Mobile Data Fusion

Automatische Ermittlung der Fahrgastnachfrage aus AFZS-, WLAN-, Bluetooth- und Verbindungsdaten

Gemeinsamer Schlussbericht

Federführer

WVI Prof. Dr. Wermuth Verkehrsforschung und Infrastrukturplanung GmbH
Nordstraße 11, 38106 Braunschweig

Verbundpartner

BLIC - Beratungsgesellschaft für Leit-, Informations- und Computertechnik mbH
Rheinstraße 45, 12161 Berlin

INIT Innovative Informatikanwendungen in Transport-, Verkehrs- und Leitsystemen GmbH
Käppelestraße 4-10, 76131 Karlsruhe

Nordhessischer VerkehrsVerbund Verkehrsverbund und Fördergesellschaft Nordhessen mbH
Rainer-Dierichs-Platz 1, 34117 Kassel

Universität Kassel, Fachgebiet Verkehrsplanung und Verkehrssysteme
Mönchebergstraße 19, 34125 Kassel

Fördergeber

Gefördert durch: Bundesministerium für Digitales und Verkehr aufgrund eines Beschlusses des Deutschen Bundestages

Bundesministerium für Digitales und Verkehr über den Modernitätsfond
Robert-Schumann-Platz 1, 53175 Bonn

Mobile Data Fusion

Automatische Ermittlung der Fahrgastnachfrage aus AFZS-, WLAN-, Bluetooth- und Verbindungsdaten

Gemeinsamer Schlussbericht

Bearbeiter:

WVI	Dr. Antje-Mareike Dietrich
BLIC	Inga Dontsova, Gustav Thiesing
INIT	Dr. Jochen Wendel
NVV	Jessica Luikenga
Uni Kassel	Dr. Ramón Briegel, Dominik Bieland, Prof. Dr. Carsten Sommer

August 2022

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

1 Mobile Data Fusion im Überblick

Das Forschungsprojekt Mobile Data Fusion wurde von Dezember 2018 bis Februar 2022 im Auftrag des Bundesministeriums für Verkehr und Digitalisierung bearbeitet. Im Rahmen der Modernitätsfond (mFUND) Förderlinie 2 „Datenbasierte Forschungs- und Entwicklungsprojekte“ wurde die Arbeit an dem Vorhaben mit 1.743.242 € gefördert. Im Folgenden stellt dieser gemeinsame Schlussbericht die inhaltlichen Ergebnisse der Projektarbeit dar. Zuvor wird der Forschungsansatz in seinem Gesamtkontext gestellt und das Vorgehen bei der Projektarbeit skizziert.

1.1 Ausgangssituation

Der ÖPNV in Deutschland steht vor großen Herausforderungen. Klimaschutz und Luftreinhaltung erfordern die Stärkung des ÖPNV als Rückgrat nachhaltiger Mobilität. Veränderungen im Mobilitätsverhalten führten vor der Corona-Pandemie zu einer wachsenden Nachfrage nach ÖPNV-Dienstleistungen. Inzwischen steht der ÖPNV vor der Herausforderung, Kunden für den ÖPNV (zurück) zu gewinnen und langfristig zu binden. Wie groß das Potenzial dafür ist, zeigen die Nachfragezuwächse, die das 9-Euro-Ticket induziert.

Während mit Infrastrukturinvestitionen und Flottenerweiterungen langfristig reagiert werden kann, gilt es dieser veränderten Nachfrage in den nächsten Jahren durch eine Optimierung der bestehenden Verkehrsnetze Rechnung zu tragen. Um das ÖPNV-Angebot zielgerichtet und kundenorientiert planen zu können, braucht es deshalb dringend verbesserte Datengrundlagen.

Zur genauen Ermittlung des Bedarfs sind quantitative Erhebungen mit Automatischen Fahrgastzählsystemen (AFZS) heute bereits die gängige Praxis. Dadurch kann die Menge an Fahrgästen ermittelt werden. Unbekannt sind jedoch die Wege dieser Fahrgäste, also der Start, das Ziel und ggf. die Umstiege. Diese Nachfragedaten werden bislang aufwändig in der Regel durch Fahrgasterhebungen und Haushaltsbefragungen generiert. Trotz des großen Aufwandes sind die erhobenen Daten häufig begrenzt hinsichtlich der zeitlichen Aussagekraft (in der Regel Verhalten an einem Stichtag) und der Aktualität (Erhebungsrhythmus alle 3 bis 10 Jahre).

Der mit Mobile Data Fusion verfolgte Ansatz beruht auf der Idee, zusätzliche Daten für die Ermittlung der ÖPNV-Nachfrage zu nutzen. Als Datenquellen kommen unter anderem WLAN- und Bluetooth-Signale mobiler Endgeräte, Anfragen an die Verbindungsauskunft und die mCloud des Bundesministeriums für Verkehr und digitale Infrastruktur in Frage.

1.2 Projektziele

Ziel des Projekts ist die Entwicklung eines Verfahrens, das umfangreiche Daten aus den verschiedenen Quellen erfasst, aufbereitet und zusammenführt, um ÖPNV-Betreibern genaue und aktuelle Informationen zur Fahrgastnachfrage automatisiert bereitstellen zu können. Im Fokus stehen insbesondere Informationen über Quelle-Ziel-Verflechtungen und Umsteigeströme. Das Verfahren soll datenschutzkonform sein.

Durch die Verwendung von bereits vorliegenden Datenbeständen, die im Wesentlichen von den ÖPNV-Kunden durch Ihre Smartphone-Nutzung erzeugt werden (WLAN, Bluetooth, Anfragen an die Fahrplanauskunft), soll geprüft werden, in welchem Umfang Fahrgastbefragungen ersetzt werden können und insbesondere die aufwändige Befragung zur Reiseroute zukünftig entfallen kann. Zielsetzung ist somit die Abbildung der Quelle-Ziel-Verflechtungen einschließlich der Umsteigeströme zwischen den einzelnen Linien durch die Verknüpfung der technisch vorliegenden Datenbestände. Gegebenenfalls besteht sogar die Möglichkeit, zusätzliche Informationen zur Regelmäßigkeit und Häufigkeit der ÖPNV-Nutzung zu gewinnen, die bislang für den Planungsprozess nur in rudimentärem Umfang zur Verfügung stehen.

1.3 Stand der Wissenschaft und Technik

Informationen zu den Quelle-Ziel-Verflechtungen in einem ÖPNV-Netz sowie zu den Umsteigevorgängen an einzelnen Haltestellen werden heutzutage in der Regel über personalintensive und damit aufwändige Fahrgasterhebungen erfasst. Diese Erhebungen werden aus Kostengründen nur als Querschnittserhebung für einen Stichtag im Abstand mehrerer Jahre punktuell oder flächendeckend für ein Bedienungs- bzw. Verbundgebiet durchgeführt. Befragungen, in denen das intrapersonelle Nutzungsverhalten über einen längeren Zeitraum von Wochen oder Monaten erfasst wird, sind unüblich und aus Gründen der Belastung für die Befragten praktisch nicht möglich. Die mangelnde Aktualität der erfassten Daten führt dazu, dass diese im täglichen Betrieb nur eine geringe Rolle spielen.

1.3.1 WLAN, Bluetooth und Fahrplanauskunft

Der (kostenlose) WLAN-Zugang für ÖPNV-Kunden wird zunehmend auch im deutschen Nahverkehr möglich. Aktuell sind bereits verschiedene Regionalbuslinien bspw. in den Bedienungsgebieten des Münchner Verkehrs- und Tarifverbundes (MVV) und des NVV mit der entsprechenden Technik ausgestattet. Im Rahmen des Projektes „BayernWLAN“ sollen neben öffentlichen Verkehrsmitteln und Haltestellen auch öffentliche Plätze, Schulen und Universitäten mit WLAN-Hotspots ausgestattet werden.

Die anfallenden WLAN-Daten werden in Deutschland teilweise zur Analyse des Einkaufsverhaltens genutzt, etwa um Kundenfrequenzen, Bewegungsprofile im Ladengeschäft, Verweildauern oder Konversionsraten zu bestimmen. Im ÖPNV bleiben die Vorteile der ohnehin anfallenden Daten bislang in der Regel ungenutzt. Dabei liegen international einige Forschungsberichte vor, die die Möglichkeiten der Nutzung von WLAN-Daten beschreiben (Song & Wynter 2017, Baeta et al. 2016, Sapiezynski et al. 2015). Inzwischen wurden Erfahrungen in London durch den ÖPNV-Betreiber „Transport for London" mit Tracking in den U-Bahn-Stationen bekannt. Auch der Austausch mit den Neederlandse Spoorwege ergab, dass Fahrgastströme mittels WLAN-Sniffing an großen Umsteigeknoten ermittelt werden können.

Darüber hinaus konnten die Wege zwischen den Stationen bzw. zwischen den Bahnsteigen, das Verhalten bei Störfällen oder die Auslastung von Fahrzeugen ermittelt werden (O'Malley 2017). In einer weiteren Studie wurden die Daten zur Bestimmung der Fahrzeugauslastung in einem Shuttlebus-System der Thammasat University in Thailand verwendet (Pattanusorn et al. 2016). Obwohl die Nutzbarkeit der Daten in der Forschung unbestritten ist, mangelt es bisher an Anwendungen in der Praxis. Die fehlende Umsetzung kann zumindest teilweise auf die Anforderungen des Datenschutzes zurückgeführt werden. In Deutschland werden beispielsweise zahlreiche Diskussionen darüber geführt, ob die MAC-Adresse als personenbezogenes Datum gewertet wird, sodass die Erfassung und Speicherung datenschutzrechtliche Handlungen erfordert.

Bluetooth wird bereits zur Erfassung von Kfz-Verkehrsströmen und zur Ermittlung des Verkehrszustands genutzt, beispielsweise in den Städten Bonn und Konstanz. Bei der Identifizierung von Smartphones kommt Bluetooth aktuell eine untergeordnete Rolle zu, da Smartphones auch bei aktiviertem Bluetooth standardmäßig unsichtbar und damit nicht detektierbar sind.

Im mFund Projekt ProTrain wurden Daten aus Routensuchabfragen der Fahrplanauskunftssysteme dahingehend analysiert, ob sie für eine Belegungsprognose herangezogen werden können (ProTrain 2020). Im urbanen Umfeld erweist sich der zeitliche Vorlauf der größten Zahl der Verbindungsabfragen als unzureichend, wohingegen Verbindungsanfragen für den Wochenendausflugsverkehr zumindest in Fusion mit anderen Datenquellen eine Indikation für die zu erwartende Belegung geben können. Parallelen ergaben sich auch zum Forschungsprojekt Tariftool-XL, welches ebenfalls vom BMDV gefördert wurde (Dietrich et al. 2018). In diesem Projekt wurde geprüft, ob Daten, die bei der Anfrage an Fahrplanauskunftssysteme anfallen, für die Erzeugung künstlicher Fahrtenmuster genutzt werden können. Die Betrachtung führte zu dem Schluss, dass die in diesem Projekt verfügbaren Auskunftsdaten nicht ohne Weiteres Rückschlüsse auf die realen Verflechtungen zulassen. Dies liegt vor allem an einem sehr hohen Anteil an redundanten Abfragen und abgefragten Teilwegen (statt Gesamtwegen).

1.3.2 GSM- Daten

Die bei Netzbetreibern anfallenden Mobilfunkdaten können, wie eine Studie des Fraunhofer IAO in Kooperation mit Telefónica NEXT und Teralytics zeigt, einen positiven Beitrag zur Verkehrsplanung leisten (Schmidt & Männel 2017). Mobilfunkdaten fallen durchgehend an – bei Telefonica Deutschland werden beispielsweise täglich rund fünf Milliarden Datenpunkte von über 46 Mio. Anschlüssen erzeugt (Stichprobe) (Telefónica NEXT 2018). Somit liegen zeitlich hochaufgelöste Daten mit einer räumlichen Differenzierung auf Postleitzahlebene vor. Durch diese Datenbasis ist es möglich, verschiedene Analysen durchzuführen und somit Erkenntnisse hinsichtlich von Bewegungsströmen, Tagesganglinien oder auch Einzugsbereichen von Großveranstaltungen zu erlangen. Es ist unbestritten, dass auch die Nutzung von Mobilfunkdaten Erhebungen wie Haushaltsbefragungen ergänzen können (Bamberger et al., 2017). Telefonica NEXT kooperierte bereits in verschiedenen Projekten mit Anwendungs- und Forschungspartnern um die Nutzbarkeit der Daten, u.a. im Verkehrssektor zu verdeutlichen (Schmidt & Männel 2017, flinc GmbH 2016, ProTrain 2020).

Die Mobilfunkdaten basieren auf der GSM-Ortung, ein funkzellenbasierendes Verfahren, das den Standort eines eingeschalteten Mobiltelefons bestimmen kann. Die Genauigkeit der Standortbestimmung ist dabei von verschiedenen Einflussfaktoren, wie etwa der Ausstattung des Raums mit Basisstationen bzw. Funkmasten, gerätespezifischen Eigenschaften des Mobilfunktelefons sowie dem genutzten Verfahren zur Positionsbestimmung abhängig. Zur Positionsbestimmung werden unter anderem die

- Cell-ID-Methode und Timing Advance oder das
- EOTD-Verfahren (Enhanced Observed Time Difference)

eingesetzt. Die Cell-ID-Methode basiert auf der Ortung auf Grundlage der Funkzelle, in der das Mobiltelefon eingebucht ist. Die Reichweite der Funkzellen und damit auch die Genauigkeit der Ortung variieren von ca. 100 bis 200 m im städtischen Raum bis hin zu über 30 km in ländlichen Regionen. Dabei gilt: Je kleiner der Radius der Funkzelle, desto genauer kann die Ortung stattfinden. Zur Erhöhung der Genauigkeit kann der Parameter Timing Advance (Laufzeit) genutzt werden. Der Parameter gibt Aufschluss über den Abstand zwischen dem Mobiltelefon und der Basisstation. Das EOTD-Verfahren nutzt Signale von mehreren Basisstationen und ermöglicht auf Grundlage der Zeitdifferenzen (Ankunftszeiten der Signale von mehreren Basisstationen) eine genauere Positionsbestimmung. Die Genauigkeit der Ortung kann dadurch auf bis zu 30 m verbessert werden. Für das EOTD-Verfahren müssen die Mobilfunkgeräte jedoch ausgelegt sein. Darüber hinaus ist die Möglichkeit zur Kommunikation mit mehreren Basisstationen in vielen ländlichen Gebieten nicht gegeben. Um die Genauigkeit der Ortung zu erhöhen, wäre ein Ausbau der Netzinfrastruktur notwendig (Schelewsky 2014, Elektronik-Kompendium.de 2018).

In Abhängigkeit vom Untersuchungsziel, der Planungsebene sowie dem Verkehrszellen- und Netzmodell werden unterschiedlich Genauigkeiten der Positionsdaten benö-

tigt. Die Ermittlung von Quelle-Ziel- Matrizen auf Bundesebene (Bundesverkehrswegeplanung) hat beispielsweise geringere Ansprüche an die Genauigkeit der Ortung als vergleichbare Untersuchungen auf regionaler oder kommunaler Ebene (Sommer 2002).

Die Nutzung von Mobilfunkdaten bietet derzeit auf Ebene von Postleitzahl-Bezirken große Potentiale für die strategische Verkehrsplanung, ist jedoch für den in diesem Projekt angestrebten Anwendungsfall nicht zielführend. Verkehrsunternehmen und -verbünde statten ihre Fahrzeuge (Busse, Tram) zunehmend mit WLAN aus und schaffen somit (unbeabsichtigt) auch die Infrastruktur, um die Projektidee von Mobile Data Fusion umzusetzen. Die Verkehrsunternehmen haben so die Möglichkeit, die Daten eigenständig zu erfassen, sodass sie sich weder in eine Abhängigkeit von den Netzbetreibern begeben noch durch eine andauernde finanzielle Belastung für den Kauf von Daten belastet werden. Sie erhalten damit Längsschnittdaten zur ÖPNV-Nachfrage für sehr lange Zeiträume, mit einer höheren Genauigkeit als dies mit Mobilfunkdaten flächendeckend möglich wäre und zu einem geringeren Kostenaufwand als bei einer Nutzung von Mobilfunkdaten. Weitere Analysetools zur Abgrenzung verschiedener Verkehrsmittel in den Mobilfunkdaten sind zudem erforderlich und waren zum Start des Projektes noch nicht verfügbar (Trialog Publishers Verlagsgesellschaft 2017).

1.4 Projektdesign

Das Projektkonsortium bestand aus der Prof. Dr. Wermuth Verkehrsforschung und Infrastrukturplanung GmbH (WVI), der INIT GmbH, der BLIC GmbH, dem Nordhessischen VerkehrsVerbund (NVV) und dem Fachbereich Verkehrsplanung und Verkehrssysteme (VPVS) der Universität Kassel.

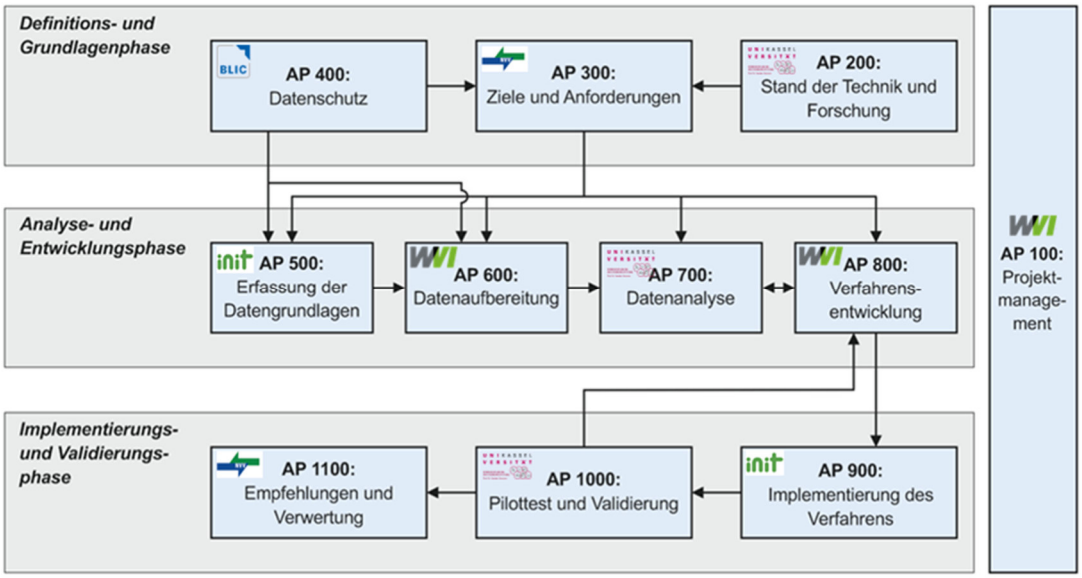


Abbildung 1: Übersicht Projektdesign

Das Forschungsprojekt war in insgesamt elf Arbeitspakete (AP) untergliedert (vgl. Abbildung 1). Dabei werden zusätzlich zum Projektmanagement (AP 100) inhaltlich drei aufeinanderfolgende, teilweise rückgekoppelte Phasen unterschieden. In der Definitions- und Grundlagenphase (AP 200 bis 400) werden unter Berücksichtigung des aktuellen Forschungsstandes Ziele und Anforderungen an das Verfahren definiert sowie das Datenschutzkonzept erstellt. Die darauffolgende Analyse- und Entwicklungsphase (AP 500 bis 800) bildet den Schwerpunkt des Forschungsvorhabens, da hier basierend auf einer umfangreichen Erfassung, Aufbereitung und Analyse der Eingangsdaten ein Verfahren für die Abschätzung der Fahrgastnachfrage entwickelt wird. In der abschließenden Implementierungs- und Validierungsphase (AP 900 bis 1100) wird das entwickelte Verfahren implementiert und in Pilottests validiert. Die Arbeitspakete unterliegen dabei einem iterativen und vernetzten Prozess, sodass die Qualität der Forschungsergebnisse erhöht wird.

Für die Bearbeitung waren als Zwischenziele die beiden folgenden Meilensteine vorgesehen:

- Meilenstein 1 (nach 9 Monaten): Die Ziele und Anforderungen an das Verfahren liegen vor (AP 300), ebenso das grundlegende Datenschutzkonzept (AP 400).
- Meilenstein 2 (nach 24 Monaten): Das Verfahren zur Berechnung der Nachfrage liegt vor (AP 800).

Die WVI übernahm die Rolle der Konsortialführerin und verantwortete AP 100 Projektmanagement, AP 600 Datenaufbereitung und AP 800 Verfahrensentwicklung. Die INIT übernahm die Hard- und Softwareentwicklung und verantwortete AP 500 Datenermittlung und AP 900 Produktivimplementierung. Die BLIC war für die Erarbeitung des Datenschutzkonzepts (AP 400) zuständig. Der Praxispartner NVV stellte seine Fahrzeuge und Haltestellen zur Verfügung, hielt den Kundenkontakt und verantwortete AP 300 Ziele und Anforderungen sowie AP 1100 Empfehlungen und Verwertung. Das VPVS begleitete die Verfahrensentwicklung von wissenschaftlicher Seite und verantwortete AP 200 Stand der Technik und Forschung, AP 700 Datenanalyse sowie AP 1000 Pilottest und Validierung.

