UNIVERSITÄT
LEIPZIG

Institut für Informatik
Fakultät für Mathematik und Informatik
Abteilung Datenbanken

Privatsphäre-erhaltende Analyse des Fahrradklimas in Leipzig

Masterarbeit

vorgelegt von:
Aruscha Kramm

Matrikelnummer:
3738552

Betreuende:
Prof. Dr. Erhard Rahm
Dr. Thomas Burghardt

Inhaltsverzeichnis

Zusammenfassung

Privatsphäre der eigenen Daten ist ein immer größer und bedeutsamer werdendes Thema. Besonders im Bereich von Applikationen, die standortbasierte Dienste bieten, ist der Schutz der Privatsphäre von großer Bedeutung. Diese Applikationen müssen den Schutz gegenüber ihren Nutzenden gewährleisten. Gleichzeitig können die von den Applikationen gesammelten Daten zur Auswertung wichtiger Informationen beispielsweise in der Stadtplanung genutzt werden. Die Herausforderung liegt im Schutz der Privatsphäre der im Datensatz enthaltenen Personen zusammen mit dem größtmöglichen Erhalt der relevanten Information. Der Kompromiss dieser beiden Gegenspieler mündet im sogenannten *Privacy Utility Trade-off*.

Die sogennanten *Privacy Preserving Mechanisms* (Privatsphäre-erhaltenden Methoden) bieten verschiedene Ansätze, Daten zu verschiedenen Zeitpunkten Privatsphäre-erhaltend zu übermitteln oder zu veröffentlichen. Über verschiedene Metriken kann anschließend die Nützlichkeit und der bewahrte Informationsgehalt quantisiert und gemessen werden.

Bezüglich des Klimawandels stellt die Verkehrswende ein unausweichliches Thema dar. Das Fahrrad ist eine klimafreundliche, alternative Transportmöglichkeit zum Auto. Die Förderung eines guten Fahrradklimas kann Personen zum Umstieg auf das Fahrrad motivieren.

In dieser Arbeit wird ein Datensatz der Firma Nextbike privatisiert und anschließend ausgewertet. Ziel der Arbeit ist es, verschiedene Privatsphäre-erhaltende Methoden vorzustellen und eine Auswahl dieser zur Privatisierung des Nextbike-Datensatzes zu implementieren. Anschließend soll anhand verschiedener Metriken der Grad der Privatsphäre und der Nützlichkeit gemessen werden.

Folgend soll der privatisierte Datensatz inhaltlich bezüglich des Leipziger Fahrradklimas ausgewertet werden. Dabei werden sowohl Hauptverkehrsrouten von Fahrradfahrenden ausgewertet, als auch verschiedene Wochentage und Tageszeiten betrachtet. Zusätzliche Informationen werden genutzt, um die Ergebnisse aus dem Datensatz mit Fragen zu Radverkehrsanlagen oder der Beschaffenheit der Straße zu verknüpfen.

Abstract

The topic of privacy of one's data is getting bigger and more important. Applications offering location-based services (LBS) should ensure the protection of privacy of their users and their data. At the same time the collected data can be very valuable when used to improve structures, like in a city planning context. The challenge lies in the so-called Privacy Utility Trade-off: Preserve privacy of users in the data while keeping the best possible utility of the data.

Privacy Preserving Mechanisms offer a variety of approaches to preserve privacy of data while they are being transmitting or going to be published. After privatising the data, its utility and privacy can be measured using different metrics.

Transformation of transport is an important aspect in matters of climate change. Using bikes as an alternative to cars depicts an environmental-friendly way of personal transport. By providing a bicycle-friendly and safe environment, a good bicycle climate, cities can motivate their inhabitants to use bicycles instead of cars.

This work is using a dataset provided by Nextbike to implement and evaluate three Location Privacy Preserving Mechanisms. It's aim is to introduce different methods before implementing a selection. Following the privatisation is the evaluation of the achieved level of privacy and utility using different metrics.

Finally the privatised dataset is used to evaluate the bicycle climate of the city of Leipzig. To do so major bicycle traffic ways are detected as well as bicycle traffic peaks concerning different days or times. Additional information is gathered to combine these results with information about bicycle lanes or conditions of most used streets.

Abkürzungsverzeichnis

API	Application Programming Interface
AS	Anonymitäts-Set
ASS	Anonymity Set Size (dt. Anonymititäts-Setgröße)
C_{AVG}	Average Equivalence Class Size Metric (Kostenmetrik)
C_{DM}	Discernability Metric (Kostenmetrik)
DP	Differential Privacy
EQ	Equivalence class (dt. Äquivalenzklasse)
GIS	Geo-Informationssystem
GPS	Global Positioning System
JSON	Java Script Object Notation
LBS	Location Based Services (dt. standortbezogene Dienste)
LPPM	Location Privacy Preserving Mechanism
OSM	Open Street Map
PIR	Private Information Retrieval
PLZ	Postleitzahl
POI	Point of Interest
SQL	Structured Query Language
QI	Quasi-identifier

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Einleitend soll die Problematik erläutert werden, mit der sich diese Arbeit befasst, sowie Ziele formuliert werden, die mit dieser Arbeit erreicht werden sollen.

1.1. Problemstellung

Location-based Services (LBS) sind Dienste, die die Umgebung von Personen auf relevante Information wie beispielsweise Geschäfte, Restaurants oder Freunde scannen. Sie benutzten dazu GPS, eine Komponente, die mittlerweile standardmäßig in jedem Smartphone enthalten ist. Die Verfügbarkeit dieser beiden Funktionalitäten hat dazu geführt, dass LBS mittlerweile fest in den Alltag integriert sind: Die schnellste Route zwischen zwei Punkten finden oder die Umgebung nach einem Supermarkt absuchen, sind nur einige der Beispiele, bei denen solche Services den Standort einer Person abfragen, um die gewünschte Information zu liefern.

Solche Dienste werden heute explizit von Personen gewollt und auch täglich benutzt, was zu einem starken Wachstum von LBS geführt hat [1]. In Bereichen wie Navigation, lokaler Geschäftssuche (Restaurant, Supermarkt, ..), Social Networks (Geo-tagged tweets), Sport, aber auch Gesundheitsapps (Schrittzähler) und Augmented Reality-Games (Pokemon-go) werden LBS bereits großflächig genutzt. Neben der aktiv von Personen genutzten Applikationen und der damit abgefragten Information, gibt es Dienste, die passiv den Standort der Person abfragen und nutzen. Dabei laufen die Applikationen nur im Hintergrund wie beim e-marketing (Coupons an Nutzende senden, wenn sie sich in der Nähe des Geschäfts befinden) oder zur Beobachtung des Autoverkehrs für aktuelle Staumeldungen [2]. Während die verschiedenen LBS zweifellos eine Bereicherung darstellen, darf nicht aus den Augen verloren werden, dass für jede Abfrage einer Information, sei es einmalig oder kontinuierlich, der eigene Standort preisgegeben wird. Je nach Aufenthaltsort kann damit womöglich private Information wie die eigene Wohnadresse oder sensible Information wie die Adresse eines besuchten Krankenhauses offen gelegt werden.

Im Laufe der Jahre hat sich die Art der Datensammlung stark verändert: Das Smartphone ist die wichtigste Datenquelle für unter anderem ortsbezogene Informationen geworden. Es sammelt leise und unauffällig eine große Menge Daten über seine Nutzenden, welche sich oft nicht bewusst sind, wie stark der Eingriff in ihre Privatsphäre dabei ist [1]. Statistiken zeigen eine immer größer werdende Sorge der Nutzenden um ihre Privatsphäre und steigende Forderung nach dieser [2]. Obwohl in den USA Privatsphäre laut der 1948 festgelegten *Universal Declaration of Human Rights* sogar ein staatlich anerkanntes Recht ist, besteht bereits eine Problematik bei der Umsetzung dieses Rechtes aufgrund der unpräzisen Definition von Privatsphäre [3].

Privacy Preserving Mechanisms (dt. Privatsphäre-erhaltende Methodiken) und im Spezialfall für ortsgebundene Daten *Location Privacy Preserving Mechanisms* (LPPM) sind schon seit circa 20 Jahren Gegenstand der Forschung, damals zunächst im Bereich Privatsphäre von Datenbanken. Heute ist das Ziel dieser Methoden, die Identität einer Person zu schützen, während sie eine Applikation nutzt, oder wenn ihre Daten bereits in einem Datensatz festgehalten wurden. Sie können in

zwei Gruppen unterschieden werden: Methoden, die während der Nutzung eines Services in Echtzeit beziehungsweise online die Privatsphäre schützen und Methoden, die einen Datensatz offline privatisieren. Es gibt bereits eine Vielzahl dieser Methoden in der Theorie, dennoch herrscht eine große Lücke zur Anwendung in der Praxis [1].

Der vorliegenden Arbeit steht ein Datensatz des Bikesharing-Anbieters *Nextbike* zur Verfügung, dessen Fahrräder nach dem *free floating*-Prinzip in Leipzig zur Verfügung stehen. Das bedeutet die Ausleihe und Abgabe sind flexibel zu unterschiedlichen Gebühren jederorts möglich. Darin sind etwa 240.000 Ausleihen im Zeitraum Juli bis September 2019 enthalten. Ausleihdatensätze dieser Art stellen eine Sammlung sensibler Daten dar, da die Start- und Endpunkte der Ausleihen fast beliebig wählbar sind. Ein Fahrrad kann also vor der eigenen Haustür ausgeliehen oder vor der Arbeitsstelle abgegeben werden. Wird ein solcher Datensatz unprivatisiert veröffentlicht, können sensible Informationen einzelner Personen enthüllt werden. Information wie Wohn- oder Arbeitsadressen werden auch als *points of interest* (POI) bezeichnet. Sie bezeichnen häufig frequentierte Orte von Personen, also Orte, die für Angreifende besonders interessant sind. Kontinuierliche Daten wie Routen, auch Trajektorien genannt, können Bewegungsmuster einzelner Personen verdeutlichen und damit Gewohnheiten von Personen offenlegen, beispielsweise joggt eine Person jeden zweiten Tag um 9:00 Uhr im Park. Im Verlauf der Arbeit werden verschiedene LPPM vorgestellt und untersucht, ob und wie diese zur Privatisierung des Nextbike-Datensatzes angewandt werden können. Anschließend werden ausgewählte LPPM für den Datensatz implementiert.

Ein zweiter großer Aspekt der Arbeit bildet die Auswertung des Datensatzes hinsichtlich der beinhalteten Informationen. Dafür sollen die privatisierten Fahrraddaten visualisiert werden, um Rückschlüsse über das Fahrverhalten und das Fahrradklima in Leipzig zu ermöglichen. In der Stadtplanung besteht eine große Aufgabe bezüglich des nicht motorisierten Verkehrs darin, die Wege von Fahrradfahrenden und zu Fuß Gehenden zu dokumentieren [4]. Verkehr und Stadtplanung werden durch den Klimawandel in großen Städten zu einem immer wichtigeren Thema. Die stetig wachsende Zahl zugelassener PKWs lässt das Verkehrsaufkommen durch PKW zunehmend steigen [5]. Einerseits hat dies einen negativen Einfluss auf die CO_2-Bilanz und andererseits führt es zu einer erhöhten Lärmbelastung für Anwohnende. Zudem nimmt es anderen Verkehrsteilnehmenden wie Fahrradfahrenden und zu Fuß Gehenden vermehrt den Raum [6]. Ein großes Ziel von Städten sollte es daher sein, Anreize für ihre Einwohnenden zu schaffen, auf alternative Transportmittel wie das Fahrrad umzusteigen. Um Menschen von der Nutzung des Fahrrades zu überzeugen, ist eine gut ausgebaute als auch sichere Fahrradinfrastruktur ein wichtiges Argument [7]. Dazu gehören das Vorhandensein eines Radfahrstreifens oder sogar eines von der Straße abgetrennten Abschnittes. Außerdem wichtig ist die Beschaffenheit beziehungsweise Qualität der Radfahrstreifen als auch der Straßen, die nicht über einen Radstreifen verfügen. Während des Fahrradfahrens möchten die meisten Menschen sich sicher fühlen, um es als eine dauerhafte Alternative zu nutzen [8].

Anhand der Visualisierung des Nextbike-Datensatzes in Kombination mit zusätzlichen öffentlichen Ressourcen wird eine exemplarische Auswertung des Fahrradklimas in Leipzig durchgeführt. Dabei sollen Fragen beantwortet werden, welche Strecken häufig gefahren werden, welche Beschaffenheit die Strecken haben und ob auf diesen ein Radstreifen vorhanden ist. Zur Durchführung dieser Analyse wird der Nextbike-Datensatz mit verschiedenen LPPM und unterschiedlich hohen Stufen privatisiert. Die aus dem Datensatz gewonnenen Erkenntnisse über das Fahrverhalten, als auch über

das Fahrradklima in Leipzig sollen dazu motivieren, die Möglichkeiten der vorgestellten Methoden und Analysen zu nutzen und die gewonnenen Informationen in Planungen einzubeziehen.

1.2. Ziele

Die Zielsetzung der Arbeit ist es, unterschiedliche, bereits existierende LPPM vorzustellen und anschließend zu prüfen, ob und wie der vorliegende Nextbike-Datensatz mit diesen privatisiert werden kann. Gleichzeitig soll eine Gesamtübersicht über die hier betrachteten Mechanismen entstehen. Aus den vorgestellten Mechanismen werden ausgewählte implementiert, sodass die Daten privatisiert werden und unterschiedlich hohe Privatsphäre-Level geschaffen werden. Anschließend soll die Effizienz der Mechanismen geprüft werden, das bedeutet die erreichte Privatsphäre und die anschließende Nützlichkeit der Daten muss evaluiert werden. Dazu werden verschiedene Metriken angewandt, um den sogenannten *Privacy Utility Trade-off*, den Kompromiss zwischen Datenverlust und Privatisierung, auszuwerten.

Im zweiten Teil der Arbeit werden die Ergebnisse in verschiedenen Visualisierungen dargestellt. Dabei werden zwei verschiedene Visualisierungsformen, Heat Map und Flow Map, gewählt, um die privatisierten Datensätze bestmöglich zu repräsentieren. Diese Visualisierungen werden anschließend genutzt, um eine Analyse des Fahrradklimas in Leipzig durchzuführen. Abschließend wird so gezeigt, welche Information weiterhin aus privatisierten Datensätze zu gewinnen ist, während die Privatsphäre der darin enthaltenen Personen geschützt wird.

1.3. Aufbau der Arbeit

Im weiteren Verlauf ist die Arbeit wie folgt aufgebaut: Zunächst folgt eine Einführung in die Grundlagen der Privatsphäre in Kapitel 2. Zur Einführung in die Grundlagen werden die Begriffe Privatsphäre, Privatsphäre ortsgebundener Daten und *Privacy Utility Trade-off* erläutert, um anschließend die verschiedenen Privatsphäre-erhaltenden Mechanismen vorstellen zu können. Anschließend erfolgt eine Betrachtung des aktuellen Forschungsstandes in Kapitel 3. Weiterführend werden in Kapitel 4 alle vorgestellten Mechanismen daraufhin beleuchtet, ob sie für die Privatisierung der Nextbike-Daten in Betracht kommen.

Bevor in Kapitel 4.2 eine Vorstellung der Implementierung der Methoden erfolgt, werden die zur Implementierung notwendigen Services und Datentypen und die erforderliche Vorverarbeitung des Datensatzes betrachtet.

In Kapitel 4.3 werden die privatisierten Daten anhand der in Kapitel 2.7 vorgestellten Metriken evaluiert. Dabei wird der Datensatz auf Privatsphäre und Nützlichkeit quantitativ untersucht und bewertet. Daran schließt eine qualitative Evaluation der Daten mithilfe von Visualisierungen an. Dabei werden die verschiedenen Privatsphäre-Level auf den Aspekt hin betrachtet, ab wann der Visualisierung nicht mehr ausreichend oder keine Information entzogen werden kann.

Um eine Auswertung der Daten und Visualisierungen durchzuführen, wird in Kapitel 5 abschließend eine Analyse des Leipziger Fahrradklimas durchgeführt, in der zusätzlich zu den privatisierten Nextbike-Daten externe Daten hinzugezogen werden.

2. Theoretische Grundlagen zu Privatsphäre

In diesem Kapitel werden die theoretischen Grundlagen erläutert. Dazu wird einleitend erklärt, was ortsgebundene Daten sind und wie diese von standortbezogenen Diensten, kurz LBS, gesammelt werden. Anschließend wird in die Begriffe der Privatsphäre und speziell Privatsphäre für ortsgebundene Daten eingeführt. Außerdem werden die Risiken beleuchtet, die bei der Offenlegung von unprivatisierten Daten entstehen können und verschiedene Metriken betrachtet, wie nach der Privatisierung das erreichte Level an Privatsphäre als auch die resultierende Nützlichkeit der Daten evaluiert werden kann. Der Hauptteil des Kapitels stellt die verschiedenen Privatsphäre-erhaltenden Mechanismen vor.

2.1. Standortbezogene Daten und Dienste

In diesem Abschnitt wird beleuchtet, welche Daten von LBS aufgezeichnet und übermittelt werden und in welcher Form ortsbezogene Daten vorliegen können. Dabei wird unter anderem die Architektur von LBS betrachtet.

Die Art und Weise Bewegungen von Personen zu verfolgen hat sich verändert: Vor circa 20 bis 30 Jahren war es nur möglich, eine Person zu verfolgen, in dem eine andere Person mit der Verfolgung beauftragt wurde. Diese Methode war teuer und barg das Risiko, entdeckt zu werden. Heute werden Informationen über Aufenthaltsorte und Bewegungen von Personen leise, unauffällig und vor allem kostenlos von Geräten gesammelt, die Menschen immer mit sich tragen: Smartphones, Kreditkarten, die Magnetstreifenkarte des öffentlichen Nahverkehrs, um nur einige zu nennen [9].

Smartphones sind in der Lage diese Standortdaten zu erfassen, da viele Applikationen standortbezogene Dienste und Funktionen anbieten, die Informationen basierend auf der Umgebung der Person bereitstellen. Indem Personen standortbezogene Dienste in ihren Alltag integrieren, werden täglich große Mengen dieser Nutzungsdaten gespeichert. Dennoch ist an dieser Stelle wichtig, solche Dienste zu unterscheiden: Zum einen existieren Dienste, die den schnellen Fortschritt der Technologie zur Datensammlung ausnutzen, wie beispielsweise eine mit Gesichtserkennung ausgestattete Kamera, die platziert im öffentlichen Raum erkennen kann, wer sich in diesem öffentlichen Raum befindet. Nichtsdestotrotz existieren auch nützliche Dienste wie Navigations-Apps, Parkuhren, die per SMS bezahlt werden können und viele andere. Vor allem letztere sind oft innovativ und von großem Nutzen, beinhalten aber auch ein Risiko für die Privatsphäre des Standortes [9].

Die Architektur von LBS besteht aus vier Hauptkomponenten (siehe Abbildung 2.1): einem mobilen Endgerät, beispielsweise einem Smartphone, Positionierungssystemen, Kommunikationsnetzwerken und einem Service Provider. Smartphonenutzende senden spezifische Anfragen an die LBS-Server des Service-Providers, beispielsweise: „Italienisches Restaurant in der Nähe“. Dabei werden ihre Standorte von GPS-Positionierungssystemen erfasst und die Anfragen sowie zugehörigen Antworten über Kommunikationsnetze, beispielsweise mobile 4/5G-Netze, zum und vom LBS-Server gesendet. LBS sind die Service-Provider, die diese Anfragen möglichst präzise beantworten [10].

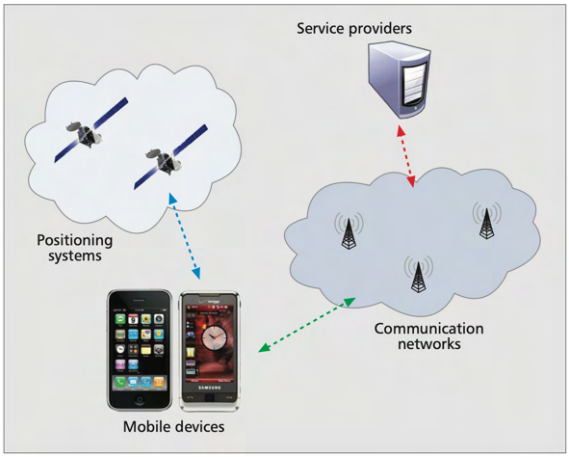


Abbildung 2.1.: Architektur eines Location Based Service, Quelle: [10]

Nutzt eine Person beispielsweise eine Smartphone-App, um die täglich beim Joggen gelaufenen Routen aufzuzeichnen, sendet die App in regelmäßigen Abständen einen Standort der Person an den LBS-Server. Der Server speichert diese Information, beispielsweise um am Ende in der Applikation die gelaufene Route visualisieren zu können. Ein Standort wird als Koordinatenpaar (Latitude, Longitude) übermittelt und Routen werden als geordnete Listen solcher Koordinatenpaare aufgezeichnet. Die Jogging-App überträgt kontinuierlich Daten an den LBS-Server, um die gelaufene Trajektorie festzuhalten. Eine Trajektorie T kann definiert werden als ein Set von n zeitlich geordneten Punkten $p_1 \rightarrow p_2 \rightarrow ... \rightarrow p_n$, wobei jeder Punkt p aus einer Koordinate (x,y) besteht und einen Zeitstempel t_i besitzt, für den gilt 1 <= i <= n. Der Aufzeichnung einer Trajektorie steht die einmalige Mitteilung des Standortes an den LBS-Server gegenüber, auch als *snapshot* (dt. Schnappschuss) bezeichnet. Ein solcher Schnappschuss entsteht beispielsweise bei einer kurzen Suchanfrage, wo sich ausgehend von der aktuellen Position das nächste Restaurant befindet [2].

Zusätzlich zu dieser räumlichen Komponente werden von der Applikation Informationen gespeichert, die Personen eindeutig von anderen unterscheidbar machen, wie beispielsweise eine ID oder Emailadresse. Weiterhin können auch bei Snapshots zeitliche Informationen wie beispielsweise ein *timestamp* (dt. Zeitstempel) enthalten sein, der die Aussage zulässt, wann eine Person sich wo befunden hat. Bei der Nutzung von LBS wird somit mindestens eine Information zur Person, eine ID, sowie die räumliche und zeitliche Information übertragen. Zudem können weitere applikationsbezogene Daten wie Sensorwerte übermittelt werden.

2.2. Definition der Privatsphäre

Alan Westin gilt heute als Begründer der modernen Datenprivatsphäre [11]. Bereits 1967 formulierte er eine Definition von Privatsphäre, die bis heute stets benutzt und zitiert wird [12]. Er definierte Privatsphäre als „*das persönliche Recht, die volle Kontrolle über Informationen über sich selbst zu haben und so auch die Entscheidung wann, wie und wie viel Information mit anderen geteilt wird*“ [12].

Viele Smartphone-Applikationen fragen nach der Installation zunächst Befugnisse ab. Dabei können solche Befugnisse die Verwendung des Mikrofons sein, als auch die Erlaubnis, den Standort abfragen zu dürfen. Werden diese Ermächtigungen nicht erteilt, sind Applikationen oft nicht in vollem Umfang nutzbar [13]. Es wird deutlich, dass die Selbstbestimmung bezüglich der eigenen Privatsphäre in der von Westin definierten Form für mobile Applikationen nur schwer möglich ist. Die Selbstbestimmung liegt eher in der Frage, ob die Applikation wirklich benötigt wird und daher die Befugnisse erteilt und dem Teilen der Daten zustimmt wird oder nicht. Hinzu kommt oft eine fehlende Kenntnis darüber, welche Daten von den Applikationen permanent gespeichert werden und was aus diesen Daten abgeleitet werden kann, vor allem bezogen auf Social Media Apps wie Facebook, Instagram oder Twitter [14]. Im Kontext von Datenbanksystemen und im Hinblick auf den massiven Anstieg von digitalen Datenmengen argumentierten Agrawal et al. bereits 2002, dass zukünftige Datenbanksysteme die Verantwortung für die Privatsphäre der Daten, die sie verwalten, übernehmen sollten [15]. Auch wenn die Autoren darlegen, dass die Verantwortung der Privatsphäre aus der Hand der Einzelpersonen in die Hände großer Firmen und Datenbanksysteme abgegeben wird, definieren sie Privatsphäre weiter nach Alan Westin. Die Pflicht solcher Datenbanksysteme liegt also darin, die Privatsphäre der in ihnen enthaltenen Daten über Einzelpersonen zu wahren. Dennoch steht fest, dass ein vollkommener Schutz der Privatsphäre sogar im alltäglichen Leben nicht möglich ist. Tritt eine Person in den öffentlichen Raum, verliert sie bereits ein Stück ihrer Privatsphäre, beispielsweise indem sie von anderen Personen gesehen wird [9].

Privatsphäre wird je nach Kontext von verschiedenen Faktoren beeinflusst. Diese Faktoren können aus der Sicht der Nutzenden definiert werden: [1]

- Ist die Übertragung der Information verschlüsselt oder nicht?
- Wie wird die Information benutzt? Wird sie intern genutzt oder an Dritte weitergegeben?
- Welche Information wird gespeichert? Handelt es sich um ein Koordinatenset? Ist die eigene Identität damit verknüpft? Ist die Information präzise oder ungenau?

2.3. Privatsphäre ortsbezogener Daten

Bevor untersucht werden kann, mit welchen Methoden die Privatsphäre von Daten gewahrt werden kann, soll Privatsphäre im Bezug auf ortsbezogene Daten beleuchtet werden. In diesem Kapitel erfolgt eine Definition von *Location Privacy*, der Privatsphäre für ortsbezogene Daten. Der Unterschied zwischen der Privatsphäre eines einzelnen Standortes wird der Privatsphäre einer Trajektorie gegenübergestellt.

Ein Problem beim Schutz der Privatsphäre des Standortes ist die ungenaue Definition dieser Privatsphäre, beziehungsweise die Unklarheit, welche Attribute geschützt werden müssen [1]. Westin hat Privatsphäre definiert als Recht einer Person, die Veröffentlichung der persönlichen Information kontrollieren und verhindern zu können. Diese Definition ist im Bezug auf die alltäglichen mobilen Applikationen zu allgemein: wird der Jogging-App der Zugriff auf den eigenen Standort verweigert,

fehlt der App die benötigte Information, um ihre Funktion auszuüben. Eine uneingeschränkte Nutzung von Applikationen muss also mit dem Schutz der persönlichen und standortbezogenen Daten kombiniert werden.

Beresford und Stajano definieren die Privatsphäre des Standortes als „*die Möglichkeit zu verhindern, dass die eigene derzeitige oder vergangene Standortinformation von anderen in Erfahrung gebracht werden kann*“ [3]. Wird diese Definition mit der Argumentation von Agrawal et al. kombiniert, liegt die Verantwortung der Privatisierung bei den Datenbanksystemen und so bei den Applikationen. Folglich müssen die Applikationen sicherstellen, dass die Übertragung oder Speicherung der Daten privatisiert erfolgt, sodass keine Standortinformationen über Personen in Erfahrung gebracht werden können. Um einen Dienst nutzen zu können, wird zwar weiterhin die eigene Standortinformation mitgeteilt, dennoch vertrauen die Nutzenden darauf, dass diese Standortinformation privatisiert weitergegeben und gespeichert wird und so nicht auf die eigene Identität zurückgeführt werden kann.

Privatsphäre des Standortes kann folglich als Schutz der drei in Kapitel 2.1 genannten Attribute UserID, räumliche und temporale Information definiert werden [1]. Einzelne Standorte beziehungsweise Trajektorien stellen dabei die räumliche Information dar. Sammelt eine Applikation einzelne Standorte ist wichtig, dass aus der Vielzahl der einzelnen Standortinformationen keine POIs wie Wohn- oder Arbeitsadressen abgeleitet werden können. Werden Trajektorien von Personen aufgezeichnet gilt zu beachten, dass die Trajektorie selbst, sowie deren Start- und Endpunkte sensible Information darstellen können und zu schützen sind. Dabei bedeutet Schutz vor Enthüllung, dass die Daten zwar übermittelt und auch gespeichert werden, aber in einer solchen Form, dass aus dieser Information nicht abgeleitet kann, welche Orte eine bestimmte Person wann oder wie häufig aufgesucht hat. Das heißt entweder, dass die Information von der Identität der Person getrennt werden muss oder, wenn dies nicht möglich oder gewünscht ist, die ortsbezogene Information der Person unpräzise gemacht werden muss.

2.4. Möglichkeiten der Re-Identifikation

Eine Problematik beim Erhalt der Privatsphäre besteht darin, dass private Informationen durch verschiedene Weisen direkt oder indirekt entschlüsselt werden können. Je nach Applikation werden unterschiedliche Daten und Attribute über Personen gespeichert, sei es der Geburtstag, die Adresse oder weitaus sensiblere Informationen wie die Sozialversicherungsnummer. Bestimmte Informationen können Personen eindeutig identifizieren. Eine solche Information wird als sogenannter *identifier* (dt. Erkennungsmarke) bezeichnet. Dieser Identifier kann beispielsweise die Sozialversicherungsnummer sein, die für jede Person einzigartig ist und so eindeutig auf diese zurückzuführen ist. Als *Quasi-Identifier* (QI) werden Informationen bezeichnet, die durch Kombination von mehreren Informationsstücken eine Person identifizieren können. Beispielsweise wird mit hoher Wahrscheinlichkeit die Kombination aus Geburtstag und Wohnadresse zu einer einzelnen Person führen [16]. Sweeney konnte 2002 nachweisen, dass 87% der Einwohnenden der USA durch QI-Attribute identifiziert werden können [17]. Die Arbeit von Douriez et al. [18] hat Taxifahrten in New York

analysiert und konnte mit Hilfe der Kombination aus Zeit und Standort einzelne Taxen und so deren Fahrende identifizieren.

Linking attacks sind Angriffe auf einen Datensatz, bei denen das Wissen aus mehreren Quellen kombiniert wird. So kann es für Angreifende möglich sein, selbst aus eigentlich privatisierten Datensätzen sensible Daten zu entnehmen [16], wenn die im privatisierten Datensatz fehlenden oder ungenauen Angaben in einem anderen Datensatz zu finden sind. Um wirklich sicherzugehen, dass ein Datensatz ausreichend privatisiert ist, müsste im Optimalfall also ein Angriff auf diesen ausgeübt werden, mit einer vorherigen Recherche, welche anderen Quellen öffentlich zu finden sind. In der vorliegenden Arbeit wird das Thema Attacken aus Kapazitätsgründen nicht betrachtet. Weiterführende Arbeiten zu diesem Thema sind die in Kapitel 3 vorgestellte Studie zu Taxi-Daten, oder die Re-Identifizierung des Gouverneurs William Weld durch k-anonymity-Begründerin Sweeney in [17].

Bezogen auf den vorliegenden Nextbike-Datensatz ist die enthaltene ortsbezogene Information von besonderem Interesse. Eine Re-Identifikation von Personen kann durch Verknüpfung der im Datensatz enthaltenen Standortinformation möglich sein. Der Nextbike-Datensatz beinhaltet die Start- und Endpunkte der Ausleihen und ordnet diese für jede gefahrene Route einer Kunden-ID zu. Stellt eine Person regelmäßig ein Nextbike an demselben Ort ab, befindet sich eine Vielzahl der gespeicherten Endpunkte der Ausleihen für diese Person gruppiert an einem Ort. Diese Gruppierung von Punkten lässt für Angreifende Rückschlüsse zu, dass die Person vermutlich an diesem Ort wohnt oder arbeitet. Wenngleich der Nextbike-Datensatz die gefahrenen Routen der Ausleihen nicht enthält, können solche Trajektorien Bewegungsmuster einzelner Personen verraten.

Abschließend sollen zwei hypothetische Beispiele erläutern, wie die Kombination der drei für die Privatsphäre des Standortes wesentlichen Punkte UserID, räumliche und zeitliche Information bei Veröffentlichung eines Datensatzes genutzt werden können:

- Szenario 1: Eine Person nutzt eine Fahrrad-Navigationsapp, bei der sie sich registriert hat, um mit dem Rad zur Arbeit zu fahren. Die App speichert die gefahrenen Routen und ordnet sie, mittels ID oder Emailadresse, einer Person zu. Gelangen diese Daten in die Hände von Angreifenden, kann das Muster abgeleitet werden, dass die Person jeden Morgen um 8:00 Uhr dieselbe Strecke fährt. Je nachdem wo die Person die Navigation der Strecke startet, kann anhand der Start- und Endpunkte ihrer Route, ihr Wohnort und ihre Arbeitsstelle ermittelt werden.

- Szenario 2: Eine Stadt plant die Umgestaltung von Flächen und Straßen und möchte mehr Radwege anlegen. Sie möchte zur Planung einbeziehen, wo diese Radwege besonders hilfreich sein könnten beziehungsweise wo sie explizit fehlen. Die Stadt kann die gespeicherten Daten der Fahrrad-Navigationsapp nutzen, die explizit für Fahrradfahrende Routen vorschlägt, um zu analysieren, wo viele Personen mit dem Fahrrad entlang fahren, es aber womöglich nicht ausreichend Radwege gibt.

Um Szenario 1 zu verhindern, aber Szenario 2 dennoch zu ermöglichen, müssen die Daten der App so privatisiert werden, dass die Nutzenden und ihre Routen nicht identifiziert, die Daten aber dennoch für Analysen genutzt werden können.

Das folgende Kapitel beleuchtet diese Herausforderung der Privatisierung, zwischen Nützlichkeit und Privatsphäre einen guten Kompromiss zu finden. Im darauf folgenden Kapitel wird zunächst erläutert, wie Privatsphäre gemessen werden kann, bevor anschließend verschiedene Privatisierungsmechanismen vorgestellt werden, um Datensätze oder auch Daten während der Verwendung von Applikationen zu privatisieren.

2.5. Privacy Utility Trade-off

Werden Daten der Öffentlichkeit zugänglich gemacht, werden sie gleichzeitig potentiellen Angreifenden zur Verfügung gestellt. Dennoch können aus verschiedenen Applikationen und Szenarios gesammelte Daten in verschiedenen Kontexten sehr nützlich sein: sei es für die Auswertung von persönlichen Leistungen oder auf höherer Ebene wie Stadtplanung. Eine ungeschützte Veröffentlichung von Standortdaten kann dazu führen, dass Verhaltens- und Bewegungsmuster einzelner Personen, aber auch deren Adressen oder andere sensible Information enthüllt werden [16]. Der zu schaffende Rahmen muss also zum einen sicherstellen, dass der Schutz der Privatsphäre gewährleistet ist, beispielsweise durch Veränderung oder Verschleierung der Daten. Zum anderen gilt es, durch den Privatisierungsprozess den Daten ihre Nützlichkeit nicht zu entziehen, sodass sie weiter für Data Mining und Informationsbeschaffung genutzt werden können.

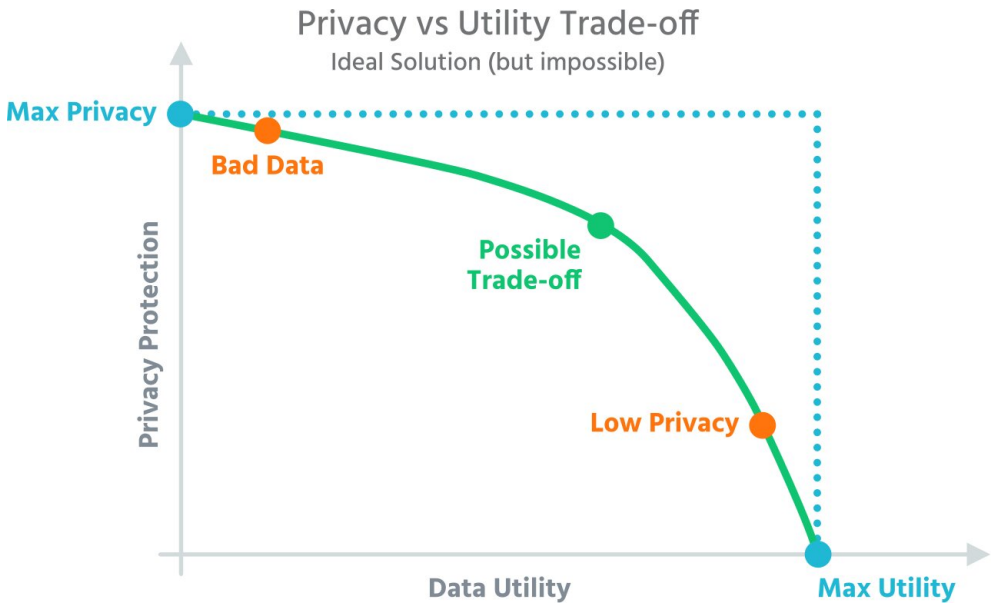


Abbildung 2.2.: Privacy Utility Trade-off, Quelle: [19]

Die Gegenspieler Privatsphäre und Nützlichkeit finden ihren Kompromiss im sogenannten *Privacy Utility Trade-off*. In Abbildung 2.2 wird die Problematik verdeutlicht: Der Fall maximale Privatsphäre und maximaler Nutzen der Daten existiert nicht. Der Begriff „Bad Data“ bedeutet, dass die Daten zwar stark verschleiert wurden, die daraus resultierende Nützlichkeit der Daten aber wenig bis keine ist. Beispielsweise lässt ein Datensatz von Patienten, bei dem die Eigenschaft „Alter“ der Personen in zu grobe Intervalle aufgeteilt wurde, keine Rückschlüsse auf das Vorkommen einer Krankheit in verschiedenen Altersstufen mehr zu. Der Fall „Low Privacy“ auf der anderen Seite repräsentiert einen schwach privatisierten Datensatz, der durch wenig Veränderung der Daten

weiterhin eine hohe Nützlichkeit aufweist. Dieser wenig privatisierte Datensatz erhöht die Chancen für Angreifende, sensible Daten zu entnehmen. Der mögliche Kompromiss zwischen Nützlichkeit und Privatisierung muss für jeden Datensatz eigens gefunden werden und befindet sich zwischen diesen beiden Szenarien.

