# Mobilitätsforschung zur Reichweitenbestimmung in der Deutschen und Schweizer Außenwerbung – Neue Wege mit GPS

**Inaugural-Dissertation**

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Dirk Ronny Hecker

aus Köln

Köln, 2013

Berichterstatter: Prof. Dr. Georg Bareth

Prof. Dr. apl. Klaus Zehner

Tag der mündlichen Prüfung: 19.10.2012

## DANKSAGUNG

Es gibt einige Personen denen ich zu danken habe, denn ohne Sie hätte diese Dissertation nicht geschrieben werden können. In erster Linie möchte ich Dr. Michael May danken, der mir den Impuls für diese Arbeit gegeben hat und mich durch zahlreiche Anregungen und viele lange Gespräche auch über die Inhalte der Arbeit hinaus betreut und unterstützt hat. Für seine wissenschaftliche Förderung danke ich ihm sehr!

Weiterhin möchte ich mich bei meinen Kollegen Robert Spindler, Hermann Streich, Renate Henkeler, Dr. Hans Voss und Hendrik Stange aus der Abteilung der Knowledge Discovery bedanken, die mir immer mit Rat und Tat, auch in schwierigen Stunden zur Seite standen.

In besondere Weise möchte ich mich bei Dr. Christine Körner und Dr. Angi Voss bedanken, die mir für jede noch so schwierige methodische Diskussion ein offenes Ohr geschenkt haben und mir durch ihre konstruktive Kritik und ihre Verbesserungsvorschläge sehr geholfen haben.

Herzlich bedanken möchte ich mich bei dem Betreuer dieser Dissertation Herrn Prof. Dr. Georg Bareth. Auch als externer Doktorand wurde ich von ihm hervorragend wissenschaftlich gefördert, unterstützt und wurde in seiner Arbeitsgruppe sehr freundlich aufgenommen. Ich danke Herrn Prof. Dr. Klaus Zehner für sein großes Interesse an diesem sehr anwendungsorientierten Thema, die Unterstützung und die Erstellung des Zweitgutachtens.

Mein ganz besonderer Dank gilt meinen Eltern und meinen Freunden Daniel Strücker, Manoj Cherumadathil, Christian Wilmsen, Andrea Rönsberg und Freddy Baum, die mich in dieser Zeit stets unterstützt und ermutigt haben. Meiner Lebensgefährtin Heike Wilz danke ich für ihre Liebe und das Verständnis während der Zeit der Dissertation.

## KURZZUSAMMENFASSUNG

Die bestmögliche Positionierung einer Plakatstelle ist seit jeher ein wichtiger Bestandteil in der Außenwerbung. Aus plausiblen Gründen werden Plakatstandorte so gewählt, dass sie häufig von Menschen passiert werden. Während jedoch die Lokalität und das Umfeld bei der Wahl von Plakatstandorten schon immer ein entscheidendes Kriterium gewesen ist, spielen die räumlichen Zusammenhänge bei der Leistungsbewertung und damit bei der Preisbildung von Plakatkampagnen erst in den vergangenen Jahren eine wichtige Rolle. Ziel der Leistungsbewertung ist es festzustellen, wer, wie oft und woher eine Person an einer Plakatkampagne vorbei gekommen ist. Dabei ist die räumliche Verteilung von Plakatstellen von entscheidender Bedeutung. Neue, GPS-basierte Messmethoden der Mobilitätserfassung erlauben die räumlich differenzierte Ausweisung von Leistungswerten für beliebig zusammengestellte Kampagnen sowie soziodemographische und räumlich ausgewählte Zielgruppen. Damit unterscheidet sich dieser GPS-Ansatz von den klassisch eingesetzten Methoden, die bisher versucht haben, über Befragungen von Testpersonen die Mobilität zu rekonstruieren, und nur Durchschnittswerte für Kampagnen anbieten konnten. So wurde z.B. für eine Stadt wie Köln nicht unterschieden, ob es sich bei einer Plakatkampagne um eine stark über die Stadt gestreute oder stark konzentrierte Kampagne handelt. Diese Dissertation widmet sich einer Gesamtschau über neuartige GPS-Verfahren in der Deutschen und Schweizer Außenwerbung und deren Anwendung in der Praxis. Sie ist ein erstmaliger Versuch, eine systematisierte Übersicht über die aktuellen Forschungsergebnisse in der Außenwerbung bzw. Mobilitätsforschung zu erstellen. Die bestehenden Publikationen zur Außenwerbeforschung (Pasquier 1997, Engel und Hofsäss 2003) stammen noch aus der Zeit vor der Mobilitätserfassung mit GPS sowie dem Einsatz von Geographischen Informationssystemen und sind somit veraltet. Zudem betrachten diese Publikationen keine geographischen Aspekte, die mit dem neuen Ansatz in den Fokus der Leistungsbewertung rücken und eine Erfolgsgeschichte für die Geographie bzw. das Geomarketing darstellen.

Zur Systematisierung dieser Arbeit zählt, geeignete Lösungswege im Umgang mit zeitlichen und räumlichen Datenlücken zu diskutieren, zu erproben, sowie Validitäts- und Robustheitsanalysen durchzuführen. Es werden geeignete Tests definiert, die z.B.

Selektionseffekte in der Rekrutierung von Probanden offen legen. Es wird die Problematik behandelt, wie Leistungswerte auf eine Neuerhebung der Mobilität reagieren und wo die Grenzen einer Leistungswertbestimmung liegen. Dabei sind die Ausgangslage und die Anforderungen in der Schweiz und in Deutschland unterschiedlich. So ist in Deutschland die GPS Stichprobengröße viel geringer als in der Schweiz. Dies hat direkte Konsequenzen auf die Modellierungsschritte und das Ergebnis. Die verwendeten Methoden sind weit über das Gebiet der Außenwerbung und Mediaplanung hinaus für die Modellierung von Mobilität von Interesse.

## ABSTRACT

The optimal positioning of a poster has all along been an important element of outdoor advertising. Clearly, poster locations are selected the way that people pass them frequently. However, while the location and surroundings of a poster have always been an important criterion for choosing poster sites, only recently spatial relationships have begun to play a role for the performance evaluation and thus the pricing of poster campaigns. The objective of such a performance evaluation is to determine how often a person of which socio-demography and from which city of residence passes a poster campaign. Key component thereby is the spatial distribution of the poster locations. Novel GPS-based measurement methods for mobility tracking allow the spatially differentiated performance indexing for arbitrary arranged campaigns as well as for socio-demographically and spatially selected target groups. Hence, the GPS-based approach differs from traditional methods, which reconstruct mobility patterns by interviewing test persons and therefore solely report mean values for specific campaigns. For example, in the city of cologne the performance of wide spread poster campaigns could not be distinguished from the performance of regionally rather dense campaigns. This dissertation gives an overview of novel GPS-based methods used in the German and Swiss outdoor advertising as well as their application in practice. It furthermore acts as an unprecedented attempt to create a systematic overview of recent findings in outdoor advertising and mobility research. Existing papers in the research field of outdoor advertising (Pasquier 1997, Engel and Hofsäss 2003) date back to the period before mobility detection with GPS-technology and the usage of geographic information systems, and are hence outdated. Furthermore, these publications do not consider geographic aspects, which move into the focus of performance evaluation in this new GPS-based approach and simultaneously depict a story of success for geography and geo-marketing.

For a holistic approach this thesis discusses and tests appropriate solutions for dealing with temporal and spatial gaps in mobility data as well as for performing validity- and robustness analyses. Suitable tests are defined, which reveal, for instance, selection effects in recruiting test persons. Furthermore difficulties are analyzed, regarding the response of performance indices to new mobility surveys and the limitations. This thesis focuses on two geographic regions, namely Switzerland and Germany, for which both the initial situation and modeling

demands vary. For example, the GPS sample for Germany is relative to the population of smaller size compared to that of Switzerland and therefore causes implications in the process of modeling and the overall findings. Nevertheless, the applied methods are of utmost interest for modeling mobility beyond the field of outdoor advertising.

# INHALTSVERZEICHNIS

# VERZEICHNISSE DER ABBILDUNGEN UND TABELLEN

## ABBILDUNGEN

## TABELLEN

## ABKÜRZUNGEN

| | |
|---|---|
| ag.ma | Arbeitsgemeinschaft Media-Analyse e.V. |
| app | Applikation Mobilfunkgerät |
| CATI | Computer Assisted Telephone Interview |
| CAPI | Computer Assisted Paper Interview |
| CAWI | Computer Assisted Web Interview |
| DGPS | Differential Global Positioning System |
| FAW | Fachverband Aussenwerbung e.V. |
| GfK | Gesellschaft für Konsumforschung |
| GIS | Geographisches Informationssystem |
| GPS | Global Positioning System |
| GRP | Gross Rating Points |
| GSM | Global System for Mobile Communications |
| IAIS | Institut für Intelligente Analyse- und Informationssysteme |
| KDD | Knowledge Discovery in Databases |
| LBS | Location Based Services |
| MAR | Missing at Random |
| MCAR | Missing Completely at Random |
| MID | Mobilität in Deutschland |
| MNAR | Missing not at Random |
| OCC | Opportunity of Contact |
| ÖPNV | Öffentlicher Personennahverkehr |
| OTS | Opportunities to See |
| POI | Point of Interest |
| REMO | Relative Motion |
| RFID | Radio Frequency Identification |
| RME | Relative Mean Error |
| RMSE | Root Mean Squared Error |
| SPR+ | Swiss Poster Research Plus |
| WGS84 | World Geodetic System of 1984 |
| W-LAN | Wireless Local Area Network |

# KAPITEL 1

## 1. EINLEITUNG

### 1.1 Motivation

> "*Viele kleine Dinge wurden durch die richtige Art von Werbung groß gemacht.*"
>
> Mark Twain (1835-1910)

Zahlreiche Branchen haben großes Interesse an Mobilitätsauswertungen. Klassisch gehören seit vielen Jahren die Kommunen und Städte dazu, die insbesondere bei Bauvorhaben Prognosen des Verkehrs erstellen (Mobilität in Deutschland, MID 2008a). Navigationsanbieter nutzen Mobilitätsinformationen zur Stauerkennung auf Autobahnen und Bundesstraßen und schlagen Alternativrouten vor. Filialisten nehmen intensive Standortzählungen vor, um bei Neuplanungen von Geschäften eine bestmögliche Entscheidung zu treffen. Eine Branche, die vor einer besonderen Herausforderung steht, ist die Außenwerbung. Sie braucht fast flächendeckend und für alle Straßenkategorien Mobilitätsinformationen. Anhand von Fragestellungen dieser Branche werden in der vorliegenden Dissertation innovative Analysemethoden der Mobilitätsforschung erforscht und Lösungen vorgestellt.

Die Außenwerbung spielt eine bedeutende Rolle in der Werbewirtschaft[1]. Jedoch gibt es neben der Außenwerbung noch eine Fülle weiterer Werbeformen, zu denen eine permanente Konkurrenzsituation besteht. Zu den klassischen Konkurrenten um Werbebudgets gehören das Fernsehen, der Hörfunk und die Printmedien. Zudem stoßen auch immer wieder neue Werbeformen hinzu, wie z. B. die Internetwerbung und das Mobile Marketing. Um bei der Vergabe von Werbeaufträgen Berücksichtigung zu finden, muss die Außenwerbung transparente, möglichst exakte und differenzierte Leistungswerte zur Verfügung stellen. Hier ist neben der reinen Passantenanzahl wichtig zu wissen, aus welchen Gebieten die Passanten stammen, wie häufig diese Personen an einer Plakatkampagne vorbeikommen und zu welcher soziodemographischen Gruppe sie gehören. Dadurch kann eine Plakatkampagne für eine soziodemographisch und geographisch eingegrenzte Gruppe speziell zugeschnitten werden.

Beim Fernsehen, Hörfunk und den Printmedien existieren bereits seit Jahren Leistungswerte, die soziodemographische Gruppen aufgeschlüsselt nach Printtiteln, Fernseh- und Hörfunksendungen ausweisen. Auch bei der Internetwerbung existieren inzwischen solche Leistungswerte (ag.ma 2012e). Die Leistungswerte der Außenwerbung waren bisher hingegen nicht ausreichend belastbar und aussagekräftig. Damit die Außenwerbung auch in Zukunft ihre Anteile am Gesamtwerbeumsatz[2] halten bzw. weiter ausbauen kann, muss sie

[1] So lag z. B. in 2010 der Umsatz der deutschen Außenwerbung bei 766 Mio. € und in der Schweiz bei 608 Mio. CHF (~497 Mio. €) (FAW 2011; Stiftung Werbestatistik Schweiz 2011).
[2] Anteile am Gesamtwerbeumsatz in Deutschland ~4% und in der Schweiz ~13%

aber vergleichbare und am Markt akzeptierte Leistungswerte liefern. Hierzu hat die Außenwerbung die Leistungswertmodellierung auf eine komplett neue empirische Grundlage gestellt. In der Schweiz und in Deutschland wurden sehr große Feldstudien mit GPS-Geräten durchgeführt. Durch die neue Empirie ergibt sich ein deutlicher Zugewinn bei der Ausweisung von Leistungswerten von selektierten Plakatkampagnen und Zielgruppen. Allerdings ist der Umgang mit den neu erfassten Daten auch deutlich komplexer. In dieser Arbeit werden die Herausforderungen der neuen Modellierung vorgestellt und mit den bisherigen Ansätzen und z. T. in anderen Ländern noch verwendeten Verfahren verglichen. Darüber hinaus werden Modellierungsschritte vorgestellt, die auch jenseits des vorgestellten Anwendungsszenarios von Bedeutung sind und generell eingesetzt werden können, wenn mobile Objekte im geographischen Raum analysiert werden.

## 1.2 Forschungsfrage und Herausforderungen der Arbeit

Die zentrale Frage der Dissertation lautet: *"Wie können GPS-Daten dazu eingesetzt werden, fundierte Leistungswerte für die Außenwerbung zu bestimmen?"*

Unter Einsatz der GPS-Feldstudien setzen Deutschland und die Schweiz als erste Länder weltweit einen neuen innovativen Erfassungsstandard für die Mediaplanung in der Außenwerbung. In der bisherigen Forschungsliteratur ist dieses Thema mit Ausnahme von eigenen Arbeiten noch nicht existent. Die bestehenden Publikationen zur Außenwerbeforschung (Pasquier 1997, Engel & Hofsäss 2003) stammen noch aus der Zeit vor der Mobilitätserfassung mit GPS sowie dem Einsatz von Geographischen Informationssystemen und sind somit veraltet. Zudem betrachten diese Publikationen keine geographischen Aspekte, die mit dem neuen GPS-Ansatz in den Fokus der Leistungsbewertung rücken. Die vorliegende Arbeit fokussiert diesen neuen Ansatz und legt den Schwerpunkt auf die Analyse, Modellierung und Anwendung von GPS-Daten. Alternative Mobilitätsdatenquellen wie Telefoninterviews, Tagebuchaufzeichnungen sowie Mobilfunkdaten werden in Abschnitt 2.2 vorgestellt, fließen jedoch nicht in die Auswertungen mit ein. Beim Einsatz von GPS-Daten bestehen folgende Herausforderungen:

**Stichprobenverzerrungen in GPS-Daten**

Die erfassten GPS-Daten umfassen neben den aufgezeichneten Trajektorien auch umfangreiche soziodemographische Daten. Eine Herausforderung liegt darin, Verzerrungen in den Daten zu erkennen. Das Augenmerk liegt dabei auf der Untersuchung von systematischen Zusammenhängen zwischen fehlenden Daten, soziodemographischen Variablen und der Mobilität von Probanden, die zu Falschaussagen führen können. Aus diesem Grund muss überprüft werden, ob „zufällige" oder „systematische" Zusammenhänge in den Daten bestehen.

**Fehlende Messdaten**

Wenn man Mobilitätsstudien durchführt, kann es leicht dazu kommen, dass an einzelnen Tagen keine Mobilität erhoben wird. In Mobilitätsstudien mit GPS gibt es eine Reihe von individuellen oder technischen Gründen, die zum Fehlen von Daten führen können. Dabei können wir zwischen drei Varianten innerhalb eines Tages unterscheiden:

1. Fehlen einer Teilstrecke eines Weges
2. Vollständiges Fehlen eines Weges
3. Fehlen eines Tages

Die Unvollständigkeit bei Teilstrecken bezieht sich auf kurze Erfassungslücken innerhalb einer Sequenz von GPS-Signalen. Mehrere Gründe können zur Entstehung solcher Datenlücken führen, z.B. Störungen des GPS-Signals durch Gebäude, Bäume oder der Signalverlust innerhalb eines Tunnels oder bei der Startphase eines GPS-Gerätes. Das vollständige Fehlen eines Weges bezieht sich auf das vollständige Fehlen einer oder mehrerer Fahrten an einem Tag. Das einmalige Fehlen eines Weges ergibt sich z.B., wenn ein Proband vergisst, das GPS-Gerät für einen kurzen Einkaufsweg mitzunehmen. Das Fehlen eines Tages kann z.B. durch leere Gerätebatterien oder technische Mängel des Gerätes entstehen. Ignoriert man an dieser Stelle die fehlenden Messdaten, unterschätzt man die Mobilität und damit die Leistungswerte von Plakatstellen. Daher ist ein zentraler Punkt dieser Arbeit, je nach Art der zeitlichen Unvollständigkeit unterschiedliche Methoden im Umgang mit GPS-Daten zu entwickeln.

**Räumliche Abdeckung**

Bei der GPS-Erfassung in der Schweiz und in Deutschland handelt es sich um eine der europaweit größten GPS-Stichproben, doch ist die Stichprobe vergleichsweise gering in Bezug auf die notwendige räumliche Abdeckung. So liegen für eine Stadt wie Köln knapp 344 GPS-Probandeninformationen vor, deren Pfade decken jedoch nicht alle Straßensegmente der Stadt ab. Würde man die unvollständige räumliche Abdeckung der GPS-Daten ignorieren, so könnten eine Vielzahl von Plakatstellen nicht bewertet werden, da sie keine GPS-Passage aufweisen.

**Robustheit der Ergebnisse**

Mobilitätsinformationen werden in Zukunft eine der zentralen Datenquellen für die Außenwerbung sein. Das heißt aber auch, dass die Außenwerbung in Zukunft ältere Mobilitätsinformationen mit jüngeren Informationen erneuern muss, um die Leistungswerte auf dem aktuellen Stand zu halten. In diesem Zusammenhang ist wichtig zu klären, wie stabil die erzielten Ergebnisse hinsichtlich einer Neuerhebung der Mobilitätsinformationen und einer evtl. kleineren Stichprobe sind. Denn die Erhebung ist sehr kostenintensiv und es entsteht die Frage, welche Auswirkungen eine kleinere Stichprobe auf die Ergebnisse hat.

**Verwandte Arbeiten**

In der Forschungsliteratur hat man sich bisher diesen beschriebenen Herausforderungen nicht gewidmet. Existierende Literatur aus dem Bereich der Trajektorienanalyse beschäftigen sich z.B. mit dem Clustering von Trajektorien (Rinzivillo et al. 2008, Nanni und Pedreschi 2006, Pelekis et al. 2007), mit dem Erkennen von relativen Bewegungsmustern (Hwang et al. 2005, Laube und Imfeld 2002, Gudmundsson et al. 2007) oder mit der sequentiellen Analyse von Bewegungen (Giannotti et al. 2007, Yang and Hu 2006, Zheng et al. 2009). Ein weiterer Schwerpunkt bei bisherigen Mobilitätsuntersuchungen liegt auf der Analyse von Bewegungen auf regionaler und nationaler Ebene. Sie werden in der Regel nach einer bestimmten Zeit zur Untersuchung der Änderungen im städtischen und nationalen Mobilitätsverhalten wiederholt. Ein Beispiel ist die Studie „Mobilität in Deutschland“ (MID 2008b) und der „Mikrozensus“ in der Schweiz (Mikrozensus 2010), die auf einem Computer Assisted Telephone Interview (CATI) und Aufzeichnungen der persönlichen Mobilität eines einzelnen Tages basieren. Die Studien werten Variablen wie durchschnittlich gefahrene Kilometer pro Tag, das Pendlerverhalten und das genutzte Verkehrsmittel aus. Die Analysen basieren auf groben räumlichen Ebenen (Bundesländer, Kreisebene) und setzen auf vergleichsweise stabile Mobilitätsaussagen. Im Vergleich dazu bedarf es bei der Anwendung zur Bewertung von individuellen Plakatstellen/Plakatkampagnen einer sehr lokalen Mobilitätsmodellierung. Aggregierte Aussagen, z.B. auf Gemeindeebene, sind im Anwendungskontext nicht feinräumig genug.

