Technische Universität Berlin

Fakultät IV Elektrotechnik und Informatik
Institut für Softwaretechnik und Theoretische Informatik
Fachgebiet Big Data Management
Einsteinufer 17, 10587 Berlin

Deutsches Zentrum für Luft- und Raumfahrt (DLR)
Institut für Verkehrssystemtechnik
Technologiefeld Datenerfassung und Informationsgewinnung
Rutherfordstraße 2, 12489 Berlin

Masterarbeit zum Thema

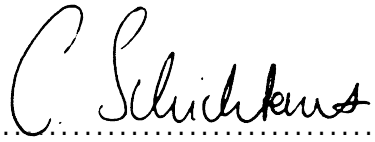
Anomalieerkennung in Straßenverkehrsdaten einer urbanen Kreuzung

Vorgelegt von:	Clemens Schicktanz
Matrikelnummer:	397889
Studiengang:	M. Sc. Wirtschaftsingenieurwesen
Abgabedatum:	22.05.2020
Erstgutachter:	Prof. Dr. Ziawasch Abedjan
Zweitgutachter:	Prof. Dr. Peter Wagner

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, den 22.05.2020



.....

Clemens Schicktanz

Abstract

Current transport science aims to reduce the amount of traffic crashes by observing behavioral patterns of traffic participants. Assuming that exceptional, abnormal situations in road traffic increase the risk of accidents, the following thesis attempts to analyze those as a prior task. Anomaly detection techniques can be used to capture such situations in large data sets. Therefore, this thesis is dedicated to the central research question, asking which and to what extent anomalies can be identified in a road traffic data set of an urban intersection. Due to the interaction with oncoming vehicles the highest amount of abnormal incidents is presumed for left turning cars. Thus, their data will be focused on primarily out of the examined data set of 430,000 trajectories.

According to their cause the anomalies detected in these data are differentiated into erroneous and exceptional data. The thesis will especially focus on the acquisition of anomalies caused by exceptional data. This enables the evaluation of the behavior of road users rather than analyzing the quality of data sets by assessing erroneous data. Besides classifying anomalies according to their cause, a distinction will be made between distance-, density- and direction-based anomalies.

To map all different types of anomalies five methods are being used for the detection of exceptional data. One of them is based on the Hausdorff metric, three on the representation of data in a discrete Euclidean space and the remaining one on the application of Gaussian mixture models. The Hausdorff metric is mainly used to identify anomalies in which the position of a road user deviates significantly from their expected route. By transforming the data in a discrete Euclidean space the number of trajectories per cell can be determined. Using this results, trajectories can be evaluated according to the number of other trajectories that are being detected at their position. Anomaly detection with regards to the change of the position of road users is divided into two different approaches – the measured heading of traffic participants provides more significant data for the anomaly detection than the transition probabilities between positions. Compared to the previously mentioned methods, Gaussian mixture models also consider and classify as “anomalies” those vehicles that are perceived by the traffic monitoring cameras while standing still.

It is concluded that the applied procedures enable the user to identify situations that are considered “exceptional” according to various criteria from the given data set. The number of anomalies in an inspected data set depends on the individual user who determines the threshold above which a trajectory is classified as abnormal.

Kurzfassung

In der Verkehrswissenschaft wird angestrebt, die Anzahl von Verkehrsunfällen zu reduzieren, indem das Verhalten der Verkehrsteilnehmer untersucht wird. Da angenommen wird, dass außergewöhnliche, anomale Situationen im Straßenverkehr das Unfallrisiko erhöhen, sollten diese vorrangig analysiert werden. Zur Detektion solcher Situationen können Verfahren der Anomalieerkennung angewandt werden. Daher wird mit dieser Arbeit der Forschungsfrage nachgegangen, welche Anomalien in welchem Umfang in einem Straßenverkehrsdatensatz einer urbanen Kreuzung vorliegen. Im zu untersuchenden Datensatz von 430.000 Trajektorien werden besonders die Positionsdaten der als PKW klassifizierten Linksabbieger untersucht, da hier aufgrund der Interaktion mit den entgegenkommenden Fahrzeugen die meisten anomalen Situationen erwartet werden.

Die in diesen Daten erfassten Anomalien werden nach deren Ursache in fehlerhafte und außergewöhnliche Daten unterschieden. In dieser Arbeit wird sich auf die Erkennung von Anomalien, die auf außergewöhnlichen Daten basieren, konzentriert. Damit wird das Verhalten von Verkehrsteilnehmern analysiert, anstatt mit der Analyse von fehlerhaften Daten die Qualität des Datensatzes zu bewerten. Neben der Einteilung von Anomalien nach deren Ursache, erfolgt auch eine Unterscheidung nach deren Art in distanz-, dichte- und richtungsbasierte Anomalien.

Für die Erkennung außergewöhnlicher Daten werden fünf Verfahren angewandt, mit denen alle Arten von Anomalien detektiert werden können. Eines der Verfahren basiert auf der Hausdorff-Metrik, drei davon auf der Repräsentation der Daten in einem diskreten euklidischen Raum und das fünfte auf der Anwendung von Gaußschen Mischmodellen. Mit der Hausdorff-Metrik werden vor allem Anomalien identifiziert, bei denen die Position eines Verkehrsteilnehmers deutlich von der erwarteten Route abweicht. Mit der Repräsentation der Daten in einem diskreten euklidischen Raum kann die Anzahl der Trajektorien pro Zelle ermittelt werden, sodass Trajektorien danach bewertet werden können, wie viele andere Trajektorien an deren Position aufgezeichnet werden. Für die Anomalieerkennung in den Werten der Positionsänderung eines Verkehrsteilnehmers, stellt die gemessene Bewegungsrichtung eine bessere Datengrundlage dar als die Übergangswahrscheinlichkeiten zwischen den Zellen. Im Vergleich zu den zuvor genannten Verfahren werden mit den Gaußschen Mischmodellen auch Fahrzeuge als Anomalie klassifiziert, die im Stehen von den Kameras erfasst werden.

Aus den Ergebnissen wird geschlussfolgert, dass die angewandten Verfahren es dem Anwender ermöglichen, Situationen die nach verschiedenen Kriterien als außergewöhnlich gelten, aus einem Datensatz zu ermitteln. Die Anzahl der Anomalien in einem Datensatz hängt vom Anwender der Anomalieerkennung ab, der den Schwellenwert festlegt, ab dem eine Trajektorie als anomal zu klassifizieren ist.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AOI	Areas of Interest
CRISP-DM	Cross-Industry Standard Process for Data Mining
DLR	Deutsches Zentrum für Luft- und Raumfahrt e. V.
EM	Erwartungs-Maximierung
Fokr	Forschungskreuzung vom DLR in Braunschweig
FPS	frames per second
LSA	Lichtsignalanlage
MRU	motorisierter Verkehrsteilnehmer

Kapitel 1

Einleitung

1.1 Motivation

Mit 2,6 Millionen Verkehrsunfällen im Jahr 2018 und einem Anstieg der Zahl um 1,5 % pro Jahr zwischen 2012 und 2018 [Sta18, S.44], ist die Verhinderung von Verkehrsunfällen nach wie vor ein ungelöstes Problem in der Verkehrswissenschaft. Um dieses Problem zu beheben, wird unter anderem versucht, das Verhalten von Verkehrsteilnehmern anhand der Analyse von Daten besser zu verstehen.

Eine Möglichkeit der Datenanalyse ist die Anomalieerkennung, bei der Daten nach außergewöhnlichen Werten überprüft werden. Die Detektion dieser seltenen Ereignisse in dem enormen Datenbestand stellt den Menschen vor einen Informationsüberfluss, der nur durch die Unterstützung von maschinell ausgeführten Algorithmen bewältigt werden kann. Diese können aus historischen Daten normales Verhalten eines Verkehrsteilnehmers erlernen und abnormales Verhalten davon unterscheiden. Die identifizierten anomalen Straßenverkehrsdaten können ein Indikator für illegale und kritische Manöver von Verkehrsteilnehmern sein. Damit ermöglicht die Detektion dieser Anomalien, proaktiv Maßnahmen zu entwickeln, um die Anzahl von illegalen und kritischen Manövern zu reduzieren und somit Verkehrsunfälle zu reduzieren. [Lax13]

1.2 Ziel und Forschungsfrage

Daher ist das übergeordnete Ziel dieser Arbeit die Erkennung von Anomalien in Straßenverkehrsdaten. Damit wird die Forschungsfrage „Welche Anomalien befinden sich in welchem Umfang im vorliegenden Straßenverkehrsdatensatz?“ verfolgt. So soll ein Beitrag zum besseren Verständnis des Verkehrsteilnehmerverhaltens an urbanen Kreuzungen geleistet werden, der letztendlich die Verbesserung der Verkehrssicherheit unterstützt. Die Forschungsfrage wird beantwortet, indem mit ausgewählten Methoden der Anomalieerkennung Verkehrsteilnehmer identifiziert werden, deren Verhalten vom erwarteten abweicht. Hierfür wird zunächst mit den jeweiligen Methoden das erwartete Verhalten eines Verkehrsteilnehmers modelliert.

Eine erfolgreiche Erreichung des Ziels ist allerdings nur gewährleistet, wenn die zu analysierenden Daten die Realität wiedergeben. Daher müssen für die Datenanalyse, die Daten, die nicht die Realität wiedergeben, sondern durch eine fehlerhafte Aufzeichnung entstanden sind, aus dem zu untersuchenden Datensatz entfernt werden. Damit soll dem entgegengewirkt werden, dass fehlerhafte Schlussfolgerungen auf Basis fehlerhafter Daten

getroffen werden.

Neben der Entfernung fehlerhafter Daten aus dem zu analysierenden Datensatz, müssen weitere Einschränkungen getroffen werden, die nachfolgend vorgestellt werden.

1.3 Abgrenzung des Themas

Zur Erreichung des zuvor beschriebenen Ziels wird ein Datensatz von Trajektorien (TJ) einer urbanen Kreuzung in Braunschweig analysiert. Demnach werden in dieser Masterarbeit nur die Daten von der Forschungskreuzung (Fokr) analysiert und kein Vergleich der angewandten Verfahren zu anderen Straßenverkehrsdaten durchgeführt. Der Umfang des zu analysierenden Datensatzes beträgt acht Tage und wird im Abschnitt 3.1.1 detailliert beschrieben.

Die Trajektorien im Datensatz werden der Kategorie multivariater Zeitreihen zugeordnet, da mehrere Datenpunkte in einem zeitlichen Verlauf erhoben werden. Die Anomalieerkennung in solchen Datentypen umfasst diverse Bereiche, da die einzelnen Zeitreihen oder Datenpunkte beliebig kombiniert werden können. Dadurch entsteht ein sehr komplexes Problem und es eröffnet sich eine Vielzahl an möglichen Fragestellungen für den Datensatz. [Gup13, cgl. S. 2]

Im Rahmen dieser Masterarbeit erfolgt daher eine Konzentration auf die Anomalieerkennung in den Positionsdaten der Verkehrsteilnehmer. Hierfür muss zunächst deren erwartetes Verhalten modelliert werden. Es wird davon ausgegangen, dass das Verhalten eines Verkehrsteilnehmers von der Objektklasse und Route abhängt. Daher werden die Trajektorien der PKW auf der Route von Westen nach Norden zum Gegenstand dieser Arbeit erklärt. Die Objektklasse der PKW ist die auf der Forschungskreuzung am häufigsten detektierte Objektklasse, weshalb durch die Analyse dieser der größte Bestandteil des Datensatzes untersucht wird. Die gewählte Route ist die einzige Route der Kreuzung, an der zwei kreuzende Verkehrsströme von PKWs nicht durch die Lichtsignalanlagen (LSA) gesteuert werden. Damit wird darauf abgezielt, Ergebnisse über die Anomalien einer Route zu erhalten, bei der ein „Einbiegen/Kreuzen-Unfall“ geschehen kann. Mit einem Anteil von 25,6 % an allen Unfällen an innerstädtischen Kreuzungen ist dies der häufigste Unfalltyp in Deutschland [Sta18, S.44]. Demnach sind weitere Erkenntnisse über diesen Unfalltyp von größter Bedeutung für die Verringerung der Unfälle an innerstädtischen Kreuzungen.

Die Anomalieerkennung in diesen Straßenverkehrsdaten kann mittels beaufsichtigter, halb- und unbeaufsichtigter Verfahren praktiziert werden. [Men19] In der vorliegenden Arbeit werden nur unbeaufsichtigte Verfahren angewandt, da keine Informationen über die Anormalität der Daten vorliegt. Weitere Informationen zu den ausgewählten Verfahren sind dem Abschnitt 2.5 zu entnehmen.

1.4 Aufbau der Arbeit

Zum Erreichen des Ziels dieser Arbeit wird sich am „Cross-Industry Standard Process for Data Mining“ orientiert, dessen Ablauf schematisch in Abbildung 1.1 dargestellt wird.

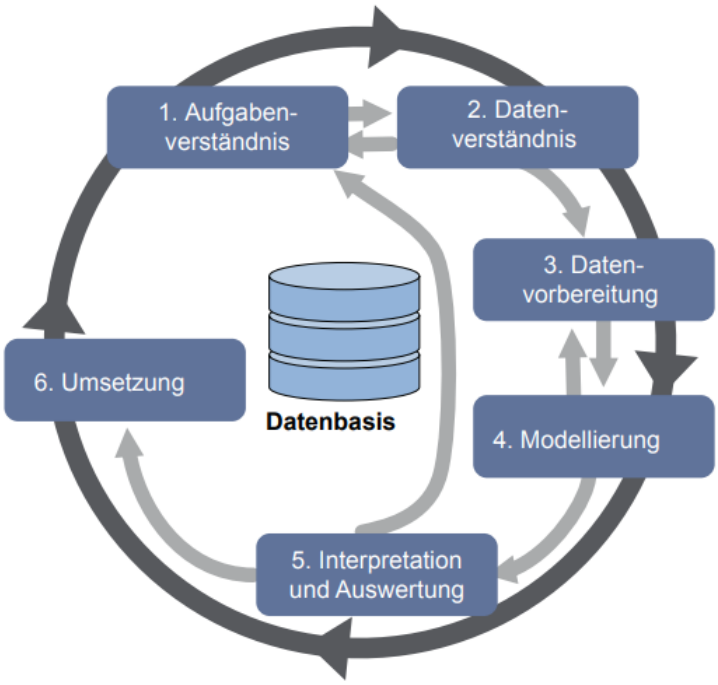


Abbildung 1.1: Ablauf des branchenübergreifenden Prozesses für Data Mining „CRISP-DM“ [Vaj09, S. 443]

Dieser beinhaltet die Schritte Aufgabenverständnis, Datenverständnis, Datenvorbereitung, Modellierung, Interpretation/Auswertung und Umsetzung. Alle in diesem Vorgehen beschriebenen Schritte werden für die Anomalieerkennung in dieser Arbeit durchlaufen.

Der Schritt des Aufgabenverständnisses wird im 2. Kapitel behandelt, in dem die theoretischen Grundlagen bezüglich der Verkehrswissenschaft und die Methoden der Anomalieerkennung vorgestellt werden. Des Weiteren werden in diesem Kapitel die Messeinrichtungen an der Fokr beschrieben. Dies kann mit dem Schritt des Datenverständnisses gleichgesetzt werden, da die Qualität der Daten und deren Anomalien von den Messeinrichtungen, die die Daten aufzeichnen, abhängig ist. Ein weiterer Bestandteil des Datenverständnisses wird in Abschnitt 3.1.1 beschrieben, da dort die zu analysierenden Daten vorgestellt werden.

Der dritte Schritt im CRISP-DM, die Datenvorbereitung, wird in der vorliegenden Arbeit im 3. Kapitel beschrieben. Dabei wird darauf eingegangen, welche fehlerhaften Daten aus dem Datensatz entfernt werden.

Anschließend können aus den bereinigten Daten Modelle aufgestellt werden, die das erwartete Verhalten eines Verkehrsteilnehmers beschreiben. Dies geschieht separat für jedes der anzuwendenden Verfahren im 4. Kapitel. Nach dem Aufstellen der Modelle werden die Daten auf das Modell angewandt, um Anomalien zu identifizieren. Diese werden sogleich interpretiert, womit der Schritt der Interpretation und Auswertung ebenso in diesem Kapitel behandelt wird. Da die Resultate verwendet werden, um die Modelle für die Anomalieerkennung zu optimieren, erfolgt auch der im CRISP-DM beschriebene Kreislauf und das Wechseln zwischen aufeinanderfolgenden Phasen.

Der Inhalt des 5. Kapitels lässt sich nicht dem CRISP-DM zuordnen, da in diesem Kapitel ein Ausblick über mögliche zukünftige Forschungsfragen gegeben wird. Diese Forschungsfragen haben sich aus der Bearbeitung dieser Arbeit ergeben. Erste Ansatzpunkte für weitere Arbeiten stellen zum Beispiel die in Abschnitt 1.3 genannten Einschränkungen des Umfangs dieser Arbeit dar.

Der letzte Schritt der Umsetzung kann auch als Bereitstellung bezeichnet werden. Dieser beschreibt die Aufbereitung der finalen Ergebnisse, um diese einem Auftraggeber vorzustellen. In der vorliegenden Arbeit wird dieser Schritt im 6. Kapitel dokumentiert, indem die Ergebnisse zusammengefasst werden. [Vaj09]

Kapitel 2

Theoretische Grundlagen

In diesem Kapitel werden die für diese Arbeit relevanten Informationen aus den Themenbereichen der Verkehrswissenschaft und Anomalieerkennung dargelegt, um eine Grundlage für das Verständnis der angewandten Methoden zu schaffen. Dazu wird in einem ersten Schritt auf verkehrswissenschaftliche Grundlagen und die Messeinrichtungen an der Fokr eingegangen. Daraufhin erfolgt die Vorstellung des theoretischen Teils der Anomalieerkennung, bevor abschließend die anzuwendenden Verfahren der Hausdorff-Metrik, des diskreten euklidischen Raums und des Gaußschen Mischmodells erläutert werden.

2.1 Verkehrswissenschaft

Die Verkehrswissenschaft stellt in der vorliegenden Arbeit den Anwendungsbereich der Anomalieerkennung dar. Daher werden nachfolgend die Größen eingeführt, die zur Quantifizierung des Straßenverkehrs für diese Arbeit relevant sind. Ebenso wird die Trajektorie definiert, da in dieser Arbeit vor allem Anomalien in diesem Konstrukt detektiert werden.

2.1.1 Definition Verkehrskenngrößen

Verkehr wird definiert als „die Ortsveränderung von Personen und Gütern“ [Leu13, S. III]. Bei der Betrachtung des Straßenverkehrs wird sowohl die Ortsveränderung des Einzelfahrzeuges, als auch Ortsveränderung von mehreren Fahrzeugen berücksichtigt. Ein Einzelfahrzeug wird durch die Beziehung von Weg, Geschwindigkeit und Beschleunigung in Abhängigkeit von der Zeit beschrieben. [Sch11, S. 25 ff.] Diese Daten werden nachfolgend „Bewegungsdaten“ genannt und sind für die vorliegende Arbeit besonders relevant, da diese Teil der Trajektoriendaten sind. Trajektorien werden im Kapitel 2.1.2 definiert. Grundlegende Kenngrößen für die Bewegung von mehreren Fahrzeugen sind die Verkehrsstärke, die Verkehrsdichte sowie die zeitlichen und räumlichen Abstände zwischen Fahrzeugen [Sch11, S. 41]. Bei der Auswahl der in dieser Arbeit zu untersuchenden Daten, wird unter anderem die Verkehrsstärke als eine Einflussgröße berücksichtigt, weshalb diese nachfolgend definiert wird. „Die Verkehrsstärke q ist der Quotient aus der Anzahl der Fahrzeuge N (Verkehrselemente) und der Zeitspanne T, während der die Fahrzeuge den Beobachtungsquerschnitt durchfahren“ [Sch11, S. 41]. Die Verkehrsstärke wird in Fahrzeugeinheiten pro Zeiteinheit angegeben und ist nach [Sch11] definiert als

$$q = \frac{N}{T}. \tag{2.1}$$

Die genannten Kenngrößen zur Quantifizierung des Verkehrs liefern eine Datengrundlage, auf welche die Anomalieerkennung in dieser Arbeit angewandt wird. Wie in Abschnitt 1.3 erläutert, werden in dieser Arbeit Trajektorien analysiert, weshalb der Begriff der Trajektorie nachfolgend definiert wird.

2.1.2 Definition Trajektorie

Eine Trajektorie repräsentiert die kontinuierliche Historie eines bewegten Objektes. Aufgrund technischer Einschränkungen der Datenaufzeichnung besteht eine Trajektorie *TJ* aus *n* zeitlich geordneten *m*-dimensionalen Datenpunkten *DP* zu diskreten Zeitpunkten. [Men19, S. 7] Die Menge der Trajektorien *TJS* beschreibt alle zu untersuchenden Trajektorien und wird definiert als

$$TJS = \{TJ_1, TJ_2, ..., TJ_N\}, \tag{2.2}$$

wobei *N* deren Anzahl angibt. Die Trajektorien bestehen wiederum aus *n* Datenpunkten, sodass diese definiert werden als

$$TJ = \{DP_1, DP_2, ..., DP_n\}. \tag{2.3}$$

In dieser Arbeit gibt *n* die Länge einer Trajektorie an. Im vorliegenden Datensatz variiert diese zwischen 1 und 2.475. Jeder Datenpunkt *DP* einer Trajektorie besteht aus *m* Merkmalen *M* und wird definiert als

$$DP = \{M_1, M_2, ..., M_m\}. \tag{2.4}$$

In Tabelle 2.1 werden alle für diese Arbeit relevanten Merkmale einer Trajektorie dargestellt. Neben den Bewegungsdaten werden im vorliegenden Fall eine eindeutige Identifikationsnummer (ID), die Art des Objektes und die Bewegungsrichtung für jeden Datenpunkt als weiteres Merkmal gespeichert.

Tabelle 2.1: Beispielhafter Datenpunkt der analysierten Daten

Merkmal	**Wert**
Identifikationsnummer	123456
Objektklasse	c
Digitaler Zeitstempel in Sekunden	1558389628,524401
X-Position der UTM Koordinaten in Metern	604753,66
Y-Position der UTM Koordinaten in Metern	5792813,37
Geschwindigkeit in X-Richtung in m/s	1,33
Geschwindigkeit in Y-Richtung in m/s	-4.84
Geschwindigkeit in Bewegungsrichtung in m/s	5,02
Beschleunigung in m/s^2	0,01
Bewegungsrichtung in Grad	285,50

Die Trajektoriendaten enthalten neben den in Abbildung 2.1 dargestellten Daten noch weitere Merkmale. Abbildung 2.1 zeigt 3.000 zufällig ausgewählte Trajektorien *TJS* vom zu untersuchenden Datenbestand. Dabei bestimmen die Merkmale X- und Y-Position den Verlauf und da Merkmal Objektklasse die Farbe der Trajektorie. Weiter Informationen über den Datenbestand, können dem Kapitel 3.1.1 entnommen werden.

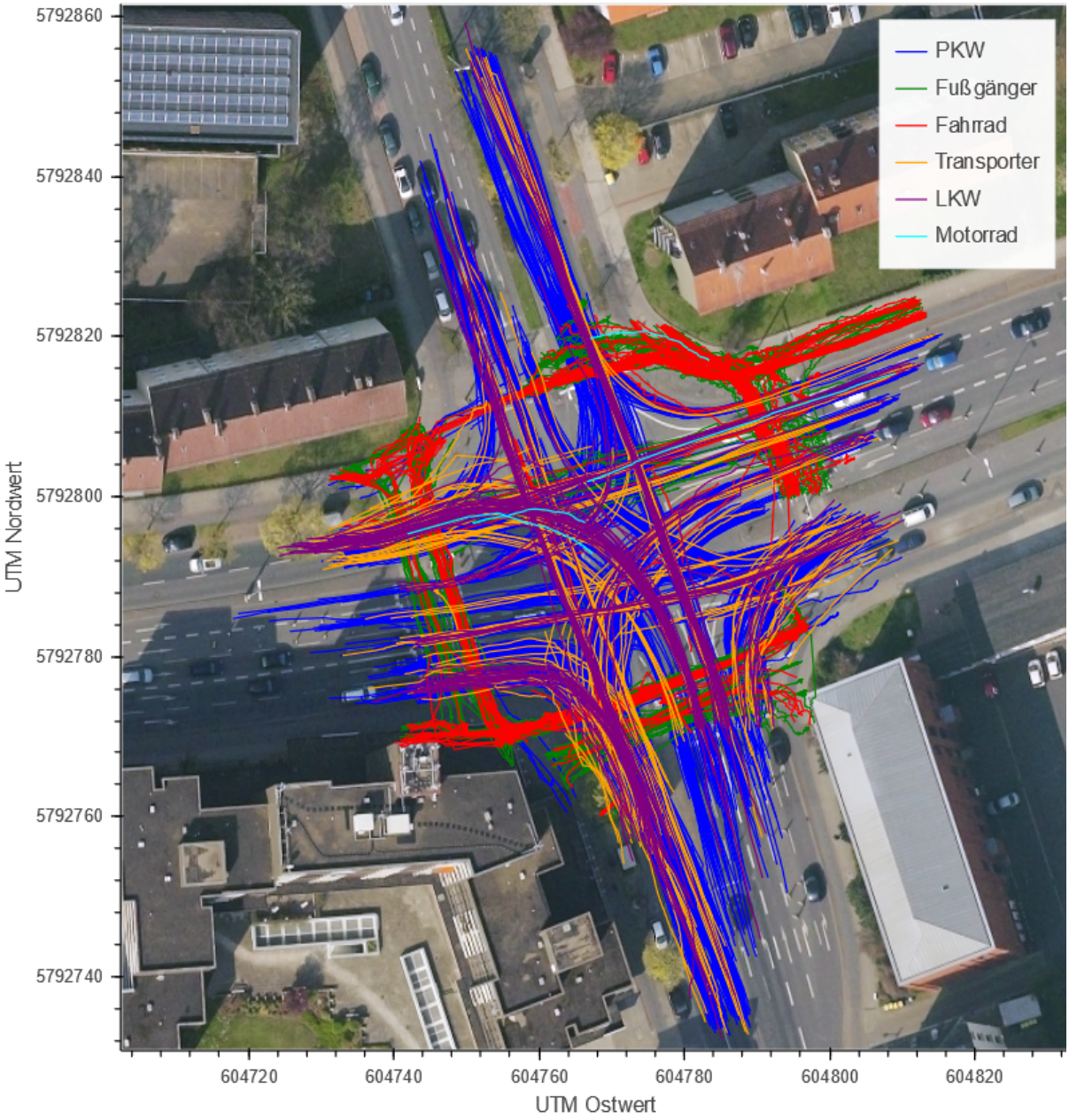


Abbildung 2.1: Trajektorien erfasster Verkehrsteilnehmer auf der Fokr. Das Luftbild im Hintergrund wird in nachfolgenden Abbildungen immer wieder verwendet und hier einmalig zitiert [Deub].

Um ein grundlegendes Verständnis dafür zu schaffen, wie die Daten aufgezeichnet werden, wird nachfolgend der technische Aufbau der Forschungskreuzung vorgestellt.

2.2 Messeinrichtungen zur Datenerhebung

In der vorliegenden Arbeit werden Daten untersucht, die an der Fokr, einer urbanen Kreuzung der Straßen Rebenring/Hagenring/Hans-Sommer-Straße/Brucknerstraße in Braunschweig aufgezeichnet werden. An dieser Kreuzung wird der Verkehr über LSA gesteuert. Neben den Geh- und Radwegen existieren 18 Fahrstreifen, auf denen die Verkehrsteilneh-

mer in die Kreuzung einfahren können und acht Fahrstreifen, auf denen die Kreuzung verlassen werden kann.

Zur Erfassung der Verkehrsdaten und 3D-Merkmalsberechnung sind derzeit drei unterschiedliche Sensor-Systeme installiert. Dies sind Mono-Video-Radar-Sensoren, Mono-Video-Sensoren und Stereo-Video-Sensoren. Ein Mono-Video-Radar-Sensor besteht aus einem Mono-Video- und einem Radar-Sensor, die parallel ausgerichtete Sichtachsen aufweisen. Mit einer Reichweite von 80 m, wird dieses System für die Erfassung des Innenbereichs der Kreuzung verwendet (siehe Abbildung 1 im Anhang). Wie in Abbildung 2 im Anhang dargestellt, erfolgt die Detektion von Verkehrsteilnehmern beim Rechtsabbiegen mit den Mono-Video-Sensoren, die eine Reichweite von 50 m aufweisen. Um die Verkehrsobjekte ununterbrochen zu verfolgen, wurden die Sensoren so ausgerichtet, dass sich die Sichtfelder überschneiden. Das dritte Sensor-System, die Stereo-Video-Sensoren, wird zur Beobachtung der Fußgängerfurt im Westen und Süden verwendet und erreicht eine maximale Detektionsreichweite von 60 m (siehe Abbildung 3 im Anhang). Alle Sensor-Systeme wurden in einer Höhe von 4,8 m an Masten montiert, die sich im Innenbereich der Kreuzung befinden. Zur Verbesserung der Sicht wurden alle Sensor-Systeme mit einem LED-Strahler ausgestattet, der bei jeder Aufnahme eines Videobildes einen Lichtimpuls im Infrarot-Spektrum aussendet. [Arn18, S. 93 ff.]

Dieses System wurde noch um zwei Stereo-Video-Sensoren (siehe Abbildung 4 im Anhang) erweitert, um eine Konfliktzone im nordöstlichen Bereich der Kreuzung mit einer durchgängig hohen Messgenauigkeit zu überwachen. Die Konfliktzone entsteht dadurch, dass die Grünphase der motorisierten Verkehrsteilnehmer von Osten nach Norden gleichzeitig mit der Grünphase der Fußgängern und Radfahrer von Osten nach Westen stattfindet. Ebenso ermöglichen diese, neben der Straße platzierten Messsysteme, eine dreidimensionale Vermessung der Fahrzeuge deutlich vor deren Eintritt in den inneren Kreuzungsbereich. [Arn18, S. 102 f.]

Die Rekonstruktion der 3D-Körper der Fahrzeuge und deren Trajektorien erfolgt mit Hilfe einer rekursiven Partikel-Akkumulation. Bei diesem Verfahren wird für jedes Videobild eine Partikel-Liste mit allen detektierten 3D-Merkmalen erzeugt. Zur Detektion der Verkehrsteilnehmer an jeder Position der Kreuzung werden die Partikel-Listen aller Sensor-Systeme fusioniert. Die Ansammlung vieler Partikel an einer Position führt dazu, dass an dieser Position ein Objekt vermutet wird und die Höhe, Länge und Breite des Objektes geschätzt werden kann. Anhand dieser Größe wird das Objekt in eine der Objektklassen Fußgänger, Fahrradfahrer, Motorradfahrer, PKW, Transporter oder LKW eingeteilt. Die Klassen Fußgänger und Fahrradfahrer werden auch als ungeschützte Verkehrsteilnehmer und die übrigen Klassen als motorisierte Verkehrsteilnehmer (MRU, englisch „motorized road users“) bezeichnet. Neben dieser Klassifizierung des Objektes, muss noch dessen Position bestimmt werden. Die Bestimmung der Position eines Verkehrsteilnehmers wird durch eine Georeferenzierung der jeweiligen Video-Sensoren ermöglicht, da bei diesem Vorgang die Kalibrierung der 3D-Position und 3D-Orientierung jedes Video-Sensors im Weltkoordinatensystem festgelegt werden. Mittels dieser Georeferenzierung können aus der Position im Bild die UTM-Koordinaten ermittelt werden. [Arn18, S. 50 ff.]

Die Geschwindigkeit wird mit Hilfe der Radar-Sensoren und des optischen Flusses aus der Ortsveränderung der detektierten Positionen eines Objektes ermittelt. Der optische Fluss stellt die Bewegungsinformation eines Objektes dar, die aus aufeinanderfolgenden Bildern eines Videos auf Pixel-Ebene erkannt werden. Damit ist eine Verfolgung von Objekten sogar möglich, wenn diese nur teilweise sichtbar sind. Allerdings nimmt die Verlässlichkeit der Positionsangaben mit abnehmender Geschwindigkeit ab. [Arn18, S. 28] Die Verfolgung der

Bewegungsmuster auf Pixel-Ebene ermöglicht ebenso die Messung der Bewegungsrichtung eines Objektes. Die Daten werden aufgrund der Bildfrequenz der Sensor-Systeme von 25 Bildern pro Sekunde (englisch:„frames per seconds“, kurz fps) alle 40 ms erhoben.

Die Beschleunigungswerte lassen sich nicht auf Basis der Videos bestimmen, weshalb diese aus den Geschwindigkeiten und Positionen mittels eines erweiterten Kalman-Filters zweiter Ordnung geschätzt werden. Detailliertere Informationen und Definitionen der Fachbegriffe zum technischen Aufbau der Fokr können dem Systemhandbuch [Arn18] entnommen werden. Der gesamte technische Aufbau wurde so konzipiert, dass keine Datenschutzgesetze verletzt werden.

In Abbildung 2.2 ist ein Videobild vom 22. Januar 2020 aus der Perspektive von vier Kameras, die auf den Innenbereich der Kreuzung ausgerichtet sind, dargestellt. Aus diesen Bildern werden die zu analysierenden Daten erhoben.

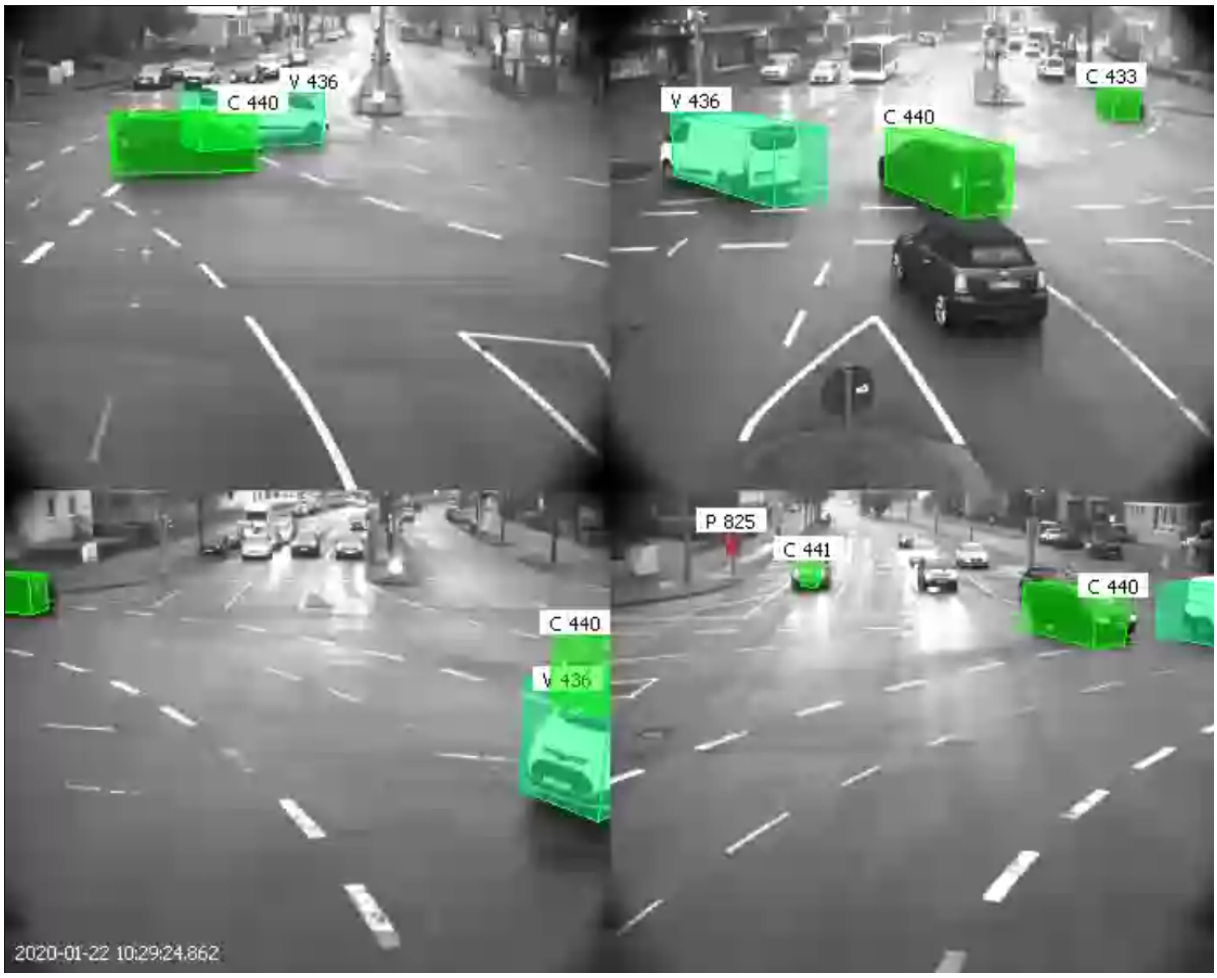


Abbildung 2.2: Bildschirmaufnahme der Videoübertragung von vier Kameras, die auf den Innenbereich der Kreuzung ausgerichtet sind. Richtung der Sichtachsen: oben links: Süden, oben rechts: Westen, unten links: Norden, unten rechts: Osten. [Deua]

Aus der Abbildung 2.2 geht hervor, dass Fahrzeuge erst spät oder gar nicht detektiert werden. Für die Schaffung eines Verständnisses, wie solche Anomalien erkannt werden können, werden nachfolgend theoretische Grundlagen der Anomalieerkennung vorgestellt.

2.3 Anomalieerkennung

In diesem Kapitel wird zunächst der Begriff Anomalie definiert. Da bei jeder Methode der Anomalieerkennung, eine Anomalie von einem Schwellenwert abhängig ist, werden anschließend mögliche Methoden vorgestellt, nach denen der Schwellenwert ermittelt werden kann. Ein solcher Schwellenwert ermöglicht die Detektion von verschiedenen Anomalien. Um den Umfang aller potenziell möglichen Anomalien zu berücksichtigen, werden daher Anomalien anschließend nach deren Ursachen unterschieden. Je nach der zu detektierenden Anomalie eignen sich verschiedene Verfahren. Daher werden zum Abschluss des Abschnitts die anzuwendenden Verfahren vorgestellt.