Die Herausforderung der Privatisierung eines Datensatzes besteht folglich aus kleinteiligen Faktoren:

- Finden eines passenden und effektiven Mechanismus zur Privatisierung des Datensatzes
- Finden guter Parameter zum Erhalt von Privatsphäre und Nutzen der Daten
- Einbeziehung von zusätzlich verfügbaren (öffentlichen) Ressourcen von Angreifenden
- Abschließende Evaluation von Privatsphäre und Nützlichkeit des Datensatzes

2.6. Bestimmung des Privatsphäre-Levels

Um Datensätze privatisieren zu können, muss definiert sein, wann ein ausreichendes Level an Privatsphäre erreicht ist. Hierfür wird ein Maß benötigt, dieses Level zu definieren. Methoden zur Privatisierung von Datensätzen haben folglich das Ziel, dieses Level zu erreichen. Im Folgenden werden zwei der am weitesten verbreiteten Auffassungen von Privatsphäre-Levels vorgestellt.

2.6.1. K-anonymity

Ist ein Datensatz k-anonym, unterscheiden sich alle darin vorkommenden Personen von mindestens k-1 anderen nicht. Mit der Auswahl von k wird also ein Risikoschwellenwert gesetzt, der selbst gewählt werden kann. Sweeney und Samarati beschreiben diesen Ansatz als „*die Bedingung, dass [..] jede Kombination von Werten von Quasi-Identifizierern uneindeutig auf mindestens k Individuen passt*“ [20]. Für einen k-Wert von 10 muss also die Kombination von Werten, beispielsweise die Eigenschaft „Krankheit“ mit dem Wert „HIV“ bei Personen im „Alter von 20 bis 25 Jahren“, in einem Datensatz auf mindestens 10 Personen passen, sodass eine einzelne Person nicht eindeutig identifizierbar ist.

Das Ziel der k-anonymity wird durch verschiedene Anonymisierungsmechanismen erreicht, darunter Generalisierung, *Cloaking*- oder *Geomasking*-Ansätze (siehe Kapitel 2.9.2). Die Ansätze setzen Privatsphäre um, indem sie den Datensatz in Äquivalenzklassen gleicher Merkmale einteilen. Welche Attribute zu Äquivalenzklassen zusammengefasst werden, ist dabei frei wählbar (beispielsweise Alter, Postleitzahlen, Geschlecht et cetera). Eine Äquivalenzklasse muss mindestens k Einträge besitzen, damit die ihr zugeteilten Einträge geschützt sind.

Ein Eintrag eines k-anonymisierten Datensatzes hat eine Re-Identifikationswahrscheinlichkeit von $\frac{1}{k}$ [21]. Je höher der k-Wert gewählt wird, umso größer ist das zu erreichende Level an Privatsphäre, da mehr Personen voneinander un-unterscheidbar werden. Gleichzeitig kann bei höherem k-Wert

mehr Information verloren gehen, da die Daten stärker verändert werden müssen, um die Ununterscheidbarkeit der Personen zu erreichen. Ist die tatsächliche Re-Identifikationswahrscheinlichkeit aber deutlich niedriger als $\frac{1}{k}$, werden eventuell Daten unnötigerweise verzerrt [21].

Der k-Wert kann bei der Privatisierung als Ziel genutzt werden, welches ein LPPM erreichen soll. Gleichzeitig kann der Wert als Maß genutzt werden, um mit Metriken auszuwerten, welches Level an Privatisierung erreicht wurde.

2.6.2. Differential Privacy

Der Ansatz *Differential Privacy* stammt aus dem Bereich der statistischen Datenbanken. Das Ziel ist es, die Daten einer einzelnen Person zu schützen, indem aggregierte Information aus der Datenbank veröffentlicht wird [22]. *Differential Privacy* bezeichnet die Bedingung, dass die Wahrscheinlichkeit, dass eine Anfrage an eine Datenbank D einen Wert v zurück gibt, verglichen mit der Wahrscheinlichkeit, bei einer Anfrage an eine Datenbank D' denselben Wert zurückzubekommen, in einem festgelegten Bereich von e^{ϵ} liegt [22].

Das Erreichen des Privatsphäre-Levels erfolgt durch das Erzeugen von synthetischen Daten beziehungsweise durch das Hinzufügen von Rauschen zur Ausgabe der Datenbank. Um Daten zu synthetisieren, werden zunächst die Originaldaten analysiert, um die statistische Verteilung dieser zu erfassen. Anschließend werden aus diesen Analysen synthetische Daten erzeugt [23].

Im Bezug auf Standortdaten ist hier auch der Begriff *Geo-Indistinguishability* (dt. Geo-Un-unterscheidbarkeit) zu erwähnen. Diese wird definiert als Schutz des Standortes einer Person in einem Radius von r, wobei das Level an Privatsphäre von r abhängig ist. Die *Geo-Indistinguishability* wird im Zusammenhang mit *Differential Privacy* oft als Metrik verwendet, um das Maß an Privatsphäre in einem Datensatz zu messen [1].

In dieser Arbeit erfolgt eine Analyse des Leipziger Fahrradklimas, daher soll der Datensatz nach der Privatisierung so originalgetreu wie möglich bleiben. Da auf *Differential Privacy* basierende Methoden mit Hinzufügen von Rauschen oder dem Erstellen von synthetischen Daten arbeiten, werden die Daten zum Erreichen des Privatsphäre-Ziels verfälscht. Daher werden Methoden mit dem Ziel der *Differential Privacy* bei der Vorstellung der LPPM im weiteren Verlauf der Arbeit ausgeschlossen.

2.7. Metriken zur Evaluation der Privatsphäre

Nachdem ein Datensatz privatisiert wurde, gilt es die erreichte Privatsphäre der Daten zu evaluieren. Hierzu werden in diesem Kapitel Metriken vorgestellt. Bislang existiert kein Standard, um das Level an Privatsphäre zu messen [1]. Viele Metriken basieren auf einem Angreifermodell und der Annahme, dass Angreifende Erfolg bei der Enthüllung von Daten haben werden. Darunter die Metriken *Certainty* (dt. Sicherheit), *Correctness* (dt. Korrektheit), *Information Gain or Loss* (dt. Informationsgewinn oder -verlust). Ein Angreifermodell beschreibt dabei die Charakteristika und Möglichkeiten die Angreifende haben, um sensible Informationen zu erlangen. Angreifende mit mehr

Vorwissen oder Ressourcen sind nach solchen Modellen eher in der Lage, Erfolg bei der Enthüllung von Information zu haben [24]. Während diese Metriken eigens zur Feststellung des Privatsphäre-Levels entwickelt wurden, können Parameter von Methoden zum Erreichen des Privatsphäre-Levels bereits als Metrik angesehen werden, so beispielsweise der k-Wert bei Methoden mit dem Ziel des Erreichens von k-anonymity.

Es gibt keinen allgemeinen Konsens, welche Aspekte eine Privatsphäre-Metrik betrachten muss: zum einen soll sie reflektieren, wie schwer es für Angreifende ist, Erfolg zu haben, zum anderen soll die Metrik einschätzen, wie effektiv der Erfolg von Angreifenden ist, einzelne Individuen zu identifizieren. Andere Metriken berechnen Wahrscheinlichkeiten, also beispielsweise die Wahrscheinlichkeit, dass Angreifende eine bestimmte Person identifizieren können [24]. Weitere Studien führen zusätzlich zur Anonymisierung eines Datensatzes eine Attacke auf diesen anonymisierten Datensatz durch (siehe [18] und [14]). Hierfür muss betrachtet werden, welche Datensätze zusätzlich öffentlich zugänglich sind oder welche Information durch Beobachtungen gesammelt werden kann.

Obgleich der unterschiedlichen Möglichkeiten Privatsphäre zu evaluieren, teilen sich die Metriken Charakteristika, die Wagner und Eckhoff [24] in vier Punkten zusammengefasst haben:

1. Ziele der Angreifenden: Die Metriken betrachten die Informationen, die bei einem Angriff enthüllt werden können. Dies können Informationen über Identität oder Eigenschaften von Personen sein.

2. Möglichkeiten der Angreifenden: Die Metriken erstellen ein Angreifermodell unter der Annahme, dass Angreifende, die mehr Wissen oder Ressourcen zur Verfügung haben, mehr Daten enthüllen können.

3. Datenquelle: Die Metriken betrachten, welche Daten geschützt werden müssen und wie Angreifende auf diese zugreifen können (Veröffentlichte Daten oder Daten durch Beobachtungen von Personen).

4. Dateninput um Privatsphäre zu berechnen: die Verfügbarkeit von Daten oder genauen Schätzungen bestimmen darüber, ob eine Metrik benutzt werden kann.

Die Ausgabe einer Metrik bezieht sich auf die Eigenschaft beziehungsweise das Attribut, deren Level an Privatsphäre eine Metrik misst. Wagner und Eckhoff teilen daher Metriken anhand ihres Outputs ein, eine Einteilung, die auch in der Studie von Liu et al. [1] vorgenommen wurde. Dabei gilt zu beachten, dass die Grenzen zwischen den Metriken oft schwammig und die Metriken womöglich nicht klar trennbar sind [24]. Es folgt eine kurzer Überblick, der einige der am meisten verwendeten Metriken vorstellt und nicht vollständig ist.

Certainty / Uncertainty (dt. Gewissheit / Ungewissheit):

Privatsphäre-Metriken dieser Kategorie machen eine eine Angabe zur Gewissheit. Sie berechnen die Gewiss- oder Ungewissheit, dass Angreifende eine eindeutige Antwort auf ihre Suche finden. Eine eindeutige Antwort kann sowohl eine Identität einer einzelnen Person, als auch räumliche oder temporale Information sein [1]. Hohe Ungewissheit korreliert hierbei mit hoher Privatsphäre. Je weniger gewisse Information Angreifende haben, desto sicherer ist die Privatsphäre der Person

beziehungsweise der Daten [24]. Um eine Angabe zur Gewissheit machen zu können, kann die *Anonymity Size Set* (ASS) (dt. Größe des Anonymitäts-Sets) berechnet werden. Dieses Set für eine Person p ist das Set von Personen, von dem Angreifende p nicht unterscheiden können. Dabei ist die Größe des Anonymitäts-Sets (AS) gleich dem Level an Privatsphäre. Es gilt also:

$$privacy = |AS| \tag{2.1}$$

wobei $|AS|$ die Größe des Anonymitäts-Sets ist. Kritisiert wird dieser Ansatz, da er nur davon abhängig ist, wie viele Personen sich in der Datenbank befinden. Er bezieht vorheriges Wissen von Angreifenden nicht mit ein.

Andere Ansätze dieser Kategorie benutzen „Entropie“ und „Shannon-Information“, um die Ungewissheit zu messen. Allgemein betrachtet misst Entropie die Unsicherheit bei der Ermittlung eines Wertes einer beliebigen Variable. Als Privatsphäre-Metrik verknüpft Entropie jedes Mitglied x_i eines Anonymitäts-Sets X mit einer Wahrscheinlichkeit p(x) für unterschiedliche Szenarien, beispielsweise dass dieses Mitglied das Zielobjekt von Angreifenden ist [24]. Bei Metriken dieser Kategorie gilt zu beachten, dass auch in ungewisser Information, also in Schätzungen, die richtigen beziehungsweise gesuchten Daten enthalten sein können. Tritt dieser Fall ein, kann es trotz eines hohen gemessenen Privatsphäre-Levels zu Datenverlust kommen.

Information Gain / Loss (dt. Informationsgewinn / -verlust):

Mit dieser Methode soll die Information gemessen werden, die Angreifende erlangen können. Dabei wird angenommen, dass das Privatsphäre-Level höher ist, je weniger Information Angreifende erlangen können [1]. Umgekehrt kann mit diesem Ansatz die Privatsphäre gemessen werden, die eine Person verloren hat.

Methoden dieser Kategorie sind ähnlich zu *Certainty*-Ansätzen, beziehen aber im Gegensatz zu diesen das bisherige Wissen von Angreifenden mit ein. Sie finden in einem breiten Bereich von Domänen bereits Verwendung, unter anderem in Social Networks, Datenbanken oder Kommunikationssystemen. Die meisten unter dieser Kategorie zu findenden Ansätze arbeiten ebenfalls mithilfe der Informationstheorie, beispielsweise indem die Menge der des Systems entwichenen Information gemessen wird (*Amount of Leaked Information*) oder indem quantifiziert wird, wie viel Information zwischen zwei zufälligen Variablen geteilt wird (*Mutual Information*). Hier wird meist die wahre Verteilung der Daten mit der Beobachtung des Gegners beziehungsweise der Verteilung im privatisierten Datensatz betrachtet [24].

Correctness / Error (dt. Korrektheit / Fehler):

Correctness berechnet die Erfolgsrate von Angreifenden beziehungsweise die Fehlerquote. Dabei wird die Wahrscheinlichkeit berechnet, dass Angreifende bei der Enthüllung von Daten Erfolg haben. Alternativ kann bemessen werden, wie hoch der prozentuale Anteil an Erfolgen bei einer großen Anzahl von Versuchen ist [24]. Weiter kann der erwartete Fehler beziehungsweise die Distanz mit distanzbasierten Metriken berechnet werden. Hierbei wird die Distanz zwischen der

wahren Information und der von Angreifenden errechneten Information berechnet [1]. Je höher der Fehlerwert, desto schwerer ist es für Angreifende an Information zu kommen und desto höher ist das Level der Privatsphäre.

Data Similarity (dt. Ähnlichkeit der Daten):

Mit diesem Ansatz kann zum einen die Ähnlichkeit der Daten in einem veröffentlichten Datensatz gemessen werden, beispielsweise durch Formung von Äquivalenzklassen. Alternativ kann die Ähnlichkeit zwischen zwei Datensätzen berechnet werden, beispielsweise dem Originalen und dem Privatisierten. Diese Metrik arbeitet ohne im Vorhinein ein Angreifermodell zu erstellen. Ähnlichkeit kann hierbei die Frequenz eines Wertes, die numerische Ähnlichkeit von Werten oder auch die Variation oder das Fehlen von Variation in einem veröffentlichten Datensatz sein [24]. Der beliebteste Ansatz dieser Kategorie ist *k-anonymity* und kann unter der Bedingung eingesetzt werden, dass ein Datensatz D in Äquivalenzklassen EQ mit mindestens k Einträgen pro Äquivalenzklasse eingeteilt werden kann. Die Privatsphäre entspricht dann k, wenn für alle Äquivalenzklassen gilt, die Anzahl der Einträge in diesen Klassen ist größer oder gleich k, also: $privacy \equiv k$, wenn $\forall EQ : |EQ| \geq k$ [24]. Dieser Ansatz ähnelt der Bildung von Anonymitäts-Sets aus der Kategorie *Certainty*.

Dem k-anonymity Ansatz allein wird vorgeworfen, bei hochdimensionalen Daten insuffizient zu sein und nicht vor Attribut-Enthüllung schützen zu können. Daher wurden Methoden entwickelt, um diesen Ansatz zu verfeinern, wie beispielsweise *(α, k)-anonymity*. Andere weiterentwickelte Ansätze zu k-anonymity werden in Kapitel 2.9.2 vorgestellt. Um der Enthüllung sensibler Attribute vorzubeugen, erweitert (α, k)-anonymity den k-anonymity Ansatz um einen Faktor α. Dieser stellt die Bedingung dar, dass die Frequenz eines sensiblen Wertes kleiner als α sein muss, wobei α ein selbst gewählter Schwellenwert zwischen 0 und 1 ist [25]. So wird überprüft, dass kein einzelnes Attribut in einer Äquivalenzklasse dominant ist [24]. Umso höher der gewählte Wert für α, umso mehr gleiche Werte können sich in einer Äquivalenzklasse befinden. Sei s in $|EQ, s|$ die Anzahl der sensiblen Werte in einer Äquivalenzklasse EQ, so errechnet sich α durch $\frac{|EQ,s|}{|EQ|}$. Für die gesamte Metrik der Data Similarity ergibt sich daher:

$$privacy_{AK} \equiv (\alpha, k), \text{wenn } \forall EQ : |EQ| \geq k \text{ und } \frac{|EQ, s|}{|EQ|} \leq \alpha \tag{2.2}$$

Re-Identifikationswahrscheinlichkeit

Ein Datum in einem k-anonymisierten Datensatz hat eine Wahrscheinlichkeit von $\frac{1}{k}$ re-identifiziert zu werden [21]. Hohe Werte für k verringern das Risiko der Re-Identifikation, führen aber unter Umständen zu einer höheren Verzerrung der Daten. Ist die tatsächliche Re-Identifikationswahrscheinlichkeit aber deutlich niedriger als $\frac{1}{k}$, können Daten unnötigerweise verzerrt werden [21].

Einträge in einem Datensatz, die dieselben Werte als Quasi-identifier haben, können in Äquivalenzklassen gruppiert werden. Sei f die Anzahl der Einträge in einer Äquivalenzklasse, die dieselben Quasi-identifier wie eine Person haben, so ist die Wahrscheinlichkeit der Re-Identifizierung dieser Person $\frac{1}{f}$ [21]. Wird der Fall betrachtet, dass eine Person im privatisierten Datensatz identifiziert werden soll, von der bekannt ist, dass sie in diesem enthalten ist, entspricht $\frac{1}{f}$ ungefähr $\frac{1}{k}$. Der Vergleich der für den Datensatz erreichten Re-Identifikationswahrscheinlichkeit mit der angestrebten von $\frac{1}{k}$ kann Rückschlüsse darüber geben, ob das erreichte Level an Privatsphäre über, unter oder gleich dem angestrebten Level von Privatsphäre ist [21]. Die erreichte Re-Identifikationswahrscheinlichkeit des Datensatzes ist höher als die angstrebte, wenn die Äquivalenzklassen weniger als k Einträge besitzen, die erreichte Privatsphäre ist also niedriger als erwünscht. Dementsprechend ist die Re-Identifikationswahrscheinlichkeit niedriger, wenn Äquivalenzklassen mehr als k Einträge besitzen.

Sei F die Anzahl der Äquivalenzklassen in einem zweiten, öffentlich zugänglichen Datensatz. Soll eine beliebige Person identifiziert werden, wäre das Identifikationsrisiko dieser Person $\frac{1}{F}$. Unter der Annahme, dass ein Gegner die Äquivalenzklasse mit den wenigsten Einträgen wählen wird, da sie die größte Wahrscheinlichkeit bietet, eine Person erfolgreich zu enthüllen, liegt die Wahrscheinlichkeit einer Enthüllung bei $\frac{1}{min(F)}$ [21].

2.8. Metriken zur Evaluation der Nützlichkeit

Neben der Betrachtung der erreichten Privatsphäre der Daten muss der Grad an Nützlichkeit ermittelt werden, den die Daten nach der Privatisierung besitzen. Der sogenannte *Privacy Utility Trade-off* beschreibt die Problematik, dass ein höheres Level an Privatsphäre eine Verringerung der Datenqualität mit sich bringt (siehe Kapitel 2.5). Um die Nützlichkeit eines Datensatzes zu bestimmen, gibt es keine standardisierte Vorgehensweise. In Studien werden oft eigene Metriken entwickelt, um die Datenqualität nach der Privatisierung zu messen [26]. In der Statistik werden verbreitete Methoden wie KL-Divergenz oder L_1-Norm verwendet, um die Veränderung der Verteilung der Daten zu bestimmen. Weiter kann die Nützlichkeit der Daten in Bezug auf spezifische Szenarien bestimmt werden. Dieser Ansatz ist allerdings schwierig zu praktizieren, da bei Veröffentlichung eines Datensatzes oft nicht bekannt ist, zu welchem Zweck er benutzt werden soll [27]. Grundsätzlich gilt, je mehr Transformation der Daten stattfinden muss, um diese zu privatisieren, umso mehr Information und somit auch Qualität der Daten geht verloren. Im Umkehrschluss bedeutet dies, je weniger Information verloren geht, umso besser ist die Datenqualität. Suppression, also das Entfernen eines Eintrages aus einem Datensatz, ist dabei die teuerste Operation im Sinne des Informationsverlustes [28].

Discernability Metric (C_{DM})

Diese von Bayardo und Agrawal [26] vorgestellte Metrik untersucht, wie un-unterscheidbar ein Eintrag in einem Datensatz von einem anderen ist. Dabei wird jedem Eintrag ein Strafwert zugewiesen [28]. Die Metrik geht von der Annahme aus, dass große Äquivalenzklassen mehr Informationsverlust repräsentieren, da die Klassen gröbere Werte enthalten, je mehr Einträge generalisiert wurden.

Der Strafwert eines Eintrages entspricht daher der Größe der Äquivalenzklasse, in der der Eintrag enthalten ist. Wurde ein Eintrag bei der Privatisierung aus dem Datensatz entfernt, entspricht der Strafwert der Größe des Originaldatensatzes.

Diese Kostenmetrik zählt den Informationsverlust, der aufgrund von Generalisierung oder Suppression entsteht [26]. Dabei repräsentieren niedrige Werte kleine Äquivalenzklassen und somit wenig Informationsverlust. Je höher der Wert der Metrik, desto höher der Informationsverlust [27]. Der Wert lässt sich berechnen durch:

$$C_{DM}(T') = \sum_{\forall EQ s.t. |EQ| \geq k} |EQ|^2 + \sum_{\forall EQ s.t. |EQ| < k} |T| * |EQ| \tag{2.3}$$

wobei T die Originaltabelle und $|T|$ die Anzahl der Einträge in dieser repräsentiert und $|EQ|$ die Größe der Äquivalenzklasse (Anzahl der Einträge) nach der Anonymisierung. Der zweite Summand bezeichnet dabei den Strafwert für einen unterdrückten Wert [26].

Average Equivalence Class Size (C_{AVG})

Bei dieser Metrik wird die Verteilung der Werte in den einzelnen Äquivalenzklassen untersucht. Der beste zu erreichende Fall ist eine gleichmäßige Verteilung aller Einträge des Datensatzes auf die Äquivalenzklassen. Es ergäben sich gleich große Äquivalenzklassen der Größe k. Das optimale Ergebnis der Metrik ist also 1 [27]. Der Wert der Metrik für eine privatisierte Tabelle T' ergibt sich aus:

$$C_{AVG}(T') = \frac{|T|}{|EQs| * k} \tag{2.4}$$

wobei T die Originaltabelle und $|T|$ die Anzahl der Einträge in dieser repräsentiert, $|EQs|$ die Anzahl der erstellten Äquivalenzklassen und k den gewählten k-Wert der Privatisierungsmethode.

2.9. Location Privacy Preserving Mechanisms (LPPM)

Im folgenden Abschnitt werden verschiedene Methoden zur Privatsphäre-Erhaltung vorgestellt. Am Ende dieses Kapitels befindet sich ein Überblick aller vorgestellten Mechanismen in Tabellenform.

Location privacy preserving mechanisms können nach verschiedenen Szenarien in Bereiche eingeteilt werden, die im Text näher erläutert werden [29]:

- Welches Ziel soll erreicht werden?
- Zu welchem Zeitpunkt soll die Privatisierung angewandt werden?

- Sollen die Daten von einer dritten Instanz (beispielsweise von einem Server) oder bereits in der verwendeten Applikation privatisiert werden?

In der Literatur werden die verschiedenen Methodiken in Kategorien eingeteilt. Shokri et al. teilen LPPM in zwei Kategorien ein: Anonymisierungs- und Vernebelungsmechanismen [29]. Um einen breiteren Überblick zu schaffen, werden in dieser Arbeit auch die Kategorien Kryptographie und Informationsreduktion vorgestellt. So erfolgt analog zu Liu et al. [1] eine Einteilung in vier Kategorien: Vernebelungs-, Anonymisierungs-, kryptographische und Reduktionsmechanismen [1]. Diese Einteilung erfolgt auf Basis der Ziele, die die Methoden erreichen sollen: Vernebelungsmechanismen sollen die gegebene Information verschleiern, also deren Präzision verringern, um die Gewissheit der Angreifenden zu verringern. Anonymisierungsmechanismen brechen die Relation zwischen der Identität einer Person und der (Standort-)information, um die Person un-unterscheidbar zu machen. Kryptographische Mechanismen sollen verhindern, dass Angreifende bei der Übertragung jedwede Information abfangen und verwenden können. Durch die Reduktion der Anzahl von Standortabfragen soll die Menge an Information, die generiert und übertragen wird, verringert werden [1].

Diese Methoden können an verschiedenen Zeitpunkten Anwendung finden. Die Unterscheidung erfolgt nach *online* und *offline*. Online beschreibt dabei die Modifizierung der Daten oder der Anfrage einer einzelnen Person während der Benutzung eines Services. Bei einem offline-Ansatz wurden bereits Daten von vielen Personen an einen Server beziehungsweise eine Datenbank übertragen. Es wird also der gesamte Datensatz privatisiert. Werden die Daten nicht von der Applikation selbst, sondern von einer dritten Instanz privatisiert, wird diese Instanz als sogenannter *trusted server* (dt. vertrauenswürdiger Server) oder „Anonymisierungsserver“ bezeichnet. Bei der Privatisierung durch eine dritte Instanz wird die Verarbeitung als *zentral* bezeichnet. Wird die Modifizierung der Daten bereits auf dem Smartphone der Person ausgeführt, arbeitet die Methodik *dezentral* oder auch *verteilt*. Wichtig bei letzterem Ansatz ist die Tatsache, dass zur Privatisierung nur die Daten der einzelnen Person und so andere Mechanismen zur Privatsphäre-Erhaltung zur Verfügung stehen [29].

Der Hauptunterschied bei der Privatisierung zwischen kontinuierlich übertragenen LBS-Daten und einem Datensatz von LBS-Daten liegt in zwei Punkten: Skalierbarkeit und globale Optimierungsmöglichkeit [2]. Daten, die kontinuierlich an den LBS-Server gesendet werden, müssen in Echtzeit, also online, privatisiert werden, während die Privatisierung offline an keinen Zeitdruck gebunden ist. Offline können also rechenintensivere Methoden verwendet werden. Zudem kann offline der gesamte Datensatz zur Privatisierung verwendet werden, was es erleichtert, Privatsphäre und gleichzeitig Nützlichkeit zu optimieren [2].

2.9.1. Kryptographische Mechanismen

Anfragen und Antworten von LBS werden über Kommunikationsnetzwerke zwischen Server und Client versendet. Auf diesen Kanälen können Angreifende lauschen beziehungsweise interferieren (sogenannte *man-in-the-middle* Attacke) [10]. Vor diesem Risiko können kryptographische Mechanismen schützen. Sie kommen daher bereits bei der Nutzung von LBS beziehungsweise bei der Übertragung der Daten zum Server zum Einsatz. Es handelt sich also um Methoden, die Anfragen und Antworten vom und zum Server vor Enthüllung schützen. Ihr Ziel ist es zu verhindern, dass Angreifende durch Lauschen Information abfangen können.

In der Literatur werden unter anderem drei Ansätze genannt:

1. *Private Information Retrieval* (PIR)
2. Aufteilung der geheimen (Standort-)information
3. Teilen eines gemeinsamen Geheimnisses zwischen zwei Parteien (*secret*)

Alle drei Ansätze vertrauen nicht darauf, dass die Kommunikationswege oder die Server die Daten geheim halten können.

PIR verfolgt das Ziel, Information aus einer Datenbank abzurufen, ohne dass diese weiß, um welche Information es sich handelt. Im LBS-Kontext kann folglich ein Client eine Anfrage senden und der Dienst diese beantworten, ohne zu wissen, welche Information der Client angefordert hat [10].

Der zweite Ansatz von Marias et al. [30] verfolgt die Idee, die angeforderte Information in mehrere Teile zu splitten und über verschiedene LBS-Server zu verteilen. Um die Information zu erlangen, muss ein Client dann mit verschiedenen Servern kommunizieren.

Die dritte Idee wird auf Applikationsebene angewandt und wurde unter dem Aspekt eines sozialen Netzwerkes entwickelt, welches darauf hinweist, wenn sich befreundete Personen in der Nähe aufhalten. Die Positionen der Personen wird nicht an den LBS-Server weitergegeben, sondern durch ein *secret* (dt. Geheimnis), welches sich beide Parteien teilen, direkt weitergegeben [31].

Ein Hauptproblem der kryptographischen Methoden ist die Komplexität und dementsprechend benötigte Rechenleistung beziehungsweise das Vorhandensein entsprechender Server. Die Ausführung kryptographischer Privatisierungsmethoden kann zu einem serverseitigen Overhead führen, der die Performanz des gesamten Systems beeinflusst [10]. Bislang gibt es noch keinen kryptographischen Ansatz, der sich in der Praxis durchgesetzt hat [1].

2.9.2. Anonymisierungsmechanismen

Das Ziel von Anonymisierungsmechanismen ist das Brechen der Relation zwischen Identität und (Standort-)information. Eine Person soll un-unterscheidbar von anderen werden [16]. Um Anonymität zu erreichen, wurden verschiedene Ansätze entwickelt, die im unteren Teil des Abschnittes erläutert werden. Ziel der meisten Verfahren dieser Kategorie ist das Erreichen von *k-anonymity*, wobei k eine selbst zu wählende Zahl ist (siehe Kapitel 2.6.1).

Generalisierung / Aggregation

Aggregation verfolgt das Ziel der Anonymität, indem QIs wie Geburtsjahr, Alter oder Start- und Endpunkte vergröbert werden und so deren Genauigkeit verringert wird. Indem beispielsweise exakte Werte auf ein Intervall abgebildet werden, werden Gruppen mit gleichen QIs gebildet. Einträge in einem anonymisierten Datensatz, die dieselben QIs haben, gehören derselben Äquivalenzklasse an und sind voneinander un-unterscheidbar [21]. Die Eigenschaften eines einzelnen Datums werden zwar ungenauer, sind aber dennoch konsistent zu ihrem Originalwert [16]. Beispiel: in einem Datensatz wird für Personen nicht mehr das genaue Alter gespeichert, sondern ein Intervall wie 18-24 Jahre, 25-35 Jahre und folgende. Diese Konsistenz zum Originalwert, auch als *faithfulness* (dt. Treue) bezeichnet, ist eine Problematik bei Generalisierungsansätzen, da die neu zugewiesene Wertspanne noch immer den Originalwert widerspiegelt. Dies ist auch für Angreifende nachvollziehbar.

Desweiteren ist das Erreichen von k-anonymity allein bei hochdimensionalen Daten nicht ausreichend, um einzelne, sensible Attribute in einer Äquivalenzklasse zu schützen. Sind bei Einträgen einer Äquivalenzklasse für ein Attribut nur wenige unterschiedliche Werte enthalten, kann k-anonymity allein nicht vor der sogenannten Attribut-Enthüllung schützen. Im schlimmsten Szenario besitzen alle Einträge der Äquivalenzklasse den gleichen Wert für das sensible Attribut, sodass Angreifende, können sie eine Person aufgrund von Vorwissen einer Äquivalenzklasse zuordnen, deren sensible Information sofort erfahren. Ist beispielsweise die Äquivalenzklasse Personen im Alter von 20-30 Jahren und ein sensibles Attribut Krankheit, dessen Wert für alle Personen der Äquivalenzklasse „HIV“ ist [24], können Angreifende, die wissen, dass ihre Zielperson 20 bis 30 Jahre ist, deren Krankheit sofort erfahren.

Um diesen Schwächen zuvor zu kommen, haben sich Verfeinerungen für k-anonymity ergeben, beispielsweise *l-diversity* und *t-closeness*. L-Diversity (dt. Diversität) soll sicherstellen, dass eine Äquivalenzklasse mindestens l Werte für sensible Attribute hat. T-closeness (dt. Nähe) soll gewährleisten, dass die Distanz der Verteilung eines sensiblen Wertes in einer Äquivalenzklasse und die Verteilung dieses Wertes im gesamten Datensatz einen Schwellenwert von t nicht überschreitet [2].

Aggregation beziehungsweise Generalisierung kann für nominale Werte wie beispielsweise Geschlecht, als auch diskrete oder kontinuierliche Attribute wie Alter oder Größe angewandt werden. Auch für Standortdaten kann Aggregation zur Privatisierung genutzt werden. Verschiedene Standorte, die in einer gemeinsamen Region liegen, können einem Zentroid, einem gemeinsamen Mittelpunkt, zugeordnet werden. Dieser Mittelpunkt kann beispielsweise ein Stadtteil oder auch ein Wohnblock sein. Zur Anonymisierung werden anschließend alle Standortdaten des Datensatzes auf den entsprechenden Zentroid abgebildet. Dabei müssen für jeden Zentroid mindestens k Einträge existieren. Indem die Standortinformation einer Person auf einen Zentroid abgebildet wird, wird der einzelne Standort innerhalb einer Gruppe un-unterscheidbar und so ist die Identität der Person ausreichend versteckt [32]. Speziell für Standortdaten existieren außerdem eigens dafür entwickelte Anonymisierungsmethoden, die im weiteren Verlauf des Kapitels erläutert werden.

Mix Zone

Sollen Standortdaten anonymisiert werden, existieren im Datensatz Positionen oder Trajektorien, die mit einer Information zu einer Person verknüpft sind. Der Ansatz *Mix Zone* (dt. gemischte Zone) kann nur online, bei Verwendung eines LBS angewandt werden. Hierfür wird eine Zone aufgespannt, in der zwei Bedingungen herrschen, wenn Nutzende einer Applikation diese Zone passieren: zum einen wird keine Person in der Zone ihren Standort aktualisieren, zum anderen bekommt jede Person ein neues Pseudonym, wenn sie die Zone verlässt. Es existieren also keine eindeutigen IDs für einzelne Personen. Diese Methode soll die Möglichkeit für Angreifende schwächen, Pseudonyme auf Personen zurückzuführen [10].

Cloaking

Weiter existieren für Standortdaten die von Gruteser und Grunwald vorgestellten *Cloaking* (dt. Tarnen) Ansätze [33]. Hier kann sowohl die räumliche (*spatial cloaking*) als auch die zeitliche (*temporal cloaking*) Instanz anonymisiert werden. Dabei wird das Ziel der k-anonymity verfolgt. Zur räumlichen Tarnung wird das entblößte Areal vergrößert, indem sich eine Person gerade aufhält. Die Genauigkeit der Standortinformation wird also reduziert. Um k-anonymity zu erreichen, muss das Areal so weit vergrößert werden, dass in der vergrößerten Region mindestens k andere Nutzende anwesend sind. Der Ansatz der zeitlichen Tarnung kann räumliche Information genauer wiedergeben, indem die zeitliche Instanz vergröbert wird [33]. So wird eine Anfrage einer Person an den Server so lange zurückgehalten, bis k andere Personen dasselbe Areal besucht haben [34]. Beide Methoden kommen online bei der Übermittlung der Daten an den Server zum Einsatz, sodass nicht zurückverfolgt werden kann, von welcher Person eine Anfrage gesendet wurde [10].

Probleme bei diesen Methoden können in Gegenden entstehen, in der nur eine geringe Bevölkerungsdichte herrscht oder die nur von wenig Personen besucht werden. Hier kann es sein, dass der Wert für k sehr niedrig gesetzt werden muss, damit die Anfragen an den Server gesendet werden können. Allerdings ist mit einem niedrigen k-Wert die Wirksamkeit der Methode beschränkt.

Auch ein Datensatz kann offline und zentral durch temporale oder räumliche Tarnung oder einer Kombination aus beidem anonymisiert werden [18]. Bei der zeitlichen Tarnung wird die Auflösung der Zeiten verringert. Nach der Privatisierung beinhaltet der Datensatz keine genauen Zeitpunkte, sondern Zeitintervalle (beispielsweise von 15 Minuten, also 10:05 - 10:20 Uhr). Durch räumliche Tarnung werden genaue Positionen auf Straßen, Postleitzahlen oder andere gröbere Standortinformationen abgebildet. Werden diese Tarnansätze offline angewandt, verschwimmen die Grenzen zum bereits vorgestellten Ansatz der Generalisierung beziehungsweise Aggregation.

Geomasking

Eine weitere speziell für Geodaten entwickelte Methode ist das sogenannte *Geomasking*. Ziel ist hierbei, die originalen Standorte zu verstecken oder zu verschieben, ohne die räumlichen Strukturen oder geographischen Verteilungen zu verändern. K-anonymity ist dann gegeben, wenn sich ein anonymisierter, verschobener Punkt von k anderen nicht mehr unterscheidet [14].

Beim Ansatz *random perturbation* (dt. zufällige Störung) wird ein Punkt räumlich um eine zufällige Distanz und Richtung verschoben. Hierbei wird angenommen, dass die Bevölkerung homogen verteilt ist, um herauszufinden, wie weit die Information eines Individuums verschoben werden muss, bis ein bestimmtes Level an Privatsphäre erreicht ist. Es wird ein Schwellenwert festgelegt, um eine maximale Verschiebedistanz festzulegen. Um eine höhere Konsistenz zum Originalwert zu gewährleisten, kann das sogenannte *Donutmasking* (dt. Donut Maskierung) verwendet werden. Hierbei wird ein minimaler und ein maximaler Radius für die Verschiebung festgelegt. Durch den minimalen Radius wird sichergestellt, dass der verschobene Punkt nicht zu nah am Originalstandort ist. Der maximale Radius gewährleistet, dass der verschobene Punkt nicht zufällig gewählt wird. Die Verschiebung eines Punktes wird so lange durchgeführt, bis beide Bedingungen erfüllt sind [14].

Problematisch bei Random Perturbation und Donutmasking ist die Bedingung, dass optimaler Schutz nur bei homogener Verteilung der Daten möglich ist. Ist diese nicht gegeben, kann es unter Umständen zu unzureichendem Schutz für einzelne Personen kommen [32]. Beispielsweise wenn die Methode zur Privatisierung von Wohnadressen in ungleichmäßig besiedelten Gebieten verwendet wird. Auch beim Geomasking kann es zu einem erhöhten Verlust der Nützlichkeit kommen, je größer die Verschiebedistanz ist [32].

2.9.3. Vernebelungsmechanismen

Vernebelungsmechanismen beabsichtigen die Verringerung der Präzision von (Standort-)information, beispielsweise durch Hinzufügen von sogenannter „Dummy“-Information oder Hinzufügen von Rauschen [1]. Manche Ansätze dieser Kategorie können sowohl online als auch offline zum Einsatz kommen, sodass in manchen Fällen eine vertrauenswürdige dritte Instanz notwendig ist, in anderen nicht.