## 1.3 Beitrag der Dissertation

Der Beitrag der Dissertation besteht in der erstmaligen wissenschaftlichen Systematisierung, Diskussion und vergleichenden Bewertung der Leistungswert-Modellierung in der Außenwerbung auf Basis von GPS-Daten und unter geographischen Gesichtspunkten. Dabei sind die Ausgangslage und die Anforderungen in der Schweiz und in Deutschland unterschiedlich. So ist z.B. in Deutschland die GPS-Stichprobengröße viel geringer als in der Schweiz. Dies hat direkte Konsequenzen für die Modellierungsschritte und die Ergebnisse. Aufeinander aufbauend werden die methodischen Schritte und die möglichen Lösungswege für eine Leistungsbewertung in Deutschland und der Schweiz auf ihre Robustheit, Validität und Einsatzgrenzen analysiert.

Unter anderem werden hierfür geeignete Lösungswege im Umgang mit zeitlichen und räumlichen Datenlücken diskutiert und erprobt sowie Validitäts- und Robustheitsanalysen durchgeführt. Darüber hinaus werden geeignete Tests definiert, die z.B. Selektionseffekte in der Rekrutierung von Probanden offenlegen. Es wird die Problematik behandelt, wie Leistungswerte auf eine Neuerhebung der Mobilität reagieren und wo die Grenzen einer Leistungswertbestimmung liegen. Zum Abschluss werden die Vorteile von räumlich differenzierten Plakatkampagnen vorgestellt.

Eine Grenze zieht diese Dissertation bei der Wahrnehmbarkeit von Werbung. Dieser sogenannte qualitative Bereich in der Außenwerbung, der sich mit der Wahrnehmungs- und Erinnerungsleistung von Werbebotschaften und Plakaten beschäftigt, ist nicht Teil dieser Arbeit. Es sei jedoch darauf hingewiesen, dass dies für die Leistungswertbestimmung in allen Mediengattungen eine wichtige Determinante ist und entscheidenden Einfluss auf die Ausweisung von Leistungswerten hat (vgl. Abschnitt 2.1).

Die Arbeit ist stark interdisziplinär und verwendet z.B. Techniken aus der Geoinformatik, der Statistik und des Data Mining. Die verwendeten Methoden sind weit über das Gebiet der Außenwerbung und Mediaplanung hinaus für die Modellierung von Mobilität auf feinräumiger Ebene von Interesse.

## 1.4 Einordnung der Arbeit in die Geographie, Time Geography und Geoinformatik

Das Abbilden des menschlichen Handelns in Zeit und Raum in Form von „Zeitpfaden“ ist bereits seit den 50er Jahren ein wichtiger Bestandteil der Geographie. Als Begründer der klassischen „Time Geography“ gilt Hägerstrand von der Universität Lund in Schweden, der sich mit der Auswertung von Daten beschäftigte, die über die Bewegung von Individuen gesammelt wurden. Aufgrund der Tatsache, dass Menschen inzwischen mittels GPS zu jeder Tageszeit verortet werden können, lassen sich sowohl räumliche als auch zeitliche Verläufe über Tage, Wochen und Monate speichern. In den Abbildungen von Hägerstrand werden Raum und Zeit in drei Dimensionen dargestellt, wobei auf die Raumdimension der Höhe verzichtet und diese durch die Zeit ersetzt wurde. In den folgenden Jahren wurde der zeitgeographische Ansatz von Hägerstrand und seinen Mitarbeitern (Ellegard, Martensson, Tornquist, u. a.) weiterentwickelt und später als „Lund Schule“ bezeichnet. Parkes und Thrift (1980) nannten die Darstellungsart der Lund Schule „dynamic maps“. Vor dem Hintergrund neuer Technologien zur Erfassung von Mobilität (GPS-, RFID, Mobilfunk) ist die Time-Geography ein spannendes Forschungsfeld. Sie verknüpft Raum und Zeit, exploriert und entdeckt Muster in den Daten. Sie nutzt die semantische Annotation von Trajektorien in vielfältiger Weise und macht die Daten nutzbar für verschiedenste Anwendungen.

Eine dieser Anwendungen findet sich in der Außenwerbung, denn ihr Leistungswert, die Reichweite, befasst sich mit der räumlichen Verteilung von Plakatstellen und der Zeit, wann Passanten an ihnen vorbeikommen. Es soll geklärt werden, wann (zu welchem Zeitpunkt), wie (mit welchem Verkehrsmittel), wo (an welchem/n Plakat/en) und wie häufig eine Person an einer bestimmten Plakatkampagne vorbeigekommen ist. Hierfür kann die klassische Darstellung der Zeitpfade nach Hägerstrand genutzt werden (Abbildung 1.1). Als Säule wird jeweils das Zeitintervall dargestellt, in dem sich die Person für einen längeren Zeitraum an einem Ort aufhält (Arbeit, Wohnung, Einkauf). Die Pfade stellen die Mobilität zwischen den einzelnen Aufenthaltsorten dar. Die Z-Achse gibt hierzu die Information über die zeitliche Dauer der Aktivitäten. Die X-Achse und Y-Achse spannen den geographischen Raum auf.

Die Grafik A zeigt einen fiktiven idealtypischen Zeitpfad eines Tages. Eine Person verlässt am Vormittag ihre Wohnung und fährt zur Arbeit. Auf diesem Weg passiert diese Person zwei Plakatstellen. Nach der Arbeit nimmt sie zwei Freizeitaktivitäten an unterschiedlichen Orten wahr und passiert wieder eine Plakatstelle. Bei Grafik B ist der gleiche Weg visualisiert, jedoch fehlen einzelne Teilstrecken durch Aufzeichnungsprobleme zwischen den jeweiligen Aufenthaltsorten. Zwei von drei Plakatpassagen konnten aufgrund der Lücken nicht aufgezeichnet werden. In Grafik C fehlt der komplette Tag, da entweder ein technisches Problem vorlag oder die Person das GPS-Gerät an diesem Tag vergessen hat. Drei von drei tatsächlich stattgefundenen Plakatpassagen konnten nicht aufgezeichnet werden. In Grafik D sind nun die vollständigen Wege eines Probanden wieder dargestellt. Dieser Proband steht in seinem Quartier für eine Menge von Personen mit gleicher Soziodemographie und gleicher Altersklasse. Folgt diese Person nur ihren eigenen typischen Wegen, kann sie nicht die räumliche Bewegung aller Personen repräsentieren. Rot abgebildete Sterne bedeuten an dieser Stelle eine aufgezeichnete Plakatpassage, grün dargestellt sind Plakatstellen, die keine Passage erzielten. Das bedeutet, dass von sieben Plakatstellen vier im Erfassungszeitraum keine Passage erzielten. Daraus folgt, dass für insgesamt vier Plakatstellen kein Leistungswert ausgewiesen werden kann. Wie reagiert man nun auf Zeitpfade, die unterbrochen werden und eine zu geringe räumliche Abdeckung aufweisen? Diese Frage lässt die Time Geography unbeantwortet. Sie wird in der vorliegenden Dissertation beantwortet.

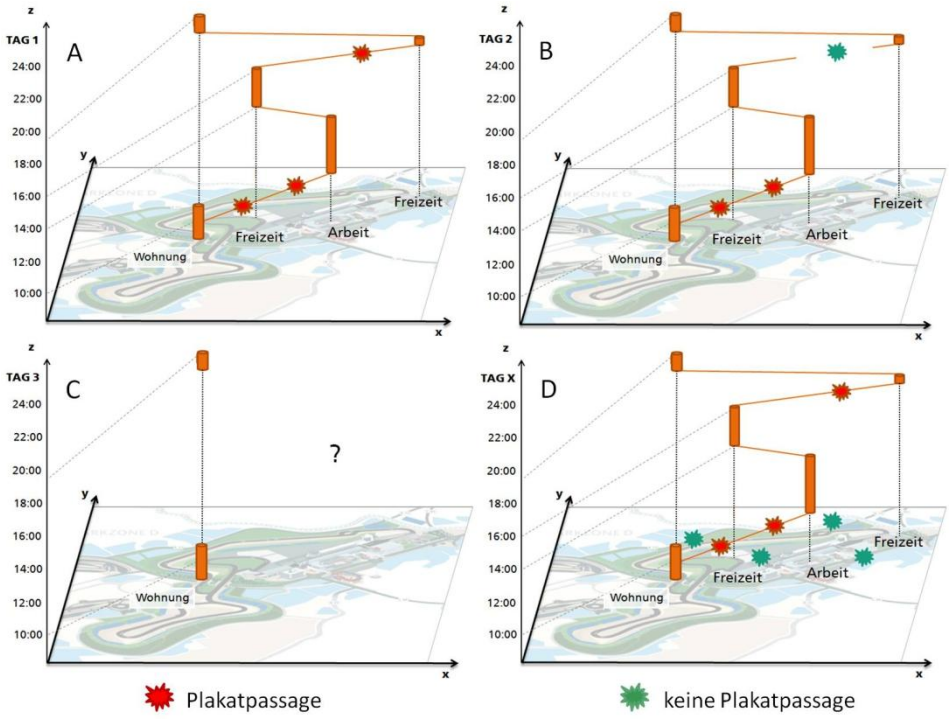



Abbildung 1.1: Zeitpfade im Anwendungskontext

Auf den Anwendungskontext Außenwerbung bezogen sind insbesondere die Wege zwischen den Aufenthaltsorten interessant, denn während eines Weges besteht potenziell die Chance, dass eine Person eine Plakatpassage erzeugt. Diese muss erkannt, aufgezeichnet und in später beschriebenen Verarbeitungsschritten aufbereitet werden. Hägerstrand liefert an dieser Stelle den theoretischen Hintergrund für die Geographie, in dem er Raum und Zeit miteinander in Verbindung bringt, denn für den Anwendungskontext ist der betreffende Zeitpunkt des Eintretens eines Ereignisses (Plakatpassage) einer der wichtigen Faktoren. Daten über Umfelder (z.B. Plakatstandorte), in denen Mobilität stattgefunden hat, zu analysieren und in die Anwendung zu bringen ist Aufgabe der anwendungsorientierten geographischen Wissenschaften. Diese Dissertation bewegt sich mit ihren Forschungsfragen in der Schnittmenge zwischen der Geographie und der Geoinformatik. Dabei liefert die Geoinformatik Methoden bei der Bearbeitung raumbezogener Informationen. Insbesondere die Analyse und Auswertung der enormen Datenmengen, die bei der GPS-Erfassung in der Schweiz und in Deutschland anfallen, greifen auf Methoden der Datenbankspeicherung, der Datenschnittstellen (Verknüpfung von POI-, Polygon- und Segmentinformationen) und der Datenbankanalyse zurück.

## 1.5 Aufbau und Struktur der Arbeit

Die Arbeit ist wie folgt strukturiert: Zu Beginn wird der Anwendungskontext der Arbeit vorgestellt, und es werden die wichtigsten Grundlagen und Anforderungen aus der Außenwerbeforschung beschrieben. Im Anschluss werden die in dieser Arbeit verwendeten Daten und Methoden vorgestellt. Im nächsten Schritt werden die z.T. sehr aufwändigen und speziellen GPS-Aufbereitungsschritte im Hinblick auf den Anwendungskontext erläutert. Es folgt die eigentliche Berechnung der Leistungswerte für die Außenwerbung. Hier werden aufgrund der unterschiedlichen Stichprobenstrukturen in der Schweiz und in Deutschland unterschiedliche Vorgehensweisen vorgestellt und auf ihre Robustheit untersucht. Die Ergebnisse der Modellierung werden für ausgewählte Städte in Deutschland und der Schweiz vorgestellt und im Anschluss diskutiert. Abschließend wird die Arbeit zusammengefasst, und es wird ein Ausblick auf zukünftige Entwicklungen gegeben.

Im Detail gliedert sich die Arbeit wie folgt:

**Kapitel 2** gibt einen Überblick zum gewählten Anwendungskontext Außenwerbung. Hierzu werden die wichtigsten Grundlagen der Medialeistungswerte erläutert. Im Anschluss werden die Verfahren der Leistungswertbestimmung in Deutschland und der Schweiz vor der Einführung von GPS vorgestellt und bewertet.

**Kapitel 3** ist ein weiteres Grundlagenkapitel. Hier werden die verwendeten Daten und Methoden der Dissertation vorgestellt. Zu den verwendeten Methoden gehören insbesondere Techniken des Data Mining und der räumlichen Analyse. Zu den eingesetzten Daten gehören neben Straßennetz- und GPS-Daten auch Datenprodukte wie der Frequenzatlas.

**Kapitel 4** beschreibt die z.T. aufwändige Aufbereitung von GPS-Daten und Plakatdaten. Insbesondere Unschärfen, Lücken und Oszillationen von GPS-Daten werden vor dem Hintergrund des Anwendungskontextes erläutert. Im Anschluss wird die geometrische Erstellung von Plakatsichtbarkeitsräumen vor dem Hintergrund der Verschneidung mit GPS-Trajektorien vorgestellt. Die Zusammenführung dieser beiden Datenquellen dient als Input für das folgende Kapitel der Reichweitenberechnung. Zusätzlich wird eine Überprüfung auf Stichprobenverzerrung in den GPS-Daten mittels Data Mining Techniken vorgestellt.

**Kapitel 5** stellt sich der Herausforderung, mit einer zeitlich und räumlich unvollständigen Datenmenge umzugehen und diese für die Modellierung von Reichweiten nutzbar zu machen. Hierzu werden geeignete Verfahren auf den Anwendungskontext adaptiert sowie

Verfahren vorgestellt, die mit der unzureichenden räumlichen Abdeckung von GPS-Probanden umgehen können.

**Kapitel 6** untersucht, wie robust die Ergebnisse hinsichtlich einer neuen oder veränderten kleineren GPS- Stichprobe sind.

**Kapitel 7** führt die Ergebnisse der einzelnen Modellierungsschritte aus Kapitel 5 zusammen. Für einzelne Teststädte werden räumlich differenzierte Reichweiten vorgestellt. Hierzu werden die Reichweiten nach soziodemographischen Gesichtspunkten und nach unterschiedlichen Einzugsgebieten berechnet.

**Kapitel 8** greift die Ergebnisse von Kapitel 4-7 auf und diskutiert diese. Es werden noch einmal die ursprünglich formulierten Ziele der Arbeit zusammengefasst und vor dem Hintergrund der Ergebnisse bewertet. Dabei werden sowohl die Vor- als auch die Nachteile der angewendeten Methodik sowie der Datengrundlage diskutiert.

**Kapitel 9** schließt mit einer Zusammenfassung und einem Ausblick auf zukünftige Entwicklungen die Arbeit ab.

## 1.6 Publikationen

Auszüge dieser Dissertation wurden bereits auf folgenden Konferenzen und Workshops sowie in folgenden Büchern veröffentlicht:

- D. Hecker, C. Körner und M. May. Robustness Analyses for Repeated Mobility Surveys in Outdoor Advertising. In Proc. of the 1th International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM'11), pages 148-153. 2011.

- D. Hecker, C. Körner, H. Stange, D. Schulz und M. May. Modeling micro-movement variability in mobility studies. In Lecture Notes in Geoinformation and Cartography, Volume 1, Part 2, pages 121-140, 2011.

- D. Hecker, C. Körner und M. May. Challenges and Advantages of using GPS- Data in Outdoor Advertisement. In Proc. of the 3th Conference on Geoinformatik - Geochange, pages 257-260. Akademische Verlagsgesellschaft. 2011.

- D. Hecker, H. Stange, C. Körner, M. May: Sample Bias due to Missing Data in Mobility Surveys. IEEE International Conference on Data Mining Workshops - ICDM 2010, pages 241-248, 2010.

- D. Hecker, C. Körner und M. May. Räumlich differenzierte Reichweiten für die Außenwerbung. In Angewandte Geoinformatik 2010, Beiträge zum 22. AGIT Symposium Salzburg, pages 194-203, 2010.

- D. Hecker, C. Körner, H. Streich und U. Hofmann. A Sensitivity Analysis for the Selection of Business Critical Geodata in Swiss Outdoor Advertisement. In GIScience 2010, Extended Abstracts Volume, 2010.

- C. Körner, D. Hecker, M. May und S. Wrobel. Visit potential: A Common Vocabulary for the Analysis of Entity-location Interactions in Mobility Applications. In Proc. of the 13th International Conference on Geographic Information Science (AGILE'10), 2010.

- C. Körner, D. Hecker, M. Krause-Traudes, M. May, S. Scheider, D. Schulz, H. Stange und S. Wrobel. Spatial Data Mining in Practice: Principles and Case Studies. In C. Soares und R. Ghani, editors, Data Mining for Business Applications. IOS Press, 2010.

- M. May, C. Körner, D. Hecker, M. Pasquier, U. Hofmann und F. Mende. Handling Missing Values in GPS- Surveys using Survival Analysis: a GPS-Case Study of Outdoor Advertising. In ADKDD '09: Proceedings of the Third International Workshop on Data Mining und Audience Intelligence for Advertising, pages 78-84, New York, NY, USA, 2009.

- M. May, C. Körner, D. Hecker, M. Pasquier, Urs Hofmann, und Felix Mende. Modelling Missing Values for Audience Measurement in Outdoor Advertising using GPS- Data. In GI Jahrestagung, Volume 154 of LNI, pages 3993-4006. GI, 2009

- M. Pasquier, U. Hofmann, F. H. Mende, M. May, D. Hecker und C. Körner. Modelling and Prospects of the Audience Measurement for Outdoor Advertising based on Data Collection using GPS-Devices (electronic passive measurement system). In Proceedings of the 8th International Conference on Survey Methods in Transport, 2008.

- M. May, D. Hecker, C. Körner, S. Scheider und D. Schulz. A Vector-Geometry Based Spatial knn-algorithm for Traffic Frequency Predictions. In Proc. of the 2008 IEEE International Conference on Data Mining Workshops (ICDMW '08), pages 442-447. IEEE Computer Society, 2008.

- D. Wegener, D. Hecker, C. Körner, M. May und M. Mock. Parallelization of r-programs with GridR in a GPS-trajectory Mining Application. In Proc. of the 1st Ubiquitous Knowledge Discovery Workshop (UKD'08), 2008.

Weitere Veröffentlichungen, die mit dem Thema verwandt sind, wurden auf folgenden Konferenzen und Workshops publiziert:

- N. Andrienko, G. Andrienko, H. Stange, T. Liebig, D. Hecker. Visual Analytics for Understanding Spatial Situations from Episodic Movement Data Journal KI Künstliche Intelligenz, Themenheft Spatiotemporal Modeling and Analysis, im Erscheinen, 2012.

- J. Schreinemacher, C. Körner, D. Hecker, G. Bareth. Analyzing Temporal Usage Patterns of Street Segments Based on GPS-Data – A Case Study in Switzerland. In Proc. of the 15th International Conference on Geographic Information Science (AGILE'10), im Erscheinen, 2012.

- T. Ellersiek, T. Liebig, D. Hecker und C. Körner. Analyse von raum-zeitlichen Bewegungsmustern auf Basis von Bluetooth-Sensoren. In Angewandte Geoinformatik 2012, Beiträge zum 24. AGIT Symposium Salzburg, im Erscheinen, 2012.

- H. Stange, T. Liebig, D. Hecker, G. Andrienko, und N. Andrienko. Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness, pages 51-58, 2011.

- T. Liebig, H. Stange, D. Hecker, M. May, C. Körner und U. Hofmann. A General Pedestrian Movement Model for the Evaluation of Mixed Indoor-Outdoor Poster Campaigns. In Proc. of the Third Workshop on Pervasive Advertising and Shopping, 2010.

- M. May, S. Scheider, R. Rösler, D. Schulz, D. Hecker. Pedestrian flow prediction in extensive road networks using biased observational data. Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, November 5-7, 2008, Irvine, USA, 1-4, 2008.

- D. Hecker, F. Warmelink. Die Einsatzmöglichkeiten von Geoinformationssystemen (GIS) im Verlagswesen. Eine Einführung zum Thema GIS und Geomarketing. BVDA, pages 1-74, 2005.

- D. Hecker, M. May, S. Scheider und A. Voss. Punktwerbung. In GeoBiT, No.8, pages 30-31, 2004.