1.5 Projektmanagement

Für die Bearbeitung der Arbeitspakete war der in Abbildung 2 grau hinterlegte zeitliche Ablauf geplant. Aufgrund der Pandemie-Situation sowie technischer und zum Teil organisatorischer Hürden musste im Projektverlauf flexibel reagiert werden. Im Wesentlichen hatte dies zur Folge, dass die Datenanalyse, die Verfahrensentwicklung sowie die Implementierung parallel bearbeitet wurden und dabei iterativ vorgegangen wurde. Diese Herangehensweise erwies als zielführend. Beide Meilensteine konnten während der Projektlaufzeit erreicht werden und der Pilottest, inklusive Befragung durchgeführt werden.

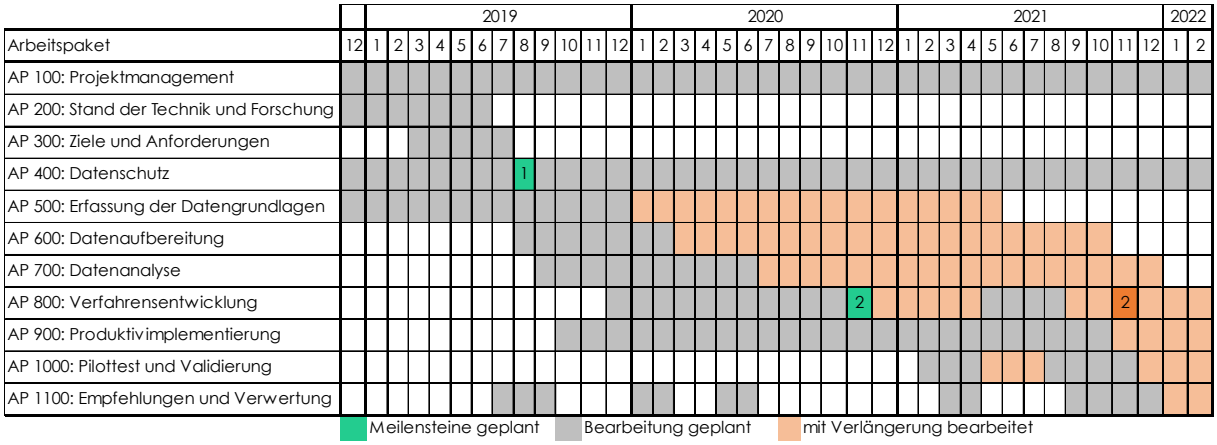


Abbildung 2: Zeitliche Bearbeitung der Arbeitspakete

Die erfolgreiche Zusammenarbeit wurde durch das Projektmanagement unterstützt, indem zu Projektbeginn eine Abstimmung der Arbeitsweise und Rollen der Partner im Projekt erfolgte. Wie in Tabelle 1 skizziert, wurde eine regelmäßige Kommunikation im Gesamtprojekt sowie in kleinen Arbeitsgruppen von Beginn an gelebt. Es wurde zudem eine gemeinsame Ablage in der WVI-eigenen Cloud angelegt. Auf eine gemeinsame Projekthomepage wurde verzichtet, da bereits die mFUND-Begleitforschung einen Webauftritt für jedes Projekt vorsah.

(https://www.bmvi.de/SharedDocs/DE/Artikel/DG/mfund-projekte/mobiledatafusion.html).

Name	Turnus	digital	Vor Ort	
Lenkungskreis	halbjährlich	x	x	
Gesamtprojekttreffen	halbjährlich	x	x	
Telko Gesamtprojekt	monatliche	x		
Telko Datenerfassung	wöchentlich	x		
Telko Verfahrensentwicklung	monatlich	x		
Telko Implementierung	zwei-wöchentlich	x		
Spezifische Themen-Workshops	bei Bedarf	x	x	

Tabelle 1: Übersicht Zusammenarbeit

Zusätzlich zur inhaltlichen Projektarbeit war ein wesentlicher Bestandteil der Projektarbeit der Austausch und die Vernetzung innerhalb des Förderprogramms sowie darüber hinaus mit der interessierten (Fach-)Öffentlichkeit. Das Projekt wurde durch das Projektkonsortium bei diversen Veranstaltungen präsentiert. Die wichtigsten Veranstaltungen sind in Tabelle 2 aufgeführt. Neben den hier genannten Präsentationen im Rahmen

von Veranstaltungen hat das Projektkonsortium bereits die folgenden Zeitschriftenartikel mit Bezug zum Forschungsprojekt „Mobile Data Fusion“ veröffentlicht:

- Automatische Erfassung der Fahrgastnachfrage, Nahverkehrspraxis, Ausgabe 6-2019, S. 34-36, von Carsten Sommer und Dominik Bieland.
- Aktuelle Verkehrsnachfragedaten für eine flexiblere ÖPNV-Planung, Nahverkehrspraxis, Ausgabe 3-2021, S. 40-41, von Jochen Sauer.
- Multimodale Kunden brauchen ein flexibles Angebot – Hilft die Digitalisierung dem ÖPNV weiter?, DER NAHVERKEHR, 4/2021, S. 44-46, von Antje-Mareike Dietrich.
- Automatisch und aktuell – Datengrundlagen für eine flexiblere ÖPNV-Planung, Internationales Verkehrswesen, 02/2022, von Antje-Mareike Dietrich.

Darüber hinaus befinden sich die folgenden Veröffentlichungen in der Planung:

- Dissertation Dominik Bieland am FG VPVS der Universität Kassel: Ableitung von Quelle-Ziel-Matrizen auf Basis von WLAN- und Bluetooth-Daten am Beispiel ausgewählter Buslinien im NVV.

Name	Turnus	digital
Kick-Off Hypermotion	21.11.2018	Frankfurt
Bus2Bus	20.03.2019	Berlin
mFUND-Forum Standardisierung	28.08.2019	Bonn
6. Netzwerktreffen Verkehrserhebungen in Verbünden	16.09.2019	Kassel
3. mFUND-Konferenz	27.09.2019	Berlin
mFUND-Forum Datenschutz	07.10.2019	Bonn
Hypermotion	26.11.2019	Frankfurt
mFUND-Forum Datenschutz	05.12.2019	Berlin
WIK Fachgespräch Mobilität im ländlichen Raum	27.02.2020	Berlin
mFUND-Forum Datenschutz & Compliance	31.03.2020	digital
mFUND-Forum Standardisierung & mCloud	15.06.2020	digital
mFUND-Fachaustausch Schienenverkehr	07.07.2020	digital
mFUND-Fachaustausch App-gestützte Erhebungen	20.10.2020	digital
mFUND-Konferenz Mit Dateninnovation zur Mobilität der Zukunft	19.10.2021	digital
Init-WVI-Webinar Nutzung von WLAN- und Bluetooth-signalen zur Abbildung von Fahrgastströmen	30.11.2021	digital

Tabelle 2: Öffentlichkeitsarbeit.

2 Inhaltliche Ergebnisse

Im Folgenden werden die inhaltlichen Ergebnisse der Projektarbeit im Detail vorgestellt. Bei der Darstellung wird sich an dem in Kapitel 1.4 skizzierten Projektdesign orientiert. Die Verantwortlichkeiten für die jeweiligen Arbeitspakte können dort bei Bedarf nachgeschlagen werden. Die Darstellung vieler wesentlicher Ergebnisse (Abschnitte 2.1, 2.2, 2.6.1.1, 2.6.1.2, 2.6.1.4, 2.6.2, 2.7.4) ist der Dissertation von Dominik Bieland am FG VPVS der Universität Kassel entnommen (Bieland 2022); dort sind auch weitere Details zu finden.

2.1 Stand der Technik und Forschung

Während der Projektbearbeitung wurden fortlaufend Recherchen unternommen. Im Fokus standen die ÖPNV-Erhebungsmethoden zur Ermittlung der Fahrgastnachfrage im Allgemeinen sowie im Speziellen die nutzbaren Datenquellen für den in Mobile Data Fusion angedachten Bearbeitungsansatz.

2.1.1 Datenquellen zur Erfassung der Verkehrsnachfrage im ÖPNV

Zu Beginn der Projektarbeit wurden die unterschiedlichen Datenquellen zur Erfassung der Verkehrsnachfrage im ÖPNV hinsichtlich ihrer wesentlichen Vor- und Nachteile gegenübergestellt und bewertet.

„Klassische“ Datenquellen zur Ermittlung der Verkehrsnachfrage sind:

- manuelle Fahrgastzählung
- Fahrgastbefragungen
- Verkauf von (Papier-)Fahrausweisen.

In Zeiten der Digitalisierung sind neue Datenquellen hinzugekommen:

- Automatische Fahrgastzählsysteme (AFZS)
- Verkauf elektronischer Fahrausweise / Elektronisches Fahrgeldmanagement (EFM)
- elektronische Verbindungsauskunft
- Tracking-Daten (GSM-Mobilfunkdaten, WLAN- und Bluetooth-Daten)

Ein gemeinsamer Nachteil manueller Erfassungsmethoden (manuelle Fahrgastzählungen und Fahrgastbefragungen) liegt darin, dass diese Erhebungsarten einerseits einen relativ hohen zeitlichen und finanziellen Aufwand erfordern, andererseits aber mit einem Stichprobenfehler und unter Umständen zusätzlich mit systematischen Fehlern (mangelnde Repräsentativität) behaftet sind. Sie liefern nur eine Momentaufnahme des in der Regel kurzen Erhebungszeitraums. Aufgrund des hohen Aufwands werden

solche Erhebungen nur in großen zeitlichen Abständen durchgeführt, so dass in vielen Fällen keine aktuellen Daten vorliegen.

Demgegenüber besitzen alle anderen Datenquellen, da sie mit kontinuierlicher, automatisierter Erfassung arbeiten, den Vorteil, mit geringen laufenden Kosten stets aktuelle Daten zu liefern, die auch längere Zeiträume abdecken. Eine automatisierte Erfassung erfordert allerdings in vielen Fällen einen hohen Investitionsaufwand für die Hardwareausstattung.

Manuelle Erfassungsmethoden sind stets fehleranfällig; bei automatischen (d.h. maschinellen) Erfassungsmethoden hängt die Datenqualität entscheidend von der technischen Zuverlässigkeit und Robustheit der verwendeten Hardware sowie der Qualität der Software zur Verarbeitung der Daten ab.

Manche automatisch erfassten Datenquellen erfassen nur einen Teil der Fahrgastfahrten. Fahrgäste stellen nicht für jede Fahrt eine Verbindungsanfrage, so dass für viele Fahrten keine Daten aus der Verbindungsauskunft existieren. Verkaufsdaten erfassen nur bei Einzelfahrkarten die einzelnen Fahrgastfahrten (und dabei oft auch nur die Preisstufe ohne konkreten Start- und Zielort), die Verkäufe von Zeitkarten lassen nur mittels zusätzlicher Befragungen zur Nutzungshäufigkeit eine Schätzung der Fahrtenzahl zu. Fahrgäste, die kein eingeschaltetes Smartphone oder ähnliches Endgerät mit sich führen, können vom Tracking nicht erfasst werden.

Umgekehrt enthalten die automatisch erfassten Daten aus manchen Quellen auch Störddaten, das heißt Daten, die nicht von der interessierenden Grundgesamtheit stammen. Bei angefragten Verbindungen ist unbekannt, ob die anfragende Person die entsprechende Fahrt auch tatsächlich (so) durchgeführt hat. Es sei denn, es wurde im Zusammenhang mit der Verbindungsauskunft auch gleich eine Fahrkarte gekauft. Bei WLAN- und Bluetooth-Daten werden auch Signale von Personen bzw. Geräten erfasst, die sich außerhalb des ÖPNV-Fahrzeugs befinden. Die automatische Erkennung und Filterung solcher Störddaten stellt eine Hürde für die Nutzung solcher Datenquellen dar (vgl. Kapitel 2.1.3).

AFZS-Daten haben hingegen kaum mit derartigen Problemen von Unter- und Übererfassung (Störddaten) zu kämpfen, sondern erfassen zielgenau und meist relativ zuverlässig die Anzahl der an den einzelnen Haltestellen ein- und aussteigenden Fahrgäste. Daraus lässt sich die Belegung auf jedem Abschnitt zwischen zwei Haltestellen und auch die gesamte Verkehrsleistung berechnen. Die für die Angebotsplanung besonders wichtigen Informationen zu Quelle und Ziel der Fahrgastfahrten lassen sich aus AFZS-Daten jedoch nicht ableiten, sondern können nur durch Befragungen oder Tracking-Daten ermittelt werden. GSM-Mobilfunkdaten weisen dabei den Nachteil einer fallweise großen räumlichen Unschärfe auf, da die Ortung auf der Einbuchung der Smartphones in Funkzellen beruht, deren Größe im ländlichen Raum 10 km deutlich überschreiten kann.

Fazit

Zusammenfassend lässt sich festhalten, dass nur Befragungen die Möglichkeit bieten, diverse Informationen von den Fahrgästen zu erfassen, die durch automatische Erfassungen prinzipiell nicht ermittelt werden können, wie z.B. den Wegezweck und demografische Daten. Sie sind allerdings mit den Nachteilen erheblicher Kosten, mangelnder zeitlicher Abdeckung und fraglicher Repräsentativität behaftet. Für alle Anwendungsfälle, die *nicht* auf derartige Zusatzinformationen angewiesen sind, stellen AFZS-Daten in Kombination mit WLAN- und/oder Bluetooth-Tracking-Daten hingegen im Prinzip eine im Betrieb kostengünstige, stets aktuelle Datenquelle für räumlich und zeitlich hoch aufgelöste Quelle-Ziel-Informationen dar.

2.1.2 Technische Grundlagen zur Erfassung von WLAN- und Bluetooth-Signalen

WLAN

Ein Wireless Local Area Network (WLAN) ist ein drahtloses lokales Funknetzwerk, in dem elektronische Geräte Daten mittels elektromagnetischer Wellen austauschen können. Zur Vermeidung von Interferenzen beim Datenaustausch durch mehrere Netzwerke bzw. Geräte in räumlicher Nähe und somit zur Sicherstellung einer fehlerfreien Übertragung nutzen die Geräte verschiedene Kanäle – d.h. schmale Frequenzbereiche – innerhalb der gesamten für WLAN zugeteilten Frequenzspektren: Beispielsweise stehen im 2,4-GHz-Frequenzspektrum 13 Kanäle zur Verfügung. Ein einzelnes Gerät kann dabei nacheinander im Wechsel verschiedene Kanäle nutzen. Dies tun Smartphones insbesondere dann, wenn sie versuchen, sich neu mit einem WLAN-Access Point zu verbinden: Das Smartphone versendet in schneller Folge so genannte *Probe Requests* auf verschiedenen Kanälen, um Netzwerke in der Nähe aufzufinden. Diese Probe Requests können von geeigneten Sensoren erfasst und aufgezeichnet werden. Die Probe Requests enthalten unter anderem folgende Informationen:

- **MAC-Adresse** (media access control address): Eine dem Endgerät (theoretisch) weltweit eindeutig zugeordnete Identifikationsnummer. Diese wird jedoch von den Smartphone-Betriebssystemherstellern aus Datenschutzgründen zunehmend *randomisiert*, das heißt es wird nicht die „wahre", sondern eine in häufigem Wechsel zufällig erzeugte neue MAC-Adresse übermittelt (siehe Kapitel 2.1.4).
- **Information Elements**: Das sind Informationen darüber, welche technischen Eigenschaften und Fähigkeiten das Smartphone hat, was für den Verbindungsaufbau und Datenaustausch zwischen dem Smartphone und dem Access Point von Relevanz ist. Die Information Elements gliedern sich in *HT Capabilities*, *Extended Capabilities* und *Supported Rates*.
- **Vendor**: Die Herstellerfirma des Smartphones (in vielen Fällen „unknown").
- **Sequence Number**: Einen Zähler, der mit jedem versendeten Probe Request um 1 erhöht wird. Bei Erreichen eines Maximalwerts (4096) wird der Zähler zurückgesetzt.

- Zusätzlich zu den genannten versandten Informationen kann beim Empfang durch den Sensor die empfangene Signalstärke (**RSSI**, relative signal strength indicator) aufgezeichnet werden.

Mit den in den Probe Requests übermittelten Informationen kann der Access Point überprüfen, ob ein Datenaustausch entsprechend den Spezifikationen des Endgeräts grundsätzlich möglich ist. Probe Requests werden in der Regel zeitlich gebündelt in Form so genannter *Bursts* ausgesandt. Innerhalb eines Bursts werden die gleichen Informationen nacheinander auf verschiedenen Kanälen versandt. Die Probe Requests innerhalb eines Bursts werden innerhalb einiger Millisekunden ausgesandt. Zwischen den Bursts als Ganzen können hingegen sehr unterschiedliche Zeitabstände von weniger als einer Sekunde bis hin zu mehr als 15 Minuten liegen. Der Zeitabstand zwischen den Bursts hängt stark von diversen Faktoren ab:

- Betriebssystem(-Version)
- aktuelle Nutzungsart (Standby, Entsperren des Geräts durch Anschalten des Bildschirms, Starten von Apps, Internetnutzung)
- Verbindungsstatus (mit Access Point verbunden oder nicht)
- Ladezustand des Akkus
- Anzahl bekannter Netzwerke

Da diese Faktoren bei der Erfassung von Probe Requests unbekannt sind, gibt es keine Möglichkeit, die Zeitabstände zwischen Bursts a priori abzuschätzen.

Bluetooth

Bluetooth bezeichnet eine standardisierte Funktechnik zur kabellosen Übertragung von Daten zwischen unterschiedlichen Geräten. Bluetooth nutzt das gleiche Frequenzband wie WLAN, allerdings sind die einzelnen Kanäle hier schmaler, so dass es hier mehr davon (79) gibt.

Der Aufbau einer Verbindung zwischen Geräten gliedert sich in den *Inquiry-Prozess* und den *Paging-Prozess*. Im Zuge des Inquiry-Prozesses erstellt ein Gerät, das den Verbindungsaufbau initiiert (*Master*), eine Liste von Geräten in der Nähe (*Slaves*), mit denen der Master sich im Rahmen des nachfolgenden Paging Prozess verbinden kann. Der Paging-Prozess bereitet den geplanten Austausch von Nutzdaten unmittelbar vor. Hierbei wird der Slave, mit dem die Verbindung eingegangen werden soll, mit seiner MAC-Adresse angesprochen.

2.1.3 Filterung von Stör- und Nutzdaten

Ein in einem ÖPNV-Fahrzeug installierter Sensor erfasst alle WLAN- und Bluetooth-Signale von allen Endgeräten in der Reichweite des Empfängers. Somit werden unvermeidlich sowohl die Daten von Endgeräten, die sich im Fahrzeug befinden (Fahrgastdaten), als auch von Endgeräten, die sich nicht im Fahrzeug befinden („Störddaten“),

erfasst. Die Stördaten können durch Endgeräte von anderen Verkehrsteilnehmenden, aber auch von Geräten bzw. Personen in Gebäuden in der Nähe der Fahrtroute des Erhebungsfahrzeugs ausgesendet werden. Daraus ergibt sich die Notwendigkeit, diese Stördaten mittels geeigneter Filter zu selektieren.

In der Literatur wird hierzu auf Filtermethoden anhand von Spannweiten einzelner Indikatoren zurückgegriffen (Dunlap et al. 2016; Mishalani et al. 2016). Zu diesen Indikatoren gehören unter anderem:

- die Dauer der Erfassung der MAC-Adressen
- die empfangene Signalstärke
- die Anzahl an empfangenen WLAN- bzw. Bluetooth-Datenpunkten (Probe Requests, Inquiry-Response-Nachrichten)
- die Entfernung zu der letzten bzw. nächsten Haltestelle für die erste und letzte Erfassung eines Datenpunkts einer MAC-Adresse.

2.1.4 Lösungsansätze zur Arbeit mit randomisierten Daten

Die Randomisierung der MAC-Adresse erschwert die Identifizierung und Verfolgung eines Smartphones über mehrere Bursts hinweg massiv oder macht sie in manchen Fällen sogar unmöglich. In der Forschung gab es daher erste Versuche, die MAC-Randomisierung zu umgehen („Derandomisierung"). Dabei wurden folgende Ansätze zur Identifikation verwendet:

- die von einem Smartphone versandten Information Elements; sie ändern sich in vielen Fällen auch bei Randomisierung nicht und können zur Identifikation des Geräts genutzt werden.
- die Reihenfolge der Kanäle, auf denen die Probe Requests versendet werden
- der zeitliche Abstand zwischen den Bursts
- radiometrische Merkmale der emittierten Signale, das heißt physikalische Eigenschaften der elektromagnetischen Strahlung
- Datenaustausch von Apps im Mobilfunk

Bedingt durch die Vielzahl und Unkontrollierbarkeit der Einflussfaktoren auf diese Merkmale ist bisher jedoch kein Verfahren bekannt, das in der Lage wäre, randomisierte MAC-Adressen unter Realbedingungen einzelnen Endgeräten fehlerfrei zuzuordnen. Die Zuordnung ist somit nur mit einer zu bestimmenden Wahrscheinlichkeit richtig.

2.2 Ziele und Anforderungen

Ziel des Vorhabens war die Entwicklung eines Verfahrens, welches basierend auf den Input-Daten unterschiedlicher Quellen Informationen zu Quelle-Ziel-Verflechtungen

und zum Umsteigeverhalten der Fahrgäste (Output-Daten) automatisiert generiert. Darüber sollte das Fahrgastverhalten im ÖPNV möglichst umfassend abgebildet und dem NVV für verschiedene Anwendungen bereitgestellt werden.