2.3.1 Definition Anomalie

Anomalieerkennung, auch Ausreißererkennung genannt, beschreibt ein Vorgehen, bei dem versucht wird, aus Daten sogenannte Anomalien oder Ausreißer zu identifizieren. Eine Anomalie ist eine Beobachtung, die sich von den anderen Beobachtungen so deutlich unterscheidet, dass der Verdacht entsteht, die Beobachtung sei von einem anderen Mechanismus generiert worden [Haw80, S.1]. Anomalieerkennung findet in vielen Bereichen wie zum Beispiel der Text- und Bildverarbeitung, Betrugserkennung oder der Medizin Anwendung. Um diesem Umfang von Problemen gerecht zu werden, bestehen eine Vielzahl an Verfahren, die zur Anomalieerkennung verwendet werden. [Cha09, S. 15:5]

Bei Verfahren zur Anomalieerkennung wird wie folgt in drei Schritten vorgegangen: Zu Beginn wird der zu untersuchende Datensatz von fehlerhaften Daten bereinigt. Im zweiten Schritt wird ein Modell aus den Daten aufgestellt, um die zu erwartende Verteilung der Daten zu bestimmen und davon anomale Daten unterscheiden zu können. Der dritte Schritt beinhaltet die Anomalieerkennung, bei der neue Daten mit dem aufgestellten Modell verglichen werden. Sollten sich die neuen Werte stärker als um einen vordefinierten Schwellenwert θ von der bekannten Verteilung unterscheiden, liegt eine Anomalie mit dem Anomalie-Wert A vor. [Agr15, S. 709]

Dieser Anomalie-Wert wird in Bezug auf Trajektorien je nach Verfahren entweder für Subtrajektorien, also Teile einer Trajektorie oder die gesamte Trajektorie berechnet. [Men19] Daher muss festgelegt werden, ab wie vielen abnormalen Werten eine Trajektorie als Anomalie zu klassifizieren ist. In der vorliegenden Arbeit wird jede Trajektorie, die mindestens einen Datenpunkt enthält, der von der Normalität abweicht, als Anomalie klassifiziert. Die Klassifizierung durch einen Anomalie-Wert, kann entweder binär erfolgen oder den Grad der Anomalie berücksichtigen. Beide Verfahren werden in der vorliegenden Arbeit angewandt. Bei einer binären Klassifikation von Datenpunkten, muss festgelegt werden, ob der Schwellenwert den minimalen oder maximalen Wert darstellt. Je nachdem ist der Anomalie-Wert Eins, wenn der Datenpunkt x den Schwellenwert θ über- oder unterschreitet und Null, wenn nicht. Für den Fall, dass der Schwellenwert den maximalen Wert darstellt, gilt

$$A = \begin{cases} 1 & \text{wenn } x > \theta \\ 0 & \text{sonst.} \end{cases} \tag{2.5}$$

Entsprechend gilt für den Fall, dass der Schwellenwert den minimalen Wert darstellt

$$A = \begin{cases} 1 & \text{wenn } x < \theta \\ 0 & \text{sonst.} \end{cases} \tag{2.6}$$

Wenn Anomalien nach deren Grad unterschieden werden sollen, wird ein Anomalie-Wert verwendet, mit dem das Verhältnis vom gemessenen Wert zum Schwellenwert angegeben wird [Bre00]. Dafür ändert sich die Berechnung des Anomalie-Wertes wie folgt.

$$A = \frac{x}{\theta} \tag{2.7}$$

In jedem Fall wird für die Anomalieerkennung ein Schwellenwert benötigt. Wie dieser bestimmt werden kann, wird in dem folgenden Abschnitt vorgestellt.

2.3.2 Bestimmung eines Schwellenwertes

Die Festlegung eines Schwellenwertes geschieht durch den Anwender der Anomalieerkennung, da dieser angeben muss, ab welchem Schwellenwert ein Wert eine Anomalie darstellt. Sollte dieser kein fachspezifisches Wissen aufweisen, nachdem der Schwellenwert ermittelt werden kann, können verschiedene Statistiken verwendet werden, um einen geeigneten Schwellenwert zu bestimmen. Die Statistiken basieren darauf, den Schwellenwert aus den vorliegenden Daten zu berechnen. [Yan19] Welche Statistik in dieser Arbeit verwendet wird, wenn der Schwellenwert berechnet werden muss, wird nachfolgend vorgestellt.

In einer Stichprobe der im Zeitraum von 2016 bis 2018 bei Google Scholar erfassten Studien zur Anomalieerkennung wird zur Bestimmung eines Schwellenwertes am häufigsten der Z-Wert verwendet. Deshalb dieser auch in der vorliegenden Arbeit angewandt. Weniger verwendete Statistiken sind die mittlere absolute Abweichung vom Median oder der Interquartilsabstand [Yan19]. Der Z-Wert [Fah16, S. 296] gibt an, wie viele Standardabweichungen s ein Wert x vom arithmetischen Mittel $\bar{x}$ entfernt ist und ist daher definiert als

$$z = \frac{x - \bar{x}}{s}. \tag{2.8}$$

Dabei ist das arithmetischen Mittel $\bar{x}$ [Fah16, S. 53] definiert als

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.9}$$

und die Standardabweichung s ergibt sich aus der Wurzel der Varianz s^2 [Fah16, S. 70], welche definiert ist als

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{2.10}$$

Durch Verwendung des Z-Wertes wird nicht der Wert selbst sondern die Repräsentation des Wertes als standardisierter Z-Wert für die Anomalieerkennung mit einem Schwellenwert verglichen. In den meisten Studien wird ein Schwellenwert θ von drei verwendet, wodurch bei einer Normalverteilung die beiden äußeren 0,13 % der Werte als Anomalien klassifiziert werden. Damit werden insgesamt 0,26 % der Daten eines normalverteilten Datensatzes als Anomalie bewertet. [Yan19] Die binäre Klassifikation aus der Formel 2.5 verändert sich

unter Verwendung des Z-Wertes wie folgt.

$$A = \begin{cases} 1 & \text{wenn } |z| > \theta \\ 0 & \text{sonst} \end{cases} \tag{2.11}$$

Sollte die den zu analysierenden Daten zugrundeliegende Verteilung sich deutlich von einer Normalverteilung unterscheiden, führt der Z-Wert zu keinem sinnvollen Ergebnis. In diesem Fall wird der oben genannte Anteil von 0,26 % der Werte als Richtwert für die Anzahl der Anomalien im Datensatz verwendet und auf Basis dieses Prozentsatzes der Schwellenwert bestimmt.

Die Anwendung des Schwellenwertes ermöglicht es, Anomalien im Datensatz zu identifizieren. Nachfolgend wird vorgestellt, wie die detektierten Anomalien unterschieden werden können.

2.3.3 Unterscheidung von Anomalien

Um eine Aussage darüber treffen zu können, welche Anomalien mit den in dieser Arbeit angewandten Verfahren detektiert werden und welche Anomalien in den Daten weiterhin vorhanden sein können, werden die Anomalien nachfolgend nach deren Ursache und Art unterschieden.

Ursache von Anomalien

Anomalien werden einer von zwei möglichen Ursachen zuordnet. Zum einen können Anomalien durch verfahrenstechnische Fehler entstehen. Dies sind nicht der Realität entsprechende Werte, die durch einen Fehler bei der Aufzeichnung entstehen. Diese werden im Folgenden als *fehlerhafte Daten* bezeichnet. Für die zu untersuchenden Daten sind diverse Umstände denkbar, die zu fehlerhaften Daten führen können. Ein Beispiel hierfür ist die Bestimmung der Position eines Verkehrsteilnehmers, die mehr als bisher bekannt von der wahren Position abweicht. Die zweite Kategorie der Ursachen einer Anomalie umfasst alle Werte, die das Geschehen realitätsgetreu abbilden, aber in einem Bereich auftreten, in dem die Daten bisher nicht aufgetreten sind. Diese Anomalien werden im Folgenden als *außergewöhnliche Daten* bezeichnet. [Mir10, S. 6] Eine Anomalie dieser zweiten Kategorie ist zum Beispiel ein Fahrzeug, das die Kreuzung außergewöhnlich schnell überquert und damit außergewöhnlich hohe Werte in der Geschwindigkeit aufweist.

Um einer Anomalie die korrekte Ursache zuordnen zu können, muss der reale Sachverhalt bekannt sein, von dem die Daten aufgezeichnet werden. Im vorliegenden Fall können die Videoaufnahmen mit den aufgezeichneten Daten verglichen werden. Damit kann überprüft werden, ob die Positionsbestimmung fehlerhaft erfolgt ist oder die Realität wiedergibt. Sollte letztes der Fall sein, ist die Anomalie aufgrund außergewöhnlicher Daten entstanden. Da den Videoaufzeichnungen nicht die tatsächliche Geschwindigkeit und Beschleunigung entnommen werden kann, werden diese Daten in dieser Arbeit nicht auf außergewöhnliche Daten überprüft.

Arten von Anomalien

Speziell für Anomalien in Trajektorien existiert eine weitere Unterscheidung nach der Art der Anomalie. Die in dieser Arbeit detektierten Anomalien können demnach auch in distanz-,

dichte- und merkmalsbasierte Anomalien unterschieden werden. *Distanzbasierte* Anomalien bezeichnen Trajektorien, deren Distanz zu den meisten Objekten in einem Datensatz einen vordefinierten Schwellenwert einer Distanzmetrik überschreiten. *Dichtebasierte* Anomalien entstehen, wenn eine Trajektorie sich an einer Position befindet, an der die Anzahl der auftretenden Trajektorien einen vordefinierten Schwellenwert unterschreitet. Bei *merkmalsbasierten* Anomalien weicht nur ein bestimmtes Merkmal wie zum Beispiel die Geschwindigkeit oder Bewegungsrichtung stark von einer erwarteten Verteilung ab. [Men19]

Welche dieser Anomalien mit welchem Verfahren detektiert werden, wird im Abschnitt 2.5 vorgestellt. Bevor auf die Verfahren eingegangen wird, werden nachfolgend die Methoden zur Datenvorbereitung beschrieben.

2.4 Datenvorbereitung

Bei der Datenvorbereitung werden Anomalien, die eindeutig durch fehlerhafte Daten entstanden sind, aus dem Datensatz entfernt. In der vorliegenden Arbeit erfolgt die Bestimmung der fehlerhaften Daten durch Analyse der Länge, Aufzeichnungsfrequenz, Beschleunigung, Geschwindigkeit, Position und Route einer Trajektorie. Nachfolgend werden die dafür verwendeten Verfahren vorgestellt.

2.4.1 Trajektorien der Länge Eins

Zunächst wird überprüft, wie viele Datenpunkte für eine Trajektorie aufgezeichnet wurden. Wenn eine Trajektorie nur aus einem Datenpunkt besteht, wird diese aus dem Datensatz entfernt. Die Anomalieerkennung wird demnach auf die Länge aller im Datensatz enthaltenen Trajektorien mit einem vordefinierten Schwellenwert von Eins angewandt.

2.4.2 Aufzeichnungsfrequenz

Bei dieser Methode wird die Differenz der digitalen Zeitstempel zweier aufeinanderfolgender Datenpunkte einer Trajektorie untersucht. Sollte diese um mehr als einen Schwellenwert abweichen, wird die Aufzeichnungsfrequenz als anomal klassifiziert. Eine Trajektorie, die eine anomale Aufzeichnungsfrequenz aufweist, wird aus dem Datensatz entfernt. Demnach wird die Anomalieerkennung auf die Differenzen der digitalen Zeitstempel angewandt. Der dafür zu verwendende Schwellenwert wird in Abschnitt 3.2.2 bestimmt.

Des Weiteren werden die Positions-, Geschwindigkeits- und Beschleunigungswerte auf die Einhaltung der physikalischen Gesetze überprüft, da eine Abweichung von den physikalischen Gesetzen eindeutig auf fehlerhafte Daten schließen lässt.

2.4.3 Beschleunigung

Als erstes wird untersucht, ob in den Beschleunigungswerten die minimal und maximal mögliche Beschleunigung nach den physikalischen Gesetzen unter- bzw. überschritten wird. Die dafür benötigten Schwellenwerte werden vom Anwender festgelegt und ersetzen die Verwendung einer Statistik. Da die minimalen und maximalen Werte je nach Objektklasse unterschiedlich sind, werden in der vorliegenden Arbeit an diesem Schritt der Datenvorbereitung nur die Trajektorien von einer Objektklasse, der PKWs, untersucht. Autotests ergeben für trockene Fahrbahn eine maximale Bremsbeschleunigung a_{min} für einen beladenen PKW von 11,9 m/s^2 [Bur09, S. 385 ff.]. Da an den Tagen, an denen der Datensatz

aufgezeichnet wurde, kein Niederschlag, außer an einem Tag 0,3 l/m^2 fiel [Wet19], wird dieser Wert auf den gesamten Datensatz angewandt. Gleiches gilt für die maximale Beschleunigung a_{max}, welche beim Anfahren im Geschwindigkeitsbereich von 0 bis 60 km/h bei 8,3 m/s^2 liegt [Bur09, S. 379 ff.]. Damit ergibt sich für die Anomalieerkennung in den Beschleunigungswerten a folgende Formel.

$$A = \begin{cases} 1 & \text{wenn } a < -11{,}9 \text{ m/s}^2 \\ 1 & \text{wenn } a > 8{,}3 \text{ m/s}^2 \\ 0 & \text{sonst} \end{cases} \tag{2.12}$$

2.4.4 Geschwindigkeit

Um Anomalien, die auf fehlerhaften Daten beruhen, in den Geschwindigkeitswerten zu identifizieren, werden jeweils zwei aufeinanderfolgende Geschwindigkeitswerte einer Trajektorie betrachtet. Die an einem bestimmten Zeitpunkt ermittelte Geschwindigkeit gibt an, mit welcher Geschwindigkeit ein Objekt bis zur nächsten Messung unterwegs sein wird. Das bedeutet, dass zu einem Zeitpunkt i der nächste Geschwindigkeitswert v_{i+1} aus dem jetzigen Geschwindigkeitswert v_i, dem Beschleunigungswert a_i sowie der Zeitdifferenz Δt_i als Wert $\hat{v}_{i+1}$ geschätzt werden kann (in dieser Arbeit werden die Werte $\hat{v}$, $\hat{x}$ und $\hat{y}$ als *berechnete* Werte und die Werte v, x und y als *gemessene* Werten bezeichnet). Aus diesem Grund müssen alle Geschwindigkeitswerte aus physikalischer Sicht folgende Formel erfüllen.

$$\hat{v}_{i+1} \approx v_i + a_i \Delta t_i \tag{2.13}$$

Dabei ist Δt_i die Differenz von zwei aufeinanderfolgenden digitalen Zeitstempeln

$$\Delta t_i = t_{i+1} - t_i \tag{2.14}$$

und sollte aufgrund der konstanten Aufzeichnungsfrequenz von 25 Hz immer 40 ms betragen. Wie die Einhaltung dieser konstante Aufzeichnungsfrequenz sichergestellt wird, ist in Kapitel 3.2.2 beschrieben.

Zur Anomalieerkennung in den Geschwindigkeitswerten wird die Differenz zwischen den gemessenen und errechneten Werten mit einem vordefinierten Schwellenwert verglichen. Überschreitet der Betrag der Differenz den Schwellenwert, liegt eine Anomalie vor.

$$A = \begin{cases} 1 & \text{wenn } |\hat{v}_{i+1} - v_{i+1}| > \theta \\ 0 & \text{sonst} \end{cases} \tag{2.15}$$

2.4.5 Position

Das Verfahren, dass zur Überprüfung der Geschwindigkeitswerte angewandt wird, kann für die Positionswerte adaptiert werden. Demnach gilt für die berechneten Werte der X-Position x

$$\hat{x}_{i+1} \approx x_i + v_{x_i} \Delta t_i \tag{2.16}$$

und der Y-Position y

$$\hat{y}_{i+1} \approx y_i + v_{y_i} \Delta t_i. \tag{2.17}$$

Dabei beschreibt v_x die Geschwindigkeit in die X-Richtung und v_y die Geschwindigkeit in die Y-Richtung. Die Formeln 2.16 und 2.17 gelten nur unter der Annahme, dass die Geschwindigkeit im Zeitintervall Δt_i konstant ist und damit die Beschleunigung vernachlässigt werden kann. Aufgrund der Kürze des Zeitintervalls Δt_i von 40 ms, wird die Beschleunigung als konstant angenommen. Die Anomalieerkennung erfolgt nach dem gleichen Verfahren, dass für die Geschwindigkeitswerte verwendet wird. Dabei bildet die Differenz der berechneten und gemessenen Positionswerte die Datengrundlage für die Anomalieerkennung. Die Berechnung der Anomalie-Werte erfolgt analog zur Formel 2.15 für die X-Positionen mit

$$A = \begin{cases} 1 & \text{wenn } |\hat{x}_{i+1} - x_{i+1}| > \theta \\ 0 & \text{sonst} \end{cases} \tag{2.18}$$

und für die Y-Positionen mit

$$A = \begin{cases} 1 & \text{wenn } |\hat{y}_{i+1} - y_{i+1}| > \theta \\ 0 & \text{sonst.} \end{cases} \tag{2.19}$$

Zur Vorbereitung der Daten zählt auch der Schritt, bei dem jeder Trajektorie einer Route zugeordnet wird. Dies ist notwendig, da die Anwendung der Verfahren zur Anomalieerkennung von außergewöhnlichen Daten Trajektorien einer Route betrachtet. Alle Trajektorien, die keiner Route zugeordnet werden, werden aus dem Datensatz entfernt. Hierbei ist zu beachten, dass die aus dem Datensatz entfernten Trajektorien nicht zwingend auf Basis fehlerhafter Daten entstanden sein müssen. Ebenso kann eine außergewöhnlich Handlung eines Verkehrsteilnehmers dazu geführt haben, dass die Trajektorie keiner Route zugeordnet werden kann.

2.4.6 Klassifizierung der Trajektorien nach Routen

Für die Zuordnung einer Trajektorie zu einer Route wird untersucht, über welche der vier Einfahrten ein Verkehrsteilnehmer in den Innenbereich der Kreuzung eintritt und über welche der vier Ausfahrten er diesen verlässt. Auf Grundlage dieser Untersuchung wird jeder Trajektorie eine von 16 Routen zugeordnet. Wie dies umgesetzt wird, kann dem Abschnitt 3.2.7 entnommen werden. Es wird erwartet, dass ein Verkehrsteilnehmer in den Innenbereich der Kreuzung über eine Einfahrt eintritt und diesen über eine Ausfahrt verlässt. Folgende Fälle werden als Anomalie klassifiziert:

1. Eine Trajektorie schneidet keine Einfahrt
2. Eine Trajektorie schneidet keine Ausfahrt
3. Eine Trajektorie schneidet mehr als eine Einfahrt
4. Eine Trajektorie schneidet mehr als eine Ausfahrt
5. Eine Trajektorie schneidet eine Ein- und keine Ausfahrt
6. Eine Trajektorie schneidet eine Ein- und mehrere Ausfahrten
7. Eine Trajektorie schneidet keine Ein- und eine Ausfahrt
8. Eine Trajektorie schneidet mehrere Ein- und eine Ausfahrt

9. Eine Trajektorie schneidet keine Ein- und keine Ausfahrt

Um diese Anomalien zu erkennen, werden für jede Trajektorie acht boolesche Variablen definiert, die angeben, ob eine Trajektorie eine Ein- bzw. Ausfahrt schneidet. Die jeweils vier Variablen für die Ein- bzw. Ausfahrten, werden durch Summieren zu jeweils einer Variable aggregiert. Die resultierenden zwei Variablen geben an, wie viele Ein- bzw. Ausfahrten eine Trajektorie schneidet. Auf Basis dieser Variablen kann die Anomalieerkennung durchgeführt werden.

Die Ursache aller mit diesem Verfahren detektierten Anomalien kann nicht eindeutig einer der Klassen von fehlerhaften oder außergewöhnlichen Daten zugeordnet werden, ohne das Videomaterial gesichtet zu haben. Dies gilt auch für die Verfahren der Anomalieerkennung, die im folgenden Abschnitt vorgestellt werden.

2.5 Auswahl und Vorstellung der anzuwendenden Verfahren

Im Vergleich zu den bisher vorgestellten Methoden zur Erkennung fehlerhafter Daten, wird in diesem Abschnitt auf die Auswahl der Verfahren zur Erkennung außergewöhnlicher Daten eingegangen. Anschließend werden die nötigen theoretischen Grundlagen für diese Verfahren erläutert.

Die Verfahren zur Anomalieerkennung werden in klassifikations-, distanz-, statistik-, dichte- und clusterbasiert unterschieden [Men19]. Die klassifikationsbasierten Verfahren, werden dem überwachten Lernen, einem Teilgebiet des maschinellen Lernens, zugeordnet, da hierfür Daten vorliegen müssen, für die bereits bekannt ist, welche Trajektorien eine Anomalie darstellen [Agr15]. Da diese Informationen im vorliegenden Fall nicht bestehen, werden klassifikationsbasierte Verfahren nicht angewandt. Alle weiteren Verfahren können auf die vorliegenden Daten angewandt werden, auch wenn keine Informationen über den Grad der Anomalie vorliegen. Daher wird in dieser Arbeit für jede dieser Verfahrensgruppen ein Verfahren ausgewählt und auf die Daten angewandt. Die Qualität der Anomalieerkennung kann bei solchen Verfahren nicht, wie bei Verfahren des überwachten maschinellen Lernens möglich, anhand der Ergebnisse quantitativ beurteilt werden. Daher wird die Qualität der Modelle qualitativ durch Analyse der detektierten Anomalien bewertet. Die Auswahl der Verfahren, die zur Anomalieerkennung angewandt werden sollen, wird nachfolgend vorgestellt.

Bei distanzbasierten Verfahren wird davon ausgegangen, dass Objekte, die eine große Entfernung zu den meisten Objekten aufweisen, eine Anomalie darstellen. Entscheidend für ein distanzbasiertes Verfahren ist daher die Metrik, mit welcher der Abstand zwischen zwei Trajektorien gemessen wird. Häufig werden die euklidische Distanz, die Hausdorff-Metrik, die längste gemeinsame Teilsequenz oder die dynamische Zeitnormierung verwendet. Für die vorliegende Arbeit wird die Hausdorff-Distanz als Abstands-Metrik ausgewählt, da diese für unterschiedlich lange Trajektorien berechnet werden kann und deren Eigenschaft, einzelne weit entfernte Datenpunkte stark zu berücksichtigen, für die Erkennung von distanzbasierten Anomalien geeignet ist. [Men19]

Bei statistikbasierten Methoden wird ein Wahrscheinlichkeitsmodell erstellt, nachdem Daten, deren Wahrscheinlichkeit nach dem Modell einen Schwellenwert unterschreitet, als Anomalie klassifiziert werden. Als statistikbasiertes Verfahren für die Anomalieerkennung in Trajektorien wird die Markow-Kette ausgewählt, da mit dieser die zeitliche Abfolge der

Bewegungsdaten einer Trajektorie modelliert werden kann [Gup13]. Da die Erkennung von Anomalien dabei auf der Bewegungsrichtung eines Verkehrsteilnehmers basiert, werden die detektierten Anomalien als merkmalsbasierte, speziell richtungsbasierte Anomalien eingeteilt. Allerdings führt die Anwendung dieses Verfahrens zu keinem zufriedenstellenden Ergebnis. Daher wird ein weiteres statistikbasiertes Verfahren angewandt, um richtungsbasierte Anomalien zu detektieren. Bei diesem Verfahren wird der kontinuierliche Raum in ein diskretes Gitter aufgeteilt, sodass die Bewegungsrichtungswerte jeder Zelle des Gitters auf Anomalien überprüft werden können. [Ge10]

In diesem diskreten Raum kann ebenso die Anzahl der unterschiedlichen Trajektorien in jeder Zelle gemessen werden. Mit der Anzahl der unterschiedlichen Trajektorien kann die Dichte in jeder Zelle ermittelt, werden. Daher stellt dieses Verfahren ein dichtebasiertes Verfahren dar, mit dem dichtebasierte Anomalien detektiert werden. [Ge10]

Die zuletzt vorgestellte Gruppe von Verfahren stellen die clusterbasierten Verfahren dar. Bei diesen wird angenommen, dass Objekte, die keine Ähnlichkeit mit einem Cluster aufweisen, Anomalien darstellen [Men19]. In der vorliegenden Arbeit wird für die sogenannte Clusteranalyse ein Gaußsches Mischmodell mit dem Erwartungs-Maximierungs-Algorithmus erstellt. Dieses Verfahren wird ausgewählt, weil damit eine höhere Genauigkeit bei der Gruppierung der Daten erreicht werden kann als mit anderen clusterbasierten Verfahren wie zum Beispiel dem k-Means-Algorithmus [Agr15]. Da bei dem ausgewählten Verfahren die Trajektorien als dreidimensionaler Datenpunkt abstrahiert werden, können die detektierten Anomalien keiner Art von Anomalie zugewiesen werden. Die vorgestellten Arten gelten nur für anomale Trajektorien. Das Vorgehen zur Detektion von Anomalien mit diesem Verfahren wird für diese Arbeit angepasst, indem die Daten zunächst nur den Clustern zugeordnet werden. Da diese Cluster die Routen darstellen, kann die Zuordnung anschließend mit dem Ergebnis der Klassifizierung der Trajektorien nach Polygonen verglichen werden. Trajektorien, die mit beiden Verfahren nicht dem gleichen Cluster zugeordnet werden, werden als Anomalie klassifiziert.

In Kapitel 5 erfolgt ein Ausblick über die mögliche Anwendung weiterer Verfahren auf die Daten. Nachfolgend werden die theoretischen Grundlagen für die ausgewählten Verfahren zur Erkennung außergewöhnlicher Daten beschrieben.

2.5.1 Hausdorff-Metrik

Dieses Verfahren wird in der vorliegenden Arbeit nach der Hausdorff-Metrik benannt, da mittels dieser im Datensatz eine durchschnittliche Trajektorie ermittelt wird. Die durchschnittliche Trajektorie stellt die Grundlage für die Anomalieerkennung dar, da anschließend Datenpunkte, deren Entfernung zur durchschnittlichen Trajektorien ein Schwellenwert übersteigt, als Anomalie klassifiziert werden. Nachfolgend werden die für dieses Verfahren notwendigen Definitionen aufgestellt. Diese sind die Definition für

- den Abstand zweier Punkte,
- den Abstand zweier Trajektorien und
- die durchschnittliche Trajektorie.

Zur Berechnung der Distanzen zwischen Punkten wird ein Ähnlichkeits- bzw. Abstandsmaß benötigt. Das am häufigsten verwendete Abstandsmaß für mehrdimensionale Punkte ist die euklidische Distanz. Bei diesem Abstandsmaß wird die Distanz zwischen den jeweili-

gen Dimensionen der beiden Punkte paarweise verglichen. Allgemein ist die euklidische Distanz $d_e(\boldsymbol{p}, \boldsymbol{q})$ zwischen zwei n-dimensionalen Punkten $\boldsymbol{p}$ und $\boldsymbol{q}$ definiert als [Bra06, S.345]

$$d_e(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}. \tag{2.20}$$

Mit der euklidischen Distanz können die Abstände zwischen den jeweiligen Punkten einer Trajektorie berechnet werden. In der vorliegenden Arbeit müssen allerdings Abstände zwischen gesamten Trajektorien gemessen werden, die aus mehreren Punkten bestehen. Dafür wird nachfolgend die Hausdorff-Metrik (oder Hausdorff-Distanz) eingeführt, welche den maximalen Wert der zwei direkten Hausdorff-Distanzen von zwei Trajektorien beschreibt. Die direkten Hausdorff-Distanzen geben die maximale Distanz der minimalen Distanzen zwischen den Punkten von zwei Trajektorien an. Demnach gelten zwei Trajektorien TJ_i und TJ_j als nahe beisammen, wenn jeder Punkt jeder Trajektorie nahe einem Punkt der anderen Trajektorie ist. Dabei stellen i und j den Index der Trajektorie TJ im Datensatz TJS dar und können demnach Werte zwischen 1 und N annehmen. Wie in Abschnitt 2.3 definiert, ist N die Anzahl der in Datensatz TJS enthaltenden Trajektorien. Formal ist die Hausdorff-Distanz $d_H(TJ_i, TJ_j)$ definiert als

$$d_H(TJ_i, TJ_j) = \max(h(TJ_i, TJ_j), h(TJ_j, TJ_i)). \tag{2.21}$$

Die direkte Hausdorff-Distanz $h(TJ_i, TJ_j)$ ist definiert als

$$h(TJ_i, TJ_j) = \max_{a \in TJ_i}(\min_{b \in TJ_j}(d_e(a, b))), \tag{2.22}$$

wobei a und b die jeweiligen Datenpunkte einer Trajektorie darstellen. Die Hausdorff-Distanz ist extrem anfällig gegenüber Rauschen. Diese Eigenschaft kann für die Detektion von Anomalien genutzt werden, da diese einen großen Abstand von der durchschnittlichen Trajektorie aufweisen. [Men19]

Der Begriff einer „durchschnittlichen Trajektorie“ kann unterschiedlich interpretiert werden. In dieser Arbeit wird die durchschnittliche Trajektorie TJ_d definiert als Trajektorie, bei der die Summe der Distanzen zu den anderen Trajektorien minimal ist. [Buc13, S. 598] Dafür muss für jede Trajektorie der Abstand zu jeder anderen Trajektorie berechnet und die Summe dieser Distanzen gebildet werden. Für die Abstandsberechnung wird in der vorliegenden Arbeit die Hausdorff-Metrik verwendet. Im finalen Schritt wird der minimale Wert d_{min} der Summen ausgewählt. Die Trajektorie mit diesem Wert stellt die durchschnittliche Trajektorie dar.

$$d_{min} = \min_{i \in TJS} \sum_{j=1}^{N} d_H(TJ_i, TJ_j) \tag{2.23}$$

2.5.2 Markow-Kette

Für die Erkennung von richtungsbasierten Anomalien wird in der vorliegenden Arbeit eine Markow-Kette verwendet. Daher wird nachfolgend das Modell der Markow-Kette vorgestellt. Da die Anwendung dieses Modells einige Annahmen voraussetzt, werden diese anschließend überprüft.

In der vorliegenden Arbeit werden die Trajektoriendaten als zeitdiskrete Markow-Kette mit endlichem Zustandsraum beschrieben. Das Modell der Markow-Kette ist ein stochas-

tischer Prozess, welcher für die Modellierung und Analyse komplexer Systeme verwendet wird. Grundlegende Größen dieser Kette sind der Zustandsraum S mit N Zuständen s_n

$$S = \{s_1, s_2, ..., s_N\} \tag{2.24}$$

und die Übergangswahrscheinlichkeiten p_{ij}, welche die Wahrscheinlichkeit dafür angeben, vom Zustand s_i in den Zustand s_j überzugehen. Die Wahrscheinlichkeiten können in einer $N \times N$ Matrix $\mathbf{P}$ dargestellt werden.

$$\mathbf{P} = \{p_{ij}\} = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1N} \\ p_{21} & p_{22} & \dots & p_{2N} \\ \vdots & & & \\ p_{N1} & p_{N2} & \dots & p_{NN} \end{bmatrix} \tag{2.25}$$

Die Übergangswahrscheinlichkeiten p_{ij} werden mittels der Wahrscheinlichkeit P definiert als

$$p_{ij} = P(s_{n+1} = j \mid s_n = i). \tag{2.26}$$

Da bei jedem Übergang ein neuer Zustand erreicht wird, muss gelten

$$\sum_{j=1}^{N} p_{ij} = 1 \quad i = 1, 2, ..., N. \tag{2.27}$$

Die Definition der Übergangswahrscheinlichkeiten und die Definition des vorliegenden Prozesses als Markow-Kette gelten, wenn die Markow-Annahme erfüllt ist. Diese und weitere Annahmen werden nachfolgend begründet. Die Markow-Annahme sagt aus, dass nur der jeweils letzte Zustand den aktuellen Zustand beeinflusst. [How12, S. 1 ff.] Es wird vermutet, dass diese Annahme für Trajektoriendaten nicht gilt, da ein Zustand jeweils von mehreren vorherigen Zuständen abhängig sein kann. Um diese Annahme trotzdem als erfüllt betrachten zu können, werden die Verkehrsteilnehmer im vorliegenden Fall immer nur von einer zur nächsten Position untersucht. Somit können Anomalien in Subtrajektorien der Länge zwei detektiert werden.

Damit die Markow-Kette als zeitdiskret betrachtet werden kann, muss die Annahme erfüllt sein, dass aufgrund der technischen Ausstattung die Datenaufzeichnung immer mit der gleichen Frequenz stattfindet. Diese Annahme wird in Abschnitt 3.2.2 untersucht. Sollte eine Trajektorie mindestens einen Datenpunkt enthalten, der von der Frequenz abweicht, wird die Trajektorie aus dem Datensatz entfernt. Damit wird die zeitdiskrete Eigenschaft des zu erstellenden Modells sichergestellt.

Die dritte getroffene Annahme ist, dass der Zustandsraum endlich ist. Da die Messinfrastruktur Objekte nur einem bestimmten Bereich detektieren kann und dieser in ein diskretes Gitter aus endlichen Zellen eingeteilt wird, ist auch diese Annahme begründet.

2.5.3 Gaußsches Mischmodell

Bei diesem Verfahren des unüberwachten maschinellen Lernens wird der Erwartungs-Maximierungs-Algorithmus (EM-Algorithmus) mit dem Ziel angewandt, die wahrscheinlichste Repräsentation der Daten als Gaußsches Mischmodell (GMM) zu modellieren. Die „wahrscheinlichste Repräsentation“ beschreibt die Anpassung der Parameter eines Gaußschen Mischmodells, bis dieses die Daten so genau wie möglich repräsentiert. Nachfolgend wer-

den zunächst GMM und der EM-Algorithmus mathematisch definiert. Für eine detaillierte Herleitung der verwendeten Formeln und Zwischenschritte kann auf die Referenz [Bis06] zurückgegriffen werden. Abschließend wird erläutert, wie das GMM für die Anomalieerkennung verwendet werden kann. Dafür wird im vorliegenden Fall jede Trajektorie als dreidimensionaler Datenpunkt im euklidischen Raum abstrahiert und definiert als

$$\boldsymbol{TJ} = \{x_1, x_2, x_3\}. \tag{2.28}$$

Definition Gaußsches Mischmodell

Ein Gaußsches Mischmodell stellt eine lineare Kombination mehrerer Gaußscher Verteilungen, die auch Normalverteilungen genannt werden, dar (im Folgenden werden die einzelnen Normalverteilungen des GMM als Komponenten bezeichnet). Durch Anpassung der einzelnen Parameter der Komponenten ermöglicht das GMM die Repräsentation von nahezu jeder stetigen Wahrscheinlichkeitsverteilung. Grundlage des Modells ist eine multivariate Normalverteilung $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ eines D-dimensionalen Vektors $\mathbf{x}$ (Variablen, die „fett“ formatiert sind, beschreiben Vektoren), welche definiert ist als

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{2.29}$$

Dabei bezeichnen $\boldsymbol{\mu}$ den D-dimensionalen Vektor der Mittelwerte, $\boldsymbol{\Sigma}$ die $D \times D$ Kovarianz-Matrix und $|\boldsymbol{\Sigma}|$ deren Determinante. [Bis06, S. 78]

In einem GMM (siehe Abbildung 2.3) wird die Wahrscheinlichkeit $p(\mathbf{x})$ für die Werte der Variable $\mathbf{x}$ als Kombination von K Komponenten mit unterschiedlichem Mittelwert $\boldsymbol{\mu}_k$ und Kovarianz $\boldsymbol{\Sigma}_k$ angegeben als

$$p(\mathbf{x}) = \sum_{k=1}^{K} m_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2.30}$$

Dabei sind m_k die jeweiligen Mischkoeffizienten, die den Anteil einer Komponente am Mischmodell angeben und demnach in Summe Eins ergeben müssen [Bis06, S. 111]

$$\sum_{k=1}^{K} m_k = 1. \tag{2.31}$$

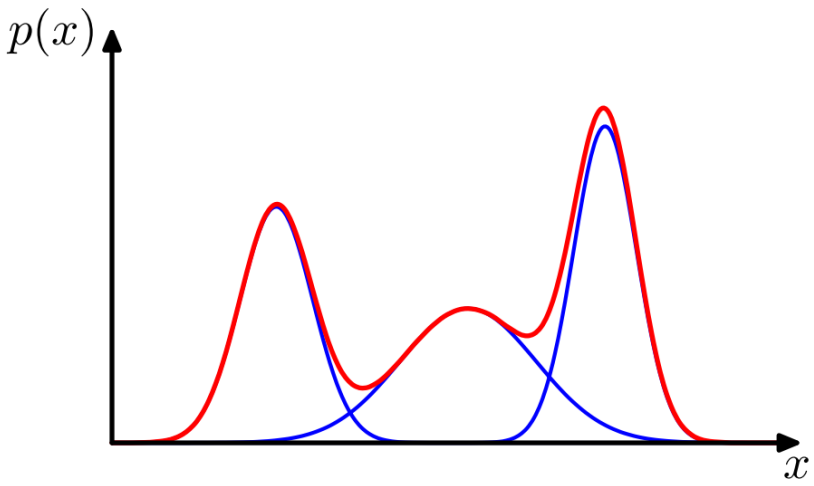


Abbildung 2.3: Beispielhafte Wahrscheinlichkeitsverteilung eines Gaußschen Mischmodells für eine eindimensionale Variable x und deren Dichtefunktion $p(x)$. In blau sind die drei einzelnen Gaußschen Verteilungen dargestellt, in rot deren Summe als Gaußsches Mischmodell. [Bis06, S. 111]

Die Gleichung 2.30 kann auch formuliert werden als

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k), \tag{2.32}$$

wobei $m_k = p(k)$ die Wahrscheinlichkeit angibt, dass die k-te Komponente einen Punkt generiert. $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\mathbf{x}|k)$ gibt an, wie wahrscheinlich es ist, dass die Daten $\mathbf{x}$ von der Komponente k generiert wurden. $p(\mathbf{x}|k)$ wird auch Likelihood-Funktion genannt und ist nach dem Satz von Bayes definiert als

$$p(\mathbf{x}|k) = \frac{m_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{J} m_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad \text{mit } J = K. \tag{2.33}$$

$p(\mathbf{x}|k)$ gibt also an, wie gut die Komponente k, definiert durch die Parametern $\boldsymbol{\mu}_k$ und $\boldsymbol{\Sigma}_k$, die Daten repräsentiert. Da diese Repräsentation der Daten als Gaußsches Mischmodell Ziel der Anwendung von Gaußschen Mischmodellen ist, führt das Maximieren dieser Funktion zum Erreichen des Ziels. Die zu maximierende Funktion wird daher auch als Zielfunktion bezeichnet (dies geschieht durch den EM-Algorithmus, der im folgenden Abschnitt erläutert wird). Als Zielfunktion für die Gaußschen Mischmodelle wird der Logarithmus der Likelihood-Funktion verwendet, der definiert ist als

$$\ln p(\mathbf{X}|\mathbf{m}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} m_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}, \tag{2.34}$$

mit $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_N$. [Bis06, S. 111 ff.]

Bevor auf den EM-Algorithmus eingegangen wird, muss noch definiert werden, wie die einzelnen Parameter $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ und m_k einer Komponente k berechnet werden. Da der jeweilige Parameter die Zielfunktion maximieren soll, wird dieser am Maximum der Zielfunktion berechnet. Das Maximum der Zielfunktion für $\boldsymbol{\mu}_k$ tritt auf, wenn die erste Ableitung der Zielfunktion nach den Mittelwerten $\boldsymbol{\mu}_k$ der Komponenten gleich Null ist und die zweite Ableitung

kleiner Null ist. Die erste Ableitung gleich Null gesetzt ergibt

$$0 = -\sum_{n=1}^{N} \frac{m_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{2.35}$$

mit

$$\gamma(z_{nk}) = \sum_{j=1}^{J} m_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \tag{2.36}$$

Umformungen ergeben

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{2.37}$$

mit

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{2.38}$$

Zum Berechnen der Maxima der Zielfunktion für die Werte $\boldsymbol{\Sigma}_k$ und m_k kann ebenso die erste Ableitung der Zielfunktion nach dem jeweiligen Wert gleich null gesetzt werden. Daraus resultiert

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \tag{2.39}$$

und

$$m_k = \frac{N_k}{N}. \tag{2.40}$$

Nachdem nun die Berechnung der einzelnen Parameter der Komponenten definiert ist, kann die Zielfunktion mittels des EM-Algorithmus maximiert werden. [Bis06, S. 435 ff.]