Dummy Locations

Dieser Ansatz kann mit kleinen Unterschieden in der Umsetzung sowohl online als auch offline angewandt werden. Bei der online-Verwendung senden Nutzende zusätzlich zu ihrer korrekten Standortinformation zufällige, falsche „Dummy-Positionen“ mit der Anfrage an den Server. So können Angreifende nicht genau zuordnen, an welchem dieser Punkte sich die Person wirklich befindet [1]. Bei der offline-Anwendung dieser Methode auf einen Datensatz werden zusätzliche Dummy-Informationen generiert und eingefügt, die semantisch in den Datensatz passen müssen. Das bedeutet im Falle eines Datensatzes mit Routeninformationen, dass die Dummy-Routen so generiert werden müssen, dass die geographische Verteilung der Routen im Datensatz im Gesamten nicht verändert wird. Weiter muss die zusätzlich generierte Information in einer solchen Form sein, dass sie von Angreifenden nicht als Falschinformation enttarnt werden kann. Beispielsweise deuten im Kreis oder sprunghaft verlaufende Routen auf Falschinformation hin.

Standort-Vernebelung

Während der Ansatz der Dummy-Locations unter vielen Dummies den exakten Standort oder die wirklich gefahrene Route einer Person dennoch enthält, ist das Ziel der Vernebelung die vorsätzliche Reduktion der Präzision der Standortinformation. Anstelle eines genauen Standortes übermittelt eine Person während der Benutzung einer Applikation ein kreisförmiges Areal in ihrer Anfrage an den Server. Ähnlich dazu kann die GPS-Koordinate vor dem Absenden rotiert oder verschoben werden [1]. Wird der Standort bereits bei der Anfrage vernebelt, ist es für den Server schwierig eine genaue Antwort auf die Anfrage zu finden. Dieses Szenario eignet sich also nicht für Anwendungen, deren Serverantworten von präziser Eingangsinformation abhängig sind, wie beispielsweise Navigation. Standortinformation kann auch offline in einem Datensatz vernebelt werden, indem eine exakte Position auf ein größeres oder kreisförmiges Gebiet abgebildet wird. Auch hier ist die Abgrenzung zum Generalisierungsansatz schwammig.

2.9.4. Reduktion der Standortinformationsfreigabe

Dieser Ansatz ist ein reiner online-Ansatz und kommt nur für wenige Anwendungsfälle in Betracht. Der Vollständigkeit halber soll dieser dennoch kurz vorgestellt werden. Ziel dieser Methode ist es, dem Server so wenig Anfragen wie möglich zu stellen, um die Notwendigkeit der Übermittlung des Standortes von Personen zu reduzieren. Zur Umsetzung werden Daten, die für die Nutzenden von Interesse sind und abgefragt werden können, vor Erreichen der Ziel-Region heruntergeladen und im Cache gespeichert. Nachteile dieser Methode sind vor allem der hohe Speicherbedarf. Außerdem lässt sich diese Vorgehensweise nur umsetzen, wenn das Ziel einer Person im Vorhinein klar ist. Ein Anwendungsbeispiel ist Navigation in einer Kartenanwendung, bei der das Ziel klar ist und POIs am Zielort, wie Restaurants oder Tankstellen, vor Erreichen des Ziels im Cache gespeichert werden können [1]. Weiter kann die Reduktion der Information erreicht werden, in dem Nutzende derselben Applikation die Information untereinander teilen. So kann die Information, bevor sie vom Server abgefragt wird, von anderen Nutzenden der Applikation abgefragt werden.

Zuletzt sei an dieser Stelle nochmals darauf hingewiesen, dass die Grenzen zwischen den Ansätzen nicht klar definiert sind und sich die Methoden teilweise mehr als einer Kategorie zuordnen lassen. Es handelt sich lediglich um eine Einteilung in Kategorien, um die Auswahl eines Mechanismus zu vereinfachen. Zudem können zur Privatisierung eines Datensatzes mehrere Methoden zum Einsatz kommen, wenn dies methodisch und leistungstechnisch möglich ist.

Als Abschluss des Kapitels werden nochmals alle in diesem Abschnitt vorgestellten Methoden in einer kurzen Zusammenfassung tabellarisch dargestellt:

Kategorie	Methode	Wann	online/ offline
Kryptographie	Geheimnis teilen	Übertragung, Nutzung von LBS	online
	PIR (Private information retireval)	Übertragung, Nutzung von LBS	online
Anonymisierung (mit Ziel k-anonymity)	Suppression	Speicherung/Veröffentlichung	offline
	Generalisierung /Aggregation	Speicherung/Veröffentlichung	offline
	Geomasking (Random Perturbation, Donutmasking)	Speicherung/Veröffentlichung	offline
	Spatial Cloaking	Übertragung, Nutzung von LBS, Speicherung/ Veröffentlichung	online/ offline
	Temporal Cloaking	Übertragung/ Nutzung von LBS, Speicherung/ Veröffentlichung	online/ offline
Anonymisierung	Mix-Zone	Übertragung/ Nutzung von LBS	online
Vernebelung	Hinzufügen von Dummy-Locations / Rauschen	Übertragung, Nutzung von LBS, Speicherung/ Veröffentlichung	online/ offline
	Standort-Vernebelung	Nutzung von LBS, Übertragung	online
Reduzierung der Informationsanfragen	Caching	Nutzung von LBS	online

Tabelle 2.1.: Übersicht der in diesem Kapitel vorgestellten Location Privacy Preserving Mechanisms (LPPM)

3. Aktueller Forschungsstand zu LPPM

Nachdem das vorangegangene Kapitel in die theoretischen Grundlagen eingeführt hat, soll im Anschluss der aktuelle Forschungsstand im Bereich Privatsphäre für ortsgebundene Daten betrachtet werden.

Privatsphäre-Erhaltung bei Veröffentlichung und Nutzung von Daten ist in jüngerer Vergangenheit zu einem immer bedeutsameren Thema geworden. Daher sind vor allem in der Theorie viele Studien zu *Privacy Preserving Mechanisms* veröffentlicht worden. Dennoch ist die Thematik der Privatsphäre an sich älter, als die Forderung nach Privatsphäre. Bereits 1998 beziehungsweise 2002 stellt Sweeney das Modell der k-anonymity vor, welches auch heute noch zur Privatisierung von Daten verwendet wird [17].

3.1. Privatisierung von Datensätzen

Die Privatisierung der Datensätzen ist vor allem dann wichtig, wenn Datensätze verkauft oder veröffentlicht werden sollen, sodass sie für andere Personen zugänglich sind. Hierbei ist oft unklar, zu welchen Zwecken sie weiter verwendet werden. In diesem Abschnitt werden nur Studien vorgestellt, die sich mit ortsbezogenen Daten befasst haben, also Daten, die Informationen zu Adressen, Standorten oder Routen beinhalten.

Unter anderem ist hier die Arbeit „Anonymizing NYC Taxi Data: Does it matter?“ von Douriez et al. aus dem Jahr 2016 [18] zu nennen. Die Studie befasst sich mit der Frage, ob perfekte Anonymität möglich ist. Um Antwortmöglichkeiten zu finden, wird der von TLC bereitgestellte Taxifahrten-Datensatz zunächst privatisiert und anschließend Attacken auf diesen ausgeübt. Zur Ausführung der Attacken wird zuvor, wie von Angreifenden, Zusatzwissen durch Observation gesammelt. Der anschließend durchgeführte Angriff wird als sogenannte *Linking Attack* (dt. Verbindungsattacke) bezeichnet. Dabei wird versucht, das durch Observation erlangte Wissen zu benutzen, um in Verbindung mit dem privatisierten Datensatz private Information enthüllen zu können. Das Ziel der Studie ist es herauszufinden, wie erfolgreich die angewandten Privatisierungsmethoden gewesen sind.

Die Arbeit von Allshouse et al. [32] untersucht speziell für Geodaten Privatisierungsmechanismen, mit denen Personen in einem Datensatz geschützt werden können. Das Augenmerk liegt dabei auf dem Schutz der Adressen von Personen, die in einer Datenbank mit Gesundheitsdaten liegen. Das Ziel der Privatisierungsmechanismen ist das Erreichen von k-anonymity. Dabei wird besonders die Problematik behandelt, dass Geodaten nicht immer homogen verteilt sind. Im Beispiel dieser Arbeit sind dies dichter und weniger dicht besiedelte Areale einer Stadt. Die Arbeit untersucht, welche Probleme bei der Privatisierung solcher Daten aufkommen können.

Weiterhin existiert eine Vielzahl von Studien, die eine bestimmte einzelne Methode beziehungsweise eine Kategorie an Methoden auf einen Datensatz anwenden, um die Effektivität der Methoden zu untersuchen. Hier kann beispielsweise die Arbeit von Song Gao et al. [14] genannt werden, die die Methodik des Geomasking anwendet oder die Arbeit von A. Hasan, Q. Jiang und C. Li [16], in der ein Generalisierungsansatz angewandt und evaluiert wird.

Während die genannten Arbeiten sich mit der Privatisierung von einzelnen Attributen oder Standorten beschäftigen, untersuchen andere Studien die Privatisierung von Trajektorien. Hier ist die Arbeit von Jin et al. [23] zu nennen, die zunächst verschiedene Mechanismen zur Privatisierung von Trajektorien vorstellt. Anschließend werden diese auf Datensätze verschiedener Größen angewandt. Dabei evaluieren sie mehrere Aspekte. Zum einen betrachten sie den Erfolg der Methode, aber auch die benötigte Zeit des Algorithmus oder die Sensitivität der Methode bezogen auf die Größe des Datensatzes.

3.2. Privatisierung von Visualisierungen

Andere Studien fokussieren sich auf die Erstellung von privatisierten Visualisierungen, allem voran in Form von sogenannten *Heat Maps*. Die 2015 publizierte Arbeit von Oksanen et al. [4] verfolgt das Ziel anhand verschiedener Berechnungsmethoden k-anonymity zu erreichen. Es werden mit drei verschiedenen Kalkulationen drei verschiedene Heat Maps generiert und anschließend deren Level an Privatsphäre verglichen. Alle drei Methoden liefern ungefähr gleich gute Ergebnisse mit verschiedenen Kompromissen. Der Anspruch dieser Arbeit liegt in der temporären Privatisierung der Daten, sodass die entstandene Visualisierung dieser Daten privatisiert ist.

Auch die Studie von Sainio et al. [35] befasst sich mit der Generierung von Heat Maps von häufig benutzten Routen. Die Konzentration liegt dabei auf der Schaffung einer Pipeline, die Heat Maps schnell und effizient erzeugen kann. Die Studie wird motiviert von der hohen Rechenleistung, die für die zur Privatisierung erforderliche Datenfilterung beziehungsweise -vorverarbeitung benötigt wird. Das Augenmerk dieser Studie liegt daher auf der Vorverarbeitung der Daten und Routen, sodass der LPPM schnell angewandt werden kann. Die generierte Heat Map zeigt anschließend nur Strecken, die k-anonym sind.

Die Privatisierung von Daten stellt ein immer größer und wichtiger werdendes Forschungsfeld dar. Bis dato besteht eine große Lücke zwischen Anwendung und Theorie. Viele Studien beschäftigen sich mit der Suche nach Privatisierungsmethoden, und nur wenige mit den Möglichkeiten, diese Methoden in echten Anwendungen zu implementieren [1]. Oft werden in Studien eigene Methoden entwickelt oder Methoden auf einzelne Datensätze angewandt. Verschiedene Datensätze enthalten eine Vielzahl unterschiedlicher, zu schützender Attribute. Dies erschwert es, eine Implementierung für einen Datensatz ohne Änderungen auf einen anderen anzuwenden. Es fehlt weiter an Frameworks, die es ermöglichen Implementierungen auf verschiedene Datensätze anzuwenden.

4. Implementierung und Evaluation der LPPM

Nachdem in Kapitel 2 und 3 die theoretischen Grundlagen erläutert und der aktuelle Forschungsstand vorgestellt wurden, leitet dieses Kapitel die praktische Anwendung von Privatisierungsmethoden auf den in der Arbeit verwendeten Datensatz von Nextbike ein. Dazu soll zunächst der Datensatz mitsamt der beinhalteten Eigenschaften und gespeicherten Informationen vorgestellt werden. Weiter werden die für Angreifende interessanten Aspekte des Datensatzes betrachtet. Im weiteren Verlauf dieses Kapitels wird der Datensatz darauf hin geprüft, welche der in Kapitel 2.9 vorgestellten Privatsphäre-erhaltenden Methoden für diesen zur Privatisierung genutzt werden können. Anschließend wird die Implementierung der ausgewählten Ansätze beleuchtet und eine Evaluation durchgeführt, in der die durch die Mechanismen erreichte Privatsphäre und die resultierende Nützlichkeit des Datensatzes bemessen werden sollen.

4.1. Eignung der LPPM für den Datensatz

Um zu untersuchen, welche der vorgestellten LPPM für die Privatisierung des Nextbike-Datensatzes implementiert werden können, müssen zunächst die zu schützenden Eigenschaften des Datensatzes herausgearbeitet werden. Für die mit dem Datensatz geplante Fahrradklimaanalyse müssen spezifische Informationen, wie die genaue Route, erhalten bleiben. Daher wird jeder Ansatz daraufhin geprüft, ob der resultierende Datensatz die für die durchzuführenden Analyse benötigten Informationen erhält und der Ansatz daher sinnvoll zu implementieren ist.

4.1.1. Vorstellung des Datensatzes

Nextbike ist ein Bikesharing-Anbieter in Leipzig, dessen Fahrräder an beliebiger Stelle ausgeliehen und abgestellt werden können. Der Datensatz enthält Ausleihdaten der Stadt Leipzig im Zeitraum Juli bis September 2019. Für diesen Zeitraum wurden circa 240.000 Ausleihen vermerkt. Nextbikes sind Leihfahrräder, die nach dem *free floating*-Prinzip gemietet werden können. Das bedeutet sie können an fast beliebiger Stelle ausgeliehen und abgestellt werden. Fast beliebig bedeutet in diesem Fall, dass Nextbike die Stadt in Zonen einteilt, in denen das Abstellen kostenfrei ist beziehungsweise eine Gebühr kostet. Das ermöglicht den Nutzenden viel Flexibilität und macht den Datensatz wertvoll als auch angreifbar, da die Genauigkeit der Daten zur Analyse sehr hilfreich ist, die Daten allerdings auch dazu genutzt werden können, um die Nutzenden und ihre üblichen Wege zu identifizieren.

Im Datensatz sind Routen-ID, Beginn- und Endzeit der Ausleihe, Datum, Start- und Endkoordinaten der Route, Dauer der Fahrt sowie weitere, für diese Arbeit nicht relevante Parameter enthalten (siehe Abbildung 4.1). Der Datensatz enthält nicht die tatsächlich gefahrenen Routen. Um verschiedene LPPM nicht nur auf die Start- und Endkoordinaten der Route, sondern auch auf Trajektorien anwenden zu können, wurden aus den Start- und Endpunkten Routen generiert. Die Start- und Endpunkte liegen in Form von GPS-Punktkoordinaten vor. Die Route wird als Liste von GPS-Punktkoordinaten generiert. Eine detaillierte Vorstellung der Routengenerierung und

Vorverarbeitung des Datensatzes erfolgt in Kapitel 4.2.3.

rental_id	bike_id	start_date	start_place_id	start_lat	start_lng	end_date	end_place_id	end_lat	end_lng	duration_sec
77143245	72340	2019-07-03 14:30:04	4660971	51.32461483	12.33692229	2019-07-03 15:05:25	32710	51.34397007	12.3831743	2121
77816271	93450	2019-07-09 14:31:08	18792999	51.33998667	12.37128889	2019-07-09 14:34:23	18799440	51.33955556	12.37639556	195
77821728	93450	2019-07-09 15:23:01	18799440	51.33955556	12.37639556	2019-07-10 11:32:38	32097	51.34457657	12.37962842	72577

Abbildung 4.1.: Ausschnitt des Nextbike-Datensatzes nach Entfernen für die Untersuchung unrelevanter Attribute

User-IDs, die die Zuordnung von Routen zu Einzelpersonen zulassen würden, wurden vor Herausgabe des Datensatzes von Nextbike entfernt. Damit enthält der Datensatz nur Routen-IDs als eindeutigen Identifizierer. Enthält ein Datensatz zusätzlich Attribute, mit der einzelne Personen eindeutig identifiziert werden können, müssen diese bei der Privatisierung priorisiert betrachtet werden. In dieser Arbeit wird aufgrund des Fehlens einer User-ID die Routen-ID wie eine User-ID behandelt. Im ursprünglichen Datensatz sind einer User-ID mehrere Routen-IDs zugewiesen.

Im Nextbike-Datensatz stellen die Start- und Endkoordinaten sowie die Trajektorien der Routen die durch den Privatisierungsmechanismus zu schützenden Objekte dar. Start- und Endpunkte der Routen werden sich an frequentiert aufgesuchten Orten wie der Wohn- oder Arbeitsadresse der Personen befinden. Auch für kontinuierlichen Daten wie gefahrene Routen ist es wesentlich, die Privatsphäre zu schützen, da sich aus Trajektorien Bewegungsmuster und Gewohnheiten von Personen ableiten lassen. Bewegungsmuster können dabei wiederholt benutzte Routen, beispielsweise zur Arbeit oder Universität, sein. Bei unzureichender Privatisierung eines Datensatzes lassen solche Bewegungsmuster auf einzelne Personen rückschließen [2]. Für die am Ende der Arbeit durchgeführte Fahrradklima-Analyse ist Letzteres nicht von praktischem Belang, da die untersuchten Routen generiert wurden und nicht den exakt gefahrenen Strecken entsprechen.

Im Folgenden wird betrachtet, welche der in Kapitel 2.9 vorgestellten Methoden zur Privatisierung des Datensatzes angewandt werden können. Ziel der Arbeit ist es, den vorliegenden Datensatz zu privatisieren und den resultierenden, privatisierten Datensatz visuell darzustellen und bezüglich des Fahrradklimas zu analysieren. Hierbei ist anzumerken, dass zur Privatisierung eines Datensatzes nicht nur einzelne Methoden angewandt werden können, sondern die Methoden gegebenfalls auch miteinander kombiniert werden können [24].

4.1.2. Kryptographische Mechanismen und Informationsreduktion

Die in Kapitel 2.9.1 vorgestellten kryptographischen Mechanismen schützen vor Enthüllung der Information auf dem Übertragungsweg. Die Anfrage des Clients an den LBS-Server wird während der Übermittlung an den Server privatisiert, um zu verhindern, dass Angreifende durch Interferenz Information ableiten können. Der Ansatz der Informationsreduktion (siehe Kapitel 2.9.4) arbeitet mit Caching, um so die Notwendigkeit einer Serveranfrage zu reduzieren. Die am Zielort benötigte Information wird vor Erreichen des Ziels oder von anderen Nutzenden der Applikation in der Umgebung abgefragt. Beide Ansätze können nur während der Nutzung einer Applikation, also online, angewandt werden.

Da der Nextbike-Datensatz bereits eine Sammlung von Daten vieler Personen darstellt, können Methoden, die während der Nutzung einer Applikation durch Einzelpersonen angewandt werden, bereits bereits an dieser Stelle ausgeschlossen werden. Für die weitere Betrachtung kommen daher nur Anonymisierungs- oder Vernebelungsmechanismen in Frage, da diese auch offline zur Privatisierung eines Datensatzes anwendbar sind. Daher werden im Folgenden die verschiedenen Ansätze dieser beiden Methoden detaillierter betrachtet.

4.1.3. Anonymisierungsmechanismen

Bei den in diesem Abschnitt betrachteten Methoden ist das Ziel der Anonymisierung des Datensatzes das Erreichen von k-anonymity.

Aggregation / Generalisierung

Um einzelne Personen in einem Datensatz un-unterscheidbar zu machen, werden deren Alleinstellungsmerkmale gruppiert und in Oberkategorien aggregiert. Dabei wird detaillierte Information in gröberer Information, die auf mehrere Personen zutreffen kann, versteckt. Für die Anonymisierung der Nextbike-Daten müssen hierbei sowohl Start- und Endpunkte als auch die Trajektorien betrachtet werden. Die Start- und Endpunkte einer Route sind dann k-anonym, wenn mindestens k-1 Start- und Endpunkte anderer Routen auf denselben Punkten liegen. Trajektorien sind dann k-anonym, wenn mindestens k-1 andere Personen denselben Weg gefahren sind.

Um zunächst eine Methode zu implementieren, die auf den ursprünglichen Nextbike-Datensatz ohne Trajektorien angewandt werden kann, wird zuerst die Anonymisierung der Start- und Endpunkte betrachtet. Hierbei können die Start- und Endpunkte auf Zentroide aggregiert beziehungsweise abgebildet werden. Ein solcher Zentroid stellt einen Mittelpunkt oder Schwerpunkt dar und kann in diesem Fall die Postleitzahl oder der Stadtteil sein, in der die Route startet beziehungsweise endet. Für jeden Start- oder Endpunkt einer Route muss also das zugehörige Postleitzahlgebiet oder der zugehörige Stadtteil gefunden werden. Diese Aggregation überführt die detaillierte Information des Routenstarts/-endes in eine gröbere Information. Um anschließend das Ziel der k-anonymity zu erreichen, muss außerdem für jede Route geprüft werden, ob mindestens k-1 andere Routen die gleiche Kombination aus Start- und Endzentroid besitzen. Ist diese Bedingung nicht erfüllt, muss überlegt werden, was mit diesen nicht anonymen Routen passieren soll.

Ein Nachteil dieses Ansatzes ist der hohe Informationsverlust. Es entsteht eine starke Unausgeglichenheit zwischen Nützlichkeit und Privatsphäre der Daten. Die Aggregation der Routeninformation auf Zentroide macht die Information sehr unpräzise, der resultierende Datensatz verliert folglich an Informationsgehalt. Dennoch kann durch die unpräzise gemachte Information das Erreichen eines hohen Privatsphärelevels vermutet werden. Eine detaillierte Betrachtung der Effizienz der Methode erfolgt in Kapitel 4.3. Der bei diesem Ansatz resultierende Datensatz kann trotz seines vergröberten Informationsgehaltes für Visualisierungen wie *Flow Maps* (dt. Flusskarten) genutzt werden. Flow Maps sind Kartenvisualisierungen, die Strömungen zwischen Punkten hervorheben und werden in Kapitel 4.2.2 genauer erläutert.

Weiter soll die Anonymisierung ganzer Trajektorien betrachtet werden. Trajektorien stellen gefahrene Strecken dar, die für diese Untersuchung generiert worden sind. Die Wegpunkte der Strecken sind als punktgenaue GPS-Koordinaten festgehalten. Die Koordinaten des ursprünglichen Datensatzes als auch die der generierten Routen haben fünf bis sechs Nachkommastellen und sind somit auf 11 bis 111 cm genau [36]. Um k-anonymity zu erreichen, muss jeder Wegpunkt einer Strecke von mindestens k-1 anderen Personen passiert worden sein. Da die Wegpunkte beziehungsweise deren Koordinaten sehr präzise sind, werden die Trajektorien zunächst in gröbere Segmente unterteilt (siehe Abbildung 4.2). Dieser Schritt vereinfacht das Bestimmen des k-Wertes für einen Wegpunkt, da so zwei Punkte, die nur wenige Zentimeter voneinander entfernt liegen, auf dasselbe Segment abgebildet werden. Ein Segment stellt einen kurzen, circa 30-50 Meter langen Abschnitt einer Straße dar und fasst mehrere Wegpunkte einer Route zusammen. Es wird für jeden Wegpunkt der Route geprüft, auf welchem Segment er liegt und durch dieses ersetzt. Eine genaue Definition von Segmenten und wie diese ermittelt werden, wird in Kapitel 4.2.1 gegeben.

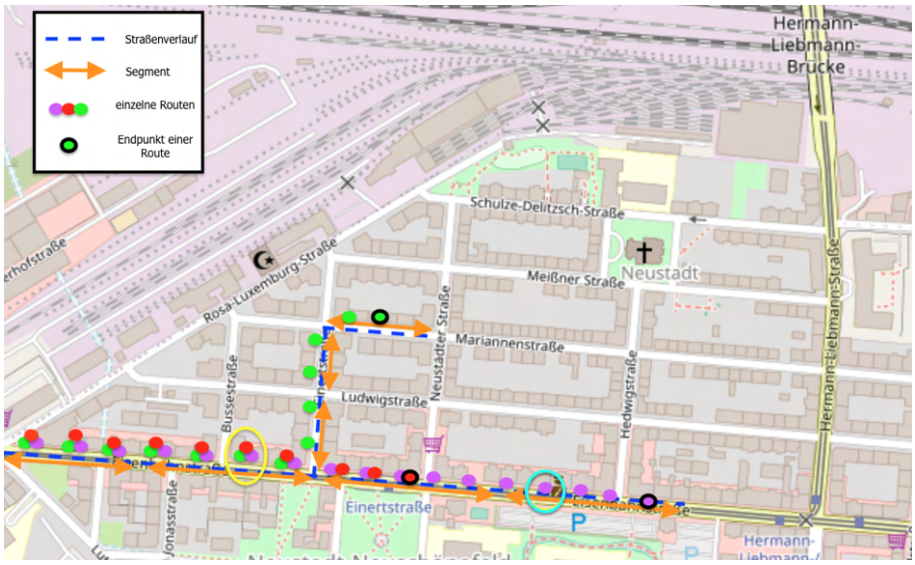


Abbildung 4.2.: Routenverläufe von drei verschiedenen Routen mit eingezeichneten Segmenten. Die Anzahl der Routen pro Segment für den gelben Kreis ist drei und für den blauen Kreis eins. Das gelb eingekreiste Segment ist folglich 3-anonym, das blau eingekreiste 1-anonym.

Die einzelnen Segmente können anschließend auf k-anonymity überprüft werden, indem gezählt wird, wie viele Routen im Datensatz ein bestimmtes Segment passiert haben (siehe Abbildung 4.2). Ein Segment ist k-anonym, wenn mindestens k Routen über dieses verlaufen. Hierbei gilt es einen k-Wert zu finden, der resultierende Nützlichkeit und erreichte Privatsphäre in ein Gleichgewicht bringt. Für nicht k-anonyme Segmente muss überlegt werden, was mit diesen Segmenten passiert.

Der Nextbike-Datensatz soll als Visualisierung dargestellt werden. In der Studie von Oksanen et al. wenden die Autoren vor der Visualisierung ihrer Daten einen Vorverarbeitungsschritt, ein sogenanntes *privacy filtering* (dt. Privatsphäre-Filterung), an. Dieser Schritt filtert Daten heraus, die nicht k-anonym sind. In der Visualisierung werden letztlich nur Daten gezeigt, die den k-Schwellenwert erreicht haben [4]. Für die Visualisierungen der Nextbike-Daten müssen nicht k-anonyme Daten also aus der Visualisierung ausgeblendet werden. Dies gilt für die oben vorgestellte Aggregation

auf Zentroide, als auch für die hier vorgestellte Aggregation auf Segmente. Soll der privatisierte Nextbike-Datensatz veröffentlicht werden, müssen die nicht k-anonymen Segmente aus dem Datensatz gelöscht werden.

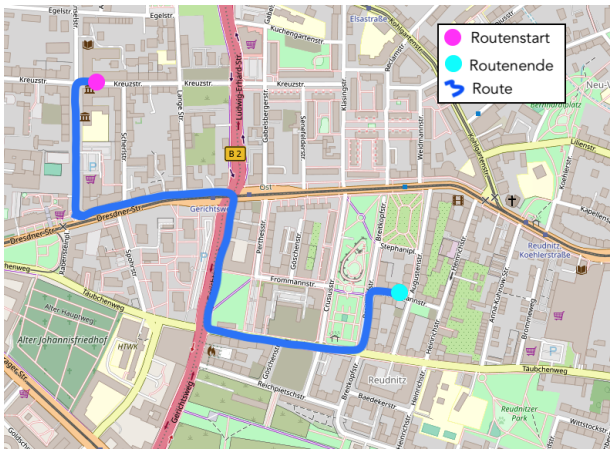
Mit der Aggregation auf Segmente werden vor allem Trajektorien privatisiert. Die Start- und Endpunkte einer Route werden nicht explizit daraufhin untersucht, ob andere Routen auch an diesen Punkten beginnen oder enden. Um die Start- und Endpunkte einer Trajektorie zu privatisieren, kann der Ansatz des *Geomasking* verwendet werden.

Geomasking

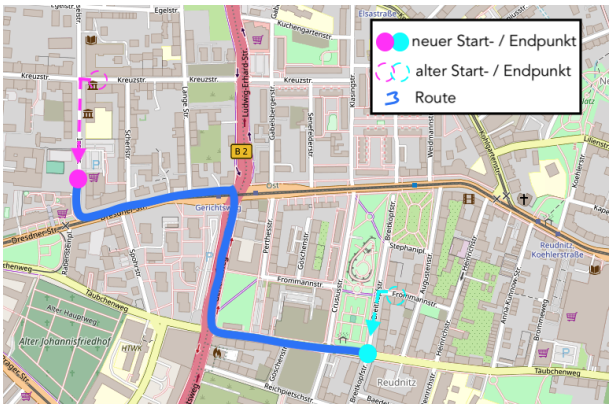
Die Start- und Endpunkte einer Route sind dann k-anonym, wenn mindestens k-1 andere Routen auf diesen Start- und Endpunkten starten beziehungsweise enden. *Geomasking* verschiebt oder verändert Punkte so, dass diese anonymisiert und in einer Gruppe von Punkten nicht zu unterscheiden sind. Donutmasking stellt dabei die zusätzliche Bedingung, dass die Verschiebung nicht zufällig sein darf, sondern die Semantik der Daten erhalten werden muss. Diese Bedingung wird durch eine minimale und eine maximale Verschiebedistanz definiert. Im Fall der Nextbike-Daten ist die Semantik der Daten der Routenverlauf. Dieser wird benötigt, um bei Analysen beispielsweise darstellen zu können, welche Strecken und Straßen von vielen Personen gefahren wurden. Würden Start- und Endpunkte zufällig und unkontrolliert verschoben werden, würde diese Information verfälscht werden.

Um die Start- und Endpunkte mithilfe des Donutmasking zu privatisieren und die Routeninformation nicht zu verfälschen, werden die Start- und Endpunkt nicht zufällig, sondern entlang der Route verschoben (siehe Abbildung 4.3, (a) und (b)). Zur Ermittlung der k-anonymity von Start- und Endpunkten werden die Routen wieder in Segmente unterteilt, die Verschiebung der Punkte erfolgt nacheinander. Hierbei kann die minimale Verschiebedistanz definiert werden als der erste Punkt in der Route, an dem der Start- oder Endpunkt k-anonym ist. Die maximale Distanz berechnet sich aus dem Anfang und dem Ende der Route. Ein Startpunkt wird dann in Richtung Routenende verschoben und ein Endpunkt Richtung Routenanfang. Für die Privatisierung der Nextbike-Daten bedeutet dies, dass nicht k-anonyme Start- oder Endpunkte solange entlang der Route nach vorne oder hinten verschoben werden, bis sie auf einem Segment mit mindestens k-1 anderen Start- beziehungsweise Endpunkten liegen. Bei dieser Vorgehensweise wird nur der Beginn beziehungsweise das Ende der Route betrachtet, der Verlauf der Route zwischen diesen Punkten wird nicht weiter analysiert. Das Donutmasking schützt also nicht die Privatsphäre der Trajektorie.

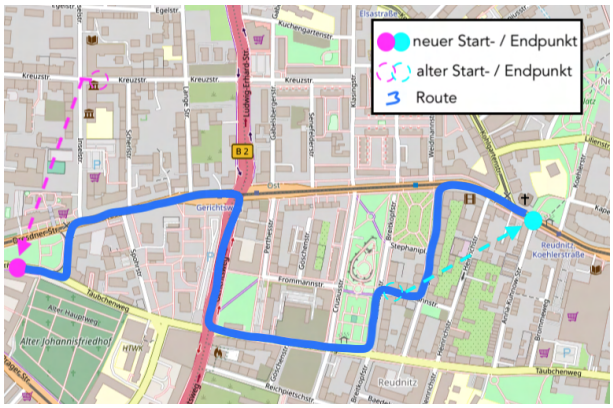
Für die später folgenden Analysen ist der Erhalt der Routeninformation von hoher Wichtigkeit. *Random Perturbation* ist daher ein Mechanismus, der für die Nextbike-Daten wenig geeignet ist. Bei diesem Ansatz werden Punkte in einem bestimmten Radius zufällig verschoben (siehe Abbildung 4.3, (c)). Wird ein Start- oder Endpunkt einer Route zufällig verschoben, geht nicht nur die genaue Routeninformation verloren, sie wird sogar verfälscht. Die hierbei entstehenden Routen würden Betrachtungen wie die Suche nach viel gefahrenen Strecken nicht mehr ohne Verzerrung zulassen.



(a) Ursprünglich gefahrene Route



(b) Durch Donutmasking verschobene Route



(c) Durch Random Perturbation verschobene Route

Abbildung 4.3.: Verschiebung von Routenstartpunkte. Donutmasking erhält die ursprünglich gefahrene Route, während Random Perturbation die Verschieberichtung zufällig wählt.

4.1.4. Vernebelungsmechanismen

Vernebelungsmechanismen privatisieren einen Datensatz, indem die Standortinformation vernebelt, also unpräzise gemacht wird. Dies wird erreicht, indem zusätzliche Falschinformation beziehungsweise Rauschen zum Datensatz hinzugefügt wird. Ziel ist es, durch die unpräzise oder zusätzliche Falschinformation vor Re-Identifikation zu schützen, indem POIs nicht mehr eindeutig einer Person zugeordnet werden können.

Hinzufügen von Dummy Locations

Eine Möglichkeit, Vernebelung zu schaffen, ist das Hinzufügen von Falschinformation, sogenannten *dummies*. Im Fall des Nextbike-Datensatzes handelt es sich bei den zu generierenden *dummies* um Dummy-Routen, die zum Datensatz hinzugefügt werden müssen, um die echten Daten unter einer Menge falscher Informationen zu verstecken. Die Dummy-Routen müssen für eine unverfälschte Analyse aus der tatsächlichen geographischen Verteilung der Routen des Datensatzes abgeleitet werden. Eine weitere Herausforderung in der Umsetzung dieses Ansatzes besteht darin, die hinzugefügten Routen nicht als Falschinformation erkennbar zu machen. Die zu generierenden Start- und Endpunkte beziehungsweise Trajektorien müssen so gewählt werden, dass diese nicht als falsch

identifizierbar sind, wie beispielsweise im Kreis oder sprunghaft verlaufende Routen. Die Methode schützt vor Re-Identifikation, indem die Dummy-Routen gleichzeitig einer User-ID zugewiesen werden, um so beispielsweise zu verhindern, dass sich für einzelne Personen Fahrmuster erkennen lassen. Da Nextbike die User-ID vor Herausgabe des Datensatzes entfernt hat, gibt es keine Personen-ID, denen die Routen zugewiesen werden können. Zudem soll der Nextbike-Datensatz in dieser Untersuchung auf häufig benutzte Strecken beziehungsweise Straßen analysiert werden. Durch Hinzufügen von Dummy-Routen wird trotz Beibehaltung der geographischen Verteilung die Häufigkeit der gefahrenen Strecken verfälscht. Es befinden sich nach Hinzufügen der Falschinformation mehr Routen im Datensatz als tatsächlich gefahren wurden. Um die gefahrenen Strecken so unverfälscht wie möglich analysieren zu können, wird das Hinzufügen von Dummy-Routen für diese Untersuchung nicht betrachtet.

Standort-Vernebelung

Bei der Nutzung von LBS kann Standort-Vernebelung zum Einsatz kommen, um die genaue Position einer Person zu vernebeln. Dies kann umgesetzt werden, indem, anstelle der genauen Koordinate, ein kreisförmiges Areal als Polygon gesendet oder die Punktkoordinate verschoben oder rotiert wird. Es handelt sich hierbei um ein Verfahren, welches online während der Nutzung einer Applikation zur Anwendung kommt. Da zur Privatisierung des in dieser Arbeit verwendeten Nextbike-Datensatzes eine offline-Methode benutzt werden muss, kann die Standort-Vernebelung nicht verwendet werden. Dennoch sind die Ansätze des Rotierens beziehungsweise Verschiebens der Punktkoordinate ähnlich der bereits vorgestellten Methode des *Geomasking*, die für diese Arbeit im weiteren Verlauf implementiert und evaluiert wird.

4.1.5. Zusammenfassung der zu implementierenden Methoden

Zusammenfassend werden im weiteren Verlauf dieser Arbeit folgende LPPM auf die vorhandenen Daten angewandt und analysiert:

- Aggregation der Start- und Endpunkte auf einen Zentroid
- Aggregation der Routendaten auf Straßensegmente
- Donutmasking von nicht k-anonymen Start- und Endpunkten

Die vorliegende Arbeit fokussiert sich trotz der bereits vorgestellten Schwächen von k-anonymity auf Methoden mit dem Ziel des Erreichens von k-anonymity. Obwohl Studien gezeigt haben, dass k-anonymity allein bei bestimmten Attacken nicht immer ausreichenden Schutz gewährleisten kann (siehe *Linking Attack*, in Kapitel 2.4), hat auch jedes daraufhin vorgeschlagene Modell zur Verbesserung dieser Schwierigkeiten Schwächen an anderen Stellen [27]. Um nach Anwendung der Methoden prüfen zu können, wie stark diese Schwächen nach der Privatisierung im Datensatz ausgeprägt sind und ob weiterer Privatisierungsbedarf besteht, werden speziell für k-anonymity entwickelte Methoden zur Evaluierung der Privatsphäre angewandt (siehe Kapitel 2.7).

4.2. Implementierung der LPPM

In diesem Kapitel werden zunächst zur Implementierung benutzte Services und Grundbegriffe erläutert. Anschließend werden die verwendeten Visualisierungsformen vorgestellt und die Vorverarbeitung des Datensatzes beleuchtet, um im Anschluss daran die Implementierung der Ansätze betrachten zu können. Abschließend wird ausgewertet, wie erfolgreich die Privatisierungsmethoden waren.

Für die Fahrradklima-Analyse wird der Nextbike-Datensatz unter anderem auf häufig gefahrene Strecken, deren Beschaffenheit und deren Radwegsituation hin analysiert. Für diese Analyse wird der Datensatz privatisiert. Dabei werden fünf verschiedene k-Werte gewählt, die eine unterschiedliches Level an Privatsphäre repräsentieren. Im Fall dieser Untersuchung ist bereits vor der Privatisierung klar, welche Eigenschaften und Informationen des Datensatzes bei der Privatisierung erhalten bleiben müssen, um die Analyse durchführen zu können. Dies kann so bei der Implementierung der Methoden berücksichtigt werden.