- H. Voss, S. Scheider, D. Hecker und A. Voss. GIS-Visionen: Fast alle Daten werden verortet sein. In GeoBiT, No.11, page 18, 2004.

# KAPITEL 2

## 2. ANWENDUNGSKONTEXT AUßENWERBUNG

> *„On my way here I passed a local cinema and it turned out you were expecting me after all, for the billboards read: The Mummy Returns".*
>
> Margaret Thatcher (British Politician and Prime Minister, 1925-)

Der Raum ist seit jeher ein wichtiger Bestandteil der Außenwerbung. Aus plausiblen Gründen werden Plakatstandorte so gewählt, dass sie möglichst häufig von Menschen gesehen werden. Während jedoch der Raum bei der Wahl von Plakatstandorten schon immer ein entscheidendes Kriterium gewesen ist, spielt er bei der Leistungsbewertung und damit bei der Preisbildung von Plakatstandorten erst seit den vergangenen Jahren eine wichtige Rolle. In diesem Kapitel werden die Leistungswerte der Außenwerbung vorgestellt und definiert (Abschnitt 2.1). Darauf folgend wird in Abschnitt 2.2 am Beispiel der Schweiz und Deutschlands die ursprüngliche Methodik vor der Einführung von GPS beschrieben. Im Anschluss werden diese Verfahren in Abschnitt 2.3 bewertet, und der Abschnitt 2.4 schließt mit der Zusammenfassung des Kapitels.

### 2.1 Leistungswerte in der Außenwerbung

Die Außenwerbung ist ein passives Medium und damit anders als alle anderen Werbemedien. Sie wird nicht gezielt wie von Radiohörern oder Fernsehzuschauern ausgewählt, und man beschäftigt sich nicht in irgendeiner Weise mit ihr, wie es z.B. beim Fernsehen, beim Radio Hören, beim Zeitschriften Lesen oder auch bei der Internetwerbung der Fall ist. Daraus ergibt sich für die Außenwerbung eine erschwerte Leistungswertbestimmung, da die Nutzung nicht direkt über das Medium erfasst werden kann. Bei der Internetwerbung können z.B. Clicks auf Werbebanner über die IP Adresse abgespeichert und Click-Raten und Webseitenaufrufe mit geloggt werden. Beim Fernsehen wird z.B. bei einer Anzahl von Testpersonen mit einer speziellen Box permanent verfolgt, welches Programm aktuell läuft. Über diese Stichprobe wird auf die Grundgesamtheit der deutschen Fernsehzuschauer und bestimmte Fernsehsendungen hochgerechnet (Koschnick 2011).

Die Außenwerbung steht erst dann zur Verfügung, wenn man in ihre Sichtweite gelangt, und dies ist in der Regel der öffentliche Raum. Dabei gibt es unterschiedlichste Formate und Formen von Plakaten. Zu den klassischen deutschen Werbeformen zählt die Litfaßsäule, die von Herrn Litfaß in Berlin erfunden und im Jahre 1855 das erste Mal aufgestellt wurde. Inzwischen zählen zu den dominierenden städtischen Werbeträgern Megalight Poster und Allgemeinstellen (siehe Abbildung 2.1).

Abbildung 2.1: Typische Plakatformate in Deutschland, links: Megalight Poster, rechts: Allgemeinstelle (Ströer 2012)

Eins haben jedoch alle unterschiedlichen Plakatstellen gemeinsam: Nur mit Leistungswerten für jede individuelle Stelle kann eine Bepreisung stattfinden. Ein Plakat hat in der Regel einen hohen Wert, wenn viele Personen an diesem vorbeikommen. Die Anzahl der Passagen im Sichtbarkeitsraum eines Plakates ist also eine der entscheidenden Größen. Zusätzlich ist relevant, wie oft Passanten ein Plakat in einer bestimmten Zeit wahrgenommen haben und welche Soziodemographie diese haben. Diese Kriterien sind medienübergreifende Definitionen, die auch für die Außenwerbung gelten. Im Folgenden werden sie beschrieben und vorgestellt. Dazu wurden folgende Quellen verwendet: Swiss Poster Research Plus (2011), Arbeitsgemeinschaft Media Analyse (ag.ma 2012b), Sissors und Baron (2002) und Koschnick (2011):

**Reichweite (RW):** Der Begriff Reichweite beschreibt, wie viele Personen eine bestimmte Anzahl von z.B. Plakatstellen oder Werbespots in einer bestimmten Zeit gesehen haben. Die Reichweite ist eine der wichtigsten Kenngrößen in der Werbebranche. Sie dient dazu, unterschiedliche Mediengattungen miteinander zu vergleichen: Mit welchem Werbemedium erreiche ich mehr oder weniger unterschiedliche Personen in einer bestimmten Zeit? Die Reichweite kann in Kontaktklassen ausgewertet werden, indem eine bestimmte Mindestgröße für die Reichweite definiert wird. Das bedeutet, dass eine bestimmte Kontakthäufigkeit pro Person vorausgesetzt wird, bevor die Person als erreicht gilt und damit in den Reichweitenwert mit einfließt. Der Reichweitenwert steigt typischerweise kontinuierlich über die Zeit und wird in Prozent angegeben.

$$RW = \frac{erreichte\ Personen}{Population} * 100$$

**Opportunities to see (OTS):** Der OTS beschreibt die durchschnittliche Anzahl der Plakatkontakte/Werbekontakte aller Personen in einem bestimmten Zeitraum. Auch der OTS kann in Kontaktklassen ausgedrückt werden. Mit steigender Kontaktklasse steigt typischerweise auch der OTS Wert.

$$OTS = \frac{Bruttokontakte\ erreichter\ Personen}{erreichte\ Personen}$$

**Gross rating points (GRP):** Der GRP wird auch gerne als Kontaktdruck auf die Zielgruppe bezeichnet. Er beschreibt die durchschnittliche Menge der Kontakte, die 100 Personen der

anvisierten Zielgruppe produzieren. Er kann als OTS bei der Kontaktklasse 0 multipliziert mit 100 interpretiert werden.

$$GRP = \frac{Bruttokontakte}{Population} * 100$$

**Interpretation von Plakatkontakten**

Grundsätzlich ist die Definition eines Kontaktes in der Plakatforschung eine sehr schwierige Aufgabe und eine eigene Forschungsrichtung. Es besteht die Frage, ob man die Definition des Kontaktes auf den Werbeträger (Plakatstelle) oder auf die konkret geschaltete Werbung (Werbemittel) bezieht. Denn dass jemand an einer Plakatstelle vorbeikommt, also eine Kontaktchance mit dem Plakat hat, bedeutet nicht, dass er auch die Werbung wahrgenommen und einen werbewirksamen Kontakt erzeugt hat (vgl. Scheier, 2005, S.279). Vor diesem Hintergrund wird in der Außenwerbung mit der sogenannten Kontaktklasse gearbeitet:

**Kontaktklasse (KK):** Die Kontaktklasse beschreibt eine festgelegte Mindestgröße an Kontakten, die erreicht werden müssen, damit eine Person als erreicht gilt (z.B. Kontaktklasse = 5, eine Person muss mindestens fünfmal eine Plakatwerbung wahrgenommen haben, damit sie zur Reichweite beiträgt).

Es existieren zwei unterschiedliche Möglichkeiten Plakatkontakte zu interpretieren.

1. Als Dosismodell, in dem jeder einzelne Plakatkontakt kumulativ über die Zeit aufaddiert wird (Werte zwischen 0 und 1).
2. Als ein Wahrscheinlichkeitsmodell. Jede Person, die ein Plakat passiert, hat eine bestimmte Wahrscheinlichkeit, das Plakat zu sehen (Wert entweder 0 oder 1). Der Kontakt wird entsprechend der Wahrscheinlichkeit simuliert.

Die Gewichtung ist abhängig von den individuellen Wahrnehmungsfaktoren des Standortes. Beim Wahrscheinlichkeitsmodell ist eine aufwändigere Modellierung als beim Dosismodell notwendig. Denn dieser Ansatz benötigt ein zugrunde liegendes Simulationsmodell, dass bestimmt, ob eine Passage zu einem Kontakt geführt hat oder nicht. Beim Wahrscheinlichkeitsmodell gilt eine Person erst dann als erreicht, wenn sie für die Kontaktklasse 1 einen Wert von 1 erzielt hat. Im Dosismodell gilt eine Person bereits bei einem Wert $> 0$ als erreicht. In Abbildung 2.2 werden beide Varianten miteinander verglichen. Es kommen jeweils insgesamt 10 unterschiedliche PKW-Fahrer mit einer Gewichtung von 0,3 an einem Plakat vorbei. Beim Dosismodell werden alle 10 Passagen mit 0,3 gewertet, was zu einer Gesamtkontaktmenge von 3 Kontakten führt. Bei der Berechnung mit Kontaktwahrscheinlichkeiten haben bei einer Gewichtung von 0,3 im Mittel 3 PKW-Fahrer einen Kontakt von 1 und 7 PKW-Fahrer von 0 und fließen somit nicht in die Bewertung mit ein. Ein wesentlicher Unterschied ergibt sich bei der Berechnung der Reichweite für Kontaktklasse 0 und 1. Bei der Berechnung der Kontaktdosen ergibt sich für die Kontaktklasse 0 eine Reichweite von 100%, da alle Fahrer erreicht worden sind. Für Kontaktklasse 1 ergibt sich eine Reichweite von 0%, da kein Fahrer einen Wert > 1 erreicht hat. In der Reichweitenberechnung mit Kontaktwahrscheinlichkeiten entfällt die Kontaktklasse 0, da entweder ein Vollkontakt oder kein Kontakt entsteht. Von insgesamt 10 PKW-Fahrern haben 3 einen Vollkontakt. So ergibt sich eine Reichweite bei Kontaktklasse 1 von 30%.

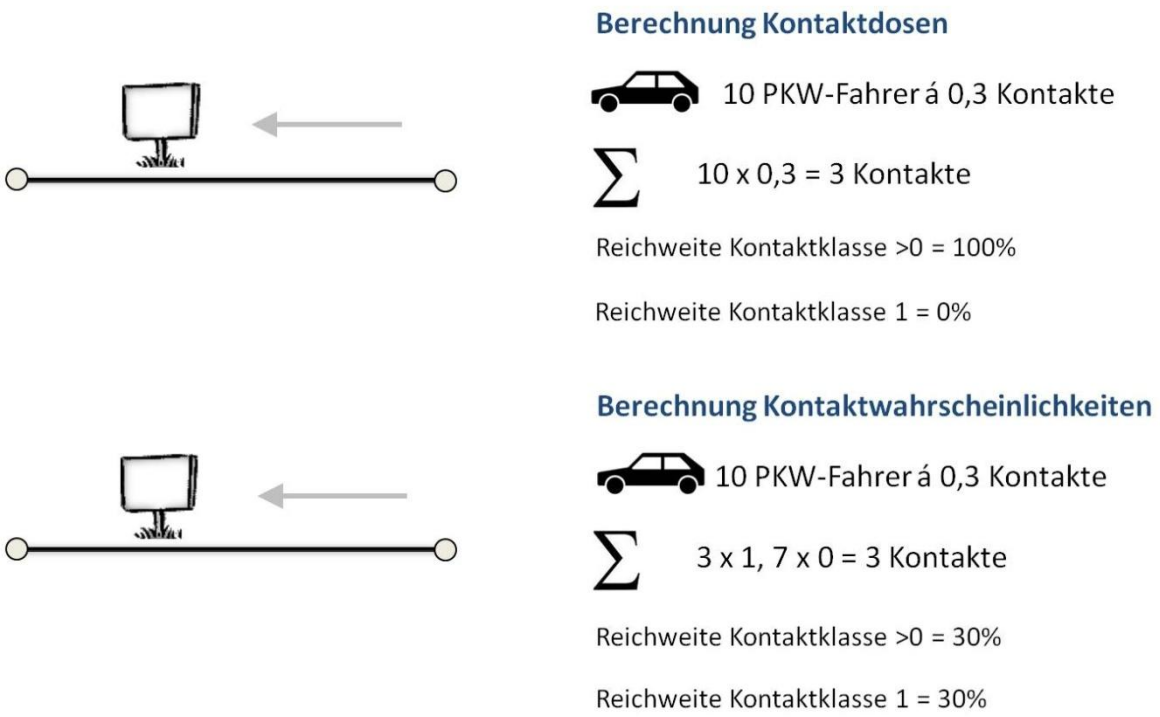



Abbildung 2.2: Kontaktwahrscheinlichkeit vs. Kontaktdosis

**Kontaktgewichtung Außenwerbung**

Zur Bestimmung des Plakatkontaktes wurden und werden aktuell in vielen Ländern Sichtbarkeitsstudien durchgeführt (Deutschland, Österreich, Frankreich, Südafrika). Zu den bekanntesten Sichtbarkeitsstudien gehört das britische Postar-Modell von Paul Barber (Koschnick 2011). In dieser Studie werden Probanden unter Laborbedingungen insgesamt 72 unterschiedliche Situationen auf Bildern gezeigt. Diese enthalten unterschiedliche Plakatformate bzw. Standorte in unterschiedlichen Verkehrssituationen. Jedes Bild wird den Probanden insgesamt 6 Sekunden gezeigt und ihre Blickverläufe werden in dieser Zeit aufgezeichnet. Als Hintergrund der Studie wurde den Probanden mitgeteilt, dass es sich um eine verkehrspsychologische Untersuchung handelt. Sie sind also nicht bereits auf spezielle Bildelemente fokussiert. Unterschieden wird die Bewertung in Fußgänger, PKW-Fahrer und Beifahrersicht. Es zeigte sich, dass bei den Blickverläufen der Probanden folgende Faktoren eine wichtige Rolle spielen:

1. die Größe, die Entfernung und der Winkel der Fläche zum Betrachter,
2. ablenkende Faktoren (Sichthindernisse, Anzahl weiterer Werbeflächen in direkter Nähe),
3. Beleuchtung der Plakatstelle,
4. Dauer der Passage.

Die Bestimmung der einzelnen Faktoren wird für fast jede Plakatstelle vor Ort vorgenommen. Dabei gilt die Regel, je sichtbarer eine Stelle ist, desto höher sind die Faktoren und desto höher ist die Kontaktchance. In Europa wurde in den letzten Jahren eine Vielzahl von Sichtbarkeitsstudien mit zum Teil unterschiedlichen Erfassungsmethodiken durchgeführt. Zum Teil variiert die Anzahl der Faktoren, jedoch verwenden fast alle Länder die oben

genannten Faktoren. Für eine spätere Modellierung der Leistungswerte muss die Möglichkeit bestehen, Sichtbarkeitsfaktoren in das Modell zu integrieren.

## 2.2 Leistungswertbestimmung in Deutschland und der Schweiz vor GPS

In der Deutschen und Schweizer Außenwerbung waren lange Verfahren im Einsatz, die auf dem Einsatz von telefonischen Befragungen (Computer Assisted Telephone Interviews) beruhen. Sie erfragen das Verhalten von Probanden an zurückliegenden Tagen und basieren damit auf einer Erinnerungsleistung der Interviewten. Einen Sonderfall in der Leistungsbewertung stellt Deutschland dar. Hier wurden seit den 90 Jahren neben den Telefonbefragungen auch Frequenzzählungen durchgeführt. Diese sind seit 2003 Datenbasis für ein deutschlandweites Verkehrsmodell. Im Folgenden werden für Deutschland und die Schweiz die jeweiligen Verfahren und ihre Limitationen in der Berechnung vorgestellt. Am Ende des Kapitels werden die Verfahren bewertet.

### 2.2.1 Leistungswertbestimmung in Deutschland

Im folgenden Abschnitt werden die Plakat-Media-Analyse (PMA) zur Reichweitenbestimmung und der G-Wert zur Kontaktbestimmung am Plakat vorgestellt. Die PMA war die deutsche Reichweitenforschung für das Medium Plakat, die bis zum Jahre 2003 durch den Fachverband Außenwerbung (FAW) in Auftrag gegeben wurde. Ab dem Jahr 2004 hat die Arbeitsgemeinschaft Mediaanalyse (ag.ma) die Forschung übernommen und die Studie unter der Bezeichnung „ma 2004 Plakat" veröffentlicht. Aufgabe und Ziel der PMA war es, intermedial vergleichbare Leistungswerte bereit zu stellen und für den Mediaplaner nutzbar zu machen (ag.ma 2012d). Die PMA ermittelt die relevanten Leistungswerte wie Reichweite, OTS und GRP. Kernmethodik der PMA war die von Gunda Opfer entwickelte Methode zur „Abfrage anhand erinnerter Wege". Auf der Grundlage von insgesamt fast 22.000 Interviews wurden Probanden über ihre Wege außer Haus am vorherigen Tag befragt. Im Interview versuchte man, die Wege über folgende Teilschritte zu zerlegen:

**Anlass:**

1. Aus welchen Anlässen hat man das Haus in der letzten Woche verlassen?

**Häufigkeit:**

2. Wie verlaufen Hin- und Rückweg pro Anlass?
3. Welche Verkehrsmittel werden benutzt?
4. Wie viele Plakatstellen liegen an den einzelnen Abschnitten des Weges?

Als Erinnerungshilfen dienten beim Interview 15 Karten mit Anlässen, 10 Karten mit Verkehrsmitteln sowie Bilder typischer Plakatstellen (Opfer, 2005, S.297). Mit dieser Methode misst man die Anzahl der wahrgenommen Plakatstellen unabhängig davon, was gerade plakatiert ist. Die Probanden wurden gebeten, ihre erinnerten Wege nach und nach zu rekonstruieren und sich an die wahrgenommenen Plakatstellen zu erinnern. Dies bedeutet, dass mit jeder Wiederholung eines typischen Weges ein weiterer Kontakt mit einer Plakatstelle entsteht. Summiert man im Anschluss alle gewonnenen Informationen der Interviews, so erhält man für eine Vollbelegung die Information, wie viele unterschiedliche Personen im Durchschnitt diese Belegung, aufgeteilt nach unterschiedlichen Plakatformaten, gesehen haben. Eine Vollbelegung bedeutet an dieser Stelle, dass alle Plakate eines bestimmten Plakatformates mit der identischen Werbung geschaltet sind. Teilbelegungen, die

sich über Bevölkerungsquoten von z.B. 1 Plakat pro 3000 Einwohner berechnen, werden über einen bestimmten Schlüssel am prozentualen Anteil von der Vollbelegung kalkuliert. Für die Berechnung der Leistungswerte für einzelne Belegungsquoten in Städten wurden jeweils alle Interviews aus einem betreffenden Bundesland und einer ähnlichen Ortsgröße stellvertretend für den zu bestimmenden Ort, wo es keine Probandeninformationen gab, herangezogen. Die resultierenden Leistungswerte wurden anhand der Einwohnerzahl des belegten Ortes skaliert. Die PMA dient dazu, durchschnittliche Leistungen für Werbeträger zu bestimmen, eine Bewertung von einzelnen Plakatstellen ist mit dieser Bewertungsmethode nicht möglich.

**G-Wert und Frequenzatlas Deutschland**

Der G-Wert ist im Gegensatz zu den durchschnittlichen Werten der PMA ein Verfahren zur Plakateinzelstellenbewertung, die keine Reichweiten und OTS liefert, sondern einen Bruttowert der Kontakte. Die Bezeichnung „G-Wert" steht für Gesamtwert und soll ausdrücken, dass er alle relevanten Verkehrsströme im Umfeld eines Plakates erfasst. Der G-Wert wurde von der GfK Marktforschung in Nürnberg ursprünglich für den Zigarettenkonzern Reynolds Tobacco entwickelt und später vom Fachverband Außenwerbung weitergeführt (Pasquier 1997, Engel & Hofsäss 2003). Dabei unterteilt sich der G-Wert in zwei Bereiche:

1. **Quantitativer Bereich**: Er gibt an, wie viele Personen insgesamt an einer Plakatstelle vorbeikommen.

2. **Qualitativer Bereich:** Er stellt die qualitative Gewichtung der Passagen anhand von Sichtbarkeitsfaktoren am Plakat dar.