Zu Beginn der Projektbearbeitung wurde deutlich, dass der praktische Zugang zu den voraussichtlich zu verarbeitenden Daten gegeben sein muss. Deshalb wurde zusätzlich zu den Anforderungen der aufgestellten Anwendungsfälle geprüft, welche der vorgesehenen Datenquellen prinzipiell für die Anwendung nutzbar sind.

2.2.1 Prüfung der Datennutzbarkeit

Die folgenden Daten wurden im Rahmen des Projektes als Input-Daten näher betrachtet:

- Daten zu Ein- und Ausstiegen je Fahrzeug, die durch das automatische Fahrgastzählsystem (AFZS) des NVV erzeugt werden
- Daten, die durch die Suche von WLAN-Netzen oder Bluetooth-Geräten ohne Zutun des Smartphone-Nutzers entstehen (Probe Requests [WLAN], Inquiry-Nachrichten [Bluetooth]) (passive Datenerzeugung)
- Daten zur Quelle, zum Ziel und zur zeitlichen Lage einer Verbindung, die durch die Anfrage an eine Fahrplanauskunft des NVV über mobile und stationäre Endgeräte entstehen
- Wetterdaten aus dem Angebot der mCloud

Die vorhandenen Input-Daten wurden hinsichtlich ihrer Verfügbarkeit und Nutzbarkeit für die Verfahrensentwicklung geprüft. Ziel war es, die wirtschaftlichen und rechtlichen Rahmenbedingungen für Ihre Bereitstellung bzw. Erfassung zu ermitteln.

Es konnte festgestellt werden, dass die Daten zu Ein- und Aussteigern, die durch das AFZS generiert werden, sowie die Anfragen der Kunden an die Fahrplanauskunft des NVV im Eigentum des NVV sind. Somit können diese Daten unmittelbar für das Verfahren genutzt werden und es entstehen keine weiteren Aufwendungen. Gleichzeitig stellte sich heraus, dass die vom Auskunftssystem des Dienstleisters HaCon Ingenieurgesellschaft mbH dargestellten Fahrtmöglichkeiten (als Antwort einer Anfrage an die Fahrplanauskunft) im Eigentum des Dienstleisters sind und somit nur mit weiteren umfangreichen finanziellen Aufwendungen des NVV im Projekt hätten genutzt werden können. Deshalb wurden diese Daten in diesem Forschungsprojekts nicht verwendet.

Die passiv erfassbaren WLAN- und Bluetooth-Daten werden an Haltestellen im Gebiet des NVV sowie in den Fahrzeugen der jeweiligen Verkehrsunternehmen generiert. Da die Unternehmen im Auftrag des NVV ihre Verkehrsleistungen erbringen, die Erfassungshardware im Eigentum des NVV ist und die Verkehrsunternehmen nicht an der Datenverarbeitung beteiligt sind, können auch diese Daten genutzt werden. Um diese

Daten zugänglich zu machen, bedurfte es der Entwicklung und Installation einer geeigneten Hardware in den Fahrzeugen und an den Haltestellen des NVV sowie der Entwicklung einer Software zur Datenverarbeitung.

Bei den Wetterdaten der mCloud handelt es sich um öffentlich zugängliche Daten, deren Nutzung ausdrücklich erwünscht ist. Somit konnten diese Daten ebenfalls im Projektverlauf verwendet werden.

Die Nutzung der Daten aus den Anfragen der Kunden an die Fahrplanauskunft und der passiv erfassbaren WLAN- und Bluetooth-Daten wurde anschließend datenschutzrechtlich geprüft (siehe Kapitel 2.3)

2.2.2 Anforderungen an die Output-Daten

Im Anschluss an die Prüfung der Nutzbarkeit der vorhandenen Input-Daten wurde untersucht, welche Output-Daten zur Beschreibung des Nachfrageverhaltens zukünftig bereitgestellt werden sollen. Folgende Anwendungsmöglichkeiten des Verfahrens wurden identifiziert:

- Angebotsplanung
- Einnahmenaufteilung
- Tarifgestaltung
- Verkehrsmanagement
- Prognose des Besetzungsgrads für die Verbindungsauskunft

Für diese Anwendungsgebiete wurde jeweils ein anwendungsspezifischer Anforderungskatalog an die Output-Daten des Verfahrens erstellt. Die Anforderungen wurden nach den folgenden Dimensionen differenziert:

- Untersuchungseinheit (Weg oder Etappe im ÖPNV)[1]
- Räumliche Auflösung
- Zeitliche Auflösung
- Verfolgungsdauer (Zeitraum, über den ein Gerät bzw. eine Person verfolgt werden kann, als Basis für die Ermittlung und Auswertung intrapersoneller Nutzungsmuster)
- zeitliche Verfügbarkeit (Verzögerung zwischen Erfassung der Input-Daten und Verfügbarkeit der Output-Daten des Verfahrens)

[1] Die Fahrt eines ÖV-Fahrgastes von der Quelle- zur Ziel-Haltestelle mit ggf. mehreren Umstiegen wird in diesem Text als Weg bezeichnet. Die Nutzung einer Linienfahrt von der Einstiegs- bis zur Ausstiegshaltestelle wird als Etappe bezeichnet.

Aus den Anforderungskatalogen ergaben sich die Ziele der Verfahrensentwicklung im Projekt und die Grundlage für eine Bewertung der Nutzbarkeit des Verfahrens hinsichtlich der verschiedenen Anwendungen.

Die Anwendungsgebiete wurden anschließend aus Sicht des Projektkonsortiums anhand ihres potenziellen Nutzens, der jeweiligen Nutzbarkeit des Verfahrens und der Umsetzbarkeit aus Sicht des Datenschutzes priorisiert. Den Anwendungen „Angebotsplanung“ und „Einnahmenaufteilung“ wurde dabei die höchste Priorität zugeordnet, der Anwendung „Tarifgestaltung“ mittlere Priorität und den Anwendungen „Verkehrsmanagement“ und „Prognose des Besetzungsgrads“ die niedrigste Priorität. Die Prognose des Besetzgrads gewann durch die Pandemielage im Laufe der Projektbearbeitung an praktische Relevanz und wurde nachträglich mit aufgenommen.

Die folgende Abbildung stellt die Anwendungsmöglichkeiten und deren Priorisierung zusammengefasst dar.

Abbildung 3: Ziele und Anwendungsfälle in der Projektbearbeitung

In den nachfolgenden Abschnitten werden die Anwendungen genauer beschrieben und die wichtigsten entsprechenden Anforderungen dargestellt.

2.2.2.1 Angebotsplanung

Ziel der Angebotsplanung ist es, ein nachfrageorientiertes Verkehrsangebot zu erstellen, das den Vorgaben und Zielen übergeordneter Planwerke (Nahverkehrsplan, strategische Planungen) entspricht. Die Angebotsplanung umfasst die Streckennetzplanung (Haltestellen und Fahrwege), die Linienplanung, die Planung der Bedienungshäufigkeit sowie die Kapazitätsplanung. Detaillierte und stets aktuelle Daten zu den Quelle-Ziel-Verflechtungen im Bedienungsgebiet, wie sie das Verfahren liefern kann, bilden eine verbesserte Datengrundlage für die Angebotsplanung.

Untersuchungseinheit für die Angebotsplanung ist idealerweise der Weg, da die Auswertung von Umstiegen insbesondere für die Planung des Streckennetzes und der Linien relevant ist. Für eine passgenaue Angebotsplanung ist im Idealfall eine haltestellenscharfe räumliche Auflösung erforderlich.

Hinsichtlich der zeitlichen Auflösung ist für die Streckennetz- und Linienplanung eine Differenzierung nach Schulzeit/Schulferien, Wochentagstyp und Stundengruppe ausreichend. Für die Planung der Bedienungshäufigkeit und Kapazitäten ist hingegen eine stündliche Auflösung notwendig, in der Hauptverkehrszeit sogar in 20-Minuten-Intervallen, da die Anforderungen hinsichtlich des maximalen Besetzungsgrades für solche Zeitintervalle formuliert sind (VDV-Schrift 4). Allerdings werden für diese Planungen keine Quelle-Ziel-Matrizen benötigt, sondern lediglich Daten zur Besetzung der einzelnen Linienfahrten.

2.2.2.2 Einnahmenaufteilung

Im Verbundraum des NVV agieren verschiedene Verkehrsunternehmen, die vor dem Hintergrund einfacher Tarifstrukturen für die Fahrgäste die von den Verbundpartnern ausgegebenen Fahrausweise wechselseitig akzeptieren. Die Zuordnung der Erlöse an die beteiligten Verkehrsunternehmen obliegt dem NVV und benötigt eine belastbare Datengrundlage. Ein wesentlicher Parameter für die Einnahmenaufteilung im NVV sind die erbrachten Personenkilometer und/oder Fahrzeugkilometer.

Die passende Untersuchungseinheit ist hier die Etappe, da die Beförderungsleistung auf den einzelnen Etappen eines Weges von unterschiedlichen Verkehrsunternehmen erbracht werden können. Für die Ermittlung der Verkehrsleistung als Aufteilungsschlüssel ist eine haltestellenscharfe räumliche Auflösung erforderlich. In anderen Verbünden kann fallweise eine geringere räumliche Auflösung ausreichen, falls ein entsprechendes Aufteilungsverfahren beschlossen wurde.

2.2.2.3 Tarifgestaltung

Im Rahmen der Tarifgestaltung im ÖPNV werden die Preise zwischen Start- und Zielhaltestelle der Fahrgäste festgelegt. Zu den grundsätzlichen Zielen und Anforderungen an die Tarifgestaltung gehören Leistungsgerechtigkeit (Orientierung des Fahrpreises an der in Anspruch genommenen Beförderungsleistung), Zumutbarkeit (Sicherstellung eines für die Fahrgäste zumutbaren Preisniveaus), Ergiebigkeit (Sicherstellung eines auskömmlichen Erlöses für die Verkehrsunternehmen bzw. Aufgabenträger), Verständlichkeit (aus Sicht der Fahrgäste) und Praktikabilität (aus Sicht der Verkehrsunternehmen bzw. Aufgabenträger). Darüber hinaus können mit der Tarifgestaltung spezifischere Ziele verfolgt werden, insbesondere eine zielgruppenspezifische, an bestimmte Nutzungsmuster angepasste Attraktivierung und eine räumliche und/oder zeitliche Lenkung der Fahrgäste.

Nutzungsmusterspezifische Tarife

Während für Selten- und Gelegenheitsnutzer der Bartarif (Einzelfahrkarten, Tageskarten) und für Vielnutzer das Zeitkartensegment (Wochen-, Monats-, Jahreskarten) bestehen, fehlen derzeit etablierte Angebote für Gelegenheitsnutzer des ÖPNV (5 bis 15 Fahrten pro Monat). Mithilfe der neuen Datengrundlage zum Nachfrageverhalten der Gelegenheitsnutzer können unter anderem auf diese Kundengruppe zugeschnittene Tarifangebote entwickelt werden.

Passende Untersuchungseinheit ist hier der Weg. Eine räumliche Auflösung auf Ebene der Tarifzonen ist ausreichend, es sei denn, es soll auch eine Neuaufteilung der Tarifzonen geprüft werden. Die zeitliche Auflösung sollte mindestens nach Wochentagstyp und Stundengruppe differenzieren, um die zeitliche Gültigkeit neuer Tarifangebote passgenau zuschneiden zu können. Die Erfassung typischer Nutzungsmuster über hinreichend lange Zeiträume (zum Beispiel einen Monat) sollte möglich sein.

Tarife zur Fahrgastlenkung

Um (punktuelle) Überlastungen der ÖV-Fahrzeuge in Spitzenstunden zu vermeiden oder abzumildern, können durch entsprechende Gestaltung des Tarifs preisliche Anreize gesetzt werden. Damit sollen Fahrgäste ihre Abfahrtszeit und Route von überlasteten auf weniger stark ausgelastete Linienfahrten verlagern, so dass die Verkehrsnachfrage räumlich und/oder zeitlich gleichmäßiger verteilt wird.

Die Fahrgastlenkung stellt hohe Ansprüche an die räumliche und zeitliche Auflösung: Erstere sollte (zumindest in städtischen Gebieten) haltestellenscharf sein, letztere in 20-Minuten-Intervallen oder fahrtenscharf.

2.2.2.4 Betriebsplanung

Im Falle von Betriebsstörungen oder temporären Überlastungen sollten die ÖPNV-Unternehmen in der Lage sein, zeitnah mit geeigneten Maßnahmen – etwa durch Einsatz von Verstärkerfahrten oder Ersatzverkehren – zu reagieren, um negative Auswirkungen auf die Fahrgäste und die Betriebsqualität zu minimieren. Bei Großveranstaltungen sollten möglichst bereits im Vorfeld zusätzliche Fahrten eingeplant werden (Eventverkehre).

Für die Überwachung bzw. Prognose der Belegung von Fahrzeugen ist die passende Untersuchungseinheit die Etappe. Für die Empfehlung von Alternativrouten ist jedoch der Zusammenhang der Etappen im Rahmen von Wegen von Bedeutung. Eine hohe räumliche und zeitliche Auflösung (haltestellenscharf, 5 Minuten) sowie eine möglichst geringe Verzögerung zwischen Datenerfassung und Verfügbarkeit der Output-Daten (nahe an Echtzeit) sind für eine zielgenaue und zügige Reaktionsmöglichkeit essenziell.

2.3 Datenschutz

Der Schwerpunkt der Bearbeitung lag auf der Begleitung der Forschungsarbeit aus Datenschutzsicht und der Sicherstellung der datenschutzkonformen Systemrealisierung. Hierfür wurden die datenschutzrechtlichen Anforderungen analysiert und in ein Datenschutzkonzept überführt.

2.3.1 Datenschutzkonzept

Das Datenschutzkonzept diente dabei als Grundlage und als Rahmen für die Systementwicklung. Entsprechend dem Prinzip „privacy by design“ wurde die rechtmäßige Datenverarbeitung und der erforderliche Schutz von Daten durch das eigentliche Systemdesign sichergestellt. Als eine der wesentlichen Schutzmaßnahmen wurde dabei die Anonymisierung der erhobenen personenbezogenen Daten vor deren Speicherung festgelegt und softwaremäßig implementiert.

Als weitere Maßnahmen wurden darüber hinaus spezifiziert und systemtechnisch umgesetzt:

- eine Opt-Out-Funktion als Widerspruchsmöglichkeit für Fahrgäste zur Nutzung ihrer Daten für die Routenerhebung
- die turnusmäßige Änderung des Anonymisierungshashwerts für die Begrenzung der zeitlichen Auflösung der Datengrundlage sowie
- die Umsetzung von Löschfristen für verschiedene Datenstämme entsprechend deren Verarbeitungszwecken.

Bei der Entwicklung des Datenschutzkonzepts wurden von Beginn an unterschieden zwischen der Phase des Forschungsvorhabens - mit zum Teil weniger stringenten Anforderungen - sowie einer späteren Weiternutzung des Systems in einem operationellen Betrieb beim Verkehrsverbund NVV.

Das Datenschutzkonzept stellte dabei sicher, dass die erforderlichen statistischen Outputergebnisse anonym bleiben und so uneingeschränkt von Systemanwendern, das heißt Verkehrsunternehmen oder Verbünden für deren Zwecke genutzt werden können.

Damit wurde eine Lösung erarbeitet, welche die Ergebnisverwertung außerhalb des Forschungsprojektes sicherstellt. Der konzeptionelle Ansatz ist in Abbildung 4 skizziert.

Abbildung 4: Übersicht der Datenverarbeitungsvorgänge und Schutzmaßnahmen

Im Projektverlauf wurden für das Datenschutzkonzept weitere Verarbeitungsvorgänge analysiert.

Speziell wurde das Hinzuziehen von Daten aus Anfragen an die Verbindungsauskunft, die potenzielle Fahrgäste in der NVV-App und auf der NVV-Webseite generieren, datenschutztechnisch bewertet und die Anforderungen an die Verarbeitung dieser zusätzlichen Datensätze aufgestellt. Die datenschutzrechtlichen Maßnahmen sehen auch in diesem Fall eine sofortige Datenanonymisierung und Löschung aller (nicht anonymisierten) Quelldatensätze vor. Die für die Auswertungen zur Fahrgastnachfrage erforderlichen Datenauswertungen werden ausschließlich auf Grundlage von anonymisierten Daten durchgeführt.

2.3.2 Öffentlichkeitsarbeit

Ergänzend zur Aufstellung des Datenschutzkonzeptes für die Entwicklung des technischen Systems wurden begleitende Maßnahmen in Bezug auf Informationen für die Fahrgäste und Kommunikation zum Forschungsvorhaben gemäß Anforderungen der DSGVO im Arbeitspaket geplant und durchgeführt.

Zur Information der Fahrgäste über das Projekt und die Datenerfassung wurden vom Praxispartner NVV alle vorhandenen Kommunikationskanäle gegenüber seinen Kun-

den genutzt. So können sich die Fahrgäste über eine eigens eingerichtete Landingpage des NVV über das Projekt informieren und ihr Widerspruchsrecht zur Datenerhebung in Anspruch nehmen: https://www.nvv.de/service-projekte/mobiledatafusion/

Weiterhin wurden Aushänge an allen in den Praxistest eingebundenen Haltestellen und Fahrzeugen angebracht (siehe Abbildung 5) und die Öffentlichkeit mittels Pressemitteilungen und Social-Media-Auftritt des NVV über das Projekt informiert.

Über eine speziell eingerichtete E-Mail-Adresse haben die Kunden zudem die Möglichkeit, bei Fragen Kontakt zum NVV aufzunehmen.

Abbildung 5: Information der Fahrgäste in NVV-Fahrzeugen

Ergänzend wurde die Datenschutzerklärung für die Webdienste des NVV (Web und App) in Bezug auf neue Datenverarbeitungsvorgänge angepasst und veröffentlicht. In diesem Zusammenhang wurde das fortgeschriebene Datenschutzkonzept durch

die juristische Abteilung sowie durch den externen Datenschutzbeauftragten des NVV geprüft und freigegeben.

2.3.3 Verifizierung des Konzepts

Vor Start des Praxistests (siehe Kapitel 2.8) wurde das Datenschutzkonzept in enger Abstimmung mit dem Partner INIT in Bezug auf den finalen Systemstand verifiziert.

Abschließend wurde das im Vorhaben erarbeitete Datenschutzkonzept gegen das Standard-Datenschutzmodell, veröffentlicht von der Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder (2020), geprüft. Dieses Datenschutzmodell ermöglicht eine Selbst-Überprüfung der im Vorhaben ausgearbeiteten datenschutztechnischen Anforderungen anhand einer unabhängigen standardisierten Methode.

Als praktische Hilfe für die Verifizierung wurde das Tool Safe-Data-Governance Framework (Teil Datenschutz) von der iRights.Lab GmbH (2020) verwendet. Das Self-Data-Governance Framework wurde im Rahmen der mFund-Forschungsinitiative aufbauend auf dem Standard-Datenschutzmodell entwickelt und steht Organisationen und Projekten für die Durchführung von Selbstevaluationsprozessen zur Verfügung.

Die darüber erfolgte Evaluation der datenschutzrechtlichen Vorgaben im Vorhaben im Dezember 2021 hat bestätigt, dass die durch das Standard-Datenschutzmodell postulierten Gewährleistungsziele erfüllt sind und das Schutzniveau den rechtlichen Anforderungen entspricht.

Im Ergebnis ist es gelungen, unter Beachtung aller datenschutzrechtlichen Anforderungen ein Konzept für eine Erfassung der Fahrgastnachfrage und Routenwahl zu spezifizieren und das entsprechend entwickelte System sicher auszulegen und damit die rechtmäßige Datenverarbeitung für die angestrebten Projektzwecke zu ermöglichen. Die datenschutzrelevanten Anforderungen sind in die Systemspezifikation (Systemarchitektur) eingeflossen und in einem Datenverarbeitungsverzeichnis dokumentiert.

2.4 Erfassung der Datengrundlage

Für die Erfassung der Datengrundlage war die Entwicklung der Systemarchitektur sowie die Entwicklung der Hardware für die Erfassung der WLAN- und Bluetooth-Signale an den Haltestellen und im Fahrzeug notwendig. Es wurden Testszenarien durchgeführt, um Aufschluss über die Datenqualität zu gewinnen. Zudem wurden Fahrgastbefragungen durchgeführt, um Referenzdaten für die Datenanalyse zu sammeln.

2.4.1 Entwicklung eines skalierbaren Dateninfrastruktur

Ein Hauptziel von Mobile Data ist die Verknüpfung verschiedener Datenbestände (siehe Abbildung 6).