4.2.1. Verwendete Services und Datentypen

Im Nextbike-Datensatz ist keine Routeninformationen enthalten, ausschließlich Start- und Endkoordinaten. Aus Mangel eines Datensatzes, der gefahrene Streckeninformationen enthält, wurden die Strecken aus den Start- und Endpunkten generiert. Hierfür wurde der Dienst „Open Route Service“ [37] genutzt. Eine Instanz dieses Dienstes wurde lokal installiert, um Limitierungen bezüglich der Anfragezahl zu umgehen. Um aus einem Koordinatenpaar eine Route zu generieren, werden die Parameter Startkoordinate und Endkoordinate übergeben. Der Dienst bietet die Wahl zwischen verschiedenen Bewegungsprofilen, darunter verschiedene Fahrradprofile (Mountainbike, regulär, elektrisch), Autoprofile als auch Laufprofile. Es wurde das Bewegungsprofil *bicycle-regular* gewählt. Die Ausgabe des Dienstes erfolgt im geoJSON-Format. Dabei wird die Route in Form eines sogenannten *LineString* zurück gegeben, welcher eine geordnete Liste von Koordinatenpaaren darstellt. Außerdem werden Informationen zur Dauer und Distanz der Route, als auch eine Wegbeschreibung im Textformat bereitsgestellt.

GeoJSON ist ein Format, welches auf der JavaScript Object Notation (JSON) beruht und für die Repräsentation von geographischen Daten entwickelt worden ist. Es wird genutzt, um eine Vielzahl von geographischen Datenstrukturen auf Karten abzubilden, darunter Punkte, Routen oder Polygone [38]. Für diese Arbeit von Interesse ist der *LineString*, welcher Routen in Form von Koordinaten in einer geordneten Liste speichert.

Um, wie in Kapitel 4.1.3 erläutert, Trajektorien in Segmente einzuteilen, muss für jeden Punkt einer Route betrachtet werden, auf welchem Segment einer Straße er liegt. Segmente stellen Straßenabschnitte mit einer Länge von 30 bis 50 Meter dar. Open Street Map (OSM) stellt lizenzfreies Kartenmaterial als auch eine Datenbank mit Hintergrundinformation zur Verfügung. In dieser Datenbank sind unter anderem Informationen über Straßen, Häuser oder Flüsse enthalten. Um Information der Datenbank abzufragen, bietet OSM den Dienst Nominatim [39] an. Dieser kann genutzt werden, um durch die Eingabe von Koordinaten ein entsprechendes Segment zu ermitteln.

Dieses Vorgehen wird auch als *Reverse Geocoding* bezeichnet (siehe Abbildung 4.4). Auch dieser Dienst wird zum Umgehen von Anfrage-Limitierungen lokal installiert.

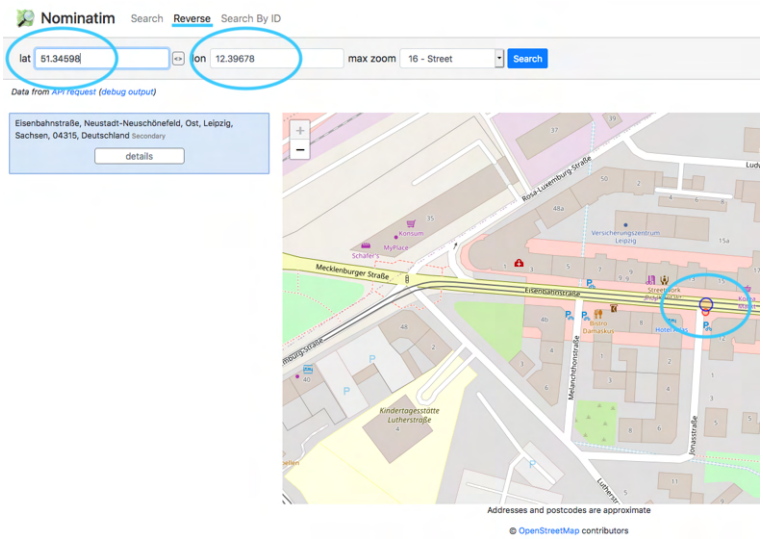


Abbildung 4.4.: Nominatim Reverse Geocoding. Durch Eingabe eines Koordinatenpaares kann der Dienst Information aus der Open Street Map Datenbank abfragen.

Mithilfe von *Reverse Geocoding* können Informationen wie POIs, Postleitzahlen, Geschwindigkeiten oder sogar die Art des Straßenbelages durch Angabe der Koordinaten gesucht werden (siehe Abbildung 4.5). Nominatim wertet die eingegebenen Koordinaten zu einer OSM-ID aus, die intern vom OSM-System vergeben wird. Es gibt verschiedene Elemente, die eine OSM-ID haben können. Wichtig sind in dieser Arbeit vor allem Straßensegmente, die in der OSM-Datenbank als Typ *way* oder *relation* gespeichert sind (s. auch Abbildung 4.5).

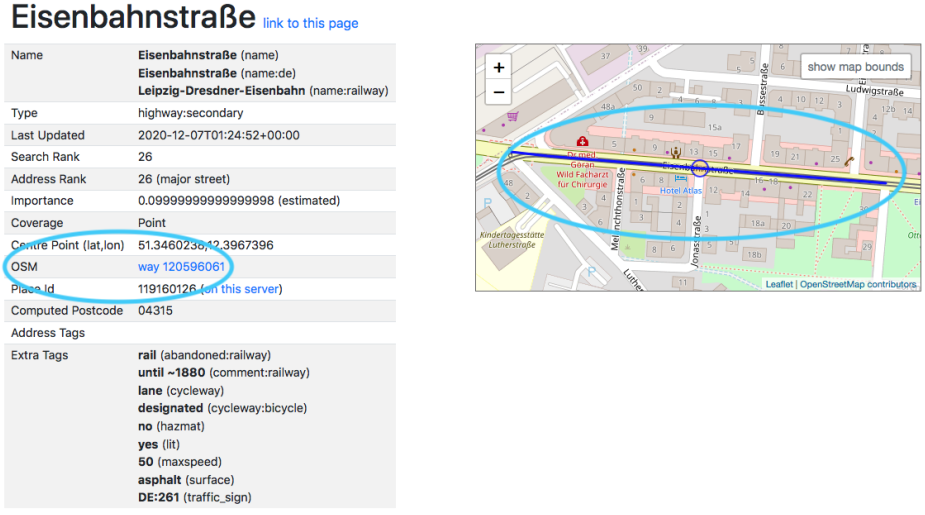


Abbildung 4.5.: Nominatim Ermittlung eines Segments für Koordinatenpaar. Das Segment mit der ID *way120596061* ist rechts im Bild zu sehen. Zusätzlich werden in den *tags* Informationen zum Straßenbelag oder zum Vorhandensein eines Radweges bereitgestellt.

Ein Segment ist ein Abschnitt einer Straße, der aus mehreren Koordinaten, den sogenannten *nodes*, besteht. Für jedes Segment können die Koordinaten, die es umfasst, anhand dieser Beziehung abgefragt werden. Ein weiterer Dienst, der *Reverse Geocoding* anbietet und für diese Arbeit getestet

wurde, ist Pelias [40], dessen Ergebnisse allerdings nicht genau genug sind und der daher für diese Arbeit nicht verwendet werden kann.

Die originalen wie auch die privatisierten Daten werden in einer PostgreSQL-Datenbank gehalten, die mit einer PostGIS-Datenbank für Geo-Daten verknüpft ist. Um die Daten anschließend in einem Frontend (Browser) visualisieren und aus der Datenbank abrufen zu können, wird das Webframework Flask verwendet. Zur Darstellung der Karten wird Mapbox verwendet.

4.2.2. Verwendete Visualisierungsformen

Für die Analyse des Fahrradklimas in Leipzig sollen die gefahrenen Strecken auf folgende Fragestellungen hin visualisiert und anschließend ausgewertet werden:

- Gibt es unterschiedlich viel Fahrradverkehr in und zwischen bestimmten Stadtteilen oder Postleitzahlgebieten?
- Welche Strecken werden besonders häufig gefahren?
- Welche Strecken verfügen über Radwege oder eine gute beziehungsweise schlechte Straßenbeschaffenheit?
- Gibt es Strecken, die zu unterschiedlichen Tageszeiten häufiger frequentiert werden?

Zur Visualisierung des Nextbike-Datensatzes werden Flow Maps und Heat Maps (dt. Hitzekarte) verwendet. In diesem Abschnitt werden die beiden Visualisierungsformen kurz erläutert. Anhand der Visualisierungen der zu analysierenden Fragen soll ermöglicht werden, vor allem in der Stadt- und Verkehrswegeplanung Fragen nach der Sicherheit oder des Bedarfs an Radverkehrsanlagen beantworten zu können. Die hier analysierten Strecken wurden aus Start- und Endpunkten des Nextbike-Datensatzes generiert. Die entstandenen Heat Maps dienen somit nur als Beispiel, wie solche Streckenanalysen visualisiert und welche Informationen aus ihnen abgeleitet werden können.

Flow Map

Die Visualisierung als Flow Map soll Strömungen zum Vorschein bringen. Sie wird häufig dafür verwendet, den Fluss beziehungsweise die Bewegungen von beispielsweise Geld oder Migration von einem Startpunkt zu einem Endpunkt zu visualisieren [41]. In vielen Darstellungen dieser Form wird die Dicke oder auch Intensität der Linien oder Pfeile als proportional zur Menge des Flusses dargestellt, sodass die Ableitung von Information aus der Darstellung intuitiv erfolgen kann. Der Typ der Flow Map, der in dieser Arbeit verwendet wird, ist die sogenannte „Ursprung-Ziel-Karte“. Sie zeigt Verbindungen zwischen zwei Punkten als Pfeil, wobei die Dicke der Pfeile ein Volumen beziehungsweise eine Menge repräsentiert [42].

Die Software, welche in dieser Arbeit zur Erstellung der Flow Map verwendet wird, ist FlowMap.blue [43]. Die Visualisierung der Nextbike-Daten als Flow Map hat die gröbste Auflösung in Bezug auf die Datenqualität. Das bedeutet, die hier gezeigten Daten wurden am meisten in ihrer Präzision verringert. Die Strömungen werden zum einen für Stadtteile betrachtet und zum

anderen für Postleitzahlgebiete. Das bedeutet die punktgenaue Information über den Routenstart beziehungsweise das Routenende wird darauf reduziert, in welchem Stadtteil oder welchem Postleitzahlgebiet die Route begonnen oder geendet hat.

Heat Map

Eine Heat Map visualisiert Geo-Daten auf einer Karte, indem Farben dazu benutzt werden, unterschiedlich hohe Konzentrationen von Punkten zu repräsentieren [44]. Heat Maps werden häufig angewandt, wenn große Mengen von Daten involviert sind. Sie wird beispielsweise im Feld der Mensch-Computer-Interaktion zum Verfolgen von Augenbewegungen verwendet [4]. Die in dieser Arbeit verwendete Form der Heat Map wird in manchen Fällen auch als *Intensity Map* (dt. Intensitätskarte) bezeichnet [44]. Intuitiv wird hohe Intensität oder eine hohe Konzentration verstanden, indem eine Farbe wie rot gewählt wird, die Hitze repräsentiert. Niedrige Intensität wird intuitiv mit kalten Farben verbunden. Der Verlauf des „heißer werdens“ wird somit automatisch mit der Erhöhung der Intensität der Farbe verbunden und interpretiert [44].

Um zu analysieren, welche Strecken sehr häufig und welche weniger häufig gefahren wurden, werden die Nextbike-Daten als Heat Map visualisiert. Dafür wird aus zwei Farben ein Farbgradient generiert, welche jeweils einen Extrempunkt darstellen: eine Farbe für das untere Ende (die kleinste Anzahl an Routen), eine Farbe für das obere Ende (die meisten Anzahl an Routen). Je öfter eine Strecke gefahren wurde, desto intensiver ist die Farbe an dieser Stelle. Innerhalb der zwei Extrempunkte des Farbgradienten wird die entsprechende Farbe für einen Punkt aus der Information berechnet, wie viele Routen aus dem Datensatz diesen Punkt passiert haben und wie viele Strecken im Datensatz insgesamt enthalten sind.

4.2.3. Vorverarbeitung des Nextbike Datensatzes

Im Folgenden werden die Vorverarbeitungsschritte kurz erläutert, bevor der Datensatz privatisiert werden kann. Um einen Überblick und ein grobes Verständnis über den Datensatz zu bekommen, wurden die Daten zunächst in Form von Histogrammen visualisiert (siehe Abbildung 4.6). Aus diesen Histogrammen lässt sich die Anzahl der Ausleihen pro Tag und deren durchschnittliche Länge ableiten. Es ist erkennbar, dass es pro Tag ungefähr 2.500 Ausleihen gibt, die circa fünf bis zehn Minuten dauern.

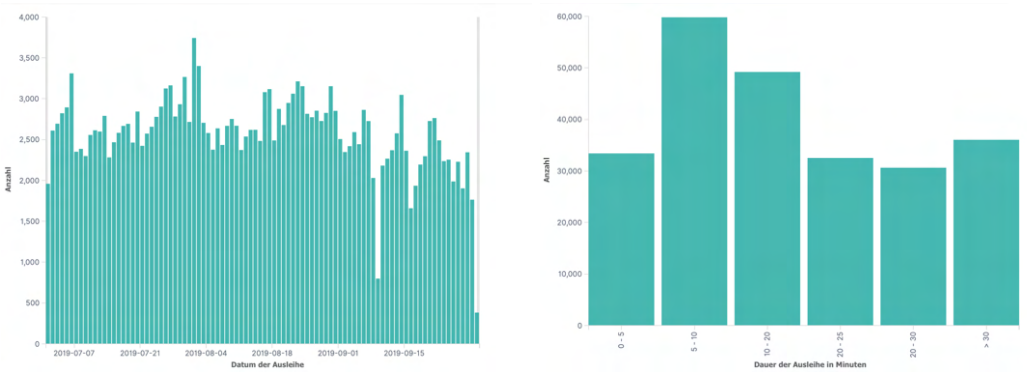


Abbildung 4.6.: Histogramme zum Nextbike-Datensatz. Links: Anzahl der Ausleihen pro Tag, rechts: Länge der Ausleihen gruppiert.

Auf Basis der gewonnenen Kenntnis wurden drei Vorverarbeitungsschritte durchgeführt:

1. Löschen von Ausleihen von und zur Station „nextbikeHQ“, da es sich hier um Testfahrten von Nextbike Mitarbeitenden handelt (693 Einträge).

2. Löschen von Ausleihen mit einer Dauer von weniger als 100 Sekunden, da mit Strecken unter 100 Sekunden keine Analysen durchgeführt werden können, beziehungsweise es sich auch um defekte Fahrräder handeln kann (6291 Einträge).

Der Originaldatensatz enthält das Datum sowie Start- und Endzeiten einer Ausleihe. Datum und Uhrzeit stellen sensitive Attribute dar, die zusammen mit der gefahrenen Strecke Bewegungsmuster vertiefen können. Beispielsweise fahren Personen eine bestimmte Strecke zu einer bestimmten Uhrzeit. Für Analysen zu häufig gefahrenen Strecken, Vorhandensein von Radwegen oder Straßenqualität sind Datum und Uhrzeit nicht von großer Relevanz. Sie werden daher vor der Privatisierung gelöscht. Dieser Schritt minimiert die Anzahl der zu schützenden Attribute im Datensatz und vereinfacht so die Privatisierung. Sollen dennoch Fahrradmengen zu bestimmten Tageszeiten analysiert werden, kann aus dem Originaldatensatz eine Teilmenge des Datensatzes mit der gewünschten Tageszeit erstellt werden. Diese Teilmenge kann anschließend privatisiert und analysiert werden. Dieses Vorgehen wird in Kapitel 5 nochmals genauer erläutert.

Um Strecken analysieren zu können, wurden anhand der Start- und Endpunkten Routen generiert. Indem das Bewegungsprofil *bicycle-regular* gewählt wurde, versucht der verwendete Routing-Dienst, fahrradfreundliche, sichere Routen zu generieren [45]. Die berechneten Routen werden als geordnete Liste von Koordinatenpaaren im GeoJSON-Format zurückgegeben. Die Koordinaten haben dabei eine Präzision von fünf bis sechs Nachkommastellen. Die Rückgabe der API wird anschließend bis auf die folgenden Punkte bereinigt: Distanz und Dauer der Route sowie die Wegpunkte.

Aus den aus der Vorverarbeitung verbliebenen Einträgen des Nextbike-Datensatzes konnten 967 Einträge nicht zu Routen umgewandelt werden, da die Start- und/oder Endkoordinaten fehlerhaft waren und entweder eine oder beide Angaben gefehlt haben. Weitere 45 konnten nicht verwendet werden, da der Routing-Dienst für diese Koordinaten keine Route finden konnte. Der Datensatz enthält folglich nach allen durchlaufenen Vorverarbeitungsschritten noch 233.213 Einträge. Um die einzelnen Implementierungen durchführen zu können, waren weitere, spezifische Vorverarbeitung notwendig, die in den entsprechenden Kapiteln erläutert werden.

4.2.4. Aggregation der Start- und Endkoordinaten auf einen Zentroid

Die erste Implementierung eines LPPM ist die Aggregation auf einen Zentroid. In dieser Arbeit stellt dies bezüglich der Datenqualität die größte Veränderung der Daten dar. Die Daten werden hierbei von genauen Punktkoordinaten auf einen Zentroid reduziert. Dieser wird entweder durch einen Stadtteil oder ein Postleitzahlgebiet repräsentiert. Da es fast unmöglich ist, von einem Stadtteil auf einen einzelnen Punkt rückzuschließen, kann hier ein hohes Maß an Privatsphäre erwartet werden.

Die Routen werden darauf reduziert, in welchem Stadtteil oder welchem Postleitzahlgebiet sie begonnen und geendet haben. Dafür wird aus den bereits vorverarbeiteten Daten (siehe Kapitel 4.2.3) mit den Attributen *rental_id*, *start_zip* (Start-PLZ-Gebiet), *end_zip* (End-PLZ-Gebiet), *start_suburb* (Start-Stadtteil), *end_suburb* (End-Stadtteil) eine neue Datenbanktabelle angelegt. Die *rental_id* wird dabei aus dem Originaldatensatz übernommen. Die entsprechende Information über Stadtteil oder Postleitzahlgebiet ist nicht im Datensatz enthalten und wird über den in Kapitel 4.2.1 vorgestellten *Reverse Geocoding*-Ansatz ermittelt. Der aus der Aggregation resultierende Datensatz enthält folglich noch die Information der Ausleih-ID sowie den Start- und Endpunkt der Ausleihe in Form von Postleitzahlgebiet oder Stadtteil (siehe Abbildung 4.7).

	rental_id integer	start_zip text	start_suburb text	end_zip text	end_suburb text
1	77143245	04229	Plagwitz	04103	Zentrum-Ost
2	83358831	04109	Zentrum-Nordwest	04109	Leipzig-Zentrum
3	85986086	04109	Leipzig-Zentrum	04109	Zentrum-West
4	80971507	04107	Zentrum-Süd	04109	Leipzig-Zentrum
5	82547783	04109	Zentrum-West	04277	Connewitz

Abbildung 4.7.: Ausschnitt der Datenbank, die für jede Ausleihe eine ID, sowie die Start- und Endpunkte als Zentroide (Stadtteil oder Postleitzahlgebiet) enthält.

Um nun die Privatsphäre durch Erreichen von k-anonymity sicherzustellen, werden die Start- und Endzentroide jeder Route im Datensatz für die k-Werte 10, 20, 50, 100 und 1000 überprüft. Diese k-Werte sollen hierbei den Vergleich unterschiedlicher Privatsphäre-Level ermöglichen. Nur Einträge, für die mindestens k-1 andere Ausleihen mit gleichen Start- oder Endzentroiden existieren, verbleiben im Datensatz. Ausleihen, die diese Bedingung nicht erfüllen, werden aus dem Datensatz gelöscht (siehe Abbildung 4.8). Dabei wird die Überprüfung für die beiden Zentroide, Postleitzahlgebiete oder Stadtteile, getrennt vorgenommen und die k-anonymen Routen in einer Tabelle entsprechend dem k-Wert gespeichert.

```
# start_value = either start_zip or start_suburb
# end_value = either end_zip or end_suburb
rental = [rental_id, start_value, end_value]
all_rentals = [rental1, rental2, ...]
k_value = 10

for start_value, end_value in all_rentals:

    start_k_anonym = test_k_anonymity(start_value, k_value)

    if start_k_anonym:
        end_k_anonym = test_k_anonymity(end_value, k_value)

        if end_k_anonym:
            save_to_database(rental_id, start_value, end_value)
    else:
        continue
```

Abbildung 4.8.: Pseudocode zur Beschreibung der Methode Zentroidaggregation

Der resultierende, auf Zentroide aggregierte Datensatz kann nun mithilfe der in Kapitel 4.2.2 vorgestellten Flow Map visualisiert werden. Diese bringt die Strömungen der Ausleihdaten zwischen zum einen Stadtteilen und zum anderen Postleitzahlgebieten zum Vorschein.

4.2.5. Aggregation der Routendaten auf Straßensegmente

Die im vorigen Abschnitt vorgestellte Aggregation der Start- und Endpunkte auf einen Zentroid stellt einen großen Eingriff in die Daten dar. Die Balance zwischen Privatsphäre und Nützlichkeit verschiebt sich hierbei stark in Richtung Privatsphäre, die Daten verlieren an Nützlichkeit. Eine Möglichkeit die Daten zu privatisieren und dabei weniger Information zu verlieren, ist die Aggregation auf Segmente. Hierbei bleibt die Routeninformation erhalten.

Die Routen sind als *LineString*, also als geordnete Liste von Koordinaten, in der Datenbank gespeichert. Um die Aggregation auf Segmente anzuwenden, wird zu jeder Route in Koordinatenform eine Route in Segmentform angelegt. Dies ist eine geordnete Liste von OSM-Segment-IDs, auf denen sich die entsprechenden Koordinaten der Route befinden (siehe Abbildung 4.9). Beide Listen haben damit die gleiche Anzahl an Elementen. Das zum Koordinatenpaar gehörige Segment wird über den Dienst Nominatim [39] durch *Reverse Geocoding* gefunden. Mehrere Koordinatenpaare eines Routenabschnittes werden so dem gleichen Segment zugeordnet. Nach der Aggregation auf Segmente liegt die Route zusätzlich in Segmentform vor. Der Datensatz enthält somit die Eigenschaften *rental_id*, Route als *geojson* und Route als *segments_list*.

Um das Ziel der k-anonymity für die gesamte Route zu erreichen, wird nun jedes Segment für die verschiedenen k-Werte 10, 20, 50, 100 und 1000 geprüft. Ein Segment ist k-anonym, wenn mindestens k-1 weitere Routen über dieses Segment verlaufen. Falls weniger als k-1 andere Personen das Segment passiert haben, wird es aus der Route entfernt und die Route an dieser Stelle durchtrennt und aufgeteilt. Sind eine größere Anzahl an Segmenten innerhalb der Route nicht k-anonym, wird

	rental_id [PK] integer	geojson jsonb	segments_list bigint[]
1	86154583	{"type": "Feature", "geometry": {"type": "LineString", "coordinates": [[12.373048, 51.363209], [12.373094, 51.363192], [12.373069, 51.3...	14639114,314639114,314639114,314639114,30312522,30312522,30312522,20294093,20294093}
2	83575085	{"type": "Feature", "geometry": {"type": "LineString", "coordinates": [[12.365645, 51.362401], [12.365781, 51.362369], [12.366617, 51.3...	{554579593,96836158,96836158,96836158,96836158,96836158,96836158}
3	82688644	{"type": "Feature", "geometry": {"type": "LineString", "coordinates": [[12.373383, 51.339481], [12.373333, 51.339728], [12.373278, 51.3...	{217286074,217286074,217286074,192089792,192089792}
4	79743352	{"type": "Feature", "geometry": {"type": "LineString", "coordinates": [[12.378462, 51.324116], [12.378563, 51.324179], [12.378588, 51.3...	{346577384,346577384,346577384,22737828,22737828,149828234,22737828}
5	85087222	{"type": "Feature", "geometry": {"type": "LineString", "coordinates": [[12.405496, 51.3463], [12.405998, 51.346272], [12.406043, 51.346...	05,780652205,780652205,780652205,780652205,7951605,7951605,337039712,7951605,7951605}
6	79712164	{"type": "Feature", "geometry": {"type": "LineString", "coordinates": [[12.413806, 51.344731], [12.414029, 51.344873], [12.414082, 51.3...	{337039728,337039728,30791930,32675739,32675739,185060532}

Abbildung 4.9.: Ausschnitt der Datenbank, die für jede Ausleihe eine ID, die Route als GeoJSON LineString und entsprechend als Segment-Liste enthält.

die Route also in zwei oder mehr Teilstücke aufgeteilt. Auf diese Weise erfolgt eine Privatisierung der gesamten Trajektorie.

Ein Segment wird auf k-anonymity überprüft, indem die Anzahl der Routen im Originaldatensatz gezählt werden, die dieses Segment beinhalten. Ist die Anzahl gleich oder größer dem gewählten k-Wert, gilt das Segment als k-anonym. Für diese Operation ist die Einrichtung eines temporären Caches vorteilhaft, in dem bereits geprüfte Segmente und ihre k-anonymity als *Boolean* gespeichert werden. So muss für Segmente, die bereits überprüft worden sind, nicht wiederholt die gesamte Datenbank abgefragt werden. In Abbildung 4.10 wird der Ablauf der Segmentaggregation als Pseudocode vereinfacht dargestellt. Das Codefragment soll dabei nur die verschiedenen Schritte darstellen und achtet nicht auf Parameter wie Effizienz. In der Implementierung werden nicht k-anonyme Segmente zunächst mit dem Wert „-1“ markiert, bevor sie aus der Route entfernt werden.

```
rental = {'rental_id': rental_id, 'coordinates': [[lat1, lon1],[lat2, lon2],..]}
all_rentals = [rental1, rental2, ...]
k_value = 10

# iterate all rentals
for rental in all_rentals:
    segments_list = []

    # get osm-segment for each coordinate pair of a rental
    for coordinate in coordinates:
        segment = get_segment(coordinate)
        segments_list.append(segment)

    # test all segments for k-anonymity
    for segment in segments_list:
        k_anonym = test_k_anonymity(segment, k_value)
        if k_anonym:
            do nothing
        else:
            segment = -1

    route = []
    temp_route = []

    # if segment marked "-1" split route
    for segment in segments_list:
        temp_route.append(segment)

        if segment == -1:
            route.append(temp_route)
            temp_route = []
```

Abbildung 4.10.: Pseudocode zur Beschreibung der Methode Segmentaggregation

Mit der Visualisierung als Heat Map werden häufig zurückgelegte Strecken hervorgehoben, da sie durch eine höhere Intensität farblich betont dargestellt werden. Nach der Aggregation auf Segmente und der Prüfung dieser Segmente auf k-anonymity, kann aus den nun privatisierten Daten bereits eine Heat Map erstellt werden. Zur Erstellung der Heat Maps muss eine Route als *LineString* (GeoJSON) an das Mapbox-Frontend (siehe Kapitel 4.2.2) weitergeleitet werden. Da die privatisierte Route nun als Liste von Segmenten vorliegt und die Segmente nicht von Mapbox verarbeitet werden können, müssen sie wieder in Listen von Koordinaten konvertiert werden. Dafür ergeben sich zwei Möglichkeiten:

Wie bereits in Abschnitt 4.2.1 erläutert, stellt ein Segment einen kurzen Abschnitt einer Straße dar und besteht aus mehreren einzelnen Koordinatenpunkten. Für jedes Segment der Segmentliste können die zugehörigen Koordinaten, bei OSM als *nodes* bezeichnet, aus der PostGIS-Datenbank abgefragt werden. Die Route wurde zuvor sowohl als Koordinatenliste als auch als Segmentliste gespeichert. Alternativ zur ersten Möglichkeit die Segmentliste durch Koordinaten zu ersetzen, kann die Liste der Koordinaten angepasst werden. Für jedes nicht k-anonyme Segment in der Route wurde der Wert „-1“ in der Segmentliste vermerkt. So können alle Koordinaten, die auf einem nicht k-anonymen Segment liegen, aus der Koordinatenliste entfernt werden. Dafür können die Indizes benutzt werden, an denen in der Segmentliste der Wert „-1“ vermerkt wurde.

Wichtig ist bei der Generierung der Routen für das Frontend zu beachten: Ein Segment kann zwischen 30 und 50 Meter lang sein. Die Start- oder Endpunkte der Routen können damit an verschiedenen Positionen auf dem Segment liegen. Um zu verhindern, dass Start- und Endpunkte sehr nah ein einem POI einer Person liegen, sollten die ersten und letzten Koordinatenpaare einer Route durch den Mittelpunkt des Segmentes, auf dem sie liegen, ausgetauscht werden. Dieser Austausch wird automatisch durchgeführt, werden die Routen durch Abfrage der entsprechenden Segment-Koordinaten aus der PostGIS-Datenbank generiert. Daher wird für diese Untersuchung die Umwandlung von Segmenten in Koordinaten gewählt. Um diesen Arbeitsschritt zu beschleunigen, wird ein *segments-nodes-index* angelegt, der für jedes bereits abgefragte Segment die entsprechenden Koordinaten in einem Cache speichert. Die abgefragten Segment-Koordinaten können zu einer Route verknüpft werden.

4.2.6. Donutmasking von nicht k-anonymen Start- und Endpunkten

Eine Aggregation auf Segmente mit Erreichen der k-anonymity stellt sicher, dass mindestens k Personen über ein Segment gefahren sind. Da allerdings keine Unterscheidung getroffen wird, ob es sich um ein Segment handelt, auf dem eine Route startet beziehungsweise endet oder ob es sich um ein Segment in der Mitte einer Route handelt, kann nicht sichergestellt werden, dass POIs wie der Start- und Endpunkt einer Route auch genügend geschützt sind. Mit einem weiteren Ansatz, dem Donutmasking, wird der Datensatz privatisiert, indem für jeden Eintrag des Datensatzes geprüft wird, ob die Start- und Endpunkte der Route k-anonym sind. Start- oder Endpunkte sind dann k-anonym, wenn mindestens k-1 andere Routen auf diesen starten oder enden.

Um k-anonymity zu erreichen, werden nicht k-anonyme Start- und Endpunkte so weit entlang der Route verschoben, bis sie auf einem k-anonymen Segment liegen. Anders als bei dem zuvor vorgestellten Ansatz in Kapitel 4.2.5 werden nur die Start- und Endpunkte geprüft und nicht die dazwischen liegenden Teile der Route. Die Start- und Endpunkte dürfen nur auf Segmente verschoben werden, für welche die Bedingung der k-anonymity erfüllt ist. Ausleihen, für die kein Segment k-anonym ist, werden aus dem Datensatz entfernt. Wie in den bereits vorangegangenen Methoden wird der Algorithmus für die k-Werte 10, 20, 50, 100 und 1000 angewandt.

```
rental = {'rental_id': rental_id, 'coordinates': [[lat1, lon1],[lat2, lon2],..]}
all_rentals = [rental1, rental2, ...]
k_value = 10

# iterate all rentals and get segment for each coordinate pair of a rental
for rental in all_rentals:
    segments_list = []

    for coordinate in coordinates:
        segment = get_segment(coordinate)
        segments_list.append(segment)

    # while segment is not k_anonym, iterate through segments_list,
    # until k-anonym condition is True
    start_segment_k_anonym, end_segment_k_anonym = False, False

    # iterate from start to end
    start_counter = 0
    while not start_segment_k_anonym:
        start_segment_k_anonym = test_k_anonymity(segments_list[start_counter],
                                                  k_value)
        if start_segment_k_anonym:
            break

        start_counter += 1

    # iterate from end to start
    end_counter = -1
    while not end_segment_k_anonym:
        end_segment_k_anonym = test_k_anonymity(segments_list[end_counter],
                                                k_value)
        if end_segment_k_anonym:
            break

        end_counter -= 1

    # cut off all not k-anonym segments from start and end of list using indices
    segments_list = segments_list[start_counter : end_counter]
```

Abbildung 4.11.: Pseudocode zur Beschreibung der Methode Donutmasking

Die k-anonymity eines Segments wird in dieser Methode geprüft, indem die Anzahl der Routen des Originaldatensatzes gezählt werden, die dieses Segment auch als Start- beziehungsweise Endpunkt haben. Ist die Zahl gleich oder größer dem gewählten k-Wert, gilt das Start- beziehungsweise Endsegment als k-anonym. Konnte der Start- beziehungsweise Endpunkt einer Route nicht auf ein k-anonymes Segment verschoben werden, bedeutet dies, dass keines der Segmente der Route k-anonym ist. Diese Routen werden aus dem Datensatz entfernt (siehe Abbildung 4.11).

Um die Daten anschließend in Form der Heat Map visualisieren zu können, werden auch hier wieder für jedes Segment die entsprechenden Koordinaten (*nodes*) abgerufen und verknüpft. Die Route wird erneut als *LineString* gespeichert und an das Frontend übergeben.

4.3. Evaluation von Privatsphäre und Nützlichkeit

Nachdem im vorangegangenen Kapitel die Implementierung aller Ansätze erläutert wurde, wird in diesem Kapitel für alle anonymisierten Datensätze eine Evaluation der Nützlichkeit und der entstandenen Privatsphäre durchgeführt. Damit kann die Performanz der implementierten LPPM evaluiert werden. Die Ansätze und Metriken zur Evaluierung dieser beiden Aspekte wurden bereits in Kapitel 2.7 und 2.8 vorgestellt. In diesem Abschnitt sollen sie auf die privatisierten Daten angewandt werden. Dazu wird zuvor nochmals anhand zweier Szenarien beschrieben, welche Aspekte bei der Evaluation der Privatsphäre von besonderer Bedeutung sein müssen. Im anschließenden Kapitel wird eine Evaluation der Privatsphäre und Nützlichkeit anhand von Visualisierungen durchgeführt, bevor daran anknüpfend die Visualisierungen zur Auswertung des Fahrradklimas vorgestellt werden.

4.3.1. Finden eines anzustrebenden Privatsphäre-Levels

Zunächst soll mit einem best-case Szenario, sowie einem worst-case Szenario dargestellt werden, welche Informationen Angreifende suchen und erlangen können, erhalten sie Zugriff auf den vorliegenden Nextbike-Datensatz. Diese theoretischen Szenarien sollen verdeutlichen, in welchen Kontext die Ergebnisse der Privatisierungs-Metriken gesetzt werden müssen.

Betrachtet wird folgender Fall: Eine Person leiht jeden Morgen ein Nextbike aus, um damit zur Arbeit zu gelangen. Angreifende können diese Person beim Ausleihen des Rades beobachten. Da die Person regelmäßig beim Ausleihen des Rades zur selben Uhrzeit zu beobachten ist, kann die Vermutung aufgestellt werden, dass sie womöglich einen oft frequentierten Ort mit dem Fahrrad aufsucht, wie beispielsweise den Arbeitsplatz. Angreifende haben also das Wissen, in welchem Umkreis die Route der Person beginnt und zu welcher Uhrzeit. Ziel der Angreifenden ist es ihre Vermutung zu bestätigen und herauszufinden, wo die Person, die jeden Morgen beim Ausleihen zu beobachten ist, mit dem Fahrrad hinfährt.

Der privatisierte Nextbike-Datensatz enthält je nach Privatisierungsmethode Routen-IDs sowie Start- oder Endpunkte der Route beziehungsweise den Routenverlauf als GPS-Koordinaten. Datum und Uhrzeit der Ausleihe sowie Kundeninformation ist in den privatisierten Datensätzen nicht enthalten, wobei hier angemerkt werden muss, dass Nextbike im Originaldatensatz eine Kunden-ID hält, diese aber vor Herausgabe des Datensatzes gelöscht hat. Datum und Uhrzeit wurden vor der Privatisierung gelöscht, da sie sensible Attribute darstellen. Das Wissen über den Zeitpunkt der Ausleihe ist so für potentielle Angreifende keine verwendbare Information. Mit dem vorliegenden privatisierten Datensatz gelingt es im best-case Szenario nicht, die entsprechende Route anhand ihres Startpunktes im Datensatz zu finden und somit den Ort zu ermitteln, den die Person nach dem Ausleihen des Fahrrades aufsucht.

Dies kann je nach Privatisierungsmethode auf mehrere Gründe zurückzuführen sein: Bei der Aggregation auf Zentroide wurden die Start- und Endpunkte einer Route auf einen Stadtteil beziehungsweise ein Postleitzahlgebiet abgebildet. Folglich ist der genaue Standort des Rades bei der Ausleihe nicht mehr im Datensatz vermerkt. Die Information des beobachteten, genauen Startpunktes der

Route wird damit nutzlos. Mithilfe des Donutmaskings wurden die Start- und Endpunkte einer Route an einen anderen Punkt verschoben, an welchem, je nach k-Wert, 10, 20, 50, 100 oder 1000 weitere Routen starten oder enden. Angreifende müssen die Route der von ihnen beobachteten Person also aus mindestens k-1 anderen Routen differenzieren können. Im besten Fall enden Routen, die an demselben Punkt starten, an verschiedenen Punkten in der Stadt. Es wird also kein Muster deutlich, dass Routen vermehrt an derselben Stelle enden. Die von Angreifenden vermutete, oft frequentierte Route der Person lässt sich somit nicht finden.

Im worst-case Szenario ist der Startpunkt der Route der beobachteten Person zwar in einer Gruppe von k-1 Startpunkten gut geschützt, allerdings lässt sich ein deutliches Muster erkennen, dass eine Vielzahl der an diesem Punkt startenden Routen an einem bestimmten Ort enden. Das Muster wird deutlich, da ein Strecke deutlich häufiger gefahren wurde, als die übrigen Strecken. Hieraus können Angreifende mutmaßen, dass es sich um die morgens beobachtete Person handeln muss. Problematisch ist außerdem, dass zur Privatisierung des Nextbike-Datensatzes eine Routen-ID und keine Personen-ID benutzt wurde. Generiert also eine Person viele Routen in einem bestimmten Bereich, kann der erreichte k-Wert für ein Segment zwar sehr hoch sein, doch ist er im Endeffekt trügerisch, da es sich immer um dieselbe Person handelt, die den k-Wert steigen lässt.

Mithilfe der Charakteristika von Privatsphäre-Metriken sollen im anschließenden Kapitel die folgenden Fragen beantwortet werden, um festlegen zu können, wie hoch das Level an Privatsphäre mindestens sein muss.