Grundlage für die quantitativen Auswertungen ist der Frequenzatlas. Er gibt an, wie viele Personen pro Stunde durchschnittlich an einer Plakatstelle vorbei kommen. Dabei ist die durchschnittliche Stunde auf die Tage Montag bis Freitag und 7 bis 19 Uhr beschränkt. Aufgeteilt nach den drei Verkehrsmodiarten PKW, Fußgänger und ÖPNV ist seit 2004 der Frequenzatlas Deutschland die Basis für die Berechnung der Passagenmenge. Der Frequenzatlas wurde vom Fraunhofer Institut IAIS entwickelt und liefert inzwischen für alle circa 6,9 Millionen Navteq Straßensegmente in Deutschland Verkehrsfrequenzwerte (May 2008a, May 2008b). Diese Verkehrsfrequenzwerte werden im Anschluss zusammen mit den qualitativen und individuellen Plakatwirkungsfaktoren zum G-Wert verrechnet. Zu den Plakatwirkungsfaktoren des G-Wertes und damit zum qualitativen Bereich gehören folgende Variablen:

- Winkel des Plakates zum jeweiligen Verkehrsstrom,
- Kontaktchancendauer,
- Entfernung des jeweiligen Verkehrsstroms zum Plakat,
- Grad der Verdecktheit durch Sichthindernisse (parkende Autos, Büsche, etc. im Sichtbarkeitsraum des Plakates),
- Umfeldkomplexität (Wie stark konkurrieren andere visuelle Reize und weitere Plakatstellen die betreffende Plakatstelle um Aufmerksamkeit?),
- Situationskomplexität (Wie viel Aufmerksamkeit kostet die Bewältigung der Verkehrssituation?),
- Höhe des angebrachten Werbeträgers,

- Beleuchtungsverhältnisse.

Zur Festlegung der einzelnen Bewertungsfaktoren wurden alle Merkmale einer Werbefläche in einer Serie von Passantenbefragungen systematisch untersucht und auf ihre Auswirkungen auf die Wahrnehmbarkeit getestet. Als Maß für die Wahrnehmbarkeit diente der Prozentsatz der Passanten, die sich in einem Wiedererkennungstest an ein bestimmtes Plakatmotiv erinnern konnten (Schlossbauer, 1997). Hierzu wurden im Anschluss an die Passage Testpersonen drei Plakatmotive gezeigt (Autofahrer an einer nahegelegenen Tankstelle). Folgende Fragen mussten beantwortet werden: „Sie sind gerade an einem dieser Plakatmotive vorbeigekommen? Können Sie sich daran erinnern, welches dieser Plakatmotive Sie gesehen haben?". Als Testplakat wurde dabei, bezogen auf Aufmerksamkeit und Wirkung, ein eher durchschnittliches Plakatmotiv ausgewählt. Auf Grundlage der oben genannten Wahrnehmbarkeitsvariablen wird nun zur Berechnung des G-Wertes für jeden einzelnen Verkehrsstrom der Erinnerer-Anteil bestimmt. Dabei werden die einzelnen Variablen multiplikativ miteinander verrechnet. Summiert man im Anschluss die Anzahl aller Erinnerer für alle Verkehrsströme einer Plakatstelle, so erhält man den G-Wert. Zur Erfassung der Wahrnehmbarkeitsvariablen hat die GfK in den letzten Jahren im Rahmen einer Qualitätsoffensive der Außenwerbung einen Großteil aller deutschen Plakatstellen vor Ort bewertet. Im Anschluss wurden die Wahrnehmbarkeitsvariablen mit dem Frequenzatlas verrechnet. Ergebnis ist eine individuelle Stellenbewertung für alle erfassten Plakatstellen, die die Grundlage für die Preisbestimmung in Deutschland darstellt. Dabei handelt es sich bei dieser Einzelstellenbewertung rein um einen Gesamtkontaktwert. Reichweiten, OTS oder GRP Werte können mit der Methode zur G-Wert Bestimmung nicht ausgewiesen werden.

### 2.2.2 LEISTUNGSWERTBESTIMMUNG IN DER SCHWEIZ

In der Schweiz war bis zum Jahre 2008 das sogenannte Copland Modell Grundlage zur Leistungswertbestimmung. Ausgangspunkt war eine Befragung von Testpersonen zu ihrem Mobilitätsverhalten. Die Testpersonen wurden gebeten, ihre am Vortag zurückgelegten Wege auf Stadtplänen einzuzeichnen. Diese aufgezeichneten Wege, die auf der Erinnerungsleistung der Probanden beruhten, wurden anschließend mit den Plakatstellen in den Städten verschnitten. Ziel der Erhebung ist es, den sogenannten *AWert* zu bestimmen, der die durchschnittliche Anzahl von Kontaktchancen mit Plakaten pro Stadtgebiet bestimmt (vgl. PASQUIER 1997). Durch Addition der hochgerechneten Kontaktchancen von den Testpersonen ergibt sich die Gesamtkontaktchancenanzahl für die erhobene Stadt.

$$AWert = \frac{Gesamtkontaktchancen}{Population * Plakatstellen}$$

Aus der Berechnung ergibt sich eine enge Beziehung zwischen den Gesamtkontaktchancen und der Bevölkerungsanzahl in den untersuchten Stadtregionen. Der *AWert* wird umso kleiner, je größer die Stadtregion ist. Damit wird der Annahme entsprochen, dass in einer kleinen Stadt die Bevölkerung eine höhere Wahrscheinlichkeit besitzt, einem Großteil der Werbeträger zu begegnen, als in einer Großstadt. In einer Großstadt findet die alltägliche Bevölkerungsmobilität hingegen nur im direkten Wohn- und Arbeitsumfeld statt, jedoch nicht in allen Bezirken der Stadt. Somit besteht nur für eine begrenzte Anzahl von Werbeträgern eine Kontaktchance. Mit dieser Berechnung kann für jede erhobene Stadtregion eine durchschnittliche Kontaktanzahl pro Plakatstelle berechnet werden. Der *AWert* dient als wichtiger Faktor für das Reichweitenmodell nach Copland. Neben dem *AWert* gehören zu den Faktoren der Reichweitenbestimmung die Anzahl der Plakatstellen einer Kampagne (*S*), die Aushangdauer (*T*) und ein Immobilitätsfaktor (*b*).

$$Reichweite = \frac{AWert * T * S}{AWert * T * S + b}$$

Dabei ist der Immobilitätsfaktor (*b*) ein konstanter Wert, der ausdrücken soll, dass eine bestimmte Anzahl der Bevölkerung mit einer Plakatkampagne selten oder gar nicht erreicht werden kann. Der Immobilitätsfaktor ist nach Copland ein konstanter Mittelwert für eine durchschnittliche Stadt eines Landes.

## 2.3 Bewertung der vorgestellten Verfahren

Die Nachteile der PMA und des Copland Modells sind offensichtlich. Bei beiden Ansätzen handelt es sich um eine reine Erinnerungsleistung der Probanden. Zudem beruhen die Berechnungen durchweg auf durchschnittlichen Plakatkampagnen. Eine Plakatkampagne, die breit über die ganze Stadt gestreut ist, und eine Kampagne, die nur in einer Straße plakatiert ist, haben den gleichen Reichweitenwert. Der Beitrag eines einzelnen Plakates an der Leistung kann nicht identifiziert und somit auch nicht durch die Außenwerbung bepreist werden. Individuellen Sichtbarkeitskriterien eines Plakates, wie der Abstand zur Fahrbahn, der Winkel des Plakates zur Fahrbahn oder der Beleuchtung werden durch die Verdurchschnittlichung nicht Rechnung getragen. Zudem bilden die aufgezeichneten Wege der Probanden das Mobilitätsverhalten nur eines Tages ab. Die typische Buchungsdauer in der Außenwerbung beträgt jedoch mindestens 7 Tage. Hierzu wird der AWert mit der Aushangdauer multipliziert, was zu einem linearen Anstieg der Kontaktchancen führt und annimmt, dass z. B. die Montagsmobilität auch gleichzeitig einen Sonntag repräsentiert. Somit wird die Varianz in der Mobilität eines Probanden innerhalb einer Woche massiv unterschätzt. Der geographische Raum spielt bei beiden Ansätzen keine Rolle. In Abbildung 2.3 ist eine typische Plakatkampagne für Köln mit insgesamt 100 Plakatstellen dargestellt. Es handelt sich um eine gestreute Plakatkampagne, die nicht auf ein bestimmtes Gebiet in Köln konzentriert, sondern flächig verteilt ist. Auf der rechten Seite der Abbildung ist die tatsächliche Reichweitekurve der Kampagne dargestellt. Die X-Achse stellt die Anzahl der Tage von 1-7 dar und die Y-Achse die Reichweite in Prozenten. Beim PMA und Copland Modell steigt aufgrund der Vervielfachung der Befragungsdaten die Reichweitenkurve linear über die Zeit.

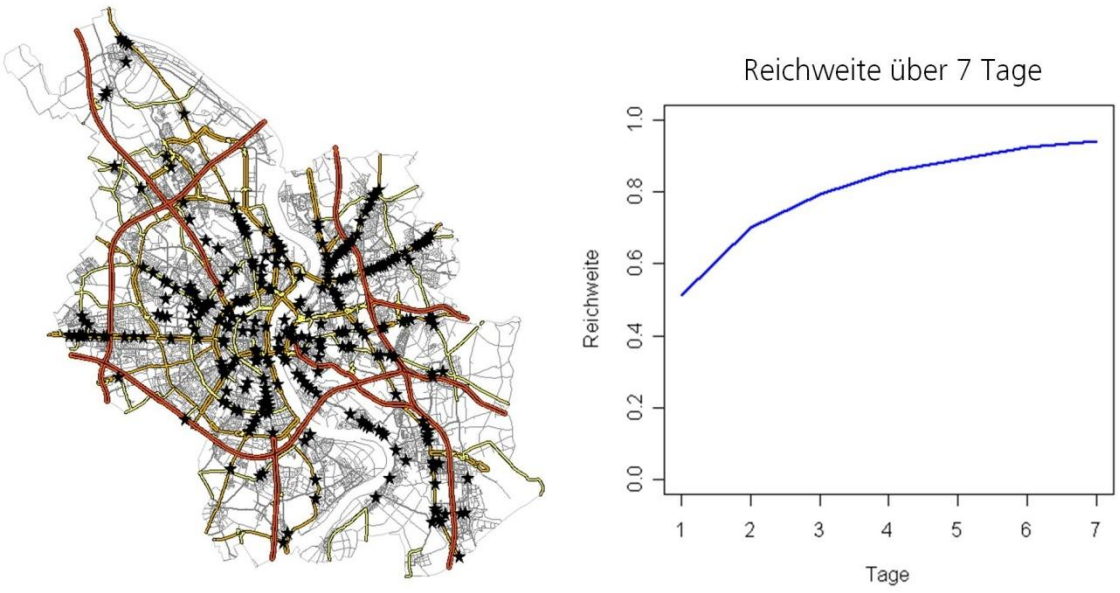



Abbildung 2.3: Reichweiten für eine gestreute Kampagne (Hecker 2010b)

Die reale Reichweite entwickelt sich jedoch, wie in der obigen Grafik dargestellt, in einer Kurvenform, die nach einer gewissen Aushangdauer in die Sättigung geht. So ist ab dem fünften Tag kein starker Anstieg der Reichweitenkurve mehr zu verzeichnen, da schon ein Großteil der Bevölkerung erreicht wurde. Am siebten Tag wird eine Gesamtreichweite von 85% erreicht. Diesen realen Reichweitenverlauf können die PMA und das Copland Modell nicht abbilden. Zudem unterscheidet sich bei der räumlichen Betrachtung eine gestreute Plakatkampagne mit hundert Werbeträgern stark von einer lokal aufgestellten Plakatkampagne mit der gleichen Anzahl von Werbeträgern. Die Annahme ist, dass bei einer Plakatkampagne, die nur in einem Stadtviertel geschaltet ist, Personen aus anderen Stadtvierteln eher eine geringe Chance haben, dieses Plakatnetz zu sehen. Jedoch wird im PMA- und Copland Modell an dieser Stelle kein Unterschied gemacht. Die Vorteile des PMA und des Copland Modelles liegen in ihrer Einfachheit. So braucht man beim Copland Modell lediglich drei Variablen und eine Konstante zur Bestimmung der Reichweite. Mit der Variable Kontakte pro Person (Zielgruppe), der Population und der Anzahl der Plakatstellen lassen sich für jeden Zeitraum Leistungswerte bestimmen.

Der in Deutschland eingesetzte G-Wert hat im Unterschied zum AWert und der PMA den Vorteil einer einzelstellenbezogenen Bewertung. Individuellen Faktoren wird in Form der Wahrnehmbarkeitsfaktoren Rechnung getragen. Jedoch kann der G-Wert solche Leistungswerte wie Reichweite, OTS und GRP nicht liefern. Der zugrundeliegende Frequenzatlas liefert nur eine reine Bruttofrequenz und keine Differenzierung hinsichtlich Wiederholerkontakten oder Soziodemographie.

Zusammenfassend kann festgestellt werden, dass die vorgestellten Verfahren keine exakten und differenzierten Leistungswerte zur Verfügung stellen. Damit steht die Außenwerbung im Vergleich zu den anderen Mediengattungen hinsichtlich der Ausweisung von Reichweiten undifferenzierter dar. In den folgenden Kapiteln werden neue Datengrundlagen und Methoden vorgestellt, die umfassendere Leistungswertaussagen zulassen.

## 2.4 Zusammenfassung

In diesem Kapitel wurde der Anwendungskontext der Arbeit vorgestellt. Hierzu wurden die wichtigsten Medialeistungswerte Reichweite, OTS und GRP definiert und erläutert. Zusätzlich wurden die unterschiedlichen Möglichkeiten einer Kontaktdefinition in der Außenwerbung dargestellt: das Kontaktdosismodell und das Kontaktwahrscheinlichkeitsmodell. Im Anschluss wurden die deutschen Modelle zur Leistungswertbestimmung vor der Einführung von GPS vorgestellt. Hierzu zählen die Plakatmediaanalyse und der G-Wert. Beide Verfahren haben unterschiedliche Ziele und Aussagen und greifen jeweils auf eine andere Datengrundlage zurück. Für die Schweiz wurde das AWert Modell präsentiert. Die PMA und der AWert sind pragmatische Modelle, aber berücksichtigen nur Durchschnittswerte und Erinnerungsleistungen. Der G-Wert ist hier differenzierter, kann aber nicht die für die Außenwerbung relevante Zielgröße der Reichweite liefern. Daraus wird deutlich, dass die bisherigen Verfahren keine räumlich differenzierten Leistungswerte zur Verfügung stellen.

# KAPITEL 3

## 3. GRUNDLAGEN - DATEN UND METHODEN

"*We are drowning in information, but starving for knowledge*".

John Naisbett (amerikanischer Zukunftsforscher, 1929-)

Im vorigen Kapitel wurde der Anwendungskontext der Arbeit vorgestellt. Er motiviert diese Dissertation und liefert mehrere herausfordernde Forschungsfragen im Umgang mit der Analyse von Mobilitätsdaten. Das Verständnis von geographischen Daten und Mobilitätsdaten haben für die Leistungsbewertung von Plakatstellen einen entscheidenden Einfluss. Aus diesem Grund werden in diesem Kapitel die Eigenschaften, der Umgang und die Analyse von diesen Daten beschrieben. Dieses Kapitel ist vornehmlich an die Leser gerichtet, die nicht vertraut sind mit dem Umgang von geographischen Daten, Mobilitätsdaten und Data Mining Techniken.

Detaillierter beschrieben ist das Kapitel wie folgt strukturiert: Abschnitt 3.1 erläutert geographische Datenmodelle, Strukturen sowie Charakteristiken von Geodaten. Insbesondere die topologische und geometrische Modellierung werden vor dem Hintergrund der späteren anwendungsorientierten Analysen und räumlichen Verschneidungen dargestellt. Im Anschluss fokussiert Abschnitt 3.2 die Beschreibung von Mobilitätsdaten. Hierbei werden Trajektorien definiert, ihre Datenstrukturen erläutert und Analysemethoden vorgestellt. Zusätzlich werden Möglichkeiten der Mobilitätsdatenerhebung und Charakteristiken der menschlichen Mobilität präsentiert. Abschnitt 3.3 stellt Verfahren der Wissensentdeckung in Datenbanken vor. Diese spielen insbesondere bei der späteren Robustheitsanalyse von Reichweitenberechnungen eine wichtige Rolle. Abschnitt 3.4 beschreibt die deutsche und Schweizer GPS-Feldstudie, die als Datenbasis für die spätere Berechnung der Reichweiten in Deutschland und der Schweiz dient. Desweiteren werden die zugrundeliegenden Straßennetzdaten Vector25 und NavTeq sowie der Frequenzatlas für Deutschland vorgestellt. Abgeschlossen wird das Kapitel mit einer Zusammenfassung der wichtigsten Fakten.

### 3.1 Geographische Daten

Dieser Abschnitt stellt die Eigenschaften von geographischen Daten vor. Dabei wird wie folgt vorgegangen: Zuerst werden möglichen Datenquellen von geographischen Daten vorgestellt. Im Anschluss werden topologische und geometrische Relationen von geographischen Objekten erläutert. Der geometrische Lagevergleich und die räumliche Aggregation schließen diesen Abschnitt ab.

### 3.1.1 GEODATENQUELLEN

Geographische Daten behandeln den Übergang von einem Ausschnitt der realen Welt und den damit verbundenen Informationen zu einer abstrakten Datenwelt. Dieser Übergang vollzieht sich in mehreren Schritten und mit unterschiedlichen Modelltypen. Ziel ist es jeweils, Objekte und Flächen auf der Oberfläche der Erde zu verorten. Dabei wird unterschieden zwischen Rohdaten und interpretierten Daten (Bartelme, 2005). Rohdaten, wie z.B. die Positionserfassung durch GPS, stellen direkte Bezüge zur Oberfläche her. Oft werden sie durch nachgelagerte Bearbeitungsschritte mit zusätzlicher semantischer bzw. interpretierter Information angereichert. So wird z.B. bei der Analyse von GPS-Trajektorien erst im Nachhinein die Fortbewegungsart (PKW, Fußgänger, ÖPNV) über die aufgezeichnete Geschwindigkeit ermittelt. Bei interpretierten Daten stammen die Informationen auch oft aus vorher manuell digitalisierten Plänen und Karten. Hier ergibt sich einerseits das Problem, dass man bereits mit Sekundärdaten arbeitet und zweitens diese interpretieren muss. Oft werden unter dem Gesichtspunkt eines bestimmten Anwendungszweckes Generalisierungen der Informationen vorgenommen, und es entstehen dadurch zum Teil Lücken in der Erfassung. Um reale oder abstrakte Objekte unserer Umwelt abzubilden, können wir mehr oder weniger genau vorgehen. In der Abbildung 3.1 ist ein Fluss in unterschiedlichen Generalisierungsstufen visualisiert (A-D). Die beiden oberen Vektordarstellungen besitzen jeweils eine unterschiedliche Menge von Stützpunkten und geben so den Flussverlauf unterschiedlich genau wieder (A, B). In der Abbildung C wird der Flussverlauf mit Rasterzellen wiedergegeben und in Abbildung D wird der Flussverlauf in einer sehr detailliert wiedergegeben. Auch der Umriss z.B. eines Gebäudes kann in sehr detaillierter Form als 3D Polygon mit Einbuchtungen, Rundungen und Höhe dargestellt werden, oder als standardisiertes, planares Rechteck. Für den Anwendungskontext der Dissertation hat die unterschiedliche Erfassung und Generalsierung von Geodaten insbesondere dann eine gewichtige Rolle, wenn es um die räumliche Verschneidung von Datenquellen unterschiedlicher Herkunft geht. So ist es bei der Verschneidung von GPS-Daten und Straßendaten ein Unterschied, ob ein Kreisverkehr als einzelner Punkt, oder als Kreisverkehr mit mehreren Segmenten und verknüpfenden Punkten erfasst worden ist. An dieser Stelle muss bereits bei den ersten Datenaufbereitungsschritten große Sorgfalt gewährleistet sein, um nicht bereits sehr früh Probleme für die spätere Modellierung zu bekommen.