Definition Erwartungs-Maximierungs-Algorithmus

Die Anwendung des EM-Algorithmus auf ein Gaußsches Mischmodell zielt darauf ab, die Likelihood-Funktion zu maximieren, sodass das Gaußsche Mischmodell die Daten so genau wie möglich repräsentiert. Das Vorgehen des Algorithmus lässt sich in vier Schritten beschreiben. Bei der nachfolgenden Erklärung der Schritte wird auf die Bilder a-f der Abbildung 2.4 und die zuvor vorgestellten Formeln Bezug genommen.

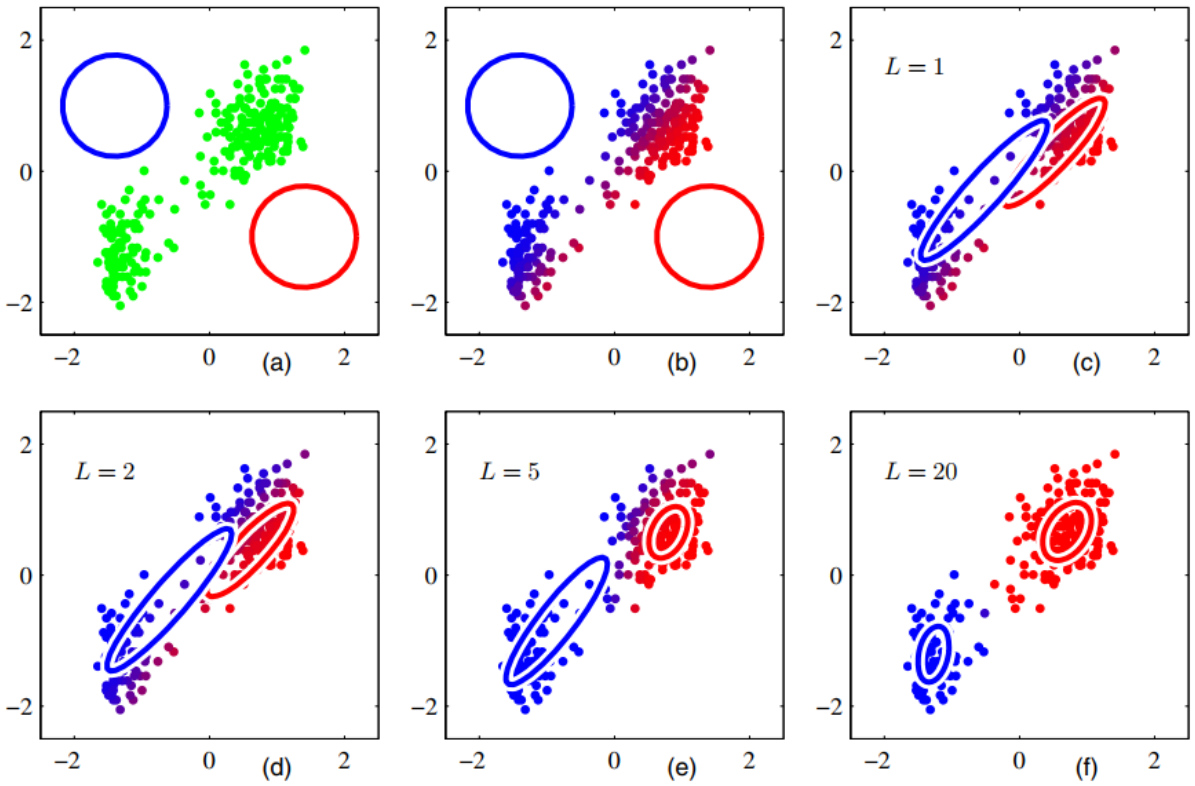


Abbildung 2.4: Ablauf des EM-Algorithmus für einen beispielhaften zweidimensionalen Datensatz, der in zwei Gaußsche Verteilungen (rot oder blau) aufgeteilt werden soll. Die Kreise zeigen jeweils den Bereich, an dem die Standardabweichung vom Mittelpunkt der Verteilung Eins ist. Die Farbe der Punkte zeigt an, welcher Verteilung dieser Punkt zugeordnet wurde. Die grünen Punkte sind die zu Beginn noch nicht klassifizierten Punkte. Violette Punkte haben eine hohe Wahrscheinlichkeit von beiden Verteilungen generiert worden zu sein. L gibt den Iterationsschritt des Algorithmus an. [Bis06, S. 437 f.]

Im ersten Schritt werden die zu erlernenden Parameter $\mathbf{m}$, $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ zufällig initialisiert (siehe a) und der Wert der Zielfunktion aus Formel 2.34 für diese Werte berechnet. Anschließend folgt der Erwartungsschritt, bei dem die Punkte den jeweiligen Clustern mittels der Formel 2.33 zugeordnet werden (siehe b). Daraufhin werden beim Maximierungsschritt die Parameter der Komponenten nach den Formeln 2.37, 2.39 und 2.40 angepasst (siehe c). Im vierten Schritt wird der Wert der Zielfunktion für die neuen Parameter berechnet. Sollten entweder die Parameter oder die Zielfunktion sich nicht mehr als um einen vordefinierten Schwellenwert verändern, terminiert der Algorithmus. Ansonsten wird dieser ab Schritt zwei wiederholt (siehe d bis f). [Bis06, S. 436 ff.]

Kapitel 3

Datenvorbereitung

Nach der Abhandlung der theoretischen Grundlagen, werden diese auf die zu analysierenden Daten angewandt. Im folgenden Abschnitt werden die zu analysierenden Daten und die bereits bestehenden Erkenntnisse über Anomalien in den Daten beschrieben. Anschließend werden fehlerhafte Daten aus dem Datensatz entfernt, um diesen für die Anomalieerkennung von außergewöhnlichen Daten vorzubereiten.

3.1 Datenbestand

In diesem ersten Abschnitt werden zunächst die zu untersuchenden Daten ausgewählt. Anschließend wird vorgestellt, welche Erkenntnisse über die Qualität der Daten bereits vorliegen.

3.1.1 Auswahl zu untersuchender Daten

Seit der Montage der im Kapitel 2.2 beschriebenen Infrastruktur an der Fokr im Jahr 2012 werden nahezu ununterbrochen Daten aufgezeichnet. In der vorliegenden Arbeit werden Daten vom 21. bis 29. Mai 2019 mit Python-Skripten auf Anomalien überprüft (der Zugriff auf den Quellcode kann über die Website `www.github.com/ChicDance` beim Autor per E-Mail angefordert werden). Dies sind 429.977 Trajektorien bestehend aus 109.582.458 Datenpunkten. Die Verfahren, bei denen ein Schwellenwert durch den Anwender festgelegt wird, können auf den gesamten Datensatz angewandt werden. Für die anderen Verfahren wird der Datensatz in einen Trainings- und einen Testdatensatz aufgeteilt. Der Testdatensatz besteht aus 60.555 Trajektorien vom 21. Mai 2019 und der Trainingsdatensatz aus den übrigen 369.422 Trajektorien vom 22. bis 29 Mai 2019. Der Trainingsdatensatz wird für die Erstellung der Modelle und Festlegung eines Schwellenwertes verwendet. Um mögliche Unterschiede in den Daten zwischen Wochentagen bei der Erstellung der Modelle zu berücksichtigen, beinhaltet der Trainingsdatensatz die Daten von einer Woche. Da für die vorliegenden Daten keine Beschriftung bezüglich deren Anomalität vorliegt, kann die Qualität der jeweiligen Modelle nicht mit dem Kreuzvalidierungsverfahren beurteilt werden. Daher wird der Testdatensatz immer auf Basis der mit dem Trainingsdatensatz erstellten Modelle und Schwellenwerte auf Anomalien überprüft.

Bei den anzuwendenden Verfahren werden jeweils unterschiedliche Daten analysiert. Eine erste Aufteilung des Datensatzes erfolgt bei der Anomalieerkennung in den Beschleunigungswerten im Abschnitt 3.2.3. Da sich die Schwellenwerte für die minimale und ma-

ximale Beschleunigung je nach Objektklasse unterscheiden, werden ab diesem Abschnitt nur Trajektorien von PKW analysiert (die Verkehrsstärke aller Verkehrsteilnehmer sowie die Anzahl der PKW im Zeitraum vom 21. bis 29 Mai 2019 kann der Abbildung 5 im Anhang entnommen werden). Damit verringert sich die Größe des zu untersuchenden Datensatzes ab diesem Abschnitt. Des Weiteren werden für die Anomalieerkennung in Kapitel 4 je nach anzuwendenden Verfahren unterschiedliche Datensätze verwendet.

Es wird vermutet, dass die Qualität der Trajektoriendaten vom Wetter beeinflusst wird. Daher gelten alle in dieser Masterarbeit getroffenen Aussagen über die Anomalieerkennung nur für die im Untersuchungszeitraum vorherrschenden Bedingungen. In der Zeit, in der Trainings- und Testdaten aufgezeichnet wurden, fiel kein Niederschlag, außer an einem Tag 0,3 l/m^2. Die durchschnittliche Sonnenscheindauer lag bei 6 Stunden und die Temperatur schwankte zwischen 6-23°C [Wet19].

3.1.2 Bisherige Erkenntnisse über die Datenqualität

Um die Ursache fehlerhafter Daten erklären zu können, werden nachfolgend die Erkenntnisse über die Datenqualität vorgestellt, die aus dem technischen Aufbau und der bisherigen Analyse der Daten bereits bekannt sind.

Aus dem technischen Aufbau der Messinfrastruktur folgt, dass eine höhere Geschwindigkeit zu einer verlässlicheren Bestimmung der Bewegungsrichtung führt. Im Umkehrschluss ist die detektierte Bewegungsrichtung und Position von Fahrzeugen mit einer geringen Geschwindigkeit weniger verlässlich. [Arn18, S. 31 f.] Des Weiteren geht aus dem Systemhandbuch der Fokr [Arn18] hervor, dass das Messrauschen der Abstandsmessung mit linear wachsender Entfernung quadratisch zunimmt. Ebenso wichtig für die vorliegende Arbeit ist die Information, dass die 3D-Position der Verkehrsteilnehmer mit einer Auflösung von 25 cm erfolgt. [Arn18, S. 44 f.] Zudem ist aus bisherigen Analysen der Daten bekannt, dass ein LKW manchmal als zwei Objekte detektiert wird und demnach zwei Trajektorien für ein Objekt gespeichert werden. Auch die Detektion von Motorradfahrern ist fehleranfällig, sodass diese teilweise mit Fahrradfahrern oder PKW verwechselt werden.

Nachfolgend wird vorgestellt, welche fehlerhaften Daten aus dem Datensatz entfernt werden, um diesen für die Analyse außergewöhnlicher Daten vorzubereiten.

3.2 Entfernung fehlerhafter Daten

Die fehlerhaften Daten können in vier Kategorien eingeteilt werden. Die erste Kategorie sind Trajektorien, die nur aus einem Eintrag bestehen und damit die Länge Eins aufweisen. Diese werden nicht weiter analysiert, da dies keine Trajektorien sondern Punkte sind. Ebenso wird überprüft, ob die Aufzeichnungsfrequenz von 25 fps eingehalten wurde. Die dritte Kategorie umfasst alle Verfahren mit denen die Einhaltung der physikalischen Gesetze überprüft wird. Der letzten Kategorie werden Trajektorien zugeordnet, denen keine Route zugewiesen werden kann. Das dafür verwendete Verfahren und dessen Anwendung wird zuletzt vorgestellt.

3.2.1 Trajektorien der Länge Eins

Zunächst werden Trajektorien, die nur aus einem Eintrag bestehen aus den Daten gefiltert. Damit wird die Anomalieerkennung auf die Länge der Trajektorien angewandt. Der

dafür benötigte maximale Schwellenwert wird vom Anwender auf Eins festgelegt, da Trajektorien die keine größere Länge als Eins aufweisen nicht wie die anderen Trajektorien verarbeitet werden können. Zum Beispiel kann keine Aufzeichnungsfrequenz, die für den folgenden Abschnitt relevant ist, ermittelt werden. Andere Trajektorien, die deutlich unter der durchschnittlichen Länge der Trajektorien von 255 Datenpunkten liegen, werden vorerst im Datensatz behalten, da zu kurze Trajektorien bei der Klassifizierung der Trajektorien nach Routen in Abschnitt 3.2.7 entfernt werden. Anschließend wird analysiert, in welchem Umfang die Trajektorien der Länge Eins auftreten. Abbildung 3.1 zeigt die jeweiligen einzelnen Positionswerte der 153 anomalen Trajektorien des gesamten Datensatzes von 429.977 Trajektorien auf der Kreuzung.

Abbildung 3.1: Trajektorien, die aus einem Datenpunkt bestehen

Wie die Abbildung vermuten lässt, treten ca. 89 % der Anomalien bei der Objektklasse Fahrradfahrer auf. Die Anomalien entstehen vor allem auf der Route von Osten nach Westen. In Abbildung 3.2 ist dieser Fehler aus der Sicht der Kamera, die im Westen auf den östlichen Innenbereich der Kreuzung ausgerichtet ist, zu erkennen.

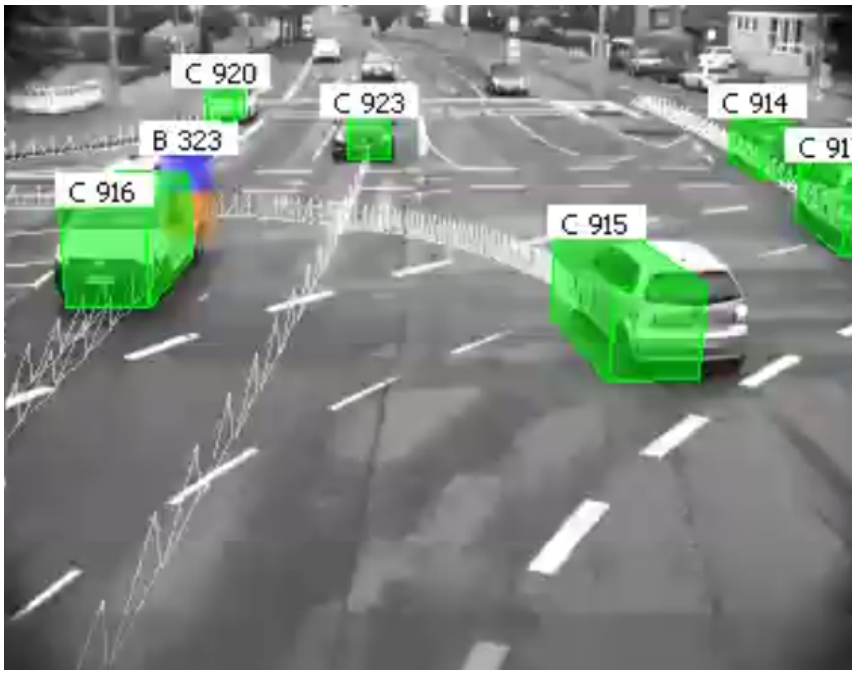


Abbildung 3.2: Kameraaufnahme einer Trajektorie der Länge Eins, die für den Fahrradfahrer B 323 aufgezeichnet wurde.

Damit wurde eine wiederkehrende Anomalie identifiziert, dessen Ursache in der Objektdetektion von vor allem Fahrradfahrern liegt. Weitere fehlerhafte Daten können aus der Analyse zweier aufeinanderfolgender digitaler Zeitstempel einer Trajektorie auftreten.

3.2.2 Aufzeichnungsfrequenz

Da jedem Videobild ein digitaler Zeitstempel zugeordnet wird, kann jedem Datenpunkt einer Trajektorie ein Zeitpunkt zugeordnet werden. Die Differenz der Zeitstempel zweier aufeinanderfolgender Datenpunkte sollte 40 ms betragen, weil im vorliegenden Fall die Daten mit einer Frequenz von 25 fps aufgezeichnet werden. Diese Eigenschaft der Daten wird für die spätere Anwendung der Markow-Kette angenommen. Eine Überprüfung der Daten auf diese Frequenz stellt eine Möglichkeit der Anomalieerkennung dar.

Die Analyse der Differenz zweier aufeinanderfolgender digitaler Zeitstempel einer Trajektorie ergibt, dass 99,97 % der 109.152.475 Zeitdifferenzen im Intervall von 39,998 ms bis 40,002 ms liegen. Daher wird der Schwellenwert für die Anomalieerkennung in der Aufzeichnungsfrequenz vom Anwender auf ±2s festgelegt. Alle Abweichungen von mehr als 2 s entsprechen einem Vielfachen von 40 ms. Bei den Vielfachen wurde ebenso eine Abweichung von ±2s detektiert. Dass die Abweichungen immer Vielfache von 40 ms darstellen, resultiert daraus, dass Verkehrsteilnehmer teilweise nicht in jedem Videobild erfasst werden. In Abbildung 3.3 wird die Position aller 33.156 identifizierten fehlerhaften Daten auf der Kreuzung dargestellt.

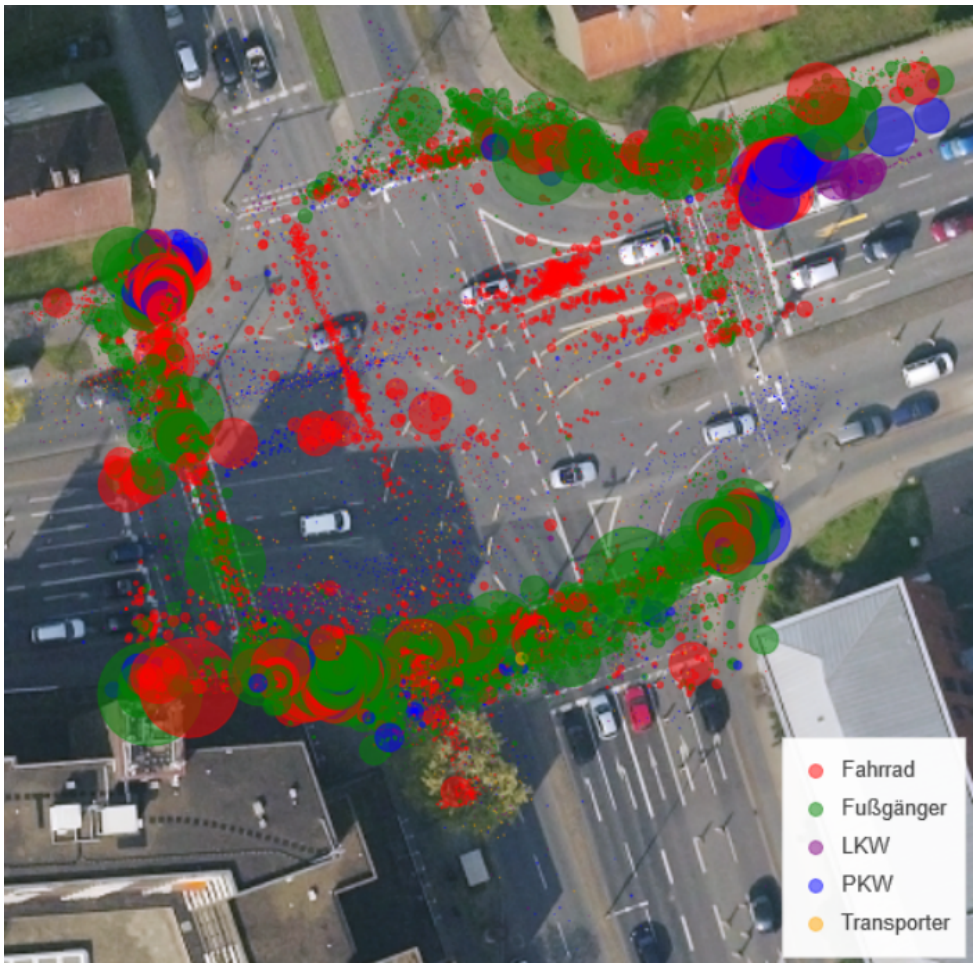


Abbildung 3.3: Datenpunkte mit fehlerhafter Differenz der digitalen Zeitstempel auf der Fokr. Die Größe des Kreises gibt die Differenz der digitalen Zeitstempel an. Der kleinste Kreis entspricht einer Differenz von 0,08 s und der größte Kreis einer Differenz von 62 s.

Die Anomalien in der Aufzeichnungsfrequenz treten bei 17.415 unterschiedlichen Verkehrsteilnehmern auf, weshalb diese aus dem Datensatz entfernt werden. Jeweils 40 % der Fehler lassen sich den Objektklassen der Fußgänger und Fahrradfahrer zuordnen. Bei Analyse des Videomaterials wurde festgestellt, dass die Anomalien in den Aufzeichnungsfrequenzen der Fußgänger und Radfahrer aus der fehlerhaften Bestimmung der Objektklasse resultieren. Damit werden die bisherigen Erkenntnisse über die Datenqualität aus Abschnitt 3.1.2 bestätigt, dass vor allem Motorradfahrer häufig als Fahrradfahrer klassifiziert werden. Der Wechsel zwischen den Klassen erzeugt Trajektorien, bei denen die Differenz zweier aufeinanderfolgender Datenpunkte nicht durchgängig 40 ms beträgt. Als weitere Ursache für diese Anomalie wurde das Anhalten von Verkehrsteilnehmern ermittelt. Diese Ursache kann auf der Abbildung 3.3 an der östlichen Einfahrt der Kreuzung vermehrt bei MRU beobachtet werden. Hier wurden MRU an der Haltelinie vor der LSA zunächst detektiert, beim Stehen nicht mehr erkannt und beim Anfahren wieder identifiziert. Eine umfangreichere Analyse der detektierten Anomalien wird an dieser Stelle nicht durchgeführt, da der Fokus dieser Arbeit auf außergewöhnlichen Daten liegt, welche nicht durch die Analyse der Aufzeichnungsfrequenz ermittelt werden können.

Eine weitere Methode zur Detektion von fehlerhaften Daten ist die Kontrolle der Einhaltung der physikalischen Gesetze. Wie in den Abschnitten 2.4.3, 2.4.4 und 2.4.5 beschrieben,

kann für jeden Datenpunkt überprüft werden, ob die erfassten Werte für Position, Geschwindigkeit und Beschleunigung physikalisch möglich sind. Die Ergebnisse der Überprüfung werden nachfolgend vorgestellt. Begonnen wird mit den Beschleunigungswerten, da die Anomalieerkennung in diesen Werten allein durch Anwendung eines vordefinierten Schwellenwertes stattfinden kann. Außerdem müssen zuerst die Beschleunigungswerte überprüft werden, da diese für die Berechnung der Geschwindigkeitswerte relevant sind.

3.2.3 Beschleunigung

Zur Überprüfung der Beschleunigungswerte auf dessen physikalische Machbarkeit werden die Beschleunigungswerte nach den in Kapitel 2.4.3 genannten Schwellwerten gefiltert. Hierbei werden nur die Werte der Objekte untersucht, die als PKW klassifiziert werden. Durch die zuvor angewandten Methoden hat sich der Datensatz auf 412.409 Trajektorien verringert, wovon 323.595 Verkehrsteilnehmer als PKW klassifiziert werden. In diesem Datensatz wird eine Trajektorie identifiziert, die sechs Beschleunigungswerte enthält, die den maximalen Schwellenwert überschreiten. Die Analyse des Beschleunigungsverlaufs dieser Trajektorie ergibt, dass die Beschleunigung des Verkehrsteilnehmers innerhalb von 120 ms von 2 auf 9,6 m/s^2 ansteigt. Dieser Anstieg erfolgt beim Anfahren des Verkehrsteilnehmers, bei dem eine Geschwindigkeit von 1,4 m/s ermittelt wird. Daraus wird geschlussfolgert, dass die Ursache der Anomalie aus einer in Abschnitt 2.2 beschriebenen Tatsache folgt. Diese beschreibt, dass die Verlässlichkeit der Positions- und Geschwindigkeitswerte, welche die Grundlage für die Ermittlung der Beschleunigungswerte bilden, bei geringer Geschwindigkeit abnehmen.

Im nächsten Schritt wird die Anomalieerkennung auf die Geschwindigkeitswerte angewandt.

3.2.4 Geschwindigkeit

In diesem Abschnitt wird die Geltung der Gleichungen 2.13 überprüft, indem die berechneten mit den gemessenen Geschwindigkeitswerten der Trajektorien verglichen werden. Die berechneten Werte werden ermittelt, indem für alle Datenpunkte einer Trajektorie, ausgenommen dem letzten, die Geschwindigkeit im darauffolgenden Datenpunkt nach der Gleichung 2.13 berechnet wird. Die berechneten Werte werden von der im darauffolgenden Zeitpunkt gemessenen Geschwindigkeit subtrahiert. Um in diesen Differenzen Anomalien zu identifizieren, muss ein Schwellenwert festgelegt werden. Dieser wird nach der in Abschnitt 2.3.2 vorgestellten Statistik, dem Z-Wert, auf Basis der verbleibenden 279.021 Trajektorien der PKW im Trainingsdatensatz ermittelt. Die Anwendung des Z-Wertes mit einem Schwellenwert von 3 Standardabweichungen (0,45 m/s) ergibt, dass 2,3 % der Werte und 91,1 % der Trajektorien als Anomalien klassifiziert werden (jede Trajektorie wird als Anomalie klassifiziert, wenn diese mindestens einen anomalen Datenpunkt enthält). Damit würden sich die Anomalien wie in Abbildung 6 im Anhang dargestellt auf nahezu der gesamten Kreuzung verteilen. Da aufgrund der Menge der Anomalien diese Datenpunkte nach der Definition einer Anomalie, nicht mehr als solche bezeichnet werden können und die Verteilung der Werte eher einer Laplace- anstatt einer Normalverteilung entspricht (siehe Abbildung 3.4), muss ein größerer Schwellenwert verwendet werden. Für diese Anpassung stellt eine Orientierung an dem Z-Wert, wenn dieser auf eine Normalverteilung angewandt wird, eine Möglichkeit dar. In einem solchen Fall werden 0,26 % der Werte als Anomalie klassifiziert. Um diesen Anteil an Anomalien in den Trajektorien zu erreichen, müssen alle Werte, die

mehr als 12,2 Standardabweichungen (1,8 m/s) vom Mittelwert entfernt sind, als Anomalie klassifiziert werden. Die Verteilung der Geschwindigkeitswerte des Trainingsdatensatzes und die verwendeten Schwellenwerte sind in Abbildung 3.4 dargestellt.

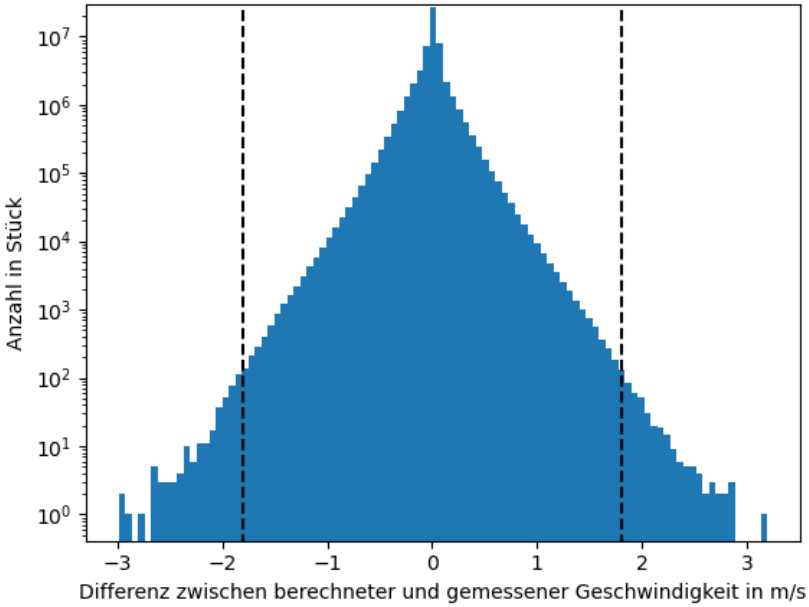


Abbildung 3.4: Histogramm der Differenzen zwischen gemessenen und berechneten Geschwindigkeitswerten aller PKW im Trainingsdatensatz. Die schwarz gestrichelten Linien stellen den minimalen bzw. maximalen Schwellenwert von 12,2 Standardabweichungen dar. Die Werte außerhalb der gestrichelten Linien werden als Anomalie klassifiziert.

Mit diesen Schwellenwerten können die Anomalien in den Testdaten bestimmt werden. Der Testdatensatz hat sich von 60.555 Trajektorien um 22 Trajektorien der Länge Eins und 2.842 Trajektorien mit Anomalien in der Differenz der digitalen Zeitstempel auf 57.691 Trajektorien verringert. Davon werden 44.573 Trajektorien der Objektklasse PKW zugeordnet. Die Beschleunigungswerte dieses Datensatzes über- oder unterschreiten keine Schwellenwerte für die Beschleunigung eines PKWs. Daher bilden die 44.573 Trajektorien die Grundlage für die Anomalieerkennung in den Differenzen der gemessenen und berechneten Geschwindigkeitswerten. Die Anwendung der Schwellenwerte auf den Datensatz ergibt, dass 129 von 9.290.013 Differenzen und 127 von 44.573 Trajektorien als Anomalie klassifiziert werden. Die als Anomalie klassifizierten Differenzen treten auf der Kreuzung an den in Abbildung 3.5 dargestellten Positionen auf.

Abbildung 3.5: Verteilung der Anomalien in den Differenzen der gemessenen und berechneten Geschwindigkeitswerte von PKW des Testdatensatzes unter Verwendung eines Schwellenwertes von 1,8 m/s.

Es wird deutlich, dass sich diese an der nördlichen und westlichen Ausfahrt konzentrieren. Der höchste Anomalie-Wert wird zu Beginn der Route von Westen nach Norden gemessen. Auf eine weitere Analyse der fehlerhaften Daten wird verzichtet, da der Fokus dieser Arbeit auf der Analyse der außergewöhnlichen Daten liegt.

Mit der Anwendung des Schwellenwertes für die Geschwindigkeitswerte auf den Trainingsdatensatz werden 757 von 279.021 Trajektorien aus diesem entfernt. Die übrigen Trajektorien werden zur Bestimmung der Schwellenwerte für die Differenz zwischen den berechneten und gemessenen X- und Y-Positionen verwendet. Dazu werden nachfolgend die Geltung der Gleichungen 2.16 und 2.17 überprüft, wobei mit den Werten der X-Positionen begonnen wird.

3.2.5 X-Position

Die Anwendung des Z-Wertes mit einem Schwellenwert von drei Standardabweichungen (0,11 m) auf die Differenzen der X-Positionen im Trainingsdatensatz ergibt, dass 2,3 % der Werte und 76 % der Trajektorien als Anomalie klassifiziert werden. Daher wird wie bei den Geschwindigkeitswerten der Schwellenwert über den angestrebten Anteil an Anomalien von

0,26 % festgelegt. Der Schwellenwert, mit dem dieser Anteil an Anomalien im Trainingsdatensatz (753 von 278.264 Trajektorien) erreicht wird, liegt bei 16,7 Standardabweichungen (0,62 m). Mit der Anwendung dieses Schwellenwertes auf den Testdatensatz werden 128 von 9.253.807 Differenzen (107 von 44.466 Trajektorien) als Anomalie klassifiziert.

3.2.6 Y-Position

Um Anomalien in den Differenzen der gemessenen und berechneten Y-Position zu finden, wird das gleiche Vorgehen wie für die X-Positionen angewandt. Der Anteil der Anomalien von 0,26 % wird mit einem Schwellenwert von 17,8 Standardabweichungen (0,67 m) im Trainingsdatensatz erreicht. Mit diesem Schwellenwert werden im Testdatensatz 168 Differenzen (127 Trajektorien) als Anomalie bewertet.

Für den folgenden Schritt der Datenvorbereitung werden die identifizierten Anomalien nach der Differenz der berechneten und gemessenen X- und Y-Position aus dem Datensatz entfernt. Von den insgesamt 234 identifizierten anomalen Trajektorien weisen sieben sowohl nach der X-Position als auch nach der Y-Position Anomalien auf. Daher reduziert sich der Testdatensatz von 44.466 um 227 auf 44.219 Trajektorien (9.248.346 Datenpunkte).

Im folgenden Abschnitt wird vorgestellt, welche Trajektorien nach der Klassifizierung der Route als Anomalie bewertet werden und welche Anomalien mit diesem Verfahren im Datensatz detektiert werden.

3.2.7 Klassifizierung der Trajektorien nach Routen

Die zu untersuchenden Objekte, die als PKW klassifiziert werden, können an der Fokr aus jeder der vier Richtungen in jede der vier Richtungen weiterfahren. Damit ergeben sich 16 unterschiedliche Routen, nach denen die Verkehrsteilnehmer klassifiziert werden können. Die Routen werden nachfolgend durch zwei Buchstaben abgekürzt, von denen der erste Buchstabe für die Himmelsrichtung steht, aus der ein Verkehrsteilnehmer in die Kreuzung eingefahren ist und der zweite Buchstabe für die Himmelsrichtung steht, in die ein Verkehrsteilnehmer die Kreuzung verlassen hat (die Himmelsrichtungen Norden, Osten, Süden und Westen werden mit den Buchstaben N, E, S und W abgekürzt). Für die Klassifizierung werden optische Schleifen verwendet, die sogenannte „Areas of Interest“ (AOI) darstellen. Dies sind bestimmte Bereiche auf der Kreuzung, an denen untersucht wird, ob ein Fahrzeug diesen Bereich passiert. Sollte eine Trajektorie jeweils einen Schnittpunkt mit einer Ein- und Ausfahrt aufweisen, so wird dieser Trajektorie die Route von der Einfahrt zur Ausfahrt zugewiesen. Als Einfahrt in die Kreuzung wird zunächst der Bereich unmittelbar hinter der Radfahrerfurt gewählt. Äquivalent dazu werden die AOI für die Ausfahrten direkt vor der Radfahrerfurt platziert. Dieses Verfahren erfordert nicht nur das Erstellen von AOI, sondern auch deren optimale Positionierung auf der Kreuzung, sodass so viele Trajektorien wie möglich richtig klassifiziert werden. Mit der Anpassung der zuvor beschriebenen Positionierung der AOI wie in Abbildung 7 im Anhang dargestellt, werden 33.065 von 44.219 Trajektorien (75 %) erfolgreich einer Route zugeordnet. Der Grund dafür, dass 11.154 Trajektorien keiner Route zugeordnet werden, wird in den Abbildungen 8 und 9 im Anhang deutlich. Diese zeigen jeweils eine Heatmap für die Verteilung der ersten bzw. letzten Datenpunkte eines Verkehrsteilnehmers auf der Fokr. Auf diesen ist zu erkennen, dass diverse Trajektorien erst im Innenbereich der Kreuzung beginnen oder enden.

Um möglichst viele Trajektorien mit der Klassifizierung mittels AOI zu erfassen, müssen diese angepasst werden. Die AOI dürfen allerdings nicht im Bereich von Routen liegen,

deren Verkehrsteilnehmer nicht der AOI zugeordnet werden sollen. Eine zu weite Verschiebung einer AOI in Richtung der Kreuzungsmitte erzeugt fehlerhafte Routenzuweisungen, da dadurch Schnittpunkte der AOI mit Trajektorien anderer Routen entstehen. Die Platzierung der AOI, wie in Abbildung 3.6 zu sehen, ermöglicht die erfolgreiche Zuordnung von 40.879 von 44.219 Trajektorien (92 %) zu einer Route.

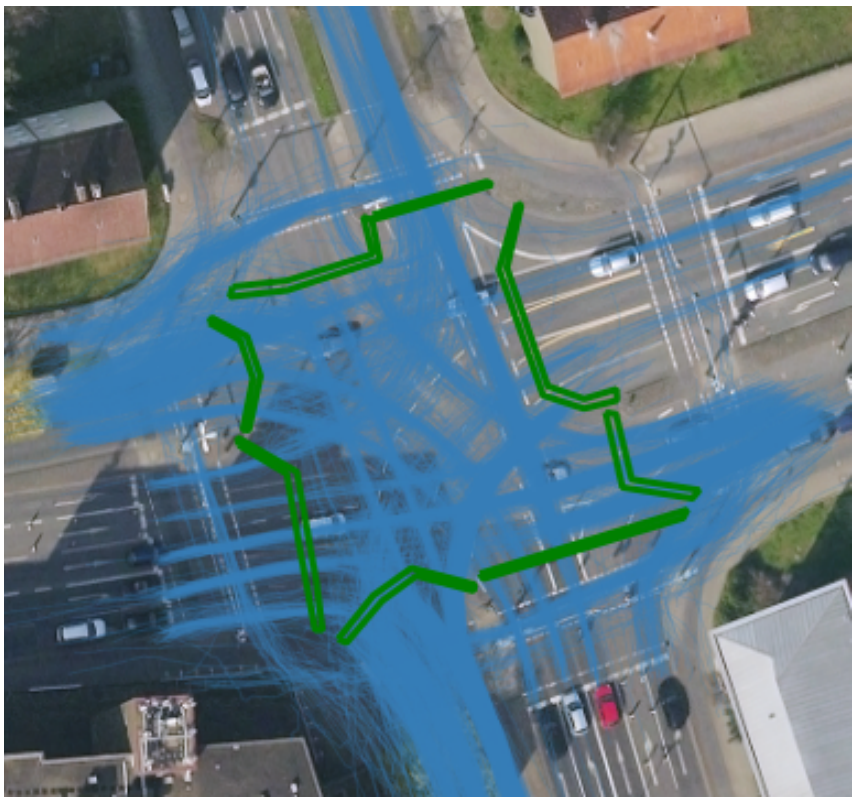


Abbildung 3.6: Trajektorien (blau), denen mit den angepassten Polygonen (grün) keine Route zugewiesen werden kann.

Die 3.340 Trajektorien, die keiner Route zugeordnet werden und damit eine Anomalie darstellen, werden nachfolgend nach den in Abschnitt 2.4.6 ersten vier genannten Kategorien unterschieden. Dabei werden die Trajektorien analysiert, für die mehrere Ein- oder Ausfahrten erfasst werden.

Von den 3.340 anomalen Trajektorien weisen 3.011 Trajektorien keinen Schnittpunkt mit einem Polygon einer Einfahrt und 1.186 Trajektorien keinen Schnittpunkt mit dem Polygon einer Ausfahrt auf. Dies ist auf die Tatsache zurückzuführen, dass ein Teil der Verkehrsteilnehmer erst auf der Kreuzung detektiert werden oder schon vor der Ausfahrt nicht mehr detektiert werden.

Keine Trajektorie wies Schnittpunkte mit mehr als einem Polygon einer Ausfahrt auf. Allerdings werden für zwei Trajektorien mehr als eine Einfahrt vermerkt. Diese werden nachfolgend mit dem Ziel analysiert, zu entscheiden, ob die Anomalien aus fehlerhaften oder außergewöhnlichen Daten entstanden ist. Hierzu wird das Videomaterial gesichtet, in dem die beiden Verkehrsteilnehmer erfasst werden. In Abbildung 3.7 ist eine der Anomalien in Form der zuletzt detektierten Position eines Fahrzeuges dargestellt, dass die Fahrbahn verlassen hat und auf einen Gehweg gefahren ist.