1. Welche Ziele können Angreifende verfolgen?
   Angreifende, die sich für einen Datensatz wie den vorliegenden interessieren, suchen womöglich POIs der Personen wie Wohnadresse, Arbeitsadresse oder regelmäßig gefahrene Strecken (Fahrmuster) oder die Uhrzeiten, zu denen die Strecken zurückgelegt werden. Diese Informationen reichen bereits aus, um herauszufinden, an welchen Orten sich Personen aufhalten.

2. Welche Möglichkeiten haben Angreifende?
   Angreifende können durch Beobachtungen von Personen beim Ausleihen eines Rads Hintergrundwissen erlangen. Ferner können sie durch andere öffentliche Ressourcen (mit überschneidenden Inhalten, beispielsweise Navigations-Apps) Verknüpfungen zum vorliegenden Datensatz schließen und zusätzliche Informationen erlangen.

4.3.2. Quantitative Evaluation der Privatsphäre anhand von Metriken

Die sensiblen Attribute des vorliegenden Datensatzes stellen die Trajektorie einer Ausleihe und deren Start- und Endpunkte dar. Der Datensatz enthält keine Informationen über einzelne Personen wie eine Kunden-ID oder Adressen. Daher wird das Hauptmerkmal der Evaluation darauf gelegt, ob die Standortdaten einzelner Personen im Datensatz ein hohes Maß an Privatsphäre besitzen.

Zunächst soll aus den in Kapitel 2.7 vorgestellten Metriken eine Auswahl getroffen werden, mit denen die Daten evaluiert werden. Viele der zuvor vorgestellten Metriken arbeiten mit einem Angreifermodell, welches Möglichkeiten und Charakteristika der Angreifenden beschreibt. Für diese Untersuchung wird kein Angreifermodell erstellt, da es derzeit keine öffentlichen Ressourcen mit

ausreichend thematischer Überschneidung gibt, die Angreifende sinnvoll mit dem in dieser Arbeit entstandenen privatisierten Datensatz verlinken können. Daher fallen einige Metriken bereits aus den Betrachtungen. Wichtig bei der Wahl der Metriken ist außerdem die Betrachtung der Eigenschaften, die mithilfe der Privatisierung geschützt werden sollen. Werden, wie in der vorliegenden Arbeit, Geo-Daten betrachtet, muss eine Metrik bewerten, ob Angreifende eine Person anhand des Standortes oder den Standort anhand der Person identifizieren können [24] (siehe hierzu auch die vorgestellten Szenarien in Kapitel 4.3.1).

Anonymity Set Size aus der Oberkategorie *Certainty* berechnet die Größe der Gruppe, von der eine einzelne Person un-unterscheidbar ist (siehe Kapitel 2.7). Diese Metrik wird für diese Evaluation verwendet, um die Anzahl der Start- und Endpunkte auf einem Segment zu berechnen. Mithilfe der Re-Identifikationswahrscheinlichkeit wird geprüft, ob das erreichte Level an Privatsphäre über, unter oder gleich dem angestrebten Level ist. Laut dem in Kapitel 4.3.1 beschriebenen best-case Szenario enden Routen, die an demselben Punkt starten, an verschiedenen Punkten in der Stadt. Diese geographische Verteilung der Daten wird mit der Metrik *Data Similarity* im folgenden überprüft.

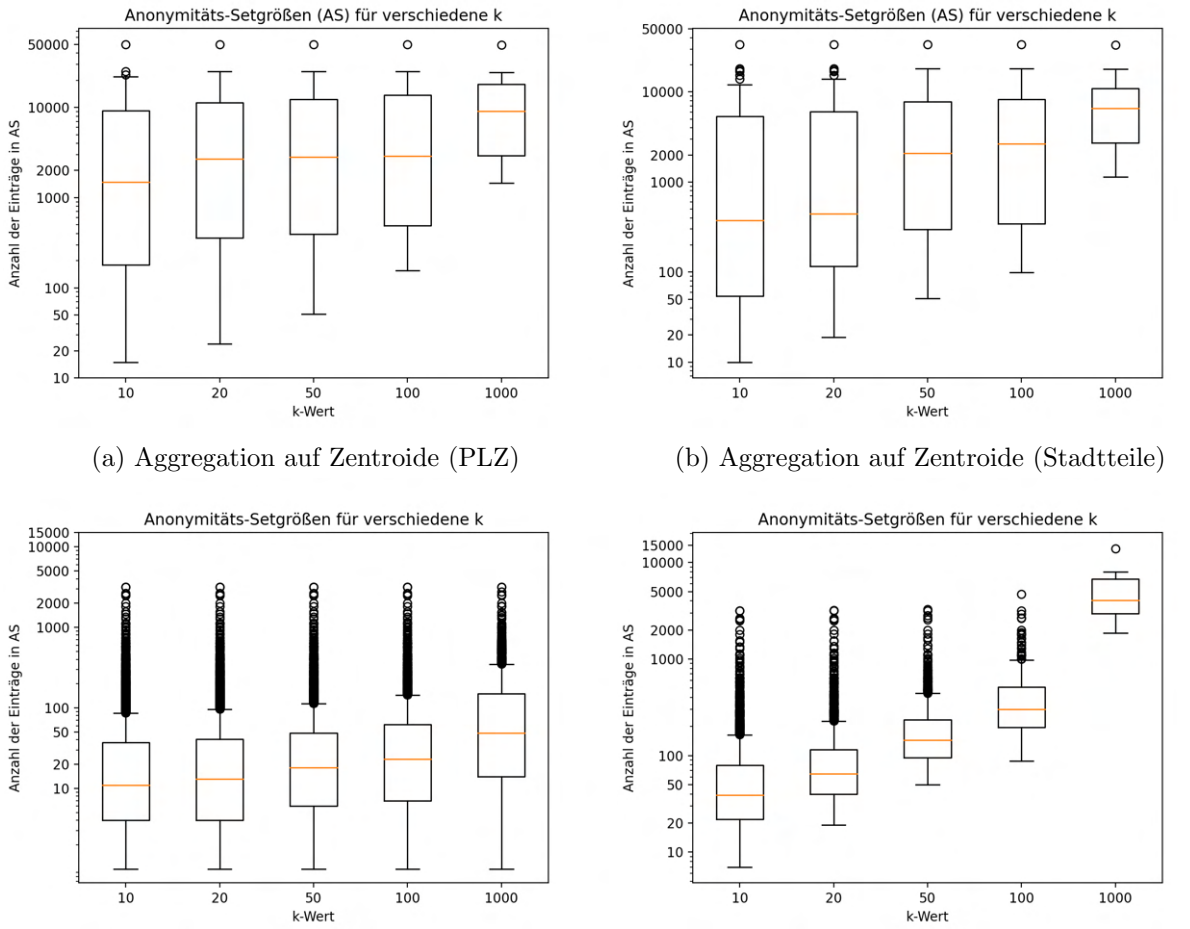
Bevor im anschließenden Abschnitt die Ergebnisse der Metriken vorgestellt werden, erfolgt eine kurze Einschätzung, welche Ergebnisse zu erwarten sind. Hierbei werden Vermutungen aufgestellt, die im Anschluss geprüft werden. Da die Methode der Aggregation auf Zentroide die Daten stark vergröbert, kann von dem dabei entstehenden Datensatz anschließend ein hohes Maß an Privatsphäre erwartet werden. Durch die Aggregation auf Segmente und das Donutmasking bleibt die Routeninformation erhalten. Daher werden die Ergebnisse der Privatsphäre-Metriken hier schlechter ausfallen. Interessant ist der Vergleich der Privatsphäre zwischen den Ansätzen mit und ohne Erhaltung der Routeninformation, da die Aggregation auf Zentroide zwar einen hohen Einfluss auf die Qualität der Daten hat, dafür aber ein deutlich höheres Maß an Privatsphäre zeigen sollte.

Anonymity Set Size

Die Größe des Anonymitäts-Sets für Startpunkte spiegelt wider, wie viele Routen an demselben Ort starten. Wissen Angreifende den ungefähren Routenstartpunkt einer Person, ist die Größe des Anonymitäts-Sets entscheidend, ob Angreifende beim Finden der Zielperson erfolgreich sind oder nicht (siehe beschriebene best- und worse-case Szenarien). Umso größer das Anonymitäts-Set ist, umso schwerer ist die Unterscheidung der darin enthaltenen Startpunkte. Ist das Anonymitäts-Set zu klein, ist die Wahrscheinlichkeit größer, dass Angreifende ihre Zielperson finden. Auch für Trajektorien kann die Größe des Anonymitäts-Sets gemessen werden. Dabei wird für jedes Segment einer Trajektorie geprüft, wie viele andere Personen dieses Segment passiert haben. Enthält die Trajektorie Segmente, deren Anonymitäts-Set zu klein sind, enthält diese Trajektorie und so auch der Datensatz selten gefahrene Routen, die somit Rückschlüsse auf Bewegungsmuster einzelner Personen zulassen.

Um das angestrebte Privatsphäre-Level zu erreichen, muss ein Anonymitäts-Sets für privatisierte Daten mindestens den bei der Privatisierung gewählten k-Wert an Einträgen zu besitzen. Der originale Nextbike-Datensatz enthält sowohl für Routenstartpunkte als auch für Trajektorien Einträge, für die der k-Wert = 1 ist. Die sensiblen Attribute dieser Einträge sind folglich ungeschützt.

Nach der Privatisierung müssen diese in Anonymitäts-Sets mit dem gewählten k-Wert der Methode enthalten sind oder aus dem Datensatz entfernt worden sind. Sind Anonymitäts-Sets mit weniger als k Einträgen im Datensatz enthalten, war die Privatisierung nicht erfolgreich. In einem Datensatz, der Personen-IDs (wie beispielsweise Kunden-IDs) enthält, würde das Anonymitäts-Set ein Set aus Personen repräsentieren. Da im vorliegenden Datensatz keine Kunden-ID vorhanden ist, enthält das Anonymitäts-Set Standorte. Anonymitäts-Sets werden anschließend für Startpunkte einer Route sowie deren Trajektorie untersucht. Das Anonymitäts-Set für einen Startpunkt einer Route s ist das Set von Startpunkten, von denen Angreifende s nicht unterscheiden können. Ein Anonymitäts-Set entspricht einer Äquivalenzklasse.



(a) Aggregation auf Zentroide (PLZ)

(b) Aggregation auf Zentroide (Stadtteile)

(c) Aggregation auf Segmente

(d) Donutmasking von Start- und Endpunkten

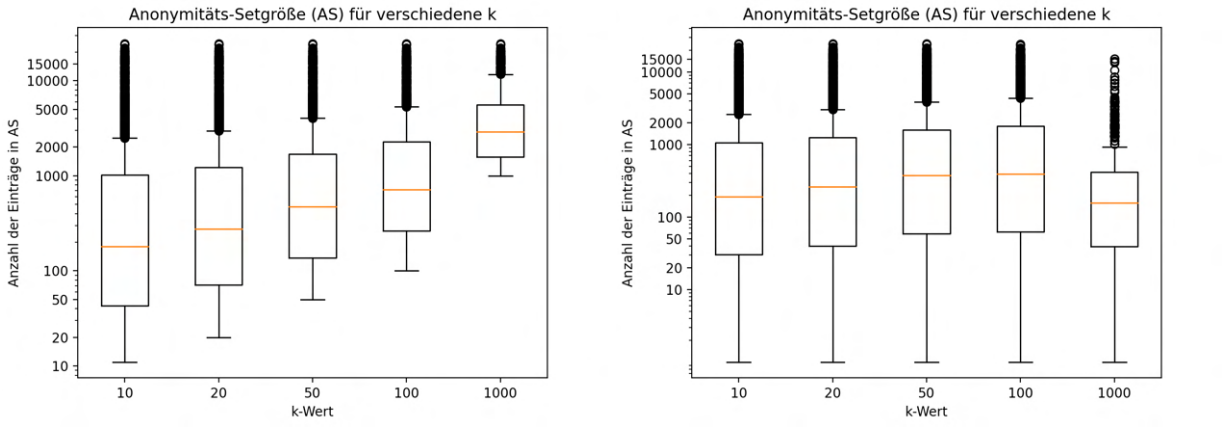
Abbildung 4.12.: Anonymitäts-Setgrößen für Startpunkte der verschiedenen LPPM. Logarithmische Darstellung der Achsen.

Das Anonymitäts-Set für Startpunkte zeigt, dass die Größe der Anonymitäts-Sets im Interquartilbereich für die meisten der angewandten LPPM höher als der gewählte k-Wert ist und so ein hohes Maß an Privatsphäre aufweist (siehe Abbildung 4.12). Bezüglich der Aggregation auf Segmente ist auffällig, dass für jeden k-Wert Sets existieren, in denen die Anzahl der Einträge im Set deutlich unter k liegt (siehe Abbildung 4.12, (c)). Dies ist auf den Algorithmus zurückzuführen, der jedes Segment auf die Gesamtanzahl seiner „Passierenden“ überprüft und nicht auf die Anzahl der startenden oder endenden Routen auf dem Segment. Für die Aggregation auf Segmente machen die Ergebnisse der Metrik deutlich, dass POIs, wie Wohnadressen, nicht ausreichend geschützt werden,

da viele Anonymitäts-Sets kleiner als der gewählte k-Wert sind. Ab k > 50 liegt die Größe der Sets deutlich unter k.

Bei der Aggregation auf Zentroide werden im Schnitt größere Anonymitäts-Sets gebildet als für die Segmentaggregation und das Donutmasking. Dies ist unter anderem darauf zurückzuführen, dass es weniger Postleitzahlgebiete und Stadtteile als Straßensegmente und Startpunkte gibt. Daher existieren bei der Aggregation auf Zentroide weniger Äquivalenzklassen, auf die die Routen aufgeteilt werden können. So besitzen die einzelnen Klassen beziehungsweise Sets mehr Einträge. Dennoch spricht ein größeres Anonymitäts-Set für mehr Privatsphäre. Für die k-Werte 10, 20, 50 und 100 ist die erreichte Privatsphäre somit deutlich höher als für die Segmentaggregation oder das Donutmasking, da die Anzahl der un-unterscheidbaren Einträge höher ist.

Alle angewandten LPPM produzieren mit steigendem k-Wert größere Anonymitäts-Sets, was für steigende Privatsphäre spricht. Festzustellen ist, dass die Aggregation auf Segmente diesen Trend zwar spiegelt, dennoch hierbei auch Sets entstehen, die weniger als k Einträge besitzen. Wie erwartet erzeugen die am stärksten vergröbernden Methoden der Aggregation auf Zentroide große Anonymitäts-Sets. Doch auch das Donutmasking schafft Sets der Mindestgröße k. Im Schnitt liegt hierbei die Größe der Anonymitäts-Sets sogar über k.



(a) Aggregation auf Segmente

(b) Donutmasking von Start- und Endpunkten

Abbildung 4.13.: Anonymitäts-Setgrößen für Trajektorien. Logarithmische Darstellung der Achsen.

Es gilt außerdem die Größe der Anonymitäts-Sets für Trajektorien zu bestimmen (siehe Abbildung 4.13). Da für die Aggregation auf Zentroide die Trajektorien aus dem Datensatz entfernt wurden, kommt diese Metrik nur für die Methoden Segmentaggregation und Donutmasking zur Anwendung. Es wird deutlich, dass die beiden Methoden sehr unterschiedliche Level an Privatsphäre für Trajektorien bieten. Während für die Segmentaggregation mit steigendem k-Wert auch die Größe der Anonymitäts-Sets steigt, bleiben diese für das Donutmasking weitestgehend gleich groß. Für die Segmentaggregation wird mindestens der gewählte k-Wert der Methode erreicht, die Größe der Anonymitäts-Sets im Interquartilbereich liegt hier allerdings deutlich über dem k-Wert. Diese Methode erreicht für die Privatsphäre der Trajektorien folglich sehr gute Ergebnisse. Für das Donutmasking existieren für jeden k-Wert Anonymitäts-Sets, deren Anzahl an Einträgen kleiner als k ist. Die Methode bietet wenig bis keinen Schutz der Privatsphäre der Trajektorien. Dieses Ergebnis bildet das Gegenstück zu den in Abbildung 4.12 gezeigten Ergebnissen der Anonymitäts-Setgrößen

für Startpunkte für Segmentaggregation und Donutmasking. Während die Segmentaggregation für Startpunkte wenig Schutz liefert, zeigt sie für den Schutz der Trajektorien umso bessere Ergebnisse. Dieses Resultat lässt sich für das Donutmasking umkehren: Dieses bietet für Startpunkte einen hohen Schutz, kann aber die Privatsphäre für Trajektorien nicht wahren.

Mit steigenden k-Werten nimmt die Anzahl der verbleibenden Routen im Datensatz ab, da viele Einträge unterdrückt beziehungsweise aus dem Datensatz entfernt werden. Die Studie von Jin et al. [23] hat analysiert, wie sich die Größe des Originaldatensatzes auf die erreichte Privatsphäre auswirkt. Indem sie eine sogenannte *linkage accuracy* (dt. Exaktheit der Verknüpfung) untersuchen, prüfen sie die Möglichkeit, einer originalen Trajektorie D eine Trajektorie D' aus dem anonymisierten Datensatz zuordnen zu können. Je höher diese Exaktheit ist, umso schlechter ist der Schutz der Privatsphäre des Datensatzes. Ihre Studie ist zu dem Ergebnis gekommen, dass in größeren Datensätzen tendenziell eine niedrigere Exaktheit der Verknüpfung zu beobachten ist. Ein größerer Datensatz kann folglich ein höheres Level an Privatsphäre erreichen. Existieren im privatisierten Datensatz ausreichend große Anonymitäts-Sets, ist es schwierig, eine erfolgreiche Verknüpfung zwischen originalem und anonymisierten Datensatz zu finden [23]. Je höher der k-Wert eines LPPM, umso größer sind die kreierten Anonymitäts-Sets und umso mehr Einträge stellen einen potentiellen Kandidat für die Verknüpfung mit dem Original dar. Handelt es sich um einen großen Originaldatensatz, können höhere k-Werte zur Anonymisierung eingesetzt werden, ohne Einträge aus dem Datensatz entfernen zu müssen. Für k = 1000 müssen bei dem hier verwendeten Nextbike-Datensatz viele Einträge gelöscht werden, um k-anonymity zu erreichen (siehe auch Kapitel 4.3.4)

Re-Identifikationswahrscheinlichkeit

Die maximal zu erreichende Re-Identifikationswahrscheinlichkeit einer Person in einem privatisierten Datensatz berechnet sich durch $\frac{1}{k}$, die tatsächlich erreichte durch $\frac{1}{f}$, wobei k der k-Wert des Privatisierungsmechanismus und f die durchschnittliche Anzahl der Einträge der Anonymitäts-Set nach der Privatisierung ist. Die erreichte Re-Identifikationswahrscheinlichkeit ist höher als die maximal zu erreichende, wenn Anonymitäts-Sets weniger als k Einträge haben. Die Privatsphäre ist dann unzureichend geschützt.

Methode	k = 10	k = 20	k = 50	k = 100	k = 1.000
Aggregation auf PLZ	4.775	4.773	4.770	4.766	4.592
Aggregation auf Stadtteile	2.436	2.434	2.428	2.423	2.323
Aggregation auf Segmente	44	47	53	62	133
Donutmasking	82	114	217	422	4.979

Tabelle 4.1.: Durchschnittliche Größe der Anonymitäts-Sets für die verschiedenen Methoden zur Berechnung der Re-Identifikationswahrscheinlichkeit

Wird die durchschnittliche Re-Identifikationswahrscheinlichkeit für die Größe der Anonymitäts-Sets der Startpunkte betrachtet, liegt diese für die Mechanismen Aggregation auf Zentroide und Donutmasking weit unter dem Maximum: Für die Aggregation auf Postleitzahlen bei einem k-Wert von

10 ist die maximal erlaubte Re-Identifikationswahrscheinlichkeit entsprechend dem k-Wert $\frac{1}{10}$, also 10 %. Die tatsächliche Re-Identifikationswahrscheinlichkeit ist allerdings deutlich niedriger und liegt im Durchschnitt bei $\frac{1}{4775}$, also 0,021 %. Die Zahl im Nenner ergibt sich aus der durchschnittlichen Anzahl der Einträge im Anonymitäts-Set (siehe Tabelle 4.1). Für die Aggregation auf Segmente liegt dieser Wert ab k = 50 über der maximal zu erreichenden Wahrscheinlichkeit. Dies wird für k = 1000 am deutlichsten: Die zu erreichende maximale Re-Identifikationswahrscheinlichkeit liegt bei $\frac{1}{1000}$, also 0,1 %, die tatsächlich erreichte bei $\frac{1}{133}$. Sie liegt mit 0,8 % folglich deutlich darüber. Es ist daher leichter, Personen in diesem Datensatz anhand der Startpunkte zu re-identifizieren.

Analog zu den Ergebnissen der *Anonymity Set Size* besitzen die Methoden Aggregation auf Stadtteile und Postleitzahlgebiete durchschnittlich eine sehr hohe Anzahl an Einträgen in ihren Äquivalenzklassen. Daher liegt die Re-Identifikationswahrscheinlichkeit hier deutlich unter der maximal zu erreichenden. Dies zeigt zum einen eine sehr hohe Privatsphäre der Daten. Die Wahrscheinlichkeit, eine Person innerhalb der großen Äquivalenzklassen zu re-identifizieren, liegt für k = 10 bei 0,021 %. Im Vergleich mit den Methoden Aggregation auf Segmente und Donutmasking macht es aber deutlich, wie stark die Information bei der Zentroidaggregation vergröbert wurde und wie viel Information verloren gegangen ist, da die Anzahl der Einträge für die Zentroidaggregation im Vergleich mindestens 10-fach höher ist. Indem bei Segmentaggregation und Donutmasking auf kleinteiligere Abschnitte der Route, die Segmente, aggregiert wird, kann mehr Informationsgehalt erhalten werden. Beim Donutmasking bleibt die erreichte Re-Identifikationswahrscheinlichkeit für alle k-Werte unter der maximal zu erreichenden. Bei der Segmentaggregation ist bereits der k-Wert 50 kritisch, da hierbei die maximale Re-Identifikationswahrscheinlichkeit von 2 % mit 1,9 % nur knapp unterschritten wird und alle folgenden k-Werte die maximale Re-Identifikationswahrscheinlichkeit überschreiten. Hier wird deutlich, dass für die Segmentaggregation ab k = 50 die Wahrscheinlichkeit gegeben ist, anhand eines Startpunktes eine Person re-identifizieren zu können.

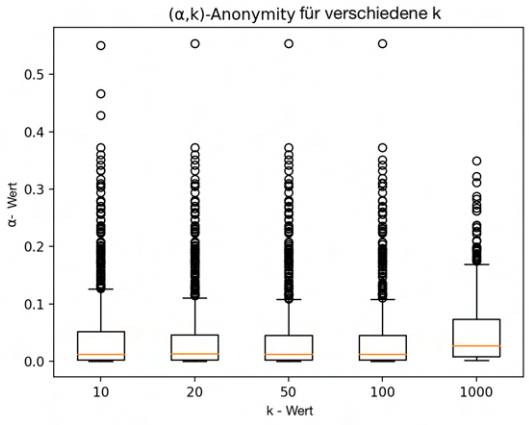
Data Similarity

Wie im worst-case Szenario beschrieben, können viele Routen an demselben Punkt starten, was den einzelnen Startpunkt un-unterscheidbar macht. Enden eine Vielzahl dieser Routen allerdings an einem Ort, lässt sich daraus ein Muster erkennen, dass diese Routen womöglich von einer Person gefahren werden. Das bedeutet, ein großes Anonymitäts-Set für Startpunkte allein ist nicht ausreichend, um die Privatsphäre der Personen zu gewährleisten. Ist das Anonymitäts-Set aus Startpunkten $\geq k$, aber die geographische Verteilung der zu den Startpunkten gehörenden Endpunkte nicht, ist dennoch eine oft gefahrene Route beziehungsweise eine einzelne Person aus dem Anonymitäts-Set deutlich erkennbar. Um diese geographische Verteilung der Endpunkte zu überprüfen, wird zusätzlich die Metrik *(α, k) - anonymity* verwendet. Dieser Ansatz gruppiert Startpunkte zu Äquivalenzklassen mit mindestens k Einträgen. Anschließend prüft er die Verteilung der Endpunkte innerhalb der Äquivalenzklasse.

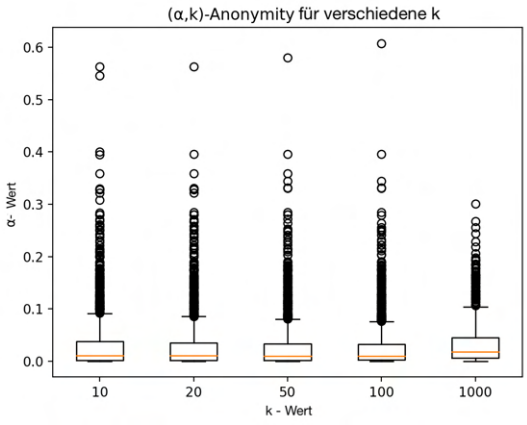
Äquivalenzklassen bestehen aus Einträgen eines Datensatzes, die dieselben Werte als *Quasi-identifier* besitzen. Für diese Betrachtung werden die Startsegmente zur Bildung der Äquivalenzklassen benutzt. Die Größe einer Äquivalenzklasse entspricht folglich der Anzahl der

Routen, die das gleiche Startsegment besitzen. Um anschließend innerhalb der Äquivalenzklassen die geographische Verteilung der Endpunkte zu prüfen, wird die Verteilung der Endsegmente betrachtet. Um ein gutes Ergebnis zu erreichen, darf kein Endsegment häufiger als ein anderes auftreten. So kann verhindert werden, dass einzelne Routen und so Personen erkennbar werden.

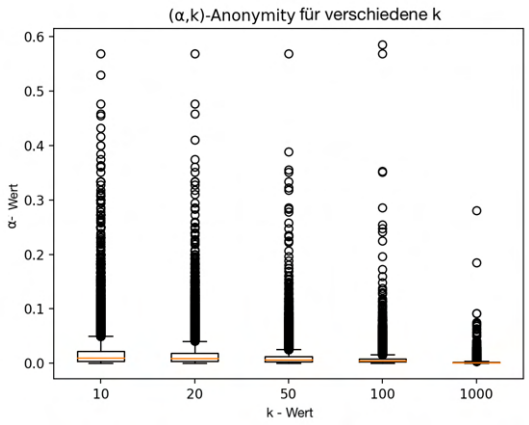
Die *(α, k) - anonymity* errechnet die Frequenz der Endsegmente mit einem α-Wert zwischen 0 und 1 und summiert die Ergebnisse für alle Einträge in der Klasse auf. So ergibt sich die Gesamtbilanz der Äquivalenzklasse. Dabei gilt, je häufiger ein Endsegment in einer Äquivalenzklasse vorkommt, umso niedriger ist der Schutz des sensiblen Wertes, in diesem Fall des Routenendpunktes, und umso höher wird der Wert der Metrik. Es gilt einen niedrigen Wert für α anzustreben.



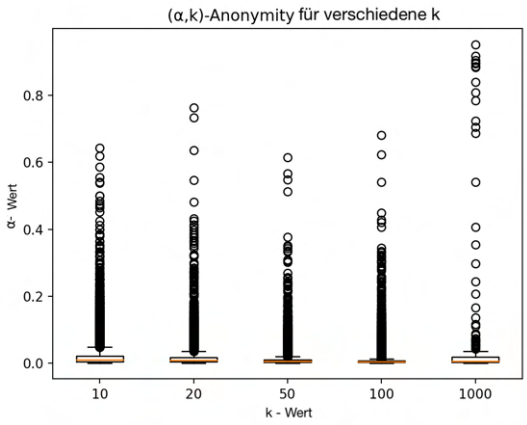
(a) Aggregation auf Zentroide (PLZ)



(b) Aggregation auf Zentroide (Stadtteile)



(c) Aggregation auf Segmente



(d) Donutmasking von Start- und Endpunkten

Abbildung 4.14.: (α, k)-anonymity für die vorgestellten LPPM ohne festgelegten Grenzwert für α.

Das Donutmasking erreicht als einziger LPPM den α-Maximalwert von 1 für k = 1000 (siehe 4.14, (d)). Alle anderen LPPM erreichen maximal einen α-Wert von 0,6. Verteilungen über einem α-Wert von circa 0,2 werden bereits als Ausreißer gekennzeichnet (siehe Abbildung 4.14, (a)). Je niedriger der Wert für α, umso homogener ist die Verteilung. Das bedeutet, umso mehr unterschiedliche Endsegmente existieren innerhalb einer Äquivalenzklasse. Der Interquartilbereich der meisten Boxplots befindet sich zwischen den Werten 0,0 und 0,05. Dies spricht für eine gute Verteilung der Endsegmente in den Äquivalenzklassen und somit für eine geringe Gefahr, dass aus Routenstartpunkten einzelne Routen erkennbar werden.

Für die Aggregation auf Segmente ist wichtig zu beachten, dass viele Startsegmente im Datensatz erhalten bleiben, die nicht k-anonym sind (siehe Abbildung 4.12, (c)). Die nicht k-anonymen Segmente werden bei der Auswertung der Verteilung durch diese Metrik nicht betrachtet, da sie die Bedingung nicht erfüllen, für alle Äquivalenzklassen mindestens k Einträge zu besitzen (siehe 2.7, Formel 2.2). Dieser Aspekt muss beim Betrachten der Abbildungen berücksichtigt werden.

Dennoch zeigt sich eine breite Verteilung der Werte in den Äquivalenzklassen. Die hohe Diversität von Routenendpunkten ist hierbei eine Eigenschaft des Datensatzes, welche dieser bereits im Vorhinein besaß und welche nicht gezielt durch den Privatisierungsmechanismus erreicht wurde. Wird bei der Privatisierungen noch nicht auf diesen Aspekt geachtet, ist es wichtig den Datensatz nach der Privatisierung auf diese Eigenschaft zu überprüfen. Nur so kann die Enthüllung von einzelnen Attributen verhindert werden. Hätte die Metrik eine schlechtere Verteilung der Werte aufgezeigt, so hätte eine Anpassung des Privatisierungsmechanismus erfolgen müssen beziehungsweise es hätte eine Lösung für die entsprechenden Daten gefunden werden müssen.

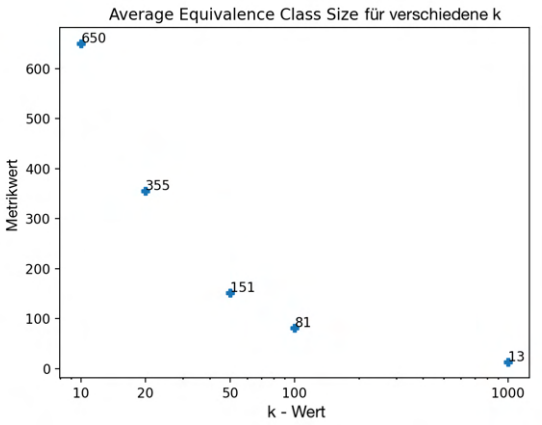
4.3.3. Quantitative Evaluation der Nützlichkeit anhand von Metriken

Um den Trade-off zwischen Privatisierung und Nutzen der Daten auszuwerten, muss zunächst die Nützlichkeit der Daten nach der Privatisierung evaluiert werden. Die Metriken für diese Kalkulation wurden in Kapitel 2.8 vorgestellt. Die *Average Equivalence Class Size* wird verwendet, um die Verteilung der Originaldaten auf Äquivalenzklassen in den privatisierten Datensätzen zu berechnen. Im besten Fall werden dabei gleichmäßig verteilte Daten in Äquivalenzklassen der Größe k erreicht. Um den entstandenen Informationsverlust messen zu können, wird die *Discernibility Metric* angewandt. Diese arbeitet mit der Grundannahme, je größer eine Äquivalenzklasse ist, umso mehr Information ist verloren gegangen beziehungsweise umso gröber ist die Information geworden.

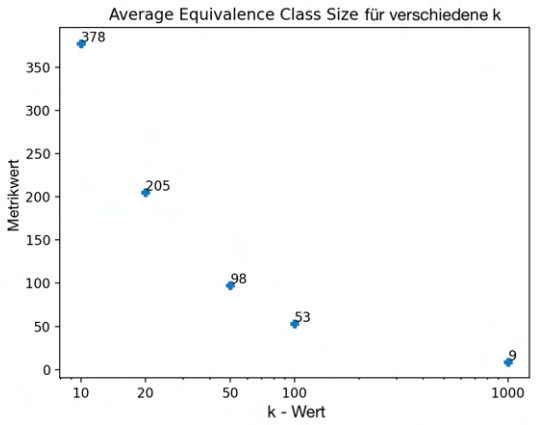
Average Equivalence Class Size

Diese Metrik strebt eine gleichmäßige Verteilung von Daten auf Äquivalenzklassen an. Somit würde der bestmögliche Wert für die Metrik in einem Ergebnis von 1 resultieren. Höhere Werte als 1 deuten darauf hin, dass die Einträge auf wenige Äquivalenzklassen verteilt wurden. Werte unter 1 bedeuten, dass die Einträge auf sehr viele Äquivalenzklassen aufgeteilt wurden. Die Äquivalenzklassen stellen dabei die Start- oder Endsegmente einer Route dar.

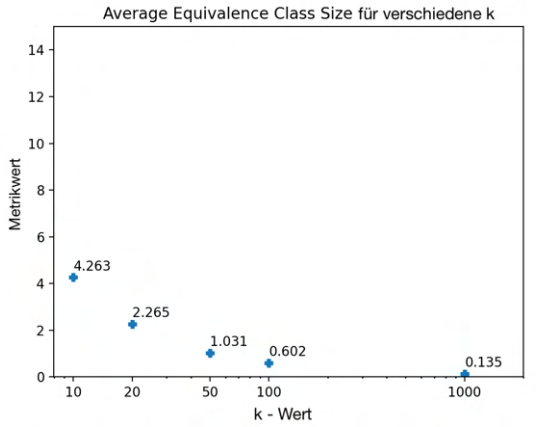
Für die Nützlichkeit der Daten nach der Aggregation auf Zentroide ist ein interessanter Verlauf zu erkennen (siehe Abbildung 4.15, (a) und (b)): Zunächst ist sehr auffällig, dass sich die Ergebnisse der Metrik sehr weit vom Optimalwert 1 befinden und sich lediglich das Ergebnis für k = 1000 dem Optimum nähert. Die hohen Metrikwerte können mit der festen Zahl der Äquivalenzklassen begründet werden. Wird auf Postleitzahlen oder Stadtteile aggregiert, sind die möglichen Äquivalenzklassen, denen ein Wert zugeteilt werden kann, entweder Postleitzahlbereiche oder Stadtteile. Die Anzahl der möglichen Äquivalenzklassen hat also eine obere Grenze, unabhängig vom k-Wert der Privatisierungsmethode. So gibt es in Leipzig 96 Stadtteile und 49 Postleitzahlcodes. Je mehr Einträge der Originaldatensatz hat, desto mehr Einträge werden auf dieselbe Zahl an Äquivalenzklassen aufgeteilt. Um für die niedrigen k-Werte bei der Zentroidaggregation einen besseren Metrikwert zu



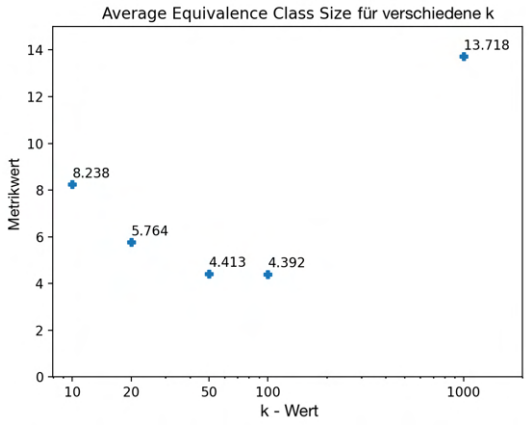
(a) Aggregation auf Zentroide (PLZ)



(b) Aggregation auf Zentroide (Stadtteile)



(c) Aggregation auf Segmente



(d) Donutmasking von Start- und Endpunkten

Abbildung 4.15.: Durchschnittliche Größe der Äquivalenzklassen *(Average Equivalence Class Size)* für Startpunkte. Logarithmische Darstellung der Achsen.

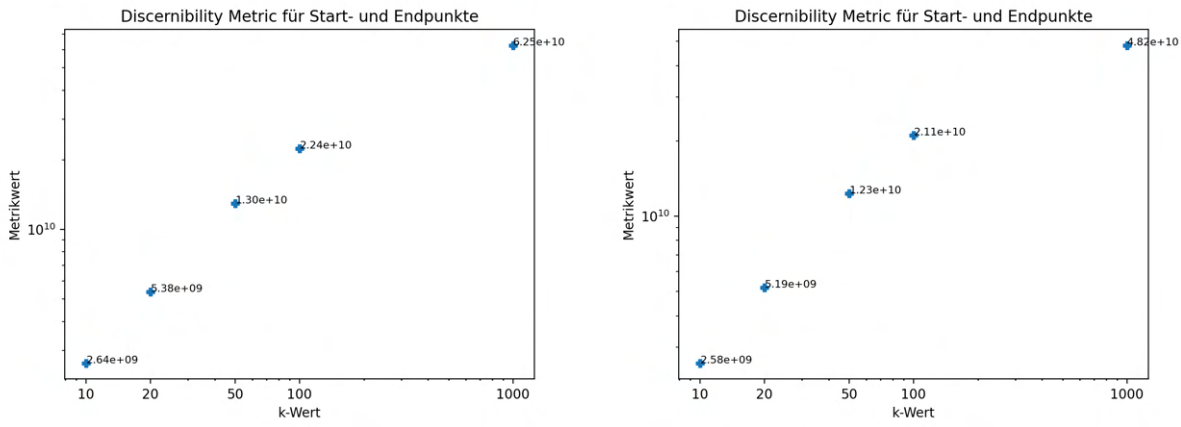
erhalten, müssten also mehr Äquivalenzklassen existieren, auf die die Einträge des Datensatzes aufgeteilt werden können, deren Anzahl ist in diesem Szenario begrenzt. Für höher werdende k-Werte wie 100 und 1000 müssen für einen guten Metrikwert wenig Äquivalenzklassen mit vielen Einträgen existieren. Für die Zentroidaggregation werden für k = 1000 zunächst viele Einträge des Datensatzes in Äquivalenzklassen eingeordnet, deren Größe anschließend nicht mindestens k entspricht. Diese Äquivalenzklassen werden aus dem Datensatz entfernt, so sinkt die Anzahl der Äquivalenzklassen. Die verbleibenden Äquivalenzklassen besitzen viele Einträge, sodass sich ein besseres Ergebnis der Metrik ergibt. So kann der absteigende Verlauf der Kurve begründet werden.