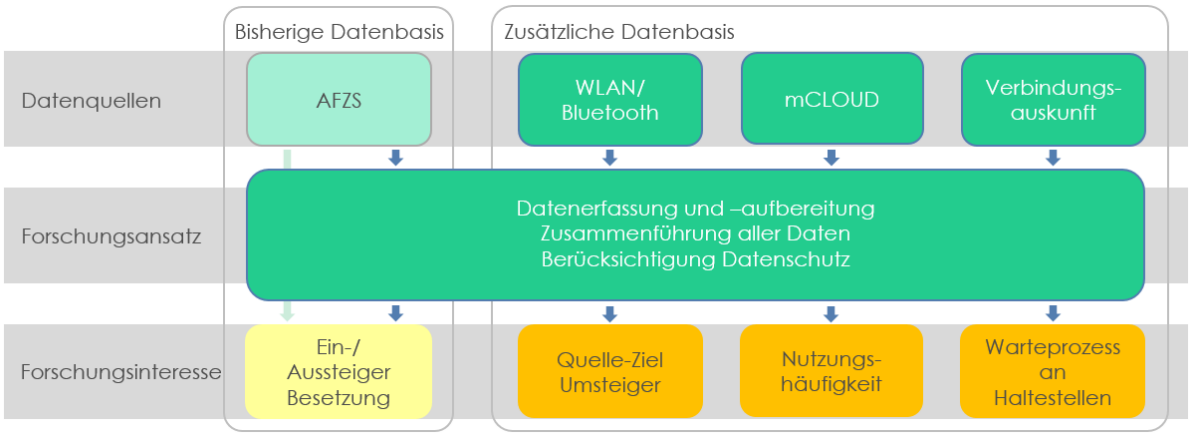


Abbildung 6: Fusion verschiedener Datenquellen in Mobile Data Fusion

Auf Basis der bisher üblichen Vorgehensweise mittels AFZS Daten zur Ermittlung der Ein- und Aussteiger sowie der Besetzgrade im Fahrzeug werden in Mobile Data Fusion zusätzliche Datenquellen wie WLAN- und Bluetooth-Signale, Daten aus der mCLOUD, offene Wetterdaten (Temperatur und Niederschlag) sowie Abfragen aus der Verbindungsauskunft fusioniert. Die Daten werden nach der Erfassung aufbereitet und zusammengeführt. Dieser Prozess erfolgt im Einklang des Datenschutzes, wie in Kapitel 2.3 beschrieben. Das Ziel ist es, neben der Erfassung der Ein-/Aussteiger und der Besetzgrade, Quelle-Ziel-Beziehungen, die Nutzungshäufigkeit sowie Informationen zum Warteprozess an den Haltestellen zu ermitteln.

Diese Vorgehensweise verlangt eine skalierbare Dateninfrastruktur, die es ermöglicht alle Fragestellungen des Projektes zu berücksichtigen. Um die Fusion der verschiedenen Datenquellen zu gewährleisten, wird daher im Projekt ein Big Data Ansatz verfolgt. Er ermöglicht es, skalierbar große Datenmengen zu erheben, zu speichern und zu analysieren. Bei diesem Ansatz werden alle gesammelten Daten unter Berücksichtigung des Datenschutzkonzeptes in Betracht gezogen. In einem zweiten Schritt werden relevante Daten zur weiteren Bearbeitung und Analyse nach dem BigData Paradigma selektiert, ohne dabei die gesamte Datenmenge zu dezimieren. Für den Begriff Big Data liegen mehrere Definitionen vor. Mobile Data Fusion fokussiert sich auf den 5V Ansatz von Gartner (Demchenko et al., 2014), welcher Big Data nach Volume, Variety, Velocity, Veracity und Volume kategorisiert (Abbildung 7).

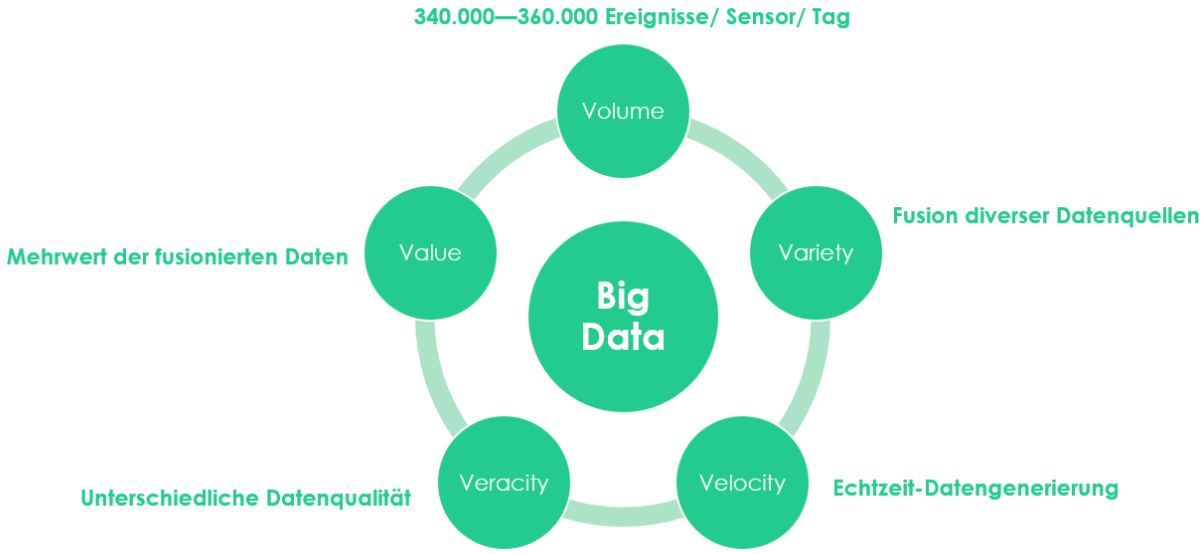


Abbildung 7: Big Data im Projekt Mobile Data Fusion

In Mobile Data Fusion beschreiben diese 5V von Big Data folgende Eigenschaften:

Value: Der Mehrwert der fusionierten Daten besteht in der neuartigen zusammenstellen der Daten, die es ermöglicht Quelle-Zielmatrizen sowie Umsteigeverbindungen zu generieren. Diese Daten bieten einen Mehrwert zur Optimierung der ÖPNV-Planung. Sie können für die Angebotsplanung, Einnahmenaufteilung, Tarifgestaltung sowie Betriebsplanung genutzt werden. In der Zukunft können auch Besetzgrade bereitstellt werden.

Veracity: Das zweite V steht im Projekt für Datenqualität. Die Qualität der Daten wird im Projekt durch verschiede Filterkriterien und Validierungsmethoden gewährleistet (siehe Kapitel 2.5). Bei diesem Prozess werden Daten verschiedener Datenqualitäten fusioniert. Die Datenqualität bezieht sich dabei hauptsächlich auf die zeitliche sowie räumliche Erfassungsrate der Daten. Ein Beispiel hierfür ist die unterschiedliche Erfassungsrate der mobilen Endgeräte verschiedener Hersteller aber auch Datenquellen wie Wetterdaten, die durch unterschiedliche zeitliche und räumliche Auflösungen kategorisiert sind.

Velocity: In Mobile Data Fusion werden die Bluetooth- und WiFi-Signale in nah-Echtzeit erfasst. Die Datenauswertung sowie die Generierung der Quelle-Zielmatrizen erfolgt retrospektiv und wird einmal pro Tag bereitgestellt.

Variety: Neben den WLAN- und Bluetooth-Signalen werden verschiedene weitere Datenquellen in Betracht gezogen. Bei diesen Daten handelt es sich um AFZF-Daten, verschiedene Daten aus der mCloud, Temperatur und Niederschlagsdaten vom Deutschen Wetterdienst (DWD), Anfragen an die Verbindungsauskunft und des Buchungssystems sowie WLAN- und Bluetooth-Signalen an den Haltestellen.

Volume: Im Projekt werden pro Sensor ca. 50 MB pro Tag generiert, diese beinhalten nur die Bluetooth- und WiFi-Signale. Bei einer Ausstattung des gesamten NVV-Netzes entspricht dies ca. 280 GB pro Tag und ca. 9TB an Datenerfassungssignalen pro Monat.

In diesem Projekt dient Big Data nicht als Beschreibung von großen Datenmengen, sondern als komplexe Analyse, Aufbereitung und Verwertung von Datenmengen aus unterschiedlichen Quellen. Dabei geht es um das Erkennen von statistischen Korrelationen, Mustern und Zusammenhängen hinsichtlich der Routenwahl von Nutzern innerhalb des ÖPNV-Netzes. Der Hauptbeitrag liegt dabei auf „variety of data", einem der 5 V nach Gartner (Demchenko et al., 2014).

Das von der INIT entwickelte Back-End System basiert auf einer Apache Kafka Dateninfrastruktur und dient als Datenextraktion, Datentransformation und als Datenspeicherungs-Pipeline. Als quelloffene Software zum Übertragen und Speichern großer Datenströme agiert Apache Kafka als Datenbroker zwischen den Datenproduzenten (Eingangsdaten) und den Datenkonsumenten (IT-Partnersysteme). Apache Flink wird dabei zur Transformation von Datenströmen sowie zum Verknüpfen der Datenquellen eingesetzt. Apache Beam dient als einheitliche Programmierschnittstelle (API), um die Algorithmen in verschiedenen Prozessen einsetzen zu können. Ein großer Vorteil dieser eingesetzten Technologien ist die breite Skalierbarkeit sowie die Möglichkeit der Echtzeitbearbeitung, mit der wir in der Lage sind, einen kontinuierlichen Strom an Ereignissen in einem Verkehrsnetz zu analysieren und in Echtzeit Maßnahmen der Verkehrssteuerung und –lenkung zu ergreifen.

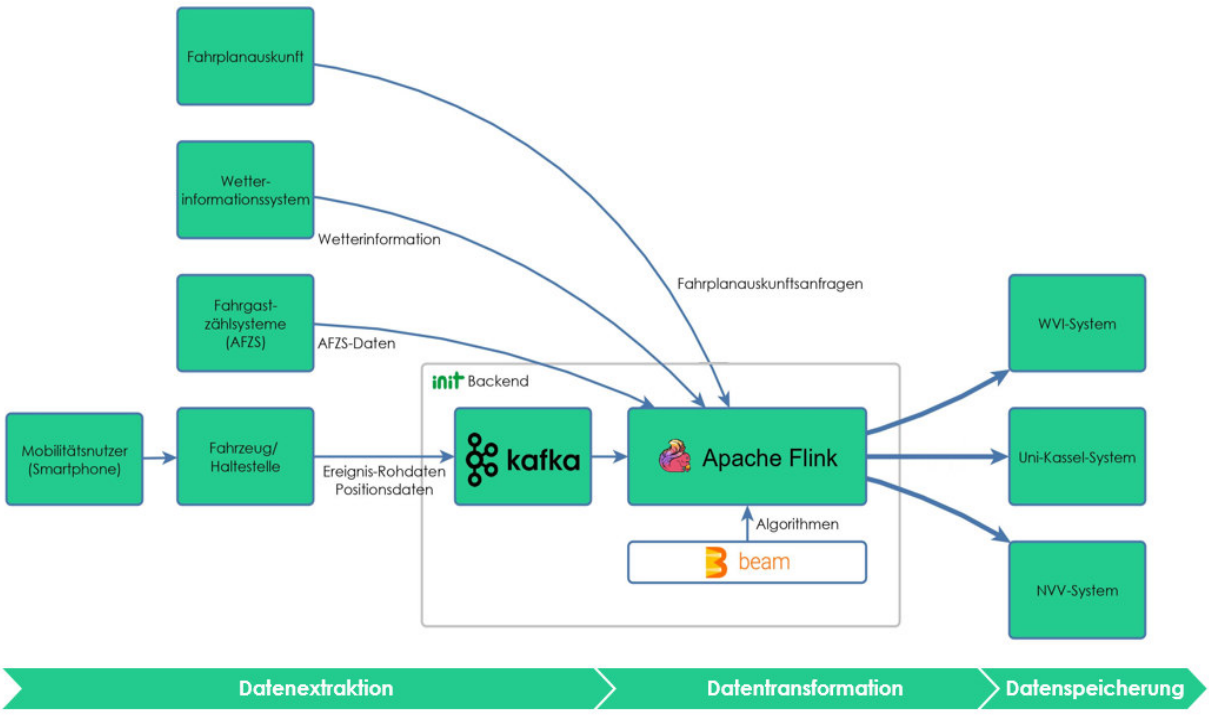


Abbildung 8: Systemarchitektur der skalierbaren Data Pipeline

Neben den Mobilitätsdaten (WLAN und Bluetooth) wurden die Wetterdaten als fester Bestandteil in die AFZS-Zähldaten integriert. Zusätzlich zu den in der mCLOUD referenzierten Daten des DWD wurden auch eine erst im Jahr 2020 in Kassel neu eingerichtete Wetterstation sowie frei verfügbare Temperaturdaten der offenen Plattform OpenWeather API[2] berücksichtigt.

Die Fahrplanauskunftsanfragen aus der NVV-App und den Verkaufsgeräten der Firma HaCon wurden aufgearbeitet und planmäßig im System abgelegt. Eine größere Herausforderung stellte dabei die zuverlässige Zusammenführung der Haltestellennummern (Stop-IDs) der verschiedenen Datensätze da. Diese Daten wurden aus diesem Grund nicht in der Analyse sowie in der Verfahrensentwicklung berücksichtigt.

2.4.2 Hardware zur Erfassung von WLAN/Bluetooth-Signalen

Um die WLAN- und Bluetooth-Erfassung umzusetzen, wurden im Rahmen der Projektarbeit verschiedene Hardwarelösungen entwickelt. Die Hardwareentwicklung teilte sich dabei in zwei Phasen auf.

In der ersten Projektphase wurde ein Prototyp auf Basis eines RaspberryPis auf 8 Testlinien in über 50 Fahrzeug und 40 Haltestellen im NVV Netz eingesetzt und erprobt (Abbildung 9). Diese Hardware-Lösung war sowohl an den Haltestellen als auch im Fahrzeug im Einsatz. Da das Projekt einen „Privacy by Design"-Ansatz verfolgt, wurde bereits in Phase 1 bei der Implementierung der sensorseitig eingesetzten Software ein Hashing- und Salting-Algorithmus berücksichtigt. Weitere Anforderungen aus dem Datenschutzkonzept, wie die Einrichtung eines MAC-Blacklist-Verfahrens für Fahrgäste, die nicht erfasst werden möchten, wurden in Zusammenarbeit mit einem externen Auftragsdatenverarbeiter und entsprechenden Anpassungen in technischer und datenschutzrechtlicher Hinsicht erfolgreich umgesetzt.

Abbildung 9: Hardware-Lösung zur Bluetooth- und WiFi-Erfassung basierend auf RaspberryPis in Projektphase 1.

[2] https://openweathermap.org/

Zur Validierung der Ergebnisse wurden Fahrgastbefragungen mit WVI-eigenen ELFE-Smartphones auf 200 Fahrten durch die Uni Kassel durchgeführt sowie Testszenarien zum Sendeverhalten der verbauen Geräte durchgeführt (siehe Kapitel 2.4.3 und 2.4.4).

Nach einer erfolgreichen ersten Projektphase im Jahr 2020 wurde in der zweiten Entwicklungsphase eine Erweiterung des INIT Bordrechners COPILOTpc mit zusätzlichen WLAN und Bluetooth-Modulen, die zuvor auf dem RaspberryPi getestet wurden, entwickelt. Ein eigens auf Mobile Data Fusion angepasstes Linux System wurde auf den Bordrechnern installiert, um vollen Zugriff auf die Hardwaremodule und deren Optimierung zur WLAN- und Bluetooth-Erfassung zu ermöglichen. Nach erfolgreicher E1 Zulassung der erweiterten COPILOTpc Bordrechner durch das Kraftfahrt-Bundesamtes im August 2022 ist dieses System im Rahmen des Projektes im Produktivbetrieb in 21 Fahrzeugen beim NVV im Einsatz. Der Pilottest wird bis 2024 fortgeführt. Nach jetzigem Stand ist diese Erweiterung des Bordrechners COPILOTpc zur Erfassung von WLAN- und Bluetooth-Signalen die erste Lösung auf dem Markt, die über solch eine Zulassung verfügt. Neben dem Einsatz auf den COPILOTpc Bordrechnern können auch weitere Systeme, wie z.B. die Flee 4 von CarMediaLab oder den EVENDpc der INIT unter Berücksichtigung der E1-Zulassung erweitert werden (Abbildung 10).

Abbildung 10: Hardwareentwicklung basierend auf dem INIT Bordrechner COPILOTpc in Phase 2

Die aufgezeichneten WLAN- und Bluetooth-Signale werden mit den hinterlegten Fahrten aus dem Statistik Tool MOBILEstatistics der INIT im Backend verknüpft und können so den Fahrten und Umläufen eines Fahrzeuges zugeordnet werden. Die aufgezeichneten Erfassungsereignisse zeigen die Anzahl der erfassten Geräte und spiegeln nicht die Fahrgäste, sondern alle erfassten Geräte wider. Die Fahrgastzahlen werden in einem späteren Schritt vom AFZS bereitgestellt. Mit diesen Daten können dann aggregierte Daten zu Ein- und Aussteigern bereitstellt werden und an die Datenaufbereitung, Datenanalyse und an die Verfahrensentwicklung (siehe Kapitel 2.6 bis 2.7) übergeben werden. Lücken in den Erfassungsrohdaten konnten in vielen Fällen durch Anpassung von Filtern, Fehlerkorrekturen und Verbesserung der Datenzuordnung zu den Haltevorgängen geschlossen werden. Insbesondere für letztere Fehlerquelle entwi-

ckelte INIT eigens den Algorithmus QADABRA (Quantum-mechanics inspired Approach Driven Automatic Block Recognition Algorithm), mit den Erfassungsereignisse bei Lücken in den ÖPNV-Kontextdaten nachträglich mit einer hohen Wahrscheinlichkeit den korrekten Fahrplanfahrten zugeordnet werden können. Dieser Algorithmus wurde direkt in die Big-Data-Pipeline integriert und ermöglicht dauerhaft eine bessere Datenqualität.

Die kombinierte Erfassung der WLAN und Bluetooth -Signale erlaubt es, eine komplette Abdeckung der Signale zu erzielen und gleichzeitig das unterschiedliche Sendeverhalten sowie die unterschiedlichen Erfassungsraten auszugleichen. Abbildung 11 zeigt beispielhaft die Line 54. Es ist zu erkennen, dass ein kombiniertes Verfahren eine deutliche Verbesserung der Datenerfassungsrate bietet und so die Erfassungseigenschaften von WLAN- und Bluetooth-Signalen vereint.

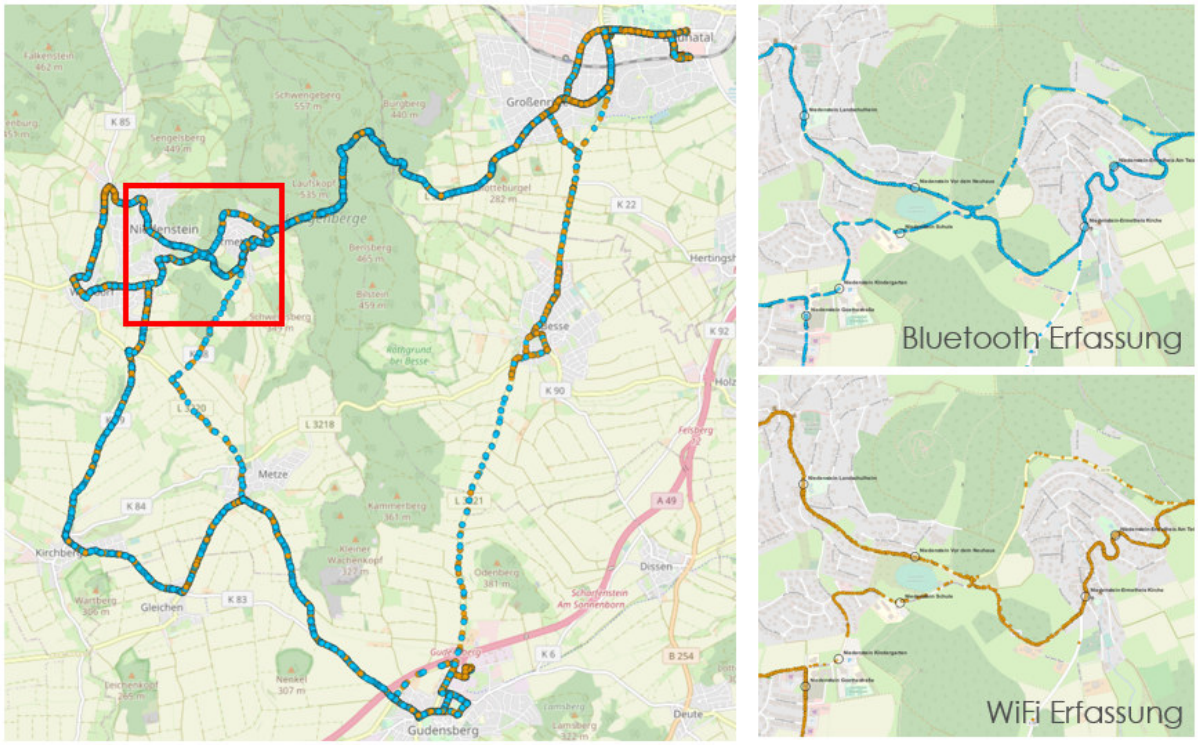


Abbildung 11: Heatmap der erfassten Bluetooth- und WiFi-Signale
(Ein Punkt spiegelt ein Erfassungsereignis wieder.)