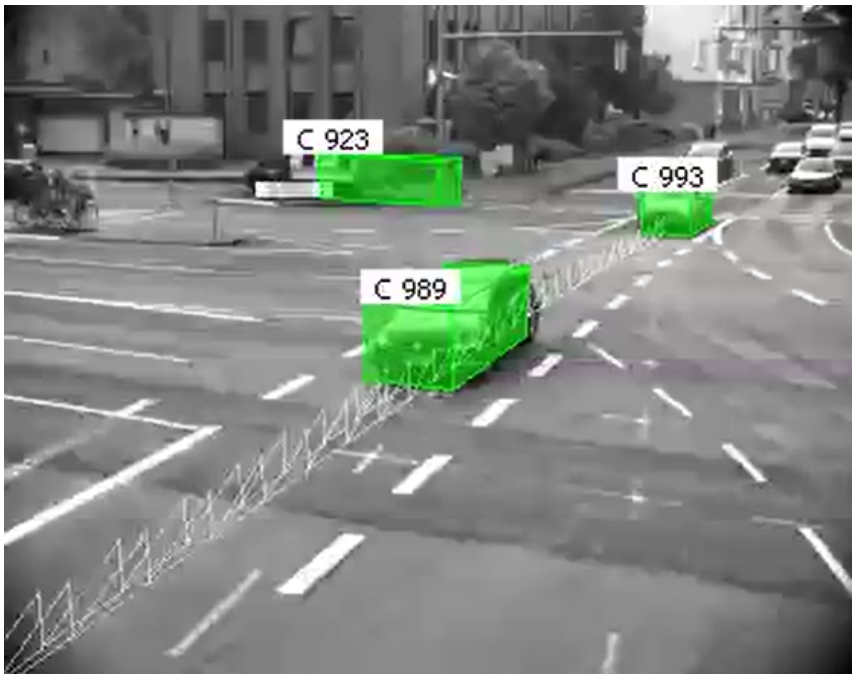


Abbildung 3.7: Anomalie Objekt C 923 (grün): Die Trajektorie des Objektes schneidet zwei Polygone, die sich an einer Einfahrt der Kreuzung befinden, da es die Fahrbahn verlassen hat und auf einen Gehweg gefahren ist.

Die Trajektorie des zuerst untersuchten anomalen Objektes wird demnach als Anomalie bewertet, weil außergewöhnliche Daten aufgezeichnet werden. Dies trifft nicht auf die zweite Anomalie zu. Wie der Abbildung 10 im Anhang entnommen werden kann, ergibt sich diese Anomalie aus der fehlerhaften Positionsbestimmung eines Objektes auf der Route von Osten nach Süden.

Auswahl der zu untersuchenden Route

Alle in dieser Arbeit angewandten Verfahren zur Anomalieerkennung, außer den Gaußschen Mischmodellen, erfordern eine Analyse von Trajektorien, die die gleiche Route befahren. Des Weiteren lassen sich die Verfahren auf die Trajektorien von jeder der 16 möglichen Routen anwenden. Da eine Untersuchung aller Routen den maximalen Umfang dieser Arbeit übersteigt, wird eine zu untersuchende Route ausgewählt. In dieser Arbeit werden beispielhaft die Ergebnisse der Route von Westen nach Norden vorgestellt. Da sich auf dieser Route die Verkehrsteilnehmer von eine auf zwei Spuren verteilen, werden die Trajektorien nochmals nach diesen Spuren unterschieden. Von den 1.060 Trajektorien für die Route von Westen nach Norden, werden 743 Trajektorien der rechten Spur (siehe Abbildung 11 im Anhang) und 317 Trajektorien der linken Spur (siehe Abbildung 12 im Anhang) zugeordnet. Die Zuordnung erfolgte auf Basis der AOI, die direkt hinter der Fußgängerfurt platziert sind, sodass auch die kürzeste Trajektorie dieser Route klassifiziert werden kann. Damit werden die Trajektorien nach deren Verlauf soweit unterschieden, dass die Berechnung einer durchschnittlichen Trajektorie ein sinnvolles Ergebnis liefern kann. Nachfolgend wird das abstandsbasierte Verfahren auf die 743 Trajektorien auf der Route WN angewandt, die in der rechten Spur enden. Mit den statistik- und richtungsbasierten Verfahren werden die 1.060 Trajektorien der gesamten Route WN untersucht. Abschließend werden für die Anwendung des Gaußschen Mischmodells alle 40.770 Trajektorien von PKWs, die am 21.05.2019 aufgezeichnet

werden, verwendet. Die Ergebnisse dieser Verfahren werden nachfolgend vorgestellt und analysiert.

Kapitel 4

Anwendung der Verfahren

In diesem Kapitel werden die Verfahren zur Erkennung von Anomalien, die aus außergewöhnlichen Daten entstanden sind, auf die jeweils zu untersuchenden Daten angewandt und die detektierten Anomalien analysiert. Die angewandten Verfahren sind die Hausdorff-Metrik, der diskrete euklidische Raum und das Gaußsche Mischmodell.

4.1 Hausdorff-Metrik

Die Hausdorff-Metrik wird bei diesem Verfahren verwendet, um die Distanz zwischen zwei Trajektorien zu messen. Nachfolgend wird der Ablauf des anzuwendenden Verfahrens vorgestellt und dieses anschließend auf die zu untersuchenden Daten angewandt. Dazu werden die durchschnittliche Trajektorie des Trainingsdatensatzes und die zu erwartenden Abstände entlang der durchschnittlichen Trajektorie ermittelt. Mit diesem Modell werden zum Abschluss Anomalien im Testdatensatz detektiert und analysiert.

4.1.1 Ablauf des Verfahrens

Mittels der in Abschnitt 2.5.1 vorgestellten Hausdorff-Metrik wird zu Beginn des Verfahrens die durchschnittliche Trajektorie TJ_d des Trainingsdatensatzes TJS ermittelt. Anschließend wird für jeden Datenpunkt einer Trajektorie die minimale euklidische Distanz zu den Datenpunkten der TJ_d berechnet. Die Bewertung eines Datenpunktes auf Anomalität erfolgt durch den Vergleich des Abstandes mit einem Schwellenwert. Dieser Schwellenwert wird für jeden Datenpunkt der durchschnittlichen Trajektorie individuell mit dem Z-Wert ermittelt. Ein für alle Datenpunkte gültiger Schwellenwert empfiehlt sich hierfür nicht, da mit diesem die Verteilung der Datenpunkte nicht berücksichtigt werden würde. Um die Verteilung zu berücksichtigen, werden für jeden Datenpunkt der TJ_d, die Datenpunkte der Trajektorien im Trainingsdatensatz ermittelt, die zu dem Datenpunkt den minimalen Abstand aufweisen. Anschließend wird für jeden Datenpunkt der TJ_d die Verteilung der Abstände der zugeordneten Datenpunkte des Trainingsdatensatzes bestimmt. Für die Anwendung des Z-Wertes auf die Abstandswerte wird angenommen, dass sich die einem Datenpunkt der durchschnittlichen Trajektorie zugeordneten Datenpunkte des Trainingsdatensatzes auf einer Achse orthogonal zur Fahrtrichtung verteilen. Die Annahme wird begründet durch die Tatsache, dass der maximale Abstand zwischen zwei aufeinanderfolgenden Datenpunkten der durchschnittlichen Trajektorie 0,46 m beträgt. Die resultierende Verteilung stellt das

Modell dieses Verfahrens dar, auf dessen Grundlage für jeden Datenpunkt aus dem Testdatensatz ein Anomalie-Wert berechnet wird. [Lax13, S. 1160]

4.1.2 Ermittlung der durchschnittlichen Trajektorie

Wie im Abschnitt 2.5.1 beschrieben, wird bei diesem Verfahren zunächst die durchschnittliche Trajektorie des Trainingsdatensatzes TJS der Größe N ermittelt. Die Größe des Trainingsdatensatzes verringert sich durch die Entfernung der Trajektorien, die fehlerhafte Daten enthalten, auf 276.822 Trajektorien verringert. Davon werden 6.297 Trajektorien der zu untersuchenden Route WN zugeordnet. Des Weiteren ist für die Anwendung der Hausdorff-Metrik die Unterscheidung der Trajektorien nach Spuren notwendig. Nach diesem Kriterium werden 4.604 Trajektorien der rechten und 1.687 Trajektorien der linken Spur zugeordnet. Nachfolgend werden die Trajektorien der rechten Spur untersucht. Da für die Ermittlung der durchschnittlichen Trajektorie für jeden Datenpunkt einer Trajektorie der Abstand zu allen anderen Datenpunkten im Datensatz berechnet wird, ist dieses Verfahren rechenintensiv. Allerdings wird festgestellt, dass eine durchschnittliche Trajektorie auch aus der Analyse einer Stichprobe des Datensatzes im Umfang von 200 zufällig ausgewählten Trajektorien ermittelt werden kann. Demnach wird diese Stichprobe ermittelt und die Entfernung zwischen zwei Trajektorien in einer 200×200 anstelle einer 4.604×4.604 Matrix gespeichert. Der Wert der Matrix in Zeile i und Spalte j stellt die Hausdorff-Distanz zwischen den Trajektorie TJ_i und TJ_j dar. Das Summieren der Werte über alle Spalten, erzeugt einen Vektor, der für jede Trajektorie die Summe der Distanzen zu allen anderen Trajektorien enthält. Der Trajektorie, der der minimale Wert des Vektors zugeordnet wird, wird als durchschnittliche Trajektorie ausgewählt.

Mit diesem Vorgehen wird die ausgewählte durchschnittliche Trajektorie in Bereichen, in denen die Trajektorien eine vergleichsweise hohe Abweichung von der durchschnittlichen Route aufweisen, zentral verlaufen. In Bereichen, in denen eine geringe Abweichung vorliegt, wird die ermittelte durchschnittliche Trajektorie nicht zwingend mittig verlaufen. Dies resultiert aus der Anwendung der Hausdorff-Metrik. Nach dieser wird der Abstand zweier Trajektorien nur auf Basis des größten Abstandes ermittelt. Dies kann weiter unten im Dokument der Abbildung 4.1 entnommen werden, in der deutlich wird, dass die ermittelte durchschnittliche Trajektorie im Bereich der Einfahrt nicht mittig im Vergleich zu den anderen Trajektorien verläuft. Allerdings liegt die Trajektorie sehr zentral im Bereich der Überquerung der entgegenkommenden Fahrbahn, in welchem sich die Verläufe der Trajektorien entlang der Route am meisten unterscheiden. Dies stellt eine Schwäche dieses Verfahrens dar. Um dieser Schwäche entgegenzuwirken kann die durchschnittliche Trajektorie für zufällige Stichproben des Datensatz mehrmals berechnet werden und anschließend die durchschnittliche Trajektorie aus den durchschnittlichen Trajektorien ermittelt werden. Aufgrund des akzeptablen Ergebnisses für die Ermittlung der durchschnittlichen Trajektorie ohne mehrmalige Berechnung, wird diese in der vorliegenden Arbeit nicht angewandt.

4.1.3 Erstellung des Modells

Nachfolgend wird die in Abschnitt 4.1.2 ermittelte durchschnittliche Trajektorie verwendet und für jeden Datenpunkt des Trainingsdatensatzes der minimale Abstand zu dieser berechnet. Aufgrund der Nähe aufeinanderfolgender Datenpunkte der durchschnittlichen Trajektorie, besteht die Gefahr, dass einem Datenpunkt nur wenige Datenpunkte des Trainingsdatensatzes zugeordnet werden. Damit könnte die berechnete Verteilung der Abstände zur

durchschnittlichen Trajektorie an Aussagekraft verlieren. Im vorliegenden Fall betrug die minimale Anzahl der Datenpunkte die einem Datenpunkt der durchschnittlichen Trajektorie zugeordnet wurden 263, sodass dieser Fall nicht eingetreten ist. Das finale Modell wurde aus den 4.606 Trajektorien des Trainingsdatensatzes erstellt und wird nachfolgend zur Bewertung des Testdatensatzes verwendet.

4.1.4 Analyse der Anomalien

Für die Bewertung des Testdatensatzes wird für jede der 743 Trajektorien für jeden Datenpunkt der Abstand zur durchschnittlichen Trajektorie ermittelt und anhand des vorliegenden Modells dem ermittelten Abstand ein Z-Wert zugeordnet. Dieser Z-Wert kann anschließend zu Anomalieerkennung verwendet werden. Die Verwendung des üblichen Schwellenwertes von drei Standardabweichungen für die Anomalieerkennung klassifiziert die in Abbildung 13 im Anhang dargestellten Datenpunkte als Anomalien. Mit diesem Schwellenwert werden 2.937 von 464.986 Datenpunkten (0,63 %) und 197 von 743 (26,51 %) Trajektorien als Anomalie klassifiziert. Durch Anpassung dieses Schwellenwertes kann bestimmt werden, ab welchem Grad der Anomalität ein Datenpunkt als Anomalie klassifiziert werden soll. Die Anpassung des Z-Schwellenwertes auf fünf (siehe Abbildung 4.1) verringert die Anzahl der Anomalien in den Datenpunkten auf 705 (0,15 %) und in den Trajektorien auf 34 (4,48 %).

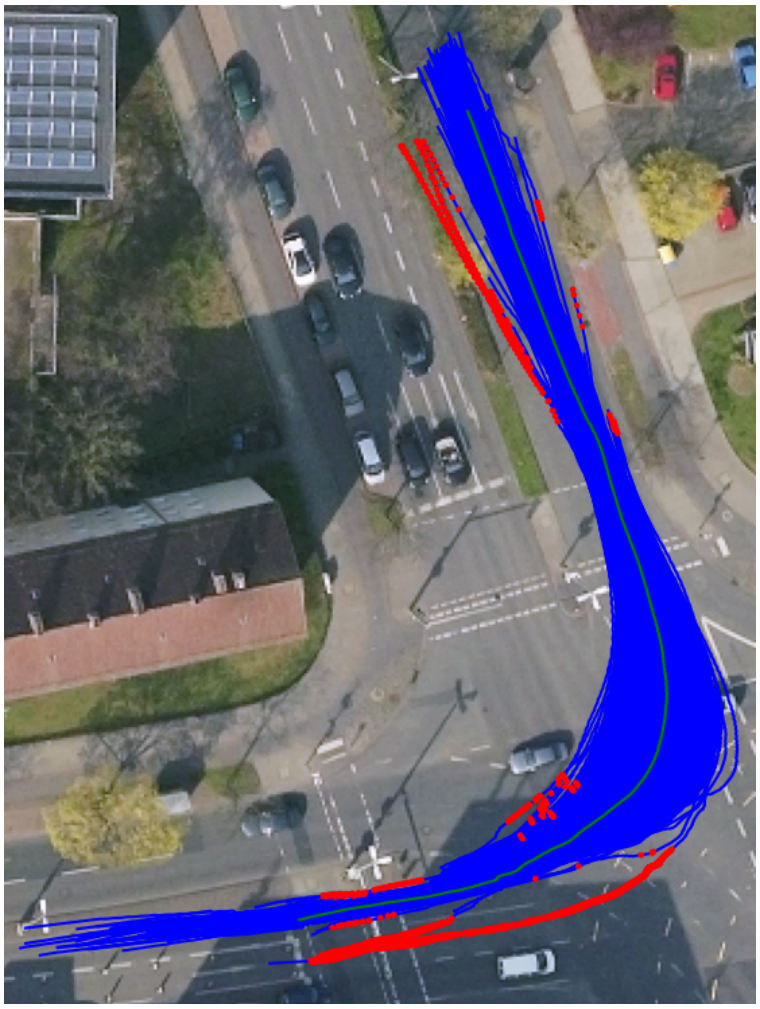


Abbildung 4.1: Mit der Hausdorff-Metrik und einem Z-Schwellenwert von fünf bestimmte anomale Datenpunkte (rote Kreise). Die blauen Linien sind der zugrunde liegenden Testdatensatz von Trajektorien, die von Westen nach Norden auf die rechte Spur führen. Die grüne Linie stellt die durchschnittliche Trajektorie dar.

Die in der Abbildung 4.1 zu erkennenden detektierten Anomalien werden nachfolgend nach deren Ursachen in Anomalien unterschieden, die aufgrund fehlerhafter oder außergewöhnlicher Daten entstanden sind. Letzteres stellen die Trajektorien von Verkehrsteilnehmern dar, die von der Fahrbahn für Geradeausfahrer in die Kreuzung einfahrend, anschließend die Route nach Norden gewählt haben. Diese weisen eine anomale Abweichung zur mittleren Trajektorie von der Haltelinie bis zu Einbiegen des Verkehrsteilnehmers auf. Die Datenpunkte vor der Haltelinie werden nicht als Anomalie klassifiziert, da diese dem ersten Datenpunkt der durchschnittlichen Trajektorie zugeordnet werden und für diesen die gemessene Abweichung noch nicht als anomal gilt. Nach Sichtung des Videomaterials für die zwei detektierten Trajektorien wird die Vermutung bestätigt, dass dies wirklich außergewöhnliche Daten darstellen und nicht aufgrund eines Fehlers entstanden sind. Ein Videobild einer dieser Trajektorien kann der Abbildung 4.2 entnommen werden.

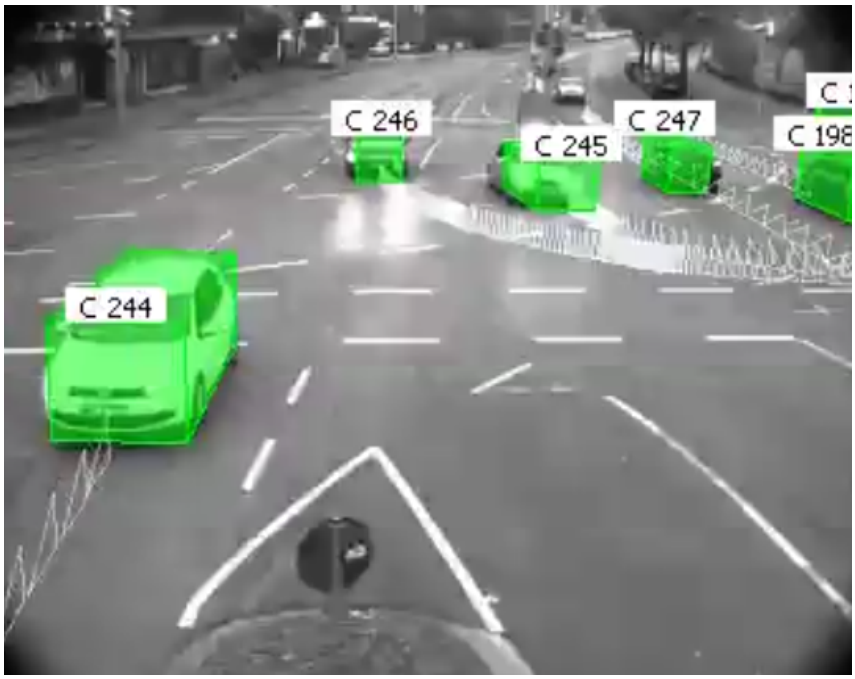


Abbildung 4.2: Videobild der Kamera, die im Osten auf den Innenbereich der Kreuzung nach Westen ausgerichtet ist. Das Fahrzeug C 246 wird nach der Hausdorff-Metrik als Anomalie klassifiziert, da für diesen Verkehrsteilnehmer eine Abweichung von der durchschnittlichen Trajektorie detektiert wurde.

Bei allen verbleibenden detektierten Anomalien werden in den Videoaufzeichnungen Abweichungen zwischen der detektierten und realen Position des Verkehrsteilnehmers deutlich. Die verbleibenden Anomalien wurden demnach alle wegen einer fehlerhaften Datenaufzeichnung als Anomalie klassifiziert und werden nachfolgend in vier Kategorien aufteilt und ausgewertet.

Die erste Kategorie umfasst die anomalen Datenpunkte am Ende der Route, welche aufgrund deren Abweichungen von der durchschnittlichen Trajektorie sowohl auf dessen rechten als auch dessen linken Seite als Anomalie klassifiziert werden. Die fehlerhafte Datenaufzeichnung an diesen Bereichen der Kreuzung ist darauf zurückzuführen, dass das Messrauschen mit linear wachsender Entfernung quadratisch zunimmt. Da diese Datenpunkte nur noch im Sichtbereich der Mono-Video-Kamera liegen, die am Süden der Kreuzung platziert ist, besteht eine große Entfernung zwischen den Verkehrsteilnehmern und der Kamera. Ein Beispiel für diese Kategorie kann der Abbildung 14 im Anhang entnommen werden.

Die zweite Kategorie beinhaltet die Anomalien, die sich im Bereich der Fußgänger- und Radfahrerfurt auf der linken Seite der durchschnittlichen Trajektorie befinden. Beide anomale Trajektorien werden gegen 21:30 Uhr aufgezeichnet, zu einer Zeit, in der die Qualität der Sichtverhältnisse aufgrund der Dunkelheit gering war. Es wird vermutet, dass dies ein Grund für die fehlerhafte Positionsbestimmung ist. Ein Beispiel dieser Kategorie kann der Abbildung 15 im Anhang entnommen werden.

Die Anomalien auf der rechten Seite der durchschnittlichen Trajektorie im Bereich der Fußgänger- und Radfahrerfurte und des Innenbereichs der Kreuzung werden der dritten Kategorie zugeordnet. Bei dieser Kategorie ist die fehlerhafte Positionsbestimmung darauf zurückzuführen, dass die detektierten PKWs Motorradfahrer sind. Demnach liegt in diesen

Fällen auch eine fehlerhafte Objektklassifizierung vor. Die Trajektorien der Motorradfahrer dürfen nicht mit der durchschnittlichen Trajektorie von PKWs verglichen werden, da diese eine abweichende durchschnittliche Trajektorie besitzen können. Dies wird in Abbildung 16 im Anhang deutlich, in der der Motorradfahrer C 714 im rechten Bereich der Spur erkennbar ist. Sollten mehrere Motorradfahrer im rechten Bereich der Spur fahren, würde auch die durchschnittliche Trajektorie weiter rechts verlaufen und diese Trajektorie nicht als Anomalie klassifiziert werden.

Zu der letzten Kategorie werden die Anomalien von Verkehrsteilnehmern gezählt, die im Bereich der Route NS eine enge Kurvenfahrt auf der linken Seite der durchschnittlichen Trajektorie aufweisen. Alle detektierten Anomalien dieser Kategorie weisen eine fehlerhafte Positionsbestimmung auf, wie diese beispielhaft in Abbildung 17 im Anhang dargestellt ist. Demnach wird geschlussfolgert, dass in diesem Bereich der Kreuzung die Positionsbestimmung grundsätzlich fehlerhaft ist.

Damit wurden alle 34 detektierten anomalen Trajektorien analysiert. Da Ziel des Verfahrens die Erkennung von Anomalien ist, die aus außergewöhnlichen Daten entstanden sind, könnte der Schwellenwert für die Anomalieerkennung angepasst werden, um nur solche Trajektorien zu identifizieren. Für die Detektion der außergewöhnlichen Daten müsste ein Schwellenwert von elf statt von fünf verwendet werden.

Die zuvor analysierten Anomalien umfassen die Datenpunkte, die als Anomalie klassifiziert werden. Des Weiteren können Datenpunkte existieren, die nicht als anomal erkannt werden, allerdings nach der Definition des Anwenders eine Anomalie darstellen. Diese nicht detektierten Anomalien werden nachfolgend in zwei Kategorien unterschieden.

Zum einen sind dies Verkehrsteilnehmer, die vor dem Abbiegen nach Norden weit in die Kreuzung eingefahren sind, aber nicht als Anomalie bewertet wurden. Die Sichtung des Videomaterials ergab, dass diese Daten die Realität wiedergeben. Mit einer Verringerung des Schwellenwertes könnten diese Trajektorien auch als Anomalie klassifiziert werden. Da allerdings aus diesen Trajektorien keine sicherheitsrelevanten Schlussfolgerungen gezogen werden können, wird der Schwellenwert nicht verringert.

Die zweite Kategorie nicht detektierter Datenpunkte, umfasst die am frühesten und am spätestens detektierten Datenpunkte der Route. Diese wurden nicht als Anomalie klassifiziert, obwohl diese einen großen absoluten Abstand zu dem ersten bzw. letzten Punkt der durchschnittlichen Trajektorie aufweisen. Diese Datenpunkte würden auch nicht als Anomalie klassifiziert werden, wenn diese deutlich orthogonal zur Fahrtrichtung abweichen. Der Grund dafür ist, dass die Datenpunkte dem jeweils ersten bzw. letzten Datenpunkt der durchschnittlichen Trajektorie zugeordnet werden. Somit wird bei der Berechnung des jeweiligen Z-Wertes auch die Entfernung entlang der Route berücksichtigt. Die Annahme, dass sich alle einem Datenpunkt der durchschnittlichen Trajektorie zugeordneten Datenpunkten auf einer Achse parallel zur Fahrtrichtung befinden, ist für den ersten und letzten Datenpunkt der durchschnittlichen Trajektorie nicht erfüllt. Demnach kann die Anomalieerkennung für diese Datenpunkt nicht als verlässlich beschrieben werden und die Abhängig dieses Verfahrens von der Bestimmung der durchschnittlichen Trajektorie wird deutlich. Bei dem Verfahren, das im folgenden Kapitel vorgestellt wird, besteht diese Abhängigkeit nicht.

4.2 Diskreter euklidischer Raum

Bei diesem Verfahren werden die Trajektoriendaten von kontinuierlichen in diskrete umgewandelt, sodass ein Gitter aus mehreren gleich großen Zellen entsteht. Diese Daten können

für drei Verfahren zur Anomalieerkennung verwendet werden. Da alle Verfahren das Gitter verwenden, wird zu Beginn dieses Kapitels das verwendete Gitter definiert. Anschließend werden die drei Verfahren vorgestellt, auf die Daten angewandt und die detektierten Anomalien analysiert.

4.2.1 Bestimmung der Größe einer Zelle

Durch Runden der Positionswerte auf bestimmte Zahlen werden die diskreten Daten für die vorliegenden X- und Y-Positionen der Trajektorien des zu untersuchenden Datensatzes gebildet. Damit können die vorliegenden kontinuierliche Daten in diskrete überführt werden und es entsteht ein Gitter, welches aus Zellen besteht, in denen mindestens eine Trajektorie detektiert wurde. Die Größe dieser Zellen wird nachfolgend auf Basis unterschiedlicher Anforderungen bestimmt.

Eine dieser Anforderungen wird durch den Anwender definiert, da dieser angibt, ab welcher Größe eine Abweichung als unwahrscheinlich klassifiziert werden soll. Außerdem bestehen Anforderungen an die minimale und maximale Größe. Für die minimale Größe einer Zelle ist relevant, dass der in den Daten enthaltenen Fehler berücksichtigt wird. Aus Abschnitt 3.1.2 ist bekannt, dass die detektierte Position des Verkehrsteilnehmers um 0,25 m von der realen Position abweichen kann. Die maximale Größe einer Zelle wird durch die Anforderung bestimmt, dass die Position des Verkehrsteilnehmers innerhalb der 2,75 m breiten Fahrbahnen unterschieden werden soll. Auf Basis dieser Anforderungen wird die Zellengröße auf 0,5 m mal 0,5 m festgelegt.

Die Umformung der Positionsdaten in dieses Gitter erfolgt durch Runden auf 0,5 m, sodass die Ränder einer Zelle bei den Nachkommastellen 25 und 75 der UTM-Koordinate liegen. Für die zu analysierenden Trajektorien von Westen nach Norden ergeben sich 3.558 unterschiedliche Zellen, durch die mindestens eine Trajektorie verläuft. Dabei kann jede Zelle aus einer X- und einer Y-Koordinate beschrieben werden. Das daraus resultierende Gitter, bestehend aus einzelnen Zellen, wird in Abbildung 4.3 dargestellt.

Nachdem die Datenpunkte des Datensatzes einer Zelle im Gitter zugeordnet worden sind, können die Modelle auf Basis des Trainingsdatensatzes erstellt und anschließend die Testdaten bewertet werden.

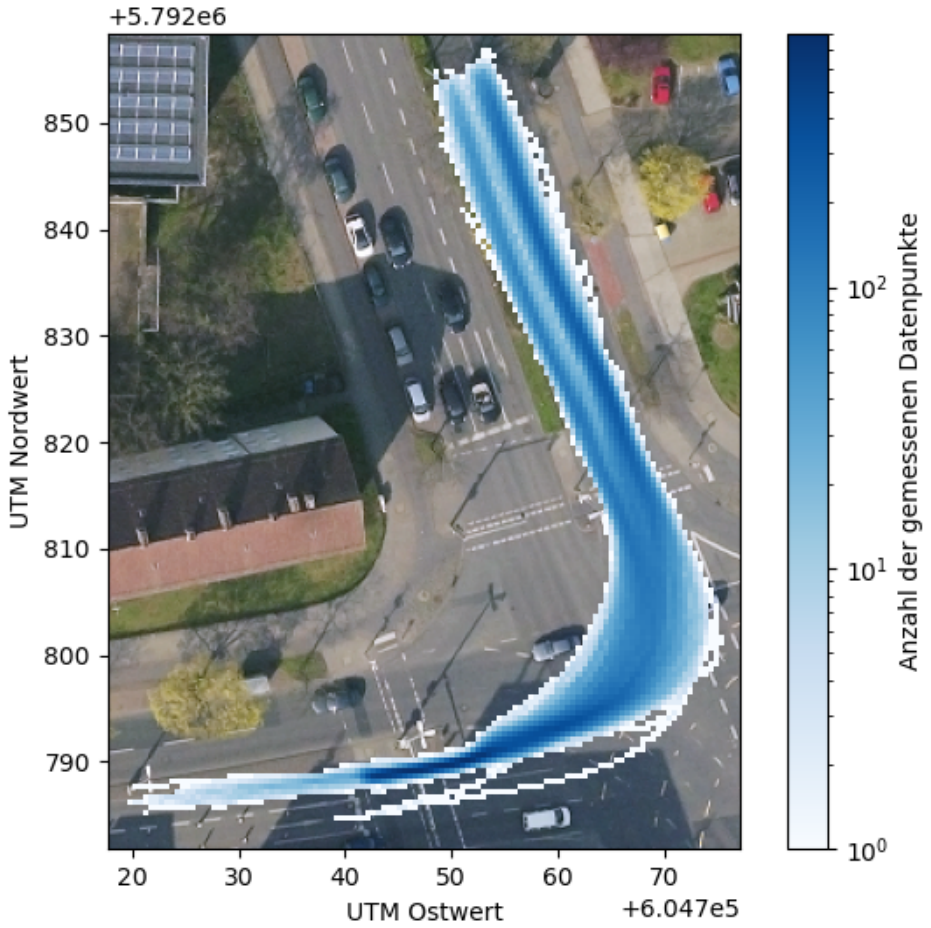


Abbildung 4.3: Darstellung der Trajektorien des Testdatensatzes im diskreten euklidischen Raum. Die Farbe der Zellen gibt die Anzahl der gemessenen Datenpunkte in dieser Zelle an.

4.2.2 Unterscheidung der Modelle

In diesem Kapitel werden drei unterschiedliche Modelle beschrieben, mit denen Anomalien detektiert werden können. Die Unterscheidung der Modelle erfolgt nachfolgend auf Basis der Daten, die für die Anomalieerkennung verwendet werden. Diese Daten sind die Übergangswahrscheinlichkeiten von einer zur nächsten Zelle, die Wahrscheinlichkeit der Bewegungsrichtung in einer Zelle und die Anzahl der unterschiedlichen Trajektorien in einer Zelle. Im Folgenden werden diese Verfahren auf die zu untersuchenden Daten angewandt und die detektierten Anomalien ausgewertet.

4.2.3 Übergangswahrscheinlichkeiten

Für das erste anzuwendende Verfahren werden die Trainingsdaten als eine Markow-Kette, wie in Abschnitt 2.5.2 beschrieben, modelliert. Dazu wird jede Zelle im Gitter als ein Zustand im Zustandsraum der Markow-Kette interpretiert. Die Anzahl der Zellen, in denen mindestens ein Datenpunkt aufgezeichnet wird, stellt demnach die Größe des Zustandsraumes $N = 3.558$ dar. Neben dem Zustandsraum müssen für eine Markow-Kette zudem Übergangswahrscheinlichkeiten berechnet werden. Ein Übergang bezeichnet im vorliegen-

den Fall zwei aufeinanderfolgende diskrete Datenpunkte einer Trajektorie, also den Übergang zwischen zwei Zellen. Jedem Übergang wird für die Erstellung des Modells eine Wahrscheinlichkeit zugeordnet, indem für jeden Zelle untersucht wird, welcher Anteil der Verkehrsteilnehmer sich aus einer Zelle auf die anderen Zellen im jeweils nächsten Zeitpunkt verteilen. Von der Darstellung der Übergangswahrscheinlichkeiten p_{ij} in einer $N \times N$ Matrix $\mathbf{P}$ wird jedoch abgesehen, da nur 13.991 von den möglichen 12.659.364 (N^2) Übergängen im Trainingsdatensatz detektiert werden. Da für 3.558 Zellen 13.991 Übergänge bestehen, kann geschlussfolgert werden, dass sich ein Verkehrsteilnehmer im Durchschnitt aus einer Zelle in eine weitere von 3,93 Zellen bewegt. Alle anderen Zellen wären demnach mit einer null in der Übergangswahrscheinlichkeitsmatrix belegt. Einen besonderen Fall stellen die Zustände dar, in denen nur eine Trajektorie endet. Diese Zustände werden im Zustandsraum berücksichtigt, besitzen aber keine Übergangswahrscheinlichkeit zu einem anderen Zustand. Um die Bedingung des Markow-Modells laut Formel 2.27 einzuhalten, wird eine Übergangswahrscheinlichkeit von Eins für das Verbleiben in dieser Zelle im Modell gespeichert. Der spezielle Fall, in dem ein Verkehrsteilnehmer in einer Zelle verbleibt, wird allerdings wie nachfolgend zu erläutern, aus dem Modell entfernt.

Für die Aufstellung dieses Modells muss beachtet werden, dass stehende Fahrzeuge sehr viele Datenpunkte generieren, deren folgender Datenpunkt in der gleichen Zelle liegt. Daher führt das beschriebene Vorgehen zu einer Markow-Kette, in der Zustände existieren, die mit einer Wahrscheinlichkeit von unter 1 % wieder verlassen werden. Trajektorien von Fahrzeugen, die sich aus einer Zelle bewegen, in der viele stehende Fahrzeuge detektiert wurden, werden demnach mit einer sehr geringen Übergangswahrscheinlichkeit bewertet. Daher werden Datenpunkte, dessen nächster Datenpunkt sich in derselben Zelle befindet, bei der Berechnung des Modells nicht berücksichtigt. Damit wird die Bewegung eines Verkehrsteilnehmers nur auf die Übergänge zwischen unterschiedlichen Zellen beschränkt. Wie in den Abbildungen 18 und 19 im Anhang zu erkennen ist, verbessert sich damit vor allen die Übergangswahrscheinlichkeit von Trajektorien an Positionen, an denen Verkehrsteilnehmer sonst anhalten. Mit dem Entfernen der Datenpunkte von stehenden Fahrzeugen aus dem Trainingsdatensatz von 6.297 Trajektorien der Route WN, verringert sich dessen Größe von 3.664.265 auf 1.106.463 Datenpunkte und die Anzahl der Übergänge von 13.991 auf 11.080. Auf Basis dieser Daten kann die Markow-Kette erstellt werden. Nach Aufstellung dieses Modells mittels den Trainingsdaten, kann allen Übergängen des Testdatensatzes deren Wahrscheinlichkeit laut dem Modell zugeordnet werden (eine Verteilung aller gemessenen Übergangswahrscheinlichkeiten kann im Anhang der Abbildung 20 entnommen werden).

Die Größe des Testdatensatzes von 1.060 Trajektorien verringert sich durch das Entfernen der Datenpunkte von stehenden Fahrzeugen von 658.033 auf 186.045 Datenpunkte. Die diesen Testdaten zugeordneten Übergangswahrscheinlichkeiten werden anschließend auf Anomalien überprüft. Dabei werden alle Übergänge als Anomalie klassifiziert, die geringere Übergangswahrscheinlichkeiten aufweisen als der zuvor definierte Schwellenwert [Han19]. Hierfür wird ein Schwellenwert benötigt, ab dessen Unterschreitung ein Übergang als Anomalie klassifiziert wird. Dieser Schwellenwert kann entweder durch den Anwender festgelegt oder mittels des Z-Wertes bestimmt werden.

Bei diesem Verfahren führt die Verwendung des Z-Wertes zur Ermittlung eines Schwellenwertes zu keinem sinnvollen Ergebnis. Grund dafür ist, dass die Daten zwischen 0 und 1 skaliert sind und mit den Statistiken neben den außergewöhnlich niedrigen auch außergewöhnlich hohe Werte, also Übergangswahrscheinlichkeiten nahe 1 als Anomalie klassifiziert werden würden. Es sollen allerdings nur niedrige Übergangswahrscheinlichkeiten als Anomalie klassifiziert werden. Auch die Orientierung am Anteil der Anomalien bei Anwendung

des Z-Wertes auf eine Normalverteilung führt zu keinem sinnvollen Ergebnis. Unter Verwendung des Z-Wertes werden werden für eine Normalverteilung 0,26 % der Werte als Anomalie klassifiziert. Demnach würden 483 von 186.045 Testdatenpunkten als Anomalie bestimmt werden. Damit wäre der Schwellenwert die 483 niedrigste Übergangswahrscheinlichkeit im Testdatensatz und läge somit bei 0,75 %. Dies würde bedeuten, dass eine Trajektorie, die einen Übergang enthält, der nur einer von 100 Verkehrsteilnehmern aufweist, nicht als anomal bewertet wird.

Da dies nicht dem Verständnis einer Anomalie entspricht, wird die Statistik zur Ermittlung des Schwellenwertes nicht berücksichtigt, sondern der Schwellenwert vom Anwender auf 10 % festgelegt. Damit wird jeder Übergang, der in der jeweiligen Zelle von weniger als jedem zehnten Verkehrsteilnehmer gefahren wird als Anomalie klassifiziert. Mit diesem Schwellenwert werden 3.802 von 10.822 (35,13 %) Übergänge, 12.386 von 186.045 (6,65 %) der Übergänge (siehe Abbildung 21 im Anhang) und alle 1060 Trajektorien des Testdatensatzes als Anomalie eingestuft (eine Trajektorie wird als Anomalie klassifiziert, wenn diese mindestens einen anomalen Wert aufweist). Dieses Ergebnis deutet daraufhin, dass im vorliegenden Datensatz viele geringe Übergangswahrscheinlichkeiten bestehen oder das Modell die Daten unrealistisch repräsentiert. Unter Beachtung der im Abschnitt 4.1.2 ermittelten durchschnittlichen Trajektorie in Abbildung 4.4, wird deutlich, dass auch als normal scheinende Übergänge mit einer geringen Wahrscheinlichkeit bewertet werden.

