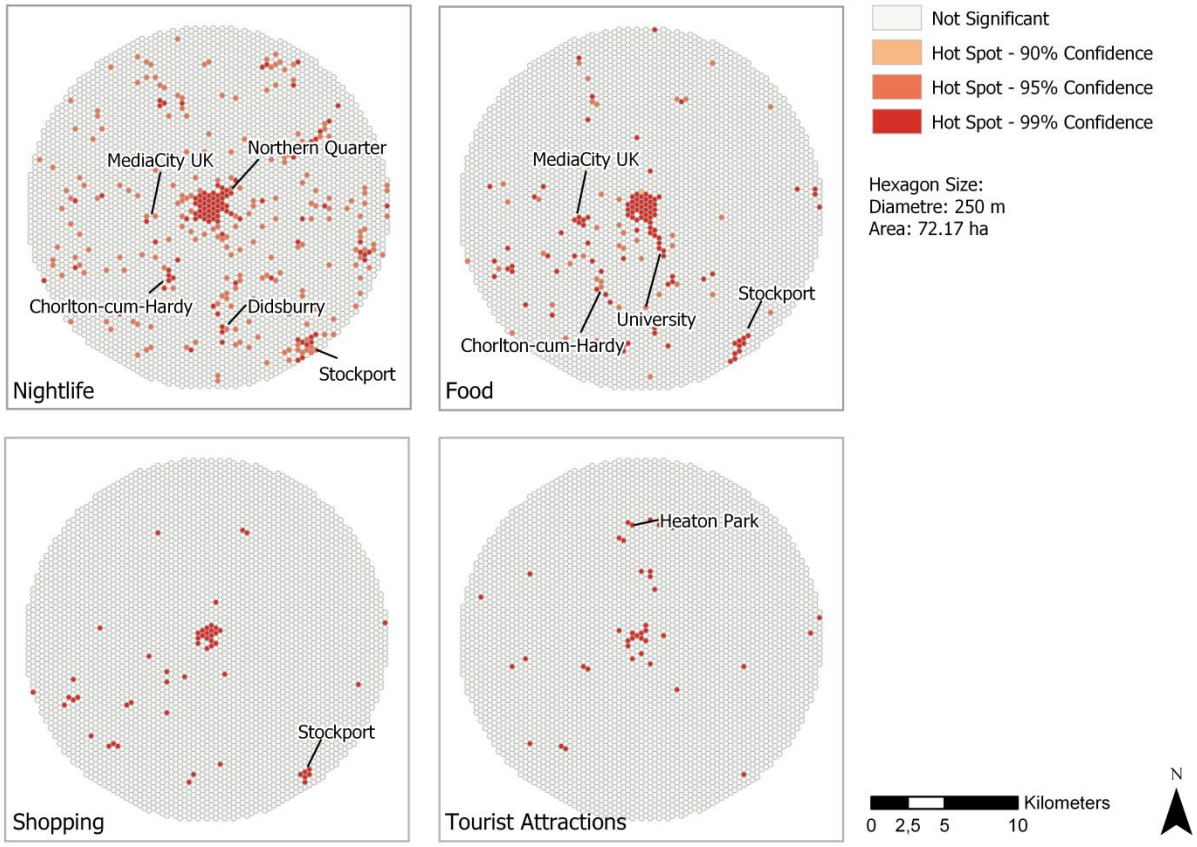


Abbildung 19: Local Getis-Ord Gi* Statistiken für OpenStreetMap-Daten in Manchester

Die Häufigkeit der OSM-Punktobjekte je Zelle und Kategorie wird als Vergleichswert den Häufigkeiten der Tweets vergleichbarer Themenfelder gegenübergestellt. Dazu werden alle Zellen ausgewählt, die mindestens 5 Tweets enthalten. Somit können 1.681 von 21.406 Zellen im Untersuchungsgebiet berücksichtigt werden. Die Häufigkeiten der Tweets für diese Themen werden addiert. Im Ergebnis liegt die Zahl der Tweets je Hexagonzelle und je Themenbereiche vor. Für die Zähler der ausgewählten Tweets sowie der Punkte aus OSM erfolgt eine Studentisierung der Werte je Hexagonzelle. Dafür wird das arithmetische Mittel subtrahiert und das Ergebnis durch die Standardabweichung dividiert. Diese normalisierten Werte sind somit besser vergleichbar, als die sehr unterschiedlich verteilten Variablen und können auf eine Korrelation hin geprüft werden.

Abbildung 20 visualisiert die Ergebnisse mittels Scatterplots. Für die Themen Nachtleben und Essen liegt eine positive Korrelation nach Pearson von 0,43 bis 0,41 vor. Das R^2 liegt bei 0,18 bzw. 0,16 für diese beiden Themenfelder. Das Ergebnis zeigt also eine positive Korrelation zwischen den standardisierten Zählern der Tweets sowie der Anzahl der Points of Interest aus OSM. Im Themenfeld Einkaufen ist ebenfalls eine positive Korrelation, jedoch ein deutlich schwächerer Zusammenhang festzustellen. Die Korrelation nach Pearson beträgt nur 0,24 und das R^2 liegt bei unter 0,06. Die touristischen Sehenswürdigkeiten korrelieren praktisch gar nicht mit der Anzahl der vorliegenden

Tweets. Hier ist die Diskrepanz zwischen den wenigen Standorten aus OpenStreetMap sowie den Themen aus Twitter, die zugleich nicht ausschließlich touristische Anziehungspunkte behandeln, am höchsten.

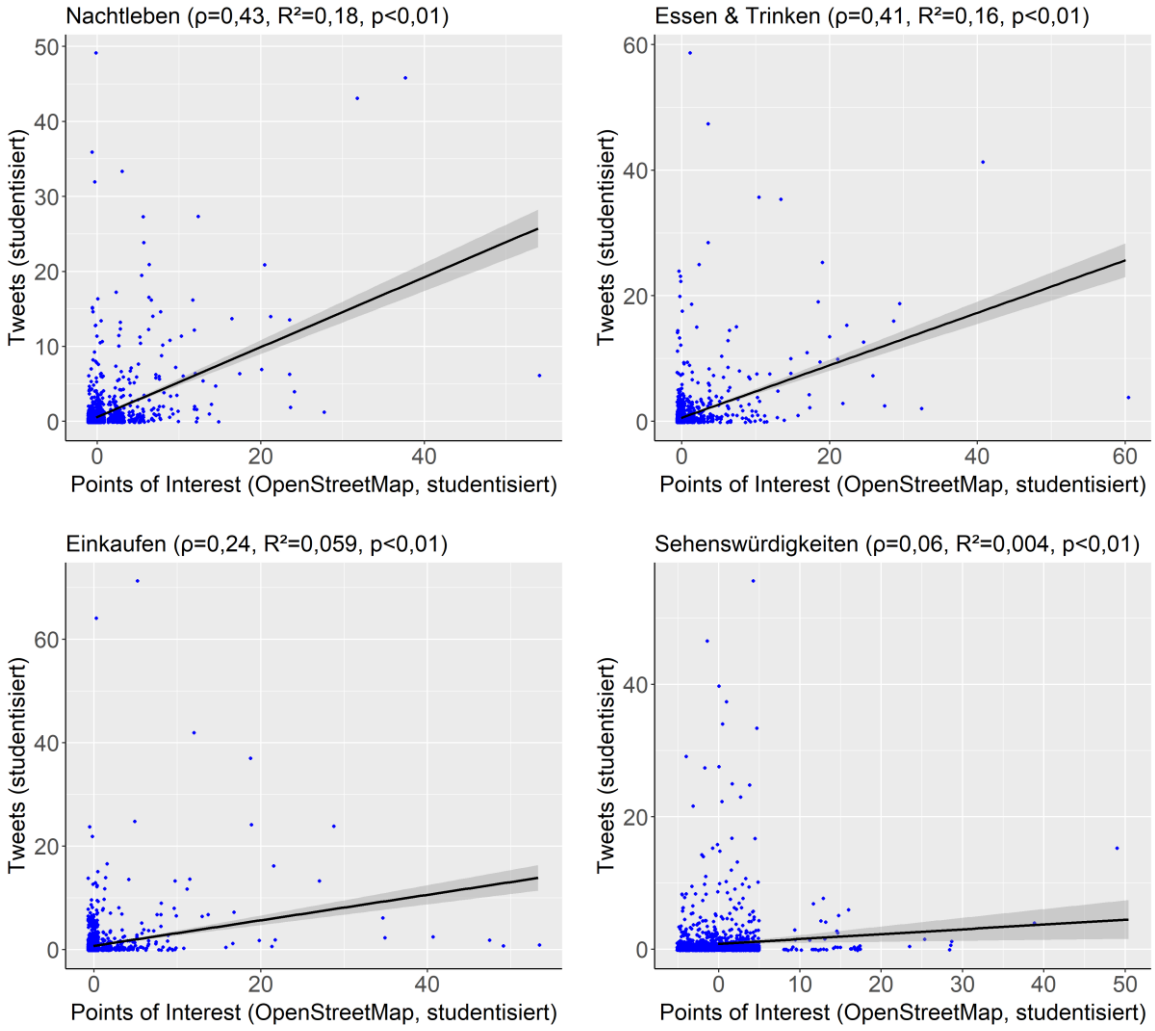


Abbildung 20: Korrelation zwischen Tweet-Häufigkeiten und OSM-Häufigkeiten für ausgewählte Themen in Manchester

Abschließend erfolgt für den Themenbereich Nachtleben ein Vergleich zwischen der Häufigkeit der Tweets je Hexagonzelle sowie den identifizierten Hot Spots für Bars, Pubs und Nachtclubs aus OSM, um dieses Phänomen räumlich weiter einzugrenzen. Dieses Thema eignet sich für den Vergleich am besten, da hier die Korrelation zwischen Tweets und OSM-Punkten am höchsten ist. Abbildung 21 zeigt in Schrägansicht das Untersuchungsgebiet um Manchester. Die Höhe der dargestellten Säulen basiert auf der absoluten Anzahl an Tweets aus den Themen [1, 10, 22, 32, 28].

Eine hohe Korrelation hinsichtlich der Themen zeigt sich bei den Säulen im Northern Quarter von Manchester. Dieser Stadtteil stellt einen wichtigen Anlaufpunkt für kulturelle Veranstaltungen und für das Nachtleben in Manchester dar (siehe dazu Kapitel 3.1.1). Hier gibt es sowohl ein großes Angebot an Pubs, Bars und Nachtclubs als auch eine besonders hohe Anzahl an Tweets im

betrachteten Themenbereich. Dies gilt auch für die südwestlichen Teile des Zentrums, in dem es ebenso ein sehr großes korrespondierendes Angebot sowie eine hohe Anzahl an Tweets gibt. Weitere Punkte mit hohen Dichten an Tweets und korrespondierenden Angeboten sind die Zentren von Chorlton cum Hardy, Didsbury, Stockport, Denton, Fallsworth, Middleton, Prestwich und die MediaCity UK.

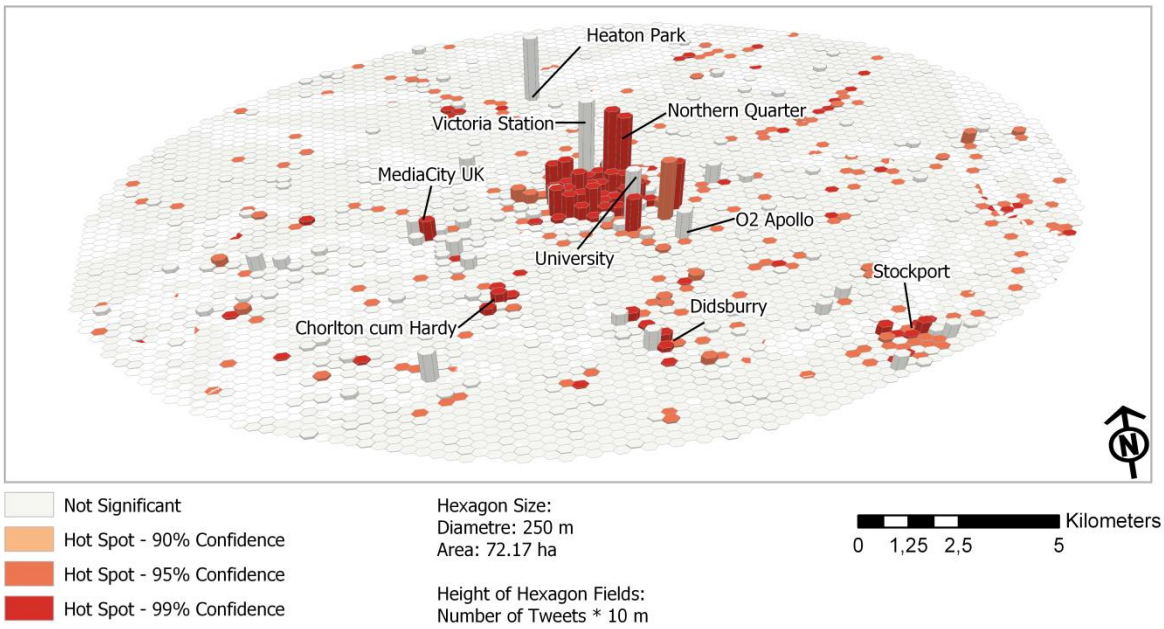


Abbildung 21: Local Getis-Ord Gi* Statistiken für OpenStreetMap-Daten und Häufigkeit korrespondierender Tweets in Manchester (Höhe der Säulen) für den Themenbereich Nachtleben

Es besteht für die dargestellten Themen eine positive Korrelation zwischen den Standorten, die in OSM kartiert wurden, sowie den Tweets der ausgewählten Themenfelder. Zugleich gibt es aber auch eine hohe Anzahl an Zellen, in denen zwar keine Punkte kartiert wurden, jedoch eine vergleichsweise hohe Anzahl an Tweets vorliegt. Zudem existieren eine ganze Reihe an Zellen, die zwar Punktdaten aus OSM, jedoch keine oder nur eine geringe Zahl an Tweets enthalten. Die höchste Säule mit der maximalen Anzahl an Tweets liegt im Bereich der Victoria Station und weist zugleich keinen Hot Spot für den Bereich Nachtleben aus. Im Bereich des Bahnhofs gibt es keine Punkte in den OSM-Daten. Eine weitere Säule mit einer hohen Anzahl an Tweets aus dem Themenbereich, aber ohne Hot Spot mit Bars oder Pubs, befindet sich auf dem Gebiet der Universität Manchester. Eine dritte, auffällige Säule ist an der Stelle zu finden, an der das Heaton Park Festival stattfindet (siehe dazu auch Kapitel 3.1.5). Eine weitere Säule dieser Art ist am Standort der Konzerthalle O2 Apollo zu finden.