2.4.3 Testszenarien

Ziel der Testszenarien war die Ermittlung und Prüfung der Erfassungsgenauigkeit der mittels WLAN- und Bluetooth-Sniffing erfassten Daten. Die konzipierten Testszenarien wurden auf die neuen Rahmenbedingungen angepasst, die sich durch die Auswirkungen der Corona-Pandemie auf den ÖPNV und technische Hürden bei der Datenerfassung und -übertragung ergaben. Folgende Testszenarien wurden durchgeführt:

- Tests unter Laborbedingungen zum **Sendeverhalten** von Smartphones (Bieland 2022):
  Zur Untersuchung des Sendeverhaltens wurden die Endgeräte entsprechend den Ergebnissen der Literaturrecherche hinsichtlich der Einflussfaktoren auf das Sendeverhalten getestet (siehe Kapitel 2.1.2): in verschiedenen Zuständen nach Nutzung (aktiv, inaktiv) und Verbindungsstatus (aktive Verbindung zu einem Access Point, Access Point in der Nähe (ohne Verbindung), kein Access Point in der Nähe) für die Dauer von jeweils 15 Minuten (je Smartphone und je Zustand). Stand der Untersuchung ist April 2020. Die Erfassung der Signale wurde manuell in einem Raum durchgeführt, der für Funkwellen nicht durchlässig ist. Es wurden Daten der folgenden vier Testhandys erfasst (in Klammern jeweils das verwendete Betriebssystem):
  - Samsung Galaxy S 7 (Android 7)
  - Apple iPhone 6S (iOS 12.0.1)
  - CAT S 61 (Android 9)
  - CAT S 61 (Android 10)
- Tests zur Evaluierung der Eignung der **RSSI**-Filtermethode (siehe Kapitel 2.1.3) zur Differenzierung zwischen Signalen von Geräten innerhalb vs. außerhalb eines Fahrzeugs (Nutzdaten vs. Störddaten): Auf dem Betriebshof der Bad Wildunger Kraftwagenverkehrs- und Wasserversorgungsgesellschaft wurde der RSSI für verschiedene Smartphones in verschiedener Entfernung und bei unterschiedlichem Zustand der Türen (offen/geschlossen) gemessen (Bieland 2022).
- Tests im ÖPNV-Realbetrieb zur **Nachverfolgbarkeit** von Endgeräten: Zum Nachweis des grundlegenden Forschungsziels wurden Tests durchgeführt, bei denen die Erfassbarkeit von Fahrgästen, des Ein- und Ausstiegsort sowie Umstiegen aufgezeigt werden. Dabei wurden die durch Haltestellen erfassten Daten gesondert betrachtet, um den Nutzen einer Haltestellenausstattung abwägen zu können. An vier Tagen im Mai und Juni 2020 führte eine studentische Hilfskraft Testfahrten auf den Testlinien mit mehreren Umstiegen mit 6 mitgeführten Smartphones in wechselnden Verbindungszuständen und Nutzungsarten durch. Ziel dieser Fahrten war die Prüfung der folgenden Fragen:
  - Können die Signale der Test-Smartphones im Fahrzeug und an der Haltestelle inmitten der sonstigen aufgezeichneten Signale (von Geräten anderer Fahrgäste sowie von Nicht-Fahrgästen) identifiziert und verfolgt werden?
  - Erhöht die zusätzliche Nutzung der Daten, die von den an den Haltestellen angebrachten Sensoren empfangen werden, die Genauigkeit der Erfassung von Ein- und Ausstiegshaltestelle?

Die Ergebnisse werden im Kapitel 2.6.1 beschrieben.

2.4.4 Erste Fahrgastbefragung

Die Referenzdaten für die Verfahrensentwicklung und -kalibrierung wurden im Rahmen einer Fahrgastbefragung ermittelt. Die Fahrgastbefragung zielte darauf ab, für ausgewählte Linienfahrten der ausgestatteten Linien der ersten Phase die realen Quelle-Ziel-Beziehungen aller Fahrgäste der Linienfahrt zu erfassen. Die Befragung musste wegen des Ausbruchs der Corona-Pandemie um ca. ein halbes Jahr verschoben werden und wurde im Zeitraum 17.9. bis 13.11.2020 durchgeführt.

Für die erste Phase wurden drei Buslinien des NVV mit unterschiedlichen Charakteristika ausgewählt: Die Linie 500 (Kassel <-> Bad Wildungen über Gudensberg und Fritzlar) gilt als eine der stärksten Regionallinien im NVV-Gebiet mit etwa 4.000 Fahrgästen pro Tag[3], wohingegen die Linie 54 (Baunatal <-> Gudensberg über Niedenstein) eine ländliche Linie mit etwa 300 Fahrgästen pro Tag[1] repräsentiert. Zwischen diesen beiden Linien besteht in Gudensberg eine Umsteigebeziehung. Mit der Linie 100 (Kassel <-> Calden) wurde zusätzlich eine typische Stadt-Umland-Linie mit etwa 1.000 Fahrgästen pro Tag[1] ausgewählt.

Bei der Auswahl der Linienfahrten für die Fahrgastbefragung wurden die Merkmale Linie, Richtung, Wochentagstyp und Zeitschicht (Stundengruppe) berücksichtigt. Für jede Kombination dieser Merkmale sollten zwei Linienfahrten erhoben werden. Daraus ergab sich eine Anzahl von 100 geplanten Erhebungsfahrten. Aufgrund diverser technischer und pandemiebedingter Ausfälle wurden schließlich 79 Erhebungsfahrten durchgeführt. Dabei wurden nach Möglichkeit stets alle Fahrgäste der Linienfahrten befragt (Vollerhebung).

Die Befragungen wurden durch geschultes Personal als CAPI (Computer Assisted Personal Interview) unter Verwendung eines standardisierten Fragebogens mit Smartphones vorgenommen. Die Inhalte der Befragung wurden auf die Ziele der Befragung (Referenzmatrix, Vollerhebung) abgestimmt. Die Befragung bestand aus einer Erhebung der

- Einstiegshaltestelle des Fahrgasts auf der Linienfahrt,
- Ausstiegshaltestelle des Fahrgasts auf der Linienfahrt sowie
- einer Filterfrage zu weiteren Linienfahrten des Fahrgastweges (Umstiege), sofern weitere in Phase 1 ausgestattete Linien genutzt wurden.

Die Befragungsdaten wurden für die Verfahrensentwicklung aufbereitet (siehe Kapitel 2.5.2) und dienten zur Kalibrierung des entwickelten Verfahrens zur Datenfusion (siehe Kapitel 2.7).

[3] Fahrgastzahlen vor Corona-Pandemie

2.5 Datenaufbereitung

Ziel der Datenaufbereitung war es, die technisch ermittelten Daten für die Datenanalyse und die Verfahrensentwicklung zugängig zu machen. Die Vorgehensweise wird im Folgenden für die technisch bereitgestellten AFZ-, WLAN- und Bluetooth_Daten sowie die manuell erfassten Befragungsdaten im Einzelnen ausgeführt.

2.5.1 AFZ-, WLAN- und Bluetooth-Daten

Sowohl die AFZ- als auch die WLAN- und Bluetooth-Daten wurden durch die INIT regelmäßig bereitgestellt. Hierfür wurde ein Server der Universität Kassel verwendet. Hierdurch konnte gewährleistet werden, dass alle Partner gleichermaßen Zugriff auf die Datenbestände erhielten. Für die Datenübertragung erhielten die INIT und die WVI VPN-Zugänge. Mit Beginn der Erfassung von WLAN- und Bluetooth-Daten Mitte 2020 wurden die aufgezeichneten Daten sowie die mittels AZFS aufgezeichneten Zähldaten wöchentlich auf dem Uni Server bereitgestellt.

Für den umfangreichen Datenimport entwickelte WVI ein eigenes Softwaretool, welches die Datensätze einer Linienfahrt direkt beim Import in die Datenbank miteinander verknüpft. Um das Datenvolumen der WLAN- und Bluetooth-Daten zu reduzieren, werden bei diesem Verarbeitungsschritt die Datensätze bereits um Störddaten reduziert und aggregiert (siehe Kapitel 2.1.2). An dieser Stelle werden zunächst in der Vorselektion alle Datensätze entfernt, die keiner Fahrt zugeordnet werden können. Zusätzlich werden alle randomisierten Datensätze (siehe hierzu Kapitel 2.1.4) aussortiert und von der weiteren Verarbeitung ausgeschlossen, da diese ohne weitere Bearbeitung für das Verfahren nicht nutzbar sind. Im nächsten Schritt werden alle Fahrten ohne zugehörige AFZ-Daten aussortiert und gelöscht. Außerdem werden die Bursts auf einem Fahrtabschnitt zusammengefasst. Durch diese Selektions- und Aggregationsschritte kann das Datenvolumen bereits um ca. 90% reduziert werden. Die verbliebenen Daten finden Eingang in den weiteren Arbeitsprozess. Das Vorgehen ist in Abbildung 12 skizziert.

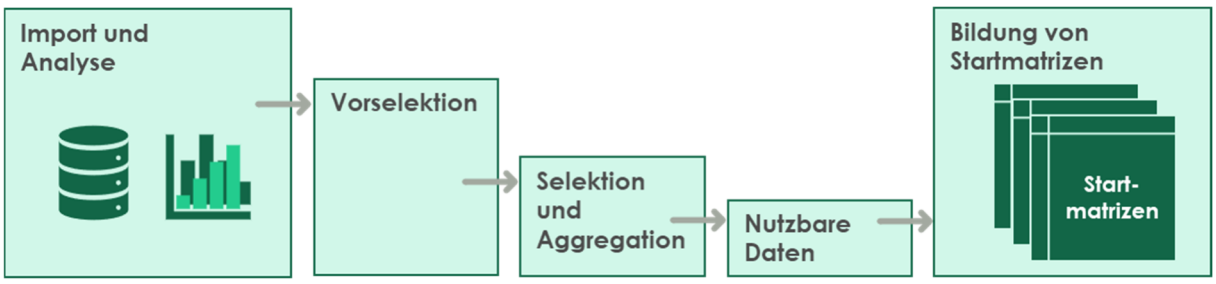


Abbildung 12: Selektion und Aggregation beim Datenimport

2.5.2 Befragungsdaten

Die Aufbereitung der Daten aus der Fahrgastbefragung aus Phase 1 erforderte zunächst manuelle Korrekturen und Ergänzungen der Daten. Ursachen dafür waren manuell erfasste Beobachtungen der Interviewenden und manuelle Korrekturen unterschiedlicher Schreibweisen der Haltestellennamen zur eindeutigen Zuordnung. Die Transformation der Befragungsdatensätze in Quelle-Ziel-Matrizen erfolgte dann mit Hilfe eines am VPVS eigens für diesen Zweck entwickelten R-Skripts. Anschließend wurden die Quelle-Ziel-Matrizen zur Vervollständigung der bei der Befragung nicht erfassten Fahrgäste auf Basis der AFZS-Daten ergänzt. Dieser Schritt umfasste je nach Umfang der Lücken bei der Befragung entweder manuelle Korrekturen oder einen Hochrechnungsschritt.

Es wurden nun mehrdimensionale Kriterien dafür entwickelt, welche der durchgeführten Erhebungsfahrten eine ausreichende Datenqualität für die weitere Verwendung im Rahmen der Entwicklung und Kalibrierung des Verfahrens aufwiesen. Falls eine der folgenden Bedingungen für eine Erhebungsfahrt zutraf, wurde sie als ungeeignet angesehen und von der weiteren Verwendung ausgeschlossen:

- Fehlende WLAN/Bluetooth-Daten (Ausfall des Sensors im Fahrzeug)
- Fehlende AFZS-Daten (Ausfall des AFZS)
- AFZS-Qualitätsindikator („APC Quality“)[4] kleiner als 80 %
- Absolute Abweichung der Anzahl der bei der Befragung erfassten Fahrgäste von der Anzahl der vom AFZS erfassten Fahrgäste größer als 5
- Relative Abweichung der Anzahl der bei der Befragung erfassten Fahrgäste von der Anzahl der vom AFZS erfassten Fahrgäste größer als 10 %
- Triviale Lösbarkeit: Bei sehr geringen Fahrgastzahlen existiert in vielen Fällen nur eine Matrix, die die vorgegebenen Randsummen einhält, so dass das Verfahren mit Nutzung der WLAN-/Bluetooth-Daten gar keinen zusätzlichen Nutzen bringen kann.

Auf Basis dieser Kriterien konnten von den 79 durchgeführten Erhebungsfahrten 29 Fahrten uneingeschränkt als Referenzdaten für die Entwicklung und Kalibrierung des Verfahrens genutzt werden. Weitere 18 Fahrten sind eingeschränkt nutzbar.

2.6 Datenanalyse

Im Rahmen der Datenanalyse wurden die Ergebnisse der Testszenarien sowie die Daten aus dem Pilotbetrieb ausgewertet.

[4] Der AFZS-Qualitätsindikator basiert auf der relativen Abweichung zwischen der vom AFZS gezählten Gesamtanzahl aller Einsteigenden und aller Aussteigenden.

2.6.1 Ergebnisse der Testszenarien

2.6.1.1 Labortests zu Sendeeigenschaften

Aus den Labortests lassen sich folgende qualitative Aussagen ableiten:

- Das Sendeverhalten (mittlerer zeitlicher Abstand zwischen den Bursts) hängt sehr stark von der Art der Nutzung (standby, aktiv) und vom Verbindungsstatus ab (mit Access Point verbunden bzw. nicht verbunden). Im Standby-Zustand und im verbundenen Zustand werden sehr viel weniger Probe Requests ausgesandt.
- Allerdings gilt: Selbst wenn diese Bedingungen konstant gehalten werden, senden die Smartphones in unregelmäßigen Abständen.
- Es zeigen sich deutliche Unterschiede in Abhängigkeit vom Gerätetyp.

Tabelle 3 gibt einen Überblick über die zahlenmäßigen Ergebnisse.

Zustand	**Aktiv**	x	x	x			
	Inaktiv / Standby				x	x	x
	Verbindung zu AP	x			x		
	AP in der Nähe		x			x	
	kein AP in der Nähe			x			x
iPhone 6S	Anzahl Bursts	1	17	68	4	1	0
	Minimaler Abstand	zu wenig Daten	00:00:06	00:00:02	00:02:45	zu wenig Daten	zu wenig Daten
	Maximaler Abstand		00:04:05	00:01:50	00:04:00		
	Mittlerer Abstand		00:00:55	00:00:13	00:03:15		
Samsung S7	Anzahl Bursts	21	11	11	1	0	0
	Minimaler Abstand	00:00:05	00:00:03	00:00:04	zu wenig Daten	zu wenig Daten	zu wenig Daten
	Maximaler Abstand	00:04:03	00:02:50	00:03:51			
	Mittlerer Abstand	00:02:25	00:01:18	00:01:27			
CAT S61 (A9)	Anzahl Bursts	8	55	26	0	32	34
	Minimaler Abstand	00:00:20	00:00:08	00:00:10	zu wenig Daten	00:00:18	00:00:10
	Maximaler Abstand	00:03:01	00:02:01	00:03:01		00:03:00	00:02:00
	Mittlerer Abstand	00:01:08	00:00:59	00:01:14		00:01:01	00:00:57
CAT S61 (A10)	Anzahl Bursts	10	15	8	1	20	19
	Minimaler Abstand	00:00:13	00:00:03	00:01:58	zu wenig Daten	00:00:05	00:00:07
	Maximaler Abstand	00:03:32	00:03:17	00:02:07		00:01:43	00:01:06
	Mittlerer Abstand	00:01:33	00:01:04	00:02:01		00:00:45	00:00:51

Tabelle 3: Ergebnisübersicht zum Sendeverhalten der Test-Smartphones.

2.6.1.2 Ergebnisse der Feldtests zum RSSI

In einer weiteren Untersuchung wurde die Wirkung der Fahrzeugkarosserie auf den RSSI gemessen (Bieland & Briegel 2021). Die Betrachtung der Signalstärke einzelner Endgeräte zeigt, dass keine eindeutige Unterscheidung möglich ist. Die Spannweiten der RSSI-Werte von Signalen einzelner Smartphones, die sowohl innerhalb (bei geschlossenen Türen) als auch außerhalb des Fahrzeugs in verschiedenen Abstandsklassen getestet wurden, überschneiden sich zu großen Teilen. Es liegen keine signifikanten Unterschiede (F-Test) zwischen den Mittelwerten der Stichproben (innerhalb / außerhalb des Fahrzeugs) vor.

2.6.1.3 Ergebnisse der Tests im ÖPNV-Betrieb

Die Auswertung der Ergebnisse wurde aufgrund von Sensorausfällen erschwert, sodass nur ein kleiner Teil der Ein- und Ausstiegsvorgänge betrachtet werden konnte. Dennoch konnte im Ergebnis festgestellt werden, dass

- die Verfolgung eines Geräts von Haltestelle zu Fahrzeug bzw. umgekehrt prinzipiell funktioniert,
- aufgrund der Randomisierung (siehe Kapitel 2.1.2) je nach Zustand nur 3 bis 4 der sechs Geräte identifizierbar waren,
- ein Genauigkeits- bzw. Sicherheitsgewinn bei Feststellung der Einstiegshaltestelle vor allem dann möglich ist, wenn der zeitliche Abstand zwischen den von einem Smartphone ausgesandten Signalen in einem mittleren Bereich von 1 bis 2 Minuten liegt.

Bei der Interpretation dieser Ergebnisse ist zu beachten, dass die einsteigenden Fahrgäste früh genug vor Abfahrt in Reichweite des Haltestellen-Sensors kommen müssen. Beim Ausstieg ist eine Erfassung an der Haltestelle unwahrscheinlich, da sich die Fahrgäste in der Regel zügig von der Haltestelle entfernen und somit die Reichweite des Haltestellen-Sensors verlassen. Ein Mehrwert der Erfassung an den Haltestellen ist deswegen insbesondere bei den Umstiegen zu erwarten, wenn sich Personen nicht unmittelbar von der Haltestelle entfernen.

Bei kürzeren Zeitabständen (deutlich unter 1 Minute) zwischen den Smartphone-Signalen ist kein Genauigkeitsgewinn zu erwarten, da dann die Ein- bzw. Ausstiegshaltestelle bereits durch die im Fahrzeug erfassten Signale eindeutig feststellbar ist. Umgekehrt ist es bei längeren Zeitabständen von mehreren Minuten zwischen den Smartphone-Signalen zunehmend unwahrscheinlich, dass während der Wartezeit des Fahrgasts an der Einstiegshaltestelle überhaupt ein Signal ausgesendet wird.

2.6.1.4 WLAN- und Bluetooth-spezifische Ergebnisse

WLAN

Die Ergebnisse aus den Testszenarien bzgl. WLAN werden an dieser Stelle stichwortartig aufgeführt:

- **Zeitabstände zwischen Signalen:** Die Analyse der Zeitabstände zwischen Signalen zeigen ein heterogenes Bild, sodass kein gerätetypisches Sendemuster abgeleitet werden kann. Die zeitlichen Abstände variieren nach Zustand bzw. aktueller Nutzungsart und nach Verbindungszustand. Darüber hinaus sind gemäß der Literaturrecherche weitere Einflussfaktoren, etwa der Ladezustand, von Bedeutung. In Abhängigkeit von Endgerät, aktueller Nutzungsart und Verbindungszustand variiert der zeitliche Abstand zwischen Signalen von wenigen Sekunden bis zu mehr als 15 Minuten. Dies hat unmittelbare Auswirkungen auf die Erfassungsgenauigkeit der Ein- und Ausstiegshaltestelle.
- **Information Elements[5]:** Bei den hier verwendeten Endgeräten waren die HT Capabilities und Extended Capabilities konstant. Diese Information Elements eignen sich somit grundsätzlich, um randomisierte MAC-Adressen einem Endgerät zuzuordnen. Da die Information Elements verschiedener Endgeräte identisch sein können und da in den WLAN-Daten aus dem Realbetrieb festgestellt wurde, dass die von ein und demselben Endgerät ausgesandten Information Elements nicht in allen Fällen konstant sind, sind weitere Merkmale (zum Beispiel Vendor, Sequence Number) bei der Zuordnung von randomisierten MAC-Adressen zu einem Endgerät einzubeziehen.
- **MAC-Adressen:** Die MAC-Adressen sind bei den getesteten Smartphones (Tests in Messeinrichtung) durchweg randomisiert, es sei denn, es besteht eine aktive Verbindung zu einem Access Point.
- **Lokales Bit:** Zur Identifikation von randomisierten MAC-Adressen bei den WLAN-Daten eignet sich das lokale Bit uneingeschränkt. Es handelt sich dabei um ein bestimmtes Bit der übermittelten MAC-Adresse, das die Unterscheidung von den nicht randomisierten Adressen ermöglicht.
- **RSSI (Signalstärke):** Die Signalstärke ist keine verlässliche Größe, um Aussagen zur Entfernung des emittierenden Endgeräts gegenüber dem Empfänger im Fahrzeug zu treffen. Es konnten keine signifikanten Unterschiede zwischen innerhalb und außerhalb des Fahrzeugs gesendeten Signalen festgestellt werden, sowohl bei geöffneten als auch bei geschlossenen Türen.

[5] Siehe hierzu Kapitel 2.1.2.

Bluetooth

Bei Bluetooth verdeutlichen die Testergebnisse, dass lediglich die „gehashte" MAC-Adresse und der Zeitstempel erfasst werden. Eine Analyse der zeitlichen Abstände zwischen den erfassten Daten zeigt, dass die Endgeräte nicht in der Häufigkeit der versendeten Inquiry-Requests (alle 10 Sekunden) erfasst werden. Auch hier variiert der zeitliche Abstand zwischen Signalen eines Endgeräts, sodass teilweise mehrere Signale pro Minute erfasst werden und andererseits große zeitliche Abstände (>15 Minuten) zwischen zwei Signalen liegen. Es ist davon auszugehen, dass das Inquiry-Response-Verhalten von weiteren, an dieser Stelle unbekannten Faktoren abhängt.