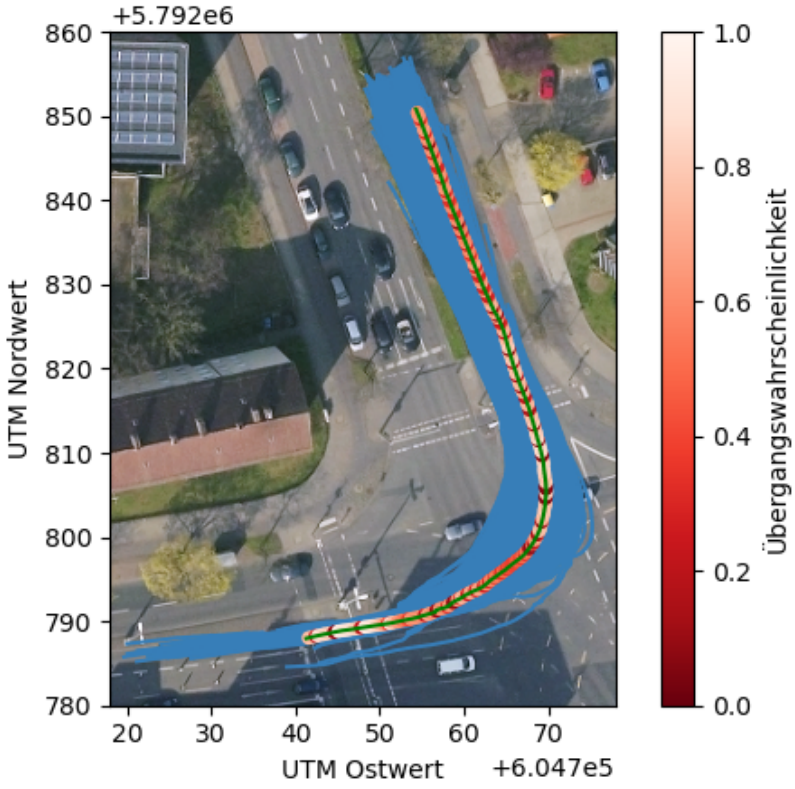


Abbildung 4.4: Bewertung der durchschnittlichen Trajektorie nach der Übergangswahrscheinlichkeit unter Verwendung eines Schwellenwertes von 10 %. Die blauen Linien stellen die Trajektorien aus dem Testdatensatz dar und werden abgebildet, um die rote Linie, die durchschnittliche Trajektorie, mit dem Datensatz vergleichen zu können. Die Farbe der Quadrate um jeden Datenpunkt der Trajektorie stellt die Übergangswahrscheinlichkeit zum nächsten Datenpunkt dar, die nach dem Modell an dieser Position detektiert wurden.

Aus Abbildung 4.4 geht hervor, dass die durchschnittliche Trajektorie zum Beginn und in der Mitte mehrere aufeinanderfolgende Übergangswahrscheinlichkeiten von über 80 % besitzt. Da die Trajektorie für den menschlichen Betrachter keine anomalen Verläufe im Vergleich zum Testdatensatz aufweist, wird erwartet, dass diese Trajektorien deutlich öfter mit einer Übergangswahrscheinlichkeit von über 80 % bewertet wird. Daraus wird geschlussfolgert, dass das Modell die Daten unrealistisch abbildet. Allein in den Bereichen, in denen die Trajektorie parallel zu einer der Achsen verläuft, wird die durchschnittliche Trajektorie mit einer Übergangswahrscheinlichkeit von über 80 % angegeben. Wie in [Han19] beschrieben, resultiert dieses Ergebnis daraus, dass die Einteilung der Kreuzung in ein diskretes Gitter nicht den Verlauf der durchschnittlichen Route berücksichtigt.

Dennoch existieren zwei weitere Ansätze, bei denen auf Basis des diskreten euklidischen Raumes Anomalieerkennung erfolgreich durchzuführen ist. [Ge10] Mit einem dieser Ansätze können dichte- und mit dem anderen richtungsbasierte Anomalien detektiert werden. Da bei dem zuvor gezeigten Modell ebenso die Bewegungsrichtung in Form der Übergangswahrscheinlichkeiten zur Anomalieerkennung verwendet wird, wird nachfolgend der richtungsbasierte Ansatz zuerst vorgestellt. Dabei werden die Trajektorien auf Basis deren Wahrscheinlichkeit der Bewegungsrichtung in jedem Datenpunkt bewertet.

4.2.4 Wahrscheinlichkeit der Bewegungsrichtung

Nach [Ge10] wird bei diesem Verfahren die kontinuierliche Bewegungsrichtung eines Objektes in eine endliche Anzahl von möglichen diskreten Richtungen transformiert. Da bei diesem Vorgehen durch die Diskretisierung ähnliche Ergebnisse wie beim zuvor vorgestellten Verfahren aufgetreten sind, wird dieser Ansatz angepasst. Die Anpassung erfolgt, indem die stetigen Bewegungsrichtungen, die in einer Zelle erfasst wurden, mittels des Z-Wertes auf richtungsbasierte Anomalien überprüft werden.

Für die im Trainingsdatensatz enthaltenen 3.670.562 Datenpunkten wird zur Ermittlung der Z-Werte ein mittlerer Winkel und die Standardabweichung aller in einer Zelle detektierten Bewegungsrichtungen berechnet. Demnach besteht das Modell aus dem durchschnittlichen Winkel und der Standardabweichung für jede Zelle. Zur Anomalieerkennung in den Testdaten werden jedem Datenpunkt für seine Position die dem Modell entsprechenden mittleren Winkel sowie die Standardabweichung zugeordnet. Nun kann für jeden Datenpunkt, der sich in einer Zelle befunden hat, in der auch im Trainingsdatensatz Trajektorien detektiert werden, ein Z-Wert für die Bewegungsrichtung berechnet werden. Anschließend werden Datenpunkte als Anomalie klassifiziert, sobald der Z-Wert der Bewegungsrichtung einen festgelegten Schwellenwert überschreitet.

Für die Zellen von 335 Datenpunkten aus dem Testdatensatz bestehen keine Daten aus dem Trainingsdatensatz, sodass diese nicht nach deren Bewegungsrichtung bewertet werden können und sich der Testdatensatz um deren Anzahl verringert. Weiterhin soll für Zellen, in denen nur wenige Datenpunkte detektiert werden, die Berechnung der Statistik nicht durchgeführt werden. Daher wird diese Methode nur auf Zellen angewandt, in denen mindestens sechs unterschiedliche Verkehrsteilnehmer detektiert werden. Dies ist bei 1.273 von 657.698 Datenpunkten der Fall, sodass der finale Datensatz aus 656.425 Datenpunkten besteht.

Zur Erkennung von Anomalien in diesem Datensatz wird zunächst für den Schwellenwert der Richtwert von drei Standardabweichungen angewendet. Damit werden 25 % der Trajektorien als Anomalie klassifiziert. Da dieser Wert über dem Anteil von 0,26 % an Anomalien in einer Normalverteilung liegt, wird der Schwellenwert auf sechs erhöht. Für diesen Schwel-

lenwert sinkt der Anteil der anomalen Datenpunkte in den Testdaten auf 40 von 656.425 (0,01 %) Datenpunkten und es ergeben sich 6 von 1.060 (0,57 %) anomale Trajektorien. Diese Anomalien werden nachfolgend analysiert. In Abbildung 4.5 werden alle Datenpunkte mit einem Z-Wert größer als sechs dargestellt, die Anwendung eines Schwellenwertes von drei ist in der Abbildung 22 im Anhang dargestellt.

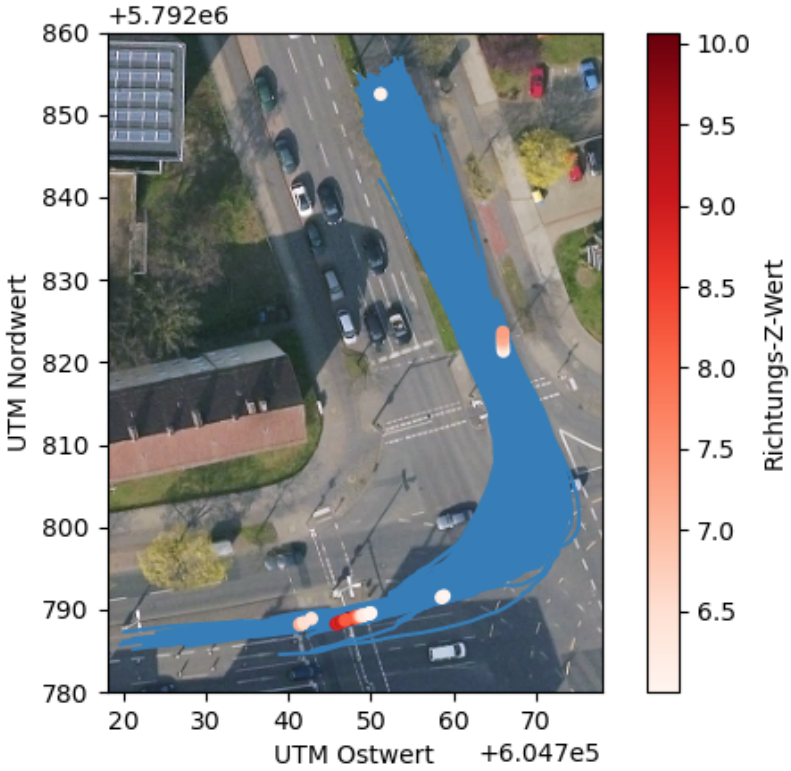


Abbildung 4.5: Richtungsbasierte Anomalien nach der Wahrscheinlichkeit der Bewegungsrichtung unter Verwendung eines Schwellenwertes von sechs Standardabweichungen. Die blauen Linien stellen die Trajektorien des Testdatensatzes dar und werden abgebildet, um die Kreise, welche einen anomalen Datenpunkt darstellen, mit dem Datensatz vergleichen zu können. Die Farbe der Kreise, gibt den Z-Wert der Bewegungsrichtung des Datenpunktes an, der auf Basis der Trainingsdaten berechnet wird.

Nachfolgend werden die in Abbildung 4.5 dargestellten Anomalien danach überprüft, ob diese aus fehlerhafter oder außergewöhnlicher Daten entstanden sind. Dabei werden die neun Trajektorien mit anomalen Werten nach ihrem maximal enthaltenen Anomalie-Wert unterschieden. Zunächst wird die Trajektorie ausgewählt, die den höchsten Anomalie-Wert (10,1) enthält. Die Sichtung des dazugehörigen Videomaterials ergibt, dass diese Trajektorie einen Verkehrsteilnehmern darstellt, der an zweiter Position auf der linken Geradeausfahrerspur vor der LSA stand und sich auf Höhe der Fußgängerfurt entschieden hat nach Norden abzubiegen (siehe Abbildung 23 im Anhang). Dieser Verkehrsteilnehmer wurde mittels der Hausdorff-Metrik allerdings noch nicht identifiziert, da dieser erst beim Spurwechsel detektiert wurde. Demnach wies dieser einen geringen Abstand zur durchschnittlichen Trajektorie auf, besaß allerdings in dieser Position eine von den anderen Verkehrsteilnehmern an dieser Position abweichende Bewegungsrichtung. Gleiches gilt für die Trajektorie mit dem fünfthöchsten Anomalie-Wert von 6,2. Diese ist eine der Trajektorien, die mittels der Hausdorff-Metrik ebenso als Anomalie klassifiziert wurden. Beide Anomalien sind wegen

außergewöhnlichen Daten als solche detektiert worden

Die Trajektorie mit dem zweithöchsten Anomalie-Wert von 7,4 wurde um 02:48 Uhr aufgezeichnet. Aus dem Referenzvideo geht hervor, dass die Position und damit die Bewegungsrichtung des Verkehrsteilnehmers fehlerhaft bestimmt werden (siehe Abbildung 24 im Anhang). Damit ist diese Anomalie auf die nicht verlässliche Positionsbestimmung bei Dunkelheit zurückzuführen.

Die Trajektorie mit dem nächsthöchsten Anomalie-Wert von 7,2 enthält sechs anomale Werte, welche die ersten Werte der Trajektorie darstellen. Nach den aufgezeichneten Werten, ändert sich die Bewegungsrichtung des Verkehrsteilnehmers in diesen sechs Werten von 29° auf 21°. Diese Änderung ist allerdings nicht in den Videoaufzeichnungen zu erkennen (siehe Abbildung 25 im Anhang). Gleiches gilt für die Trajektorien mit dem vierthöchsten und sechsthöchsten Anomalie-Wert von 6,4 bzw. 6,1 welche beim vierten bzw. drittletzten Datenpunkt der Trajektorie gemessen werden. Daher wird vermutet, dass die Bestimmung der Bewegungsrichtung zu Beginn und zum Ende der Detektion fehleranfälliger ist. Es ist bekannt, dass wie in Abschnitt 3.1.2 beschrieben, die Verlässlichkeit der Bewegungsrichtung mit sinkender Geschwindigkeit abnimmt. Da die Geschwindigkeit bei den Anomalien, die in den ersten sechs bzw. dem vierten Wert der Trajektorien auftreten, gering ist, kann dies ein weiterer Grund für die fehlerhaften Daten sein.

Nach der Analyse der Anomalien wird abschließend die Qualität des Verfahrens bewertet, indem das Ergebnis mit dem des zuvor angewendeten Verfahrens verglichen wird.

Unter der Verwendung eines Schwellenwertes von vier wird erneut der als PKW klassifizierte Motorradfahrer C 714, der auch mithilfe Hausdorff-Metrik in Abschnitt 4.1.4 als Anomalie erkannt wird und in Abbildung 16 im Anhang dargestellt ist, als Anomalie klassifiziert. Nach der Bewegungsrichtung wird dieser Verkehrsteilnehmer beim Warten vor der Überquerung der entgegenkommenden Fahrbahn als Anomalie klassifiziert, da hier die Geschwindigkeit auf 0 km/h abfällt und der Verkehrsteilnehmer über 10 Sekunden eine anomale Bewegungsrichtung beibehält, mit der dieser vor dem Stillstand detektiert wurde. Damit wird deutlich, dass die Bewegungsrichtung der Verkehrsteilnehmer robust gegen die geringen Positionsänderungen ist, die teilweise beim Stillstand eines Objektes detektiert werden.

Der Abbildung 26 im Anhang wird die durchschnittliche Trajektorie nach der Bewegungsrichtung bewertet. Der für diese Trajektorie ermittelte maximale Z-Wert liegt bei zwei Standardabweichungen, sodass diese Trajektorie nicht als Anomalie klassifiziert wird. Beim nachfolgend vorgestellten dritten Modell wird die durchschnittliche Trajektorie aus Abschnitt 4.1.2 ebenso wenig als Anomalie erkannt.

4.2.5 Anzahl unterschiedlicher Trajektorien

Bei diesem Verfahren werden dichtebasierte Anomalien anhand der Anzahl der unterschiedlichen Trajektorien in einer Zelle ermittelt. Für das aufzustellende Modell wird demnach bestimmt, wie viele unterschiedliche Trajektorien in jeder Zelle des Gitters erfasst werden. Das Modell besteht folglich aus der Anzahl der unterschiedlichen Trajektorien für jede der unterschiedlichen 3.558 Zellen in von 6.297 Trajektorien und 3.670.562 Datenpunkten des Trainingsdatensatzes. [Ge10]

Der für die Anomalieerkennung benötigte Schwellenwert gibt an, wie viele unterschiedliche Verkehrsteilnehmer mindestens in einer Zelle detektiert worden sein müssen, damit ein Datenpunkt in dieser Zelle nicht als Anomalie klassifiziert wird. Unter der Annahme, dass wie bei der Anomalieerkennung in einer Normalverteilung mittels des Z-Wertes, ungefähr

0,26 % der Werte als Anomalie klassifiziert werden sollen, muss der Schwellenwert im vorliegenden Fall auf sechs gesetzt werden. Damit werden für die zu analysierende Route im Testdatensatz 903 von 3.494 (25,84 %) Zellen, 1.636 von 658.033 (0,25 %) Datenpunkten und 40 von 1.060 (3,77 %) Trajektorien als Anomalie klassifiziert. Alle anomalen Datenpunkte werden in der Abbildung 4.6 dargestellt.

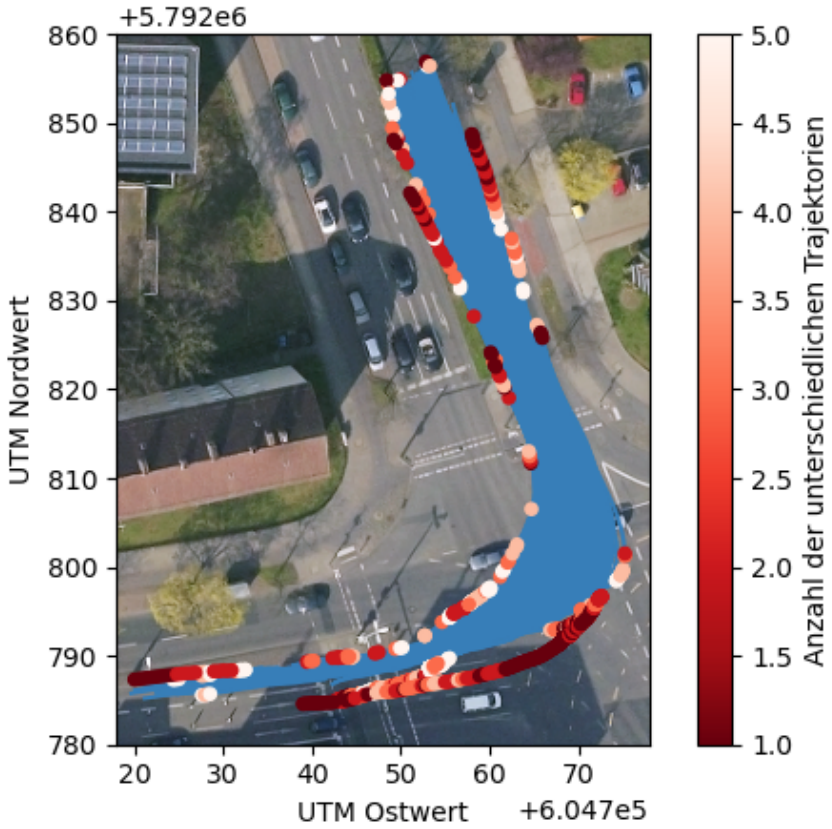


Abbildung 4.6: Dichtebasierte Anomalie auf Basis der Anzahl der unterschiedlichen Trajektorien und eines Schwellenwertes von sechs. Die blauen Linien stellen die Trajektorien aus dem Testdatensatz dar und werden abgebildet, um die Kreise, welche einen anomalen Datenpunkt anzeigen, mit dem Datensatz vergleichen zu können. Die Farbe der Kreise, gibt die Anzahl der in dieser Zelle im Trainingsdatensatz detektierten unterschiedlichen Trajektorien an.

Mit diesem Verfahren werden Verkehrsteilnehmer, die sich an einer Position befinden, an der die Anzahl der im Trainingsdatensatz detektierten Verkehrsteilnehmer einen Schwellenwert unterschreitet, als Anomalie eingestuft. Der Abbildung ist zu entnehmen, dass Datenpunkte am Rand des Testdatensatzes als Anomalie klassifiziert werden. Die Anomalien, die aus außergewöhnlichen Daten entstanden sind, werden nachfolgend vorgestellt. Von den 40 anomalen Trajektorien stellen zwei Trajektorien den Bewegungsverlauf von Verkehrsteilnehmern dar, die von der linken Geradeausfahrerspur nach Norden abbiegen. Ein weiterer Verkehrsteilnehmer wird aufgrund seines späten Einbiegens in die Fahrbahn nach Norden als Anomalie klassifiziert. Außerdem wird der mit den beiden vorherigen Verfahren detektierte Motoradfahrer C 714 als anomal bewertet. Da sich die anomalen Zellen am Rand des Testdatensatzes befinden und keine Anomalien nahe der zuvor ermittelten durchschnittlichen Trajektorie detektiert werden, wird geschlussfolgert, dass die Größe des Rasters ausreichend groß gewählt wurde. Die übrigen dichtebasierten Anomalien entstehen dadurch,

dass in einzelnen Zellen am Rand des Trainingsdatensatzes weniger als sechs Verkehrsteilnehmer erfasst werden. Dies folgt daraus, dass der kontinuierliche Raum in einen diskreten Raum aufgeteilt wurde. Von einem solchen Gitter ist das im nächsten Abschnitt vorgestellte Verfahren unabhängig.

4.3 Gaußsches Mischmodell

Nachfolgend wird das Verfahren des Gaußschen Mischmodells, dessen Theorie in Abschnitt 2.5.3 beschrieben wird, auf die vorliegenden Trajektoriendaten angewandt. Dazu wird als erstes der Ablauf des Verfahrens beschrieben, bevor die Trajektorien als mehrdimensionale Datenpunkte abstrahiert werden. Anschließend werden die Trajektorien aus dem Trainingsdatensatz durch Anwendung des Gaußschen Mischmodells nach deren Routen gruppiert. Das dabei gelernte Modell wird daraufhin auf den Testdatensatz angewandt und die detektierten Anomalien werden analysiert.

4.3.1 Ablauf des Verfahrens

Wie im Abschnitt 3.2.7 beschrieben, können Polygone dazu eingesetzt werden, um Trajektorien nach deren Routen zu klassifizieren. Eine weitere Methode ist die Modellierung der Daten als Gaußsches Mischmodell. Hier stellen anstatt der Trajektorien, deren Abstraktionen als dreidimensionale Datenpunkte im euklidischen Raum die zu untersuchenden Daten dar. Nachdem diese Abstraktionen erfolgt ist, können die Daten als Gaußsches Mischmodell modelliert werden. Es wird erwartet, dass die Repräsentationen der Trajektorien einer Route, mittels dem Gaußschen Mischmodell einer Komponente zugeordnet werden, da diese ähnliche Eigenschaften besitzen. Trajektorien, deren Repräsentation im euklidischen Raum einen geringen Abstand zu Repräsentationen anderen Trajektorien haben, werden als ähnlich betrachtet und einer Komponente zugeordnet [Eck10, S. 428]. Damit werden die Trajektorien nach zwei Methoden nach deren Route klassifiziert. Für die Anomalieerkennung wird untersucht, welche Trajektorien nach den Methoden unterschiedlich klassifiziert werden. Dabei wird die Klassifizierung durch die Polygone in dieser Arbeit als richtig angenommen. Sollte eine Trajektorie nach dem Gaußschen Mischmodell einer anderen Route zugeordnet werden, als für die Trajektorie mittels der Polygone bestimmt wurde, wird diese Trajektorie als Anomalie klassifiziert.

4.3.2 Abstraktion der Trajektorien

Wie im Abschnitt 3.2.7 beschrieben, wurden die Trajektorien nach ihren Routen mittels Polygonen klassifiziert. Ebenso können die Trajektorien mittels Gaußschen Mischmodellen nach Routen unterschieden werden. Dazu wird jede Trajektorie als mehrdimensionaler Datenpunkt abstrahiert. Da bisher die Positionsdaten zur Anomalieerkennung verwendet wurden, werden diese Daten für jede Trajektorie abstrahiert, indem das arithmetische Mittel berechnet wird. Jede Trajektorie wird also als zweidimensionaler Datenpunkt aus der durchschnittlichen X- und Y-Position dargestellt. Da diese Merkmale, wie in Abbildung 4.7 zu erkennen, keine eindeutige Gruppierung der Datenpunkte nach den jeweiligen Routen ermöglichen, wird den Datenpunkten eine weitere Dimension hinzugefügt.

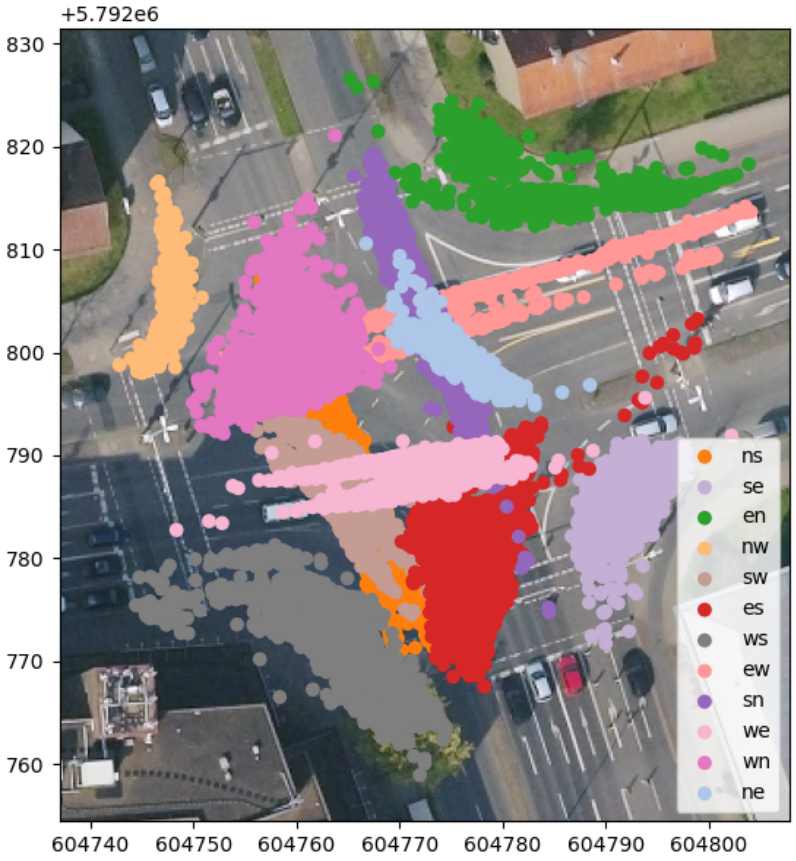


Abbildung 4.7: Ergebnis der Klassifizierung mit Polygonen von den abstrahierten Trajektorien im zweidimensionalen Raum. Die Farbe gibt die Route an, der die Trajektorie zugeordnet ist. Die Abkürzung „ns“ in der Legende steht für die Route von Norden nach Süden. Nach gleichem Schema wurden die anderen Routen mit zwei Buchstaben abgekürzt.

Um die Daten gruppieren zu können, müssen vor allem die bisher ähnlichen Daten in der dritten Dimension unterschiedlich sein. Ein Merkmal, dass eine solche Anforderung erfüllt ist der Winkel des Richtungsvektors zwischen dem zuerst und zuletzt aufgezeichneten Datenpunkt einer Trajektorie. Es wird deutlich, dass zwei entgegenkommende geradeausfahrende PKWs nach diesem Merkmal maximal unterschiedlich bewertet werden. Dies tritt in den vorliegenden Daten zum Beispiel bei den Routen WE und EW ein. Allerdings ist noch wichtiger, dass dieses Merkmal in Abbildung 4.7 ähnliche Datenpunkte unterscheidet. Dies trifft beispielsweise auf geradeausfahrende und abbiegende Verkehrsteilnehmer der Routen NW und SN zu.

Dieses Merkmal weißt allerdings auch ähnliche Werte für zwei Arten von Routen auf. Jeder Linksabbieger der aus einer Richtung kommt, hat einen ähnlichen Richtungsvektor wie der Rechtsabbieger, der aus der Richtung rechts neben dem Linksabbieger kam. Ein Beispiel dafür stellen die Routen WN und SE dar. Die Routen dieser Trajektorien sollten allerdings dennoch differenzierbar sein, da sich deren durchschnittliche Positionen unterscheiden.

Die zweite Art von Routenpaarungen, die gleiche Werte aufweisen können, sind die Geradeausfahrer mit den Verkehrsteilnehmern, die auf der links daneben liegenden Kreuzungseinfahrt einen Fahrtrichtungswechsel vollziehen. Ein Beispiel für diese Routenpaarung sind die Routen SN und WW. Hier ist allerdings wieder zu erwarten, dass sich die durchschnittli-

che Position der beiden Routen unterscheiden lässt.

Nachdem nun die Trajektorien als dreidimensionaler Datenpunkt abstrahiert wurden, können diese Datenpunkte mit dem GMM in Gruppen eingeteilt werden.

4.3.3 Erstellung des Modells

Die Anwendung von GMM bedeutet, mehrere Gaußsche Verteilungen mit Hilfe des EM-Algorithmus an die Daten anzupassen, sodass diese die Daten optimal repräsentieren. Dafür muss festgelegt werden, welche Daten verwendet und wie die Kenngrößen für den EM-Algorithmus definiert werden. Die benötigten Kenngrößen sind die Anzahl der erwarteten Komponenten, die Art der Initialisierung der Parameter, die Anzahl der Durchführungen und der Schwellenwert, der zur Terminierung des Algorithmus verwendet wird. im Folgenden wird vorgestellt welche Daten und wie die Kenngrößen für das GMM gewählt werden. Abschließend wird das Ergebnis der Anwendung des Algorithmus bewertet.

Da zur Anomalieerkennung das Ergebnis der Gaußschen Mischmodelle mit dem Ergebnis der Polygon-Klassifizierung verglichen wird, können für das Gaußsche Mischmodell nur Trajektorien verwendet werden, denen mit den Polygonen eine Route zugeordnet wurde. Von den 276.822 Trajektorien aus dem Trainingsdatensatz werden 259.603 einer Route zugewiesen. Die Anwendung der GMM auf diesen Datensatz ergibt allerdings, dass die Routen der Fahrtrichtungswechsel nicht als einzelne Gruppen identifiziert werden. Stattdessen werden die Routen NS, WS, ES und EW in jeweils zwei Gruppen unterteilt. Daher wird der Datensatz für die GMM um die Routen verringert, die einen Fahrtrichtungswechsel beschreiben. Damit entfallen insgesamt 606 Trajektorien, sodass der finale Datensatz für die Anwendung der GMM aus 258.997 abstrahierten Trajektorien besteht.

Damit kann die erste Kenngrößen für den EM-Algorithmus, die Anzahl der zu erwarteten Komponenten, festgelegt werden. Diese entspricht im vorliegenden Fall Zwölf und ergibt sich aus der Anzahl aller Routen (16) vermindert um die vier Routen von Fahrtrichtungswechseln.

Die zweite Kenngröße für die Durchführung des EM-Algorithmus ist die Initialisierung der Parameter. Wie im Abschnitt 2.5.3 erläutert, werden die zu erlernenden Parameter $\mathbf{m}$, $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ zufällig initialisiert. Dieses Vorgehen kann verbessert werden, indem die Parameter durch den k-Means-Algorithmus festgelegt werden. Bei diesem Algorithmus werden auch die Erwartungs- und Maximierungsschritte durchlaufen, mit dem Unterschied, dass die zu optimierenden Parameter nur die Mittelpunkte der Komponenten sind und die Verteilung der Daten nicht als Gauß-Verteilung dargestellt wird. Eine genaue Beschreibung des k-Means-Algorithmus kann der Referenz [Bis06, S. 424 ff.] entnommen werden. Mit dieser Initialisierung, kann das Ergebnis des Gaußschen Mischmodells deutlich verbessert werden. Zudem ergeben sich Verbesserungen der benötigten Rechenzeit. Der k-Means-Algorithmus erfordert jedoch neben der Angabe der erwarteten Komponenten ebenso eine Initialisierung der Mittelpunkte. Diese erfolgt im vorliegenden Fall zufällig, da keine Informationen über die zu erwartenden Mittelpunkte vorliegen.

Der dritte Parameter, der festgelegt werden muss, ist die Anzahl der Durchführungen des Algorithmus. Aufgrund der im vorherigen Abschnitt beschriebenen zufälligen Initialisierung der Parameter, kann der EM-Algorithmus für die gleichen Daten ein unterschiedliches Ergebnis liefern. Grund dafür ist, dass der Algorithmus bei einem lokalen Maximum der Zielfunktion terminiert ist, das globale Maximum aber nicht erreicht hat. Da daher die erneute Anwendung des Algorithmus zu einem besseren Ergebnis führen kann, wird der Algorithmus im vorliegenden Fall fünfmal durchgeführt. Der Wert, bei dem die Zielfunktion am gerings-

ten ist, wird anschließend als finales Ergebnis ausgewählt.

Als letzter Parameter wird der Schwellenwert für die Terminierung des Algorithmus festgelegt. Für den Schwellenwert von Eins terminiert der Algorithmus nach zwei Iterationen. Die geringe Anzahl an benötigten Iterationen lässt vermuten, dass mit dem zur Initialisierung der Parameter verwendeten k-Means-Algorithmus die Daten wie ein GMM gruppiert werden. Es wird erwartet, dass mit einem geringeren Schwellenwert ein besseres Ergebnis der Zielfunktion erreicht werden kann. Daher wird für die Bestimmung des finalen Schwellenwertes der Algorithmus mit den Werten 0,1, 0,01, 0,001 und 0,0001 erneut durchgeführt und anschließend der Schwellenwert gewählt, bei dem der höchste Wert der Zielfunktion erreicht wird. Dieser ergibt sich im untersuchten Fall nach fünf Iterationen für einen Schwellenwert von 0,001. Da dies nicht der geringste Schwellenwert ist, der getestet wird, wird nicht davon ausgegangen, dass mit einem geringeren Schwellenwert ein besseres Ergebnis erzielt werden kann. Daher wird der Wert 0,001 als Schwellenwert für die Terminierung des final auszuführenden Algorithmus festgelegt.

Damit sind alle Parameter für die Anwendung des Algorithmus bestimmt. Die Gruppierung der 40.770 Trajektorien aus dem Testdatensatz auf Basis des Modells ergibt, dass zehn Trajektorien einer unterschiedlichen Route als mithilfe der Polygone zugeordnet werden. Das Ergebnis der Klassifizierung mittels den Gaußschen Mischmodellen ist in Abbildung 4.8 dargestellt. Dabei sind die Trajektorien entsprechend ihrer zugewiesenen Route nach den GMM farblich markiert. Im Vergleich dazu sind die Datenpunkte in der Abbildung 27 im Anhang nach deren mit den Polygonen ermittelten Route farblich unterschieden. Die Unterschiede dieser Klassifizierungen werden in Abbildung 28 im Anhang dargestellt, indem die Trajektorien entsprechend des Anomalie-Wertes farblich markiert sind. Die detektierten Anomalien werden im folgenden Abschnitt analysiert.

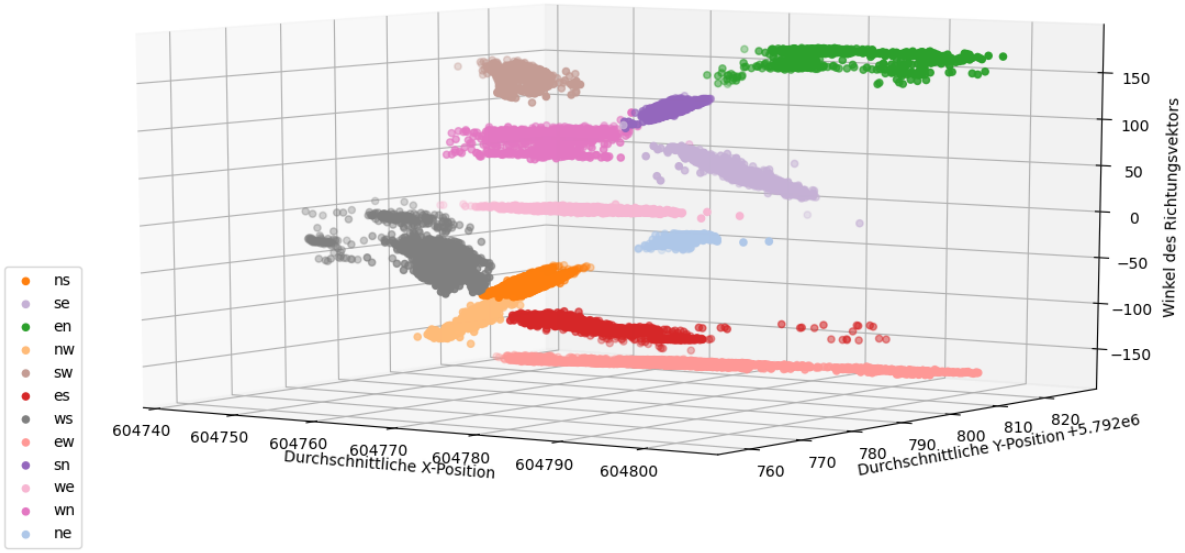


Abbildung 4.8: Ergebnis der Klassifizierung mit Gaußschen Mischmodellen von abstrahierten Trajektorien im dreidimensionalen Raum. Die Zuordnung zwischen Farben und Routen entspricht der Zuordnung in Abbildung 4.7.

4.3.4 Analyse der Anomalien

Trotz zufälliger Initialisierung der Komponentenmittelpunkte wird das soeben vorgestellte Ergebnis der Klassifizierung bei mehreren Durchführungen immer wieder erreicht. Daher werden die detektierten Anomalien nachfolgend näher untersucht.

Die identifizierten Anomalien lassen sich in zwei Gruppen einteilen. Fünf der Anomalien treten bei der Route SN und fünf Anomalien bei der Route WE auf. In der Abbildungen 4.9 sind die anomalen Trajektorien von der Route WE auf der Forschungskreuzung dargestellt.

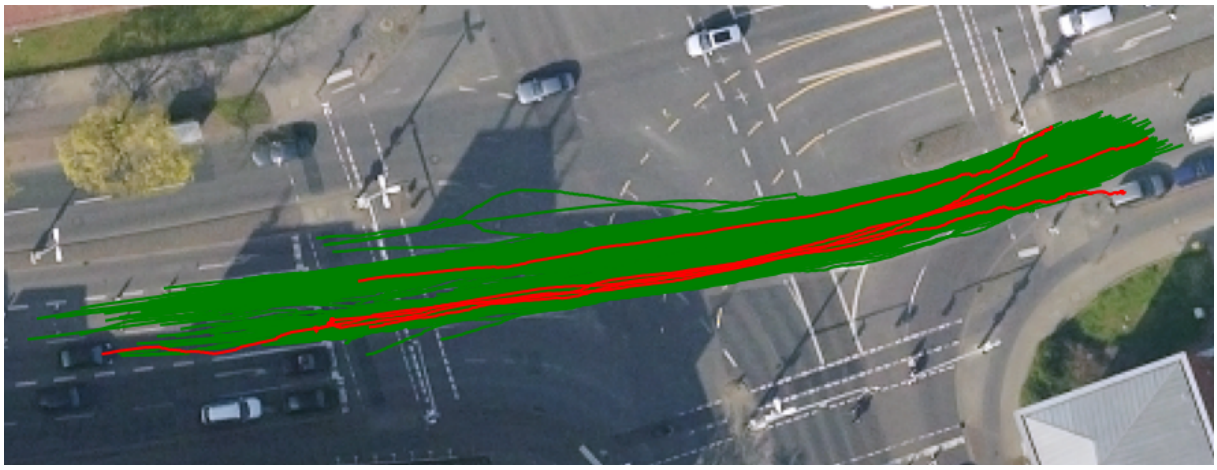


Abbildung 4.9: Trajektorien von Westen nach Osten. Grün: Trajektorien, die mit beiden Verfahren der Route zugeordnet werden. Rot: Trajektorien, die mit den beiden Verfahren unterschiedlichen Routen zugeordnet werden.

Nachfolgend werden diese Anomalien zuerst analysiert. Von den detektierten anomalen Trajektorien werden zwei TJ der Route SE und drei TJ der Route WS zugeordnet. Da im Gaußschen Mischmodell nur die durchschnittlichen Positionswerte und der Winkel des Richtungsvektors berücksichtigt wurden, können nur Unterschiede in diesen Werte zu einer fehlerhaften Klassifizierung geführt haben. Um herauszufinden wieso ein Datenpunkt nach dem GMM als Anomalie klassifiziert wurde, werden die Z-Werte für die jeweilige Variable berechnet. Daraus ergibt sich, dass bei allen Anomalien die Positionswerte hohe Abweichungen aufweisen, der Winkel des Richtungsvektors allerdings nicht. Es wird vermutet, dass diese Abweichungen in den durchschnittlichen Positionswerten daraus resultieren, dass Verkehrsteilnehmer eine lange Zeit gestanden haben. Durch die Aufzeichnung vieler Positionswerte an einer Position verschiebt sich der Mittelwert der Positionswerte zu dieser Position. Für die vorliegende Route werden damit vor allem Abweichungen in der X-Position für die Klassifizierung als Anomalie verantwortlich.