Auch die Aggregation auf Segmente folgt dem Trend: für steigende k wird der Metrikwert geringer. Dennoch sind die Werte insgesamt deutlich näher an dem Optimalwert 1, was generell als bessere Datenqualität gedeutet werden kann. Bei der Aggregation auf Segmente bleibt die Information der genauen Route erhalten und die Anzahl der Äquivalenzklassen ist nicht beschränkt. So kann der bessere Metrikwert erklärt werden. Ein fast optimaler Wert von 1,031 wird bei k = 50 erreicht. Auch für k = 100 ist der Wert 0,602 nicht weit vom Optimum entfernt. Für k = 1000 liegt der Wert fast bei 0, was darauf hin deutet, dass für die Anzahl der Einträge, die die Äquivalenzklassen im besten Fall haben sollen (= 1000), die Anzahl der Äquivalenzklassen zu hoch ist.

Die Ergebnisse der Metrik für die Methode Donutmasking zeigt einen abweichenden Kurvenverlauf: für k = 100 stagniert die vorher fallende Kurve und steigt anschließend an. Für k-Werte von 50 und 100 werden die besten Werte erreicht. Für k = 1000 werden zu wenig Äquivalenzlassen gebildet. Folglich zeigt die Metrik hier den schlechtesten Wert für diese Methode. Der Startpunkt einer Route muss hier so weit verschoben werden, bis er auf einem Segment mit 1000-1 anderen liegt. Diese Bedingung kann für viele Einträge nicht erfüllt werden, daher werden sie aus dem Datensatz entfernt. Es ergeben sich nur wenige Äquivalenzklassen, die die Bedingung erfüllen können. Der starke Anstieg für k = 1000 kann so begründet werden. Dass die Metrikwerte für Segmentaggregation und Donutmasking sich vor allem für den k-Wert von 1000 so stark unterscheiden, kann auf den Algorithmus zurückgeführt werden. Während bei der Segmentaggregation geprüft wird, wie viele andere Routen über ein Segment verlaufen sind, wird beim Donutmasking spezifisch geprüft, wie viele andere Routen auf dem Segment starten oder enden. Dementsprechend werden solange Segmente entlang der Route gelöscht, bis ein k-anonymes Segment erreicht wird. Dadurch werden mehr Daten gelöscht, je höher k wird.

Discernibility Metric

Eine weitere Metrik zur Evaluation der Nützlichkeit der Daten ist die sogenannte *Discernibility Metric* (dt. Un-unterscheidbarkeitsmetrik). Diese Metrik soll den Informationsverlust des transformierten Datensatzes messen, indem sie jedem Eintrag des Datensatzes einen Strafwert entsprechend der Größe der Äquivalenzklasse zuweist. Je höher der Wert der Metrik, umso höher ist der Informationsverlust und umso schlechter die Datenqualität.



(a) Aggregation auf Segmente

(b) Donutmasking der Start- und Endpunkte

Abbildung 4.16.: Ergebnisse der *Discernibility Metric* für Aggregation auf Segmente und Donutmasking. Logarithmische Darstellung der Achsen.

Die Kurven für die Methoden Segmentaggregation und Donutmasking verlaufen wie erwartet (siehe Abbildung 4.16): mit steigendem k-Wert steigt der Wert der Metrik. Insgesamt steigen die Werte der Metrik für beide Methoden exponentiell an, je größer der k-Wert wird. Ein hoher Metrikwert repräsentiert einen hohen Informationsverlust. Dieser Verlust entsteht durch die Vergröberung der Routen. Je höher der k-Wert wird, desto größer werden die Äquivalenzklassen, in die die Segmente oder Start- und Endpunkte eingeordnet werden und desto höher wird der Metrikwert.

Für alle k-Werte liegt der Wert der *Discernibility Metric* für die Segmentaggregation über dem Wert des Donutmasking (siehe Abbildung 4.17). Der Informationsverlust bei der Segmentaggregation ist folglich höher, als beim Donutmasking. Für niedrige k-Werte wie 10 oder 20 liegen die Ergebnisse allerdings nah beieinander, so beträgt für k = 10 das Metrikergebnis des Donutmasking $2,58e^9$ und der Aggregation auf Segmente $2,64e^9$. Mit steigendem k-Wert bewegen sich die Ergebnisse auseinander und der Metrikwert für die Segmentaggregation steigt stärker an, als der des Donutmasking.

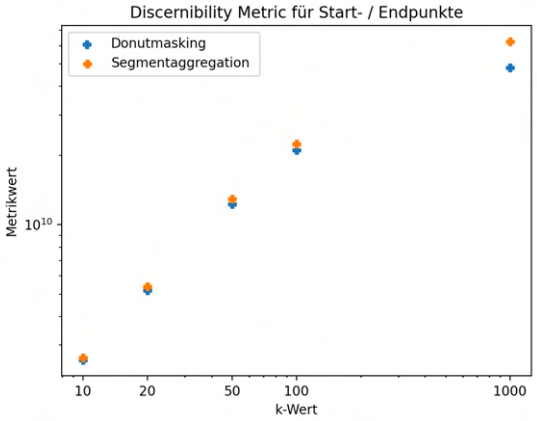
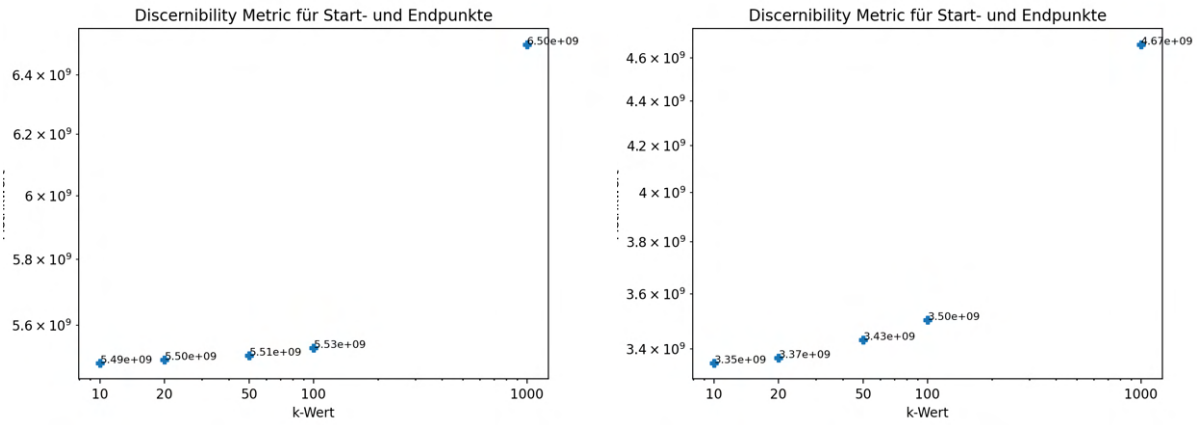


Abbildung 4.17.: *Discernibility Metric* für Segmentaggregation und Donutmasking im Vergleich

Die Kurvenverläufe der beiden Methoden für Aggregation auf Zentroide zeigen einen Anstieg des Metrikwertes mit steigendem k-Wert (siehe Abbildung 4.18). Für die k-Werte 10 bis 100 steigen die Metrikergebnisse beider Methoden nicht stark an, für Postleitzahlgebiete von $5,49e^9$ auf $5,53e^9$. Sowohl für Postleitzahlgebiete, als auch Stadtteile ist zwischen dem k-Wert von 100 und dem k-Wert von 1000 ein großer Wertunterschied sichtbar. Hier steigt der Metrikwert für Postleitzahlgebiete von $5,54e^9$ auf $6,5e^9$.



(a) Postleitzahlgebiete

(b) Stadtteile

Abbildung 4.18.: *Discernibility Metric* für Aggregation auf Zentroide

Als Erklärung kann hier das Zustandekommen des Metrikwertes herangezogen werden. Die Metrik bestraft Einträge, die in einer großen Äquivalenzklasse sind, stärker als Einträge, die kleinen Äquivalenzklassen zugeordnet werden. Zwischen k = 100 und k = 1000 ist der Abstand zwischen

den gewählten k-Werten größer, als zwischen den zuvor gewählten Werten von 10, 20 und 50. Der k-Wert gibt für die Privatisierungsmethode die Anzahl der Einträge vor, die eine Äquivalenzklasse erreichen muss, um k-anonym zu sein. Das sind für k = 1.000 mindestens 1.000 Einträge, deutlich mehr als für k = 100 mit 100 Einträgen. Der starke Sprung zwischen den Metrikwerten für die k-Werte 100 und 1000 kann also auf den starken Sprung der k-Werte zurückgeführt werden. Im Gesamten betrachtet sind die Metrikwerte für die Aggregation auf Postleitzahlen höher als für die Aggregation auf Stadtteile. Die Aggregation auf Postleitzahlen erzielt folglich ein schlechteres Ergebnis und somit eine schlechtere Datenqualität. Dieses Ergebnis lässt sich mit der festen Zahl an Äquivalenzklassen für Stadteile als auch Postleitzahlgebiete erklären. Da es weniger Postleitzahlgebiete als Stadtteile gibt, müssen die Einträge des Datensatzes auf weniger Klassen aufgeteilt werden. Die Äquivalenzklassen haben folglich mehr Einträge und bekommen höhere Strafwerte zugewiesen.

4.3.4. Qualitative Evaluation von Privatsphäre und Nützlichkeit anhand von Visualisierungen

In diesem Kapitel werden die für die unterschiedlichen k-Werte und Privatisierungsmechanismen entstandenen Visualisierungen in Form von Heat Maps und Flow Maps zunächst bezüglich des Informationsgehaltes und des Privatsphäre-Levels betrachtet und evaluiert. Anschließend werden die Visualisierungen genutzt, um das Leipziger Fahrradklima zu analysieren (siehe Kapitel 5). Die bei der Aggregation auf Zentroide entstandenen Daten werden hierbei als Flow Map visualisiert. Die Daten von Segmentaggregation und Donutmasking werden als Heat Map visualisiert. Vorgestellt werden in diesem Kapitel nur die interessantesten Visualisierungen, die die in der Analyse gemachten Beobachtungen am stärksten widerspiegeln. Zusätzlich befinden sich alle erstellen Flow und Heat Maps für alle Privatisierungsmechanismen im Anhang der Arbeit (siehe Anhang A). Besonderes Augenmerk bei der Untersuchung der Visualisierungen hat der Informationsgehalt verglichen mit dem erzielten Privatsphäre-Level. Für eine Analyse des Fahrradklimas sind Informationen relevant wie Häufigkeit der Nutzung von bestimmten Strecken oder auch Bewegungsströme durch die Stadt. Ziel der Visualisierungen ist es, mithilfe des privatisierten Datensatzes diese Fragen so gut wie möglich beantworten zu können.

Um die Datenqualität der unterschiedlichen Privatsphäre-Level zu vergleichen, werden die anonymisierten Visualisierungen mit der Visualisierung des Originaldatensatzes verglichen und geprüft, ob und wie stark die Verteilungen und entstandenen Muster vom Original abweichen. Je größer die Abweichung vom Originaldatensatz, umso weniger Nützlichkeit haben die Daten. Außerdem sollen die vier in dieser Arbeit angewandten Privatisierungsmechanismen miteinander verglichen werden. Aufgrund der unterschiedlichen Daten, die die Privatisierungsmechanismen generieren, werden die Mechanismen in zwei Gruppen aufgeteilt und innerhalb dieser verglichen: Methoden, die die Daten stark vergröbern (Aggregation auf Zentroide für Postleitzahlgebiete und Stadtteile) und Methoden, die die Routeninformation beibehalten (Segmentaggregation und Donutmasking).

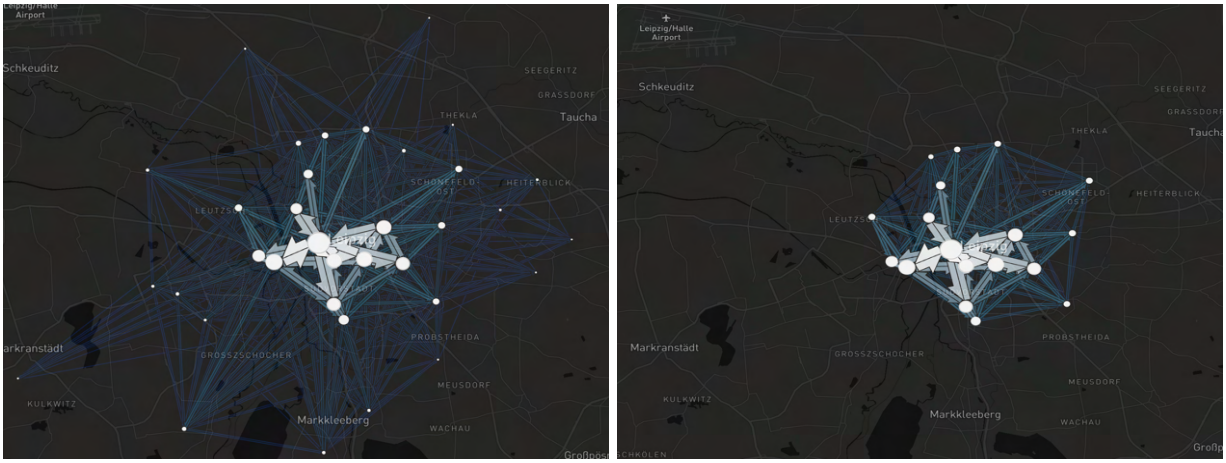
Die Visualisierungen werden zusammenfassend auf folgende Fragen hin betrachtet:

- Wo liegen die Unterschiede für verschiedene k-Werte innerhalb eines Privatisierungsmechanismus?
- Wo liegen die Unterschiede für gleiche k-Werte bei ähnlichen Privatisierungsmechanismen?
- Wie stark ändert sich die geographische Verteilung der Startpunkte für verschiedene k-Werte?

Vergleich unterschiedlicher k-Werte

Um Unterschiede für verschiedene k-Werte, also unterschiedliche Level an Privatsphäre, innerhalb eines Privatisierungsmechanismus zu finden, werden für jeden Privatisierungsmechanismus (Aggregation auf Zentroide, Aggregation auf Segmente und Donutmasking) die Visualisierungen für jeden k-Wert betrachtet.

Für alle vier Mechanismen steigt mit höher werdendem k-Wert die Konzentration der Daten auf den Stadtkern. Während für kleine k-Werte auch Strecken von und zu Vororten von Leipzig auf der Karte zu sehen sind, beschränkt sich mit zunehmendem k die Verteilung der Strecken auf die Stadtmitte (siehe Abbildung 4.19).



(a) k = 10 (b) k = 1000

Abbildung 4.19.: Flow Map: Aggregation auf Postleitzahlgebiete. Mit steigendem k-Wert zeigt sich eine Zentrierung der Daten auf Stadtkern.

Für die Methoden Segmentaggregation und Donutmasking, die die Routeninformation beibehalten, ähnelt bis zu einem k-Wert von 100 die Verteilung der Daten im Groben weiterhin dem Originaldatensatz. Ab einem k-Wert von 1000 lässt die Visualisierung allerdings keine Interpretation der Daten mehr zu, da nur Routen im Stadtkernbereich angezeigt werden. Hier ist der größte Unterschied zum Originaldatensatz zu erkennen (siehe Abbildung 4.20). Diese Beobachtung verhält sich passend zu den Ergebnissen der Nützlichkeits-Metriken aus Kapitel 4.3.3: für steigende k-Werte sinkt der Nutzen der Daten.

Bezüglich des Informationsgehaltes ist ein spannender Unterschied zwischen den Visualisierungsformen Heat Map und Flow Map zu erkennen. Für k = 1000 können häufig genutzte Strecken in der Heat Map nicht mehr abgelesen werden, sie zeigt fast nur weiß eingefärbte, also wenig genutzte

(a) k = 10 (b) k = 1000

Abbildung 4.20.: Heat Map: Donutmasking. Verringerung der Datenqualität für k = 1000. Rote Einfärbung repräsentiert intensive Nutzung der Strecke, weiße Einfärbung wenig Nutzung der Strecke.

Strecken. Aus der Flow Map für k = 1000 hingegen kann weiterhin eine Information über starke und weniger starke Bewegungsströme entnommen werden (siehe (b) in Abbildungen 4.19 und 4.20). Spannend ist diesbezüglich der Vergleich mit den Nützlichkeits-Metriken. Diese ergibt für die Methode der Aggregation auf Zentroide (Flow Map) einen insgesamt höheren Informationsverlust als für die Methode des Donutmasking (Heat Map). Wird dieser Aspekt in der Visualisierung verglichen, so erscheint der Informationsverlust beim Betrachten der Heat Map gravierender als beim Betrachten der Flow Map.

Diese Beobachtung kann damit erklärt werden, dass im Vergleich zur Heat Map die von der Flow Map gezeigten Daten stärker vergröbert werden, da hier außer den Start- und Endpunkten jegliche Information gelöscht wurde. Die Flow Map zeigt allerdings nur Strömungen zwischen Start- und Endpunkten, daher ist nicht von Belang, dass die Routeninformation nicht verfügbar ist. Bei der Privatisierung der Daten für die Heat Maps wird die detaillierte Routeninformation für hohe k-Werte so weit vergröbert, dass ganze Routen aus dem Datensatz verschwinden. So tritt bei Heat Maps der Unterschied zwischen Visualisierungen für niedrige und hohe k-Werte stärker in Erscheinung als für Flow Maps.

Vergleich zwischen Segmentaggregation und Donutmasking

Mit dem Vergleich zwischen den Methoden Segmentaggregation und Donutmasking, die beide die genaue Routeninformation enthalten, kann untersucht werden, an welchen Stellen sich die Privatisierungsmechanismen unterscheiden. Bei dieser Auswertung soll betrachtet werden, wie weit die Routenstartpunkte vom Originalstartpunkt verschoben wurden. Dafür werden die Startpunkte der Routen für jede Methode in einer Karte als Cluster, also gruppiert, visualisiert (siehe Abbildungen 4.21 und 4.22). Je mehr Punkte zu einem Cluster hinzugefügt werden, desto stärker wurden die Routendaten verändert, um k-anonymity zu erreichen. Dabei dient die geographische Verteilung der Punkte der Originaldaten als Vergleich. Je größer die Abweichung von diesen, desto stärker wurden die Daten verändert.

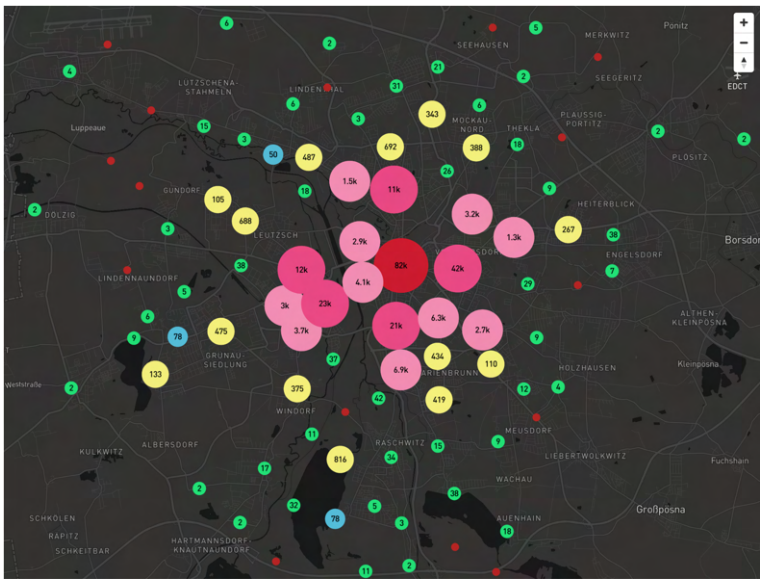
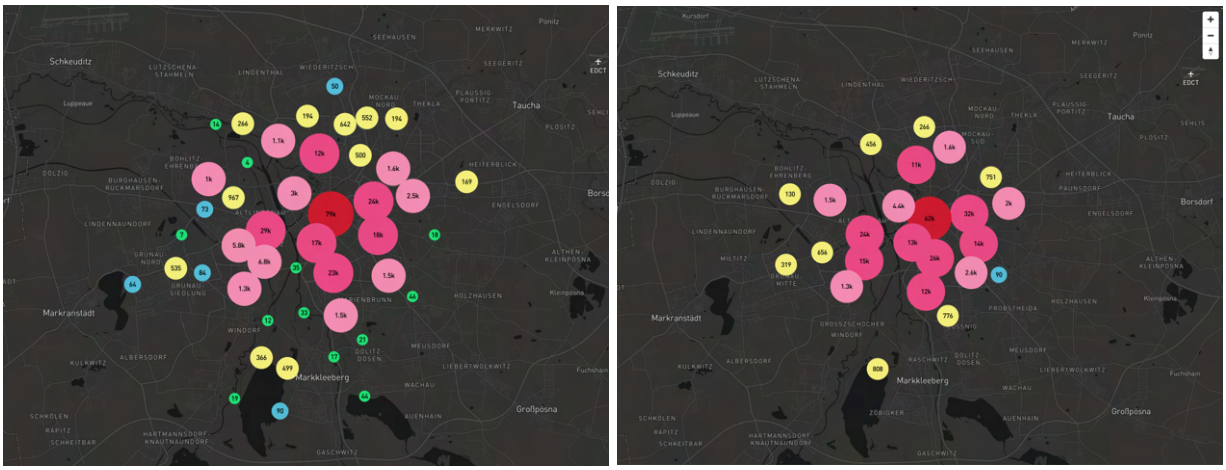


Abbildung 4.21.: Startpunkte der Originaldaten. In rosa gefärbten Kreise wurden zwischen 1.000 und 20.000 Startpunkte gruppiert. Der rote Punkt repräsentiert ein Cluster mit circa 80.000 Startpunkten. Gelbe kreise repräsentieren Cluster der Größe 100 bis maximal 1.000, blaue Kreise Cluster mit 50 bis 100 Punkten und grüne Kreise Cluster mit unter 50 Punkten.



(a) Startpunkte für Aggregation auf Segmente (b) Startpunkte für Donutmasking

Abbildung 4.22.: Verschiebung der Routenstartpunkte nach Privatisierung für k = 50. Zu erkennen ist die Minimierung der kleinen Cluster-Größen.

Die Abbildung 4.21 zeigt die geographische Verteilung der Routenstartpunkte des Originaldatensatzes. Die großen rosa gefärbten Cluster zeigen, dass pro Clusterpunkt 1.000 bis 20.000 Routen im Zentrum der Stadt starten. Gleichzeitig gibt es viele kleine, gelb oder grün gefärbte Cluster um die Stadt verteilt. Auch an diesen Punkten starten Routen, an je einem dieser Punkte maximal 1.000 Routen, im Vergleich zur Stadtmitte also deutlich weniger. Nach der Privatisierung ist für beide Methoden ein Unterschied zum Originaldatensatz zu erkennen (siehe Abbildung 4.22). Nach Anwendung des Donutmasking ist eine stärkere Vergröberung zu erkennen, als für die Segmentaggregation. In Abbildung 4.22 (b) sind wenige gelbe Cluster und keine grünen Cluster zu sehen. Das Donutmasking verschiebt Start- und Endpunkte einer Route bis sie auf k-anonymen Segmenten lie-

gen. Für den hier betrachteten Fall von k = 50 werden daher große Cluster gebildet und kaum kleine. Die Startpunkte außerhalb des Zentrums sind hier bei der Privatisierung verschoben oder gelöscht worden. Die geographische Verteilung der Startpunkte für die Segmentaggregation entspricht mehr der Originalverteilung. Es sind alle Cluster-Größen in Abbildung (a) zu sehen. Routen, die weit außerhalb der Stadt starten, sind auch hier nicht mehr in der Abbildung zu finden.

Beide Privatisierungsmethoden haben den Originaldatensatz merklich verändert. Das Donutmasking, welches explizit Start- und Endpunkte auf Anonymität überprüft, zeigt einen stärker veränderten Datensatz als die Segmentaggregation. Da in Abbildung (b) wenig bis keine kleinen Cluster zu sehen sind, kann auf eine hohe Privatsphäre der Startpunkte geschlossen werden. Im Vergleich zum Originaldatensatz sind außerhalb des Zentrums startenden Routen gelöscht oder verschoben worden, was mit entsprechend hohem Informationsverlust einhergeht. Auch die Segmentaggregation hat Routen außerhalb des Stadtkerns gelöscht, dennoch ist aus dem Vergleich zwischen Abbildung 4.22 (a) und (b) abzuleiten, dass bei der Segmentaggregation mehr Information erhalten wurde, da auch kleinere Cluster-Größen zu sehen sind. Die kleinen Cluster wiederum zeigen, dass die Startpunkte bei dieser Privatisierungsmethode ungeschützt bleiben und einzelne Routenstartpunkte auf Personen zurückgeführt werden können.

Werden die Heat Maps der beiden Methoden miteinander verglichen, sind die Unterschiede nur im Detail erkennbar. Für den Kernstadtbereich, in dem viele Routen verlaufen, lässt sich in den Visualisierungen als Heat Map kaum ein Unterschied zwischen beiden Mechanismen erkennen. Die Differenzen sind an Orten zu erkennen, an denen wenige Routen entlang führen. In Abbildung 4.23 sind die Unterschiede der Mechanismen erkennbar. In blau eingekreisten Teilen der linken Karte wurden bei der Segmentaggregation größere Teile gelöscht, die beim Donutmasking im rechten Teil der Karte zu sehen sind. Dabei handelt es sich um Mittelteile von Routen, die von sehr wenigen Personen gefahren wurden und daher aus Mangel an Privatsphäre bei der Segmentaggregation gelöscht wurden. Der gleiche Bereich wurde beim Donutmasking nicht gelöscht, da es sich nicht um Start- oder Endpunkte der Route handelt und die Prüfung nicht bis zu diesen Stellen vorgedrungen ist. Durch die Löschung von Mittelstrecken einer Route kann die Privatsphäre von Trajektorien gewährleistet werden, da sich Bewegungsmuster einzelner Personen nicht im Datensatz wiederfinden, wenn sie nicht von genug anderen Personen geteilt werden.

Im Gegensatz dazu stellen die in Abbildung 4.23 rot markierten Kreise durch das Donutmasking verschobene Start- oder Endpunkte einer Route dar, die in der linken Karte weiterhin eingezeichnet sind, während sie in der rechten Karte verschoben wurden und nicht sichtbar sind. Hier wird deutlich, dass das Donutmasking Routenstart- oder Endpunkte, die von nicht genügend anderen Personen geteilt werden, soweit verschoben hat, dass sie mit Routenstart- oder Endpunkten anderer Personen zusammenfallen.

Abbildung 4.23.: Heat Maps für k = 20 im Vergleich. Links: Segmentaggregation, rechts: Donutmasking. Blau eingekreist sind bei der Segmentaggregation (links) gelöschte Trajektorien, rot eingekreist sind beim Donutmasking (rechts) verschobene Start- oder Endpunkte.

4.3.5. Auswertung des Privacy Utility Trade-off

Um abschließend eine Aussage darüber machen zu können, für welchen LPPM und welchen k-Wert der beste Kompromiss aus Nützlichkeit der Daten und Schutz der Privatsphäre gefunden werden kann, müssen die Ergebnisse der Privatsphäre-Metriken mit denen der Nützlichkeits-Metriken verglichen werden.

Werden die Ergebnisse der Nützlichkeits-Metrik für das Donutmasking und für die Segmentaggregation betrachtet, nimmt die Nützlichkeit der Daten mit steigendem k-Wert exponentiell ab (siehe Abbildung 4.16). Die erreichte Privatsphäre für Trajektorien steigt für die Segmentaggregation exponentiell mit steigendem k-Wert an (siehe Abbildung 4.13), wie auch die Privatsphäre der Start- und Endpunkte für das Donutmasking (siehe Abbildung 4.12). Je höher der ausgewählte k-Wert, desto höher ist zwar der Datenverlust, aber auch die gewonnene Privatsphäre. Die Anonymitäts-Sets sind für die Start- und Endpunkte beim Donutmasking als auch für die Trajektorien bei der Segmentaggragation größer, als die Mindestgröße von k es vorschreibt. Die Privatsphäre ist somit

für jeden k-Wert höher, als das mindestens zu erreichende Level, insgesamt erreicht die Privatsphäre also zufriedenstellende Werte. Da mit dem Anstieg von k die Nützlichkeit der Daten sinkt, die erreichte Privatsphäre aber immer höher als die mindestens zu erreichende Privatsphäre ist, erscheint bereits ein k-Wert von 20 ein guter Kompromiss zwischen Nützlichkeit und Privatsphäre. Um nochmals zu validieren, dass die Daten der Personen im Datensatz geschützt sind, muss ein Angriff auf den privatisierten Datensatz durchgeführt werden. Aus Kapazitätsgründen kann ein solcher Angriff in dieser Arbeit nicht durchgeführt werden.

Für die Methoden der Zentroidaggregation können in der Nützlichkeit der Daten für Postleitzahlgebiete als auch Stadtteile der gleiche Trend beobachtet werden: ein steigender Informationsverlust mit steigendem k-Wert (siehe Abbildung 4.18). Werden die Ergebnisse der Privatsphäre-Metrik *Anonymity Set Size* hinzugezogen, sind für Postleitzahlen die Anonymitäts-Setgrößen für die k-Werte 10 bis 100 ungefähr gleich groß, ein starker Anstieg der Anonymitäts-Setgröße ist nur für k = 1000 zu vermerken (siehe Abbildung 4.12). Für die k-Werte 10 und 20 beinhalten die Anonymitäts-Sets bei Stadtteilen weniger Einträge als für die k-Werte 50 und 100. Auch hier hebt sich, wie schon bei den Postleitzahlgebieten, der k-Wert von 1000 mit den größten Anonymitäts-Sets von den anderen k-Werten ab.

Die Re-Identifikationswahrscheinlichkeit für Postleitzahlgebiete liegt für k = 10 bei durchschnittlich 4.775 Einträgen in einem Anonymitäts-Set bei nur 0,02 % und für Stadtteile mit mit durchschnittlich 2.436 Einträgen bei 0,04 %. Die Re-Identifikationswahrscheinlichkeit einer Person liegt folglich deutlich unter der maximal zu erreichenden Wahrscheinlichkeit von 10 %. Dies spricht für eine hohe Privatsphäre der Zentroidaggregation für k = 10. Gleichzeitig zeigt der erreichte Wert der Re-Identifikationswahrscheinlichkeit, der weit unter dem maximal zu erreichenden Wert liegt, dass die Daten womöglich stärker vergröbert wurden, als notwendig gewesen wäre. Da bereits mit k = 10 ein hohes Privatsphäre-Level erreicht wird, kann für die Zentroidaggregation dieser Wert gewählt werden.

Beim Finden des Optimums für diesen Kompromiss ist immer das Szenario zu beachten, für das die Daten am Ende verwendet werden sollen. Werden die Daten als Datensatz veröffentlicht, sollte bei Unsicherheit zwischen zwei k-Werten der Höhere gewählt werden. Handelt es sich nicht um eine explizite Veröffentlichung des Datensatzes, sondern eine implizite Veröffentlichung, beispielsweise als Visualisierung, kann auch die Grafik selber Aufschluss darüber geben, mit welchem k-Wert sie veröffentlicht werden sollte (siehe dazu Kapitel 4.3.4).

4.3.6. Resümee bezüglich der verwendeten Methoden

Als letzter Teil der Evaluation sollen erwartete Ergebnisse mit den tatsächlich eingetretenen gegenübergestellt werden und ein kurzes Resümee über die verwendeten Methoden gegeben werden.

Betrachtung der verwendeten Metriken

Die Abbildung von genauen Punktkoordinaten auf Stadtteile oder Postleitzahlgebiete stellt eine starke Vergröberung der Information dar. Diese erschwert es Angreifenden, Nutzen aus den Information zu ziehen, folglich steigt die Privatsphäre durch die Vergröberung. Die Erwartung,

dass die Aggregation auf Zentroide somit schlechte Werte für die Nützlichkeit der Daten erzielt, kann bestätigt werden. Im Vergleich mit Donutmasking oder Segmentaggregation sind die Werte der *Discernibility Metric* für die Aggregation auf Zentroide höher und zeugen so von schlechterer Datenqualität: Für einen k-Wert von 10 ist das Ergebnis der *Discernibility Metric* für Postleitzahlgebiete $5,49e^9$, für Donutmasking $2,58e^9$ (siehe Abbildungen 4.18 und 4.16). Wie erwartet erzielt eine stärkere Vergröberung der Daten aber auch ein höheres Maß an Privatsphäre. So sind die Anonymitäts-Sets für die beiden Zentroid-Methoden deutlich größer, als für die Segmentaggregation oder das Donutmasking: Bei einem k-Wert von 20 liegt die Größe der Anonymitäts-Sets im Interquartilbereich für die Aggregation auf Stadtteile zwischen 150 und 6.000 Einträgen, für das Donutmasking sind es zwischen 40 und 150 Einträge (siehe Abbildung 4.12). Bei den Methoden Segmentaggregation und Donutmasking bleibt im Gegensatz zu den Zentroid-Methoden die Routeninformation erhalten. Hier war zu erwarten, dass sich für die Methodiken weniger Privatsphäre aber höhere Nützlichkeit ergibt. Dies spiegelt sich in den Metrikergebnissen auch wieder.

Beim Donutmasking werden die Routenstart- und Endpunkte entlang der gefahrenen Route verschoben. Bei der Segmentaggregation hingegen werden alle Segmente aus der Route gelöscht, die nicht von mindestens k Personen passiert worden sind. Für das Donutmasking wurde gegenüber der Segmentaggregation eine bessere Datenqualität erwartet, da die Route präziser bearbeitet und nicht gelöscht, sondern verschoben wird. Die Werte der *Discernibility Metric* für diese beiden Methoden sind allerdings sehr ähnlich und unterscheiden sich nur stärker bei einem k-Wert von 1000: Für k = 50 ist das Metrikergebnis für die Segmentaggregation $1,30e^{10}$, für das Donutmasking $1,23e^{10}$, das Donutmasking erreicht also einen leicht besseren Wert. Segmente, auf die der Start- oder Endpunkt wegen fehlender Anonymität nicht verschoben werden konnte, werden beim Donutmasking, ähnlich wie bei der Segmentaggregation, aus der Route gelöscht. Für höher werdende k-Werte werden so auch beim Donutmasking viele Segmente aus der Route gelöscht, worauf die nah beieinander liegenden Ergebnisse mit großer Wahrscheinlichkeit zurückzuführen ist.

Die angewandte Nützlichkeits-Metrik *Average Equivalence Class Size* erscheint abschließend betrachtet für Koordinatendaten beziehungsweise Datensätze mit Geoinformation nicht besonders gut geeignet. Diese Metrik bewertet die Verteilung der Daten auf Äquivalenzklassen und erzielt ein optimales Ergebnis, wenn jede Äquivalenzklasse k Einträge hat. Die Verteilung der Daten auf Äquivalenzklassen konnte bei der vorliegenden Untersuchung allerdings nur in einem gewissen Maße selber bestimmt werden. Als Zentroide wurden Stadtteile und Postleitzahlgebiete gewählt. Diese Zentroide stellen die Äquivalenzklassen dar, auf die die Daten aufgeteilt wurden. Hierbei konnten nicht mehr Äquivalenzklassen generiert werden, als es Postleitzahlgebiete oder Stadtteile in Leipzig gibt. Für die Segmentaggregation und das Donutmasking wurden die Straßensegmente der Routen als Äquivalenzklassen gewählt. Maßgeblich hierbei ist die Routeninformation, die im Originaldatensatz enthalten ist. Die Anordnung und Häufigkeiten der Routen und somit die Aufteilung auf Äquivalenzklassen ist vom Datensatz vorgegeben. Auch hier kann nur wenig Einfluss auf die Verteilung der Daten auf die Äquivalenzklassen genommen werden.

Die Prüfung des Privatsphäre-Levels anhand der *Anonymity Set Size* erscheint passend, da Anonymitäts-Sets sowohl für kleine Bereiche, wie Segmente, als auch große Bereiche, wie Stadtteile, gebildet werden können. Zugleich sind sie sehr aussagekräftig bezüglich der Re-Identifizierungswahrscheinlichkeit. Da k-anonymity allein vorgeworfen wird, nicht vor Enthüllung sensibler

Attribute zu schützen, wurde durch die *(α-k) - anonymity* zusätzlich die Verteilung der Endpunkte innerhalb eines Anonymitäts-Sets von Startpunkten überprüft. So kann mit hoher Wahrscheinlichkeit ausgeschlossen werden, dass sich innerhalb eines Anonymitäts-Sets dennoch Lücken in der Privatsphäre befinden.

Betrachtung der verwendeten LPPM

Bezüglich der angewandten *Location Privacy Preserving Mechanisms* ist die Aggregation auf Zentroide die am einfachsten umzusetzende und bezüglich der vorab benötigten Information effektivste Methode. Bei dieser Methode werden nur zwei Punkte der Route, der Start- und Endpunkt, auf Anonymität geprüft. Diese Punkte können aus dem Datensatz entnommen werden. Die Zentroidaggregation erreicht ein sehr hohes Maß an Privatsphäre, indem POIs wie Wohn- oder Arbeitsadresse nicht mehr direkt aus dem Datensatz entnehmbar sind. Allerdings geht sie auch mit den größten Einschnitten in die Qualität der Daten einher.

Methoden wie Segmentaggregation oder Donutmasking sind rechenintensiver als die Zentroidaggregation, da zu ihrer Durchführung für jeden Punkt der Route zunächst eine fehlende Information eingeholt werden muss. Zu jedem Punkt wird das zugehörige OSM-Segment benötigt, auf dem sich der Routenpunkt befindet. An diesen Vorverarbeitungsschritt anknüpfend muss jedes Segment der Route auf Anonymität geprüft werden. Die große Menge an zu überprüfenden Punkten erfordert mehr Rechenleistung als bei der Zentroidaggregation. Aus dem Donutmasking geht im Vergleich zur Aggregation auf Zentroide ein deutlich informativerer Datensatz hervor. Zugleich kann die Privatsphäre von POIs erhaltenden werden. Für die Aggregation auf Segmente ist zu beachten, dass der Anonymitätsbegriff hier für Trajektorien und nicht für Start- oder Endsegmente gilt. Der Schutz von POIs wie Wohn- und Arbeitsadresse ist daher nicht in jedem Anwendungsfall gewährleistet. Durch Löschen der nicht anonymen Segmente aus der Route ist gegenüber dem Donutmasking ein besserer Schutz der Trajektorien gegeben.