Die ersten beiden dargestellten Beispiele mit einer hohen Anzahl an Tweets ohne räumliche Überschneidung mit entsprechenden Angeboten zeigen, dass sich die semantische Dimension der Tweets nicht in allen Fällen eindeutig mit der räumlichen Dimension überdeckt. Orte mit einer hohen

Fluktuation von Menschen weisen zugleich eine hohe Anzahl an Beiträgen auf. Die Tweets müssen sich jedoch in den einzelnen Themen nicht immer mit dem Ort decken, sondern können auch andere semantische und räumliche Bezüge aufweisen. Im Fall der Victoria Station kann zunächst vermutet werden, dass hier viele Besucher von Manchester den Ort passieren oder dort Zeit verbringen. Dabei werden dann Tweets veröffentlicht, auch wenn das eigentliche Ziel vielleicht im Zentrum liegt oder der Nutzer sich bereits auf dem Weg nach Hause befindet und der Bahnhof nur eine Zwischenstation ist. Das O2 Apollo sowie der Heaton Park zeigen darüber hinaus, dass diese im Zusammenhang von Konzerten besonders genutzten Räume zudem sehr stark mit der Dichte an Tweets korrelieren. Hier lässt sich der Zusammenhang ohne weitere Datengrundlagen herstellen.

Eine methodisch vergleichbare Analyse wurde von Steiger et al. (2015b) als Vergleich zwischen Zensus-Bevölkerung und Tweets durchgeführt. Auch hier wurde ein positiver Zusammenhang zwischen Arbeitsstätten und Wohnumfeld sowie zwischen Tweets mit Arbeitsbezug oder Wohnungsbezug festgestellt. Die Korrelation der arbeitsbezogenen Daten war dabei deutlich höher. Die Autoren stellen nämlich fest, dass viele wohnungsbezogene Tweets mit den Standorten von Bahnhöfen korrelieren. Vor diesem Hintergrund muss davon ausgegangen werden, dass Thema und Standort nicht immer vordergründig zusammenhängen, sondern eben bestimmte Themen auch mit Orten korrelieren können, die eine hohe Anzahl an Passanten vorweisen, wie dies z.B. bei Bahnhöfen der Fall ist. Die Veröffentlichung neuer Inhalte in den sozialen Medien, z.B. von Tweets, erfolgt dann „auf der Durchreise“, aber nicht an dem Ort, auf den sich der Tweet bezieht.

3.3.2 Birmingham

Im Gegensatz zu Manchester erfolgt diese Analyse hier nur für zwei Themenbereiche: Nachtleben und Essen. Für Tourismus und Shopping konnten keine eindeutigen Themen in Birmingham identifiziert werden. Points of Interest mit Bezug zum Nachtleben konzentrieren sich in den OSM-Daten im Zentrum von Birmingham. Anders als in der Region um Manchester bilden die Daten nicht die kleineren Zentren am Stadtrand und um die Stadt herum ab. Nur für Restaurants und Fast Food-Angebote erscheinen hier einzelne Hot Spots. Am auffälligsten sind die Stadtteile Small Heath und Saltley, die aber bisher in der Analyse der Tweets nicht besonders hervorgetreten sind. Die Daten vermitteln somit ein anderes räumliches Bild, als in der Region Manchester, wo neben dem Stadtzentrum auch das Umland durch eine Vielzahl von Hot Spots gekennzeichnet ist. In Birmingham konzentrieren sich die Angebote dagegen viel stärker auf das Zentrum. Der direkte Vergleich ist aber schwierig, da sich die Siedlungsstrukturen und die Größe der Kernstädte beider Regionen unterscheiden.

Abbildung 22: Local Getis-Ord Gi* Statistiken für OpenStreetMap-Daten in Birmingham

Die Häufigkeit der Punktobjekte aus OSM je Zelle und Kategorie wird auch für Birmingham als Vergleichswert den Häufigkeiten der Tweets der zugeordneten Themenfelder gegenübergesetzt. Für diesen Datensatz gilt gleichermaßen, dass nur Zellen mit mindestens 5 Tweets in die Auswertung einfließen. Somit können 3.879 von 4.493 Zellen im Untersuchungsgebiet berücksichtigt werden. Das sind deutlich mehr Zellen, als im Manchester-Datensatz. Die Häufigkeiten der Tweets für die Themen [2], [3] und [12] werden für den Bereich Nachtleben addiert. Der Bereich Essen basiert auf dem Thema [14]. Für die Zähler der ausgewählten Tweets sowie der Punkte aus OpenStreetMap erfolgt auch hier eine Studentisierung der Werte je Hexagonzelle.

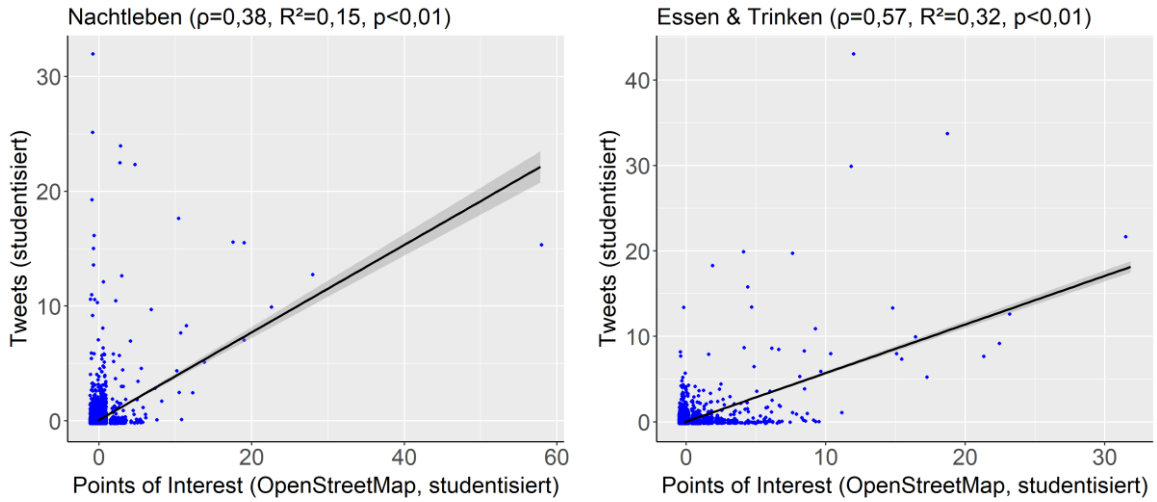


Abbildung 23: Korrelation zwischen Tweet-Häufigkeiten und OSM-Häufigkeiten für ausgewählte Themen in Birmingham

Abbildung 23 zeigt die Ergebnisse als Scatterplots. Eine positive Korrelation nach Pearson mit einem Wert von 0,38 liegt für die Themen in Bezug auf das Nachtleben vor. Das entspricht dem Wert aus der Manchester-Analyse. Das R^2 liegt bei 0,15. Im Themenfeld Essen wird eine positive Korrelation mit dem stärksten Zusammenhang bei einem Korrelationskoeffizienten nach Pearson von 0,57

gemessen. Das R^2 liegt bei 0,32. Für Birmingham kann also ein starker statistischer Zusammenhang zwischen den Standorten von Restaurants und Cafés sowie den Tweets mit Bezug auf Essensthemen festgestellt werden.

Der Vergleich von Tweets und OSM zeigt, dass es in der Region Birmingham neben dem Stadtzentrum weniger kleine Zentren mit höheren Dichten an Beiträgen bzw. Angeboten gibt. Die Datenanalyse zeigt vielmehr eine starke Konzentration auf das Zentrum der Stadt. Das wird auch anhand der folgenden Abbildung deutlich, in der für die identifizierten Cluster der OSM-Daten für Restaurants und Fast Food Angeboten den Tweethäufigkeiten aus dem Thema [14] mit Bezügen zu Essen und Trinken gegenübergestellt werden. Bis auf wenige Ausnahmen korrespondieren die Zellen mit hohen Häufigkeiten mit den Clustern und das insbesondere im Zentrum. Auch hier wird die im Vergleich viel größere Bedeutung des Zentrums gegenüber dem Umland sichtbar.

Bei den identifizierten Hot Spots mit zu gleich sehr hoher Anzahl an Tweets stechen einige Orte besonders hervor. Südwestlich der Innenstadt erscheinen die University of Birmingham und der Stadtteil Bournbrook. Im Stadtzentrum treten die höchsten Dichten an Tweets an der New Street Station bzw. dem Grand Central Einkaufszentrum, am Bullring Shopping Centre, am ODEON Broadway Plaza und am Canal Wharf auf. Das ODEON ist ein Kino mit diversen Essensmöglichkeiten und weiteren Angeboten im Umfeld. Beim Canal Wharf handelt es sich um eine Wohnanlage an einem alten Kanal im Stadtzentrum. Im Umfeld gibt es eine Vielzahl von Restaurants und auf der gegenüberliegenden Seite des Kanals liegt das Mailbox Einkaufszentrum.

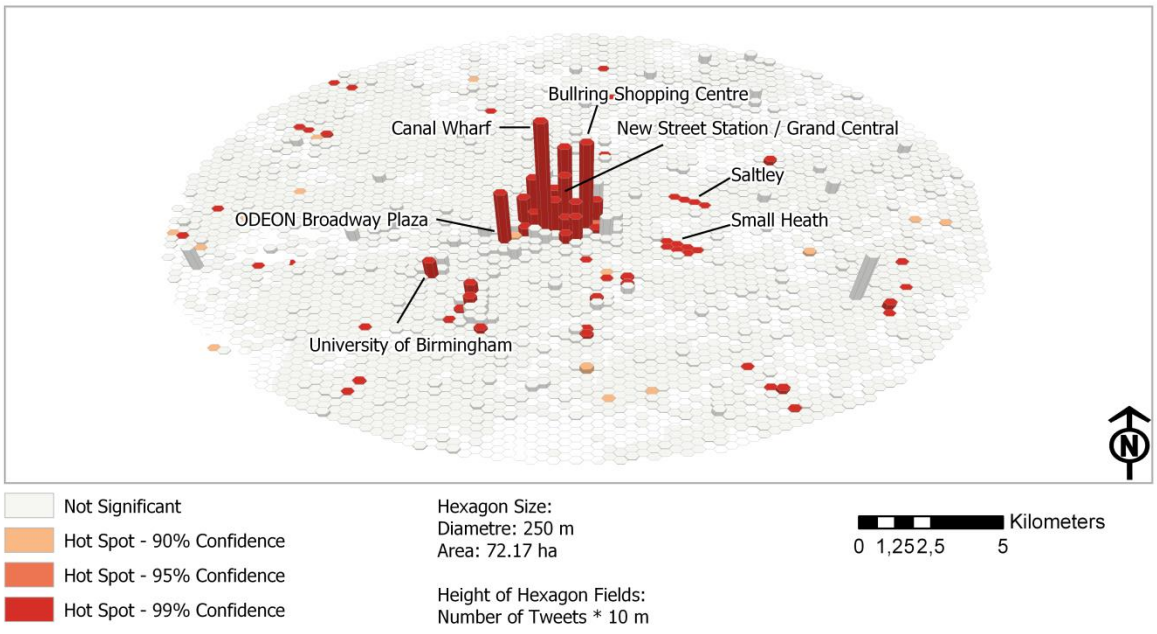


Abbildung 24: Local Getis-Ord Gi* Statistiken für OpenStreetMap-Daten und Häufigkeit korrespondierender Tweets in Birmingham (Höhe der Säulen) für den Themenbereich Nachtleben

In den Stadtteilen Small Heath und Saltley korrespondieren die OSM-Daten nicht mit den Tweets. Obwohl es hier entlang der jeweiligen Hauptstraße eine größere Anzahl an Essensangeboten gibt, liegen nur sehr wenige Tweets mit Bezug zu Essensthemen aus diesen Stadtteilen vor. Eine nähere Betrachtung der Restaurants zeigt, dass es sich hier überwiegend um migrantisch geprägte Angebote, z.B. pakistanische, afghanische oder türkische Restaurants, sowie sonstige Fast Food Angebote handelt. Das trifft auf beide Hauptstraßen in den Stadtteilen zu (Coventry Road in Small Heath, Alum Rock Road in Saltley). Möglicherweise kommen hier solche sozialen Ausschlussmechanismen zu tragen, die Li et al. (2013) am Beispiel von Kalifornien gezeigt haben. Das hätte zur Konsequenz, dass die in den beiden Stadtteilen vertretenen Milieus weniger auf Twitter vertreten sind und räumliche Nutzungsmuster somit durch die Daten nicht abgebildet werden. Die Filtermechanismen der Bereinigung mit dem Ausschluss nicht englischer Zeichen kann ebenso zu einem Teil dazu beigetragen haben, dass Tweets aus diesen Milieus weniger berücksichtigt werden.