Zuerst wird das Ergebnis der Sichtung des Videomaterials für die fünf detektierten Anomalien der Route WE, die der Route WS bzw. SW zugeordnet werden, vorgestellt. In den Videos wird deutlich, dass zwei Objekte beim Stehen an der roten Ampel detektiert werden und sich damit eine deutlich geringere X-Position ergibt. Damit sind diese Anomalien durch außergewöhnliche Daten entstanden und werden fälschlicherweise der Route WS zugeordnet. Die dritte Anomalie der Route WS stellt einen Verkehrsteilnehmer dar, der außergewöhnlich früh nicht mehr detektiert wird, wodurch sich die durchschnittliche X-Position verringert. Höhere durchschnittliche X-Positionen ergeben sich für die zwei Verkehrsteilnehmer, die der Route SE zugeordnet werden, da diese am Ende der Route beim Stehen detektiert werden. Eine dieser Anomalien ist durch fehlerhafte Daten begründet, da das Objekt

detektiert wird, obwohl der Verkehrsteilnehmer in der Realität die Position verlassen hat. Die zweite Anomalie stellt allerdings einen Verkehrsteilnehmer dar, der sofort hinter der Kreuzung geparkt hat, dabei weiterhin detektiert wird und somit außergewöhnliche Daten erzeugt hat.

Des Weiteren wurden die Videoaufzeichnungen für die Anomalien der Route SN analysiert, die in Abbildung 4.10 dargestellt sind. Von den fünf detektierten anomalen Trajektorien wird eine Trajektorie der Route SE und vier Trajektorien der Route WN zugeordnet.

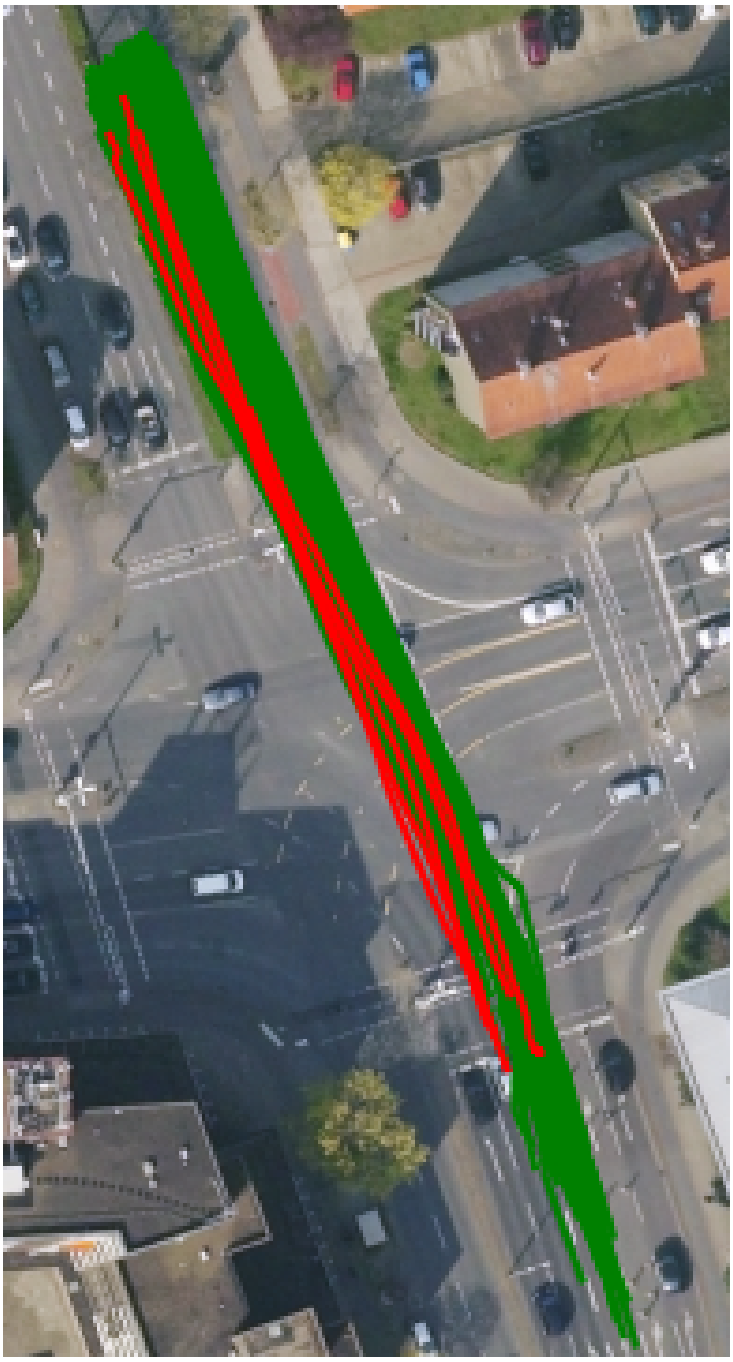


Abbildung 4.10: Trajektorien von Süden nach Norden. Grün: Trajektorien, die mit beiden Verfahren der Route zugeordnet werden. Rot: Trajektorien, die mit den beiden Verfahren unterschiedlichen Routen zugeordnet werden.

Die Analyse dieser Trajektorien ergibt, dass die vier Anomalien der Route WN Fahrzeu-

gen zugeordnet werden, die von einer Spur für Linksabbieger geradeaus gefahren sind. Bei einem solchen Routenverlauf werden die Datenpunkte weiter links als eine durchschnittliche Trajektorie detektiert, sodass die Abweichungen in der durchschnittlichen X-Position für die Klassifizierung als Anomalie verantwortlich sind. Die verbleibende Anomalie der Route SE stammt von einem Verkehrsteilnehmer, der im Stehen an der LSA detektiert wird und damit deutliche Abweichungen in der durchschnittlichen Y-Position aufweist.

Ein weiterer Einflussfaktor auf die Klassifizierung der Trajektorien ist die Nähe einer Routengruppe zur anderen Komponente. Liegt eine Komponente bezüglich einer bestimmten Variable näher an einer anderen, wird eine Abweichung bezüglich dieser Variable eher in einer falschen Klassifizierung resultieren, als wenn ein Datenpunkt eine Abweichung in einer Dimension aufweist, in der keine andere Gruppe besteht. Mit dem vorliegenden Verfahren werden daher nicht die für eine Komponente untypischen Werte ermittelt. Vielmehr wird berücksichtigt, mit welcher Wahrscheinlichkeit die Werte von anderen Komponenten generiert werden. Für die Fragestellung, welche Trajektorie in Bezug auf deren Route eine Anomalie dargestellt, müssten die Trajektorien einer Route isoliert betrachtet werden.

Insgesamt wird für dieses Verfahren festgestellt, dass die Trajektorien in einem Maße reduziert werden, dass bei der Interpretation der Ergebnisse viele Ursachen berücksichtigt werden müssen. Weiterhin beeinflusst die Auswahl der Merkmale, nach denen die Trajektorien abstrahiert werden, das Ergebnis. Gleiches gilt für die Anzahl der Trajektorien pro Route. Da die Fahrtrichtungswechsel im Vergleich zu den anderen Routen sehr selten befahren werden, werden diese mittels Gaußschen Mischmodellen nicht als Route identifiziert. Ein Vorteil des Verfahrens ist, dass die Zusammenhänge zwischen Routen näher untersucht werden können. Durch die Darstellung der abstrahierten Trajektorien werden Cluster erkannt, die den Vergleich der Routen untereinander ermöglichen. Als weiterer Vorteil dieses Verfahrens kann die Verringerung des Rechenaufwandes durch Abstrahierung der Daten angeführt werden. Allerdings besteht der Nachteil, dass abnormale Subtrajektorien bei dieser Methode nicht detektiert werden, da diese nach der Berechnung des Durchschnitts der Trajektoriendaten nicht mehr identifiziert werden können.

Damit wurden alle ausgewählten Verfahren auf den Datensatz angewandt und die detektierten Anomalien analysiert. Nachfolgend wird ein Ausblick gegeben, welchen zukünftigen Fragestellungen auf Basis der Ergebnisse dieser Arbeit nachgegangen werden kann.

Kapitel 5

Ausblick

Nachdem mit den anzuwendenden Verfahren die zu untersuchenden Daten nach Anomalien überprüft wurden, wird nachfolgend ein Ausblick gegeben, wie im Rahmen von weiterführender Forschung auf den Ergebnissen dieser Arbeit aufgebaut werden kann.

Zunächst wird vorgeschlagen, die in Abschnitt 1.3 für die Abgrenzung des Umfangs dieser Arbeit notwendigen Einschränkungen auszuweiten. Dabei können die Verfahren auf einen größeren Datensatz von Wochen oder Monaten angewandt werden, um weitere außergewöhnliche Daten zu detektieren. Darüber hinaus ist es möglich, die angewandten Verfahren, mit denen die Route WN und Objektklasse PKW untersucht werden, auf weitere Routen und Objektklassen auszuweiten. Ebenso können spezifische Fragestellungen für andere Merkmale eines Datenpunktes wie zum Beispiel die Geschwindigkeit beantwortet werden, indem diese anstelle der Positionsdaten oder zusätzlich zu diesen in den Verfahren als weitere Dimension berücksichtigt werden. Bei allen drei vorgestellten Verfahren können weitere Merkmale der Trajektorie problemlos berücksichtigt werden. Beim Hausdorff-Abstand bildet die Datengrundlage für die Anomalieerkennung die Abstandsmessung mittels der euklidischen Distanz. Diese kann nicht nur im zweidimensionalen angewendet werden, sondern auch die Geschwindigkeit und Beschleunigung oder Zeit berücksichtigen. Auch bei dem diskreten euklidischen Raum könnten weitere Merkmale berücksichtigt werden, indem der Raum möglicher Zustände von dem derzeitigen zweidimensionalen Raster in ein Gitter einer höheren Dimension umgewandelt wird. Eine weitere Dimension kann die Geschwindigkeit in Form einer diskreten Größe darstellen. Gleiches gilt für das Gaußsche Mischmodell, indem für die dreidimensionale Abstrahierung der Trajektorien andere als die in dieser Arbeit untersuchten Merkmale verwendet werden können.

Ein Merkmal, dass die Detektion von vor roten LSA stehenden Fahrzeugen verhindern kann, ist die aktuelle Schaltung der LSA. Außerdem können Verkehrsteilnehmer, die als erstes nach einer Rotphase die Kreuzung befahren, von denen unterschieden werden, die kurz vor Schaltung der LSA von Rot auf Gelb in die Kreuzung einfahren. Es wird vermutet, dass dadurch das erwartete Verhalten eines Verkehrsteilnehmers besser modelliert werden könnte. Neben der Berücksichtigung der LSA kann eine Korrelation zwischen dem Auftreten von Anomalien und der Verkehrsstärke, Tageszeit, Wochentagen oder Witterungsbedingungen untersucht werden. Dazu können diese Parameter entweder als Merkmale der Trajektorie hinzugefügt oder bei der Analyse der Anomalien berücksichtigt werden.

Für die Verbesserung der Qualität der Ergebnisse können weitere Methoden zur Datenvorbereitung entwickelt werden, da die Arten von identifizierten fehlerhaften Daten nicht zwingend allumfassend ist.

Auch die Anpassung des Vorgehens zur Bestimmung eines Schwellenwertes kann die

Qualität der Ergebnisse steigern. Zum einen kann eine andere Statistik als der Z-Wert verwendet werden. Nach [Yan19] wird neben dem Z-Wert der Interquartilsabstand („interquartile range“) häufig für die Anomalieerkennung genutzt. Zum anderen kann das Verfahren zur Anwendung der Statistik angepasst werden. Da der Z-Wert aus dem Datensatz berechnet wird, der die Anomalien enthält, kann eine Verzerrung der Statistik entstehen, wodurch das Ergebnis an Genauigkeit verliert. Um diese Gefahr der Verzerrung zu reduzieren, wird der Ansatz empfohlen, das Verfahren zwei Mal durchzuführen. Beim zweiten Durchlauf werden die zuvor detektieren Anomalien aus dem Datensatz entfernt, sodass letztendlich eine genauere Statistik erstellt werden kann. [Yan19]

Alle genannten Ergänzungen würden auf die vorliegenden Daten angewandt werden. Es besteht die Möglichkeit, makroskopische Verkehrsdaten wie die Verkehrsstärke auf Anomalien zu überprüfen. Hierzu würden alle Trajektorien pro Zeiteinheit gezählt werden, wobei diese nach der Objektart unterschieden werden können. Anschließend würde erneut eine zu erwartende durchschnittliche Verteilung der Objektarten über den Tagesverlauf erstellt werden, um zu untersuchen, inwiefern eine bestimmte Verkehrsstärke einer bestimmten Objektart zu einer bestimmten Tageszeit eine Anomalie darstellt. Als weitere Datengrundlage für die Anomalieerkennung können die Videobilder verwendet werden, sodass sich zukünftige Arbeiten mit der Anomalieerkennung in Videoaufzeichnungen befassen können [Ahm18]. Hierzu werden andere Verfahren wie zum Beispiel neuronale Netze wie das „Convolutional Long Short-Term Memory“ verwendet [Gao17]. Dieses Verfahren wird empfohlen, da auf dem Gebiet der Anomalieerkennung mit dem Verfahren häufig höhere Genauigkeiten als mit anderen Verfahren erzielt werden konnten [Ma18, S. 7].

Des Weiteren stellen Anomalien Unregelmäßigkeiten in dem Informationsgehalt von Daten dar, sodass diese auch mittels der Entropie gemessen werden können. Denn die Entropie, als Kenngröße aus der Informationstheorie, gibt für eine Verteilungen die Unsicherheit über dessen Informationsgehalt an. [Cha09, S. 15:35]

Damit konnte gezeigt werden, dass diverse weitere Verfahren zur Anomalieerkennung bestehen und dass auf den Ergebnissen dieser Arbeit aufgebaut werden kann. Eine Zusammenfassung der wesentlichen Ergebnisse und die Beantwortung der Forschungsfrage werden im folgenden Kapitel vorgestellt.

Kapitel 6

Zusammenfassung

Das Ziel dieser Arbeit besteht in der Erkennung von Anomalien in Trajektoriendaten einer urbanen Kreuzung. Dazu wird die zentrale Forschungsfrage „Welche Anomalien befinden sich in welchem Umfang im vorliegenden Straßenverkehrsdatensatz?“ untersucht. Zur Beantwortung dieser Frage werden die Anomalien in fehlerhafte und außergewöhnliche Daten unterschieden. Mit der Analyse fehlerhafter Daten kann die Qualität der Daten bewertet werden. Hingegen kann durch Untersuchung der außergewöhnlichen Daten das Verständnis über das Verkehrsteilnehmerverhalten ausgebaut werden. Da dies im Fokus der Arbeit steht, wird der zu untersuchende Datensatz von fehlerhaften Daten bereinigt und die verbleibenden Daten auf Anomalien überprüft. Die identifizierten Anomalien werden durch Sichtung der Videoaufzeichnungen, aus denen die Daten erhoben werden, analysiert und nach fehlerhaften und außergewöhnlichen Daten unterschieden.

Der zu analysierende Datensatz von acht Tagen wird für dieses Vorgehen in einen Trainingsdatensatz aus sieben Tagen und einen Testdatensatz eingeteilt, der die Daten des verbleibenden Tages umfasst. Um Anomalien zu detektieren, wird zu Beginn jedes Verfahrens ein Modell aus einem Trainingsdatensatz aufgestellt, das die zu erwartende Verteilung der Daten repräsentiert. Anschließend werden Testdaten mit dem Modell verglichen und jeder Datenpunkt, dessen Abweichung vom Modell einen Schwellenwert überschreitet als Anomalie klassifiziert.

In Tabelle 6.1 sind die in dieser Arbeit angewendeten Verfahren zur Anomalieerkennung, die Größe der Datensätze sowie die Anzahl der erkannten Anomalien aufgelistet, um einen Überblick über die Ergebnisse zu geben. Die Tabelle 1 im Anhang gibt einen Überblick über die verwendeten Trainingsdaten und die darin enthaltenden fehlerhaften Daten. Mit diesen Tabellen wird der Teil der Forschungsfrage, die sich auf den Umfang der vorliegenden Anomalien bezieht, beantwortet. Für die Beantwortung der Frage, welche Anomalien im Datensatz vorliegen, werden die Ergebnisse der Analyse der detektierten Anomalien nachfolgend vorgestellt.

Die Entfernung von ausschließlich fehlerhaften Daten erfolgt auf Basis der sechs in der Tabelle zuerst aufgelisteten Datengrundlagen. Anschließend werden ebenso im Rahmen der Datenvorbereitung die Trajektorien nach deren Routen klassifiziert. Bei diesem und den fünf zuletzt aufgelisteten Verfahren werden Anomalien detektiert, die erst nach Sichtung des Videomaterials eine der Gruppen von fehlerhaften oder außergewöhnlichen Daten zugeordnet werden können.

Zunächst wird die Anomalieerkennung auf die Länge der Trajektorien angewandt. Des Weiteren werden Trajektorien detektiert, deren Aufzeichnungsfrequenz nicht der Frequenz der Messinfrastruktur von 25 fps entspricht. Anschließend werden die Bewegungsdaten

Tabelle 6.1: Anzahl der verwendeten Testdaten und detektierten Anomalien je Verfahren

Datengrundlage	Werte		Trajektorien	
	Anzahl	Anomalien	Anzahl	Anomalien
Länge der Trajektorien	15.838.303	22	60.555	22
Aufzeichnungsfrequenz	15.838.281	5.621	60.533	2.842
Beschleunigung	9.334.586	0	44.573	0
Geschwindigkeit	9.334.586	129	44.573	127
X-Position	9.298.253	128	44.446	107
Y-Position	9.298.253	168	44.446	127
Route	9.248.346	625.011	44.219	3.340
Hausdorff-Metrik	464.986	705	743	34
Übergangswahrscheinlichkeit	186.045	12.386	1.060	1.060
Bewegungsrichtung	656.425	40	1.060	6
Anzahl der Trajektorien	658.033	1.636	1.060	40
Gaußsches Mischmodell	8.569.987	6.681	40.770	10

speziell für die als PKW klassifizierten Objekte auf fehlerhafte Daten untersucht. Auf Basis der verbleibenden Beschleunigungswerte wird die Anomalieerkennung in den Geschwindigkeitswerten durchgeführt. Dabei wird die Differenz der berechneten und gemessenen Geschwindigkeit in aufeinanderfolgenden Datenpunkten einer Trajektorie als Grundlage für die Anomalieerkennung verwendet. Ebenso wird mit den Werten der X- und Y-Position verfahren.

Anschließend werden die Trajektorien nach deren Route mittels Polygonen gruppiert, um nur die Trajektorien der Route von Westen nach Norden auf außergewöhnliche Anomalien zu überprüfen. Bei dieser Klassifizierung können allerdings nicht allen Trajektorien eine Route zugewiesen werden, da diese erst im Innenbereich der Kreuzung beginnen oder dort enden. Ebenso werden für zwei Trajektorien Schnittpunkte mit mehreren Polygonen registriert, die an einer Einfahrt positioniert sind. Die Sichtung des Videomaterials ergibt, dass eine dieser Anomalien durch die fehlerhafte Positionsbestimmung eines Objektes entstanden ist. Die zweite anomale Trajektorie beschreibt allerdings einen Verkehrsteilnehmer, der von der Kreuzung auf den Gehweg gefahren ist. Somit werden auch mit diesem Verfahren außergewöhnliche Daten detektiert.

Um weitere außergewöhnliche Anomalien unterschiedlicher Art zu identifizieren, werden in der Arbeit fünf Verfahren angewendet. Da die anomalen Daten nicht bekannt sind, können die erstellten Modelle nicht nach deren Genauigkeit bewertet werden. Die Genauigkeit der Modelle könnte ansonsten mit der Rate der richtig detektierten Anomalien quantifiziert werden. Da dies nicht möglich ist, werden die Modelle nach deren Plausibilität aus Sicht des Anwenders bewertet. Allgemein wird daher festgehalten, dass das Ergebnis einer solchen Anomalieerkennung immer vom Benutzer abhängt, da dieser den Schwellenwert festlegt, nachdem Anomalien detektiert werden. Um einen angemessenen Schwellenwert zu definierten, wird im Rahmen der vorliegenden Arbeit der Z-Wert verwendet.

Beim ersten Verfahren wird die Hausdorff-Metrik angewandt, um eine durchschnittliche Trajektorie zu bestimmen und anschließend Anomalien zu detektieren, deren Abstand zur durchschnittlichen Trajektorie einen Schwellenwert übersteigt. Da die Berechnung einer durchschnittlichen Trajektorie nur Sinn für diese Trajektorien ergibt, die von der gleichen Spur kommend in der gleichen Spur enden, werden nur solche Trajektorien untersucht. Hierfür werden die 715 Trajektorien des Testdatensatzes ausgewählt, die auf der Route von Westen nach Norden verlaufen und deren Detektion in der rechten Fahrspur abbricht. Mit

dem verwendeten Schwellenwert werden 34 anomale Trajektorien erkannt, von denen zwei Trajektorien außergewöhnliche Daten darstellen. Diese stellen Trajektorien von Verkehrsteilnehmern dar, die von der Fahrbahn für Geradeausfahrer in die Kreuzung einfahrend, anschließend die Route nach Norden gewählt haben. In dem vorliegenden Fall kann der Schwellenwert soweit erhöht werden, dass nur noch die zwei außergewöhnlichen Trajektorien als Anomalie klassifiziert werden. Die restlichen 19 Trajektorien werden aufgrund einer fehlerhaften Positionsbestimmung oder Objektklassifizierung als Anomalie detektiert und stellen demnach fehlerhafte Daten dar. Die Differenzen zwischen der Position der Verkehrsteilnehmer und deren detektierten Position kann mit schlechten Sichtverhältnissen aufgrund von Dunkelheit und der weiten Entfernung zu Kameras erklärt werden.

Auch bei dem zweiten, dritten und vierten Verfahren muss speziell eine Route untersucht werden. Hierfür wird erneut die Route von Westen nach Norden ausgewählt, wobei die Trajektorien aller Fahrstreifen untersucht werden. Diese Verfahren werden zusammen genannt, da die Anomalieerkennung jeweils auf Basis der Repräsentation der Datenpunkte in einem diskreten euklidischen Raum durchgeführt wird. Dazu wird die Kreuzung als ein Gitter bestehend aus mehreren Zellen interpretiert, sodass an jeder Position der Kreuzung das zu erwartende Verkehrsteilnehmerverhalten bestimmt werden kann. Bei dem zweiten Verfahren, werden die Trajektorien nach deren Übergangswahrscheinlichkeit zwischen unterschiedlichen Zellen bewertet. Hierfür können keine sinnvollen Ergebnisse erzielt werden, da das Gitter nicht an den Verlauf der Route angepasst ist und alle Trajektorien mindestens eine Übergangswahrscheinlichkeit enthalten, die als Anomalie erkannt wird. Daher wird ein weiteres Verfahren angewandt, bei dem der Verlauf der Position eines Verkehrsteilnehmers auf Anomalien überprüft wird. Dazu werden die aufgezeichneten Daten der Bewegungsrichtung eines Verkehrsteilnehmers verwendet und in jeder Zelle des Gitters die Verteilung der Bewegungsrichtungen im Trainingsdatensatz ermittelt. Bewegungsrichtungen im Testdatensatz die von dieser Verteilung abweichen, werden als Anomalie klassifiziert, wodurch sechs anomale Trajektorien bestimmt werden. Von diesen ist die Trajektorie des Datenpunktes, der den höchsten Anomalie-Wert aufweist, die einzige außergewöhnliche. Alle weiteren anomalen Trajektorien basieren auf fehlerhaften Daten. Die außergewöhnliche Trajektorie stellt wie bei der Anwendung der Hausdorff-Metrik einen Verkehrsteilnehmer dar, der von einer unzulässigen Spur nach Norden abgebogen ist. Dieser wird mit der Hausdorff-Metrik allerdings nicht als Anomalie identifiziert, da dieser erst nahe der durchschnittlichen Trajektorie detektiert wird, sodass dessen Distanz zu dieser nicht den Schwellenwert überschreitet, die Bewegungsrichtung an dieser Stelle allerdings schon.

Da für die Bestimmung der Bewegungsrichtung nur Zellen analysiert werden, in denen im Trainingsdatensatz mehr als fünf Bewegungsrichtungen vorhanden sind, wird eine Anomalieerkennung auf Basis der Anzahl der Trajektorien in einer Zelle durchgeführt. Dabei werden Datenpunkte, die sich in einer Zelle befinden, in der im Trainingsdatensatz die Anzahl der Verkehrsteilnehmer einen Schwellenwert unterschreiten, als Anomalie klassifiziert. Mit diesem Verfahren werden 40 Trajektorien des Testdatensatzes als Anomalie klassifiziert, die sich am Rand des Trainingsdatensatz befinden und damit nicht für die Anomalieerkennung auf Grundlage der Bewegungsrichtung verwendet werden können. Von den 40 Anomalien werden vier auf außergewöhnliche Daten zurückgeführt.

Bei dem Verfahren des Gaußschen Mischmodells werden die Trajektorien als Anomalie klassifiziert, deren zugewiesenen Route nach den Polygonen sich von der zugewiesenen Route nach dem Gaußschen Mischmodell unterscheidet. Daher wird die Anomalieerkennung mit dem Gaußschen Mischmodell auf den Anteil des Testdatensatzes angewandt, der die als PKW klassifizierten Objekte enthält, denen mittels den Polygonen eine Route zuge-

ordnet werden kann. Die Anwendung erfolgt, indem die Trajektorien als dreidimensionalen Datenpunkt repräsentiert werden. Daraus ergeben sich zehn Anomalien, von denen jeweils die Hälfte der Anomalien den Routen von Süden nach Norden und von Westen nach Osten zugeteilt wird. Die Analyse der Anomalien ergibt, dass fünf Verkehrsteilnehmer, länger als die anderen Trajektorien im Stehen detektiert werden, vier Verkehrsteilnehmer, einen unzulässigen Fahrstreifen für die Route gewählt haben und ein Verkehrsteilnehmer, der aufgrund von Dunkelheit kürzer als andere erkannt wird. Von den stehenden PKWs werden drei vor einer roten LSA aufgenommen. Ein weiterer wird beim Verlassen der Kreuzung lange an einer Position detektiert, an der sich dieser nicht befunden hat. Der verbleibende Verkehrsteilnehmer wird beim Parken seines PKWs direkt hinter der Fußgängerfurt erfasst. Demnach stellen acht der zehn Anomalien außergewöhnliche und zwei Anomalien fehlerhafte Daten dar.

Mit der Anwendung der vorgestellten Verfahren werden verschiedene außergewöhnliche Daten detektiert. Damit wird deutlich, dass die Auswahl des Verfahrens und die Festlegung der zu verwendenden Daten und Parameter genau auf die zu detektierenden außergewöhnlichen Daten abgestimmt sein müssen. Allerdings besteht selbst dann die Möglichkeit, dass die detektierten Anomalien aufgrund fehlerhafter Daten als solche klassifiziert wurden. Alles in allem wird mit den vorgestellten Ergebnissen gezeigt, dass es die angewandten Verfahren dem Anwender ermöglichen, außergewöhnliche Situationen aus einem Datensatz zu ermitteln, um diese anschließend im Detail zu untersuchen.

Literaturverzeichnis

[Agr15] Agrawal, Shikha; Agrawal, Jitendra: *Survey on Anomaly Detection using Data Mining Techniques*. Procedia Computer Science, 60:708–713, 2015.

[Ahm18] Ahmed, Sk Arif; Dogra, Debi Prosad; Kar Samarjit; Roy Partha Pratim: *Trajectory-Based Surveillance Analysis: A Survey*. IEEE Transactions on Circuits and Systems for Video Technology, 2018.

[Arn18] Arndt, Richard: *Verteiltes digitales Messen zur Erkennung von Objekten im Straßenverkehr*. rialgo realtime systems GmbH & Co. KG, 1. Auflage, 2018.

[Bis06] Bishop, Christopher M.: *Pattern Recognition and Machine Learning*. Springer, 2006.

[Bra06] Braynova, Elena: *Indexing Spatio-Temporal Trajectories with Orthogonal Polynomials*. In: *DMIN*, Seiten 343–348. Citeseer, 2006.

[Bre00] Breunig, Markus M.; Kriegel, Hans-Peter; Ng Raymond T.; Sander Jörg: *LOF: Identifying Density-Based Local Outliers*. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, Seiten 93–104, 2000.

[Buc13] Buchin, Kevin; Buchin, Maike; Van Kreveld Marc; Löffler Maarten; Silveira Rodrigo I; Wenk Carola; Wiratma Lionov: *Median Trajectories*. Algorithmica, 66(3):595–614, 2013.

[Bur09] Burg, Heinz; Moser, Andreas: *Handbuch Verkehrsunfallrekonstruktion: Unfallaufnahme, Fahrdynamik, Simulation*. Springer-Verlag, 2009.

[Cha09] Chandola, Varun; Banerjee, Arindam; Kumar Vipin: *Anomaly Detection: A Survey*. ACM computing surveys (CSUR), 41(3):15, 2009.

[Deua] Deutsche Zentrum für Luft- und Raumfahrt e. V.: *Kameraperspektive auf die Forschungskreuzung*. Zugriff am 24.03.2020.

[Deub] Deutsche Zentrum für Luft- und Raumfahrt e. V.: *Satellitenbild von der Forschungskreuzung*. Zugriff am 24.03.2020.

[Eck10] Eckstein, Peter P.: *Statistik für Wirtschaftswissenschaftler*. Springer, 2010.

[Fah16] Fahrmeir, Ludwig; Heumann, Christian; Künstler Rita; Pigeot Iris; Tutz Gerhard: *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag, 2016.

[Gao17] Gao, Weixin Luo; Wen Liu; Shenghua: *Remembering history with convolutional LSTM for anomaly detection*. 2017 IEEE International Conference on Multimedia and Expo (ICME), Seiten 439–444, 2017.

[Ge10] Ge, Yong; Xiong, Hui; Zhou Zhi-hua; Ozdemir Hasan; Yu Jannite; Lee Kuo Chu: *TOP-EYE: Top-k Evolving Trajectory Outlier Detection*. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*, Seiten 1733–1736, 2010.

[Gup13] Gupta, Manish; Gao, Jing; Aggarwal Charu C; Han Jiawei: *Outlier Detection for Temporal Data: A Survey*. IEEE Transactions on Knowledge and Data Engineering, 26(9):2250–2267, 2013.

[Han19] Han, Yutao; Tse, Rina; Campbell Mark: *Pedestrian Motion Model Using Non-Parametric Trajectory Clustering and Discrete Transition Points*. IEEE Robotics and Automation Letters, 4(3):2614–2621, 2019.

[Haw80] Hawkins, Douglas M.: *Identification of Outliers*, Band 11. Springer, 1980.

[How12] Howard, Ronald A: *Dynamic Probabilistic Systems: Markov Models*, Band 1. Courier Corporation, 2012.

[Lax13] Laxhammar, Rikard; Falkman, Göran: *Online Learning and Sequential Anomaly Detection in Trajectories*. IEEE transactions on pattern analysis and machine intelligence, 36(6):1158–1173, 2013.

[Leu13] Leutzbach, Wilhelm: *Einführung in die Theorie des Verkehrsflusses*. Springer-Verlag, Heidelberg, 2013.

[Ma18] Ma, Cong; Miao, Zhenjiang; Li Min; Song Shaoyue; Yang Ming-Hsuan: *Detecting Anomalous Trajectories via Recurrent Neural Networks*. In: *Asian Conference on Computer Vision*, Seiten 370–382. Springer, 2018.

[Men19] Meng, Fanrong; Yuan, Guan; Lv Shaoqian; Wang Zhixiao; Xia Shixiong: *An overview on trajectory outlier detection*. Artificial Intelligence Review, 52(4):2437–2456, 2019.

[Mir10] Miranda, Jackelyn Wirri: *Data Mining - Anomalieentdeckung*. Technischer Bericht, Hochschule für Technik, Wirtschaft und Kultur Leipzig, 2010.

[Sch11] Schnabel, W. und Lohse, D.: *Grundlagen der Straßenverkehrstechnik und der Verkehrsplanung: Band 1 - Straßenverkehrstechnik*. Beuth Forum. DIN Deutsches Institut für Normung e.V., 2011.

[Sta18] Statistisches Bundesamt (Destatis): *Verkehrsunfälle*, 2018.

[Vaj09] Vajna, Sandor; Weber, Christian; Bley Helmut; Zeman Klaus: *CAx für Ingenieure: Eine praxisbezogene Einführung*. Springer-Verlag, 2009.

[Wet19] WetterKontor GmbH: *Wetterrückblick Braunschweig*. `https://www.wetterkontor.de/de/wetter/deutschland/rueckblick.asp?id=29&datum0=21.05.2019&datum1=29.05.2019&jr=2020&mo=1&datum=02.06.2019&t=2&part=0`, 2019. Zugriff: 31.03.2020.

[Yan19] Yang, Jiawei; Rahardja, Susanto; Fränti Pasi: *Outlier Detection: How to Threshold Outlier Scores?* In: *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, Seiten 1–6, 2019.

Anhang

Abbildung 1: „Sichtbereiche der Sensor-Systeme zur 3D-Merkmalsberechnung im Innenbereich der Kreuzung.“ [Arn18, S. 96]

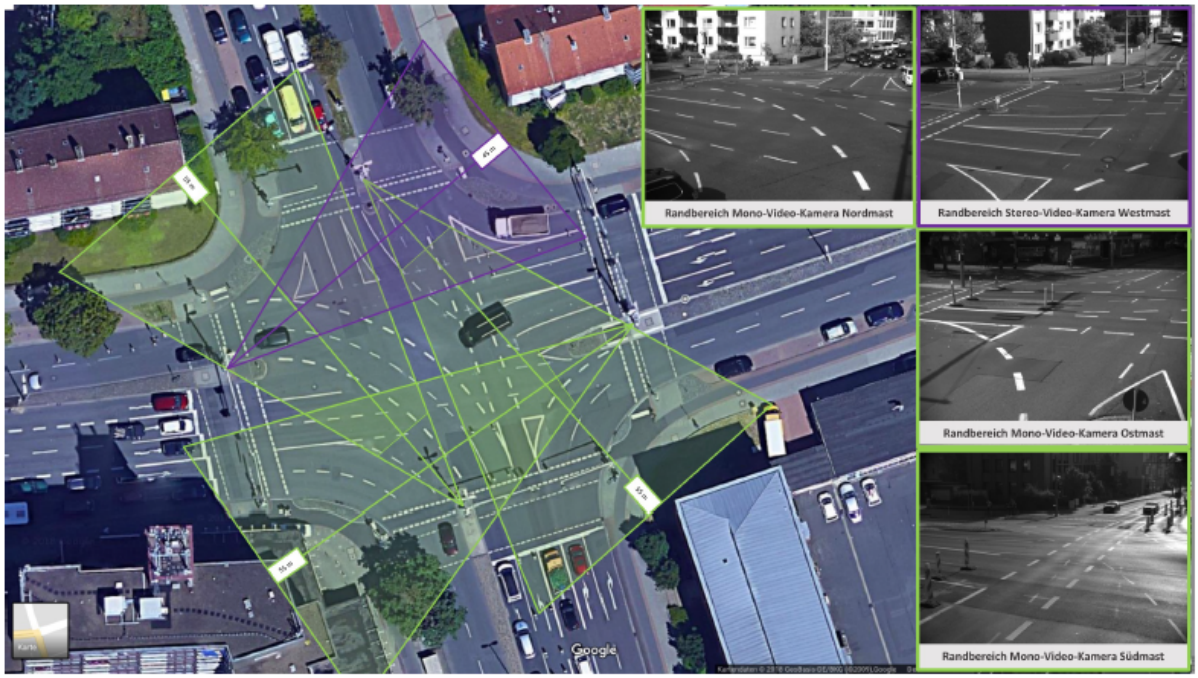


Abbildung 2: „Sichtbereiche der Sensor-Systeme zur 3D-Merkmalsberechnung am Randbereich der Kreuzung.“ [Arn18, S. 98]

Abbildung 3: „Sichtbereiche der Stereo-Kameras zur Beobachtung der Fußgängerfurten im Westen und Süden der Kreuzung.“ [Arn18, S. 101]

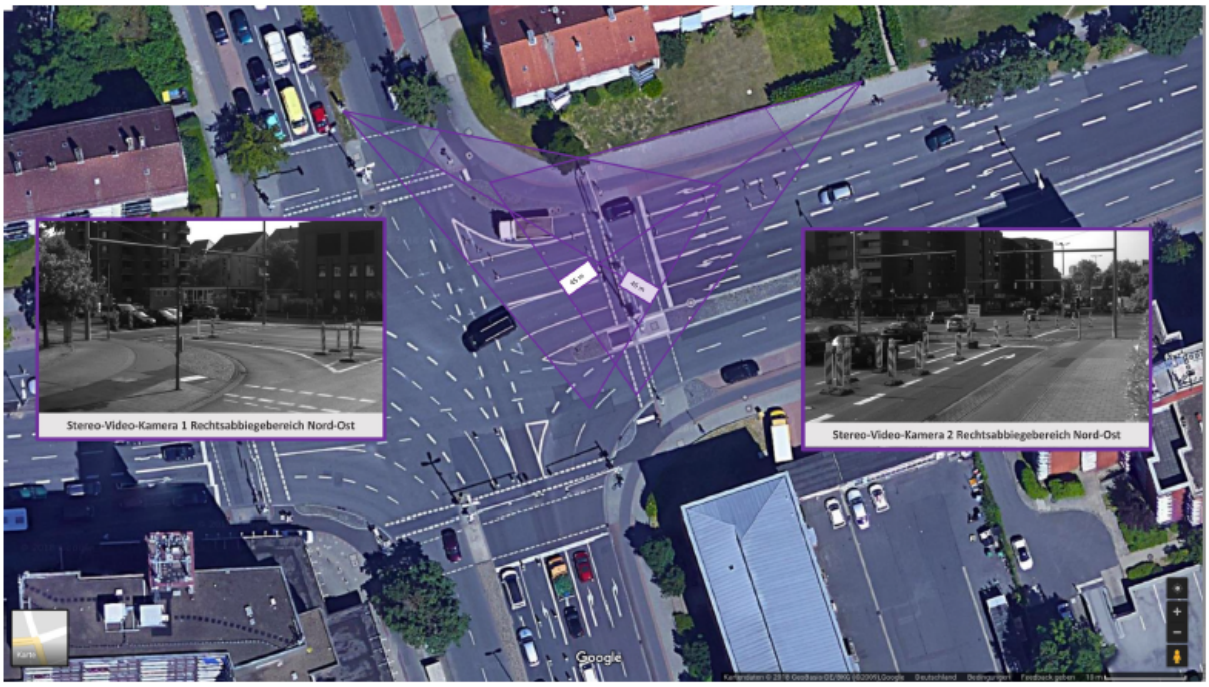


Abbildung 4: „Sichtbereiche der Stereo-Kameras zur 3D-Merkmalsberechnung im nordöstlichen Rechtsabbiegebereich.“ [Arn18, S. 103]

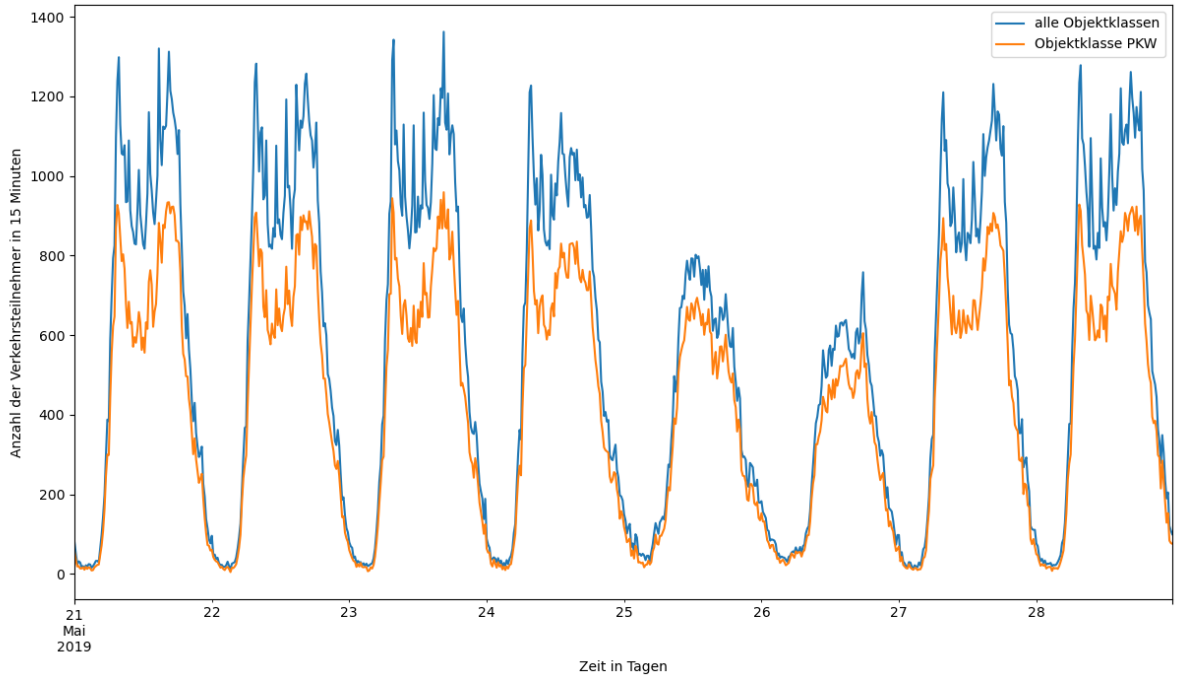


Abbildung 5: Verkehrsstärke im Zeitraum der analysierten Daten vom 21. bis 29 Mai 2019.