An diesem Punkt stellt sich die abschließende Frage, welche Privatisierungsmethoden und welcher k-Wert für das Erstellen der Visualisierungen für die Fahrradklima-Analyse geeignet sind. Die Aggregation auf Zentroide stellt bereits eine so starke Vergröberung für die Daten und so die Un-unterscheidbarkeit der dahinterstehenden Personen dar, dass ein niedriger k-Wert von 10 für die Visualisierung als Flow Map verwendet werden kann. Für die Erstellung der Heat Maps gilt zu entscheiden, ob eine plötzlich endende Route in der Visualisierung (siehe Abbildung 4.23, rote Kreise) mehr Privatsphäre enthüllt, als eine Route, die nur von wenigen Personen genutzt wird, aber auf der Karte angezeigt wird (siehe Abbildung 4.23, blaue Kreise).

Genaue Verläufe von einzelnen Routen können in der Heat Map nicht verfolgt werden. Ist der Beginn einer Route in der Heat Map sichtbar, kann der Endpunkt der Route daraus nicht abgeleitet werden. Wird ein Datensatz als Heat Map veröffentlicht, können Angreifende somit eine Route anhand des Start- oder Endpunktes nicht einer Person zuordnen. Ein durch Donutmasking privatisierter Datensatz stellt sicher, dass keine einzelnen Start- oder Endpunkte im Datensatz enthalten sind, sondern mindestens k andere Routen diese Start- oder Endpunkte teilen. Ein in der Heat Map zu erkennender Start- oder Endpunkt kann folglich auf mindestens k Personen zurückgeführt werden. Zeigt eine Visualisierung ein Bewegungsmuster einer einzelnen Person, das nicht privatisiert

wurde, lässt dies deutlich mehr Rückschlüsse zu. Eine Trajektorie, die nur von einzelnen Personen gefahren wird, ist in der Heat Map schnell zu erkennen. Sie wird sich auf wenig befahrenen Strecken befinden, die in der Heat Map farblich als solche gekennzeichnet ist. Kennen Angreifende das in der Visualisierung abgebildete Bewegungsmuster, können sie es leicht einer Person zuordnen. Für die Entscheidung, welche Form der Privatisierung für die folgende Fahrradklima-Analyse gewählt wird, ist außerdem bedeutsam, aus welcher Visualisierung mehr Information gewonnen werden kann. Im Vergleich der LPPM haben die von der Segmentaggregation privatisierten Daten mehr dem Originaldatensatz entsprochen und somit weniger Information verloren, als die durch das Donutmasking privatisierten Daten (siehe Abbildung 4.22). Daher wird die Methode der Segmentaggregation zur Erstellung der Heat Maps gewählt.

Diese oben gegebene Einschätzung zur Wahl eines Privatisierungsmechanismus beruht auf der Veröffentlichung des Datensatzes als Visualisierung. Wird der Datensatz an sich veröffentlicht, gilt es die Risiken gesondert zu betrachten. Liegt Angreifenden der Datensatz vor, können sie für jeden beliebigen Start- oder Endpunkt den zugehörigen End- oder Startpunkt einer Route ermitteln. Hier gilt es, den Datensatz durch Donutmasking mit einem selbst zu wählenden k-Wert zu privatisieren, um zu vermeiden, dass diese Ermittlung erfolgreich ist (siehe beschriebene Szenarien in Kapitel 4.3.1). Besitzen Angreifende die Möglichkeit, einen Datensatz als Karte zu visualisieren, können Bewegungsmuster von Personen schnell aus der Visualisierung erfasst werden. Für diesen Schritt benötigen Angreifende zwar erweiterte Fähigkeiten, dennoch sollte diesem Szenario vorgebeugt werden, indem der Datensatz zusätzlich mittels Segmentaggregation privatisiert wird. Für die Veröffentlichung des Datensatzes sollten sowohl Start- und Endpunkte als auch Trajektorien geschützt werden und der Datensatz so mittels Donutmasking und Segmentaggregation privatisiert werden.

5. Analyse des Leipziger Fahrradklimas

Verkehr ist vor allem in Großstädten ein immer wichtigeres Thema: meist zu Hauptverkehrszeiten sind die Straßen mit PKWs überfüllt und nehmen anderen Verkehrsteilnehmenden zunehmend den Raum. Zudem sind Anwohnende von Hauptverkehrsstraßen der Lärmbelastung durch die hohe Befahrung der Straße ausgesetzt [6]. Gleichzeitig ließen sich viele Menschen bei einer guten Fahrradinfrastruktur zum Umstieg auf das Rad überzeugen [7]. Ein solcher Umstieg kann dazu beitragen, das Verkehrsaufskommen durch PKWs zu verringern. Zu einer guten Fahrradinfrastruktur gehören neben dem Vorhandensein von Radwegen oder Radfahrstreifen auf Straßen, auch Eigenschaften wie die Beschaffenheit und Sicherheit der von Radfahrenden benutzten Radwege und Straßen [8].

In diesem Kapitel soll anhand der Kombination aus den bereits gesammelten und privatisierten Daten und öffentlichen Ressourcen das Fahrradklima der Stadt Leipzig visuell dargestellt und untersucht werden. Dafür wird der Nextbike-Datensatz zusätzlich nach Uhrzeiten ausgewertet. Um dabei weiterhin den Schutz der Privatsphäre gewährleisten zu können, wird aus dem Originaldatensatz der gewünschte Zeitraum ausgesucht, beispielsweise Werktage zwischen 8 und 10 Uhr, und anschließend der gewünschte Privatisierungsmechanismus auf diesen Datenausschnitt angewandt. Im daraus resultierenden privatisierten Datensatz werden weiterhin nur die Attribute *rental_id*, Route als *geojson* und Route als *segments_list* gespeichert.

Um eine Fahrradklima-Analyse für Leipzig durchführen zu können, werden zunächst die dafür zu beantwortenden Fragen formuliert:

1. Wie sind die Unterschiede in der Verteilung der Bewegungsströme? Gibt es Orte, an denen mehr Bewegung stattfindet?
2. Welche Routen werden am häufigsten mit Fahrrädern befahren und wie sieht die Radwegsituation an diesen Routen aus?
3. Wie viel Prozent der in dieser Arbeit am häufigsten gefahrenen Strecken befinden sich auf einem eigenen Radweg / Radverkehrsanlage?
4. Wie ist die Beschaffenheit von oft befahrenen Straßen?
5. Gibt es Unterschiede der Fahrradmengen zwischen Werktagen und Wochenende?
6. Gibt es Unterschiede der Fahrradmengen an Werktagen zu verschiedenen Uhrzeiten?

Für einige dieser Analysen müssen zunächst zusätzliche Daten beschaffen, visualisiert und ausgewertet werden. Dieser Prozess wird im Abschnitt der jeweiligen Analyse kurz erläutert.

5.1. Untersuchung der Bewegungsströme

Zunächst soll die Frage nach den Bewegungsströmen innerhalb der Stadt beantwortet werden. Dazu wird die Flow Map für einen k-Wert von 10 sowohl für Stadtteile als auch Postleitzahlgebiete betrachtet. Ein Pfeil stellt dabei einen Bewegungsstrom zwischen zwei Punkten dar. Die Dicke des Pfeils deutet auf die Anzahl der gefahrenen Routen in diesem Strom hin, dickere Pfeile repräsentieren hierbei eine höhere Anzahl gefahrener Routen.

Die Visualisierung der Postleitzahlgebiete verdeutlicht, dass die Hauptbewegungsströme zwischen den zentrumsnahen Gebieten, darunter die Postleitzahlen 04105, 04103, 04107, 04299, 04315, 04177, und dem Zentrum 04109 stattfinden. Ein solcher Bewegungsstrom, beispielsweise von 04103 nach 04109, beinhaltet zwischen 4.000 und 5.000 gefahrenen Routen. Insgesamt werden von den 233.213 im Datensatz enthaltenen Routen circa 86.000 von und zum Stadtzentrum (04109) gefahren, das sind 37 % der insgesamt gefahrenen Routen. Der restliche Teil der Routen verstreut sich auf das gesamte Stadtgebiet, die Anzahl pro Bewegungsstrom ist hierbei kleiner, meist maximal 100 gefahrene Routen zwischen zwei Punkten.

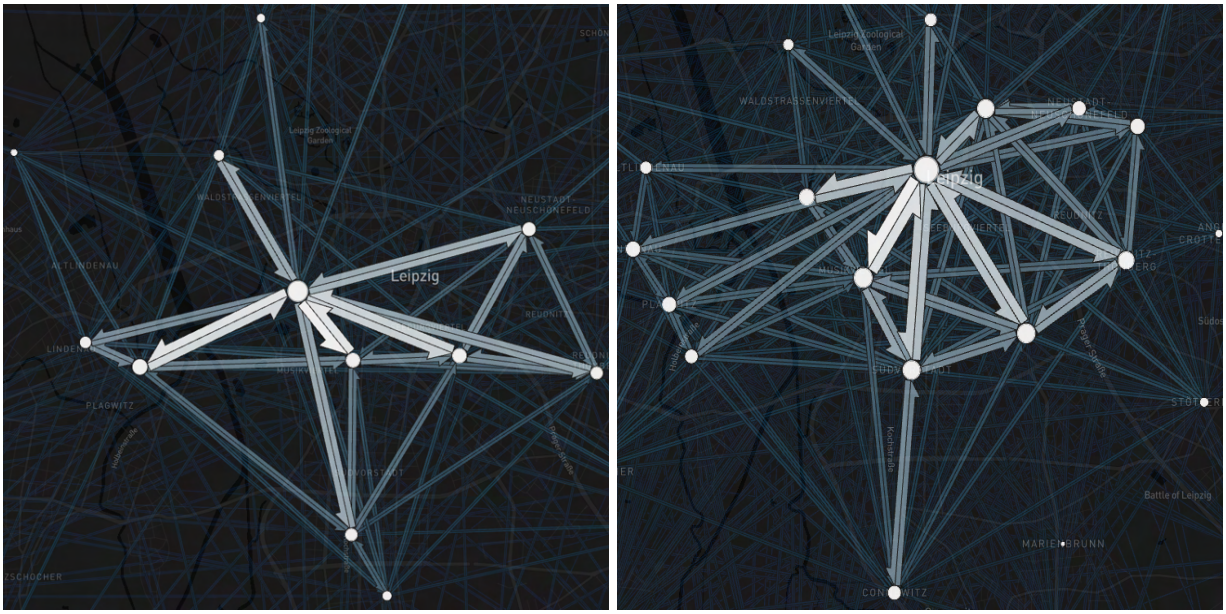


Abbildung 5.1.: Flow Maps mit k=10 zur Analyse der Verteilung der Bewegungsströme in Leipzig, links: Postleitzahlgebiete, rechts: Stadtteile

Die Betrachtung der Visualisierung der Stadtteile lässt eine detailliertere Analyse zu, da mehr Stadtteile als Postleitzahlgebiete existieren und somit die Verteilung der Routen kleinteiliger ist. Dennoch machen Routen von und zum Stadtkern mit 26 % noch immer einen Großteil der insgesamt gefahrenen Routen aus. Die Anzahl der Routen pro Bewegungsstrom ist aber insgesamt gleichmäßiger verteilt, sowohl zwischen Stadtzentrum und stadtezentrumsnahen Gebieten als auch zwischen zwei Stadtteilen außerhalb des Stadtzentrums. Während bei der Aggregation auf Postleitzahlen zwischen 4.000 und 5.000 Routen zwischen zwei Punkten im Zentrumsgebiet gefahren wurden, liegt diese Zahl für die Aggregation auf Stadtteile mit 2.500 Routen knapp bei der Hälfte. Zwischen zwei Stadtteilen, beispielsweise von Plagwitz nach Schleußig, sind es durchschnittlich circa 1.000 gefahrene Routen (siehe Abbildung 5.1).

Außerdem ist festzustellen, dass ein weiterer Teil der Routen zu Leipziger Seen führen, vor allem dem Cospudener See, Kulkwitzer See und Markkleeberger See. Diese Routen sind im vorliegenden Datensatz vermutlich unterrepräsentiert, da sich die Seen außerhalb der Abstellzone für die Ausleihfahrräder befinden und daher mit einem erhöhten Ausleihpreis gerechnet werden muss. Dennoch zeigt die Visualisierung so, dass die Wege mit dem Fahrrad neben der Fortbewegung in der Stadt auch für Freizeitaktivitäten und Ausflüge im nahen Umland gefahren werden.

5.2. Häufig gefahrene Strecken und deren Radwegsituation

Um die Information zu erlangen, auf welchen Straßen und Wegen sich Radverkehrsanlagen befinden, wird der OSM-Tag *cycleway* ausgewertet und visualisiert [46] (siehe Nominatim Dienst, Abbildung 4.5). Hierfür existiert bereits die nützliche Visualisierung „OSM-Fahrradkarte“ [47], die zusätzlich zur Radwegsituation wichtige Informationen beinhaltet wie das Tempo, das auf der Straße gefahren werden darf oder ob ein Weg mit motorisierten Fahrzeugen befahren werden darf. Zur Beantwortung der Frage nach häufigen Routen und deren Radwegsituation, werden die Heat Maps mit einem k-Wert von 20 ausgewertet und mit der OSM-Fahrradkarte verglichen (siehe Abbildung 5.2). So kann analysiert werden, ob oft befahrene Strecken auf Fahrradwegen beziehungsweise von der Autospur getrennten Wegen oder in verkehrsberuhigten 30er-Zonen entlang führen. Diese Analyse erfolgt exemplarisch und soll als Beispiel einer möglichen Analyse dienen, da die Routendaten des Nextbike-Datensatzes zwar nach bester Möglichkeit generiert sind, dennoch keine exakt gefahrenen Strecken wiedergeben.

Direkt ins Auge fällt bei der Betrachtung der Visualisierungen in Abbildung 5.2, dass sich die oft befahrenen, in der Heat Map dunkelrot gekennzeichneten Routen zu großen Teilen mit den Mustern der mit dunkelblauen Strichen eingezeichneten Fahrradanlagen in der OSM-Fahrradkarte decken. In der Fahrradkarte stellt eine durchgezogene dunkelblaue Linie einen von der Autospur getrennten Radweg und eine hellblau gestreifte Linie einen Radweg auf der Autospur dar. Die am häufigsten gefahrenen Routen verlaufen entlang der Karl-Liebknecht-Straße in Richtung Süden, entlang des Täubchenwegs in Richtung Osten, entlang der Rosa-Luxemburg-Straße über die Ludwigstraße in Richtung Nord-Osten und durch den Clara-Zetkin Park in Richtung Westen. Auch der Innenstadtring als Verbindungsstück weist ein hohes Fahrradaufkommen auf. Während die Routen gen Osten und Süden auf den Hauptverkehrsstraßen Täubchenweg und Karl-Liebknecht-Straße verlaufen, auf denen sich jeweils ein auf der Straße eingezeichneter Fahrradweg befindet, werden für die Wege in den Westen und Nord-Osten Straßen abseits der Hauptverkehrsstraßen bevorzugt: Für den Westen verläuft ein Großteil der Routen durch den Clara-Zetkin-Park, während die Routen in den Nord-Osten über die verkehrsberuhigte Ludwigstraße parallel zur Hauptverkehrsstraße Eisenbahnstraße verlaufen.

Dies kann als Indiz gedeutet werden, dass die Routen abseits der eigentlichen Hauptverkehrsstraßen Eisenbahnstraße und Jahnalle / Kätze-Kollwitz-Straße von den Radfahrenden gemieden werden, da sie beispielsweise als zu unsicher beziehungsweise zu wenig abgegrenzt von PKWs bewertet werden.

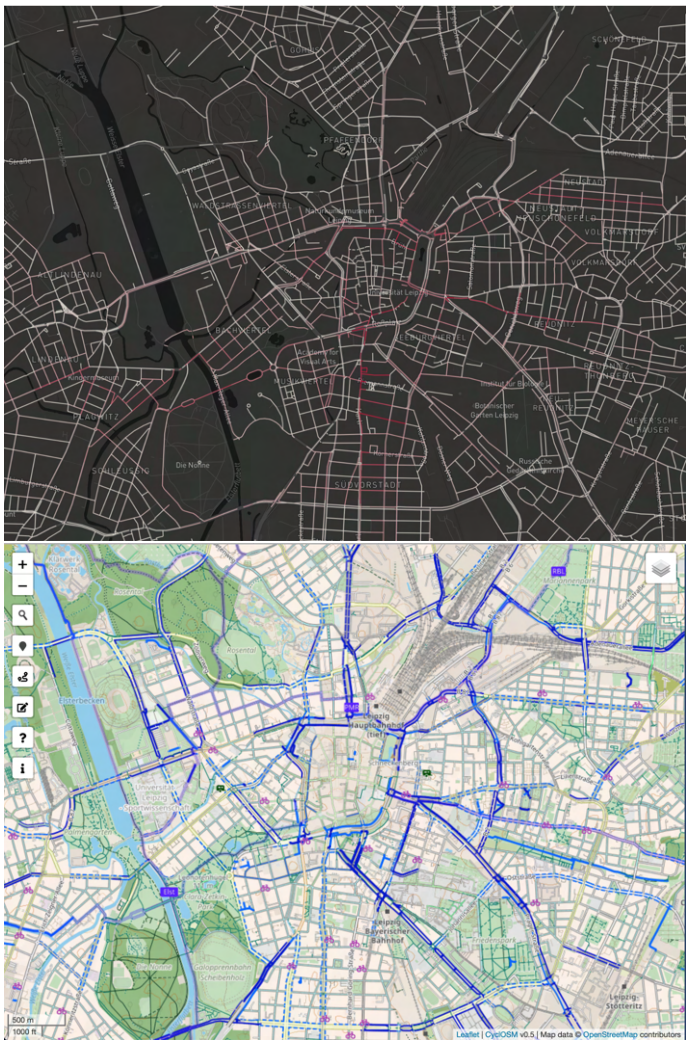


Abbildung 5.2.: Heat Map mit k = 20 (oben), Ausschnitt der OSM-Fahrradkarte (unten) zur Analyse der oft befahrenen Straßen in Leipzig

Ein Aspekt, der bei dieser Analyse miteinhergeht, ist die unter Punkt 3 gestellte Frage, wie viele der im Nextbike-Datensatz enthaltenen Routen über Radwege verlaufen. Diese kann durch einen Abgleich der Segmente in der Routendatenbank mit dem bereits erwähnten OSM-Tag *bicycle-way* beantwortet werden. Dazu wird abgefragt, ob sich auf dem Segment aus dem Nextbike-Datensatz ein Fahrradweg befindet. Von insgesamt 11.628 Straßensegmenten im Nextbike-Datensatz verlaufen 856 über einen bei OpenStreetMap vermerkten Fahrradweg, also circa 7,4 %. Dabei gilt zu beachten, dass sich auf vielen Straßen in 30er-Zonen kein Radweg befindet. Daher ist weiter interessant, wie viele Routen durch 30er-Zonen verlaufen. Indem die OSM-Datenbank für Straßen auch die Maximalgeschwindigkeit vermerkt, kann im gleichen Verfahren geprüft werden, in welcher Geschwindigkeits-Zone ein Segment liegt. Circa 30% der in dieser Arbeit vorkommenden Routen werden in Seitenstraßen mit einem Tempolimit von 30 km/h gefahren. Circa 38% werden auf Straßen mit Limit 50 km/h gefahren und für weitere circa 30% gibt es keine Angabe. Zusammen mit den auf Radwegen zurückgelegten Strecken werden insgesamt also 37,4 % der Wege auf verkehrsberuhigten Wegen oder Fahrradwegen zurückgelegt.

5.3. Beschaffenheit von oft befahrenen Straßen

Die vierte zu analysierende Frage beschäftigt sich mit der Beschaffenheit der Straße. Mit Beschaffenheit ist der allgemeine Zustand als auch die Art des Fahrbelages der Straße gemeint. Hinter dieser Analyse steht die Vermutung, dass Menschen, die auf von ihnen regelmäßig gefahrenen Wegen schlechte Straßenqualität vorfinden, es vermeiden das Fahrrad zu nutzen. Eine weitere Vermutung stellt eine erhöhte Unfallgefahr dar, wenn Schlaglöcher oder schlechte Fahrbahnteile umfahren werden müssen. Für diese Analyse wird ein Vergleich zu häufig gefahrenen Straßen mit dem Zustand dieser Straßen getätigt. Für die Bewertung des Zustandes einer Straße wird der OSM-Tag *smoothness* (dt. Gleichmäßigkeit / Glätte) [48] ausgewertet und eine Visualisierung mit Farben erzeugt. Dabei gibt es sieben unterschiedliche Stufen, in die der Zustand eingeteilt werden kann (in absteigender Reihenfolge): exzellent (hellgrün), gut, durchschnittlich, schlecht, sehr schlecht, schrecklich, unpassierbar (dunkelrot). Diese Visualisierung ist in Abbildung 5.3 zu sehen.

Abbildung 5.3.: Analyse der Straßenqualität von oft befahrenen Straßen in Leipzig. Heat Map mit k=20 (oben), Ausschnitt der Straßenzustandskarte (unten). Je grüner gezeichnet, desto besser die Straßenqualität.

Der Zustand kann einerseits der Zustand des Straßenbelages an sich sein, als auch die Art des Belages. Beispielsweise ist Asphalt glatter als Kopfsteinpflaster, daher würde Kopfsteinpflaster im gleich guten Zustand mit hoher Wahrscheinlichkeit eine schlechtere Bewertung erhalten.

Direkt zu erkennen ist die grün markierte und damit als sehr gut bewertete Hauptverkehrsstraße Karl-Liebknecht-Straße und die viel gefahrene Route durch den Clara-Zetkin-Park. Seitenstraßen sind vermehrt mit einem schlecht bewerteten, orange oder rot markierten Zustand gekennzeichnet. Für die Routen gen Osten und Nord-Osten liegen keine Bewertungen vor. Dies ist auch dahingehend für die Analyse bedauerlich, da hier eine Nebenstraße zur Hauptverkehrsstraße Eisenbahnstraße bevorzugt wird und die Frage entsteht, ob diese Bevorzugung mit der Qualität der Straße in Verbindung gebracht werden kann. Vor allem im Stadtbereich Südvorstadt und Zentrum-Süd, gibt es vermehrt Routen, die auf als schlecht markierten Seitenstraßen entlang führen.

5.4. Unterschiede der Verkehrsmengen zwischen Wochentagen

Um Unterschiede der Fahrradmengen an Werktagen und Wochenende zu analysieren, werden aus dem Originaldatensatz die Werktage getrennt von Samstagen und Sonntagen mit einem k-Wert von 20 privatisiert und anschließend als Heat Map visualisiert (s. Abbildung 5.4).

Abbildung 5.4.: Analyse von Unterschieden zwischen Werktagen und Wochenendtagen. Links: Werktage, Rechts: Samstage & Sonntage

Wird angenommen, dass am Wochenende weniger Fahrten in die Stadtmitte und mehr Fahrten zu Außengebieten oder Seen unternommen werden, beziehungsweise am Wochenende mit dem Fahrrad zur Arbeit gefahrenen Strecken wegfallen, hätte eine deutliche Unterscheidung der Einfärbung der Strecken erkennbar sein müssen. Allerdings ist auf Abbildung 5.4 keine solch deutliche Unterscheidung zu erkennen. Die am häufigsten gefahrenen Routen befinden sich weiterhin auf den Hauptverkehrsadern der Stadt.

Die Annahme, dass das Fahrradverkehrsaufkommen durch Wegfall der Berufstätigen am Wochenende niedriger als an Werktagen ist, kann auch nicht bestätigt werden. Von 233.213 Ausleihen im privatisierten Datensatz wurden unter der Woche 166.417 und am Wochenende 66.796 getätigt. Die

durchschnittliche Anzahl an gefahrenen Routen an einem Tag liegt für den Datensatz bei 33.317, sodass keine besondere Prägnanz weder für Wochenende noch Werktage zu erkennen ist.

Ein kleiner Unterschied zwischen den Karten lässt sich dahingehend erkennen, dass die Start- und Endpunkte der Routen für Werktage (linke Karte) weiter außerhalb des Stadtzentrums liegen, als für das Wochenende (rechte Karte). Es werden also insgesamt mehr außenliegende Stadtteile unter der Woche angefahren. Hier gilt es zu beantworten, ob diese Beobachtung in einem Zusammenhang zwischen Arbeitsverkehr an Werktagen und wegfallendem Arbeitsverkehr an Wochenenden steht. Diese Beobachtung kann damit interpretiert werden, dass zum einen das Fahrrad für den Arbeitsweg verwendet wird, als auch dass weite Strecken zur Arbeit mit dem Fahrrad gefahren werden.

5.5. Unterschiede der Verkehrsmengen zwischen Tageszeiten

Abbildung 5.5.: Analyse von Unterschieden zwischen Arbeitszeiten und nicht-Arbeitszeiten. Links: Werktage zwischen 08:00 und 10:00 Uhr, Rechts: Werktage zwischen 10:00 und 15:00 Uhr

Zunächst fällt auf, dass die Einfärbung der linken Abbildung deutlich roter ist, als die der rechten. Das lässt sich mit der Anzahl der Einträge und der verwendeten Generierung der Farben erklären: zwischen 8:00 und 10:00 Uhr sind 16.547 Routen im Datensatz enthalten, während zwischen 10:00 und 15:00 Uhr 39.119 Routen gefahren wurden. Die Anzahl der Farben ergibt sich aus der Gesamtzahl der zu visualisierenden Routen. Je weniger Daten enthalten sind, desto weniger Farben werden verwendet und desto geringer ist der Farbgradient zwischen den Abstufungen.

Die rotere Einfärbung der linken Heat Map enthält dennoch folgende Information: in der Zeit zwischen 8 und 10 Uhr sind pro Stunde circa 8.300 Routen gefahren worden, in der Zeit zwischen 10 und 15 Uhr pro Stunde circa 7.800. Die Anzahl der Ausleihen pro Stunde ist also für den Morgen und den kleineren Zeitraum höher, als nachmittags für den größeren Zeitraum. Beim Vergleich der gefahrenen Routen tritt zwar keine neue Route in den Vordergrund, dennoch ist die Einfärbung der Hauptverkehrsrouten für die Uhrzeiten von 8 bis 10 Uhr intensiver, als für die Uhrzeiten von 10 bis 15 Uhr. Auch führen die meisten dieser Routen in die Innenstadt, in der vermehrt Büroräume vorhanden sind. Daraus kann die Vermutung abgeleitet werden, dass viele Menschen das Fahrrad

morgens für den Weg zur Arbeit nutzen und dieser Arbeitsweg mit dem Fahrrad das Fahrradverkehrsaufkommen für diese Uhrzeiten erhöht.

5.6. Zusammenfassung der Ergebnisse

Aus der Summe der vorangegangenen Analysen lassen sich abschließend folgende Ergebnisse festhalten. Die Betrachtung der Flow Maps zeigt, dass der größte Teil der Bewegungsströme vom und zum Stadtzentrum führt. Nicht weit dahinter ist die Häufigkeit der Bewegungen zwischen aneinander angrenzenden Stadtteilen. Diese eher kurzen Bewegungsströme zwischen angrenzenden Gebieten passen auch mit der durchschnittlichen Ausleihdauer von fünf bist 15 Minuten zusammen (siehe Abbildung 4.6). Insgesamt gibt es dennoch in der gesamten Stadt von vielen Punkten aus Bewegungsströme, sodass Leipzig im Gesamten als eine mit dem Fahrrad gut zu befahrende Stadt erscheint. Leicht auffällig zeigt sich in Richtung Norden ein fehlender Strom großer Bewegung, sodass die Frage aufkommt, ob in nördlichen Stadtteilen mehr Menschen das Auto oder die öffentlichen Verkehrsmittel als das Fahrrad nutzen.

Als Hauptverkehrsrouten sind in Leipzig nicht nur die großen Verkehrsadern wie die Karl-Liebknecht-Straße oder der Täubchenweg zu finden. Für Routen in Richtung Westen und Nord-Osten erscheinen große Straßen eher gemieden und verkehrsberuhigte Seitenstraßen beziehungsweise Straßen durch den Park ohne motorisierten Verkehr favorisiert zu werden. Für Seiten- beziehungsweise Nebenstraßen ist häufig ein deutlich schlechterer Zustand oder schlechterer Belag (Kopfsteinpflaster) zu erkennen (siehe von Karl-Liebknecht abgehende Straßen oberhalb der Kurt-Eisner-Straße in Abbildung 5.3). Interessant ist hier, ob die Hauptverkehrsstraßen gefahren werden, weil die Seitenstraßen keine ausreichend gute Qualität aufweisen, um zügig voranzukommen. Weiterhin von Interesse ist hier, ob eine Verlagerung des Fahrradverkehrs auf Seitenstraßen zu einer Entlastung des gesamten Straßenverkehrs führen würde. Für weiterführende Analysen kann die Frage mitgenommen werden, ob eine bessere Straßenqualität von Nebenstraßen zu solch einer Verlagerung führen würde. Außerdem mitgenommen werden kann die Frage, ob die Entscheidung, eine verkehrsberuhigte Straße zu fahren, eine bewusste ist. Dieser Trend ist beispielsweise für die Ludwigstraße als verkehrsberuhigte Nebenstraße zur Eisenbahnstraße zu beobachten.

Der Vergleich zwischen Wochenende und Wochentagen konnte bezüglich häufig gefahrener Strecken keine gravierenden Unterschiede zum Gesamtdatensatz erkenntlich machen. Entgegen der gemachten Annahme, ist die Anzahl der am Wochenende gefahrenen Routen nicht geringer. Im Schnitt werden an jedem Tag der Woche ungefähr gleich viele Strecken mit dem Fahrrad zurückgelegt. Diese Erkenntnis zeigt bereits eine hohe Bereitschaft der Bewohnenden der Stadt Leipzig, das Fahrrad zu nutzen.

Bezüglich der Nutzung des Fahrrades für den Arbeitsweg konnten zwei Beobachtungen gemacht werden: Für die Unterscheidung zwischen Werktagen und Wochenende konnte an Werktagen ein größerer Radius für gefahrene Routen um das Stadtzentrum festgestellt werden. Dies kann als eine Nutzung des Fahrrades auch für weite Arbeitswege interpretiert werden. Die Unterscheidung der Uhrzeiten hat an Werktagen zwischen 8 und 10 Uhr ein deutlich höheres Fahrradaufkommen gezeigt, als zwischen 10 und 15 Uhr, was erstere Interpretation unterstützt.

Die Ergebnisse zeigen, dass der Fahrradverkehr sich auf Hauptbereiche beziehungsweise -straßen aufteilen lässt und es zu bestimmten Zeiten zu mehr Fahrradverkehr kommt. Um Anreize für den Umstieg auf das Fahrrad zu schaffen, können bei der Stadtplanung Ideen wie priorisierte Fahrradampeln zu Hauptzeiten des Arbeitsverkehrs überlegt werden, um das Fahrradfahren schneller und das Autofahren weniger reizvoll zu gestalten. Weiterhin ist es wichtig, an diesen Hauptverkehrsstraßen für genügend Sicherheit der Radfahrenden zu sorgen, indem solche Straßen mit einem eigenen Fahrradweg, besser mit einer getrennten Spur versehen sind, um Überholvorgänge von Autos mit dem vorgeschriebenen Mindestabstand zu ermöglichen. Vor allem in Richtung Westen erscheint der Nachholbedarf dahingehend am größten: Auch wenn der Park gerne als Straßenalternative genutzt wird, sollten auch Hauptverkehrsstraßen wie die Käthe-Kollwitz-Straße oder die Jahnallee sicheren und sich sicher anfühlenden Radverkehr möglich machen. Insgesamt betrachtet erscheint Leipzig bereits eine Stadt zu sein, in der viele Wege mit dem Rad zurückgelegt werden, was die durch die gesamte Stadt verteilten Routen zeigen.

Für die Gesamtbetrachtung dieser Analyse und Auswertung ist wichtig zu beachten, dass die Strecken aus Mangel eines Datensatzes mit Fahrradrouten aus den Start- und Endpunkten des Nextbike-Datensatzes generiert wurden und nicht das exakte Fahrradfahrverhalten abbilden. Die Interpretation der Ergebnisse ist daher vorsichtig zu betrachten. Dennoch kann diese Auswertung als Muster verwendet werden, ist ein Datensatz mit exakten Fahrradrouten verfügbar.

6. Diskussion und Ausblick

In dieser Arbeit wurden verschiedene *Location Privacy Preserving Mechanisms* (LPPM) betrachtet und drei Varianten implementiert. Diese Implementierung wurde für einen Datensatz durchgeführt, der Routendaten inklusive genauer Start- und Endpunkte beinhaltet. Die Routendaten sind um eine weitreichendere Analyse der LPPM durchführen zu können, mittels eines Routingdienstes generiert worden. Bezüglich der generierten Daten ist hier anzumerken, dass es nicht möglich war, Originaldaten von gefahrenen Fahrradstrecken zu erlangen. Mit Daten von tatsächlich gefahrenen Strecken kann die Effektivität der angewandten Methoden bezüglich der tatsächlich erreichten Privatsphäre genauer evaluiert werden. Auch der verwendete Datensatz von Nextbike wurde insofern bereits voranonymisiert, indem Nextbike die Kunden-IDs aus dem Datensatz entfernt hat. Mit einer Kunden-ID kann beispielsweise genauer untersucht werden, ob durch offengelegte Bewegungsmuster einzelne Personen identifiziert werden können oder wie stark Start- und Endpunkte der Routen einer einzelnen Person wirklich Rückschlüsse auf deren Wohnadresse, Arbeitsadresse oder sonstige sensible Information zulassen. Ein solcher Datensatz ist durch das Projekt „Stadtradeln Leipzig“ von der Forschungsgruppe movebis [49] gesammelt worden. Dieser wurde nur verzerrt veröffentlicht, sodass der Datensatz für diese Arbeit nicht verwendet werden konnte. Bei diesem Projekt konnten Fahrradfahrende in einem Zeitraum von 20 Tagen ihre Routen beim Fahren aufzeichnen, um einen Datensatz zu generieren, der anschließend Auswertungen der Standzeiten an Ampeln, aber auch häufig gefahrener Routen ermöglicht. Nach Abschluss des Projektes wurde der Datensatz für die Stadt Leipzig in Form einer Heat Map visualisiert (siehe Abbildung 6.1). Unklar ist hier, ob der Datensatz im Vorhinein privatisiert worden ist.

Es wurden drei verschiedene LPPM implementiert und zur Privatisierung des Nextbike-Datensatzes angewandt. Diese LPPM haben bei der anschließenden Evaluation der Aspekte Privatsphäre und Nützlichkeit gute Ergebnisse gezeigt. An dieser Stelle ist mitzunehmen, dass die implementierten Methoden, da nur auf den Nextbike-Datensatz angewandt, bezüglich der zu schützenden Attribute auf diesen zugeschnitten sind. Zukünftig ist interessant, andere Geo-Datensätze auf deren sensible Attribute oder Alleinstellungsmerkmale hin zu betrachten, die zusätzlich zu den Start- und Endpunkten der Routen privatisiert werden müssen, wie beispielsweise Kunden-IDs.

Ein weiterer wichtiger Aspekt, der in dieser Arbeit aus Kapazitätsgründen nicht betrachtet wurde, ist die Effizienz der implementierten Algorithmen. Wenngleich genaue Messungen und Auswertungen nicht gemacht worden sind, ist bereits der Prozess des *Reverse Geocodings* von GPS-Koordinaten auf Straßensegment sehr rechenintensiv und zeitaufwendig gewesen. Dieser Schritt erschwert es, die gesamte Pipeline von Privatisierung des Datensatzes bis zum Anzeigen der Visualisierung im Browser hintereinander ablaufen zu lassen. Indem das *Reverse Geocoding* für den Originaldatensatz vorgenommen wurde und alle folgenden Methoden die Daten aus diesem verwenden, musste der Schritt nur einmalig durchgeführt werden. Für eine Weiterentwicklung der Methoden kann die Frage mitgenommen werden, wie dieser Prozess zukünftig optimiert werden kann.

Anonymisierungsmethoden wie Aggregation oder Donutmasking werden mit dem Ziel angewandt, eine Un-unterscheidbarkeit von Personen zu erreichen und so eine *Linking Attack* verhindern zu können. Obwohl die hier implementierten Mechanismen zum Schutz der Privatsphäre in den Privatsphäre-Metriken gute Ergebnisse geliefert haben, ist weiterführend zu testen, wie gut ein Datensatz wirklich privatisiert ist, indem eine *Linking Attack* auf diesen durchgeführt wird. Bei einer solchen Attacke wird versucht, durch Verknüpfung von Information aus zwei unterschiedlichen, privatisierten Datensätzen, eine Person identifizieren zu können. Da der verwendete Nextbike-Datensatz allerdings keine Kunden-IDs und somit keine Rückschlüsse auf die Routen einzelner Personen zulässt und die zusätzliche Informationsbeschaffung von Fahrraddaten sich als schwierig herausgestellt hat, kann dieser Punkt zunächst nur für zukünftige Analysen mitgenommen werden.

Die Auswertung der visualisierten Daten hat gezeigt, dass aus den privatisierten Datensätzen weiterhin informative Rückschlüsse und Interpretationen vorgenommen werden können. Außerdem konnten durch öffentlich zugängliche Informationen, die mit den Informationen aus dem Nextbike-Datensatz kombiniert worden sind, Zusammenhänge hergestellt werden, die über den Informationsgehalt des Nextbike-Datensatzes hinausgehen. Bei der Interpretation muss dennoch Vorsicht gezeigt werden, da es sich bei dem Datensatz um Ausleihen von Leihfahrrädern in einem kurzen Zeitraum von Juli bis September 2019 handelt. Einerseits beinhaltet der Datensatz nur Monate, in denen das Radfahren wettertechnisch weitestgehend täglich erfolgen kann. Hier kann weiterführend eine Analyse hinsichtlich der Wintermonate durchgeführt werden. Andererseits handelt es sich, im Gegensatz zum Movebis-Datensatz, um Leihfahrräder und nicht um das private Rad. Das bedeutet, der Start- und Endpunkt der Route kann zwar fast frei gewählt werden, dennoch besteht die Möglichkeit, dass am Startort einer Person kein Fahrrad verfügbar ist und die Person ein Stück zum nächsten Rad laufen muss. Weiterhin ist das Leihfahrrad bezüglich der Abstellmöglichkeiten leicht eingeschränkt, da auch Zonen existieren, in denen das Rad nicht abgestellt werden darf. Insofern wird die Interpretation an dieser Stelle leicht verzerrt.