3.4 Raum-zeitliche Korrelation von Tweets in Birmingham

Der Birmingham-Datensatz ermöglicht durch die große Stichprobe weitere Detailanalysen. Dabei können die räumliche und die semantische Ebene auf einzelne Tage oder Tageszeiten heruntergebrochen werden. Der Manchester-Datensatz enthält über die gesamte Region gesehen, aber auch für die Stadt Manchester selbst auf einzelne Tage bezogen, eine deutlich geringere Anzahl an Tweets. Außerdem umfasst die Birmingham-Stichprobe Daten für ein gesamtes Jahr vom 1.1.2015 bis zum 31.12.2015. Der Manchester-Datensatz ist dagegen auf wenige Monate innerhalb des Jahres 2017 beschränkt.

Neben der räumlichen Verteilung und dem Faktor der unterschiedlichen Häufigkeit von Beiträgen je Nutzer stellt die zeitliche Dimension einen weiteren wichtigen Faktor im Datensatz dar. Wie in den folgenden Abbildungen für beide Datensätze dargestellt wird, folgt die Häufigkeit der Tweets dem typischen Tagesverlauf mit der niedrigsten Anzahl an Beiträgen in den frühen Morgenstunden, steigenden Beitragszahlen tagsüber bis in die Abendstunden und dem absoluten Peak zwischen 22 und 23 Uhr (Abbildung 25 und Abbildung 26). Die vorliegenden Datensätze zeigen somit das erwartete zeitliche Tagesmuster, wie es in anderen Studien ebenfalls gezeigt wird (Steiger et al., 2015b, Li et al., 2013). Somit können die Daten nicht nur zur Abbildung räumlicher Zusammenhänge, sondern auch für zeitbezogene Fragestellungen genutzt werden.

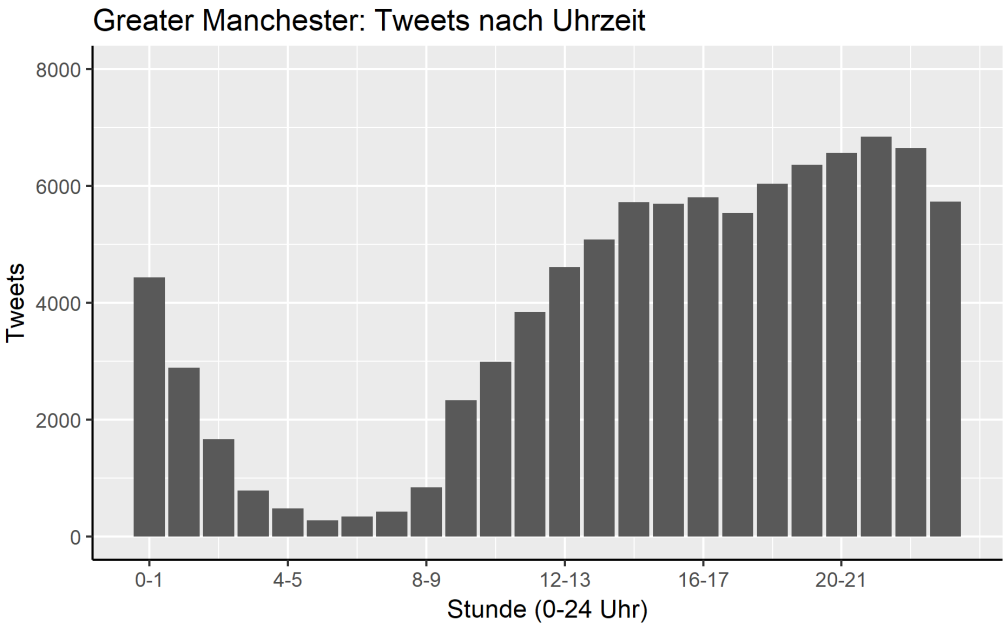


Abbildung 25: Geocodierte Tweets im Tagesverlauf in Manchester (bereinigter Datensatz)

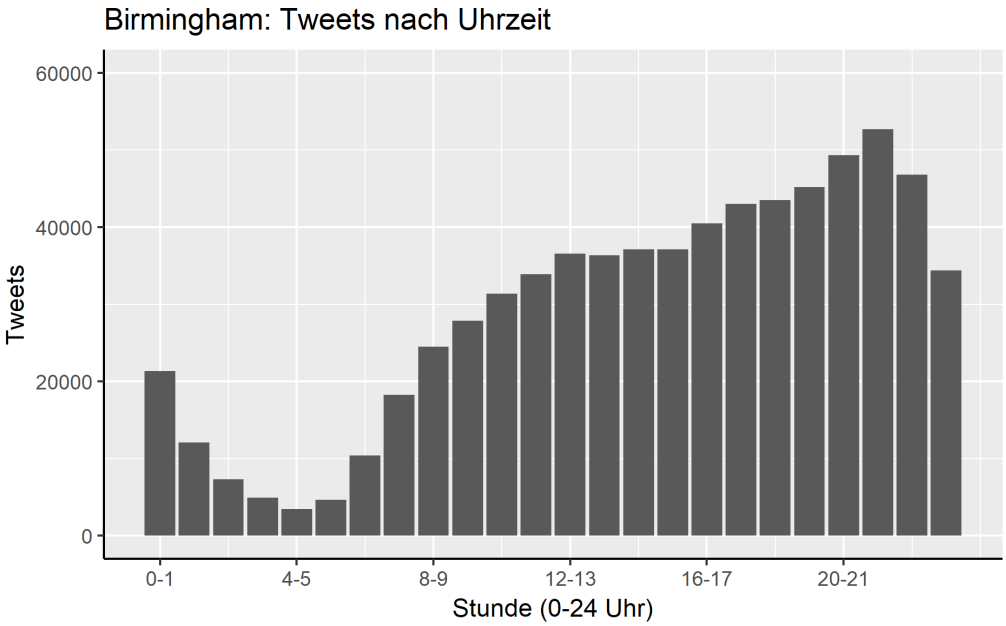


Abbildung 26: Geocodierte Tweets im Tagesverlauf in Birmingham (bereinigter Datensatz)

In Kapitel 3.4.1 erfolgt die räumlich-zeitliche Analyse im Umfeld von Bahnhöfen anhand des Tagesverlaufs über 1-Stunden-Scheiben. Dafür kann zunächst die gesamte Stichprobe untersucht werden, weil der Tagesverlauf im Vordergrund steht. Die Analyse in Kapitel 3.4.2 untersucht raum-zeitliche Zusammenhänge im Umfeld des Villa Park Stadions in Abhängigkeit von Spielen des Aston Villa Football Club. Dabei werden die Tagessummen der jeweils betrachteten Tweets als Indikator verwendet. Der Birmingham-Datensatz muss für diese Analyse in zwei Zeitfenster unterteilt werden.

Im Zeitraum 1.1.2015 bis zum 26.04.2015 liegen pro Tag im Mittel etwa 4.792 Tweets in der Stichprobe vor. Ab dem 27.04.2015 liegt das Tagesmittel dagegen nur bei rund 608 Tweets pro Tag. Da in den zweiten Zeitraum mehr Spiele fallen, aber beide Zeiträume nicht in einer Zeitreihe vergleichbar sind, wird hier der Zeitraum ab dem 27.04.2015 als Grundlage verwendet.

3.4.1 Tageszeiten im Umfeld von Bahnstationen

Die Analyse von Tweets im Tagesverlauf erfolgt im Umfeld von Bahnhaltestellen innerhalb des Untersuchungsraumes in Birmingham. Zunächst wurden auf Basis von OSM alle Bahnhaltestellen in Birmingham und Umgebung identifiziert. Die New Street Station im Stadtzentrum wurde herausgenommen, da sich direkt über dem Bahnhof das Grand Central Einkaufszentrum befindet, das zugleich die höchste Dichte an Tweets im gesamten Untersuchungsraum aufweist (siehe Kapitel 2.2). Hier wäre es ansonsten zu starken Verzerrungen bei der Analyse von Tweets im Bahnhofsumfeld gekommen.

Zunächst erfolgt eine semantische Differenzierung aller Tweets der Themen [1], [17], [18] und [30], die in Kapitel 3.2.2 als Themen mit Bezug zu Arbeit, Reisen und Verkehr beschrieben werden. Alle Tweets, die nicht diesen vier Themen zugeordnet sind, werden in einem zweiten Teildatensatz getrennt betrachtet. Zur Berücksichtigung der räumlichen Dimension erfolgte dann eine Unterscheidung zwischen Tweets, die im Umkreis von 100 m um die Haltestellen verortet sind, und allen anderen Tweets im Untersuchungsgebiet. Im Ergebnis liegen somit vier Teildatensätze vor.

Die vier Teildatensätze können somit im Hinblick auf raum-zeitliche Muster anhand der Uhrzeit, zu der die Veröffentlichung der Tweets erfolgte, untersucht werden. Abbildung 27 bildet den Tagesverlauf der vier Datensätze ab. Dabei ist auf jeden Teildatensatz bezogen der Anteil der Tweets in einem Stundenbucket (z.B. 8-9 Uhr) an allen Tweets des Teildatensatzes (0-24 Uhr) dargestellt.

zIn den Nachtstunden am Anfang des Tages befinden sich Tweets im Haltstellenumfeld auf einem Tagestiefpunkt während alle anderen Tweets in der Häufigkeit bis etwa 5 Uhr kontinuierlich zurückgehen. Im Bahnhofsumfeld steigen dagegen ab etwa 2 Uhr die Häufigkeiten wieder an und erreichen zwischen 7 und 9 Uhr das erste Tageshoch (je Stunde etwa 7,5 % aller Tweets dieses Teildatensatzes). Ab etwa 9 Uhr gleichen sich alle Kurven in der relativen Häufigkeit an (zwischen 5 und 6 % aller Tweets je Stunde). Zwischen 16 und 18 Uhr erreichen dann Tweets im Haltestellenumfeld das zweite Tageshoch (erneut rund 7,5 %) und gehen danach kontinuierlich in ihrer relativen Häufigkeit zurück. Parallel steigt die Häufigkeit der nicht verkehrs- bzw. arbeitsbezogenen Tweets außerhalb der Haltestellen allmählich auf den Höchstpunkt, der zwischen 21 und 22 Uhr erreicht wird (etwa 7,5 % aller Tweets dieses Datensatzes).

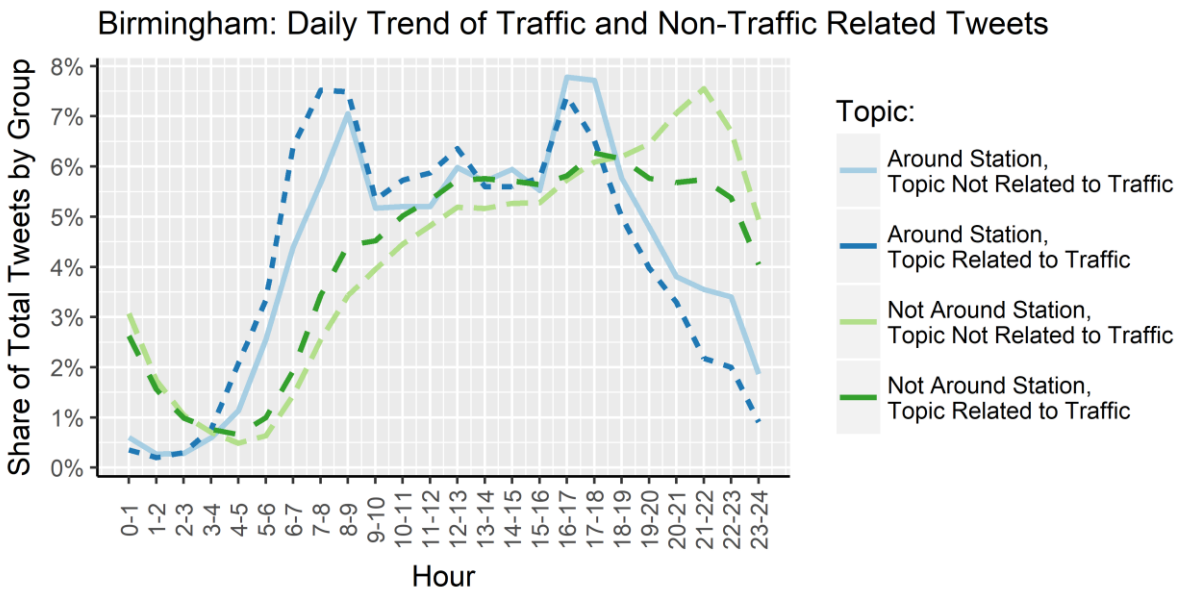


Abbildung 27: Durchschnittliche Häufigkeit von Tweets nach Uhrzeit, Verkehrs- oder Arbeitsbezug und räumliche Lage im Umfeld von Bahnhaltestellen in Birmingham

Anhand der Kurven wird deutlich, dass an Bahnhaltestellen zwei Tageshochs vorhanden sind, die sich mit den typischen Arbeitszeiten von Bürojobs zwischen 9 und 17 Uhr überschneiden. Die Kurven der Tweets im Bahnhofsumfeld, unabhängig vom semantischen Bezug, zeigen Höchststände insbesondere in den typischen Pendlerzeiten morgens und nachmittags sowie in den Stunden dazwischen. Nach dem Feierabend geht die relative Häufigkeit kontinuierlich zurück. Gleichzeitig steigt die Zahl der nicht verkehrs- oder arbeitsbezogenen Tweets außerhalb der Haltestellen abends immer weiter an. Außerhalb der Haltestellen sind Verkehrs- und Arbeitsthemen tagsüber relativ betrachtet häufiger vertreten und ab dem Feierabend weniger häufig als alle anderen Tweets im Durchschnitt.