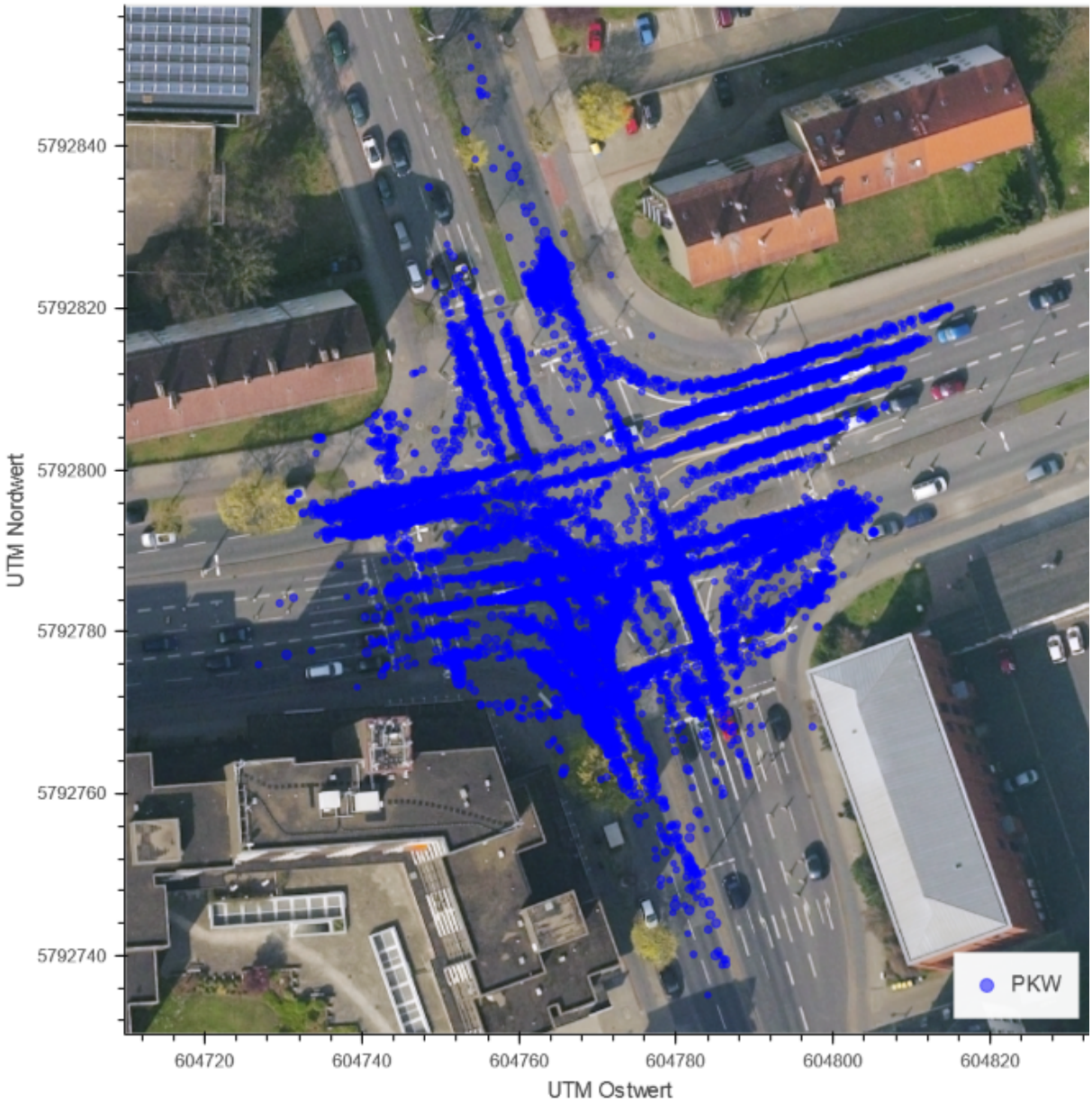


Abbildung 6: Verteilung der Anomalien in den Geschwindigkeitswerten, die für einen Schwellenwert von drei Standardabweichungen ermittelt werden.

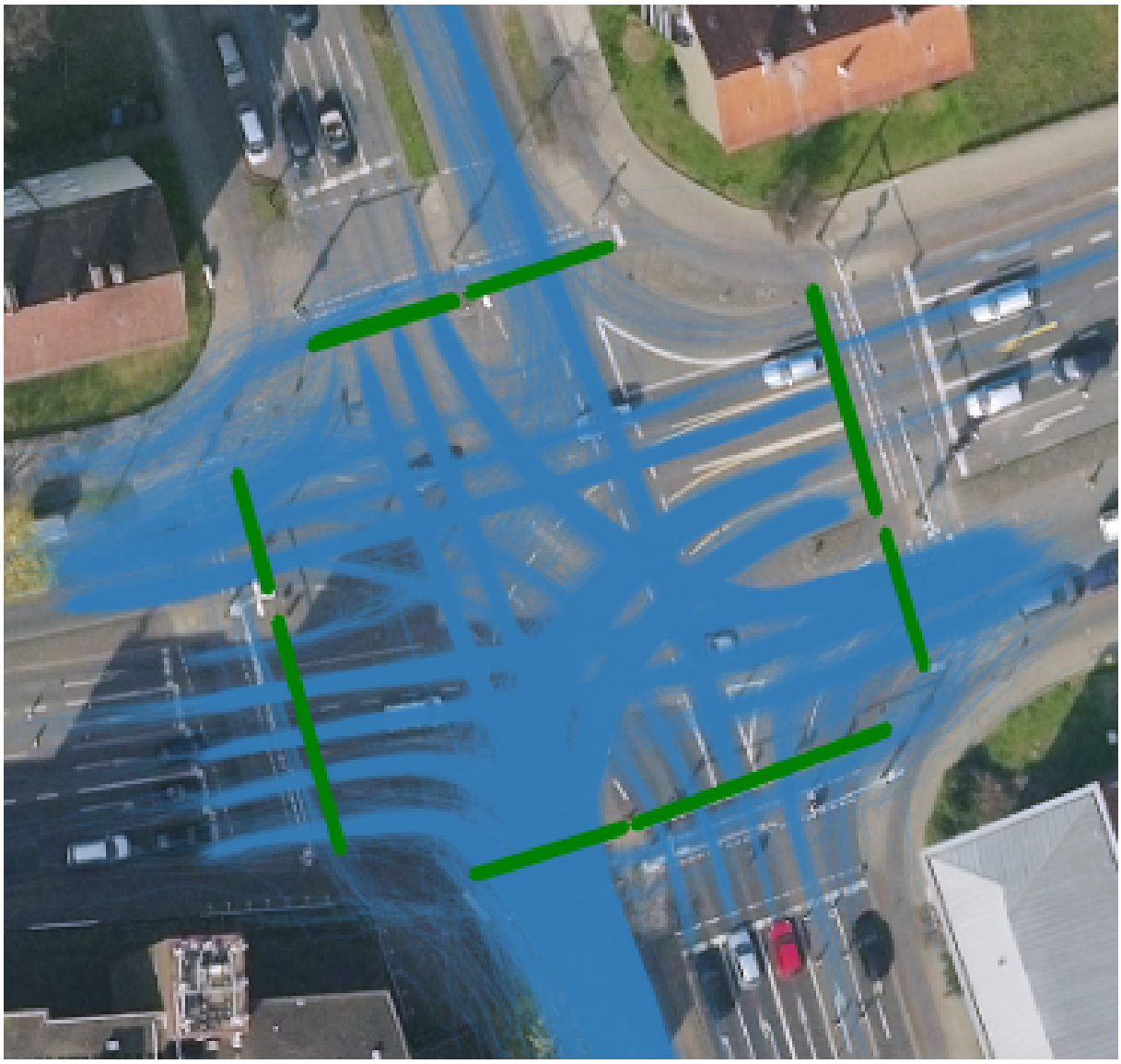


Abbildung 7: Trajektorien (blau) des Testdatensatzes, denen mit den AOIs (grün) keine Route zugewiesen werden kann.

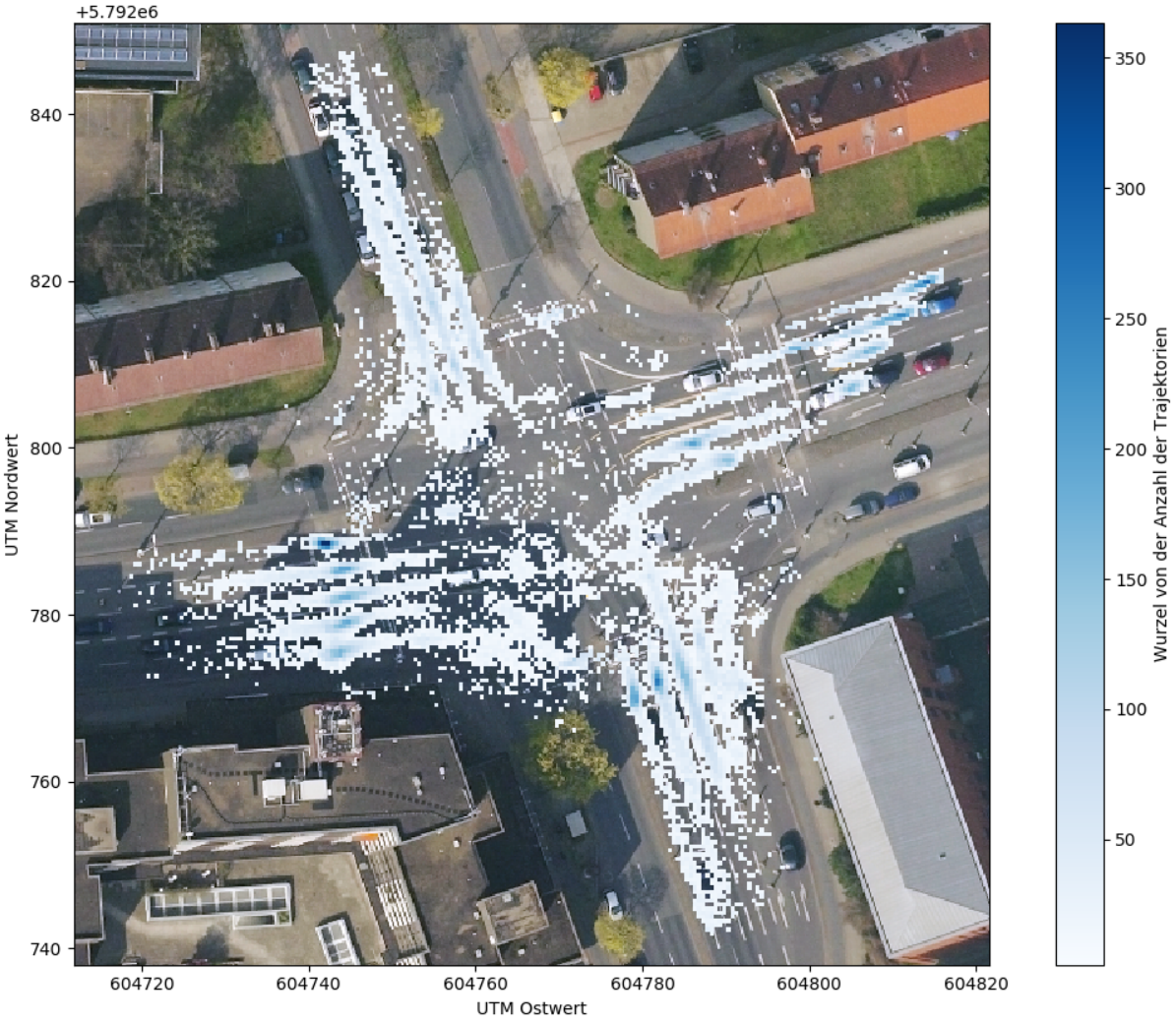


Abbildung 8: Verteilung der zuerst detektierten Position von Trajektorien

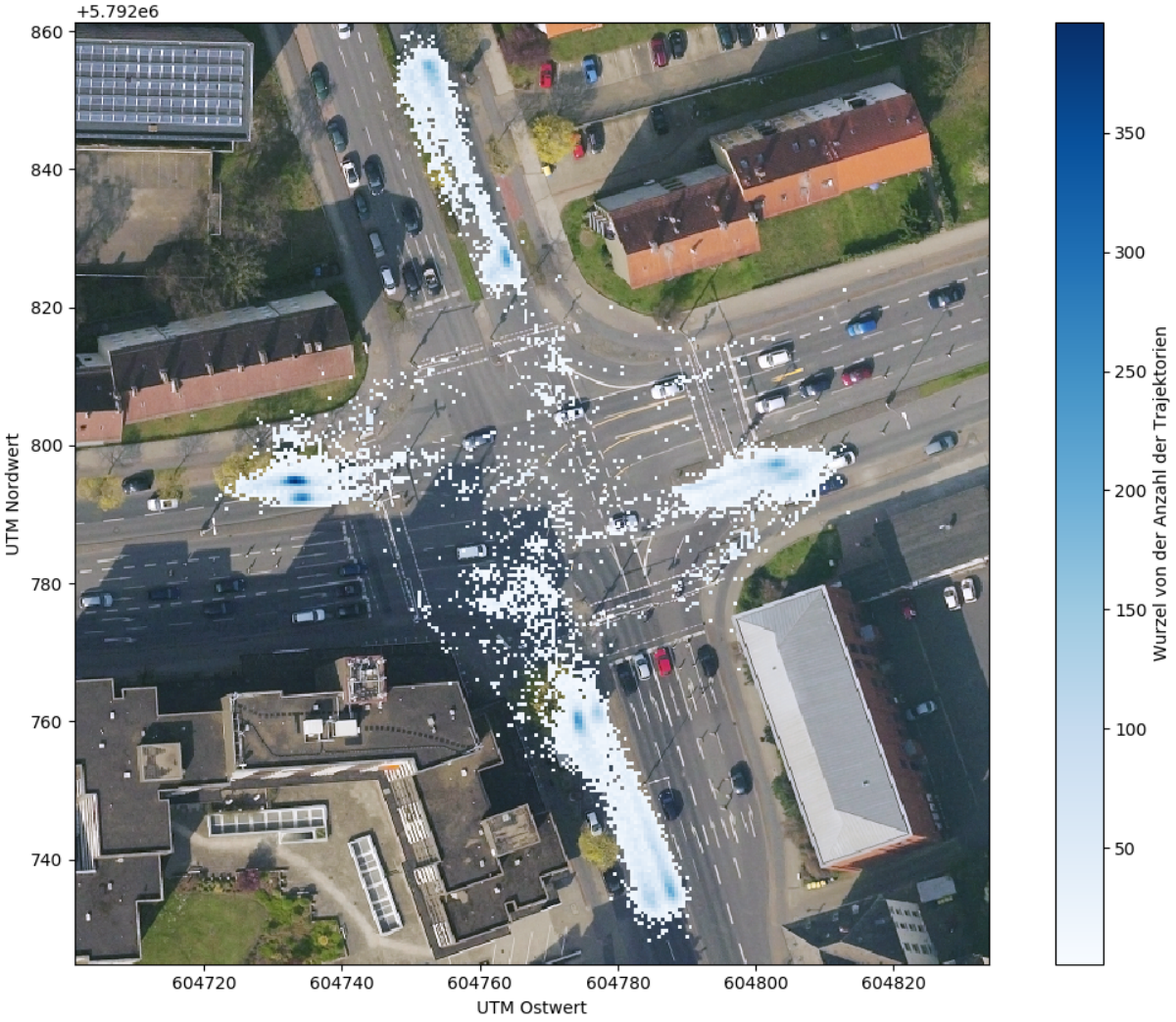


Abbildung 9: Verteilung der zuletzt detektierten Position von Trajektorien

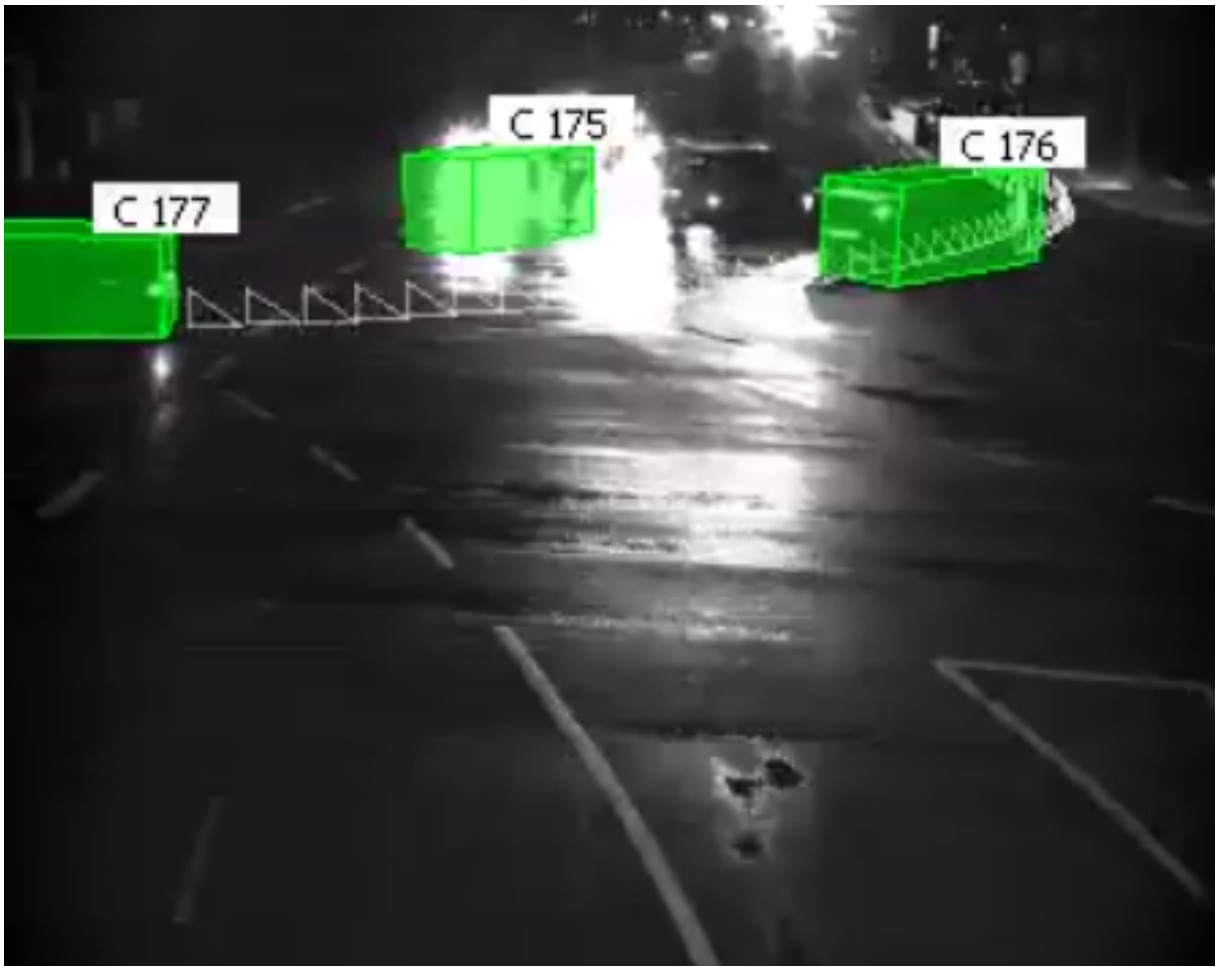


Abbildung 10: Anomalie Objekt C 175 (grün): Die Trajektorie des Objektes schneidet zwei Polygone, die sich an einer Einfahrt der Kreuzung befinden. Dies ist darauf zurückzuführen, dass die Position eines Objektes fehlerhafte bestimmt wird.

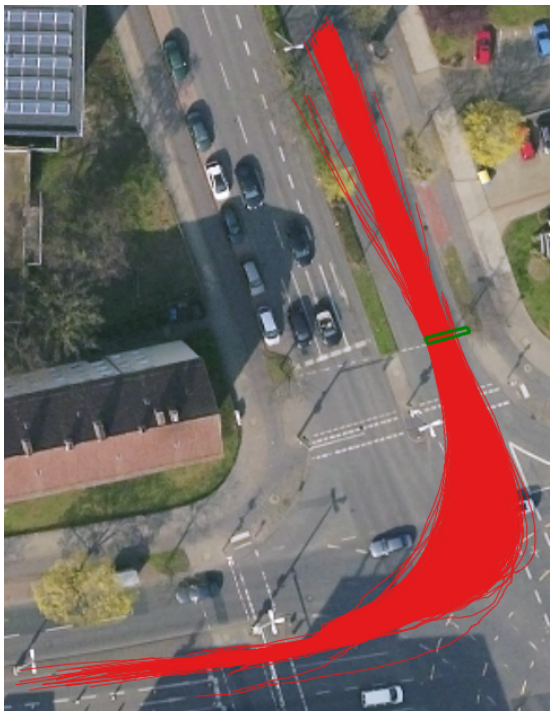


Abbildung 11: Trajektorien (rot) auf der Route von Westen nach Norden, denen mit dem Polygon (grün) die Ausfahrt der rechten Spur zugeordnet wird.

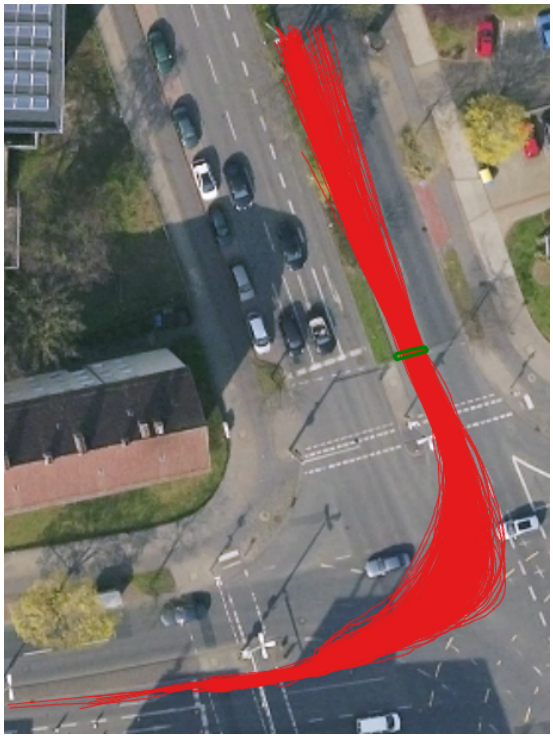


Abbildung 12: Trajektorien (rot) auf der Route von Westen nach Norden, denen mit dem Polygon (grün) die Ausfahrt der linken Spur zugeordnet wird.

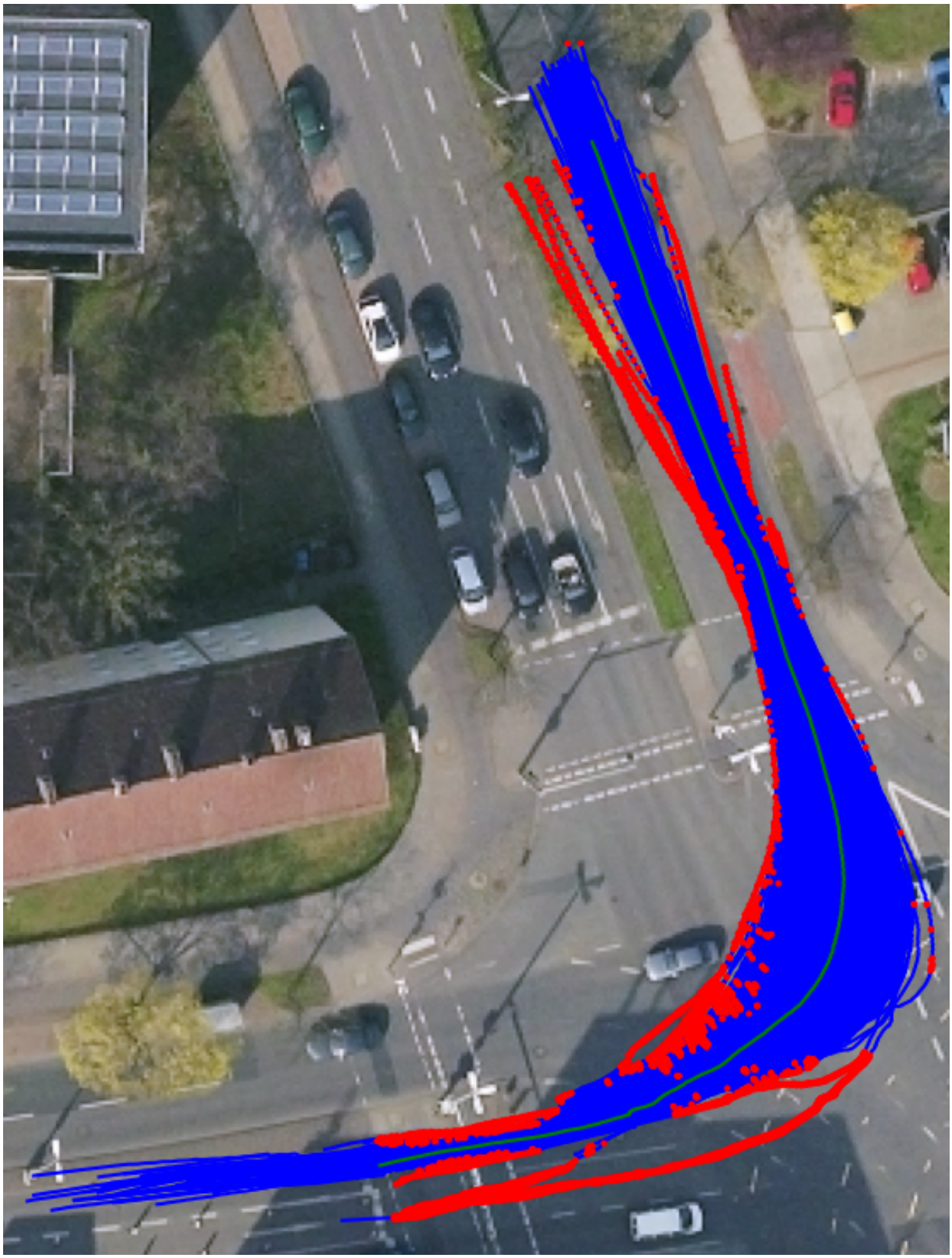


Abbildung 13: Mit der Hausdorff-Metrik und einem Z-Schwellenwert von drei bestimmte anormale Datenpunkte (rote Kreise). Die blauen Linien sind der zugrunde liegenden Testdatensatz von Trajektorien, die von Westen nach Norden auf die rechte Spur führen. Die grüne Linie stellt die durchschnittliche Trajektorie dar.

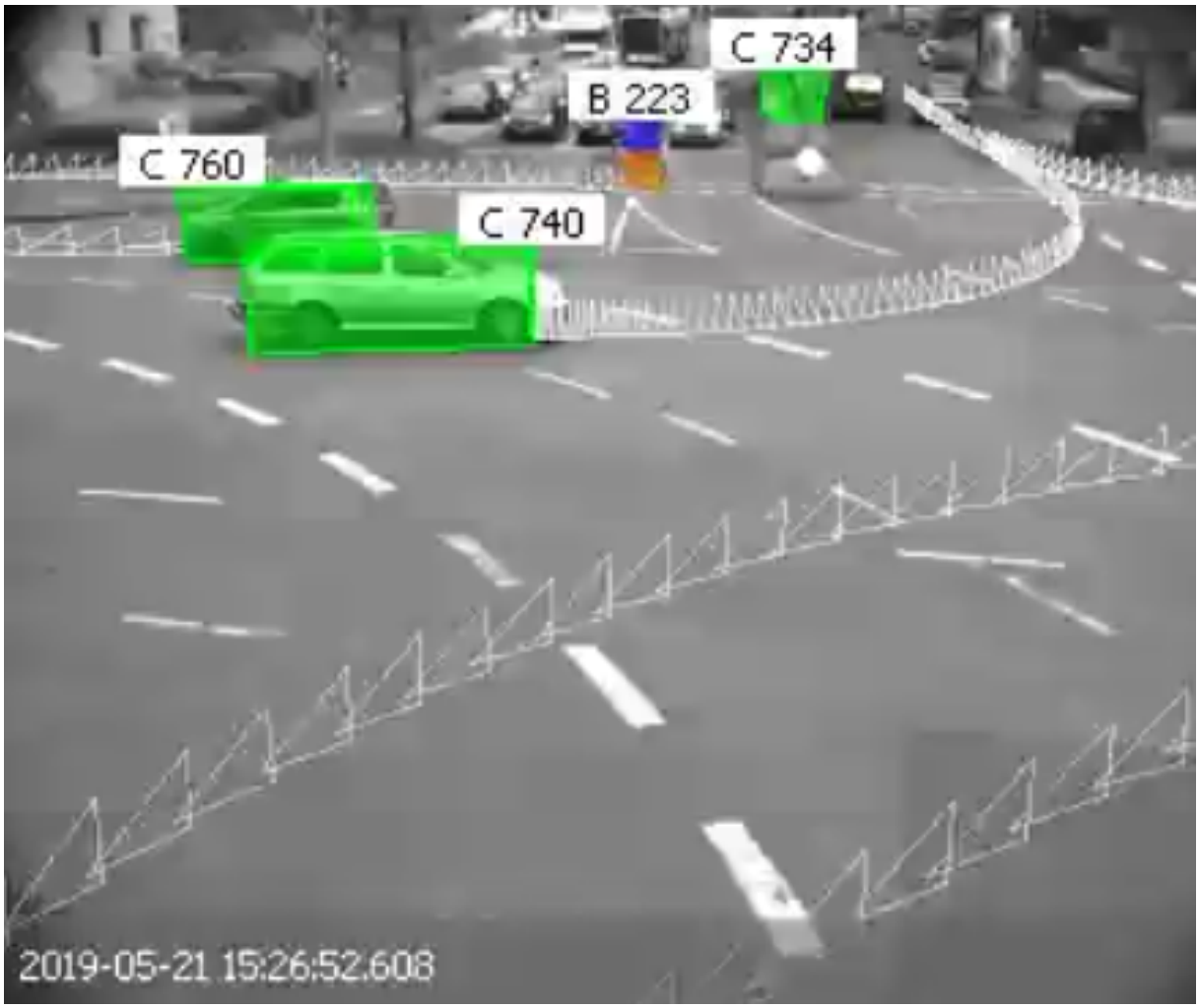


Abbildung 14: Das Fahrzeug C 734 wird nach der Hausdorff-Metrik als Anomalie klassifiziert, da für dieses Fahrzeug aufgrund einer fehlerhaften Positionsbestimmung eine Abweichung von der durchschnittlichen Trajektorie detektiert wurde.

Abbildung 15: Das Fahrzeug C 988 wird nach der Hausdorff-Metrik als Anomalie klassifiziert, da für dieses Fahrzeug aufgrund einer fehlerhaften Positionsbestimmung eine Abweichung von der durchschnittlichen Trajektorie detektiert wurde.

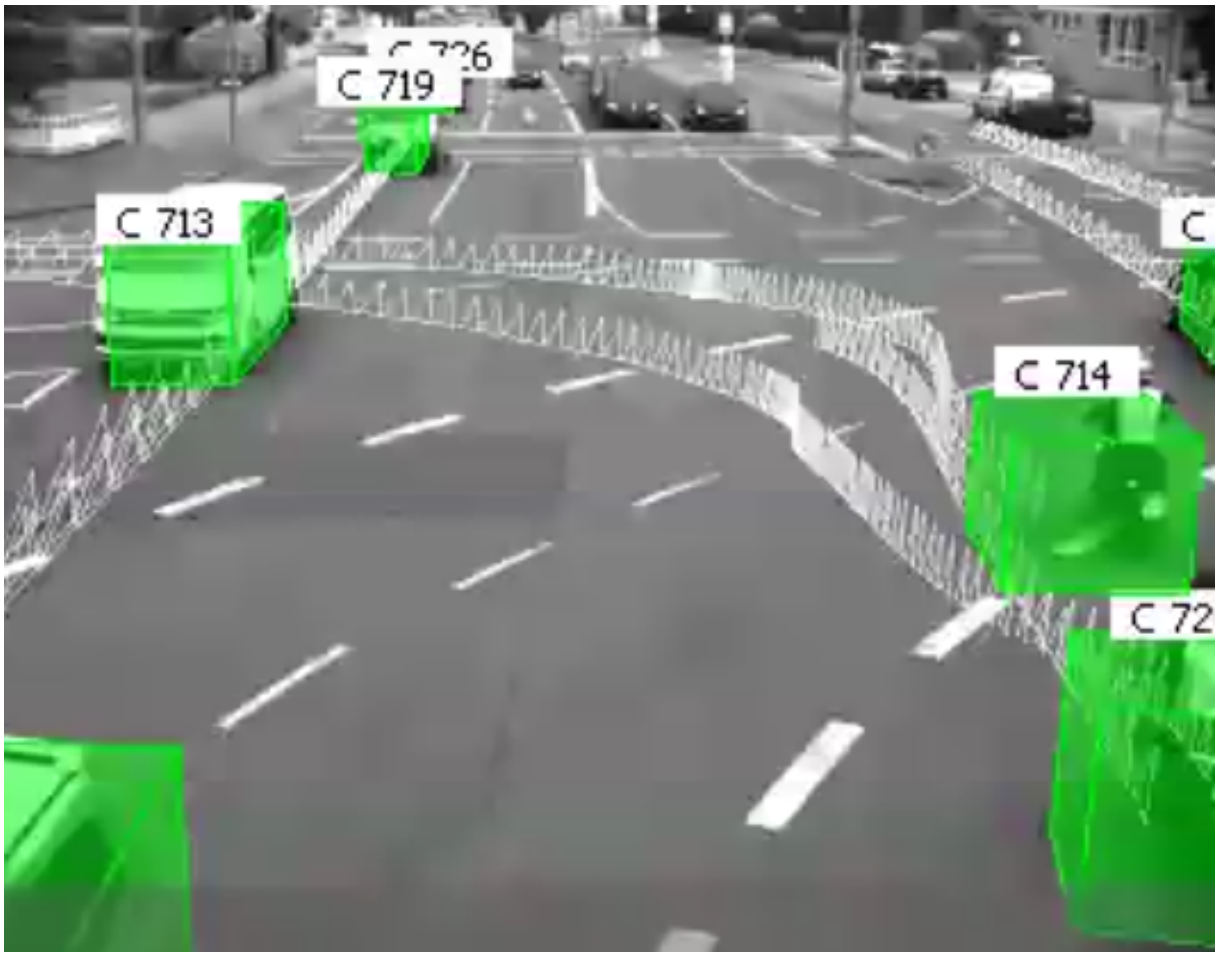


Abbildung 16: Der Verkehrsteilnehmer C 714 wird nach der Hausdorff-Metrik als Anomalie klassifiziert, da für dieses Fahrzeug aufgrund einer fehlerhaften Positionsbestimmung eine Abweichung von der durchschnittlichen Trajektorie detektiert wurde.

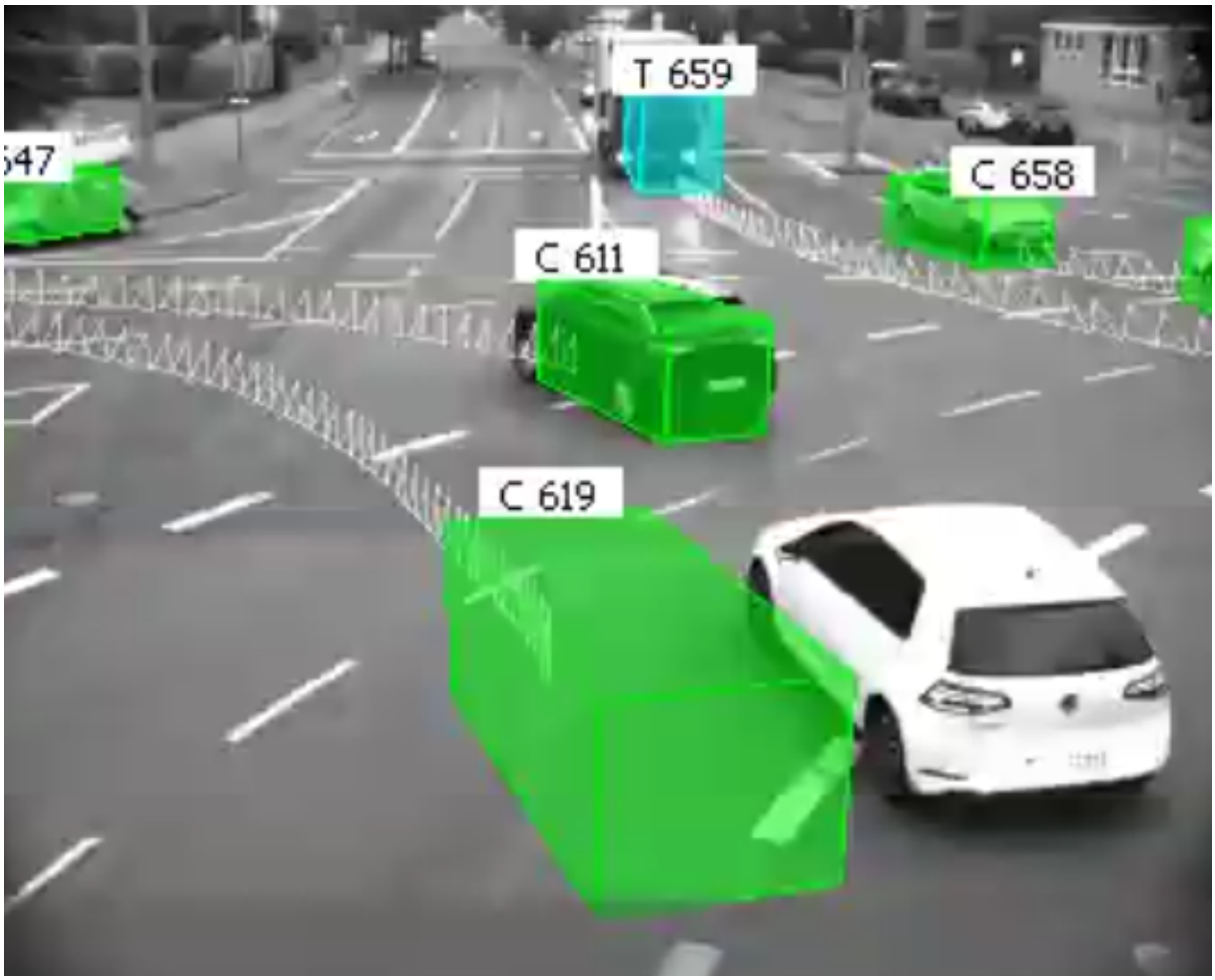


Abbildung 17: Das Fahrzeug C 358 wird nach der Hausdorff-Metrik als Anomalie klassifiziert, da für dieses Fahrzeug eine enge Kurvenfahrt aufgezeichnet wurde. Die Sichtung des Videomaterials ergab, dass die Anomalie durch fehlerhafte Daten entstanden ist.