Hinsichtlich der Fahrradklima-Analyse können weitere Aspekte analysiert werden: beispielsweise die Abstellorte der Leihfahrräder. Interessant ist hierbei, ob das Abstellen von Leihrädern vermehrt an Punkten wie Tram- oder S-Bahnstationen erfolgt und das Fahrrad so ein wichtiges Mittelstück einer insgesamt größeren zurückgelegten Strecke darstellt.

Um einen Schluss zuzulassen, wie weit entfernt die Visualisierungen und Auswertungen dieser Arbeit von tatsächlich gefahrenen Strecken liegen, wird zum Abschluss ein Vergleich mit der Visualisierung aus dem Stadtradeln 2020 hergestellt (siehe Abbildung 6.1). Auffällig ist, dass die Farben „verkehrt herum“ beziehungsweise nicht intuitiv genutzt wurden und gelb eine starke Intensität und rot eine schwache repräsentiert.

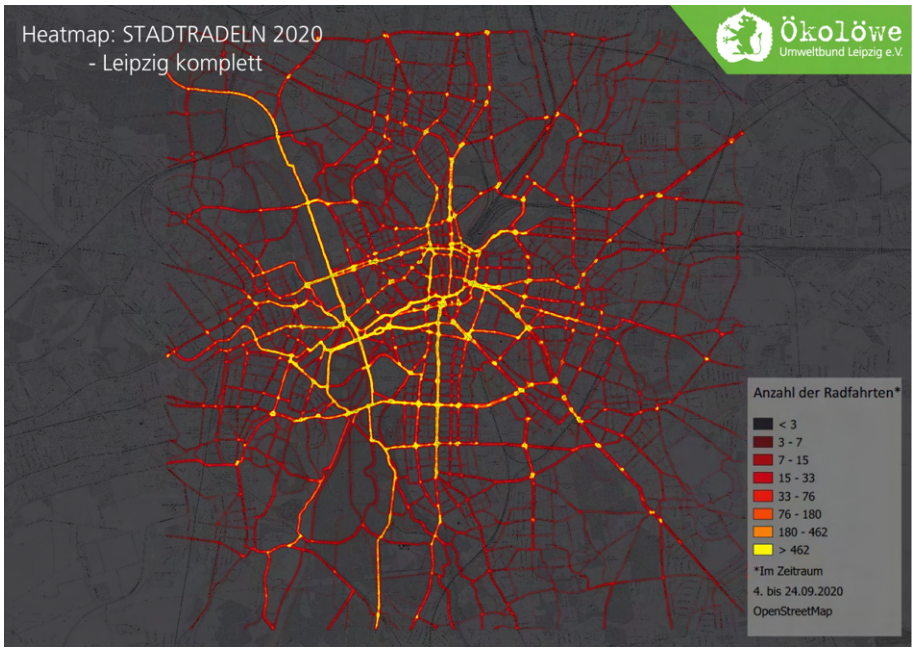


Abbildung 6.1.: Heat Map der Daten aus dem Stadtradeln im Zeitraum 04. bis 24.09.2020 von Ökolöwe Leipzig, Quelle: [50]

Laut Ökolöwe hat der Datensatz einen Schwachpunkt, da „*in dem Datensatz sehr deutlich die Route der Leipziger RADNACHT 2020 zu erkennen ist. Diese verlief um den Innenstadtring und die B2. Da diese der Auftakt zum Stadtradeln war, wurde die Strecke sehr oft mit der App aufgezeichnet und verfälscht den Datensatz an dieser Stelle.*“ [50]. Dennoch sind wie in den Ergebnissen dieser Untersuchung auch in der Heat Map des Ökolöwen die vier Hauptverkehrsstraßen Karl-Liebknecht-Straße (Richtung Süden), Clara-Zetkin-Park (Richtung Westen), Gerichtsweg (Richtung Osten), Rosa-Luxemburg-Straße (Richtung Nord-Osten) zu erkennen. Dies validiert ein Stück weit die in dieser Arbeit vorgenommenen Interpretationen. Zur vollständigen Validierung und Weiterführung der Arbeit müssen die hier vorgestellten Methoden auf Datensätze mit tatsächlich gefahrenen Strecken angewandt werden.

Literaturverzeichnis

[1] B. Liu u. a. "Location Privacy and Its Applications: A Systematic Study". In: *IEEE Access* 6 (2018), S. 17606–17624. DOI: 10.1109/ACCESS.2018.2822260.

[2] Chi-Yin Chow und Mohamed F. Mokbel. "Trajectory Privacy in Location-Based Services and Data Publication". In: *SIGKDD Explor. Newsl.* 13.1 (Aug. 2011), S. 19–29. ISSN: 1931-0145. DOI: 10.1145/2031331.2031335.

[3] A. R. Beresford und F. Stajano. "Location privacy in pervasive computing". In: *IEEE Pervasive Computing* 2.1 (2003), S. 46–55. DOI: 10.1109/MPRV.2003.1186725.

[4] Juha Oksanen u. a. "Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data". In: *Journal of Transport Geography* 48 (2015), S. 135–144. ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2015.09.001.

[9] A.J. Blumberg und P. Eckersley. "On locational privacy, and how to avoid losing it forever". In: *Electronic frontier foundation* 10.11 (2009), S. 1–7.

[10] K. G. Shin u. a. "Privacy protection for users of location-based services". In: *IEEE Wireless Communications* 19.1 (2012), S. 30–39. DOI: 10.1109/MWC.2012.6155874.

[12] Jan Holvast. "History of Privacy". In: *The Future of Identity in the Information Society*. Hrsg. von Vashek Matyáš u. a. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, S. 13–42.

[13] Benjamin Henne u. a. "Selective cloaking: Need-to-know for location-based apps". In: *2013 Eleventh Annual Conference on Privacy, Security and Trust*. 2013, S. 19–26. DOI: 10.1109/PST.2013.6596032.

[14] Song Gao u. a. "Exploring the effectiveness of geomasking techniques for protecting the geoprivacy of Twitter users". In: *Journal of Spatial Information Science* (Dez. 2019). DOI: 10.5311/JOSIS.2019.19.510.

[15] Rakesh Agrawal u. a. "Chapter 14 - Hippocratic Databases". In: *VLDB '02: Proceedings of the 28th International Conference on Very Large Databases*. Hrsg. von Philip A. Bernstein u. a. San Francisco: Morgan Kaufmann, 2002, S. 143–154. ISBN: 978-1-55860-869-6.

[16] A. S. M. Hasan, Qingshan Jiang und Chengming Li. "An Effective Grouping Method for Privacy-Preserving Bike Sharing Data Publishing". In: *Future Internet* 9 (Okt. 2017), S. 65. DOI: 10.3390/fi9040065.

[17] Latanya Sweeney. "k-anonymity: A model for protecting privacy". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), S. 557–570.

[18] M. Douriez u. a. "Anonymizing NYC Taxi Data: Does It Matter?" In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2016, S. 140–148. DOI: 10.1109/DSAA.2016.21.

[20] P. Samarati und L. Sweeney. "Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression". In: *Technical Report SRI-CSL-98-04* (1998).

[21] Khaled El Emam und Fida Kamal Dankar. "Protecting Privacy Using k-Anonymity". In: *Journal of the American Medical Informatics Association* 15.5 (Sep. 2008), S. 627–637. ISSN: 1067-5027. DOI: 10.1197/jamia.M2716.

[22] Miguel E. Andrés u. a. "Geo-Indistinguishability: Differential Privacy for Location-Based Systems". In: *Proceedings of the 2013 ACM SIGSAC conference on Computer communications security - CCS '13* (2013). DOI: 10.1145/2508859.2516735.

[23] Fengmei Jin u. a. "A Survey and Experimental Study on Privacy-Preserving Trajectory Data Publishing". In: *TechRxiv* (2021).

[24] Isabel Wagner und David Eckhoff. "Technical Privacy Metrics: A Systematic Study". In: *ACM Computing Surveys* 51.3 (Juli 2018), S. 1–38. ISSN: 1557-7341. DOI: 10.1145/3168389.

[25] Raymond Chi-Wing Wong u. a. "(, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, S. 754–759. ISBN: 1595933395. DOI: 10.1145/1150402.1150499.

[26] Roberto J Bayardo und Rakesh Agrawal. "Data privacy through optimal k-anonymization". In: *21st International conference on data engineering (ICDE'05)*. IEEE. 2005, S. 217–228.

[27] Vanessa Ayala-Rivera u. a. "A Systematic Comparison and Evaluation of K-Anonymization Algorithms for Practitioners". In: *Trans. Data Privacy* 7.3 (Dez. 2014), S. 337–370. ISSN: 1888-5063.

[28] P. Kiran und P. KavyaN. "A Survey on Methods, Attacks and Metric for Privacy Preserving Data Publishing". In: *International Journal of Computer Applications* 53 (2012), S. 20–28.

[29] R. Shokri u. a. "Quantifying Location Privacy". In: *2011 IEEE Symposium on Security and Privacy*. 2011, S. 247–262. DOI: 10.1109/SP.2011.18.

[30] G. F. Marias u. a. "Location privacy through secret sharing techniques". In: *Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks*. 2005, S. 614–620. DOI: 10.1109/WOWMOM.2005.60.

[31] Sergio Mascetti u. a. "Privacy in geo-social networks: Proximity notification with untrusted service providers and curious buddies". In: *The VLDB Journal* 20 (Juli 2010). DOI: 10.1007/s00778-010-0213-7.

[32] William B. Allshouse u. a. "Geomasking sensitive health data and privacy protection: an evaluation using an E911 database". In: *Geocarto International* 25.6 (2010). PMID: 20953360, S. 443–452. DOI: 10.1080/10106049.2010.496496.

[33] Marco Gruteser und Dirk Grunwald. "Anonymous usage of location-based services through spatial and temporal cloaking". In: *Proceedings of the 1st international conference on Mobile systems, applications and services*. 2003, S. 31–42.

[34] B. Gedik und L. Liu. "Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms". In: *IEEE Transactions on Mobile Computing* 7.1 (2008), S. 1–18. DOI: 10.1109/TMC.2007.1062.

[35] Jani Sainio, Jan Westerholm und Juha Oksanen. “Generating Heat Maps of Popular Routes Online from Massive Mobile Sports Tracking Application Data in Milliseconds While Respecting Privacy”. In: *ISPRS International Journal of Geo-Information* 4.4 (2015), S. 1813–1826. ISSN: 2220-9964.

[44] Mike DeBoer. “Understanding the heat map”. In: *Cartographic perspectives* 80 (2015), S. 39–43.

Online-Quellenverzeichnis

[5] *Statistik: Anzahl zugelassener Pkw in Deutschland von 1960 bis 2021.* https://de.statista.com/statistik/daten/studie/12131/umfrage/pkw-bestand-in-deutschland/. [Online; letzter Zugriff: 01.06.2021]. 2021.

[6] *Radverkehr in Deutschland – Zahlen, Daten, Fakten, Bundesministerium für Verkehr und digitale Infrastruktur.* https://www.bmvi.de/SharedDocs/DE/Publikationen/K/radverkehr-in-zahlen.html. [Online; letzter Zugriff: 31.05.2021]. 2020.

[7] Deutscher Städte- und Gemeinschaftsbund und Allgemeiner Deutscher Fahrrad-Club (ADFC). *Förderung des Radverkehrs in Städten und Gemeinden.* https://repository.difu.de/jspui/bitstream/difu/581435/1/doku-158-radverkehr-dstgb-adfc-komprimiert.pdf. 2021.

[8] *ADFC - Leitlinien zur Radverkehrsinfrastruktur.* https://www.adfc.de/fileadmin/user_upload/Im-Alltag/Radverkehrsfoerderung/Download/ADFC-Leitlinien-Fahrradinfrastruktur_gestaltete-Endversion.pdf. [Online; letzter Zugriff: 06.07.2021]. 2021.

[11] Luisa Rollenhagen. *Alan Westin is the father of modern data privacy law.* https://www.osano.com/articles/alan-westin. [Online; letzter Zugriff: 23.06.2021].

[19] *Bildquelle Privacy Utility Trade-off.* https://aircloak.com/history-of-data-anonymization/. [Online; letzter Zugriff: 09.07.2021].

[36] *OpenStreetMap-Wiki Genauigkeit von Koordinaten.* https://wiki.openstreetmap.org/wiki/DE:Genauigkeit_von_Koordinaten. [Online; letzter Zugriff: 31.09.2021]. 2021.

[37] *Open Route Service API Documentation.* https://openrouteservice.org/dev/#/api-docs. [Online; letzter Zugriff: 17.3.2021]. 2021.

[38] *RFC 7946 - The GeoJSON Format.* https://tools.ietf.org/html/rfc7946. [Online; letzter Zugriff: 02.04.2021]. 2016.

[39] *Nominatim API Documentation.* https://nominatim.org/release-docs/develop/api/Overview/. [Online; letzter Zugriff: 17.3.2021]. 2021.

[40] *Pelias API Documentation.* https://github.com/pelias/documentation/. [Online; letzter Zugriff: 04.06.2021]. 2021.

[41] *Flow Maps in Excel.* https://www.clearlyandsimply.com/clearly_and_simply/2020/03/geographical-flow-maps-in-excel-part-1-of-3.html. [Online; letzter Zugriff: 17.3.2021]. 2020.

[42] *Flow Map Wiki.* https://de.xcv.wiki/wiki/Flow_map. [Online; letzter Zugriff: 02.06.2021]. 2021.

[43] *Software zum Erstellen von Flow Map Visualisierungen.* https://flowmap.blue. [Online; letzter Zugriff: 02.06.2021]. 2021.

[45] *Open Route Service Profil für Generierung von Fahrradrouen.* https://ask.openrouteservice.org/t/cycle-safe-profile-direction-api/139. [Online; letzter Zugriff: 02.04.2021].

[46] *OpenStreetMap-Wiki Key:cycleway*. `https://wiki.openstreetmap.org/wiki/DE:Key:cycleway`. [Online; letzter Zugriff: 26.05.2021].

[47] *OpenStreetMap-based bicycle map*. `https://www.cyclosm.org/#map=14/51.3405/12.3915/cyclosm`. [Online; letzter Zugriff: 26.05.2021].

[48] *OpenStreetMap-Wiki Key:smoothness*. `https://wiki.openstreetmap.org/wiki/DE:Key:smoothness`. [Online; letzter Zugriff: 26.05.2021].

[49] *Movebis Auswertung von Crowdsourced-Daten zur Verbesserung der kommunalen Fahrradinfrastruktur*. `https://www.movebis.org`. [Online; letzter Zugriff: 02.06.2021]. 2021.

[50] *Leipzigs leuchtender Fahrradplan*. `https://www.oekoloewe.de/nachhaltige-mobilitaet-stadtentwicklung-detail/leipzigs-leuchtender-fahrradplan.html`. [Online; letzter Zugriff: 02.06.2021]. 2021.

Erklärung

Ich versichere, dass ich die vorliegende Arbeit mit dem Thema:

„Privatsphäre-erhaltende Analyse des Fahrradklimas in Leipzig“

selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.

Leipzig, den

_____

Aruscha Kramm

A. Anhang

(a) k = 0

(b) k = 10

(c) k = 20

(d) k = 50

(e) k = 100

(f) k = 1000

Heat Maps der Methode Donutmasking für unterschiedliche k-Werte

(a) k = 0

(b) k = 10

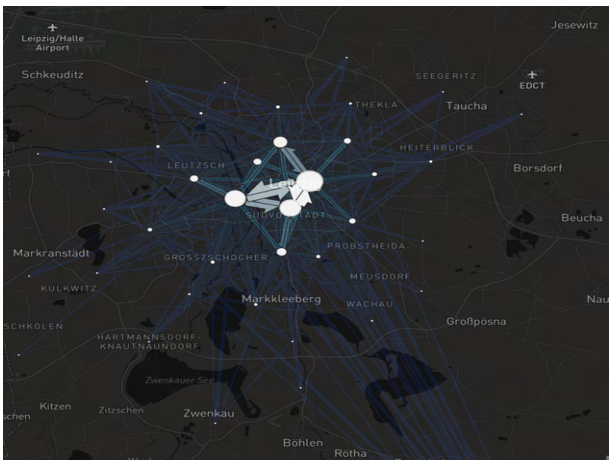
(c) k = 20

(d) k = 50

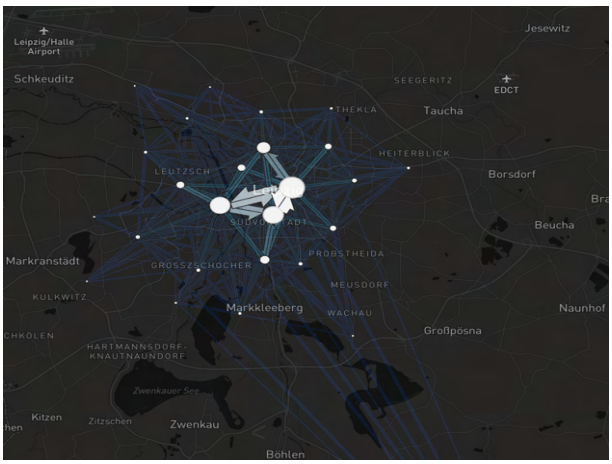
(e) k = 100

(f) k = 1000

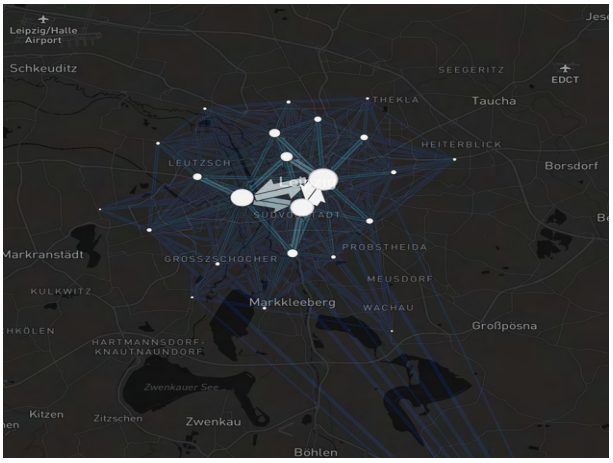
Heat Maps der Methode Segmentaggregation für unterschiedliche k-Werte



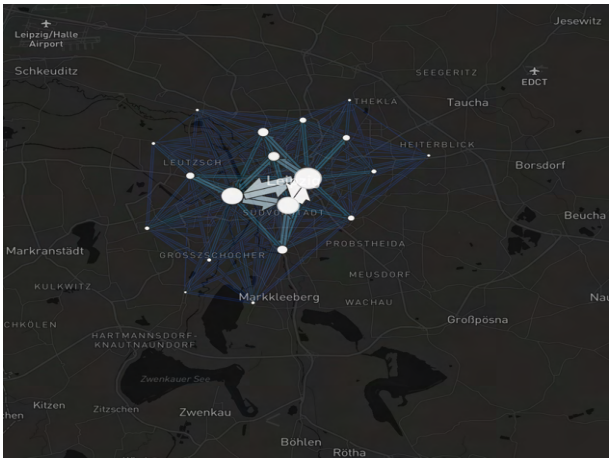
(a) k = 0



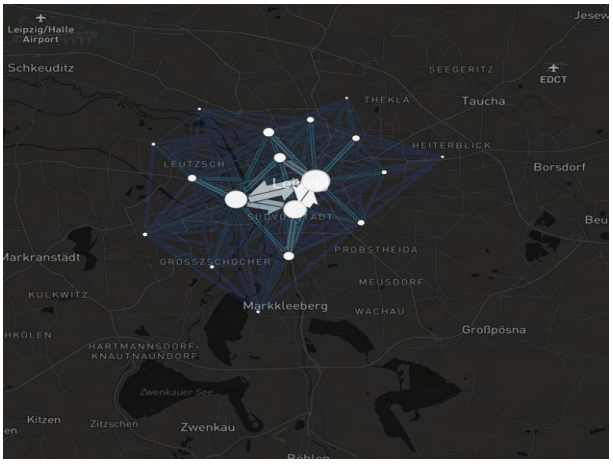
(b) k = 10



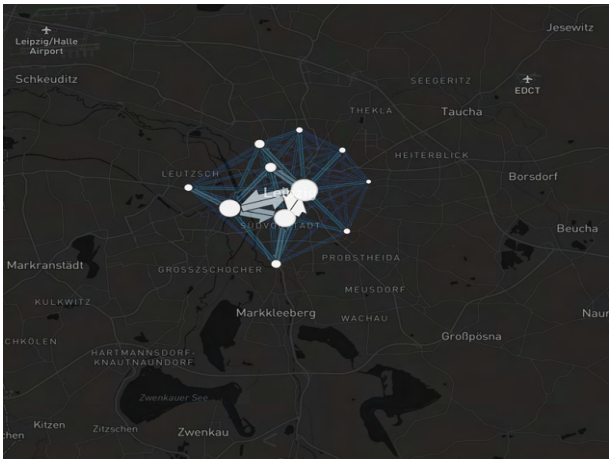
(c) k = 20



(d) k = 50

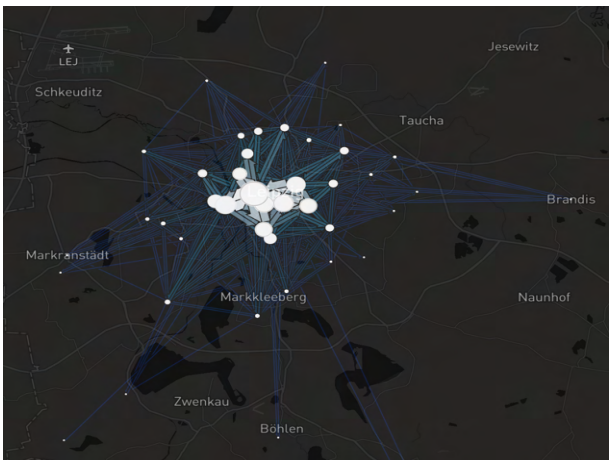


(e) k = 100

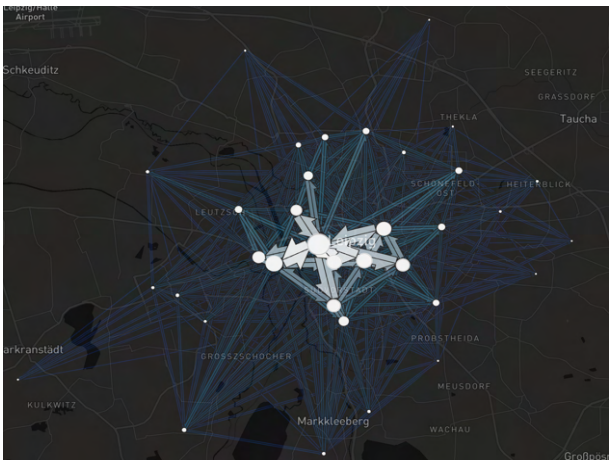


(f) k = 1000

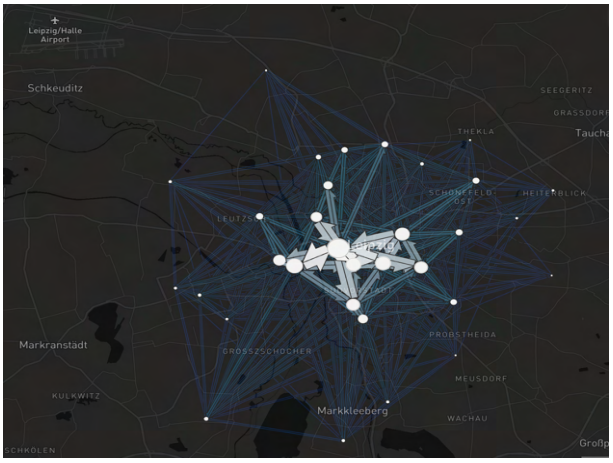
Flow Maps der Methode Zentroidaggregation (Stadtteile) für unterschiedliche k-Werte



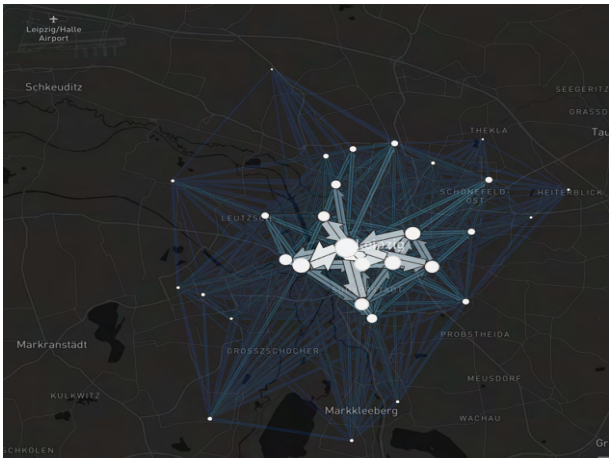
(a) k = 0



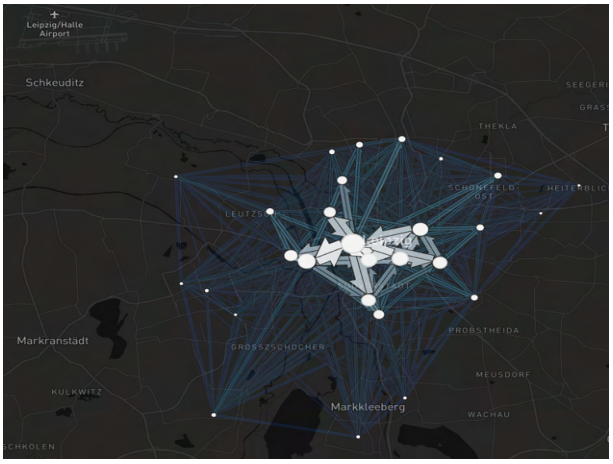
(b) k = 10



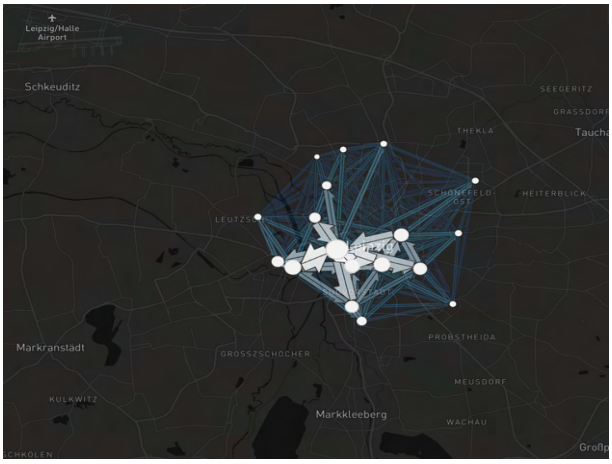
(c) k = 20



(d) k = 50

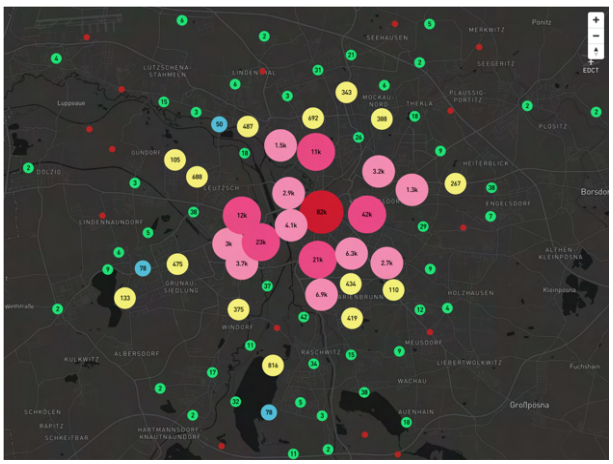


(e) k = 100

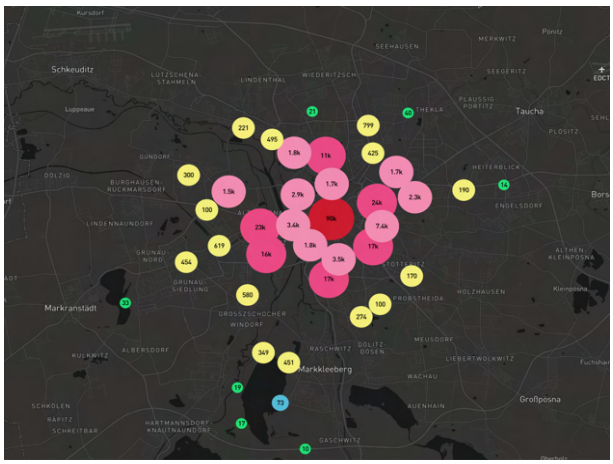


(f) k = 1000

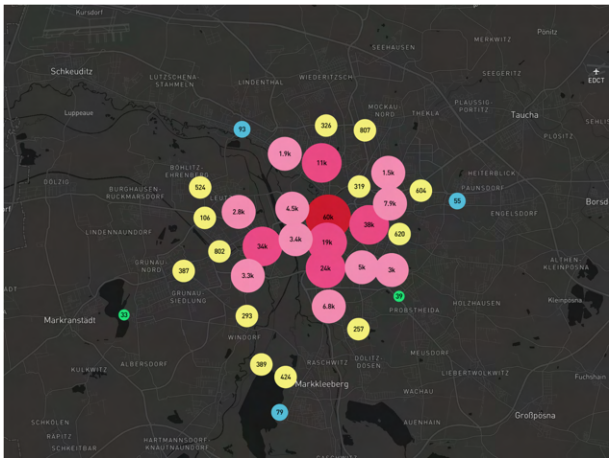
Flow Maps der Methode Zentroidaggregation (Postleitzahlen) für unterschiedliche k-Werte



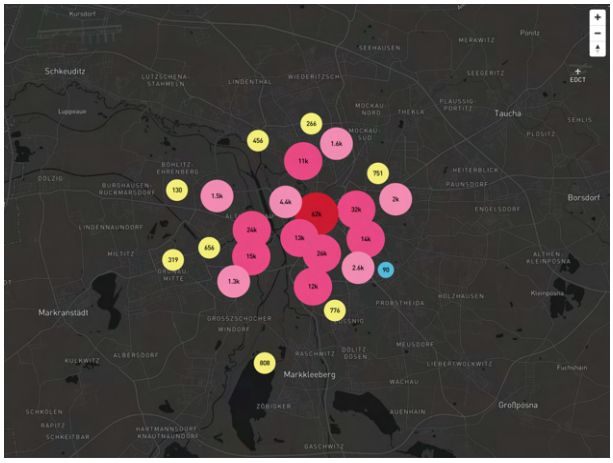
(a) k = 0



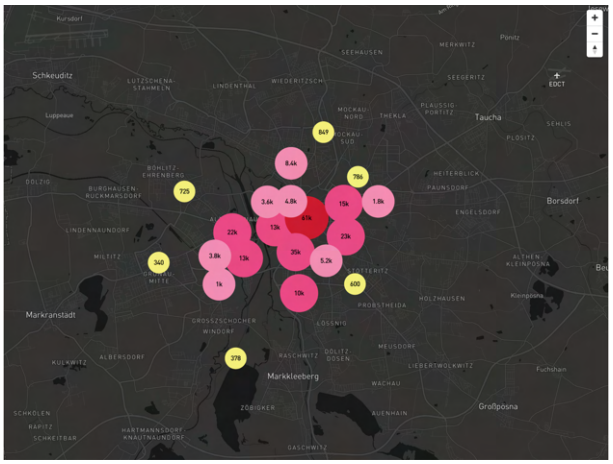
(b) k = 10



(c) k = 20



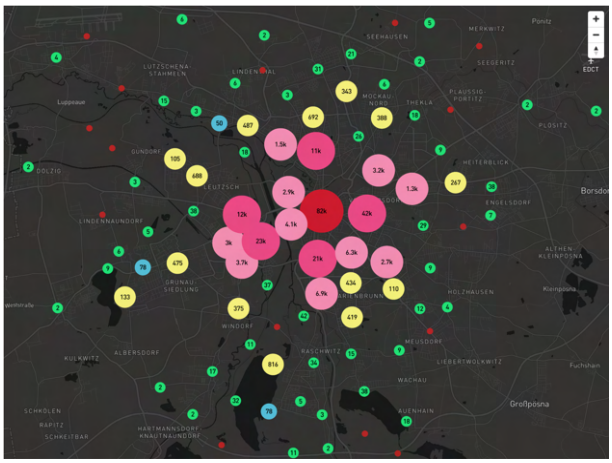
(d) k = 50



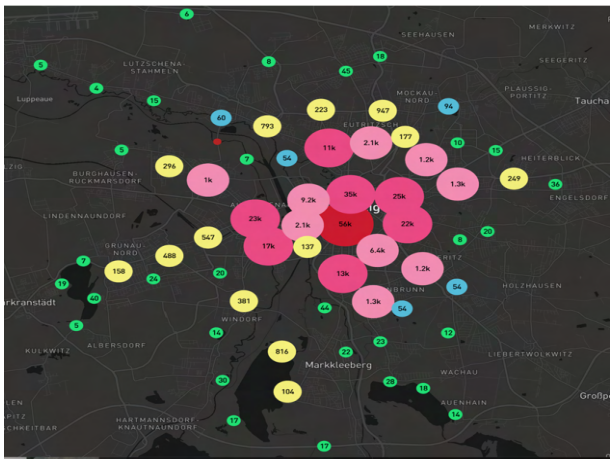
(e) k = 100

(f) k = 1000

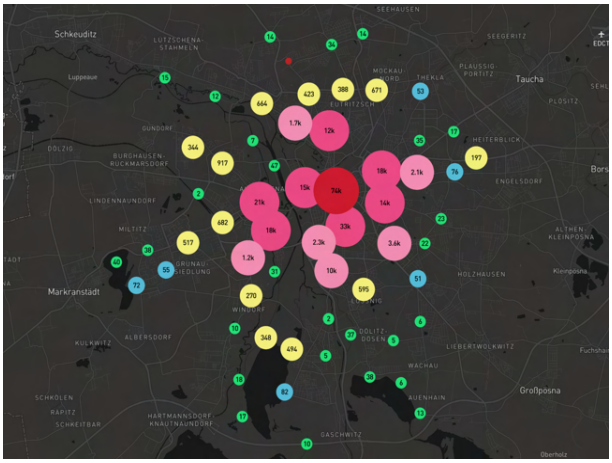
Punkt Cluster für Startpunkte der Methode Donutmasking



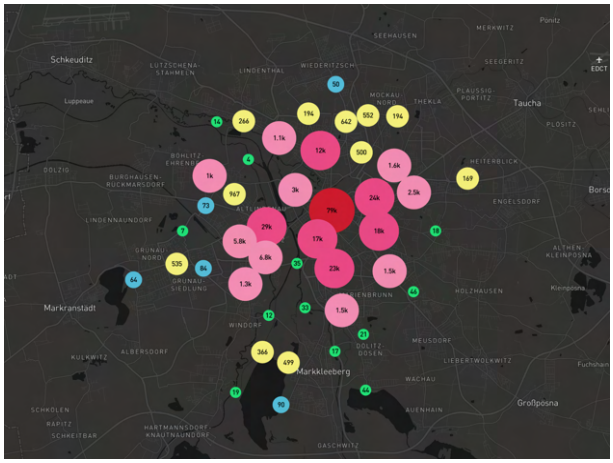
(a) k = 0



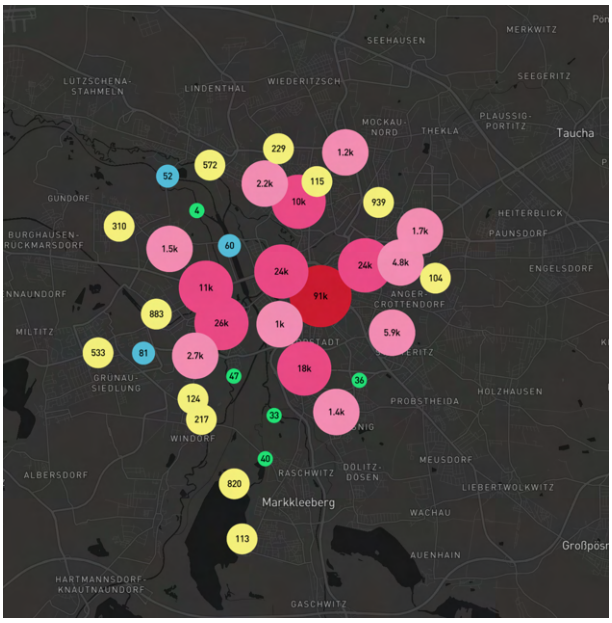
(b) k = 10



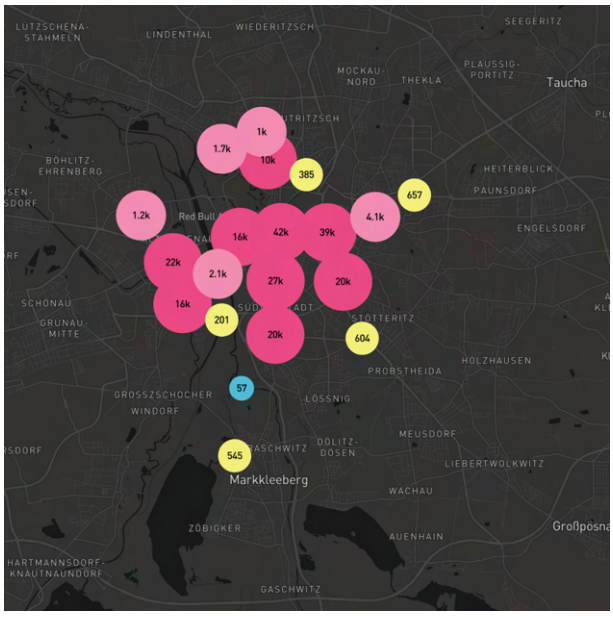
(c) k = 20



(d) k = 50



(e) k = 100



(f) k = 1000

Punkt Cluster für Startpunkte der Methode Segmentaggregation