Abbildung 28 differenziert die Daten nochmals nach Wochentagen (Montag bis Freitag) sowie Wochenenden (Samstag und Sonntag). Das grundsätzliche Muster bleibt gleich. Das Aufkommen an Tweets außerhalb von Bahnhaltestellen steigt in den Morgenstunden unter der Woche früher an. Unter der Woche abends zwischen 21 und 22 Uhr wird zudem der höchste Anteil an Tweets aus dieser Gruppe veröffentlicht. Am Wochenende sind Tweets außerhalb von Bahnhaltestellen dagegen ab dem Vormittag gleichmäßig verteilt bis in die Abendstunden. Tweets im Bahnhofsumfeld weisen unter der Woche zwei deutliche Peaks auf, die mit den täglichen Pendlerströmen zusammenpassen. Am Wochenende gibt es dagegen morgens einen breiteren Peak und dann mehrere kleine Peaks über den Tag verteilt.

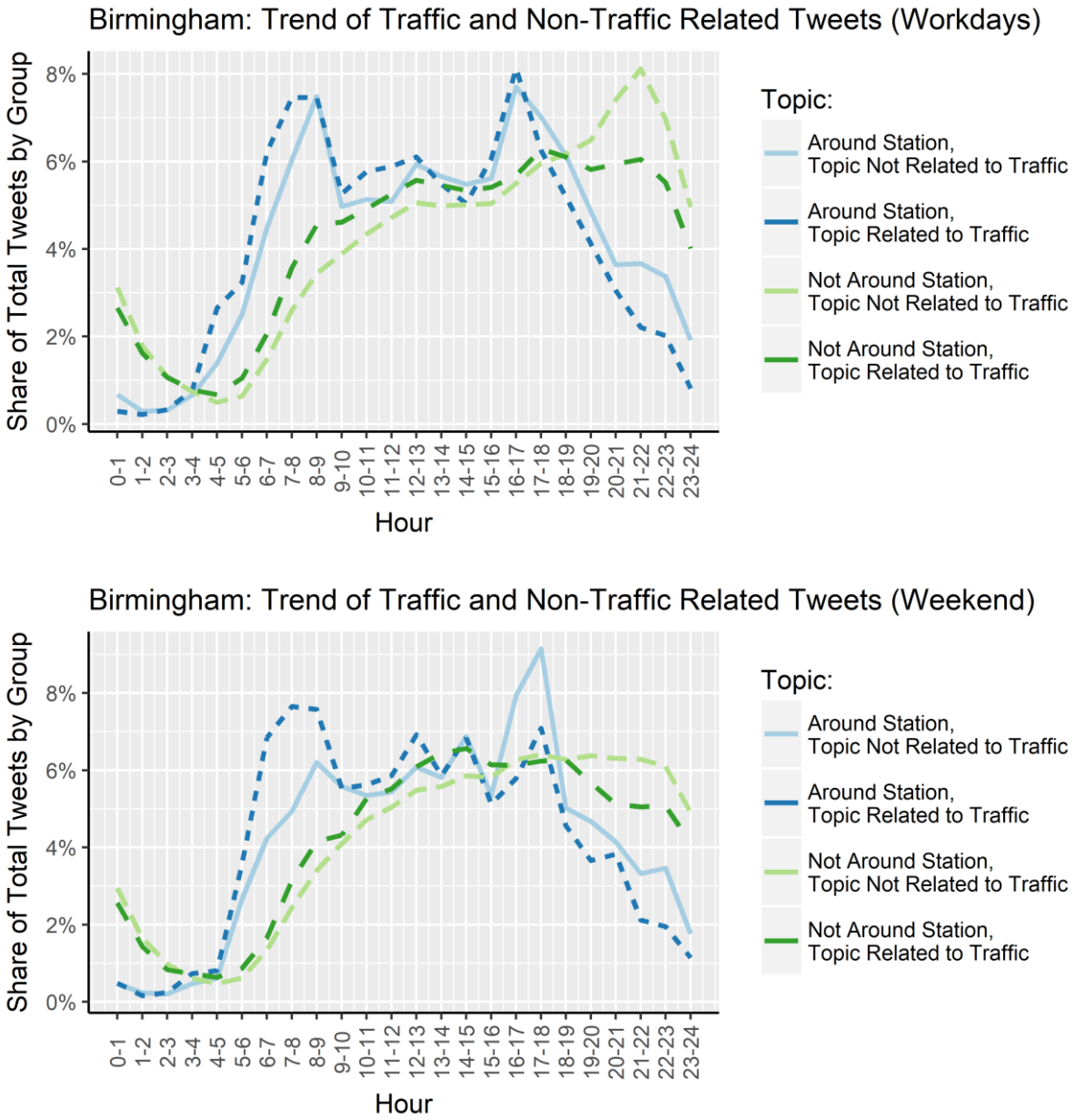


Abbildung 28: Durchschnittliche Häufigkeit von Tweets nach Uhrzeit, Verkehrs- oder Arbeitsbezug und räumliche Lage im Umfeld von Bahnhaltestellen differenziert nach Wochentagen und Wochenenden in Birmingham

Die Datenanalyse bestätigt den Zusammenhang zwischen Tweethäufigkeiten im räumlichen und zeitlichen Zusammenhang sowie menschlichen Mobilitätsmustern. Vergleichbare Muster sind auch in vorherigen Studien identifiziert worden. Steiger et al. (2016) zeigen beispielsweise vergleichbare Muster in London. Somit eignen sich Tweets neben der räumlichen Analyse auch zum Verständnis von zeitlichen Abläufen innerhalb eines Tages oder einer Woche in der Stadt.

3.4.2 Spieltage des Aston Villa Football Club

Die zweite Teilanalyse mit Betrachtung der zeitlichen Dimension basiert auf unterschiedlichen Teildatensätze, die wie folgt abgegrenzt wurden: Die Themen [19] und [20] korrelieren beide stark mit fußballspezifischen Begriffen, wobei das Thema [20] semantisch und räumlich deutlich mit dem Aston Villa Football Club verbunden ist (siehe Kapitel 3.2.3). Das Thema konzentriert sich aber auch räumlich um mehrere andere Stadien. Das Thema [19] hat dagegen einen stärkeren Bezug zu Sportveranstaltungen im Allgemeinen und soll hier insbesondere als Vergleichsdatensatz dienen.

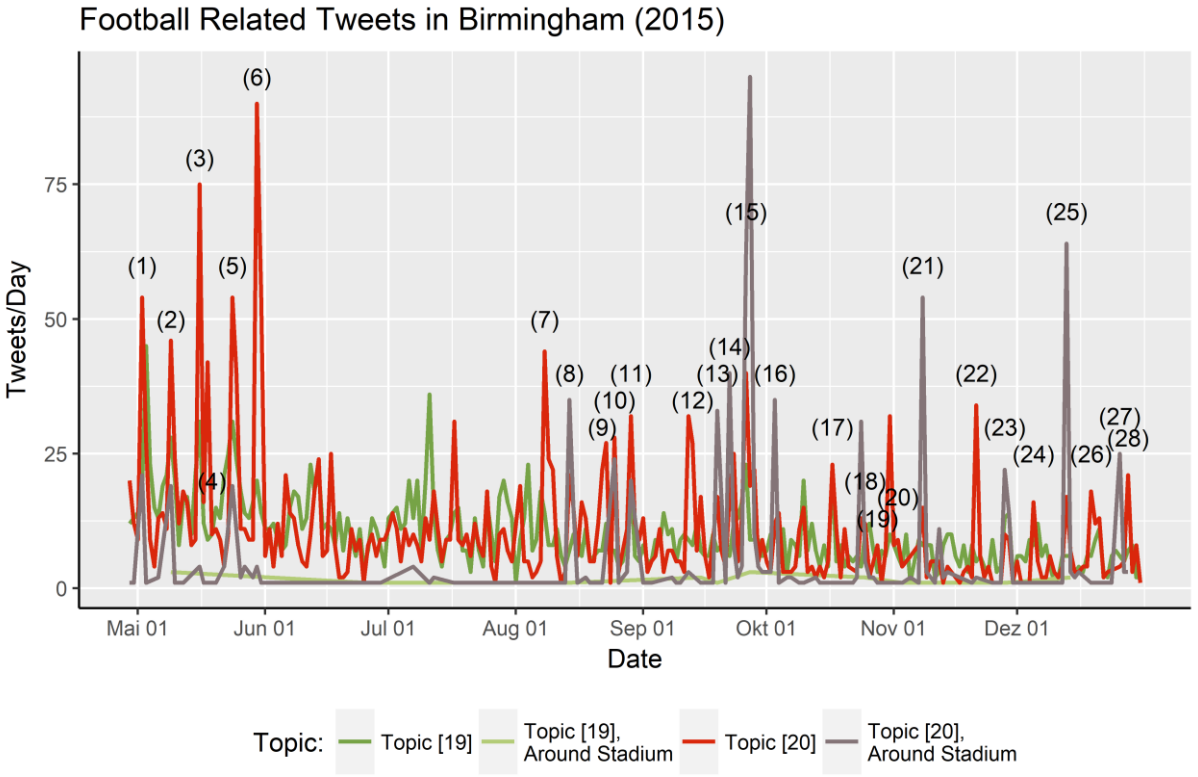
Um die Daten auf eine Korrelation mit Spieltagen des Aston Villa Football Club abzugleichen, wurden zwei Teildatensätze für die Themen [19] und [20] aus der Stichprobe herausgezogen. Beide Teildatensätze wurden wiederum unterteilt in Tweets, die in einem 500 m-Radius um das Aston Park Stadion verortet sind sowie alle anderen Tweets der Stichprobe mit diesen Themen.

Abbildung 29 spiegelt die Häufigkeit der Tweets pro Tag in den vier Datensätzen im betrachteten Zeitraum ab Ende April wieder. Deutlich zu erkennen sind einzelne Peaks, insbesondere beim Thema [20]. Diese Höchstpunkte treten sowohl im Stadionumfeld als auch im übrigen Datensatz auf und korrelieren mit den Heimspielen des Aston Villa FC. Tabelle 3 führt alle identifizierten Spiele im Zeitraum auf. Die dargestellten Nummern sind in Abbildung 29 den einzelnen Peaks zugeordnet.

Besonders auffällig sind einzelne Peaks unabhängig vom Stadionumfeld am Ende der Saison 2014/2015 im Mai 2015. Im Jahr 2015 hat der Aston Villa Football Club in der Premier League gespielt. Die Spiele (1), (2) und (5) waren Heimspiele und korrelieren mit Peaks bei der Linie der Tweets des Themas [20] im Stadionumfeld. Die Saison 2014/2015 endete am 30. Mai mit dem FA Cup Finale gegen Arsenal London (Nr. 6) und somit mit einem besonderen Saisonhöhepunkt. An diesem Datum wird das vorläufige Maximum der Tweets erreicht. Die hohe Anzahl an Tweets, nicht nur im Stadionumfeld, ist möglicherweise auf das besondere Saisonfinale zurückzuführen, bei dem erst am 16. Mai der Klassenerhalt sichergestellt werden konnte und am 30. Mai das beschriebene Finale gespielt wurde. Dieses hat dann vermutlich zu einem höheren Aufkommen an Tweets mit Bezug zum Fußballverein geführt. Zugleich fanden Halbfinale und Finale des FA Cups in London statt.

Neben den Peaks zeigt aber insbesondere auch der Verlauf der Linie für das Thema [20] im Stadionumfeld im Juni und Juli den Zusammenhang zwischen Spieltagen und Tweets. In diesem Zeitraum gibt es praktisch keine Tweets im Umfeld des Stadions. Das korreliert wiederum mit der Sommerpause zwischen den beiden Spielzeiten. Erst mit dem Saisonstart und dem ersten Heimspiel am 14. August (8) wird wieder eine Spitze bei den Tweets erreicht. Die Saison 2015/2016 begann am 8. August mit einem Auswärtsspiel in der Premier League. Die Spiele erfolgten dann wieder etwa im 1-Wochen-Rhythmus bis zum 28. Dezember. Zusätzlich gab es bis zum Oktober noch drei Spiele im Carabao Cup. Dieser Spielrhytmus wird auch durch die Peaks im Datensatz wiedergespiegelt.

Es gibt einzelne Ausnahmen mit besonders hohen Peaks in diesem Zeitraum, die nicht mit Spielansetzungen des Aston Villa FC korrelieren. Beispielsweise liegt der höchste Peak im gesamten Zeitraum am 27. September im Stadionumfeld vor. Ein Spiel des Aston Villa Football Club hat allerdings am 26. September in Liverpool stattgefunden (Peak Nr. 15). Durch weitere Recherchen stellte sich heraus, dass am 27. September im Villa Park Stadion ein Spiel der Rugby Weltmeisterschaft 2015 zwischen Australien und Uruguay stattgefunden hat.[1] Auch am 26. September fand hier ein Spiel zwischen Südafrika und Samoa statt. Somit erklärt sich die hohe Anzahl an Tweets im Stadionumfeld trotz Auswärtsspiel des Fußballvereins bei gleichzeitiger hoher Anzahl an Tweets zum Aston Villa FC am 26. September.