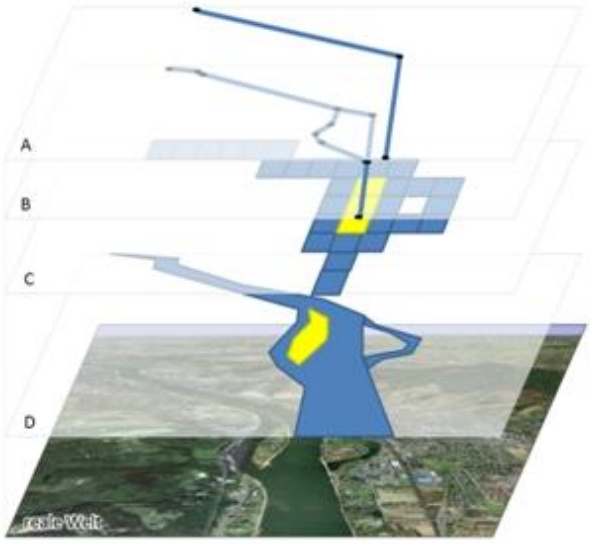



Abbildung 3.1 Geometrisches Modell und reale Welt (nach Bartelme 2005)

Neben den Möglichkeiten der geometrischen Vereinfachung gibt es auch bei der semantischen Informationstiefe unterschiedliche Stufen. So können Straßen in unterschiedliche Kategorien lokaler oder nationaler Bedeutung eingeteilt werden. Einzelne Straßensegmente können Informationen über Anzahl der Spuren, erlaubte Geschwindigkeit, ÖPNV (Ja/Nein), usw. enthalten. Neben den Erfassungsmethoden und dem Detaillierungsgrad von Geodaten sind insbesondere die topologischen Eigenschaften im Anwendungskontext von Bedeutung.

### 3.1.2 TOPOLOGISCHE RELATION

Die topologischen Beziehungen von Objekten im $\mathbb{R}^2$ stellen eine wichtige Grundlage bei vielen geographischen Fragestellungen dar. Zu diesen Fragen gehören z.B.:

- Welche Straßensegmente liegen auf dem Gebiet der Gemeinde Köln?
- Welche Personen besuchen das Gebiet A und B?
- Gibt es eine Überlappung zwischen Polygonen?

Die topologischen Beziehungen befassen sich mit Fragen der Nachbarschaften, des Enthaltenseins und des Verbundenseins. Im 9-Intersection-Modell von Egenhofer (Egenhofer, 1991) wird die Ausprägung von räumlichen Beziehungen von Objekten als Kriterium für die Bildung von Klassen eines formalen Modells gewählt. Jedes flächenhafte Objekt wird als geschlossene zweidimensionale Teilmenge des Raumes aufgefasst, das einen Innenraum (°), eine Grenze ($\partial$) und einen Außenraum (C) besitzt.

**Innenraum** (°)

Das Innere besteht aus Punkten, Linien oder Flächen, welche im Objekt liegen, aber nicht zur Grenze gehören.

**Grenze** ($\partial$)

Die Grenze besteht aus Punkten oder Linien, welche das Innere vom Äußeren trennen. Der Rand einer Linie besteht aus den Endpunkten. Der Rand eines Polygons ist die Linie, welche den Perimeter definiert.

**Außenraum** (C)

Das Äußere oder das Komplement besteht aus den Punkten, Linien oder Flächen, welche nicht zum Objekt oder zur Grenze gehören.

Die topologischen Relationen zwischen 2 Objekten $A, B$ lassen sich über die Bildung von Schnittmengen zwischen dem Außenraum, der Grenze und dem Innenraum bilden. Über das Bilden von Schnittmengen ergibt sich eine 3x3-Felder Matrix, in der jeder Eintrag entweder leer (0) oder nicht leer (1) sein kann. In der Matrix kann das 9-Intersection Modell wie folgt dargestellt werden (Tabelle 3.1).

| ∩ | B° | B $\partial$ | BC |
|---|---|---|---|
| A° | A°∩ B° | A°∩ B$\partial$ | A°∩ BC |
| A $\partial$ | A $\partial$ ∩ B° | A $\partial$ ∩ B$\partial$ | A $\partial$ ∩ BC |
| AC | AC ∩ B° | AC ∩ B$\partial$ | AC ∩ BC |

Tabelle 3.1: Matrix 9-Intersection Modell (nach Egenhofer 1991)

Es kann theoretisch zu insgesamt $2^9 = 512$ unterschiedlichen topologischen Relationen kommen. Egenhofer schließt jedoch mit Hilfe von geometrischen Überlegungen die meisten Möglichkeiten aus. Bei zwei einfachen Flächen sind nach dem 9-Intersection Modell 8 Möglichkeiten realisierbar (vgl. Abbildung 3.2).

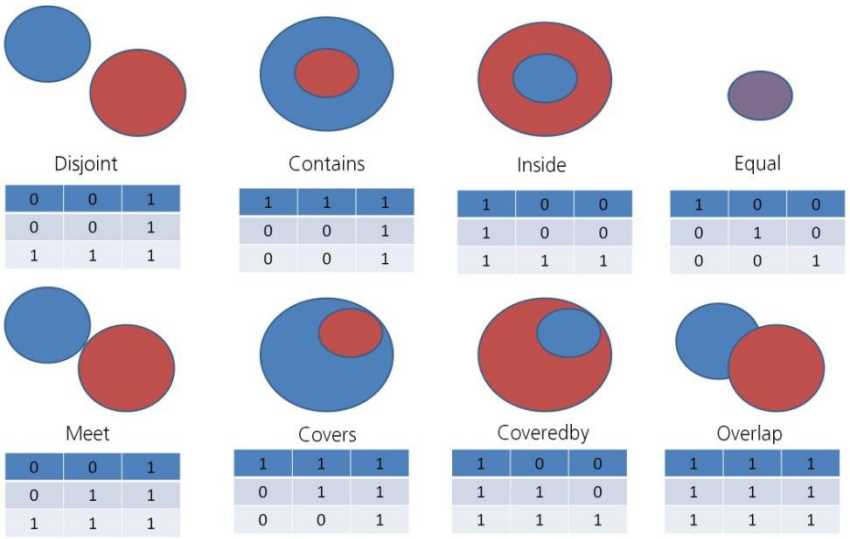



Abbildung 3.2: 9-Intersection Modell (nach Egenhofer 1991)

Das 9-Intersection Modell ist ein Konzept zur Klassifikation von topologischen Beziehungen. Allerdings hat es auch in einigen Fällen seine Grenzen. So wird keine Unterscheidung zwischen Mehrfachberührungen und Überschneidungen vorgenommen. In Abbildung 3.3 sind auf der rechten Seite zwei Beispiele von Mehrfachberührungen dargestellt. Sie haben die gleiche Klassifikation wie die Überdeckung, sind jedoch in der Anzahl ihrer räumlichen Berührungen unterschiedlich.

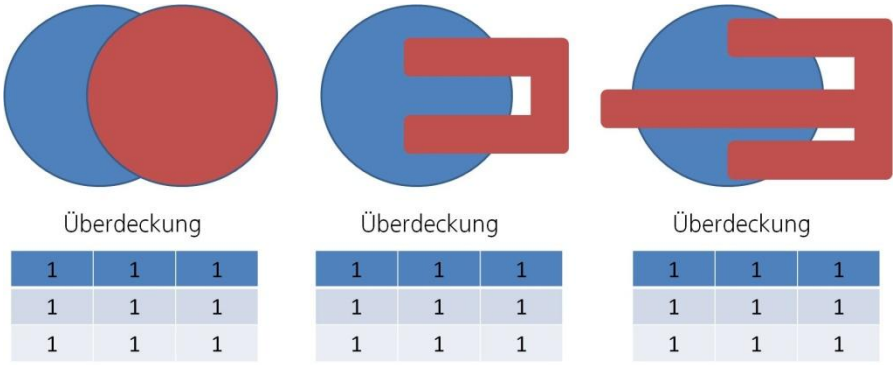



Abbildung 3.3: Mehrfachüberschneidungen beim 9-Intersection Modell (nach Bartelme 2005)

Topologische Beziehungen sind im Anwendungskontext einer der zentralen Punkte für die Berechnung von Plakatpassagen. Es ist wichtig zu wissen, ob eine Person durch den Sichtbarkeitsraum gekommen ist, ihn evtl. sogar mehrfach berührt hat, oder nicht. Aber auch die Zuordnung von Probanden der GPS-Stichprobe zu einer Gemeinde oder einem bestimmten Stadtviertel sind im Kontext der Arbeit notwendig. Egenhofer liefert an dieser Stelle die theoretischen Grundlagen für die Arbeitsschritte.

### 3.1.3 Geometrische Relation

Zu den wichtigsten Fragen in der räumlichen Analyse zählt neben den topologischen Beziehungen die Frage nach Distanzen und Entfernungen zwischen Objekten:

- Wie weit ist ein Objekt entfernt?
- Wie viele Personen erreicht man in einer Entfernung von 500 Metern?
- Wie komme ich am schnellsten von Standort A zu Standort B?

Zu den häufigsten Aufgaben gehört, die Distanz zwischen zwei Punkten $a, b \in \mathbb{R}^2$ zu berechnen. Im kartesischen Koordinatensystem kann der Abstand über die euklidische Distanz wie folgt bestimmt werden:

$$dist(a,b) = \sqrt{\sum_{i=1}^{n=2} (a_i - b_i)^2}$$

Schwieriger wird die Berechnung bei zwei Flächen oder bei zwei Straßensegmenten. Hier ist die Entscheidung zu treffen, ob für die Distanzberechnung die kürzeste Entfernung oder die Entfernung der Zentroide berechnet wird. In Abbildung 3.4 ist beispielhaft für zwei Polygone der Unterschied zwischen einer Distanzberechnung auf Basis der kürzesten Entfernung zwischen den Rändern dargestellt (a) und der Berechnung auf Basis von Zentroiden (b).

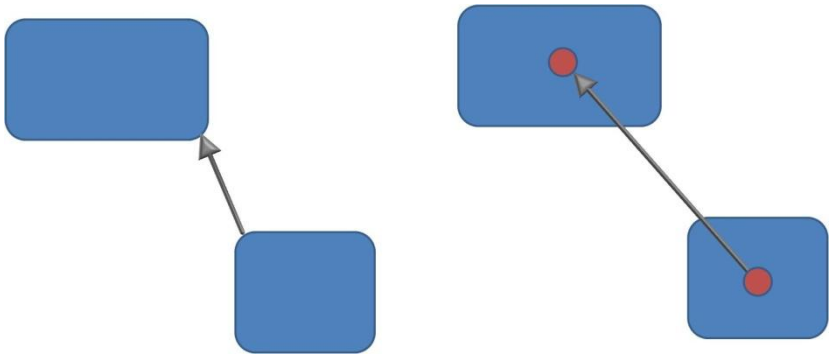


Abbildung 3.4: Distanz zwischen Objekten: kürzeste Entfernung zwischen 2 Rändern (a), Entfernung zwischen 2 Zentroiden (b)

Eine weitere wichtige Aufgabenstellung in der Distanzberechnung ist die Bestimmung von Erreichbarkeiten in Netzwerken. Wo ist z.B. der beste Standort für ein neues Einzelhandelsgeschäft mit den meisten potenziellen Kunden im Umfeld? Wo erreiche ich mit einer neuen Bushaltestelle die meisten Anwohner? Was ist der schnellste Weg zwischen zwei Punkten? Hierbei kann man zwischen zwei Möglichkeiten der Berechnung unterscheiden:

1. Welches Einzugsgebiet hat ein Einzelhandelsgeschäft in 5 Minuten Fahrzeit (Erreichbarkeit),
2. Kürzeste Wege zwischen zwei Knoten (Routing).

In Abbildung 3.5 sind beide Möglichkeiten dargestellt. Im linken Bild wird das Einzugsgebiet um einen Punkt berechnet. Im rechten Bild wird der kürzeste Weg zwischen zwei Punkten berechnet. An dieser Stelle hat das Wort „kurz“ zwei unterschiedliche Bedeutungen haben:

1. Die Länge des Weges.
2. Die benötigte Zeit des Weges.

Um diese Berechnungen durchzuführen, kommt dem zugrundeliegenden Straßennetz eine große Bedeutung zu. Die einzelnen Straßensegmente erhalten sogenannte Kantengewichte, die zur Berechnung herangezogen werden. Je schneller auf einem Segment gefahren werden kann und darf, desto „kostengünstiger“ wird es für die Berechnung.

Abbildung 3.5: Einzugsgebietsberechnung und kürzeste Wegberechnung zwischen zwei Punkten

Attribute, wie die erlaubte oder tatsächliche gefahrene Geschwindigkeit auf einem Straßensegment sind bei der Berechnung eines Einzugsgebietes oder der kürzesten Distanz zwischen zwei Punkten im Straßennetz zu berücksichtigen, denn sie geben darüber Aufschluss, welcher Weg der effektivste ist (vgl. Abschnitt 3.5.1). Um die kürzesten Wege in einem Graphen zu bestimmen, stehen unterschiedliche Algorithmen zur Verfügung. Zu den bekanntesten Algorithmen zählen der $A^*$ Algorithmus (Hart et al. 1968) und der Dijkstra Algorithmus (Dijkstra 1959).

Der Dijkstra Algorithmus, auch „kürzeste Wege" Algorithmus, baut vom Startpunkt bis zum Endpunkt einen sogenannten kürzesten Wegebaum auf (Cormen et al. 2004). Hierzu wird ausgehend vom Startpunkt der jeweils kürzeste Weg zum nächsten Nachbarn und danach wieder zum nächsten Nachbarn über einen Graph aufgebaut. Dieses Vorgehen wird als Greedy-Strategie bezeichnet. Dabei wird der jeweils kürzeste Weg abgespeichert. Aus der Menge der kürzesten Wege ergibt sich der Weg zwischen Start- und Zielknoten. Der Algorithmus ist beendet, sobald der Endpunkt erreicht worden ist. Trotzdem zählt der Dijkstra Algorithmus zu den sehr rechenaufwändigen Verfahren. Da im ersten Schritt der komplette Suchraum an Möglichkeiten untersucht wird, ist die Laufzeit bei großen Netzwerken sehr hoch. So nimmt die Rechenzeit quadratisch mit der Anzahl der Knoten zu. In Abbildung 3.6 ist das typische Vorgehen beim Dijkstra Algorithmus dargestellt.

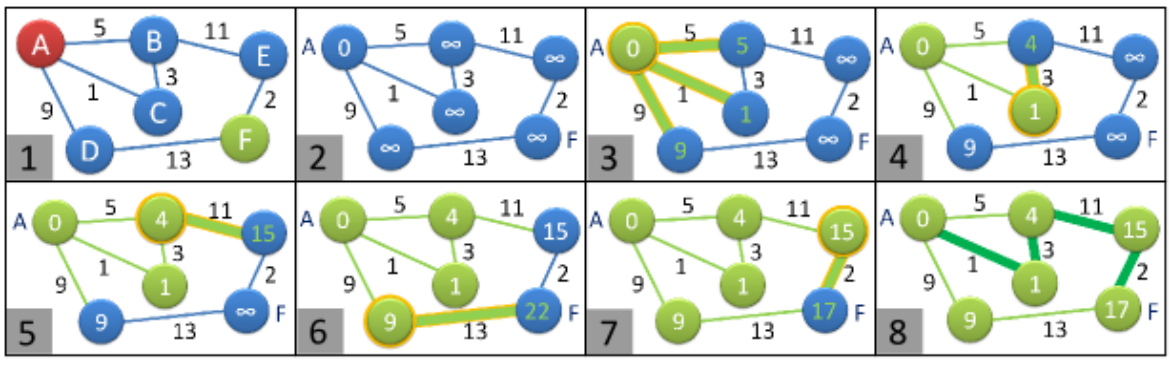



Abbildung 3.6: Vorgehen des Dijkstra Algorithmus

Ausgehend vom Startknoten (A) wird der kürzeste Weg zu jedem anderen Knoten oder zu einem zuvor definierten Zielknoten (F) berechnet. Initial wird der Startknoten mit der Wegstrecke 0 belegt, und alle anderen Knoten erhalten den Wert „unendlich". Zusätzlich erhalten alle Knoten den Status „unbesucht", welcher angibt, dass von einem Knoten aus

noch keine Berechnung zu den Nachbarknoten stattgefunden hat. Der Algorithmus wählt iterativ aus den noch unbesuchten Knoten den Knoten mit der niedrigsten Wegstrecke als aktuellen Knoten aus und berechnet für all seine unbesuchten Nachbarknoten die Wegstrecke. Diese berechnet sich je Nachbar aus der Summe der Wegstrecke des ausgewählten Knotens und dem Kantengewicht zum Nachbarknoten. Sofern die aktuelle Wegstrecke eines Nachbarknoten größer ist als der neu berechnete Wert, so wird diese ersetzt. Wenn alle Nachbarknoten abgearbeitet sind, erhält der aktuelle Knoten den Status „besucht" und verschwindet damit aus der Liste der unbesuchten Knoten. Die nächste Iteration startet, sofern der Zielknoten noch nicht unbesuchter Knoten mit minimaler Wegstrecke oder der Graph vollständig durchlaufen ist. Im Folgenden werden die Schritte 3-7 aus der Abbildung 3.6 im Detail beschrieben:

Zunächst wird in Bild 3 der Startknoten A ausgewählt. Dieser hat die drei unbesuchten Nachbarknoten B, C und D, welche jeweils die Wegstrecke ∞ besitzen. Die Wegstrecke von A ist 0, und somit sind die neuen Wegstrecken 5, 1 und 9. In Bild 4 ist aus der Liste der noch unbesuchten Knoten C mit einer Wegstrecke von 1 der niedrigste und wird daher ausgewählt. C hat mit B nur einen unbesuchten Nachbarknoten. Für diesen berechnet sich die Wegstrecke aus der Wegstrecke 1 von Knoten C und dem Kantengewicht 3. Die aktuelle Wegstrecke von B beträgt 5, da 4 kleiner ist, wird die Wegstrecke verändert. Nachdem alle unbesuchten Nachbarknoten abgearbeitet sind, erhält Knoten C den Status „besucht". Im nächsten Bild 5 ist der noch unbesuchte Knoten B mit einer Wegstrecke von 4 der niedrigste und wird daher ausgewählt. B hat mit E nur einen unbesuchten Nachbarknoten. Für diesen berechnet sich die Wegstrecke aus der Wegstrecke 4 von Knoten B und dem Kantengewicht 11. Die aktuelle Wegstrecke von B beträgt ∞, da 15 kleiner ist, wird die Wegstrecke verändert. Nachdem alle unbesuchten Nachbarknoten abgearbeitet sind, erhält Knoten B den Status „besucht". Der Knoten D ist in Bild 6 mit einer Wegstrecke von 9 der niedrigste und wird als nächster ausgewählt. D hat mit F nur einen unbesuchten Nachbarknoten. Für diesen berechnet sich die Wegstrecke aus der Wegstrecke 9 von Knoten D und dem Kantengewicht 13. Die aktuelle Wegstrecke von F beträgt ∞, da 22 kleiner ist, wird die Wegstrecke verändert. Nachdem alle unbesuchten Nachbarknoten abgearbeitet sind, erhält Knoten D den Status „besucht". In Bild 7 ist E mit einer Wegstrecke von 15 der niedrigste. E hat mit F nur einen unbesuchten Nachbarknoten. Für diesen berechnet sich die Wegstrecke aus der Wegstrecke 15 von Knoten D und dem Kantengewicht 2. Die aktuelle Wegstrecke von F beträgt 22, da 17 kleiner ist, wird die Wegstrecke verändert. Der nächste kleinste unbesuchte Knoten ist in Bild 8 der Zielknoten F ohne weitere unbesuchte Nachbarknoten. Der Algorithmus hat den kürzesten Weg vom Startknoten A zum Zielknoten F berechnet und stoppt. In diesem Fall entspricht dies gleichzeitig dem vollständigen kürzeste Wege Graphen, da alle Knoten besucht wurden und vom Startknoten zu jedem beliebigen anderen Knoten der kürzeste Weg berechnet ist.

Eine Erweiterung vom Dijkstra Algorithmus stellt der $A^*$ Algorithmus dar. Er ist inzwischen einer der am häufigsten eingesetzten Routenalgorithmen, da er insbesondere bei der Rechenlaufzeit Vorteile bietet. Der $A^*$ Algorithmus benutzt zur Routenberechnung eine Schätzfunktion, die für jeden Knoten eine untere Schranke für den noch zu erwartenden Abstand zum Zielpunkt liefert. Damit unterscheidet sich der $A^*$ Algorithmus vom Dijkstra Algorithmus, indem er Informationen, wie weit man noch vom Zielpunkt entfernt ist, ausnutzt. Knoten, die wahrscheinlicher zum Ziel führen, werden somit zuerst untersucht. Die Suche erfolgt also zielgerichteter, was z.B. die Laufzeiten bei großen Datenmengen effizienter macht. Um den vielversprechendsten Knoten zu ermitteln, wird allen bekannten Knoten $k$ jeweils ein Wert $f(x)$ über eine Schätzfunktion zugeordnet, die angibt, wie lange ein Weg vom Start zum Ziel unter Verwendung des betrachteten Knotens im günstigsten Fall ist. Der Knoten mit dem niedrigsten f Wert wird als nächstes untersucht.

$$f(x) = k(x) + t(x)$$

Dabei stellt $k(x)$ die Summe der bisherigen Kosten vom Startknoten aus dar und $t(x)$ bezeichnet die geschätzten Kosten bis zum ausgewählten Zielknoten. Für die Schätzfunktion werden an dieser Stelle Heuristiken angenommen. Im Falle des $A^*$ Algorithmus können also eine Reihe von möglichen Kanten ausgeschlossen werden, die Berechnung wird somit effizienter.