2.6.2 Analyse der Daten aus dem Pilotbetrieb

2.6.2.1 Randomisierung der MAC-Adressen

Eine Betrachtung der WLAN-Daten zeigt, dass in der ersten Erhebungsperiode in 2020 90% der erfassten Signale eine feste bzw. nicht randomisierte MAC-Adresse besitzen. Die Signale stammen sowohl von Geräten innerhalb des Fahrzeugs als auch von außerhalb. Wie in Kapitel 2.5.1 geschildert, handelt es sich bei einem Großteil der erfassten Daten um Störddaten. Diese gehen unter anderem von Geräten anderer Verkehrsteilnehmenden aus oder von Geräten in Geschäften und Wohnungen in der Nähe der Fahrtroute des Fahrzeugs.

Der Anteil randomisierter Signale WLAN/Bluetooth-Signale stieg während der Projektlaufzeit zwischen dem 4. Quartal 2020 und dem 4. Quartal 2021 signifikant an. Dies ist erklärbar durch die zunehmende Marktdurchdringung von Smartphones mit Betriebssystem-Versionen, bei denen die Randomisierung standardmäßig aktiviert ist – dies ist bei Android seit Version 10 der Fall, die im September 2019 erschien (Bieland 2022).

Hinsichtlich der randomisierten MAC-Adressen aus dem Realbetrieb wurde im Rahmen einer automatisierten Analyse der mitgesendeten Information Elements und des Herstellers ein Ähnlichkeitsmaß entwickelt. Unter Berücksichtigung von Ausschlüssen, die sich aus zeitlichen Überschneidungen zwischen den Sichtbarkeitszeiträumen der MAC-Adressen ergeben, können aus dem Ähnlichkeitsmaß Wahrscheinlichkeitsaussagen darüber abgeleitet werden, ob zwei verschiedene randomisierte MAC-Adressen zu ein und demselben Endgerät gehören. Darauf aufbauend könnte unter zusätzlicher Heranziehung der „Sequence Number" ein Tool zur stochastischen (wahrscheinlichkeitsbasierten) Identifikation von Endgeräten mit randomisierten MAC-Adressen entwickelt werden. Diese Arbeiten wurden jedoch im Rahmen der Priorisierung der verbleibenden Arbeitsschritte in diesem Forschungsprojekt zurückgestellt.

2.6.2.2 Filterung von Störddaten

Die erfassten WLAN- und Bluetooth-Daten wurden mit den AFZS-Daten zusammengeführt. Dadurch war ein quantitativer Abgleich mit den gemessenen Ein- und Ausstiegen möglich. Daraus ergab sich ein hohes Maß an Störddaten in den WLAN- und Bluetooth-Daten. Ein weiterer Schwerpunkt der Datenanalyse zielte deshalb darauf ab, geeignete Ansätze zur Filterung von Störddaten abzuleiten (siehe Kapitel 2.1.3).

- Filterung **unvollständiger Daten**: Signale, die aufgrund von Datenlücken keiner Linienfahrt eindeutig zugeordnet werden können, sind unbrauchbar und werden aussortiert. Ebenso werden Fahrzeugfahrten verworfen, zu denen keine AFZS-Daten vorliegen.
- Filter nach **Einsatzzeiten des Fahrzeugs**: Die MAC-Adressen von Signalen, die außerhalb der Zeiten empfangen wurden, zu denen das Erhebungsfahrzeug im Fahrgasteinsatz war (Standzeiten im Depot und an Endhaltestellen, Ein- und Aussetzfahrten) werden aussortiert.
- Filterung von Signalen mit **randomisierter MAC-Adresse**: Diese Signale werden aussortiert, da sie nach derzeitigem Stand nicht sicher von einem Burst zum nächsten nachverfolgt werden können.
- Filter nach **Anzahl Fahrten mit der gleichen Linie**: MAC-Adressen, die innerhalb ein und desselben Tages zu häufig in der gleichen Linie auftreten, werden aussortiert, da ein solches Fahrtverhalten für einen Fahrgast als unrealistisch angesehen wird.
- Filter nach **Signalstärke**: Datensätze, deren Signalstärke (RSSI, received signal strength indicator) einen einstellbaren Schwellenwert unterschreitet, werden aussortiert.
- **GPS-Filter**: Unterschreitet die räumliche Entfernung, die zwischen den Fahrzeugpositionen bei der ersten und bei der letzten Erfassung einer MAC-Adresse liegt, einen einstellbaren Schwellenwert, werden die entsprechenden Daten aussortiert.

2.7 Verfahrensentwicklung

Das Ziel der Verfahrensentwicklung bestand darin, die Methode zur Datenfusion zu entwickeln. Hierzu wurden die notwendigen Arbeitsschritte identifiziert und softwaretechnisch abgebildet. Darüber hinaus wurden Gütemaße für die Kalibrierung der Verfahrensparameter definiert und umgesetzt. Bei der Kalibrierung wurden zwei Ansätze verfolgt, um im Sinne der iterativen Bearbeitung zeitnah für die Produktivimplementierung eine erste Auswahl an Parametersets zu treffen, bevor die umfangreiche Kalibrierung auf Basis der erfassten WLAN-/Bluetooth-Daten abgeschlossen werden konnte.

2.7.1 Verfahrensschritte

Das entwickelte Verfahren beinhaltet vier Schritte, die in Abbildung 13 dargestellt sind. Im ersten Schritt 1 werden die erfassten Daten importiert und analysiert. Unbrauchbare Daten werden an dieser Stelle im Verarbeitungsprozess aussortiert und gelöscht. Außerdem werden die Daten aggregiert, wodurch das Datenvolumen bereits deutlich reduziert wird (siehe Kapitel 2.5.1). Daran schließen sich die Schritte Bildung von Startmatrizen, Hochrechnung der Startmatrizen und Bewertung der Hochrechnungsergebnisse an.

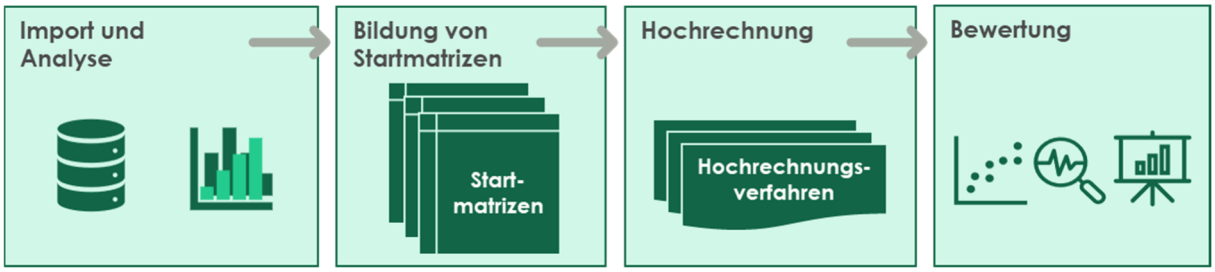


Abbildung 13: Verfahrensentwicklung (schematisch)

Für die Bildung von Startmatrizen und zur Hochrechnung entwickelte WVI geeignete Softwaretools. Die Startmatrizen können im „Startmatrizen und Filter-Tool (StauFi)“ auf Basis der unterschiedlichen Datenquellen generiert werden. Dabei kann zwischen unterschiedlichen Filtern gewählt werden. Ziel der Filterung ist, nach der groben Datenbereinigung beim Import weitere Störsignale auszusortieren. Die Daten können hier zusätzlich mittels einer Fuzzy-Methode bearbeitet werden. Das Ziel dieser Methode ist es, Unschärfen bezüglich der Ein- und Ausstiegshaltestellen zu korrigieren.

Das „Hochrechnungs-Tool (HR-Tool)“ liest die Startmatrizen und die zugehörigen AFZ-Daten ein und rechnet die Startmatrizen auf die AFZ-Daten hoch. Dabei kann zwischen verschiedenen Hochrechnungsverfahren mit variierenden Größen für die Hochrechnungsparameter gewählt werden. Für jede Hochrechnung werden diverse Gütemaße berechnet, die bei der Kalibrierung zum Einsatz kommen. Für die Berechnung der Gütemaße wird die mittels Hochrechnung erzeugte Ergebnismatrix mit der aus den Befragungsdaten abgeleiteten Zielmatrix gegenübergestellt und die Abweichungen bestimmt.

Die Bewertung der großen Anzahl an möglichen Hochrechnungsergebnissen erfolgt mit Hilfe der Gütemaße. Das Ziel der Bewertung ist es, in Abhängigkeit der Inputgrößen geeignete Hochrechnungsverfahren und -parameter zu identifizieren. Das Vorgehen wird in den folgenden Abschnitten beschrieben.

2.7.2 Gütemaße zur Bewertung

Für die Bewertung der Abweichungen zwischen Referenz- und Bewertungsmatrix wurden die folgenden Gütemaße entwickelt:

- euklidische Distanz bzw. Fehlerquadratsumme
- GEH-Wert
- Abweichung der Verkehrsleistung in drei Entfernungsklassen
- Treffermaß
- Überfüllungsfaktor

Die euklidische Distanz ist die Verallgemeinerung des üblichen Abstandsmaßes im dreidimensionalen Raum auf höherdimensionale Räume (hier den Raum der Matrizen mit der passenden Anzahl Spalten und Zeilen).

Der GEH-Wert vergleicht jeweils ein Wertepaar, hier den Eintrag in der Bewertungs- und in der Referenzmatrix für die gleiche Relation. Dabei stellt die Bewertungsmatrix das Ergebnis der hochgerechneten Startmatrix dar. Die Referenzmatrix ergibt sich aus den aufbereiteten Ergebnissen der Fahrgastbefragung Der GEH-Wert ist dann das geometrische Mittel aus absolutem und relativem Fehler.

$$GEH_{ij} = \sqrt{\frac{2 * \left(f_{ijhr} - f_{ijr}\right)^2}{\left(f_{ijhr} + f_{ijr}\right)}}$$

Mit

f_{ijhr}	Berechnete Anzahl von Fahrten (Fahrgästen) von Einstiegshaltestelle i nach Ausstiegshaltestelle j
f_{ijr}	Reale Anzahl von Fahrten (Fahrgästen) von Einstiegshaltestelle i nach Ausstiegshaltestelle j

Es wird die Anzahl der Referenzfahrten bestimmt, bei denen ein Grenzwert von 0,7 bei mindestens 95 % der Relationen und ein Grenzwert von 1,5 bei allen Relationen eingehalten wird.

Bei der Abweichung der Verkehrsleistung wird die Summe der Beträge der Abweichung der Verkehrsleistung in den einzelnen Entfernungsklassen betrachtet.

Das Treffermaß setzt die Anzahl der Relationen, die sowohl in der Referenz- als auch in der Bewertungsmatrix (mit einem Wert größer als 0) belegt sind, ins Verhältnis zu Gesamtzahl der belegten Relationen in der Referenzmatrix.

Der Überfüllungsfaktor bezeichnet den Quotienten aus der Anzahl belegter Relationen in der Bewertungsmatrix und der Anzahl belegter Relationen in der Referenzmatrix.

2.7.3 Kalibrierung der Hochrechnungsparameter mit synthetischen Startmatrizen

Im Kontext der Verfahrensentwicklung wurden verschiedene Analysen zur Kalibrierung der Hochrechnungsparameter durchgeführt. Datengrundlage für diese Analysen waren die nutzbaren Ergebnisse der Fahrgastbefragungen (siehe Kapitel 2.5.2). Aus diesen Daten wurden mit Hilfe des WVI-eigenen Softwaretools StauFi sogenannte Zielmatrizen erzeugt, die für die weitere Verfahrensentwicklung als Referenzmatrizen herangezogen wurden.

Um den Einfluss der Filter und Parameter ohne externe Einflüsse überprüfen zu können, wurden aus den Zielmatrizen synthetische Startmatrizen gebildet. Die Zielmatrizen wurden dazu per Zufallsauswahl reduziert oder überfüllt, sowie systematisch dahingehend manipuliert, dass Kurz-, Mittel- und Langstrecken über- oder unterrepräsentiert sind.

Anschließend wurden die so erzeugten synthetischen Startmatrizen mit verschiedenen Hochrechnungsverfahren und -parametern mit Hilfe des WVI-eigenen Softwaretools „HR-Tool“ hochgerechnet und die Gütemaße berechnet.

Durch dieses Vorgehen wurde ein erster Vorschlag zur Identifizierung eines First-best-Sets an Filtern, Hochrechnungsverfahren und Hochrechnungsparametern erarbeitet, der in die Produktivimplementierung Eingang fand. Die Ergebnisse wurden anschließend im Rahmen der Kalibrierung für das Verfahren mit WLAN- und Bluetooth-Daten verifiziert und bei Bedarf korrigiert.

2.7.4 Kalibrierung der Filter und Hochrechnungsparameter

2.7.4.1 Vorgehen

Grundlage für die Kalibrierung waren die 29 uneingeschränkt nutzbaren Referenzfahrten (siehe Kapitel 2.5.2). Aus den WLAN- und Bluetooth-Daten dieser Fahrten wurden Ergebnismatrizen erzeugt. Hierfür fanden 81 Parameterkonstellationen für die Filterung und 28 Parameterkonstellationen für die Hochrechnung Anwendung.

Zur Ermittlung der optimalen Parameterkombination wurde für jede Parameterkombination ein Punktwert berechnet, der sich aus einer gewichteten Summe der normierten Bewertungen anhand der einzelnen Gütemaße berechnet. Die vergebenen Gewichte sollen der Bedeutung des jeweiligen Gütemaßes für die Nutzbarkeit des Verfahrens entsprechen. Die optimale Parameterkombination ist diejenige mit dem höchsten Punktwert.

Abbildung 14 zeigt das Vorgehen zur Bewertung der Verfahrensergebnisse unter Verwendung dieser Gütemaße.

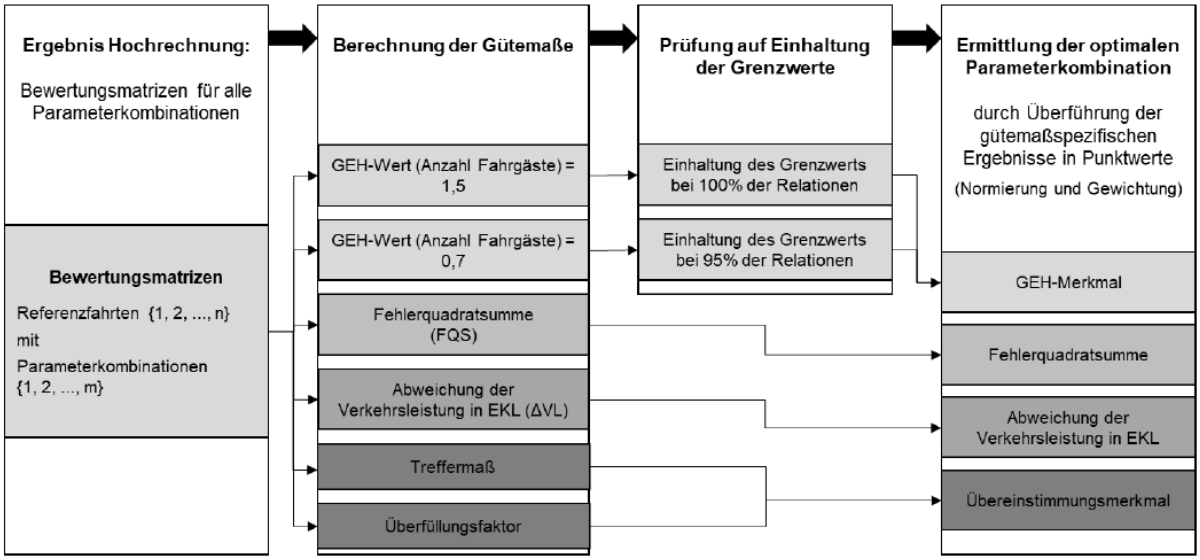


Abbildung 14: Vorgehen zur Ermittlung der optimalen Parameterkombination

2.7.4.2 Ergebnisse

Die drei optimalen Parameterkombinationen – sie erreichen alle den gleichen Punktwert 0,420 (theoretisch möglicher Maximalwert ist 1) – basieren alle auf der Kombination der Datenquellen WLAN und Bluetooth. Diese Parameterkombinationen greifen sowohl auf WLAN- als auch auf Bluetooth-Daten als Quelle zurück. Die Eingangsdaten zur Erstellung der Ausgangsmatrix werden hier durch eine Filterung auf Basis der GPS-Luftlinienentfernung reduziert und mit dem Hochrechnungs-Parameterset Fratar mit Hilfswert 0,00001 und Reduktionsschwellenwert 1,0 hochgerechnet. Sie unterscheiden sich lediglich im Filterset, nämlich GPS-Filter mit Schwellenwert 100 bzw. 400 bzw. 600 m Luftlinienentfernung zwischen der ersten und letzten Erfassung ein und derselben MAC-Adresse. Tabelle 4 listet die optimalen Parameterkombinationen mit den dabei jeweils erreichten Gütemaßen auf.

Datenquelle	Filterset	HR-Set	Fuzzy	GEH-Merkmal	FQS	Summe des Betrags der Abweichung der Verkehrsleistung in den EKL	Übereinstimmungsmerkmal	Summe (Normierung)
WLAN_BT(Sum)	P50_GPS100	FRAT_h0,00001_r0,5_----	x	0,36	0,33	0,31	0,50	0,40
WLAN_BT(Sum)	P50_GPS100	FRAT_h0,00001_r1,0_----		0,31	0,20	0,32	0,59	0,40
WLAN_BT(Sum)	P51_GPS200	FRAT_h0,00001_r0,5_----	x	0,33	0,29	0,36	0,52	0,40
WLAN_BT(Sum)	P51_GPS200	FRAT_h0,00001_r1,0_----	x	0,30	0,18	0,36	0,58	0,40
WLAN_BT(Sum)	P51_GPS200	FRAT_h1,00000_r0,5_----	x	0,36	0,28	0,28	0,51	0,40
WLAN_BT(Sum)	P51_GPS200	FRAT_h1,00000_r1,0_----	x	0,33	0,17	0,23	0,57	0,40
WLAN_BT(Sum)	P52_GPS400	FRAT_h0,00001_r1,0_----		0,33	0,16	0,27	0,57	0,40

Tabelle 4: Optimale Parameterkombinationen.

Die besten Ergebnisse werden bei Nutzung von WLAN- und Bluetooth-Daten als Datenquelle, einem auf GPS-Luftlinienentfernung basierenden Filterset (100m, 200m und 400m) und der Hochrechnung mit dem Fratar-Verfahren und einem Reduktionsgrenzwert von mindestens 0,5 erzielt. Die ermittelten Parameterkombinationen haben sich auch bei differenzierter Betrachtung der Referenzfahrten nach Linien und Nachfrageklassen als robust erwiesen. Insbesondere die Parameterkombination unter Verwendung von WLAN- und Bluetooth-Daten, dem Filterset P51_GPS200 mit Fuzzifizierung und Hochrechnung mit dem Fratar-Verfahren (Hilfswert 1,000, Reduktionsgrenzwert 0,5) liegt durchweg unter den besten 10% der betrachteten Parameterkombinationen. Darüber hinaus schneiden die ermittelten Parameterkombinationen besser ab als

- das beste Verfahren ohne Ausgangsdaten,
- das beste Verfahren ohne Filterung der Eingangsdaten sowie
- das beste Verfahren ohne Hochrechnung der Ausgangsmatrizen.

Dies bestätigt den Mehrwert des Verfahrens sowie der entwickelten Verfahrensschritte.

Wichtig für die Interpretation der Ergebnisse der besten Parameterkombinationen ist der Vergleich mit Referenzverfahren, die nur auf einer Hochrechnung basieren, ohne zusätzliche Informationen aus den WLAN- und Bluetooth-Daten zu verwenden. Für die Vergleichsverfahren hat sich gezeigt, dass das Hochrechnungsverfahren nach Li und

Cassidy auf Basis der gewählten Alternativen geringere Punktwerte erreicht als das Fratar-Verfahren.

Ergänzend zu der gerade dargestellten Optimierung anhand des Punktwerts, der alle Gütemaße zusammenfasst, ist es auch möglich, anhand der einzelnen Gütemaße zu optimieren. Bei dieser gütemaßspezifischen Optimierung ergeben sich jeweils andere optimale Parameterkombinationen.

Die gütemaßspezifischen Auswertungen zeigen, dass keine Parameterkombination mehrfach bei den besten Ergebnissen genannt wurde. Die Ergebnisse der Optimierung sind also nicht robust gegenüber der Wahl des Gütemaßes. Auffällig ist auch, dass sich unter allen Parameterkombinationen keine Kombination mit der alleinigen Datenquelle Bluetooth befindet. Die Bluetooth-Daten sind daher auf Basis dieser Ergebnisse lediglich als ergänzend zu betrachten. Regelmäßig tritt dagegen die Datenquelle WLAN und Bluetooth auf sowie das Filterset GPS 100 auf. Bei der Betrachtung der Hochrechnungsverfahren zeigt sich, dass insgesamt kleine Hilfswerte (0,00001) gegenüber dem Hilfswert 1 bessere Ergebnisse erzielen. Die Verwendung kleiner Hilfswerte führt dazu, dass durch WLAN- und Bluetooth-Daten besetzte Relationen ein (deutlich) höheres Gewicht besitzen. Bei Nutzung der Fuzzifizierung wird dieser Effekt abgeschwächt.

Die Vergleichsverfahren führen bei den meisten Gütemaßen zu Ergebnissen mit geringerer Güte als andere Parameterkombinationen.