Zahlen in Klammern verweisen auf Spiele des Aston Villa Football Club (siehe Tabelle 3).

Abbildung 29: Auftreten ausgewählter Themen zum Aston Villa Football Club und Fußball in Birmingham (Ende April bis Ende Dezember 2015)

Tweets im Thema [20] weisen zudem einzelne Peaks auf, die nicht auf Spiele zurückgeführt werden können. Vermutlich handelt es sich hierbei um besondere Ereignisse, die möglicherweise in den Medien diskutiert wurden. Diese Ereignisse konnten im Einzelnen nicht nachrecherchiert werden. Das Thema [19] zeigt weiterhin einzelne parallele Peaks. Hier schlagen sich vermutlich auch einige Tweets mit Fußballbezug an den Spieltagen nieder. Gleichzeitig gibt es aber, bis auf Einzelfälle, keine

[1] Online unter: https://www.rugbyworldcup.com/2015 (letzter Abruf: .28.05.2018).

Tweets im Stadionumfeld selber aus diesem Thema. Daraus kann gefolgert werden, dass mit dem Thema [20] ein guter Indikator für Spieltage und Fluktuation von Menschen im Stadionumfeld gefunden wurde. Der aufgezeigte Zusammenhang zwischen dem Saisonverlauf des Aston Villa Football Club und der Zahl der Tweets im Thema [20], zum Teil auch Thema [19], verdeutlicht, wie mittels Tweets und semantischer Klassifizierung neben den räumlichen Zusammenhängen auch zeitliche Zusammenhänge berücksichtigt und analysiert werden können. Die Tweets im Stadionumfeld ermöglichen Rückschlüsse auf lokale Events, auch wenn diese nicht immer eindeutig dem Fußballverein zugeordnet werden können.

Tabelle 3: Spiele des Aston Villa Football Club und korrelierende Peaks bei der Anzahl der Tweets in ausgewählten Themen in Birmingham (2015)

Nr.	Datum	Spiel	Austragungsort
1	02.05.2015	**Aston Villa** 3:2 FC Everton (Premier League)	Villa Park
2	09.05.2015	**Aston Villa** 1:0 West Ham United (Premier League)	Villa Park
3	16.05.2015	FC Southhampton 0:1 **Aston Villa** (Premier League)	Auswärts
4	19.05.2015	**Aston Villa** 2:1 FC Liverpool (FA Cup)	Wembley Stadion*
5	24.05.2015	**Aston Villa** 0:1 FC Burley (Premier League)	Villa Park
6	30.05.2015	FC Arsenal 4:0 **Aston Villa** (FA Cup Finale)	Wembley Stadion*
7	08.08.2015	AFC Bournemouth 0:1 **Aston Villa** (Premier League)	Auswärts
8	14.08.2015	**Aston Villa** 0:1 Manchester United (Premier League)	Villa Park
9	22.08.2015	Crystal Palace 2:1 **Aston Villa** (Premier League)	Auswärts
10	25.08.2015	**Aston Villa** 5:3 Notts County (Carabao Cup)	Villa Park
11	29.08.2015	**Aston Villa** 2:2 AFC Sunderland (Premier League)	Villa Park
12	13.09.2015	Leicester City 3:2 **Aston Villa** (Premier League)	Auswärts
13	19.09.2015	**Aston Villa** 0:1 West Bromwich Albion (Premier League)	Villa Park
14	22.09.2015	**Aston Villa** 1:0 Birmingham (Carabao Cup)	Villa Park
15	26.09.2015	FC Liverpool 3:2 **Aston Villa** (Premier League)	Auswärts
16	03.10.2015	**Aston Villa** 0:1 Stoke City (Premier League)	Villa Park
17	17.10.2015	FC Chelsea 2:0 **Aston Villa** (Premier League)	Auswärts
18	25.10.2015	**Aston Villa** 1:2 Swansea City (Premier League)	Villa Park
19	28.10.2015	FC Southhampton 2:1 **Aston Villa** (Carabao Cup)	Auswärts
20	02.11.2015	Tottenham Hotspur 3:1 **Aston Villa** (Premier League)	Auswärts
21	08.11.2015	**Aston Villa** 0:0 Manchester City (Premier League)	Villa Park
22	21.11.2015	FC Everton 4:0 **Aston Villa** (Premier League)	Auswärts
23	28.11.2015	**Aston Villa** 2:3 FC Watford (Premier League)	Villa Park
24	05.12.2015	FC Southhampton 1:1 **Aston Villa** (Premier League)	Auswärts
25	13.12.2015	**Aston Villa** 0:2 FC Arsenal (Premier League)	Villa Park
26	19.12.2015	Newcastle United 1:1 **Aston Villa** (Premier League)	Auswärts
27	26.12.2015	**Aston Villa** 1:1 West Ham United (Premier League)	Villa Park
28	28.12.2015	Norwich City 2:0 **Aston Villa** (Premier League)	Auswärts

* Die Spiele des FA Cup finden ab dem Halbfinale im Wembley Stadion statt in London

Quelle: eigene Recherche (Datenbasis: fussballdaten.de, flashscores.co.uk)

4 Diskussion

In diesem Kapitel werden die Ergebnisse im Zusammenhang mit den eingangs gestellten Fragen diskutiert und vor dem Hintergrund aktueller Herausforderungen in der Stadtentwicklung und Stadtplanung reflektiert. Zunächst geht es um die Diskussion der verwendeten Methoden, insbesondere der Latent Dirichlet Allocation, und der Analyse der daraus resultierenden Themen (Kapitel 4.1). Die Ergebnisse der räumlichen Analyse auf den unterschiedlichen Betrachtungsebenen und in Verknüpfung mit Daten aus OpenStreetMap sowie in einer kombinierten raum-zeitlichen Betrachtung werden in einem eigenen Kapitel diskutiert (Kapitel 4.2). Abschließend erfolgt die Diskussion der Ergebnisse mit dem Fokus auf Stadtentwicklung und Stadtplanung (Kapitel 4.3).

4.1 Qualität der identifizierten Themen

Mit der ersten Forschungsfrage wurde nach einer Methode gefragt, mit der Beiträge zur Beschreibung städtischer Räume aus geocodierten Tweets extrahiert und semantisch klassifiziert werden können. Die Klassifizierung von Dokumenten mit dem Latent Dirichlet Allocation Algorithmus wurde als geeignete Methode identifiziert und angewendet. Die Implementierung der Methode in den dargestellten Forschungsansatz resultiert sowohl in der Identifikation von plausiblen Themen mit Bezug zur Stadt, geht aber auch einher mit einigen nicht oder nur zum Teil interpretierbaren Ergebnissen.

Zwischen den zwei verwendeten Datensätzen, einem kleineren Korpus englischsprachiger Tweets aus der Region Manchester sowie einem deutlich größeren Korpus englischsprachiger Tweets aus Birmingham, zeigen sich im Ergebnis Unterschiede hinsichtlich der Qualität der resultierenden Themen. Im Manchester-Datensatz konnten anteilig mehr Themen identifiziert werden, die inhaltlich ohne Vorkenntnisse der Daten und nur anhand der wichtigsten Wörter interpretiert werden können. In der räumlichen Analyse wurden 17 von 32 identifizierten Themen aus der Region Manchester hinsichtlich der räumlichen Verteilung im Untersuchungsraum beschrieben. Für Birmingham erfolgte dagegen nur die detaillierte Betrachtung von 11 Themen aus insgesamt 30 identifizierten Themen.

Für den kleineren Korpus konnten deutlich mehr Iterationen mit dem Algorithmus durchgeführt werden als mit dem größeren Birmingham-Datensatz. Dies hat möglicherweise zu einer besseren Abbildung der latenten Themen in Manchester geführt. Bei der Verwendung der Methodik ist die Größe des Datensatzes sowie die verwendete Hardware ein wichtiger Limitierungsfaktor. Aber auch die Software kann hier ggf. eine Rolle spielen. Im Vorfeld der Master Thesis wurde jedoch kein Vergleich der unterschiedlichen Anwendungen durchgeführt. Recherchen deuten aber darauf hin,

dass Gensim in Python effizienter in der Verarbeitung größerer Datensätze gegenüber dem topicmodel Paket in R ist.

Fasst man die resultierenden Themen zu Themenfeldern zusammen, dann ergeben sich in etwa solche Gruppen, wie die von Zhou and Zhang (2016) gebildeten Kategorien (Bildung, Reisen und Transport, Draußen und Erholung, Essen und Restaurant, Geschäft und Dienstleistung sowie Nachtleben) oder die von Jenkins et al. (2016) anhand eines Abgleichs zwischen Tweets und Wikipedia-Artikeln in mehreren Städten identifizierten Themen (Politik, Wirtschaft, Bildung, Erholung, Sport und Unterhaltung). Bildung im engeren Sinne spielt allerdings in hier identifizierten Themen keine Rolle. Lediglich Universitäten erscheinen als Standorte mit deutlichen Hot Spots für unterschiedliche Themen. Die Themenfelder Unterhaltung (Musik, Film) und Konsum (Einkaufen, Essen, Trinken) nehmen dagegen den höchsten Stellenwert ein. In den hier untersuchten Datensätzen sind solche Freizeithemen dominant. Politik erscheint lediglich im Birmingham-Datensatz in einem Thema und spielt sonst bei den räumlichen Daten eine untergeordnete Rolle.

Zusätzlich ergeben sich viele Überschneidungen zwischen den identifizierten Themen, die mit der angewandten Klassifizierungsmethode nicht eindeutig abgegrenzt werden können. Das deutlichste Beispiel dafür ist das Etihad Stadion in Manchester, in dem sowohl Fußballspiele als gelegentlich auch Konzerte stattfinden. Beide Arten von Ereignissen finden sich im gleichen Thema in der Analyse wieder. Zugleich werden Tweets zu anderen Musikveranstaltungen im übrigen Stadtgebiet von Manchester ebenfalls diesem Thema zugeordnet. Ein anderes Beispiel ist das Heaton Park Festival, das gemeinsam mit anderen Tweets mit Bezug zum Wort Park in unterschiedlichen Zusammenhängen sowie Beiträgen zu anderen Konzerten in einem Thema zusammengefasst wird. Es fehlt bei diesen Beispielen an Trennschärfe zwischen den Themen, da gleiche Wörter in anderen Kontexten verwendet werden und zugleich diese Wörter im Datensatz besonders häufig vorkommen. Beim Beispiel der Stadtnamen Manchester und Birmingham sowie anderer Ortsnamen (Stadtteile, Stadien, Gebäude, etc.) tritt dieser Fehler häufiger auf. Solche Eigennamen werden in einer Vielzahl von Kontexten genutzt, wie zum Beispiel Manchester City, Manchester United, Greater Manchester, Manchester Music Hall, usw. Der Algorithmus kann bei der Kürze der analysierten Tweets und der damit verbundenen geringen Anzahl an weiteren Kontextwörtern innerhalb eines Tweets nicht hinreichend gut klassifizieren und gleichzeitig auch differenzieren zwischen den verschiedenen semantischen Zusammenhängen, in denen Ortsnamen verwendet werden.

Die Reihenfolge der Wörter spielt bei der verwendeten Methode keine Rolle. Nur das Vorhandensein von Wörtern in einem Dokument ist entscheidend. Somit kann ein Wort, das unterschiedliche semantische Bedeutungen in seinem jeweiligen Kontext hat, nicht hinreichend genau den einzelnen Themen zugeordnet werden. Eine Lösung für dieses Problem wäre die Berücksichtigung von

Begriffen aus zwei Wörtern als ein Token in der Analyse. Somit wird das gemeinsame Auftreten von bestimmten Wörtern nicht nur implizit berücksichtigt. Dieser Ansatz wurde jedoch im Rahmen der vorliegenden Master Thesis nicht erprobt. Andere Studien zeigen, dass dies ein guter Ansatz zur Verbesserung der Qualität der Ergebnisse ist (Cheng et al., 2014), der weiter verfolgt werden sollte.

Blei (2012) argumentiert, dass die Themen, die durch den Algorithmus identifiziert werden, Beobachtungen der statistischen Struktur einer Sprache und ihrer Interaktion mit dem statistischen Ansatz der LDA darstellen. Er weist aber auch darauf hin, dass es ein Problem beim Überprüfen der Modellergebnisse gibt. Aus seiner Sicht ist zu prüfen, wie die unterschiedlichen Annahmen und die Wahl des Modells im Einzelnen begründet werden können. Im Kontext der gestellten Forschungsfragen in dieser Arbeit kann eine Antwort auf die Frage nach der Überprüfung die räumliche Analyse der Themen und die Verknüpfung mit weiteren räumlichen Datenquellen sein. Die Untersuchung der geocodierten Tweets bietet einen Zugang, die identifizierten Themen in ihrer räumlichen Verteilung zu untersuchen und Zusammenhänge mit anderen Informationen über die reale Welt abzugleichen. Dadurch werden Themen in einen räumlichen Kontext gesetzt, der eine Überprüfung anhand des Wissens über die reale Welt ermöglicht. Wenn die identifizierten Themen in einem Zusammenhang mit realen Orten stehen, dann bestätigt dies die latenten, zuvor nicht sichtbaren, semantischen Muster im Korpus.

4.2 Ergebnisse der räumlichen und raum-zeitlichen Analysen

Die zweite Forschungsfrage zielt darauf ab, einen Vergleich von Themen und räumlichen Mustern durchzuführen. Eine räumliche Konzentration eines Themas stellt ein Hinweis auf zugrunde liegende städtische Strukturen oder Prozesse dar. Um das zu prüfen, erfolgte ein Abgleich von räumlichen Clustern der inhaltlich interpretierbaren und somit nachvollziehbaren Themen in Manchester und Birmingham mit der jeweiligen räumlichen Situation.