Im Kontext der Arbeit spielen Wegealgorithmen insbesondere bei der Aufbereitung von GPS-Daten eine Rolle, so können diese dazu dienen, einzelne GPS-Punkte über ein Routing miteinander zu verbinden.

### 3.1.4 Geometrischer Lagevergleich

Zu den geometrischen Grundaufgaben zählt weiter der geometrische Lagevergleich. Dieser Vergleich dient dazu, die relative Lage von Punkten, Linien und Polygonen zueinander zu beschreiben. Die Aufgabe ist z.B., ein Auto in Bezug zu einem Plakatstandort zu verorten. Man kann sagen, „das Auto steht am Plakatstandort", „es steht weit entfernt vom Plakat", „das Auto steht nahe beim Plakat". Wie die Beispiele zeigen, kann man das Auto mit vielen unterschiedlichen Möglichkeiten verorten. Um zu bestimmen, wie die Ausrichtung eines Ausgangsobjektes zu einem Referenzobjekt ist, ist ein Bezugsrahmen wichtig (Hernandez, 1994). Hierfür sind drei Bezugsrahmen geläufig (Levinson 2003):

1. Absoluter Bezugsrahmen: Man wählt unveränderliche, von jeglicher Situation unabhängige Bezugspunkte. Im klarsten Falle sind das die vier Himmelsrichtungen.
2. Deiktischer oder relativer Bezugsrahmen: Der Sprecher (das deiktische Zentrum) nimmt sich selbst nicht nur als Bezugspunkt für die rein topologische Orientierung, sondern er macht auch noch Gebrauch von der Tatsache, dass er ein strukturiertes physikalisches Objekt ist. Er hat nämlich eine Vorder- und eine Rückseite, eine linke und eine rechte Seite (sowie eine Ober- und eine Unterseite, auf die wir jedoch erst später zurückkommen).
3. Intrinsischer Bezugsrahmen: Viele Relationen haben selbst räumliche Struktur. Eine Kirche z.B. hat eine Vorder- und eine Rückseite und somit auch eine linke und eine rechte Seite.

Im vorliegenden Anwendungskontext ist der geometrische Lagevergleich mit relativen oder intrinsischen Bezugsrahmen eine wichtige Komponente, denn es ist wichtig zu unterscheiden, ob ein GPS-Proband z.B. frontal oder seitlich an einem Plakat vorbeikommt. Die relative Lage hat Einfluss auf die Gewichtung des Kontaktes. In der Regel gilt, dass ein direkter frontaler Kontakt besser gewichtet wird als ein paralleler, bzw. seitlicher Kontakt.

### 3.1.5 Räumliche Aggregation

Die räumliche Aggregation ist ein wichtiger Bestandteil in der Aufbereitung von Geodaten. Sie dient dazu, Informationen zusammenzufassen oder Informationen eines einzelnen Messpunktes auf eine Fläche zu verteilen. In Abhängigkeit von der Fragestellung können die Aggregationseinheiten sehr unterschiedliche Form und Größe haben. Zu den gängigsten räumlichen Aggregationen zählen administrative Einheiten wie Kreise, Gemeinden oder Postleitzahlgebiete. Weitere Möglichkeiten der räumlichen Aggregation stellen Pufferzonen oder Voronoi Polygone dar. Das Konzept von Voronoi Polygonen, auch Thiessen Polygone genannt, basiert auf der Idee, dass ein gegebener Raum durch den Einfluss einzelner Punkte oder Objekte in abgegrenzte Polygone aufgeteilt wird (Klein 2005). Sei $P = \{p_1, p_2, \dots p_n\}$ eine Menge von Punkten in einem 2-dimensionalen Raum. Diese Punkte werden Orte genannt

(Bartelme 2005). Zerteilt man die Fläche, indem man deren Punkte zu ihrem nächsten Ort $p_i$ zuordnet (z.B. über die euklidische Distanz), entstehen zu jedem Ort Voronoi-Regionen $V(p_i)$:

$$V(p_i) = \{x: |p_i - x \leq p_j - x \mid \text{für } i \neq j, j = 1 \dots n\}$$

Manche Punkte können mehr als einem Ort (oder auch nächsten Nachbarn) zugeordnet werden. Die Menge aller dieser Punkte bildet das Voronoi-Diagramm V(P) für die Menge der Orte. Die so entstandenen Polygon-Flächen partitionieren das gesamte Untersuchungsgebiet, ohne sich zu überschneiden. Als Grundannahme gilt, dass der Messwert des Ortes in die Fläche übertragbar ist. Die Voronois bilden insofern eine Näherung an den beobachteten Wert, an den Polygongrenzen treten Wertesprünge auf. Fließende Übergänge zwischen Polygonen werden mit dieser Methode nicht abgebildet.

Bei der Erzeugung von Pufferzonen wird um ausgewählte Geoobjekte eine Fläche generiert. Diese Geoobjekte können Punkte, Linien oder Flächen sein. Die Pufferzonen umschließen das Geoobjekt und das umliegende Gebiet innerhalb eines bestimmten Abstandes vom Geoobjekt. Dies kann ein fixer Distanzwert sein oder in Abhängigkeit von den Attributen des Geoobjektes geschehen. Die ursprünglichen Geoobjekte werden bei diesem Vorgang nicht verändert. Ein Nachteil der Pufferzonen ist, dass natürliche Barrieren wie z.B. Flüsse und Höhenzüge nicht beachtet werden.

Für den Anwendungskontext kommen Aggregationen bei der Zusammenfassung von GPS-Probanden für Einzugsgebiete und Pufferzonen (Hecker, 2010c) bei der Bildung von Plakatsichtbarkeitsräumen zum Einsatz.

## 3.2 Mobilitätsdaten

Die Mobilität macht einen großen Teil unserer täglichen Aktivitäten aus, dabei hinterlassen wir immer mehr digitale Spuren im Raum. Einhergehend mit der Verbesserung und Verbreitung von GPS- und anderen Ortungstechnologien hat sich in der Geographie, der Geoinformatik und der Informatik als Forschungsthema entwickelt, das sich mit der Speicherung, Verwaltung, Bearbeitung und Analyse großer Datensätze von Positionsdaten sich bewegender Objekten beschäftigt. Moderne Geräte können inzwischen Positionsdaten in hoher Präzision, kurzen Intervallen und über größere Zeiträume erfassen. Durch die Erfassung werden räumliche Bezugspunkte und Beziehungsnetzwerke sichtbar, die mittels Trajektorien miteinander verbunden werden. Andrienko et al. (2008b: 18) definiert eine Trajektorie wie folgt:

> *„Eine Trajektorie ist ein Pfad, den ein Objekt im Raum, in dem es sich bewegt, zurücklegt."*

Trajektorien $T$ werden durch eine Menge von Koordinaten $(x, y)$, die jeweils zu einem Zeitpunkt $(t)$ gemessen wurden, repräsentiert T=$((x_i, y_i, t_i)|i = 1 \dots n))$.

Die Erfassung einer Trajektorie besteht in der Regel aus folgenden Daten:

- Eine Identifikation des Objekts, damit die Messung dem entsprechenden Objekt zugewiesen werden kann.
- Eine Position, im Normalfall bestehend aus einer x- und einer y-Koordinate.
- Eine Trajektorie besteht aus mehreren Positionen zu einem Zeitpunkt t. Ein Zeitpunkt kommt pro Trajektorie nur einmal vor.

Trajektorien haben einen zeitlichen Anfang und ein (zumindest vorläufiges) Ende. Dabei ist eine Trajektorie kontinuierlich, doch kann sie nicht kontinuierlich erfasst werden. Aufgrund der Technik können nicht alle exakten Positionen gemessen und gespeichert werden. Aus diesem Grund werden Trajektorien diskretisiert, d.h. Trajektorien werden als Sequenzen gemessen, mit einem variablen oder festen zeitlichen Abstand in der Reihenfolge zwischen den einzelnen Messungen. Eine Trajektorie ist also eine Sequenz von Positionen, geordnet in der Reihenfolge, in welcher sie besucht wurden (Andrienko et al. 2008b: 29). Je mehr Messpunkte ein Zeitintervall beinhaltet, desto genauer kann die zeitliche Auflösung einer Bewegung abgebildet werden. Mit den Zeitpunkten der Messungen lassen sich die Dauer und die Geschwindigkeit bestimmen. Über die Kombination von Zeit, Position und Beschleunigungssensor können Geschwindigkeit und Beschleunigung des Objektes bestimmt werden. Aus aufeinander folgenden Positionsmessungen lassen sich bestimmte Charakteristiken/Eigenschaften der Trajektorie ableiten. Andrienko et al. (2008b) und Dodge et al. (2008) unterscheiden Charakteristiken von Trajektorien in zwei Kategorien: eine zeitpunktbezogene und eine gesamtheitliche Charakteristik. Die zeitpunktbezogenen Charakteristiken können für jede einzelne Trajektorie analysiert werden, während die gesamtheitlichen Charakteristiken ein Trajektorienintervall beschreiben. Beispiele für zeitpunktbezogene Charakteristiken geben Andrienko et al. (2008b):

- Räumliche Position,
- Zeit,
- Richtung und Richtungswechsel,
- Geschwindigkeit,
- Reisezeit und

- zurückgelegte Strecke.

Für die gesamtheitlichen Charakteristiken geben Andrienko et al. (2008b) folgende Beispiele für Trajektorien:

- Geometrie,
- Länge und Dauer,
- Start und Ziel der Trajektorie,
- minimale und maximale Geschwindigkeit,
- durchschnittliche Geschwindigkeit,
- Anzahl Stopps.

Anfang und Ende einer Trajektorie sind gewöhnlich Anfang und Ende eines Weges. Ein Tag besteht somit in der Regel aus mehreren Trajektorien. Trajektorien können aber auch als Reisen definiert werden. Ein Beispiel dafür ist die jährliche Migration von Zugvögeln, die per GPS überwacht werden (Spaccapietra et al. 2008). Anfang und Ende einer solchen Trajektorie ist dann Start und Ende einer solchen Reise. In den folgenden Abschnitten werden Charakteristiken von Trajektorien, Distanzberechnungen, die topologischen Beziehungen bei Trajektorien sowie die Möglichkeiten der Mobilitätserfassung vorgestellt.

### 3.2.1 Topologische Beziehungen bei Trajektorien

Topologische Beziehungen zwischen Trajektorien betrachten jeweils den Raum und die Zeit. Allen definiert 1984 insgesamt sieben zeitliche und räumliche Beziehungen für jeweils ein Zeitintervall. Ein Zeitintervall $x$ ist dabei definiert durch einen Startzeitpunkt und Endzeitpunkt. Der Startzeitpunkt wird mit $x^-$, der Endzeitpunkt mit $x^+$ bezeichnet.

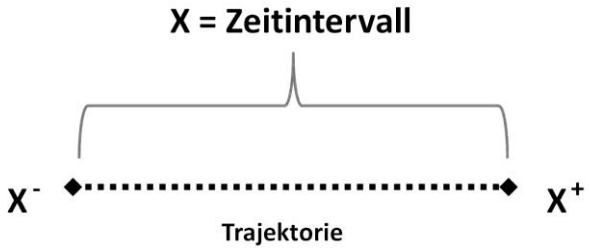



Abbildung 3.7: Zeitintervall Trajektorie

Für zwei Zeitintervalle können nun nach der Allenschen Zeitlogik sieben verschiedene Arten von Basisrelationen hergestellt werden (vgl. Abbildung 3.8).

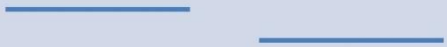
| Relation | Darstellung | Interpretation |
|---|---|---|
| X<Y |  | X ist vor Y |
| X m Y |  | X trifft Y |
| X o Y |  | X hat eine teilweise Überlappung mit Y |
| X s Y |  | X startet mit Y |
| X d Y |  | X ist in Y enthalten |
| X f Y |  | X endet mit Y |
| X = Y |  | X ist gleich Y |

Abbildung 3.8: Zeitrelationen von Zeitintervallen nach Allen (1984)

Wenn eine der beschriebenen Relationen bekannt ist, bedeutet dies, dass alle Informationen über eine Beziehung zwischen zwei Trajektorien vorliegen. Die Allensche Zeitlogik liefert hierfür das Gerüst zur Modellierung von zeitlichen Zusammenhängen.

### 3.2.2 Charakteristiken menschlicher Mobilität

Die Freiheit der menschlichen Mobilität ist eine der größten Errungenschaften, die einher geht mit der Entwicklung der öffentlichen Verkehrsmittel, der Entwicklung der Massenproduktion von Autos und der Öffnung von Grenzen. Nach dem Bundesamt für Statistik in der Schweiz (BfS 2011) legt ein Schweizer pro Jahr 19.000 Kilometer zurück, davon 10.600 Kilometer im Auto. 3600 entfallen auf den öffentlichen Verkehr (Bahn, Bus, Straßenbahn), und mit dem Flugzeug sind es noch einmal 3400 km. Die restlichen Kilometer entfallen auf Wegstrecken, die zu Fuß oder mit dem Fahrrad zurückgelegt werden. Durchschnittlich ist eine Person in der Schweiz pro Tag 37 Kilometer unterwegs. In Deutschland liegt die durchschnittliche Anzahl an Kilometern laut der Studie „Mobilität in Deutschland“ (MID 2008a) bei 41 Kilometer pro Tag. Die Mobilität nimmt stetig zu. So hat sich die Personenverkehrsleistung in der Schweiz zwischen 1970 und 2008 mehr als verdoppelt. Viele Menschen verbringen ihre Zeit an wenigen Orten, was auch nicht weiter verwundert, da sie in der Regel an einem bestimmten Ort wohnen, arbeiten oder zur Schule gehen. Das Abbilden der menschlichen Mobilität in Zeit und Raum ist bereits seit den 50er Jahren ein wichtiger Bestandteil der Geographie. Als Begründer der klassischen „Time Geographie“ gilt Hägerstrand von der Universität Lund in Schweden, der sich mit der Auswertung von Daten beschäftigte, die über die Bewegung von Individuen gesammelt wurden (vgl. Abschnitt 1.4). Bei der Analyse raum-zeitlicher Daten werden in der Time Geography bestimmte Einschränkungen und Zwänge des menschlichen Handelns beschrieben. Hägerstrand formulierte in „What about people in regional science“(1970) folgende „constraints“:

- **Capability Constraints:** „those which limit the activities of the individual because of his biological construction and/or the tools he can command.” (Seite 12)

- **Coupling Constraints:** „define where, when, and for how long the individual has to join other individuals, tools, and materials in order to produce, consume, and transact."(Seite 14)

- **Authority Constraints**: „refer to 'control areas' or 'domains'. A domain is a time-space entity within which things and events are under the control of a given individual or a given group."(Seite 16)

Eine der wesentlichen Charakteristiken, die Hägerstrand von diesen Constraints ableitet, ist die Regelmäßigkeit der menschlichen Mobilität. So sind die Interaktionen z.B. der Coupling Constraints auf eine bestimmte Personengruppe (Familie, Bekannte) und Orte (Einkauf, Arbeit, Schule) beschränkt und haben eine zeitliche Regelmäßigkeit.

In den letzten Jahren wurden einige Studien zur Mustererkennung der menschlichen Mobilität mit den z.T. beschriebenen Erfassungsmethoden durchgeführt. Schlich und Axhausen (2003) analysierten auf Basis einer sechswöchigen Tagebuchbefragung die Muster von Probanden. Kim et al. (2006) und McNett et al. (2005) untersuchten die Muster von Personen, die sich mit ihren Mobilfunktelefonen oder Laptops in Wireless-Lan Access Points eingeloggt haben, mit dem Nachteil, dass sie keine flächendeckende Abdeckung für einen großen geographischen Raum zur Verfügung stehen hatten und somit eher ein lückenhaftes Bild der Mobilität erfasst wurde. Barabasi (2010), Gonzalez et al. (2008) und Song et al. (2010) analysierten Mobilfunkdaten über mehrere Monate. Dabei wird eine Person immer nur dann auf Funkzellenebene lokalisiert, wenn sie telefoniert. Auch wenn die Datenquellen bei diesen Studien sehr unterschiedlich sind, so stellte sich heraus, dass die Personen während der Studien im Durchschnitt 60 (Schlich et al. 2003) bzw. 50 (Gonzalez et al. 2008) unterschiedliche Orte besucht haben. Dabei wurde ein Großteil der Aufenthaltsdauer an einigen wenigen Standorten verzeichnet. Schlich (2003) hielt fest, dass 70% aller Besuche auf 2-4 Standorte beschränkt sind und 90% der Besuche auf 8 Standorte. Zu einem ähnlichen Ergebnis kommt Barabasi (2010), der angibt, bei wenig reisenden Menschen zu 93% den Aufenthaltsort eines Handynutzers in einer Werktagwoche vorhersagen zu können. Datengrundlage hierfür waren Abrechnungsdaten von 50.000 Personen eines Mobilfunkanbieters in den USA, die bei jedem Gespräch die Funkzelle mit abgespeichert haben. In einem Spiegel Online Interview (Spiegel, 2011) sagt Barabasi: „Wohin wir im Laufe eines Tages und der Woche fahren, ist gut vorhersagbar". Dies liegt nach seiner Meinung zum größten Teil an Routinen: „Acht Stunden im Büro, einkaufen, zehn zu Hause, da bleibt nicht mehr viel Spielraum." Am freien Willen des Menschen zweifelt Barabasi trotzdem nicht: „Wir benutzen ihn nur nicht, wenn wir uns bewegen. Es gibt viele Zwänge, und wir machen immer das Gleiche." Bei den Aussagen von Barabasi muss man allerdings Vorsicht walten lassen. So handelt es sich bei den aufgezeichneten und vorhergesagten Standorten um Orte, an denen die Personen telefoniert haben. Diese Anzahl der Orte ist nicht vergleichbar mit der tatsächlich besuchten Anzahl an Orten einer Person. Dies mit der menschlichen Mobilität gleichzusetzen ist falsch! Trotzdem kann man festhalten, dass eine der Charakteristiken des menschlichen Mobilitätsverhaltens seine Periodizität und seine Beschränkung auf wenige besuchte Orte ist.

### 3.2.3 Annotation von Trajektorien

Mobile Geräte mit eingebautem GPS-Empfänger erzeugen massive Mengen von Geopositionen mit Informationen zur Zeit und Geschwindigkeit des Objektes. Allerdings fehlen bei den aufgezeichneten GPS-Trajektorien semantische Informationen. Die Fragen, die offen bleiben, sind u.a.: Wo befindet sich der Wohnort und der Arbeitsplatz der Person. Hierzu müssen zuerst Stopps und Bewegungen aus der aufgezeichneten Trajektorie voneinander unterschieden werden. Im Anschluss geht es darum, Orte zu identifizieren und

diese semantisch aufzufüllen, z.B. der Besuch eines Einkaufscenters, eines Restaurants oder anderer Sehenswürdigkeiten, anstatt nur zu identifizieren, dass geografische Koordinaten über einen längeren Zeitpunkt dort aufgetreten sind. Darüber hinaus ist der Zweck eines bestimmten Weges (Arbeitsweg, Freizeit) unbekannt, ebenso das gewählte Fortbewegungsmittel (PKW, Fußgänger, ÖPNV). Eine automatische semantische Extraktion von aussagekräftigen Standort- und Wegezuordnungen ist ein Thema, mit dem sich bereits einige Forschungsgruppen beschäftigt haben. Wolf und Wolf et al. (Wolf, 2000; Wolf, 2001) haben sich bspw. mit dem Thema auseinandergesetzt, traditionelle Reisetagebücher mit GPS-Aufzeichnungen zu ersetzen mit dem Fokus, den Reisezweck zu identifizieren. In einem ersten Schritt wurden getrennte Zeitintervalle (Wege) einer Person identifiziert und damit der Start und das Ende eines Weges. Diese Starts und Enden wurden mit der Landnutzung räumlich verschnitten. Eine zuvor definierte Beziehung zwischen Landnutzung und Zweck der Fahrt, zusammen mit Ausflugsziel und weiteren Informationen wie Ankunftszeit, wurden dann zur Ableitung des Zwecks der Reise genutzt.