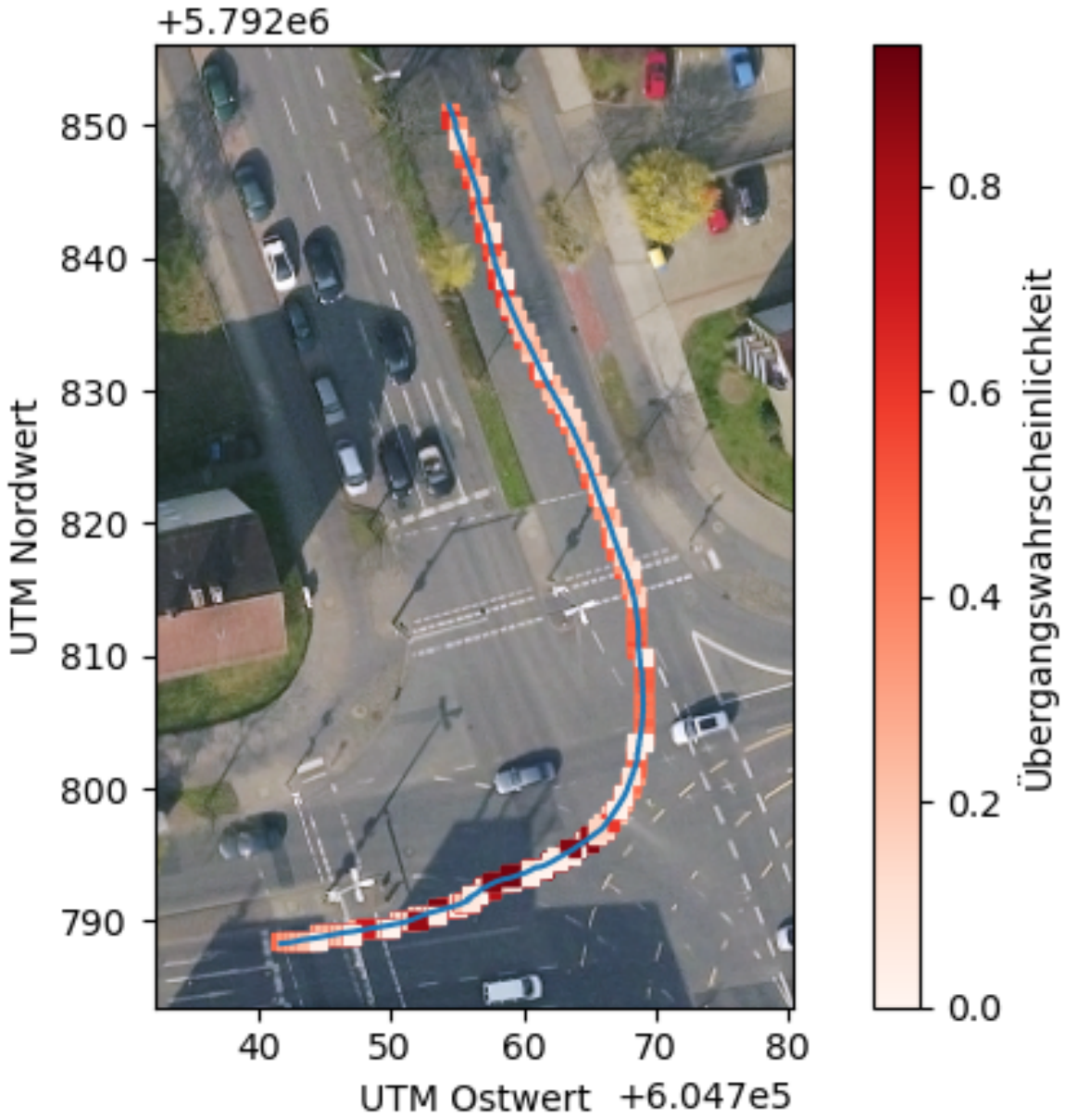


Abbildung 18: Bewertung einer beispielhaften Trajektorie (blau) mit der Markov-Kette. Die weißen bis roten Vierecke unter der Trajektorie geben die Übergangswahrscheinlichkeit von einem Datenpunkt zum nächsten an. Für das zugrundeliegende Modell wurden die stehenden Fahrzeuge berücksichtigt.

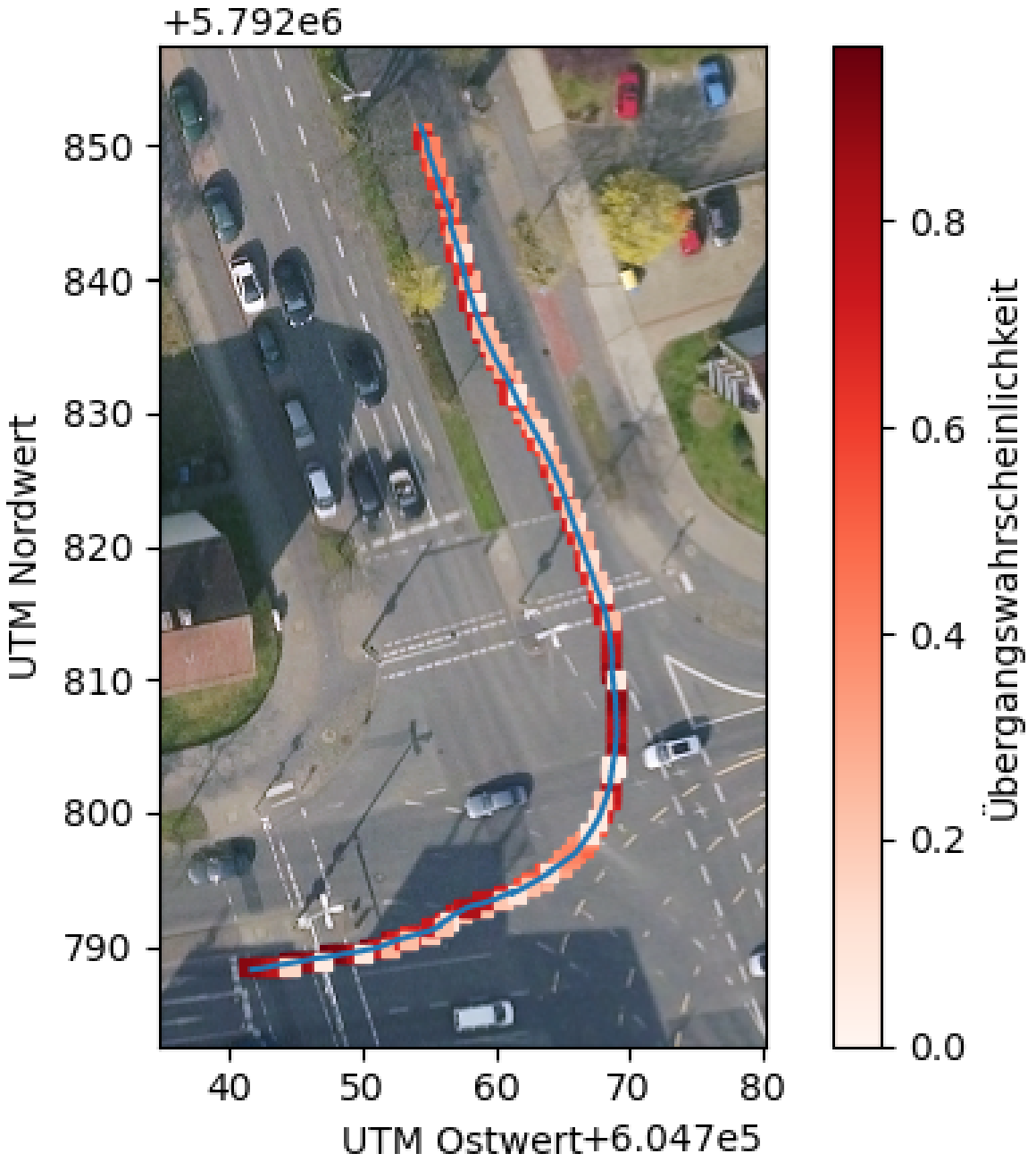


Abbildung 19: Bewertung einer beispielhaften Trajektorie (blau) mit der Markov-Kette. Die weißen bis roten Vierecke unter der Trajektorie geben die Übergangswahrscheinlichkeit von einem Datenpunkt zum nächsten an. Für das zugrundeliegende Modell wurden die stehenden Fahrzeuge nicht berücksichtigt.

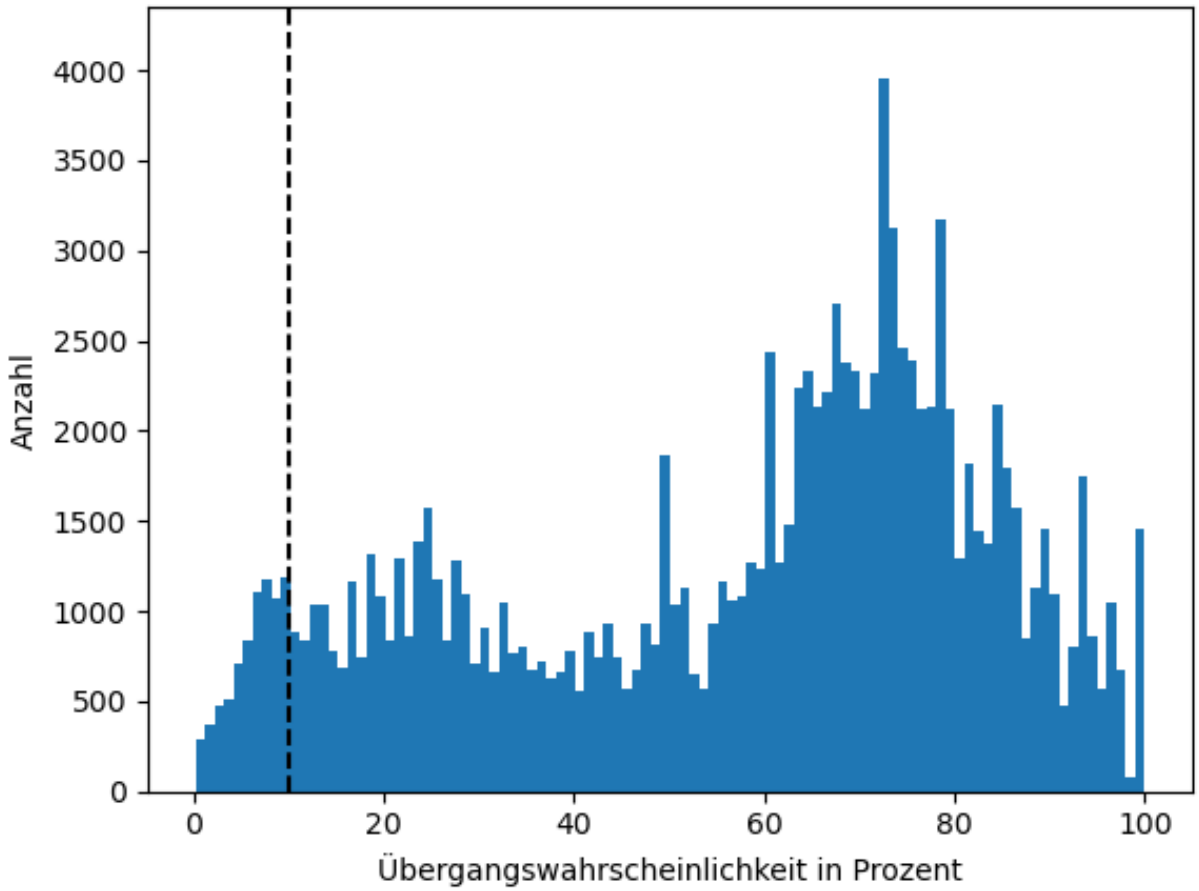


Abbildung 20: Verteilung der Übergangswahrscheinlichkeiten von den Übergängen, bei denen die Zelle gewechselt wurde. Die schwarz gestrichelte Linie zeigt den vom Anwender festgelegten Schwellenwert zur Anomalieerkennung an.

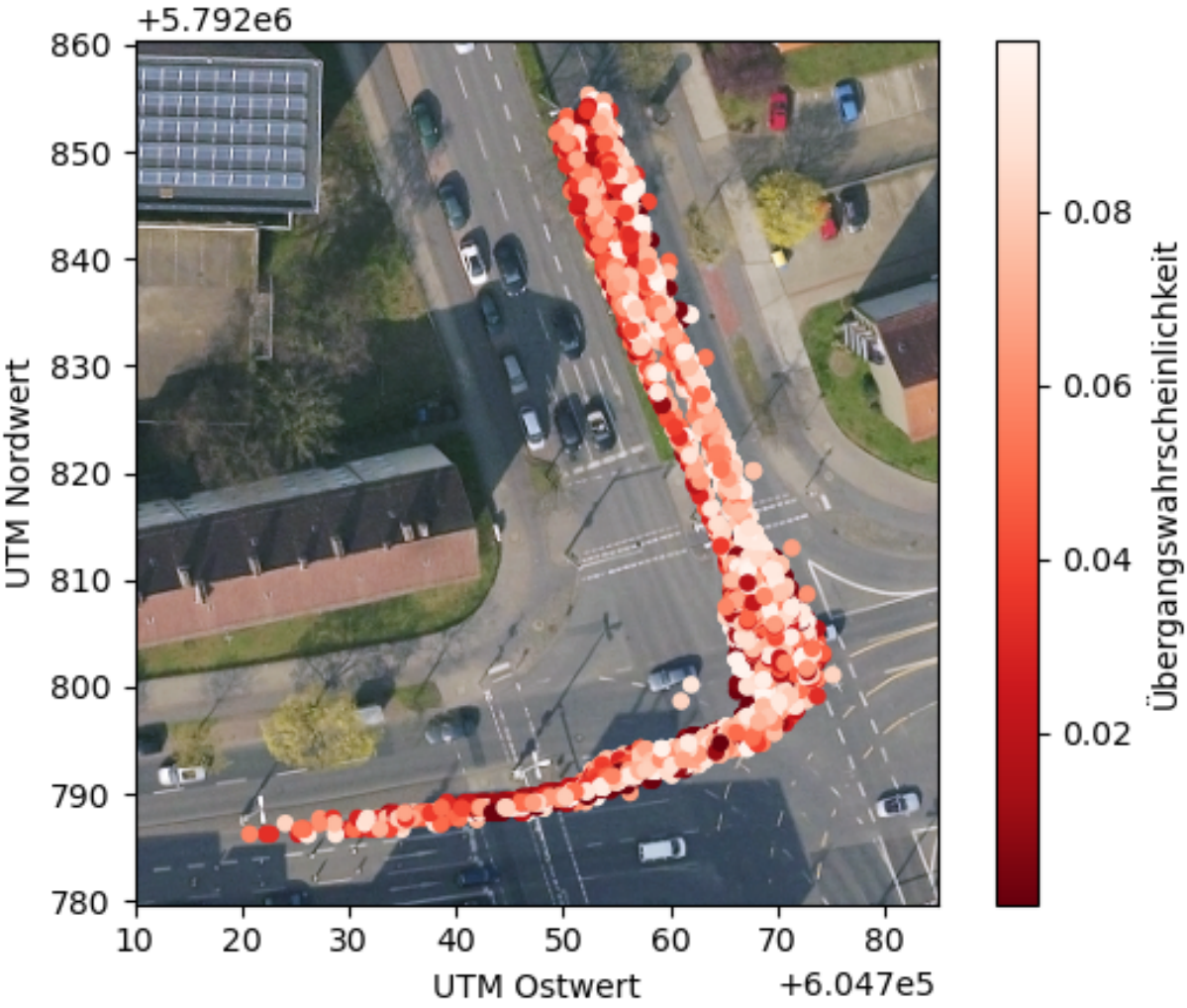


Abbildung 21: Richtungsbasierte Anomalien im Testdatensatz auf Basis der Übergangswahrscheinlichkeit zwischen Zellen unter Verwendung eines Schwellenwertes von 10%. Dargestellt ist der End-Datenpunkt des Übergangs. Die jeweiligen Start-Datenpunkte sind nicht dargestellt.

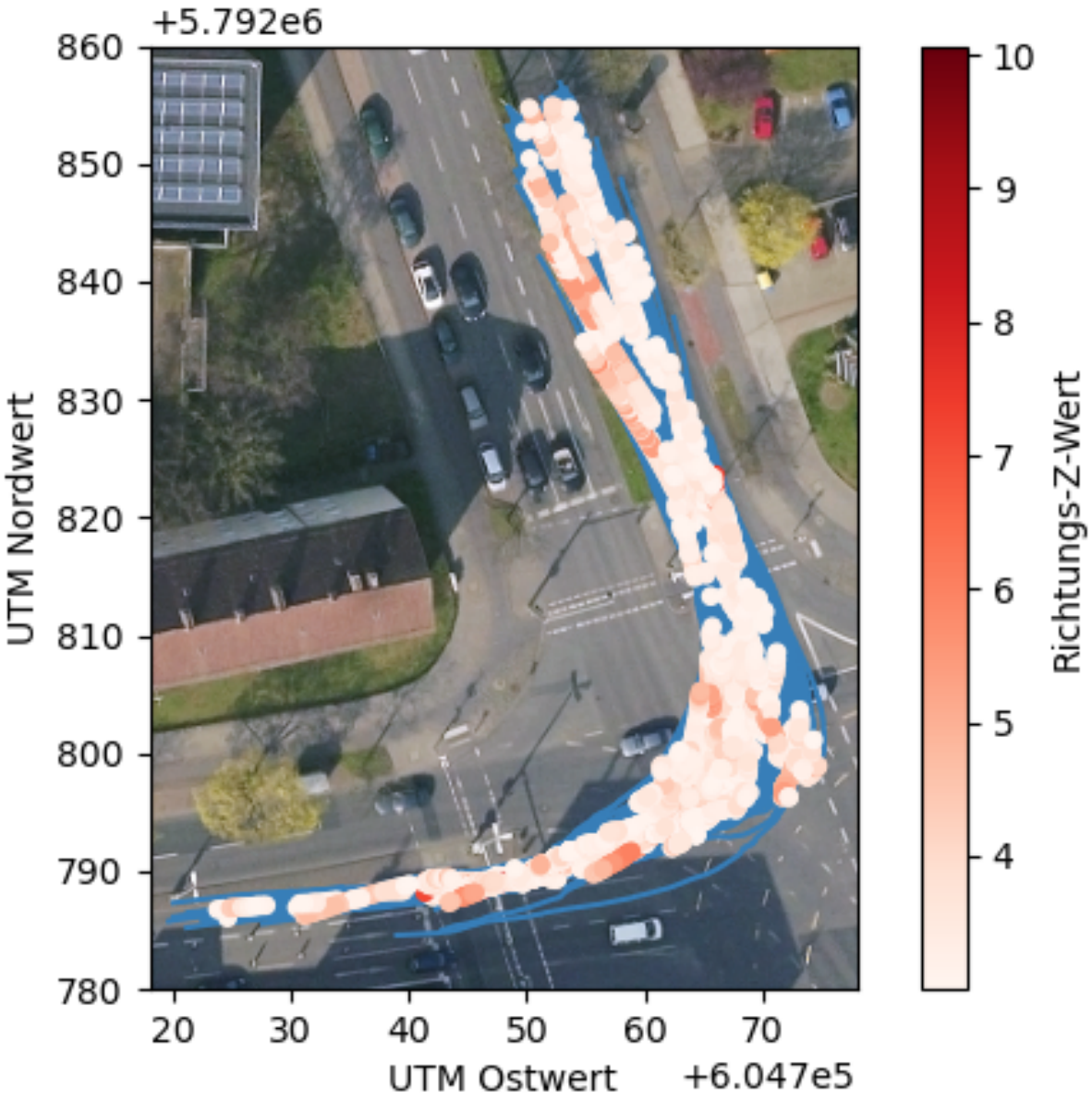


Abbildung 22: Richtungsbasierte Anomalien nach der Wahrscheinlichkeit der Bewegungsrichtung unter Verwendung eines Schwellenwertes von drei Standardabweichungen. Die blauen Linien stellen den Test-Datensatz dar, die Kreise einen anormalen Datenpunkt. Die Farbe der Kreise, gibt den Z-Wert der Bewegungsrichtung des Datenpunktes an, der auf Basis der Trainingsdaten berechnet wird.

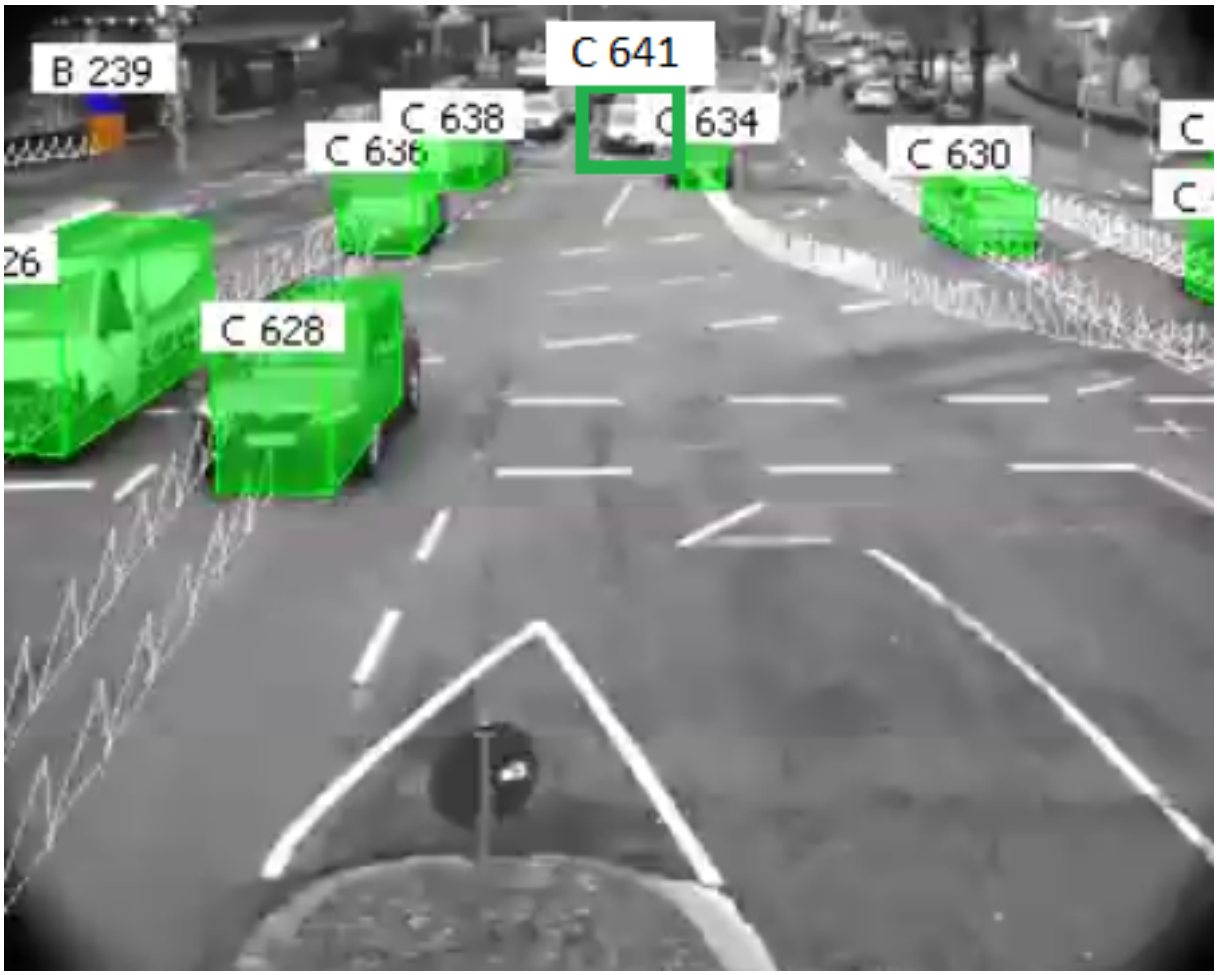


Abbildung 23: Videobild der Kamera, die im Osten auf den Innenbereich der Kreuzung nach Westen ausgerichtet ist. Das Fahrzeug C 641 wird beim Spurwechsel wie im Bild dargestellt nach der Bewegungsrichtung als Anomalie klassifiziert.

Abbildung 24: Videobild der Kamera, die im Süden auf den Innenbereich der Kreuzung nach Norden ausgerichtet ist. Der Verkehrsteilnehmer C 207 wird nach der Bewegungsrichtung in diesem Videobild als Anomalie klassifiziert, da für dieses Fahrzeug aufgrund einer fehlerhaften Positionsbestimmung eine Abweichung von der an dieser Position üblichen Bewegungsrichtung detektiert wurde.

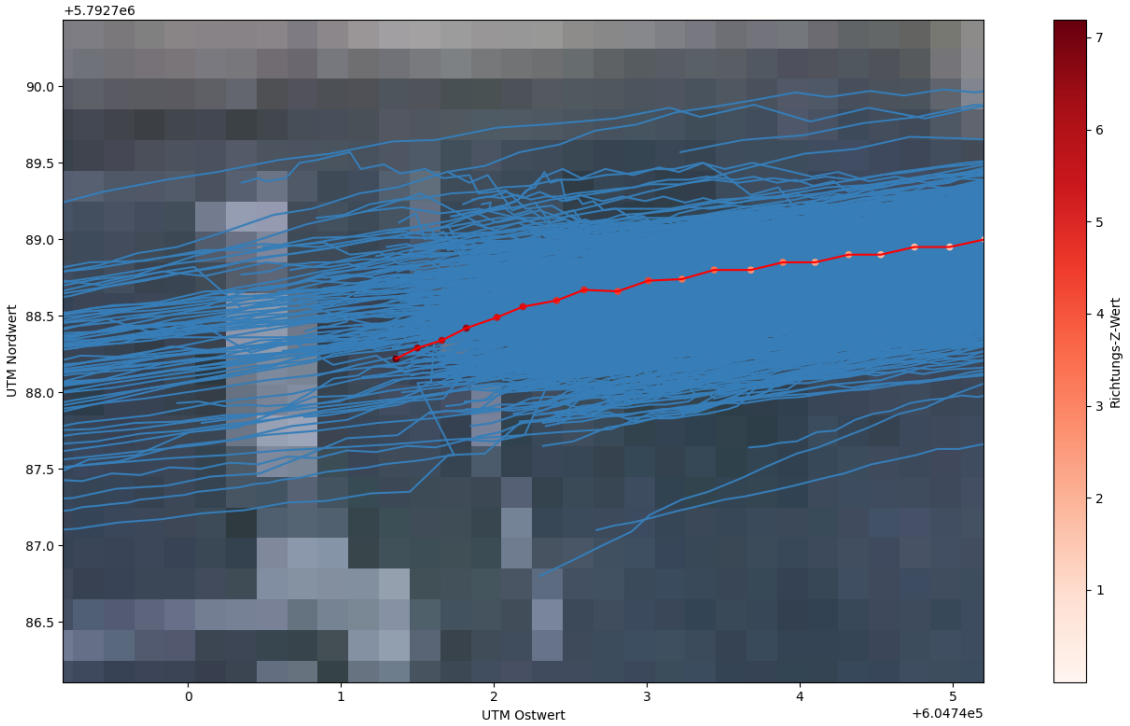


Abbildung 25: Richtungsbasierte Anomalien nach der Wahrscheinlichkeit der Bewegungsrichtung zu Beginn einer Trajektorie unter Verwendung eines Schwellenwertes von fünf Standardabweichungen. Die blauen Linien stellen den Testdatensatz, die rote Linie die anormale Trajektorie und die Kreise einen anormalen Datenpunkt dar. Die Farbe der Kreise gibt den Z-Wert der Bewegungsrichtung des Datenpunktes an, der auf Basis der Trainingsdaten berechnet wird.

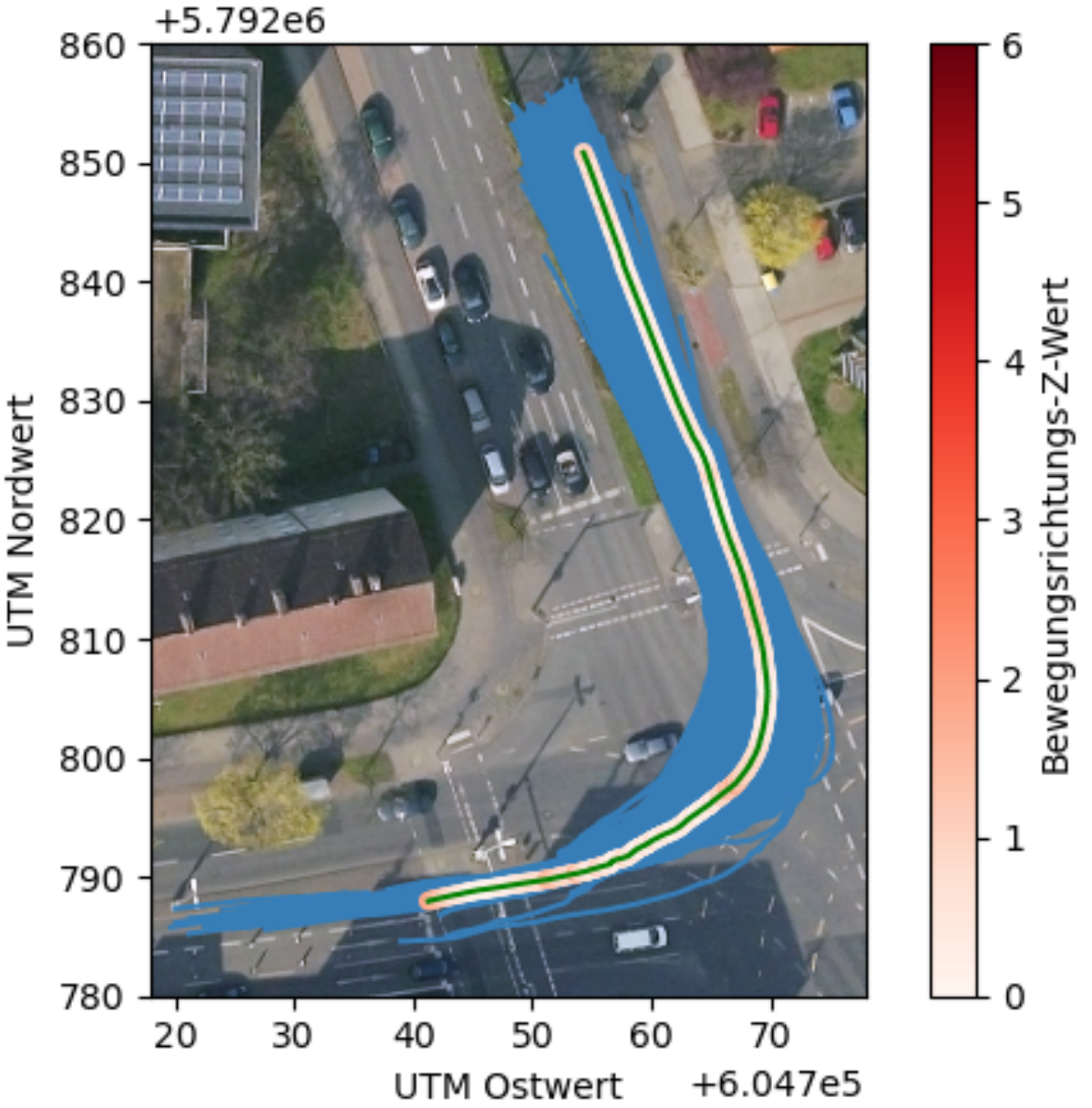


Abbildung 26: Bewertung der durchschnittlichen Trajektorie nach der Bewegungsrichtung unter Verwendung eines Schwellenwertes von fünf Standardabweichungen. Die blauen Linien stellen die Trajektorien aus dem Testdatensatz dar und werden abgebildet, um die rote Linie, die durchschnittliche Trajektorie, mit dem Datensatz vergleichen zu können. Die Farbe der Kreise um jeden Datenpunkt der durchschnittlichen Trajektorie stellt den Z-Wert der Bewegungsrichtung dar.

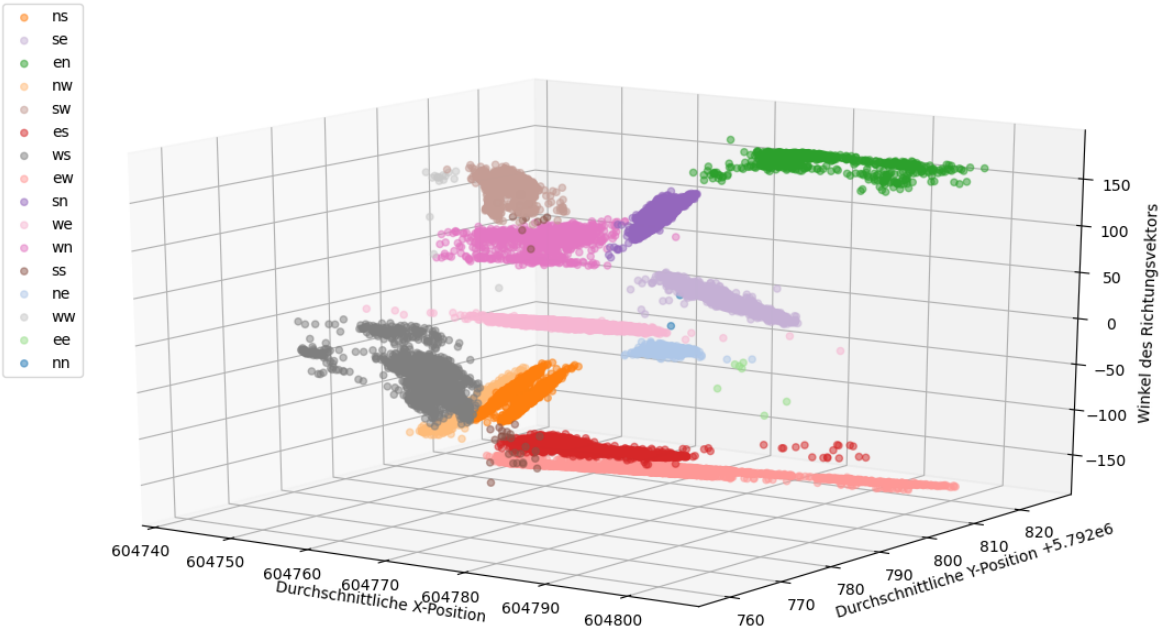


Abbildung 27: Darstellung der Trajektorien als abstrahierte Datenpunkte. Farblich gekennzeichnet nach der zugewiesenen Routen mittels Polygonen. Ohne Trajektorien, denen keine Route zugewiesen werden konnte.

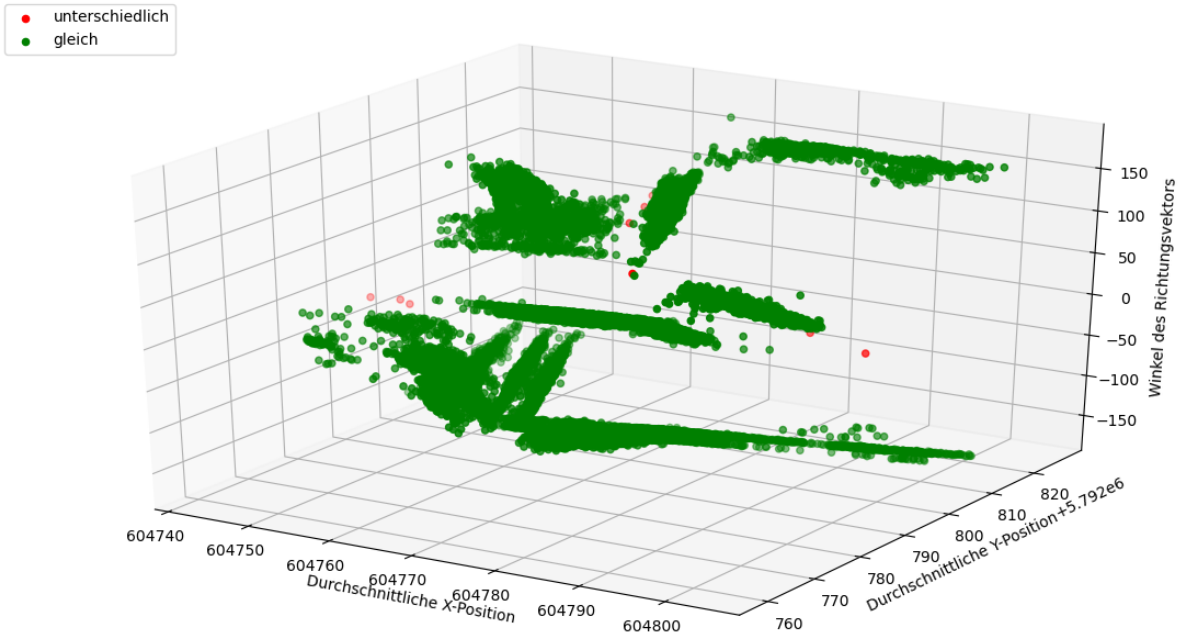


Abbildung 28: Darstellung der Trajektorien als abstrahierte Datenpunkte. Farblich gekennzeichnet, je nachdem ob für die Trajektorien mit mittels Polygonen und GMM der gleichen Route zugewiesen wurde oder nicht. Rote Datenpunkte werden als Anomalien bezeichnet.

Tabelle 1: Anzahl der verwendeten Trainingsdaten und detektierten Anomalien je Verfahren

Datengrundlage	Werte		Trajektorien	
	Anzahl	Anomalien	Anzahl	Anomalien
Länge der Trajektorien	94.044.955	131	369.422	131
Aufzeichnungsfrequenz	93.975.926	27.535	369.291	14.573
Beschleunigung	57.475.254	6	279.022	1
Geschwindigkeit	57.474.388	771	279.021	757
X-Position	57.260.302	876	278.264	753
Y-Position	57.260.302	967	278.264	754
Route	56.915.345	3.367.277	276.822	17.219
Hausdorff-Metrik	2.686.444	-	4.604	-
Übergangswahrscheinlichkeit	1.106.463	-	6.297	-
Bewegungsrichtung	3.670.562	-	6.297	-
Anzahl der Trajektorien	3.670.562	-	6.297	-
Gaußsches Mischmodell	53.255.380	-	258.997	-