Die Datenanalyse zeigt, dass eine Identifikation von räumlich-semantischen Clustern in einer Stadt möglich ist. Die in dieser Studie identifizierten Hot Spots in Manchester und Birmingham korrelieren zu einem großen Anteil mit den Standorten bestimmter Gebäude, Infrastrukturen oder frequentierter Stadtteile und somit mit typischen Anziehungspunkten oder Bereichen mit hohen Passantenfrequenzen in der Stadt (z.B. Fußballstadien, Konzerthallen, Universitäten, Stadtzentren, Parkanlagen, Einkaufszentren oder Bahnhöfe). Durch den Abgleich der räumlichen Cluster der identifizierten Themen mit weiteren Datengrundlagen sollte dann eine Möglichkeit überprüft werden, die Ergebnisse zu validieren. Hierzu wurden Daten von OpenStreetMap (OSM) verwendet, bei denen es sich um explizite Volunteered Geographic Information (VGI) handelt. Denn während georeferenzierte Tweets implizite geographische Daten darstellen, die mit einer anderen Intention

erstellt wurden, stammen die Daten aus OSM von Nutzern, die Geodaten zur Repräsentation geographischer Features aktiv erstellt haben.

Wie auch schon viele Studien vorher gezeigt haben (siehe Kapitel 1.2), eignen sich Daten aus sozialen Medien zur Analyse unterschiedlicher räumlicher Fragestellungen im Kontext von Städten. Die Analysen in dieser Master Thesis zeigen ebenfalls, dass Prozesse und Strukturen (Crooks et al., 2015) durch die Daten sichtbar gemacht werden können. Orte mit hohen Passantenaufkommen weisen zugleich höhere Dichten an Tweets auf. Die Nutzer der sozialen Medien hinterlassen als *social sensor* (Liu et al., 2015) Spuren in Form von räumlich verortbaren Beiträgen, die wiederum mit den räumlichen Gegebenheiten in der realen Welt korrelieren. Dies wird über die verwendeten Hexagon-Zellen auch unabhängig von Stadtteilen oder anderen statistischen, räumlichen Einheiten, sichtbar. Somit kann für die zweite Forschungsfrage nach der Verortung von städtischen Strukturen und Prozessen festgestellt werden, dass räumliche Cluster spezifischen räumlichen Strukturen und Prozessen zugeordnet werden können und das dabei insbesondere auch Zusammenhänge mit den semantisch klassifizierten Themen hervortreten.

Es gibt aber auch Einschränkungen, die in den Ergebnissen sichtbar werden. In Kapitel 1.1 wurden die Unterschiede zwischen klassischen, in der Regel amtlichen Daten und den neuen Datenquellen, wie Volunteered Geographic Information (VGI) und Daten aus sozialen Medien, diskutiert. Amtliche Daten sind einer Qualitätskontrolle unterstellt, die bei den neuen Datengrundlagen in vergleichbarer Form nicht existiert. Senaratne et al. (2016) nennen dafür die folgenden Gründe: Die Daten stammen von heterogenen Nutzern, die unterschiedliche Technologien verwenden mit unterschiedlicher Genauigkeit und sie verfolgen verschiedene Zwecke. Tweets sind ebenso von diesen Ungenauigkeiten betroffen. Das zeigen auch die Ergebnisse anderer Studien (Craglia et al., 2012).

Für beide Datengrundlagen – Tweets und OSM – stellt sich die Frage der Qualitätskontrolle im Hinblick auf die Eignung für die räumliche Analyse im Kontext von Stadtentwicklung und Stadtplanung. Diese Diskussion um die Qualität der geographischen Informationen wird vor allem im Zusammenhang mit der Validität von crowdbasierten Datenbanken wie OSM geführt. Die Diskussion ist aber auch im Hinblick auf implizite geographische Informationen aus sozialen Medien wichtig und sollte darum im Kontext expliziter als auch impliziter Daten geführt werden. Goodchild and Li (2012) verweisen in diesem Zusammenhang auf den geographischen Ansatz, der auf dem ersten Gesetz der Geographie von Tobler basiert: Eine mögliche Information über einen Ort solle sich mit anderen Informationen über den Ort selbst und dessen Nachbarschaft decken. Ein solcher Abgleich sei erforderlich, um die Validität der Daten zu beurteilen.

Für einen einzeln betrachteten, räumlich verorteten Tweet kann der geographische Ansatz leicht widerlegt werden. Das Thema eines zufällig ausgewählten Tweets muss sich nicht mit dem Ort und

seiner Nachbarschaft decken. Eine Äußerung auf Twitter kann sich auf ein bestimmtes Thema beziehen, dass zumindest vordergründig nicht inhaltlich mit dem Ort verknüpft ist. Dennoch wurde der Tweet mit einer impliziten, räumlichen Information veröffentlicht (z.B. über ein Smartphone mit GPS, das den Standort des Nutzers zum Zeitpunkt der Veröffentlichung gespeichert hat). Auf diese Limitierung hinsichtlich des tatsächlichen Standortes weisen auch Steiger et al. (2016) hin. Für die Analyse von Tweets ist die Validität im Sinne des geographischen Ansatzes somit erst durch statistische Analyse einer ausreichend großen Anzahl von Datenpunkten nachweisbar. Erst dann ist die kombinierte semantische und räumliche Analyse der Tweets möglich.

Doch auch in ausreichend großen Datensätzen können topologische Ausreißer zu Fehlinterpretationen führen. Westerholt et al. (2016) zeigen, dass sich die mittels Tweets gemessenen Phänomene im städtischen Raum überlagern und zugrunde liegende Cluster somit unentdeckt bleiben können. Dadurch können Fehlschlüsse über tatsächlich vorhandene räumliche Muster entstehen. Die Autoren schließen daraus, dass es sich bei sozialen Medien um einen eigenen, räumlichen Prozess handelt, der eben nicht durch soziale Interaktion in der realen Welt entsteht. Dadurch werden die räumlichen Muster, die eigentlich untersucht werden sollen, überlagert oder verdeckt. Nicht berücksichtigt wurde beispielsweise in dieser Master Thesis die Auswirkung von Bots auf die Ergebnisse. Nur offensichtliche Bots, beispielsweise mit einer deutlich überdurchschnittlichen Anzahl ähnlicher oder gleicher Beiträge, konnten aus dem Datensatz entfernt werden. Es ist aber anzunehmen, dass Bots auf Grund der hohen Anzahl an Beiträgen (Haustein et al., 2016) einen Einfluss auf die Ergebnisse semantischer und räumlicher Analysen nehmen können.

Ein weiterer Aspekt, der zu Fehlschlüssen führen kann, ist die unterschiedliche Dichte an Ereignissen im Untersuchungsgebiet. Besonders hohe Dichten von Datenpunkten bzw. Ereignissen treten erwartungsgemäß im Zentrum von Manchester bzw. Birmingham auf. Hier sind flächendeckend hohe Dichten vorzufinden. Insbesondere am Beispiel Birmingham zeigt sich, dass die Konzentration von Publikumsmagneten, wie den dort verorteten Einkaufszentren, zu einer sehr hohen Konzentration von Tweets führen. Auch bei Themen, die nicht unmittelbar mit diesen Einkaufswelten verknüpft sind, bilden sich räumliche Konzentrationen alleine durch die hohe Anzahl an Tweets am selben Ort. Im Bereich des Grand Central in Birmingham mit dem darunter liegenden Bahnhof wird die mit Abstand höchste Dichte an Tweets erreicht und daraus resultiert in nahezu allen Themen ein Hot Spot. Auf Basis der vorliegenden Daten lässt sich jedoch kein Rückschluss auf die tatsächlichen Besucherzahlen ziehen. Dazu wäre im Rahmen weiterer Studien zu prüfen, wie hoch die Korrelation zwischen tatsächlichen Besuchern und Tweets oder anderen Beiträgen in sozialen Medien ist.

Weitere räumliche Ankerpunkte mit einer hohen Anzahl an Besuchern, wie Fußballstadien, Konzerthallen, Bahnhöfe, Einkaufszentren, Gebäudekomplexe wie die MediaCity UK sowie temporäre

Ereignisse wie das Heaton Park Festival zeigen ebenfalls häufiger Hot Spots, die aber insbesondere bei gleichzeitiger isolierter Lage besser erklärbar sind. Diese Gebiete mit hohen Ereignisdichten treten punktuell auf, in der durchgeführten Analyse im Rahmen von ein bis vier Hexagonzellen bzw. im Umkreis von wenigen Hundert Metern. Einige dieser Anlaufpunkte sind nur temporär (Festival, Fußballspiel, Konzert) oder sind insbesondere für bestimmte Zielgruppen relevant (z.B. Konzerte oder Festivals mit einem hohen Anteil Besucher von außerhalb der Stadt oder der Region). Für weitere Analysen könnte ein Ansatz sein, die Tweets und ihre Nutzer zunächst nach bestimmten Merkmalen zu klassifizieren wie etwa dem Wohnort oder der Altersklasse der Nutzer. Longley and Adnan (2016) oder Garcia-Palomares et al. (2015) zeigen, dass dies anhand von Nutzern mit mehreren georeferenzierten Tweets oder Daten aus anderen sozialen Medien (z.B. Flickr) innerhalb eines Datensatzes möglich ist. Daraus ließe sich z.B. eine Differenzierung von „Alltagsaktivitäten" der Bewohner einer Stadt oder Region von „Besucheraktivitäten" der auswärtigen Gäste ableiten. Dadurch könnten die jeweiligen Ankerpunkte besser identifiziert und differenziert werden. Im Hinblick auf Fragestellungen, die auf Personen, die in einer Stadt wohnen, abzielen, wäre das ein wichtiges Kriterium für die Anwendung solcher Analysen im Bereich der Stadtplanung.

Die weitere Überprüfung der räumlichen Analyse der identifizierten Tweets erfolgte schließlich anhand von raum-zeitlichen Mustern. Die dritte Forschungsfrage zielt darauf ab, den Zusammenhang zwischen den identifizierten Themen mit lokalen Ereignissen herzustellen, die nicht nur räumlich verortet werden können, sondern auch spezifische zeitliche Muster aufweisen. Wenn ausgewählte Themen in einem räumlichen Zusammenhang mit bestimmten Orten stehen, dann müsste sich dieser Zusammenhang anhand unterschiedlicher Häufigkeiten im Zeitverlauf, z.B. innerhalb von Tagen oder Stunden, wiederspiegeln. Diese zeitlichen Muster sollten wiederum Ereignisse oder alltägliche Bewegungsmuster der Nutzer wiederspiegeln. Um das zu überprüfen wurden solche Themen und Orte ausgewählt, für die ein eindeutiger Zusammenhang hinsichtlich zeitlicher Muster angenommen werden kann und für die sowohl ausreichend Tweets im Umfeld als auch Themen mit einem inhaltlichen Bezug vorliegen: Bahnhaltestellen und Fußballstadien.

Die Analyse bestätigt den Zusammenhang zwischen semantisch klassifizierten Tweets, ausgewählten Orten und den vermuteten zeitlichen Mustern. Die dritte Forschungsfrage kann somit für die beiden gewählten Beispiele bestätigt werden. Auch Steiger et al. (2015b) zeigen für Bahnstationen vergleichbare raum-zeitliche Muster von Aktivitäten. Dadurch wird klar, dass Tweets nicht nur räumliche Muster wiederspiegeln, sondern auch zeitlich begrenzte Events abbilden können, indem sie kollektive Mobilitätsmuster der Nutzer sichtbar machen. Laut Steiger et al. (2016) können individuelle Ereignisse, wie Konzerte, Demonstrationen oder Sportveranstaltungen, mittels Twitter-Daten verortet werden. Die dargestellten Ergebnisse in dieser Master Thesis bestätigen das auch.

Eine Zielsetzung für zukünftige Untersuchungen könnte die stärkere Verschränkung der unterschiedlichen VGI Ressourcen sowie die Überschneidung mit anderen Daten darstellen (Neis and Zielstra, 2014). Der Vergleich zwischen OSM-Daten und den identifizierten semantisch-räumlichen Clustern der Tweets zeigt, dass es hier sowohl Bereiche mit eindeutigen Überschneidungen gibt als auch Bereiche, in denen eine hohe Konzentration von Tweets keine Korrelation mit korrespondierenden Angeboten, wie z.B. Bars oder Nachtclubs, aufweist. Auch Standorte aus den OSM-Daten korrelieren nicht immer mit semantischen Clustern der Tweets. Die Gründe dafür können sowohl auf semantisch falsche Klassifizierungen der Tweets, fehlende Daten zu Angeboten als auch unterschiedliche, räumliche Prozesse in diesen Bereichen zurückzuführen sein. Für die Identifikation charakteristischer Nutzungen in unterschiedlichen Stadtteilen hilft es darum, durch die Kombination verschiedener Datenquellen Schwachstellen in den einzelnen Datengrundlagen zu identifizieren. Dazu müssen aber auch die offenen Fragen der räumlichen Genauigkeit dieser Datengrundlagen geklärt werden (Elwood et al., 2012).