2.8 Produktivimplementierung

Die Implementierung umfasst die Bildung der Ergebnismatrix im Produktivsystem und deren Bereitstellung über das Statistiktool MOBILEstatistics für den NVV.

2.8.1 Produktivsystem

In der Produktivimplementierung wurden die für die Datenverarbeitung entwickelten Algorithmen und das kalibrierte Verfahren in ein Produktivsystem überführt und umgesetzt. Das Produktivsystem ermöglicht es dem NVV, die Daten sowie die Quelle-Ziel-Matrizen direkt aus dem Statistiktool einzusehen. Im Wesentlichen wurden die Arbeitsschritte implementiert, die zuvor schon in Kapitel 2.7.1 beschrieben und in Abbildung 15 nochmal skizziert werden.

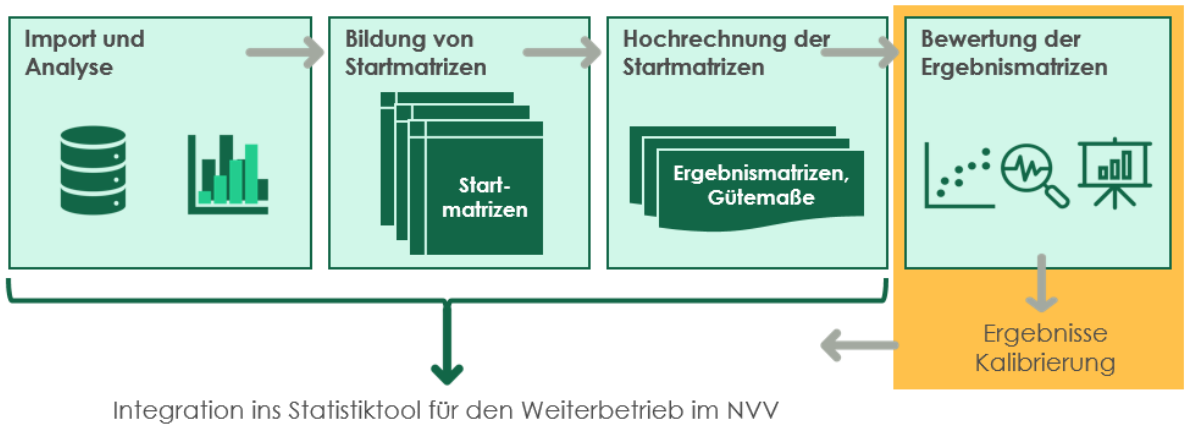


Abbildung 15: Integration der entwickelten Verfahren

Das entwickelte Verfahren zur Bestimmung der Quelle-Ziel-Matrizen wurde in die INIT Big-Data Pipeline integriert, so dass Verkehrsstromdaten linienfahrtgenau und tagesaktuell dargestellt werden können. Abbildung 16 zeigt das Verfahren im INIT Backend in dem die Daten gefiltert, bereinigt und integriert werden. Die so generierten Quelle-Ziel-Matrizen werden anschließend über einen Kafka Broker zur Weiterverarbeitung im Statistiktool MOBILEstatistics bereitgestellt.

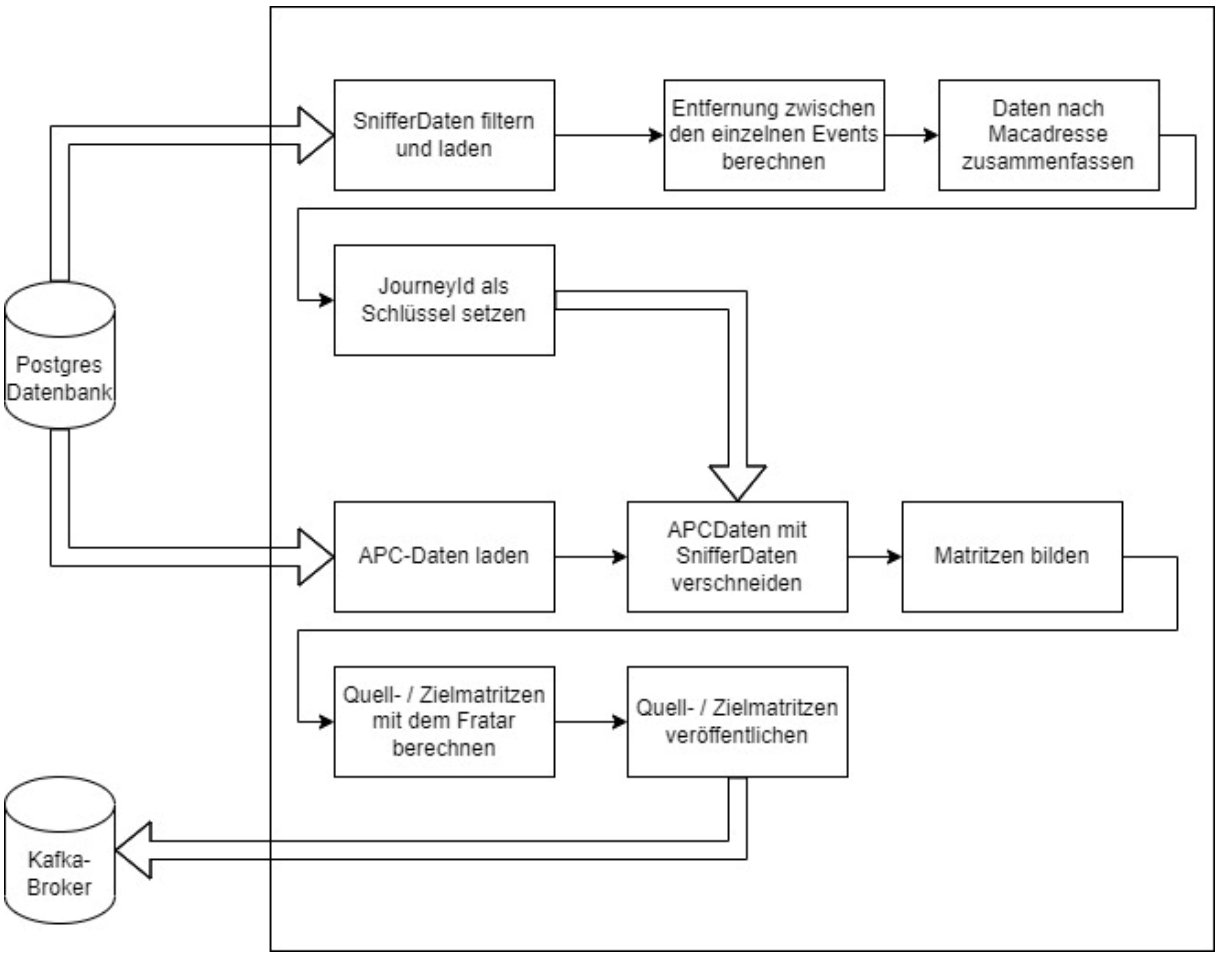


Abbildung 16: Big Data Pipeline im Produktivsystem

2.8.2 Statistiktool

Die zusätzliche Funktionalität wird in MOBILEstatistics für den Projektpartner NVV für die Dauer der Realerprobung bis 2024 umgesetzt und zur Verfügung gestellt. Die Anwendenden können dabei tagesaktuell die Quelle-Ziel-Matrizen generieren sowie die Ein- und Aussteiger pro Fahrt und Haltestelle angezeigt bekommen und zur weiteren Analyse bereitgestellt werden. Exemplarisch ist ein Auszug aus dem Statistiktool ist in Abbildung 17 zu sehen.

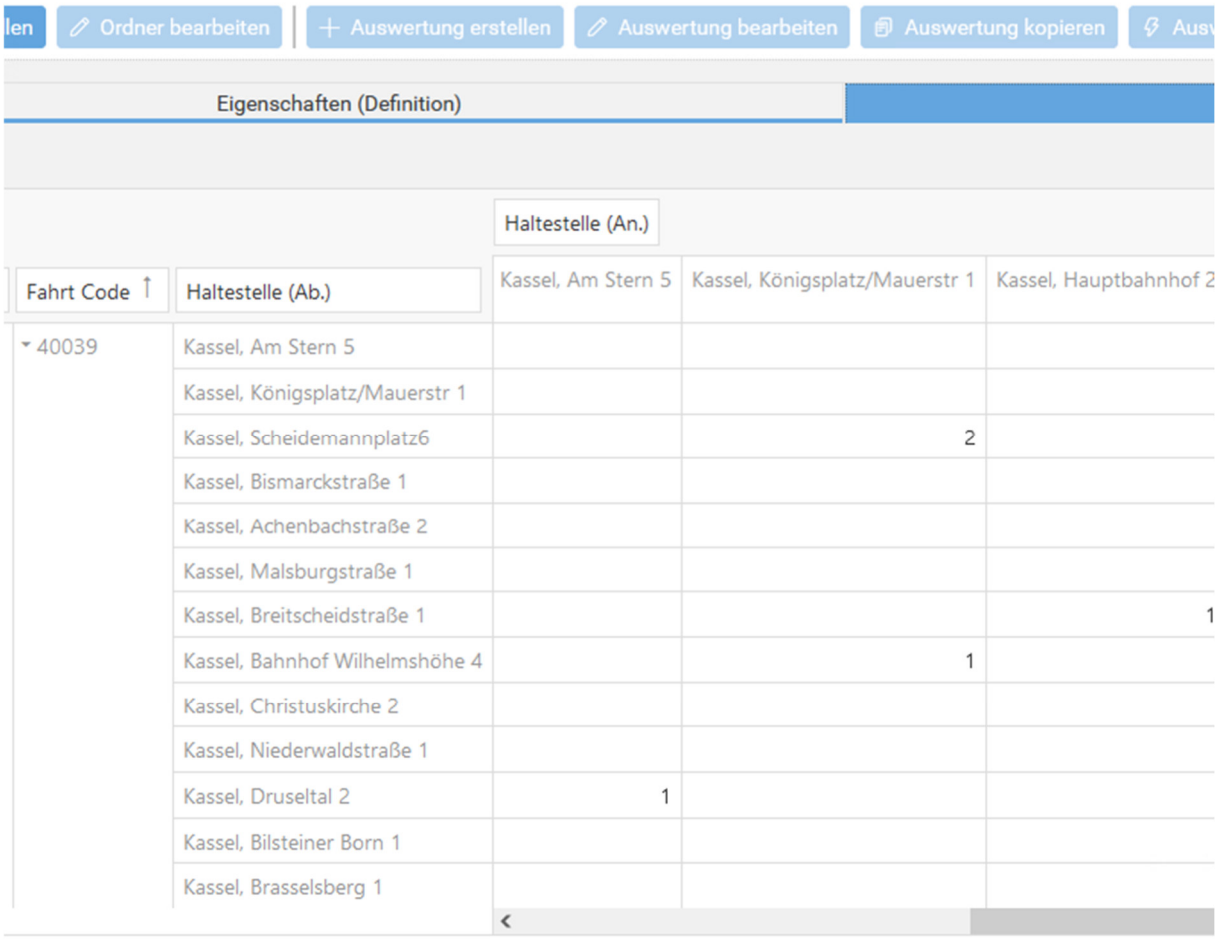


Abbildung 17: Integration der Ergebnisse im Statistiktool MOBILEstatitics der INIT

Für die Produktivimplementierung lässt sich zusammenfassend festhalten, dass es gelungen ist, das entwickelte und kalibrierte Verfahren vollständig zu implementieren und für den Pilotbetrieb bereitzustellen. Darüber hinaus ist es gelungen, die berechneten Ergebnisse über die Statistiksoftware für die Anwendung im NVV zugänglich zu machen. Das implementierte Verfahren umfasst die Schritte der Datenselektion und -aggregation, die Bildung von Startmatrizen aus den in den Fahrzeugen erfassten WLAN- und Bluetooth-Daten unter Verwendung geeigneter Filter sowie die Hochrechnung der erzeugten Startmatrizen mit einem geeigneten Verfahren.

Die berechneten Quelle-Ziel-Matrizen liegen tagesaktuell für jede erfasste Linienfahrt vor. Bei den Ergebnissen handelt es sich durch die verschiedenen Bearbeitungsschritte um statistische Größen, die keine Rückschlüsse auf Personen ermöglichen.

2.9 Pilottest

Für die zweite Projektphase, in welcher die Validierung des entwickelten Verfahrens erfolgen sollte, wurden weitere Fahrzeuge und Haltestellen mit der entwickelten Hardware ausgestattet. Zusätzlich wurde eine weitere Fahrgastbefragung durchgeführt, um Referenzdaten für die Validierung zur Verfügung zu haben.

2.9.1 Hardwareausstattung

Als eine der stärksten Stadt-Umland-Linien des NVV mit etwa 8.000 Fahrgästen pro Tag wurde die Buslinie 52 ausgewählt. Des Weiteren wurden die Linien des Stadtbusverkehrs in Bad Wildungen 590.1-4 herangezogen. Hierbei handelt es sich um ein kleinstädtisches Bussystem mit etwa 2.000 Fahrgästen pro Tag[6], welches in Vorbereitung auf das Projekt mit Zählsystemen ausgestattet wurde. Zwischen den vier Stadtbuslinien bestehen vor allem im Zentrum von Bad Wildungen, welches auch den Kreuzungspunkt dieser Linien bildet, größere Umsteigebeziehungen (siehe Abbildung 18). Da das Verfahren ausschließlich anhand von regionalen Linien entwickelt wurde, sollte der Stadtverkehr in Bad Wildungen die Möglichkeit bieten, das Verfahren auch im städtischen Bereich zu prüfen und in einem höheren Maß Umsteigebeziehungen zu erfassen.

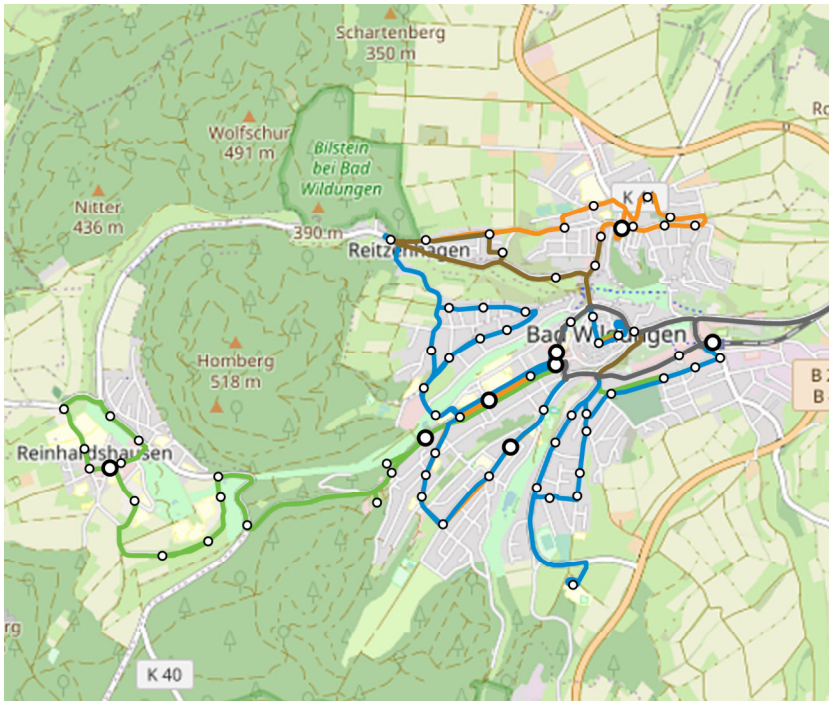


Abbildung 18: Stadtbuslinien in Bad Wildungen

[6] Alle genannten Fahrgastzahlen stammen aus der Zeit vor der Corona-Pandemie.

Alle regulär auf den genannten Linien verkehrenden Fahrzeuge wurden mit den benötigten Hardwarekomponenten ausgestattet. Auch konnte an neun Haltestellen in Bad Wildungen in Zusammenarbeit mit dem städtischen Bauhof eine Stromversorgung mittels der Vitrinenbeleuchtung und eine Ausrüstung mit der Erfassungshardware erreicht werden. Damit sollte eine Datenbasis für den Pilottest und die Validierungsphase erreicht werden.

Eine ursprünglich geplante Ausrüstung der RegioTram-Linie RT4 war im Rahmen der Projektlaufzeit nicht möglich, da sich die Ausstattung mit Fahrgast-WLAN und Zählsystem stark verzögerte und die Zuständigkeiten projektextern bei der RegioTram Gesellschaft mbH (Betreiberin der RegioTram) und der Regionalbahn Kassel GmbH (Eigentümerin der Fahrzeuge) liegen.

2.9.2 Zweite Fahrgastbefragung

Im Vorfeld der Fahrgastbefragung in Phase 2, die sowohl auf den Testlinien aus Phase 1 als auch auf den zusätzlichen Testlinien der Phase 2 durchgeführt wurde, kam es aufgrund organisatorischer und technischer Probleme beim Einbau der Hardware zu Verzögerungen. Die Hardwareausstattung der Fahrzeuge konnte erst im Laufe der Monate September und Oktober 2021 sukzessive für die einzelnen Linien fertig gestellt werden. Daher konnte die zweite Fahrgastbefragung auf den Linien der Phase 1 und Phase 2 erst im Zeitraum von 13.9. bis 27.11.2021 durchgeführt werden.

Auf den acht Erhebungslinien waren insgesamt 208 Erhebungsfahrten geplant. Aufgrund der genannten Verzögerungen, die wegen auslaufender Arbeitsverträge zu einem Mangel an Erhebungspersonal führten, konnten nur 135 Erhebungsfahrten durchgeführt werden. Anschließend wurden die Befragungsdaten wie geplant für die konzipierten quantitativen Tests aufbereitet. Hierzu kam eine vom VPVS entwickelte Java-Software zum Einsatz. Bei der Aufbereitung und Prüfung der Befragungsergebnisse wurden die gleichen Kriterien wie in Phase 1 zu Grunde gelegt (siehe Kapitel 2.5.2).

2.9.3 Ergebnisse

In Bezug auf die Haltestellen-Sensoren setzten sich die bereits in Phase 1 auftretenden Probleme in Phase 2 fort. So sendeten zunächst noch fünf der neu installierten Haltestellensensoren Daten aus, jedoch ging die Zahl im Laufe des Pilottests stark zurück. Zum Ende des Jahres 2021 konnten keine Daten mehr von den ausgestatteten Haltestellen empfangen werden. Aus technischer Sicht besteht hier weiterer Entwicklungsbedarf, der von nachfolgenden Projekten angegangen werden kann.

Die Ergebnisse der Fahrgastbefragungen sind in Tabelle 5 zusammengefasst.

Projekt-phase	Befragungs-zeitraum	Erhebungs-linien	Fahrten geplant	Fahrten durch-geführt	Fahrten prinzipiell nutzbar	Fahrten als Referenz nutzbar
1	17.9.-13.11.2020	4	100	79	73	29 (+18)
2	13.9.-27.11.2021	8	208	135	92	9

Tabelle 5: Übersicht Fahrgastbefragungen

Aufgrund von Ausfällen bei den AFZ-Systemen und Fehlern in den Daten lagen in Phase 2 nur von 92 Erhebungsfahrten vollständige und plausible AFZ-Daten vor. Eine Ursache für die fehlenden bzw. fehlerhaften AFZ-Daten waren mehrere baustellenbedingte Umleitungen auf den Untersuchungslinien, die eine korrekte Zuordnung der gezählten ein- und aussteigenden Fahrgäste zu den jeweils real angefahrenen Haltestellen verhinderten.

Von den verbleibenden 92 Fahrten konnten wegen zu großer Abweichungen zwischen den Zähldaten aus dem AFZS einerseits und der Fahrgastbefragung andererseits nur aus 38 Erhebungsfahrten Quelle-Ziel-Matrizen als Referenzdaten gewonnen werden. Schließlich führten temporäre Ausfälle der Sensoren in den Fahrzeugen für die Erfassung der WLAN- und Bluetooth-Signale dazu, dass davon lediglich für neun Erhebungsfahrten entsprechende Daten vorlagen. Da eine derart geringe Zahl von Erhebungsfahrten keine aussagekräftige Validierung der Verfahrensergebnisse erlaubt, wurde auf diese Prüfung verzichtet.

2.10 Empfehlungen und Verwertung

Auf Basis der Erkenntnisse der in diesem Bericht beschriebenen Projektbearbeitung können Empfehlungen für den weiteren Einsatz des Verfahrens gegeben werden. Zudem diskutierte das Projektkonsortium bereits projektbegleitend auch Ideen für eine mögliche Weiterentwicklung und Verwertung des Verfahrens, welche hier benannt werden sollen.

2.10.1 Empfehlungen

Im Verlauf des Projektes erschwerten einige Herausforderungen und Problemfelder die Projektbearbeitung. Aus diesen „Stolpersteinen“ können im Gegenzug entsprechende Empfehlungen für weitere Forschungsprojekte des mFund sowie für die mögliche Weiterentwicklung dieses Verfahrens abgeleitet werden.

Die Herausforderungen innerhalb des Projektes traten im Wesentlichen im Zusammenhang mit der Erfassung der Datengrundlage (siehe Kapitel 2.4) auf und wirkten sich auf die folgenden Arbeitspakete aus. So führte der notwendige Austausch von nicht ausreichend performanten Hardwarekomponenten bereits zu Beginn des Projektes zu ersten Verzögerungen aufgrund der abzuwartenden Lieferfristen. In diesem Fall hätte sich ein noch frühzeitigerer Test der ausgewählten Komponenten positiv auf die Zeitplanung ausgewirkt. Auch können zeitliche Puffer innerhalb der einzelnen Arbeitspakete für eventuell notwendige Anpassungen von Software aufgrund vorher nicht erwartbarer Umstände (hier: die Überarbeitung der WLAN-Empfänger durch den Hersteller) empfohlen werden.