Ein ähnlicher Ansatz wird von Axhausen et al. (2003) verwendet. Auch hier wird der Zweck der Fahrt über Landnutzungsinformationen und zusätzliche persönliche Informationen, wie Wohn- und Arbeitsplatz, Sportverein usw. des Reisenden identifiziert. Die vorgestellten Ansätze sind in mehrfacher Hinsicht begrenzt. Bei beiden Ansätzen werden nur wenige Reisekategorien voneinander unterschieden. Zusätzlich kommt noch ein grundlegendes Problem hinzu, die Mehrdeutigkeit von bestimmten Orten. So kann zum Beispiel ein Einkaufszentrum mit einer umfangreichen Ladenstruktur, wie Supermärkten, Ärzten, Fitnesscentern und Kino nicht eindeutig einer bestimmten Aktivität und damit dem Zweck zugeordnet werden. Darüber hinaus haben bestimmte Wege nicht direkt ein Ziel, das zugeordnet werden kann. So sind sonntägliche Spaziergänge, an denen der Start- und Zielort identisch ist, schwer zu klassifizieren.

Yan et al. (2011a) haben 2011 ein automatisiertes System zur Annotation vorgestellt. In einem nachfolgenden Schritt wurde eine Online-Annotation von Trajektorien vorgestellt (Yan et al. 2011b).

In Abbildung 3.9 werden die einzelnen Schritte der semantischen Annotation von Trajektorien dargestellt. Von oben beginnend werden in der Abbildung immer weitere Informationen erkannt und mit der Trajektorie verknüpft. Zu Beginn steht die Rohtrajektorie mit einer Menge von GPS-Punkten. In unregelmäßigen Abständen treten immer wieder Punktwolken, sogenannte Oszillationen, auf. Die Oszillationen geben einen Hinweis darauf, ob eine Person längere Zeit an einem Ort verweilt. Eine Klassifikation nach Bewegungen und Stopps ist eine der ersten semantischen Annotationen von Trajektorien. Im Anschluss werden mithilfe einer räumlichen Verschneidung die Stopps bestimmten Orten wie Wohnort, Arbeitsstätte und Einkauf zugeordnet. Im nächsten Schritt wird die Fortbewegung zwischen den Orten näher analysiert und den jeweiligen Fortbewegungsmitteln PKW, Fußgänger und ÖPNV zugewiesen. Gemeinsam mit der Ortserkennung wird den Wegen ein Reisezweck zugeordnet.

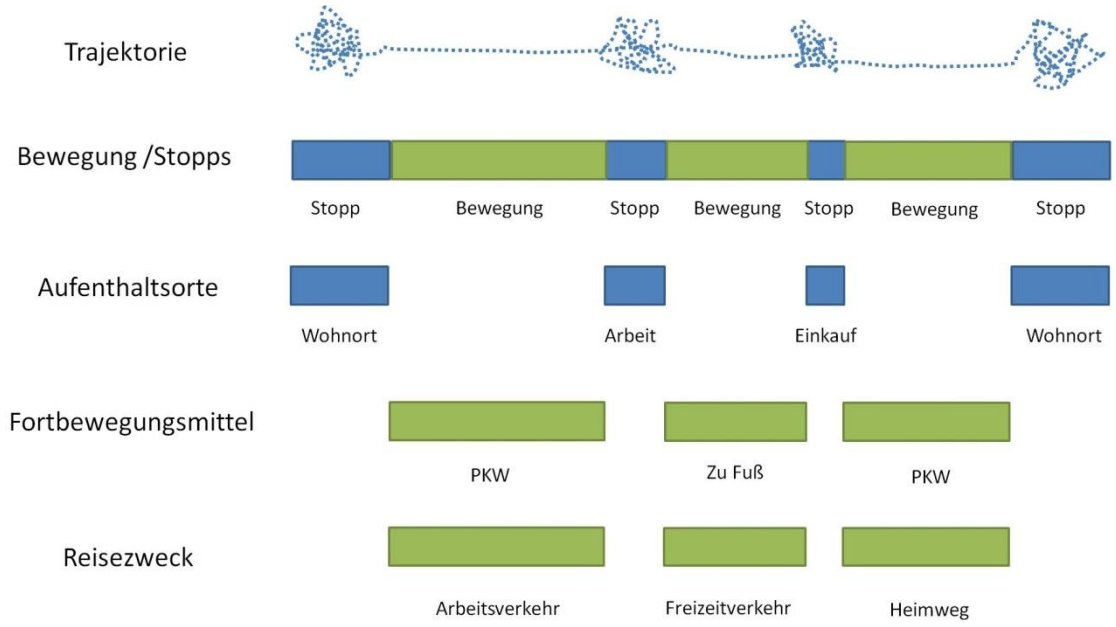



Abbildung 3.9: Annotation von Trajektorien (Körner 2012)

Um eine automatische Annotation von Trajektorien durchzuführen, ist eine angemessene Menge von Trainingsdaten notwendig. Hierzu werden in der Regel Tagebuchaufzeichnungen und GPS-Erfassungen parallel durchgeführt und Reisezweck, Verkehrsmittel und Aufenthaltsorte festgehalten. Einige Studien haben jedoch den Widerspruch zwischen GPS-Trajektorien und Reisetagebüchern aufgezeigt (Stopher 2007, Zmund 2003). Probanden vergessen häufig einzelne Wege. Dadurch hat insbesondere die Zuordnung des Verkehrsmittels zu häufigen Zuordnungsfehlern geführt. Mithilfe von Karten, den aufgezeichneten GPS-Trajektorien und des zeitlichen Kontextes ist es für den Probanden häufig einfacher, sich an die jeweiligen Aktivitäten zu erinnern und so eine semantische Zuordnung manuell durchzuführen, als mithilfe von Tagebüchern. Die semantische Annotation von Trajektorien ist weiterhin noch ein aktives Forschungsthema.

### 3.2.4 Analyse von Trajektorien

Die Bewegungen von mobilen Objekten und deren Interaktion mit ihrem Umfeld können mit Verfahren des Mobility Mining analysiert werden (Giannotti und Pedreschi 2008). Die analysierten Daten können Ergebnisse aus Telefonbefragungen, aus GPS-, oder auch aus Mobilfunkdaten sein. Dabei ist das Mobility Mining stark verwandt mit den Themen Spatial Data Mining und Spatiotemporal Data Mininig, denn insbesondere die z.T. sehr großen Mengen an Daten bzw. Koordinaten müssen in der Regel mit maschinellen Methoden analysiert werden, da eine manuelle Bearbeitung nicht mehr möglich ist. Im folgenden Abschnitt werden häufig eingesetzte Verfahren des Mobility Mining vorgestellt.

**Mustererkennung**

Die Detektion von häufigen Mustern in Trajektorien untersucht Bewegungsmuster von Personengruppen und einzelnen Probanden. Dabei unterteilt man die Mustererkennung bei Trajektorien in zwei Bereiche: 1. Das Erkennen von Mustern, 2. Das Erkennen von Instanzen, gegeben ein Muster. Im ersten Fall wird untersucht, wie häufig eine Gruppe oder eine Person z.B. an Ort A, dann Ort B und schließlich Ort C vorbeikommt. Im zweiten Fall wird untersucht, wann und wo ein spezifisches Muster auftritt und wer daran teilnimmt.

Erkennen von Mustern

Bei der Suche nach Wegemustern kann man unterscheiden zwischen der Suche nach raum-zeitlichen Mustern oder nach der Suche, die rein auf den Raum beschränkt ist. Reduziert man die Suche nur auf den Raum, so ist die Abfolge der besuchten Orte ein Muster, das zu finden ist. Aufgrund der begrenzten Batterielaufzeiten von mobilen Geräten und die z.T. unzuverlässige Übertragung oder Aufzeichnung von Koordinaten können Vorhersagen über die nächste Position auf Grundlage der bisherigen Standorte ein probates Mittel sein (Yang und Hu 2006). Nimmt man als weitere Komponente die Zeit hinzu, so wird nicht nur wichtig, ob ein bestimmter Ort besucht wurde, sondern auch zu welchem Zeitpunkt. Giannotti et al. (2006) haben hierzu das Konzept der „temporally annotaded sequences" (TAS) vorgestellt. Bei TAS unterscheidet man zwischen Sequenzen von Positionen und den zeitlichen Sequenzen zwischen diesen Postionen. Das von Giannotti vorgestellte Verfahren basiert dabei auf einer Nachbarschaftsfunktion, um raum-zeitliche Muster zu finden. Dabei ist eine der besonderen Herausforderungen der Umgang mit GPS-Daten. In der Regel wird es kaum vorkommen, dass bei einer Stichprobe an Personen die GPS-Koordinaten identisch oder fast identisch sind. Aus diesem Grund müssen die gesammelten Trajektorien generalisiert oder diskretisiert werden. Gianotti et. al (2007) haben hierzu die Trajektorien bestimmten Polygonen/Regionen zugeordnet. Cao et. al. (2005) und Kang and Yong (2010) stellten Generalisierungen mithilfe einer Gridstruktur dar. Zur Bestimmung der besuchten Orte können entweder externe Datenquellen wie POI-Informationen herangezogen werden, oder man nimmt häufig besuchte Koordinaten der Trajektorien und puffert mit einem bestimmten Umkreis (Hecker 2010c).

Instanzen von Mustern

Damit Muster detektiert werden können, muss in einem ersten Schritt festgelegt werden, nach welchen Mustern gesucht werden soll. In der Literatur am häufigsten genannt ist die Suche nach Gruppenmustern. Gruppenmuster beschreiben das Verhalten von Objekten, die einem bestimmten kollektiven Muster folgen. Diese Objekte haben sowohl einen sehr engen räumlichen Bezug in einem Zeitintervall ihrer Bewegung. Wang et al. (2003) stellte hierzu einen Algorithmus vor, der abhängig von einem räumlichen und zeitlichen Grenzwert die räumlich und zeitlich nächsten k-Trajektorien identifiziert. Der sogenannte k-Gruppen Algorithmus findet Muster, die auf Daten beruhen, die periodisch verteilte Punkte in der Zeit darstellen. Eine Methode, die aperiodische Muster für Trajektorien auf Straßenzügen untersucht, wurde von Hwang et al. 2005 vorgestellt. Im Unterschied zur Methode der raum-zeitlichen Nachbarschaft von Wang erkennt diese Methode auch Gruppen nach ihrer internen Struktur. Hierzu muss kein direkter räumlicher Zusammenhang bestehen. Zusätzlich zu der räumlichen Nähe können auch andere Faktoren zur Mustererkennung herangezogen werden. Zum Beispiel kann eine Gruppe von einem Objekt angeführt werden, das die Bewegungsrichtung der Gruppe antizipiert. Dieses Muster wurde 2002 von Laube und Imfeld vorgestellt und gehört zum Konzept der relativen Bewegungsmuster (Relative Motion – ReMo). Andere Bewegungsmuster stellen z.B. konvergierende und divergierende Muster dar. Beim Leadership folgen mehrere Trajektorien einer bestimmten Trajektorie, bei den konvergierenden Muster laufen Trajektorien an einem bestimmt Punkt zusammen, bei den divergierenden laufen die Trajektorien von einem bestimmten Punkt weg (Abbildung 3.10).

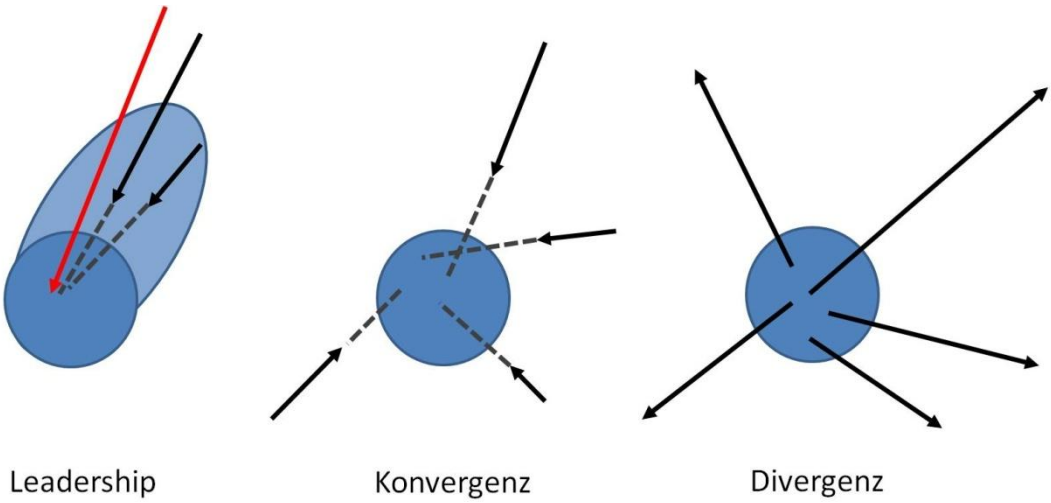



Abbildung 3.10: Relative Motion (nach Laube 2002)

**Clusterbildung**

Die Clusterbildung oder die Segmentierung von Trajektorien ist die Bildung von Gruppen mit einer ähnlichen Charakteristik. Dabei bezieht sich in der Regel die Clusterbildung auf einzelne Abschnitte einer Trajektorie und nicht auf den zeitlichen Bezugsrahmen eines einzelnen Tages oder einer gesamten Woche (Spaccapietra et al. 2008). Diese Abschnitte können semantisch interessante Bewegungen darstellen. Zur Identifizierung dieser Muster müssen Zeitintervalle festgelegt werden. Ein interessantes Muster sind z.B. Staus im Autoverkehr. Der Ort und die Geschwindigkeit bleiben für ein bestimmtes Zeitintervall annähernd konstant, während die Objekte sich regelmäßig ändern. Neue Fahrzeuge kommen hinzu, andere Fahrzeuge sind durch den Stau durch oder verlassen ihn über Abfahrten. Grundlage für die Identifizierung von Abschnitten einer Trajektorie ist die Erkennung von Starts und Stopps (siehe Abschnitt 3.2.2).

Aktuelle Verfahren der Clusterbildung setzen ihren Fokus auf eine möglichst sinnvolle Wahl der Cluster. Nanni und Pedreschi (2006) verwenden zur Clusterbildung dichte-basierte Verfahren für Trajektorien. Der Vorteil solcher Clusterverfahren liegt darin, dass sie sehr gut mit GPS-Datenpunkten umgehen können, sie robust gegenüber Rauschen in den Daten sind und keine Anzahl an Clusterklassen vorgeben. Häufig zum Einsatz kommt der von Ankerst et al. (1999) vorgestellte OPTICS Algorithmus, der diese drei wichtigen Eigenschaften zur Trajektorienanalyse mitbringt. Abhängig von den Fragestellungen können unterschiedliche Clusterkriterien bevorzugt werden. Z.B. können die räumliche Nähe, das Raum-Zeitliche, aber auch eine vorher durchgeführte semantische Annotation zur Clusterbildung herangezogen werden.

**Wegevorhersage**

Die Wegevorhersage hat in den letzten Jahren stark an Bedeutung gewonnen. So ist man mit der Verfügbarkeit von großen GPS- und Mobilfunkdaten erstmals in der Lage, auf eine große Basis von historischen Daten zurückzugreifen und diese für die Vorhersage nutzen zu können. Navigationsanbieter wie TomTom nutzen GPS- und Mobilfunkdaten, um festzuhalten, wo aktuell wie viele Personen mit dem PKW unterwegs sind. Es wird auch berechnet, wo die Fahrer im nächsten Zeitintervall sein könnten, um Staus zu prognostizieren. Auch für die Mobilfunkanbieter ist die Bewegung der Nutzer eine wichtige Information. Wenn sie wissen, wie wahrscheinlich es ist, dass eine bestimmte Person von

einer Funkzelle A in eine Funkzelle B wechselt, ist die Funkzellenübergabe zwischen diesen einzelnen Zellen einfacher und problemloser, und es kommt zu weniger Funkabbrüchen.

Um eine Wegevorhersage zu machen ist es wichtig zu wissen, mit welchem Verkehrsmittel die Person unterwegs ist. Ist eine aktuelle Position $l_c$ und ein Richtungsvektor $\mathbf{v_c}$ eines Objektes gegeben, kann die zukünftige Position nach $\Delta\mathbf{t}$ des Objektes wie folgt berechnet werden $\mathbf{l_f} = \mathbf{l_c} + \mathbf{v_c}\Delta\mathbf{t}$. Der Time-parameterized R-Tree (TPR-Tree), der von Saltenis et al. 2000 vorgestellt wurde und die optimierte Version TPR* (Tao et al. 2003) wurden dazu entwickelt, vorausschauende Bereichsabfragen zu ermöglichen. Eine der Voraussetzungen bei diesem Vorgehen ist, dass der Richtungsvektor $\mathbf{v_c}$ seine Bewegung fortsetzt. Dabei wird die Bewegung im $\mathbb{R}^2$ linear weitergeführt, was für alle nicht straßengebundenen Trajektorien eine ausreichende Vorhersage darstellt. Doch gerade bei straßengebundenen Trajektorien im städtischen Bereich kommen häufige Wechsel der Bewegungsrichtung und Geschwindigkeiten vor. Monreale et al. (2009) stellten ein Verfahren vor, das auf Basis von Häufigkeiten der besuchten Orte eines Probanden die Vorhersage des nächsten Ortes vornimmt. In einem ersten Schritt werden raum-zeitliche Muster einer Trajektorie identifiziert. Im Anschluss wird mithilfe eines Entscheidungsbaumes, der auf die gefundenen Muster zurückgreift, der wahrscheinlichste nächste Ort identifiziert. Dabei kann das Straßennetz innerhalb einer Stadt berücksichtigt werden, indem der aktuelle und der prognostizierte Ort über ein Routing miteinander verbunden werden.

In Zukunft wird die Wegevorhersage noch ein spannendes und interessantes Forschungsfeld bleiben. So sind gerade für Location Based Services solche Vorhersagen interessant, ermöglichen sie es doch, Werbung gezielt an einzelne Personen zu richten. So kann man sich vorstellen, dass gerade zur Mittagszeit Personen mit Angeboten in ein bestimmtes Restaurant gelockt werden, welches auf dem Weg liegt und keinen Umweg bedeutet. Eine der generellen Annahmen ist dabei, dass Personen gewisse Routinen entwickeln und bestimmte Wege sehr häufig in einer ähnlichen Abfolge zurücklegen. Geht man von einer typischen Woche aus, so sind die Orte an denen man häufig verweilt, die eigene Wohnung, der Arbeitsplatz, der Verein, der Freundeskreis, etc. Die Wege zwischen diesen einzelnen Orten werden nur in einem geringen Maße variieren und lassen aufgrund dessen eine Vorhersage zu. Die aktuelle Forschung beschäftigt sich mit einer solchen individuellen Vorhersage für einzelne Personen, für bestimmte Tage, Uhrzeiten und das genutzte Verkehrsmittel. Hier spielen neben der eigentlichen Analyse von Trajektorien auch Verfahren zum Datenschutz eine wichtige Rolle, um die Privatsphäre von Personen zu schützen.

## 3.3 Möglichkeiten der Mobilitätserfassung

In diesem Abschnitt werden Möglichkeiten der Mobilitätserfassung vorgestellt, ihre Schwächen und Stärken dargestellt und im Anschluss für den speziellen Anwendungskontext in der Außenwerbung bewertet. Es wird der Frage nachgegangen, warum sich gerade GPS zur Leistungswertbestimmung in der Außenwerbung anbietet.

### 3.3.1 Erfassungsmethoden, Fragestellungen und Studien zur Mobilitätserfassung

Die Pendelbewegung zum Arbeitsplatz ist wohl einer der meist getätigten Wege vieler Menschen. Dazu kommen Wege zum Einkauf, zur Freizeit und Besuchsfahrten. All diese Wege ergeben unseren sogenannten Aktivitätsraum, der in seiner Ausdehnung sehr unterschiedlich sein kann. Die Ausdehnung des Aktivitätsraumes wird nach Golledge und Stimson (1997) im Wesentlichen durch drei Faktoren beeinflusst:

- Durch die Lage des Wohnortes,
- durch die Lage der Orte, an denen sich eine Person regelmäßig aufhält,
- und durch die Wege, die diese Orte miteinander verbinden.