Trotz der statistischen Artefakte, und ohne dass alle räumlichen Muster eindeutig erklärt werden können, zeigt die Analyse, dass unterschiedliche Räume einer Stadt durch Themen aus Tweets charakterisiert werden können. Das Bild der kollektiven Wahrnehmung eines Raumes durch soziale Medien, wie es Jenkins et al. (2016) beschreiben, ist in diesem Zusammenhang plausibel trotz der dargestellten Einschränkungen. Mittels der verwendeten quantitativen Analysemethoden kann die kollektive Wahrnehmung des Raumes untersucht werden. Eine Charakterisierung städtischer Räume mittels extrahierter Themen aus sozialen Medien ist also möglich. Die zentrale These der Master Thesis ist damit unter Beachtung der dargestellten Einschränkungen zu bestätigen.

4.3 Perspektiven für Stadtentwicklung und Stadtplanung

Abschließend stellt sich die Frage, wie diese kollektive Wahrnehmung des städtischen Raumes in Form digitaler Daten weiter genutzt werden kann. Dafür erfolgt abschließend eine Diskussion der Ergebnisse im Hinblick auf unterschiedliche Anwendungsfelder, insbesondere aktueller Fragen der Stadtentwicklung und Stadtplanung. Die Zielsetzung dieser Master Thesis war die Identifizierung von städtischen Räumen mit einer hohen Frequentierung durch Menschen und die Ableitung von charakteristischen Nutzungsmustern. Die Analyse zeigt, dass mittels Extrahierung von Themen aus georeferenzierten Tweets eine solche Charakterisierung möglich ist.

Grundsätzlich stellen Tweets eine geeignete Datengrundlage für die Analyse von Fragestellungen innerhalb der Stadtforschung dar. Dies wird in einer Vielzahl aktueller Studien gezeigt. Crooks et al. (2015) gehen sogar deutlich darüber hinaus und sehen hierin ein neues Paradigma der Sammlung, Analyse und Modellierung der urbanen Morphologie auf Basis von crowdbasierten Daten. Die

Autoren bezeichnen die Nutzer der sozialen Medien und damit die Datenerzeuger als hybride, soziokulturelle Sensoren. Die *smart city* werde somit nicht erst durch die Installation von Sensoren ermöglicht, sondern die Stadt selber sei bereits ein solches intelligentes System. Das System könne wiederum abgebildet und untersucht werden durch explizite Daten (z.B. OpenStreetMap) und implizite Daten (z.B. Tweets). Dies sei wiederum Grundlage für eine neue Wissenschaft der Städte.

Laut Kitchin (2014) geht es bei der Analyse vor allem um die automatisierte Nutzung dieser Daten zur Beobachtung des städtischen Lebens und der städtischen Infrastruktur, um beispielsweise aktuelle Trends anhand der Tweets sichtbar zu machen. Ein wesentlicher Bestandteil solcher Analysekonzepte ist die Verwendung von Big Data zur Beobachtung von Prozessen in der Stadt. Dies umfasst neben einer Vielzahl an weiterer Datenrundlagen eben auch die Interaktionen von Nutzern in sozialen Medien (Kitchin, 2014). Auswertungen dieser Daten liefern wichtige Indikatoren zur Beantwortung von räumlichen Fragestellungen. Wie beispielsweise Shelton et al. (2015) anhand der Analyse von räumlicher Segregation zeigen, können Daten aus sozialen Median auch beim Verständnis räumlicher Prozesse helfen. Die Analysen sollten aber stets in einen größeren Kontext stattfinden und in parallele Betrachtungen auf Basis anderer Grundlagen eingebunden sein. In vielen Fällen könnten Volunteered Geographic Information oder soziale Medien komplementär oder sogar integriert mit amtlichen Daten zusammen geführt werden, um Forschungsfragen zu beantworten.

Für Planungsprozesse in der Stadt ist es vorteilhaft, Zugang zu Wissen von Nicht-Experten zu erhalten. Nicht-Experten können manchmal sogar zu besseren Lösungen für spezifische Problemstellungen gelangen, als Experten, wie Brabham (2009) herausarbeitet. Der Autor stellt dem aber auch die negativen Effekte von öffentlicher Beteiligung, z.B. durch Zugangsbarrieren, fehlende Transparenz und fehlende Einbindung von lokalem Wissen, gegenüber. Ziel der Beteiligung soll es darum sein, lokales Wissen der Bürger in die Entscheidungsprozesse einfließen zu lassen und damit Entscheidungen zu legitimieren (Tenney and Sieber, 2016).

Im Kontext von Volunteered Geographic Information stellt auch Goodchild (2007) fest, dass die Frage des Teilnehmerkreises einen wesentlichen Einfluss auf die Qualität der resultierenden Informationen hat. Die Frage ist also, ob durch digitale Beteiligungsprozesse und ganz spezifisch im Zusammenhang dieser Master Thesis durch den Zugang zu neuen Datengrundlagen aus sozialen Medien auch ein Zugang zu alternativen bzw. zusätzlichen Informationen möglich ist. Die empirische Analyse in dieser Forschungsarbeit zeigt, dass auf Basis von Tweets räumlich verankerte Themen extrahiert werden können. Nicht beantwortet werden kann jedoch auf Grundlage der Analysen, inwiefern Ausschlussmechanismen dazu führen, dass bestimmte Gruppen und Themen nicht berücksichtigt werden. Dies ist wiederum Gegenstand weiterer Forschungsarbeiten. Dabei stehen die selektiven Mechanismen von Algorithmen sowie die Zugangsbarrieren zu digitalen Inhalten für bestimmte

Bevölkerungsgruppen im Vordergrund (Tenney and Sieber, 2016, O'Brien et al., 2016, Craglia and Shanley, 2015).

Aus den vorliegenden Studien geht hervor, dass es in den sozialen Medien verschiedene Ausschlussmechanismen gibt. Soziale Netzwerke werden nicht gleichermaßen von allen Bevölkerungsgruppen genutzt bzw. sind nicht allen Gruppen gleichermaßen zugänglich. Nicht zu unterschätzen ist dabei die demographische und die soziale Komponente der sozialen Netzwerke. Auch wenn alle Altersklassen und Bevölkerungsschichten in sozialen Medien vertreten sind, gibt es Unterschiede hinsichtlich der Repräsentativität einzelner Gruppen (Longley and Adnan, 2016). Bestimmte Bevölkerungsgruppen und Altersklassen sind überdurchschnittlich aktiv. Das führt zu einer Überrepräsentation dieser Gruppen, beispielsweise besser verdienender und akademischer Haushalte gegenüber anderen sozialen Schichten (Li et al., 2013). Stadtplanung und öffentliche Beteiligungsprozesse sind grundsätzlich aber immer darauf ausgerichtet, alle betroffenen Bevölkerungsschichten einzubinden. Somit können Daten aus sozialen Medien wiederum nur eine Informationsquelle von mehreren darstellen. Aus den Ergebnissen dieser Master Thesis wird zudem ersichtlich, dass das vorliegende Themenspektrum nur Teilaspekte der Strukturen und Prozesse im städtischen Raum abbilden kann. Dies spricht dafür, soziale Medien als zusätzliche, ergänzende Datengrundlage zu betrachten, keinesfalls aber als neues Dogma der Stadtentwicklung.

Die Analysen zeigen auch weitere Limitierungen der Daten und Methoden. Wie in dieser Arbeit dargestellt wurde, sind Tweets unbereinigt zunächst kaum zu analysieren und enthalten viele Informationen, die im Kontext der hier bearbeiteten Fragestellung nicht auswertbar sind. Zudem gibt es eine hohe Anzahl von Beiträgen, die von Bots stammen und zunächst herausgefiltert werden müssen, um Verzerrungen zu vermeiden. Hier gibt es aber auch eine Dunkelziffer, da nicht bei allen Inhalten ersichtlich ist, ob es sich um einen menschlichen Urheber oder einen computergenerierten Inhalt handelt (Haustein et al., 2016). Zudem ergeben sich weitere Unschärfen. Diese sind bedingt durch die nicht immer genauen Ortsangaben (Koordinaten) der Tweets, durch auch maschinell schwer zu klassifizierende Inhalte oder dadurch, dass Text und Ort der Veröffentlichung nicht unbedingt in einem inhaltlichen Zusammenhang stehen müssen (siehe dazu auch die Diskussion im vorherigen Kapitel 4.2). Die Themen können dann semantisch nicht in einen räumlichen Zusammenhang gesetzt werden oder sind in sich semantisch nicht interpretierbar.

Crooks et al. (2015) fassen die resultierenden Herausforderungen im Umgang mit crowdbasierten Daten zusammen: Es gibt im Forschungsfeld der crowdbasierten Daten noch keine Standards für die Daten an sich sowie für die Metadaten, wie sie in anderen, traditionellen Bereichen der Arbeit mit geographischen Daten etabliert sind. Die Fähigkeit, crowdbasierte Daten in ihrer Menge und Heterogenität zu sammeln, zu verarbeiten und zu analysieren wird aber ein Schlüsselelement auf

dem Weg hin zur Nutzung des vollen Potentials dieser Daten darstellen. Die Dynamik der Daten und ihre heterogene Beschaffenheit, die nötigen räumlichen und zeitlichen Analysemethoden sowie die Visualisierung der Ergebnisse stellen dabei die zentralen Herausforderungen dar.

5 Fazit und Ausblick

Die Ergebnisse der Master Thesis zeigen, dass die Verbindung zwischen semantischen Strukturen in Tweets und räumlichen Strukturen in der realen Welt hergestellt werden kann. Räumlich verortbare Tweets spiegeln in ihrer Verteilung reale Orte wieder. Dabei werden jedoch nicht alle Orte und räumlichen Muster exakt abgebildet. Die Arbeit mit den Daten erfordert vielmehr die Eliminierung des „digitalen Rauschens“, das aus ungenauen Koordinaten und falschen Ortsangaben, statistisch abgeleiteten Themen mit gewissen Irrtumswahrscheinlichkeiten sowie „falschen“ Beiträgen von Bots und Nachrichten ohne für die Analyse relevante Inhalte resultiert. Trotz dieser Ungenauigkeiten und Fehler treten semantische, räumliche und zeitliche Muster aus den Daten klar hervor und können auch durch weitere Datengrundlagen und Wissen über die realen Orte bestätigt werden.

Eine naheliegende Fragestellung, die im Rahmen dieser Master Thesis nicht betrachtet wurde, ist der Vergleich verschiedener städtischer Räume innerhalb einer Stadt oder vergleichbarer städtische Räume im Vergleich unterschiedlicher Städte. In einer weiterführenden Analyse auf Basis der verwendeten Datengrundlagen wäre es beispielsweise möglich, solche Räume hinsichtlich ihrer Nutzungsintensität miteinander zu vergleichen. Damit verbunden wäre die Frage, ob stärker frequentierte Räume mit einem höheren Aufkommen an Tweets korrelieren oder ob hierfür z.B. die Aufenthaltsdauer der Nutzer oder andere Faktoren zur Erklärung dienen.

Eine weitere wichtige Datengrundlage im Bereich der Stadtentwicklung stellen Befragungen von Passanten bzw. Nutzern öffentlicher Räume dar. Hier ergibt sich auch Potenzial für weitergehende Analysen, in dem die Wahrnehmung der Befragten hinsichtlich der bevorzugten Nutzungen eines Raumes den Themen aus Tweets gegenübergestellt wird. Ob Tweets eine geeignete Datengrundlage für stadträumliche Fragestellungen sind, könnte somit auch noch einmal kritisch reflektiert werden. Dies eröffnet dann Möglichkeiten, Daten aus sozialen Medien für Stadtplanungsprozesse einzusetzen. Es wird in den kommenden Jahren wichtig sein, diesen Schritt von der Analyse zur Anwendung in der Praxis kritisch zu begleiten und stets die verwendeten Methoden zu reflektieren. Es muss sich eben erst noch zeigen, ob soziale Medien eine zusätzliche Säule im Rahmen von entsprechenden Verfahren und vorbereitenden Analysen sein können.

Auf dem Weg dahin müssen die Methoden aber auch noch weiter verfeinert werden. Das gilt für alle in dieser Master Thesis verwendeten Bausteine beginnend bei der Datenbereinigung über die semantische Klassifizierung bis hin zur räumlichen Analyse. Schwächen in den Methoden und den Datengrundlagen müssen klar kommuniziert werden. Daten aus sozialen Medien sollten weiterhin nicht für sich alleine stehen, sondern immer im Zusammenhang mit anderen Datengrundlagen betrachtet werden. Auch dies haben die Analysen in dieser Master Thesis gezeigt.

Literatur

BLEI, D. M. 2012. Probabilistic Topic Models. *Communications of the Acm,* 55**,** 77-84.

BLEI, D. M., NG, A. Y. & JORDAN, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research,* 3**,** 993-1022.

BOY, J. D. & UITERMARK, J. 2016. How to Study the City on Instagram. *Plos One,* 11.

BOYD, D. M. & ELLISON, N. B. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication,* 13**,** 210-230.

BRABHAM, D. C. 2009. Crowdsourcing the Public Participation Process for Planning Projects. *Planning Theory,* 8**,** 242-262.

CALDAS DE CASTRO, M. & SINGER, B. H. 2006. Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. *Geographical Analysis,* 38**,** 180-208.

CHANG, J. & BLEI, D. M. Relational topic models for document networks. International conference on artificial intelligence and statistics, 2009. 81-88.

CHENG, X., YAN, X., LAN, Y. & GUO, J. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering,* 26**,** 2928-2941.

CHENG, Z., CAVERLEE, J. & LEE, K. You are where you tweet: a content-based approach to geo-locating twitter users. Proceedings of the 19th ACM international conference on Information and knowledge management, 2010. ACM, 759-768.