Bezüglich der Haltestellensensoren lassen sich mehrere Problemfelder identifizieren. Zum einen kam es hardwarebedingt zum häufigen Ausfall dieser Sensoren. Außerdem konnte keine zuverlässige Stromversorgung mittels der Haltestellenvitrinen-Beleuchtung sichergestellt werden. Auch erschwerte der unzureichende Mobilfunkempfang am Einbauort den Datenempfang. Aufgrund dieser Umstände konnten die WLAN- und Bluetooth-Signale von den ausgestatteten Haltestellen leider nicht zuverlässig empfangen werden und gehen nicht in das entwickelte Verfahren ein. Sollte innerhalb weiterer Feldversuche der Nutzen dieser Daten für das Verfahren geprüft werden, so empfiehlt sich ein robusterer Hardwareaufbau mit zuverlässiger Stromquelle. Auch sollte ein möglicher Wartungsaufwand der Sensoren je nach zeitlichem Umfang dieses Versuchs in die Auswahl der Haltestellenstandorte und den Zeitplan einfließen.

Der umfassende Austausch mit der Kasseler Verkehrs-Gesellschaft (KVG) über eine schließlich nicht mehr realisierte Haltestellenausstattung im Stadtgebiet von Kassel zeigte, dass die Zuständigkeiten für Haltestellen und die jeweiligen Ansprüche an eine Hardwareausstattung sowohl für Forschungsprojekte als auch in der Praxis sehr frühzeitig geklärt und berücksichtigt werden sollten.

Zwar verzeichneten auch die Fahrzeugsensoren vereinzelt Ausfälle, jedoch konnten diese meist kurzfristig durch die INIT behoben werden. Durch die Implementierung der Fahrzeugsensoren in den Fahrzeugbordrechner Copilot der INIT inkl. der benötigten Zulassungen konnte hier zudem anders als bei den Haltestellensensoren eine Praxistauglichkeit gezeigt und sichergestellt werden.

Bei Überprüfung der Sendeeigenschaften verschiedener Endgeräte ließ sich eine starke Heterogenität ebendieser nachweisen. Eine gezielte Filterung der relevanten Signale war somit erschwert. Weiterhin steigerte sich der Anteil der Smartphones mit aktiver Randomisierung im Laufe des Projektes, da neuere Betriebssysteme diese standardmäßig durchführen. Dies verringerte die Datenbasis für die im Projekt untersuchte passive WLAN-/Bluetooth-Datenerfassung. Positiv konnte hingegen festgestellt werden, dass mit 85% der in Phase 2 befragten Fahrgäste die große Mehrheit ein mobiles Endgerät mit sich führten. Für den weiteren Einsatz bzw. die weitere Entwicklung des Verfahrens bietet sich die Hinzunahme weiterer Datenklassen (bspw. aus durch ÖPNV-Betreiber bereitgestelltem aktivem WLAN in den Fahrzeugen oder durch die Kunden

freigegebene Informationen aus der Nutzung von Mobilitäts-Apps) an, um die grundsätzlich vorhandene Datenbasis voll auszuschöpfen.

Für die Festlegung der Ziele und Anforderungen (siehe Kapitel 2.2) lässt sich festhalten, dass eine Überprüfung der vorhandenen und nutzbaren Datenbasis für ein derartiges Projekt bzw. für die Implementierung des entwickelten Verfahrens unbedingt sinnvoll und notwendig ist. Dies zeigt sich am Beispiel der Fahrplanauskunftsdaten, von denen lediglich die Anfragen an das Auskunftssystem, nicht jedoch die dargestellten Fahrtmöglichkeiten, ohne weitere Aufwendungen verwendbar gewesen wären. Weiterhin wird empfohlen, mit entsprechenden Dienstleistern von Beginn an eine mögliche Nutzung von Daten für entsprechende Projekte oder Verfahren vertraglich festzuhalten.

Während der zweiten Fahrgastbefragung stellten Ausfälle des AFZ-Systems und fehlerhaftes Haltestellen-Matching aufgrund von Baustellenfahrplänen auf den Untersuchungslinien die wesentlichen Problematiken dar. Grundsätzlich bilden Fahrgastzähldaten aus AFZ-Systemen eine wichtige Datenbasis für das entwickelte Verfahren. Zum jetzigen Zeitpunkt liefern AFZS die umfangreichsten und validesten Daten über die ÖPNV-Nachfrage. Im Laufe des Projektes konnte beobachtet werden, dass Ausfälle oder Qualitätsmängel große Auswirkungen auf die Qualität der Verfahrensergebnisse haben. Dies zeigte sich bereits zu Beginn des Projektes, als Lücken in den ÖPNV-Kontextdaten nach dem Fahrplanwechsel zu größeren Lücken in den Erfassungs-Rohdaten führten. Dies konnte unter anderem durch die Entwicklung des QADABRA-Algorithmus durch die INIT behoben werden. Obwohl AFZ-Daten eine hohe Datenqualität aufweisen können, ist im jeweiligen Projekt eine Untersuchung der Datenqualität bezüglich der Güte und Nutzbarkeit für den geplanten Verwendungszweck zu Projektbeginn und während der Projektlaufzeit empfehlenswert. Für den zukünftigen Einsatz bzw. die Weiterentwicklung des Verfahrens sowie für weitere AFZ-datengestützte Forschungsprojekte ist ein regelmäßiges Monitoring der Zähldaten empfehlenswert.

2.10.2 Verwertung

Es findet sowohl eine praktische als auch eine wissenschaftliche und wirtschaftliche Verwertung der Projektergebnisse statt.

Die Fahrzeugsensorik und Statistik wird durch den NVV und die INIT mindestens drei Jahre über den Projektabschluss hinaus weiter betrieben, damit das entwickelte Verfahren langfristig in die Arbeitsabläufe des NVV integriert werden kann. Die Bereitstellung von Daten für die mCloud wurde durch den NVV geprüft. Eine Bereitstellung von Rohdaten in die mCloud ist aus Gründen des Datenschutzes nicht möglich. Es können jedoch aggregierte Daten bereitgestellt werden. Die INIT befindet sich derzeit im Austausch mit dem BMDV, um die benötigte Datenstruktur und die technischen Voraussetzungen für die Veröffentlichung der Daten in der mCloud abzustimmen.

Das Projekt und das entwickelte Verfahren sind die Themen der Dissertation von Dominik Bieland am Fachbereich Verkehrsplanung und Verkehrssysteme der Universität Kas-

sel. Diese dient einer über das Projekt hinausgehenden wissenschaftlichen Betrachtung der vorhandenen Fragestellungen und wird nach erfolgter Promotion veröffentlicht.

Die Projektpartner möchten das Verfahren auch über das Projekt hinaus weiterentwickeln und für den Praxisbetrieb optimieren. Unter anderem besteht die Idee der Hinzunahme weiterer Datenquellen. Zudem könnte die Robustheit des Verfahrens noch erhöht sowie die Aussagekraft der Ergebnisse weiter gesteigert werden.

Durch die während des Projektes erfolgte E1-Zulassung der Fahrzeug-Erfassungssensoren im INIT Copilot PC, besteht die Möglichkeit der Hinzunahme weiterer Praxispartner zur weiteren Erprobung des Verfahrens in anderen Verkehrsräumen.

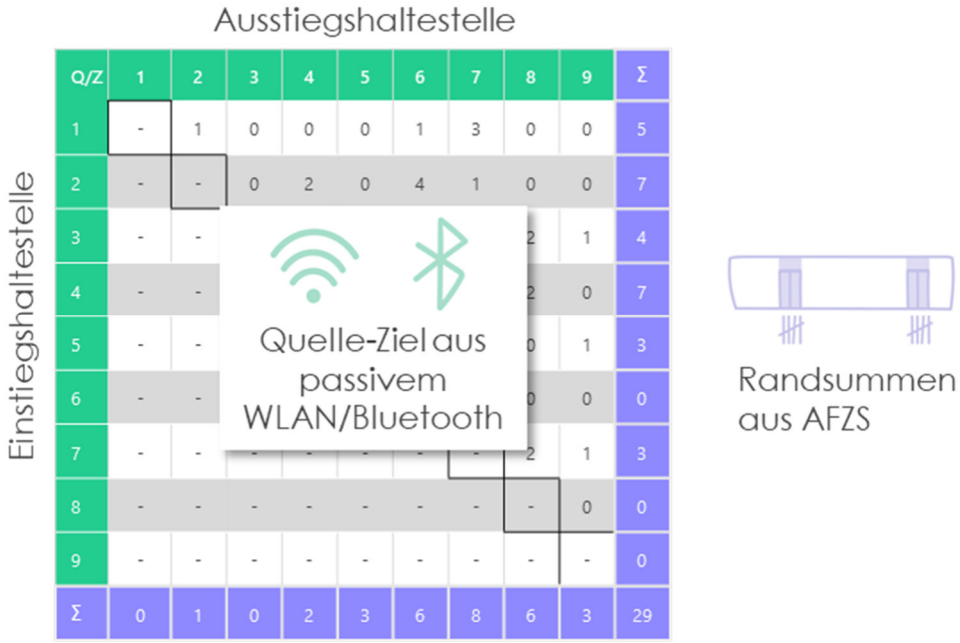
Die Projektpartner INIT und WVI haben bereits mit der Vermarktung und Akquise unter anderem im Rahmen eines gemeinsamen Webinars begonnen, welches eine sehr hohe Resonanz aufweisen konnte. Dies zeigt das hohe Interesse an den im Projekt betrachteten Fragestellungen von Seiten der Verkehrsunternehmen.

Der Projektpartner BLIC wird das entwickelte Datenschutzkonzept sowie die erworbenen Kompetenzen im Bereich Datenschutz voraussichtlich wirtschaftlich verwerten können, da diese für die meisten datengestützten Projekte und Anwendungen notwendig sein werden. Auch der NVV nutzt bereits die gesammelten Erfahrungen im datenschutzrechtlich konformen Umgang mit den Daten während dieses Projektes für den Umgang mit vergleichbaren Daten in weiteren Projekten, wie zum Beispiel dem Forschungsprojekte U-hoch-drei, im Forschungsprogramm zu Mensch-Technik-Interaktion: „Technik zum Menschen bringen“ des Bundesministeriums für Bildung und Forschung (BMBF).

3 Ausblick

Im Forschungsprojekt Mobile Data Fusion ist während der Bearbeitungszeit von gut drei Jahren die Entwicklung eines Verfahrens gelungen, das aktuelle Informationen zur Fahrgastnachfrage automatisiert bereitstellen. Das Verfahren fusioniert die mittels geeigneter Sensorik erfassten WLAN-Probe-Requests und Bluetooth Inquiry Responses mit den AFZ-Daten. Als Ergebnis des Verfahrens liegen für jede erfasste Linienfahrt tagesaktuelle, haltestellenscharfe Quelle-Ziel-Matrizen vor.

Exemplarisch ist eine Quelle-Ziel-Matrix schematisch in Abbildung 19 dargestellt. Die obere Hälfte der Matrizen bildet ab, wie viele Fahrgäste von Haltestelle A nach Haltestelle B im Fahrzeug mitgefahren sind. Die WLAN- und Bluetooth-Daten fungieren hierfür als Input-Daten. Diese verzerrten und unvollständigen Informationen werden mit den AFZ-Daten verknüpft und auf diese hochgerechnet. In der Darstellung fließen die Daten aus dem AFZS als Randsummen der Quelle-Ziel-Matrix ein.



Q/Z	1	2	3	4	5	6	7	8	9	Σ
1	-	1	0	0	0	1	3	0	0	5
2	-	-	0	2	0	4	1	0	0	7
3	-	-	-	-	-	-	2	1		4
4	-	-	-	-	-	-	2	0		7
5	-	-	-	-	-	-	0	1		3
6	-	-	-	-	-	-	0	0		0
7	-	-	-	-	-	-	2	1		3
8	-	-	-	-	-	-	-	0		0
9	-	-	-	-	-	-	-	-		0
Σ	0	1	0	2	3	6	8	6	3	29

Abbildung 19: Quelle-Ziel Matrizen (schematisch)

Das Verfahren wurde mit Hilfe von Referenzdaten aus Fahrgastbefragungen kalibriert und erfolgreich in das Produktivsystem der INIT implementiert und im Pilotbetrieb in zwei Phasen auf acht Linien getestet. Für die Hardwarelösung in den Fahrzeugen erhielt die INIT die erforderliche E1-Zulassung und ist damit aktuell die erste Herstellerin mit einer zertifizierten Lösung. Der Testbetrieb läuft nach Projektende weiter. Dazu wurden die Ergebnisse für den NVV in das Statistiktool MOBILEstatistics integriert.

Das implementierte und getestete Verfahren ist datenschutzkonform. Bei seiner Entwicklung wurde nach dem Ansatz „privacy by design“ gearbeitet und sämtliche Bearbeitungsschritte datenschutzrechtlich begleitet. Im Datenschutzkonzept wurden Lösungen für den Forschungs- und den Realbetrieb erarbeitet.

Der Pilottest bestätigte, dass das Verfahren grundsätzlich funktioniert und die entwickelten Hintergrundsysteme stabil laufen. In Hinblick auf den zukünftigen Praxiseinsatz sollte jedoch an der Robustheit des Systems gearbeitet werden. Während für die Erfassung in den Fahrzeugen eine praxistaugliche Lösung gefunden wurde, erwies sich die verbaute Technik an den Haltestellen als ungeeignet. Der Pilottest zeigte auch, dass die Robustheit des Gesamtsystems davon abhängt, dass die einzelnen Teilsysteme robust laufen. Vor der Fusion weiterer Daten sollte deshalb die Robustheit aller Systeme in Hinblick auf den geplanten Verwendungszweck geprüft werden.

Um das Verfahren in der Breite ausrollen zu können, bedarf es zudem der methodischen Weiterentwicklung. Während die Fahrzeugflotte des NVV eine Vollausstattung mit AFZS hat, ist dies bei den meisten Verkehrsunternehmen und Verbünden nicht der Fall. Das Verfahren muss an diese reale Gegebenheit angepasst werden. Das gilt ggf. auch für den Umgang mit der zunehmenden Randomisierung von MAC-Adressen. Bisher wird die MAC-Adresse als eindeutiges Merkmal zur Erfassung der Geräte verwendet. Zukünftig muss die Eindeutigkeit möglicherweise in weiteren Bearbeitungsschritten vor der Datenfusion hergestellt werden. In das vorliegende Verfahren können weitere Datenquellen eingebunden werden, welche Informationen zu Quelle-Ziel-Verflechtungen der ÖV-Fahrgäste bereitstellen.

Gegenüber den bisher etablierten, stichtagsbezogenen Befragungen besitzt das in Mobile Data Fusion entwickelte Verfahren den Vorteil, dass Nachfragedaten automatisiert und kontinuierlich erfasst werden. Diese Daten können in Zukunft genutzt werden, um zeitlich und/oder räumlich aggregierte Matrizen zu erzeugen und dadurch die Güte des Verfahrens zu verbessern. Perspektivisch könnte sich auch die Fusion weiterer automatisiert erfasster Verkehrsnachfragedaten positiv auf die Güte auswirken.

Neben der Güte entscheiden auch die konkreten Anwendungsfälle und die praktischen Vorteile der automatisierten Datenbereitstellung darüber, ob sich der Einsatz des Verfahrens in der Praxis bewährt. Dazu wird der NVV in den kommenden Jahren weitere Erfahrungen sammeln. Der Weiterbetrieb des Systems ist bis mindestens Ende 2024 geplant. Die Projektpartner haben darüber hinaus ein großes Interesse, im Rahmen von nachfolgenden Projekten an dem Ansatz weiterzuarbeiten.

4 Literatur

Baeta N., Fernandes A., Ferreira J. (2016): Tracking Users Mobility at Public Transportation. In: Bajo J. et al. (Hrsg.): Highlights of Practical Applications of Scalable Multi-Agent Systems. PAAMS 2016. Communications in Computer and Information Science, vol 616. Springer, Cham.

Bamberger D., Beige S., Schneider M., Schubert M. (2017): Anwendungsfelder anonymisierter Mobilfunkdaten im ÖPNV - Fünf regionale Beispiele für Analyse und Vorhersage von Verkehrsbedarfen auf Quelle-Ziel-Relationen. In: Der Nahverkehr, 35, S. 12 – 17.

Bieland D., Ableitung von Quelle-Ziel-Matrizen auf Basis von WLAN- und Bluetooth-Daten am Beispiel ausgewählter Buslinien im NVV. Unveröffentlichte Dissertation.

Bieland, D., Briegel, R. (2021): Internes Arbeitspapier zur Fahrgastbefragung MobiDat Phase 1 (unveröffentlicht). Kassel.

Demchenko, Y., De Laat, C., & Membrey, P. (2014, May): Defining architecture components of the Big Data Ecosystem. In: 2014 International conference on collaboration technologies and systems (CTS), Pages 104-112. IEEE.

Dietrich A.-M., Henninger J., Sauer J., Gründel T., Wagner H. (2018): Tariftool-XL Softwaresystem zur Entwicklung flexibler Tarife im öffentlichen Personenverkehr, Gemeinsamer Schlussbericht, November 2018.

Dunlap M., Li Z., Henrickson K., Wang Y. (2016): Estimation of Origin Destinatin Nformation form Bluetooth and Wi-Fi Sensing for Transit. In: Jorunal oft he Transportation Research Board, Volume 2595, Issue 1, 2016, Pages 11-17.

Elektronik-Kompendium.de (2018): Ortung und Positionsbestimmung mit Mobilfunk. https://www.elektronik-kompendium.de/sites/kom/1201061.htm.

flinc GmbH (2016): Studie zeigt: Shuttle-System sorgt für 97% weniger Autos in Hamburg. http://presse.flinc.org/2016/11/17/studie-zeigt-shuttle-system-sorgt-fuer-97-weniger-autos-in-hamburg/.

iRights.Lab GmbH (2020): Self-Data-Governance Framework, V1.0, Mai 2020, abgerufen von https://www.irights-lab.de/projekt/data-governance am 31.08.2021.

Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder (2020): Das Standard-Datenschutzmodell: Eine Methode zur Datenschutzberatung und -prüfung auf der Basis einheitlicher Gewährleistungsziele, Version 2.0b, 2020.

Mishalani A., Mishalani R. G., Sengupta R., Walker J. L. (2016): In Pursuit of the Happy Transit Rider: Dissecting Satisfaction Using Daily Surveys and Tracking Data. In: Journal of Intelligent Transportation Systems, Volume 20, Issue 4, 2016, Pages 345-362.

O'Malley J. (2017): London Underground Wifi Tracking: Here's Everything We Learned From TfL's Official Report. http://www.gizmodo.co.uk/2017/09/london-underground-wifi-tracking-heres-everything-we-learned-from-tfls-official-report/.

Pattanusorn W., Nilkhamhang I., Kittipiyakul S., Ekkachai K., Takahashi A. (2016): Passenger Estimation System Using Wi-Fi Probe Request. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7467124&tag=1.

ProTrain (2020), Verbund-Forschungsvorhaben ProTrain: ProTrain Gesamtbericht, https://www.tib.eu/de/suchen?tx_tibsearch_search%5Baction%5D=download&tx_tibsearch_search%5Bcontroller%5D=Download&tx_tibsearch_search%5Bdocid%5D=TIBKAT%3A1797328468&cHash=570604792017f5cb0669508dbed7243f#download-mark.

Sapiezynski P., Stopczynski A., Gatej R., Lehmann S. (2015): Tracking Human Mobility using WiFi signals. https://arxiv.org/abs/1505.06311.

Schelewsky M. (2014): Tracking mit Smartphones: Einführung in die Technik und Stand der Forschung. In: Schelewsky, M., H. Jonuschat, B. Bock, K. Stephan (Hrsg.)(2014): Smartphones unterstützen die Mobilitätsforschung – Neue Einblicke in das Mobilitätsverhalten durch Wege-Tracking. Springer Fachmedien, Wiesbaden, S. 5 – 24.

Schmidt A., Männel T. (2017): Potenzialanalyse zur Mobilfunkdatennutzung in der Verkehrsplanung. Abrufbar unter: https://www.iao.fraunhofer.de/lang-de/images/iao-news/telefonica-studie.pdf am 30.05.2018.

Sommer C. (2002): „Erfassung des Verkehrsverhaltens mittels Mobilfunktechnik: Konzept, Validität und Akzeptanz eines neuen Erhebungsverfahrens“. In: TU Braunschweig, Institut für Verkehr und Stadtbauwesen, Schriftenreihe des Instituts, Band 51, Shaker Verlag, Aachen, 2002.

Song B., Wynter L., (2017): Real-time public transport service-level monitoring using passive WiFi: a spectral clustering approach for train timetable estimation. https://arxiv.org/pdf/1703.00759.pdf.

Telefónica Germany Next GmbH (2018): Transport Analytics. https://next.telefonica.de/loesungen/transport-analytics.

Trialog Publishers Verlagsgesellschaft (2017): Mobilfunkdaten mit gutem Potenzial für Verkehrsplanung. https://www.internationales-verkehrswesen.de/mobilfunkdaten-zur-verkehrsplanung/.

VDV-Schrift 4 „Verkehrserschließung, Verkehrsangebot und Netzqualität im ÖPNV“, Ausgabe 01/2019.