CRAGLIA, M., OSTERMANN, F. & SPINSANTI, L. 2012. Digital Earth from vision to practice: making sense of citizen-generated content. *International Journal of Digital Earth,* 5**,** 398-416.

CRAGLIA, M. & SHANLEY, L. 2015. Data democracy – increased supply of geospatial information and expanded participatory processes in the production of data. *International Journal of Digital Earth,* 8**,** 679-693.

CROITORU, A., WAYANT, N., CROOKS, A., RADZIKOWSKI, J. & STEFANIDIS, A. 2015. Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems,* 53**,** 47-64.

CROOKS, A., PFOSER, D., JENKINS, A., CROITORU, A., STEFANIDIS, A., SMITH, D., KARAGIORGOU, S., EFENTAKIS, A. & LAMPRIANIDIS, G. 2015. Crowdsourcing urban form and function. *International Journal of Geographical Information Science,* 29**,** 720-741.

ELWOOD, S., GOODCHILD, M. F. & SUI, D. Z. 2012. Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers,* 102**,** 571-590.

FEINERER, I. 2010. Analysis and algorithms for stemming inversion. *Information Retrieval Technology***,** 290-299.

FEINERER, I. & HORNIK, K. 2017. tm: Text Mining Package. R package version 0.7-1.

GARCIA-PALOMARES, J. C., GUTIERREZ, J. & MINGUEZ, C. 2015. Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography,* 63**,** 408-417.

GETIS, A. & ORD, J. K. 1992. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis,* 24**,** 189-206.

GHOSH, D. & GUHA, R. 2013. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science,* 40**,** 90-102.

GOODCHILD, M. F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal,* 69**,** 211-221.

GOODCHILD, M. F. & LI, L. N. 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics,* 1**,** 110-120.

GRIFFITHS, T. L. & STEYVERS, M. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences,* 101**,** 5228-5235.

HAUSTEIN, S., BOWMAN, T. D., HOLMBERG, K., TSOU, A., SUGIMOTO, C. R. & LARIVIERE, V. 2016. Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter. *Journal of the Association for Information Science and Technology,* 67**,** 232-238.

HORNIK, K. & GRÜN, B. 2011. topicmodels: An R package for fitting topic models. *Journal of Statistical Software,* 40**,** 1-30.

JAVA, A., SONG, X., FININ, T. & TSENG, B. 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.* San Jose, California: ACM.

JENKINS, A., CROITORU, A., CROOKS, A. T. & STEFANIDIS, A. 2016. Crowdsourcing a Collective Sense of Place. *Plos One,* 11**,** 20.

JIANG, B., MA, D., YIN, J. J. & SANDBERG, M. 2016. Spatial Distribution of City Tweets and Their Densities. *Geographical Analysis,* 48**,** 337-351.

KIM, K.-S., KOJIMA, I. & OGAWA, H. 2016. Discovery of local topics by using latent spatio-temporal relationships in geo-social media. *International Journal of Geographical Information Science,* 30**,** 1899-1922.

KITCHIN, R. 2014. The real-time city? Big data and smart urbanism. *GeoJournal,* 79**,** 1-14.

LEE, R., WAKAMIYA, S. & SUMIYA, K. 2013. Urban area characterization based on crowd behavioral lifelogs over Twitter. *Personal and Ubiquitous Computing,* 17**,** 605-620.

LI, L., GOODCHILD, M. F. & XU, B. 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science,* 40**,** 61-77.

LIU, Y., LIU, X., GAO, S., GONG, L., KANG, C. G., ZHI, Y., CHI, G. H. & SHI, L. 2015. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers,* 105**,** 512-530.

LONG, Y. & LIU, L. 2016. Transformations of urban studies and planning in the big/open data era: a review. *International Journal of Image and Data Fusion,* 7**,** 295-308.

LONGLEY, P. A. & ADNAN, M. 2016. Geo-temporal Twitter demographics. *International Journal of Geographical Information Science,* 30**,** 369-389.

LONGLEY, P. A., ADNAN, M. & LANSLEY, G. 2015. The geotemporal demographics of Twitter usage. *Environment and Planning A,* 47**,** 465-484.

MEYER, D., HORNIK, K. & FEINERER, I. 2008. Text mining infrastructure in R. *Journal of statistical software,* 25**,** 1-54.

MORAN, P. A. P. 1950. Notes on Continuous Stochastic Phenomena. *Biometrika,* 37**,** 17-23.

NEIS, P. & ZIELSTRA, D. 2014. Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. *Future Internet,* 6**,** 76-106.

O'BRIEN, J., SERRA, M., HUDSON-SMITH, A., PSARRA, S., HUNTER, A. & ZALTZ-AUSTWICK, M. 2016. Ensuring VGI Credibility in Urban-Community Data Generation: A Methodological Research Design. *Urban Planning,* 1**,** 88-100.

OPENSTREETMAP - DEUTSCHLAND. *Was ist OpenStreetMap?* [Online]. Available: https://www.openstreetmap.de/ [Accessed 2018-02-10].

ORD, J. K. & GETIS, A. 1995. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis,* 27**,** 286-306.

PONWEISER, M. 2012. Latent Dirichlet allocation in R.

RESCH, B., SUMMA, A., ZEILE, P. & STRUBE, M. 2016. Citizen-Centric Urban Planning through Extracting Emotion Information from Twitter in an Interdisciplinary Space-Time-Linguistics Algorithm. *Urban Planning,* 1**,** 114-127.

SALESSES, P., SCHECHTNER, K. & HIDALGO, C. A. 2013. The Collaborative Image of The City: Mapping the Inequality of Urban Perception. *Plos One,* 8**,** 12.

SCHÜLLER, K. F., AGNES; 2017. Digital Literacy für die Stadt. *Informationen zur Raumentwicklung*.

SEE, L., MOONEY, P., FOODY, G., BASTIN, L., COMBER, A., ESTIMA, J., FRITZ, S., KERLE, N., JIANG, B., LAAKSO, M., LIU, H. Y., MILCINSKI, G., NIKSIC, M., PAINHO, M., PODOR, A., OLTEANU-RAIMOND, A. M. & RUTZINGER, M. 2016. Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *Isprs International Journal of Geo-Information,* 5**,** 23.

SENARATNE, H., MOBASHERI, A., ALI, A. L., CAPINERI, C. & HAKLAY, M. 2016. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science***,** 1-29.

SHELTON, T., POORTHUIS, A. & ZOOK, M. 2015. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning,* 142**,** 198-211.

STEIGER, E., DE ALBUQUERQUE, J. P. & ZIPF, A. 2015a. An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Transactions in GIS,* 19**,** 809-834.

STEIGER, E., RESCH, B., DE ALBUQUERQUE, J. P. & ZIPF, A. 2016. Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. *Transportation Research Part C: Emerging Technologies,* 73**,** 91-104.

STEIGER, E., WESTERHOLT, R., RESCH, B. & ZIPF, A. 2015b. Twitter as an indicator for whereabouts of people? Correlating twitter with uk census data. *Computers, Environment and Urban Systems,* 54**,** 255-265.

TENNEY, M. & SIEBER, R. 2016. Data-Driven Participation: Algorithms, Cities, Citizens, and Corporate Control. *Urban Planning,* 1**,** 101-113.

TOBLER, W. R. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography,* 46**,** 234-240.

TSOU, M. H. & LEITNER, M. 2013. Visualization of social media: seeing a mirage or a message? *Cartography and Geographic Information Science,* 40**,** 55-60.

TWITTER INC. *Twitter Nutzung / Fakten zum Unternehmen* [Online]. Available: https://about.twitter.com/de/company [Accessed 09-17 2017].

WESTERHOLT, R., STEIGER, E., RESCH, B. & ZIPF, A. 2016. Abundant Topological Outliers in Social Media Data and Their Effect on Spatial Analysis. *Plos One,* 11**,** 31.

ZHOU, X. L. & ZHANG, L. 2016. Crowdsourcing functions of the living city from Twitter and Foursquare data. *Cartography and Geographic Information Science,* 43**,** 393-404.

Anhang

Datensätze

Tabelle 4: Manchester-Datensatz (Metadaten)

Anzahl Tweets (unbereinigt)	n = 99.884
Anzahl Tweets (bereinigt)	n = 91.952
Erhebungsstart	12.04.2017
Erhebungsende	26.10.2017
Quelle	eigene Erhebung (Twitter Search API)

Tabelle 5: Birmingham-Datensatz (Metadaten)

Anzahl Tweets (unbereinigt)	n = 826.282
Anzahl Tweets (bereinigt)	n = 702.568
Erhebungsstart	01.01.2015
Erhebungsende	31.12.2015
Quelle	Interfakultärer Fachbereich Geoinformation – Z_GIS an der Universität Salzburg

Tabelle 6: OpenStreetMap-Datensatz (Metadaten)

Abfragedatum	07.02.2018 (21:43 Uhr)
Quelle	Daten von OpenStreetMap - Veröffentlicht unter ODbL, http://opendatacommons.org/licenses/odbl
Download Daten:	http://download.geofabrik.de/europe/great-britain/england-latest-free.shp.zip

Tabelle 7: Spiele des Aston Villa Football Club 2015 (Metadaten)

Abfragedatum	Mai 2018
Quellen	eigene Recherche, Datenbasis: www.fussballdaten.de, www.flashscores.co.uk

Tabellen

Tabelle 8: Vereinheitlichte Wörter

Ursprüngliches Wort	Vereinheitlichung	Datensatz
altrinchamtoday	altrincham today	Manchester
bankholidayweekend	bankholiday weekend	Manchester
manchesterutd	manchesterunited	Manchester
manutd		
manunited		
manchesterunited		
mancity	manchestercity	Manchester
manchestercity		
bmcr	manchester	Manchester
bhx	airport	Birmingham

Tabelle 9: Ergänzende Stoppwortliste

Stoppwörter
got
ive
still
try
dont
amp
regrann
aka
can
like
say
one
way
use
also
however
tell
will
much
need
take
tend
even
particular
rather
said
get
make
ask

Stoppwörter
come
two
often
may
might
see
something
thing
point
post
look
right
now
another
put
set

Tabelle 10: Maximale z-Werte für ausgewählte Themen in Manchester (Global Moran's I)

Thema	Maximum bei Distanz (in m)	Moran's Index	Varianz	z-score	p-value
1	1.800	0,162225	0,000001	197,960073	0,000000
2	1.400	0,081219	0,000001	79,868197	0,000000
5	1.800	0,087854	0,000001	107,356244	0,000000
6	2.000	0,080660	0,000001	105,938072	0,000000
7	1.800	0,099270	0,000001	123,638733	0,000000
10	1.500	0,062815	0,000001	69,079172	0,000000
14	1.500	0,009692	0,000001	9,790972	0,000000
16	1.800	0,031755	0,000000	46,766663	0,000000
17	2.100	0,015471	0,000000	44,015476	0,000000
19	800	0,060222	0,000002	49,025212	0,000000
22	1.600	0,093931	0,000001	103,631280	0,000000
24	1.200	0,081200	0,000001	67,005379	0,000000
26	1.800	0,037560	0,000001	46,163320	0,000000
27	1.100	0,009446	0,000001	8,811351	0,000000
28	1.700	0,029433	0,000001	38,674579	0,000000
29	1.200	0,011929	0,000001	11,868264	0,000000
30	1.700	0,034018	0,000001	41,185515	0,000000
32	1.100	0,138115	0,000001	113,723196	0,000000

Tabelle 11: Maximale z-Werte für ausgewählte Themen in Birmingham (Global Moran's I)

Thema	Maximum bei Distanz (in m)	Moran's Index	Varianz	z-score	p-value
1	1.400	0,248382	0,000003	149,647474	0,000000
2	1.200	0,142033	0,000003	76,408378	0,000000
3	1.300	0,175371	0,000003	101,276809	0,000000
5	1.700	0,151889	0,000002	108,171043	0,000000
7	1.700	0,142068	0,000002	101,548476	0,000000
8	1.700	0,118171	0,000002	84,873448	0,000000
12	1.400	0,229357	0,000003	136,423804	0,000000
14	1.100	0,260249	0,000004	128,474123	0,000000
17	1.400	0,164077	0,000003	97,451749	0,000000
18	800	0,211776	0,000007	80,889180	0,000000
19	2.000	0,110492	0,000001	91,690179	0,000000
21	1.800	0,093601	0,000001	81,333732	0,000000
23	1.500	0,167728	0,000003	103,649513	0,000000
24	1.400	0,131111	0,000003	80,702554	0,000000
26	1.500	0,052802	0,000002	36,377914	0,000000
30	1.600	0,080734	0,000001	74,912814	0,000000

Tabelle 12: Maximale z-Werte für OpenStreetMap-Daten in Manchester (Global Moran's I)

Thema	Maximum bei Distanz (in m)	Moran's Index	Varianz	z-score	p-value
Nachtleben	1.100	0,167839	0,000002	128,274957	0,000000
Essen	1.000	0,279492	0,000003	176,132446	0,000000
Einkaufen	700	0,193740	0,000005	87,064956	0,000000
Tourismus	1.100	0,035966	0,000002	27,343620	0,000000
Zusammengefasst	1.000	0,262339	0,000003	163,473231	0,000000

Tabelle 13: Maximale z-Werte für OpenStreetMap-Daten in Birmingham (Global Moran's I)

Thema	Maximum bei Distanz (in m)	Moran's Index	Varianz	z-score	p-value
Nachtleben	1.200	0,133192	0,000003	72,683192	0,000000
Essen	1.000	0,222011	0,000006	87,314789	0,000000
Zusammengefasst*	700	0,284702	0,000013	79,960638	0,000000

** einschließlich Einkaufen, Tourismus*