Es existieren eine Reihe von verschiedenen Methoden, um die Mobilität von Personen aufzuzeichnen. Sie unterscheiden sich in den Erhebungseigenschaften, in ihrer Erhebungsdauer, in computergestützt vs. Befragung und dem Zeitpunkt der Erhebung (zeitgleich, zeitversetzt). Die umfangreichsten empirischen Mobilitätsstudien wurden in der Vergangenheit mittels zeitversetzter Befragungen durchgeführt. Um die Kosten der Befragungen möglichst gering zu halten, wurden die Probanden in der Regel nur für einen bestimmten Stichtag befragt. In Deutschland wird hierzu turnusmäßig alle 6 Jahre die Studie „Mobilität in Deutschland" durchgeführt. Sie stellt eine landesweite Telefonbefragung zum Mobilitätsverhalten von Haushalten dar. Beauftragt wird die Studie vom Bundesministerium für Verkehr-, Bau- und Stadtentwicklung auf Basis von CATI Mobilitätsdaten für Planung, Politik und Wissenschaft. Dabei wird in der Erfassung nicht der exakte Weg festgehalten, sondern nur die Orte, an denen sich der Proband an dem Tag aufgehalten hat. Zusätzlich werden neben soziodemographischen Variablen der Wegezweck und das genutzte Verkehrsmittel erfasst. Auch in der Schweiz existiert eine vergleichbare Studie, der Schweizer Mikrozensus. Er erfasst turnusmäßig alle 5 Jahre per Telefonbefragung die Mobilität, jeweils an einem Stichtag und flächendeckend. Ziel beider Studien ist es, sogenannte Kenngrößen des Verkehrs abzuleiten und vergleichende Statistiken über die Jahre zu erstellen. Typische Fragestellungen der Studien sind u.a.:

- Hat die Mobilität in den vergangen Jahren zugenommen?
- Wie ist der modale Split des Verkehrs?
- Welche Pendelbeziehung zwischen den Gemeinden gibt es?

Es handelt sich hierbei also um großräumige Betrachtungen der Mobilität. Aussagen auf Straßenabschnittsebene werden hier nicht getroffen. Zukünftig werden Telefonbefragungen immer mehr auf das Problem stoßen, eine repräsentative Stichprobe zu erreichen. Lag der Versorgungsgrad der Festnetzanschlüsse in Deutschland und in der Schweiz im Jahre 2000 noch bei fast 100%, so nimmt dieser Anteil in den letzten Jahren durch die Mobilfunknutzung kontinuierlich ab. Viele, insbesondere junge Haushalte, besitzen keinen Festnetzanschluss mehr, wodurch diese nicht mehr oder nur sehr schwer durch eine Telefonbefragung erreicht werden können, da Mobilfunknummern nicht zentral registriert werden.

Eine bereits lange eingesetzte Möglichkeit der Erfassung der Mobilität stellen Tagebuchaufzeichnung dar. In einem gemeinsamen Projekt der PTV AG aus Karlsruhe und der ETH Zürich wurden im Rahmen eines Forschungsprojektes über einen Zeitraum von sechs Wochen von Probanden Tagebuchaufzeichnungen vorgenommen. In den Städten Karlsruhe und Halle an der Saale wurden insgesamt 361 Personen zu ihrer individuellen Mobilität befragt (Schlich 2003). Die Befragung zählt hinsichtlich des Erhebungszeitraumes zu einer der langen Studien dieser Art. Im Jahre 1971 wurde in Uppsala eine ähnlich lange (5 Wochen) Befragung durchgeführt. Protokolliert wurden Wegezweck, Start- und Zielort, Zeiten, Begleitsituationen (Mitfahrer), Kosten und Verkehrsmittel.

Zu den typischen Fragestellungen der Studien gehören:

- Wie häufig wurden bestimmte Orte aufgesucht?
- Wie viele Wege wiederholen sich über die Woche/Monat?
- Welche Regularitäten hat die Mobilität für unterschiedliche soziodemographische Personengruppen?

Bei den Erhebungen mittels Tagebüchern liegt eine der großen Schwierigkeiten darin, eine ausreichend große Menge an Teilnehmern zu rekrutieren. So ist insbesondere eine Anforderung, dass die Teilnehmer über den Zeitraum von 6 Wochen keine Abwesenheit einplanen, ihr Tagebuch gewissenhaft ausfüllen und ihre Wege jederzeit protokollieren. Um die Gewissenhaftigkeit der Probanden zu überprüfen, wurden die Fragebögen direkt nach Abschluss überprüft und Unplausibilitäten direkt im Anschluss telefonisch geklärt. Über die Dauer der Befragung wurde eine nachlassende Zuverlässigkeit festgestellt.

Eine weitere Erfassungsmethode ist die GPS-basierte Erfassung. Mit einer Genauigkeit von in der Regel 3-10 Metern und einer sekundengenauen Erfassung gehört sie zu den präzisesten Erfassungsmöglichkeiten der Mobilität. Neben der hohen Qualität der Erfassung ist die geringe Belastung der Probanden ein großer Vorteil. Die Probanden müssen keine Erinnerungsleistung aufbringen und müssen auch keine Tagebücher protokollieren. Es müssen auch keine zeit- und personalintensiven Interviews am Telefon durchgeführt werden. Die Wege liegen direkt digital vor. Die sonst so schwierigen Zeitangaben in Interviews lassen sich hier direkt den besuchten Orten zuordnen (Zheng 2009, Marchal 2010). Nachteile ergeben sich durch die aufwändige und z.T. anfällige Technik. So ist insbesondere die Stromversorgung bei GPS-Geräten immer noch ein schwieriger Punkt. In der Regel halten die Akkus nicht länger als 24 Stunden und müssen dann aufgeladen werden. Zudem liegt bei der GPS-Erfassung keine Annotation vor. Während sie bei Interviews direkt mit abgefragt werden kann, muss sie bei GPS mit viel Aufwand nachträglich ergänzt werden. GPS ist an dieser Stelle ein vollkommen passives Beobachten. Erfassungsprobleme bei GPS tauchen insbesondere bei engen Häuserschluchten und Tunneln auf. Auch muss der Empfänger zu mindestens drei Satelliten eine Verbindung aufgebaut haben, da sonst keine exakte Lokalisierung vorgenommen werden kann. Ein weiteres Problem stellt die sogenannte Warmup-Phase der GPS-Geräte dar. Es dauert einige Zeit, bis die GPS-Geräte eine Verbindung aufgebaut haben. Das kann dazu führen, dass die ersten Meter außer Haus nicht aufgezeichnet werden. Zu den größten GPS-Stichproben in Europa zählen Daten in der Telematik von Autos. So kann man in Italien seine Autoversicherung auf Basis der zurückgelegten Kilometer abrechnen. Hierzu wird ein GPS-Gerät in die PKW´s eingebaut, und die Daten werden via SIM Karte an die Versicherung übertragen.

Zu den typischen Fragestellungen der Studien gehören:

- Wie orientieren sich Reisende in Netzwerken?
- Wie lange halten sich Probanden an bestimmten Standorten auf?
- Wie viele Kilometer legt eine Person pro Tag und pro Woche zurück?

Ebenfalls zu den Erfassungsmethoden mit hoher zeitlicher Auflösung gehört die Erfassung mittels Global System for Mobile Communications (GSM). Fast 80% der Bundesbürger haben inzwischen ein Mobilfunkgerät. Grundlage für die mobile Telekommunikation ist die Nutzung elektromagnetischer Wellen eines Frequenzbandes. Der derzeitige GSM-Standard unterscheidet den 900er und den 1800er MHZ Frequenzbereich. Zukünftig wird noch das Frequenzband Long-Term-Evolution (LTE) stärker dazukommen. LTE unterstützt verschiedene Bandbreiten (1.4, 3, 5, 10, 15, und 20 MHz) und kann so flexibel für unterschiedliche Anforderungen eingesetzt werden. Bisher ist dies in Deutschland und der Schweiz jedoch nur in einzelnen Großstädten verfügbar. Wird ein Mobilfunkgerät eingeschaltet, so gibt es zwei Arten von Zuordnungen im Mobilfunknetz. Das zentrale System bilden miteinander verbundene Dienstvermittlungsstellen Mobile Switching Center (MSC). Die MSC haben die Aufgabe, bei einem Gesprächsaufbau oder einer Datenübermittlung (SMS, Internet) zu der Funkzelle (Base Transceiver Station - BSC) zu vermitteln, in der sich der mobile Teilnehmer gerade aufhält, oder bei einem Zellenwechsel das Gespräch von einer Basisstation zur nächsten weiterzureichen. Wir unterscheiden dabei drei Arten von Funkzellen: Makrozellen (>1km), Mikrozellen (<1km) und Pikozellen in Gebäuden. Ferner wird jeder Funkzellenstandort in zwei Typen differenziert. Ein Standort mit nur einem Sektor, der rundum strahlt, wird als Omni-Standort bezeichnet. Werden mehrere Bereiche von einem Standort aus erzeugt, wird dieser als Sector-Standort deklariert. Entfernt sich ein Gesprächsteilnehmer von seiner derzeitigen Empfangsstation, geht dies mit einer Verschlechterung der Signalstärke und damit der allgemeinen Gesprächsqualität einher. In diesem Falle erfolgt eine Weitergabe (Handover) der Funkverbindung an die nächst beste Funkzelle, sobald die Feldstärke einen Schwellenwert unterschreitet.

Möchte man Mobilitätsanalysen auf Basis von Mobilfunkdaten durchführen, so steht man vor besonderen Herausforderungen. Die MSC sind die einzigen Zelleneinheiten, die permanent die Bewegungen der Mobilfunkkunden mit aufzeichnen. Sie stellen jedoch für Mobilitätsanalysen sehr grobe räumliche Einheiten dar. So wird z.B. die Gemeinde Köln in insgesamt nur vier MSC Zellen aufgeteilt. Die kleineren BSC Zellen stellen zwar kleinere räumliche Einheiten dar, zeichnen jedoch die Mobilfunkkunden erst dann auf, wenn auch eine Übertragungsaktivität einsetzt. Eine zusätzliche Herausforderung ist die nicht permanente ID der Mobilfunknutzer. Aus Datenschutzgründen wird in Deutschland und auch in der Schweiz nach vier Stunden eine neue ID zugeordnet. Das bedeutet, dass ein Tracking einer Person über einen längeren Zeitraum nicht möglich ist. Aktuell werden Mobilfunkdaten intensiv in kommerziellen Projekten zur Stauprognose eingesetzt. Die Firma TomTom aus den Niederlanden nutzt Mobilfunkdaten, um auf dem übergeordneten Straßennetz (Autobahnen, Bundesstraßen) etwaige Staus zu erkennen. Hierzu werden die Handover Daten dazu benutzt um festzustellen, ob sich das Eintreten von Mobilfunkteilnehmern einer Funkzelle stark von dem Austreten unterscheidet. Gerade auf Autobahnen und stark befahrenen Bundesstraßen hat dieses Vorgehen zu einer Verbesserung der Stauerkennung beigetragen. Ein großer Vorteil von Mobilfunkdaten ist die hohe zeitliche Auflösung. Die Mobilfunkdaten können fast in Real Time verarbeitet und analysiert werden. Zu den typischen Fragestellungen bei Mobilfunkdaten zählen:

- Wo habe ich innerhalb einer Stadt einen aktuellen Aktivitätsschwerpunkt?
- Gibt es monatliche, saisonale Effekte in der Mobilität von Personen?
- Welche Übergangswahrscheinlichkeiten habe ich in einer Stadt?

Ein Beispiel für eine weitere Form der Mobilitätserfassung, die jedoch aus Kostengründen nur lokal begrenzt eingesetzt werden kann, stellt die Bluetooth-Ortung dar. Ungefähr 6-10% aller Mobilfunkhandys haben Bluetooth eingeschaltet und lassen sich hierüber orten. Notwendig ist hierzu ein Bluetooth Scanner, der über eine maximale Entfernung von 100 Metern Mobilfunkhandys mit eingeschalteter Bluetooth Funktionalität lokalisieren kann. Eine

exakte Positionierung der erfassten Geräte ist dabei durch Ablenkungen und Inferenzen relativ schwierig. Im optimalen Fall ist eine Ortung auf 5 Meter genau. Neben der Feldstärke, die zur Lokalisierung genutzt wird, werden noch der Zeitpunkt und die MAC-ID des erfassten Gerätes protokoliert. Die Mac-ID stellt einen eindeutigen Identifier von elektronischen Geräten/Mobilfunkgeräten dar. Sind mehrere Bluetoothscanner innerhalb eines Objektes oder Veranstaltungsortes installiert, können Bewegungsprofile zwischen einzelnen Scannern aufgezeichnet werden. Einsatzgebiete für die Bluetooth-Scanner sind typischerweise Objekte wie Bahnhöfe oder Flughäfen. Hier werden auf einem lokal begrenzten Raum Bewegungsanalysen z.B. für die Bewertung von Shopstandorten durchgeführt. Für einen flächendeckenden Einsatz in Städten oder Gemeinden kommt Bluetooth aus Kostengründen nicht in Frage. Zusätzlich können in der Regel nur Fußgänger erfasst werden, da die Karosserie und die Geschwindigkeit eines Autos eine Erfassung kaum zulässt. Zu den typischen Fragestellungen bei lokalen Erfassungen mit Bluetooth zählen:

- Wie lange ist die Aufenthaltsdauer in einer bestimmten Zone eines Einkaufsmarktes?
- Wie lange braucht ein Kunde durchschnittlich für den Einkauf?

Zusammenfassend können die Methoden zur Mobilitätserfassung in zwei Kategorien unterschieden werden: Einmal in die Methoden, die über die Erinnerungsleistung versuchen, die Mobilität von Probanden nachträglich zu rekonstruieren und zum anderen Methoden, die über eine direkte Erfassung mittels elektronischer Sensoren/Scanning Mobilität erfassen. Die Methoden unterscheiden sich neben dem Zeitpunkt der Annotation noch im Wesentlichen in der Dauer und der Aktualität der erfassten Daten. Die Abbildung 3.11 stellt dies grafisch dar. Die genannten Charakteristiken der Erfassungsmethoden sind in den jeweils grünen Boxen aufgeführt. In den blauen Boxen sind die Kategorien der befragungsgetriebenen und der sensorgetriebenen Erfassung mit ihren jeweiligen zugeordneten Erfassungsmethoden visualisiert.

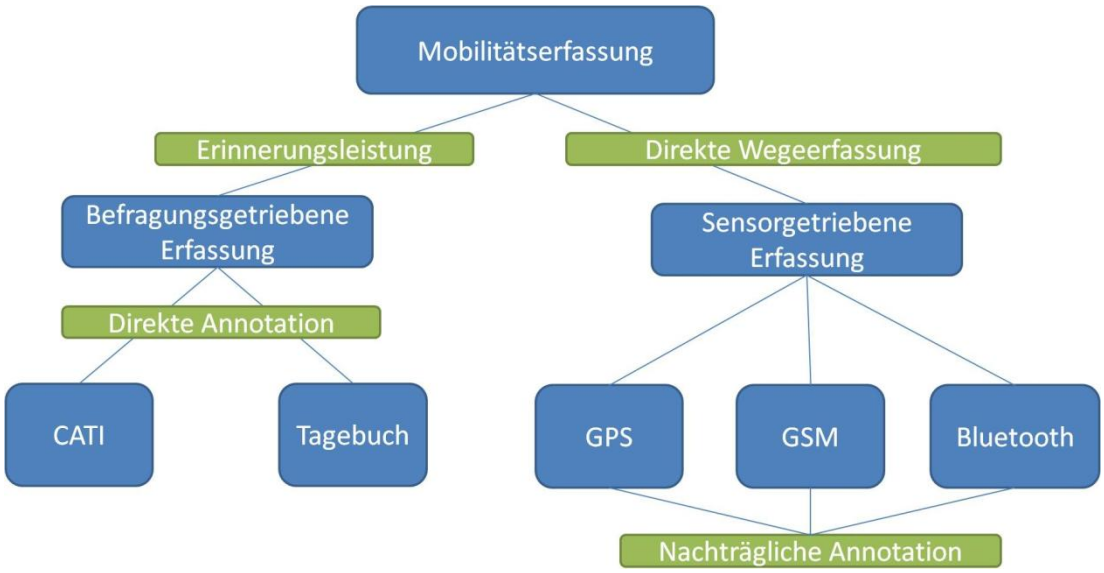



Abbildung 3.11: Methoden der Mobilitätserfassung

Je nach Anwendungskontext und Fragestellungen muss eine geeignete Erfassungsmethode gefunden werden.

### 3.3.2 Bewertung des Einsatzes von Methoden der Mobilitätserfassung hinsichtlich des Anwendungskontextes

Betrachtet man die Anforderungen der Außenwerbung, so ist es zur Bestimmung der Reichweite zwingend notwendig, Mobilitätsdaten über einen Zeitraum von 7 Tagen vorliegen zu haben (vgl. Abschnitt 2.1). Tagebuchaufzeichnung sowie CATI Interviews über diesen Zeitraum durchzuführen scheint ein überaus ambitioniertes Vorhaben zu sein, denn wenige Probanden werden das Interesse haben, über so einen langen Zeitraum an einer Studie teilzunehmen. Zudem haben verschiedene Studien gezeigt, dass zwischen der Erinnerungsleistung von Probanden und den tatsächlich zurückgelegten Wegen Unterschiede bestehen. Gerade die nicht regelmäßigen Wege, aber auch z.B. die Suche nach einem Parkplatz, fehlen häufig in der Erinnerung der Probanden (Stopher 2007, Zmund 2003). Alternativ bieten sich direkte Wegeerfassungen mittels Mobilfunk, Bluetooth und GPS an. Sehr interessant sind hierbei Mobilfunkdaten, da sie z.T. für einen sehr langen Zeitraum vorliegen. Jedoch stellt sich hier das Problem, dass keine exakten Wege aus Mobilfunkdaten abgebildet werden können, da die Daten nur auf Funkzellenebene vorliegen. Ein weiteres Problem für die Berechnung der Reichweiten ist die regelmäßige Änderung der Mobilfunk-ID´s. So ist es nicht mehr möglich, einer bestimmten Person über einen längeren Zeitraum einen Weg zuzuordnen. Dies ist aber für die Reichweitenberechnung zwingend notwendig. Ein flächendeckender Einsatz von Bluetooth Scannern für eine komplette Stadt bzw. alle Plakatstellen einer Stadt, ist aus Kostengründen nicht realisierbar. Insbesondere fehlen bei beiden Varianten der kontinuierlichen Erfassung die Zielgruppenmerkmale wie Geschlecht, Alter, Einkommen, etc. Die Wegeerfassung mit GPS ist im Anwendungskontext die genaueste und exakteste Möglichkeit, Probandenmobilität festzuhalten. Allerdings gilt es hier, wie in Kapitel 4 dargestellt, einige Herausforderungen zu lösen, sowie eine nachträgliche Annotation der Trajektorien durchzuführen. Hierauf wird im Rahmen der Arbeit noch vertieft eingegangen.

## 3.4 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) ist ein relativ junges Forschungsgebiet, das sich im Schnittpunkt von verschiedenen Wissenschaftsdiziplinen befindet. So kommen Inhalte der Statistik, des Maschinellen Lernens, des Pattern Recognition, der Datenbanktheorie und der Künstlichen Intelligenz zum Einsatz. Dabei besteht hinsichtlich der Ziele und Aufgaben eine enge Verknüpfung zur Statistik. In beiden Fällen ist die Suche und das Aufklären von Mustern das primäre Ziel (Hudec 2002). Die klassischen Methoden der angewandten Statistik werden in der Regel für einfache Hypothesen eingesetzt. Die Daten sind häufig für die Beantwortung einer spezifischen Fragestellung erhoben worden.

Bei Knowledge Discovery in Databases spricht man häufig von sehr großen Datenmengen, die weit über 1 Million Einträge hinaus gehen und nicht primär für die Beantwortung einer bestimmten Fragestellung erhoben worden sind. Vergleichbar mit den Zielen der Statistik sucht KDD nach Mustern, die in Datensammlungen bei z.B. Versicherungen, Banken oder Universitäten versteckt sind. Dabei entstand diese Forschungsrichtung vor dem Hintergrund, dass gängige Techniken zur Datenanalyse nicht mehr Schritt hielten mit dem rapiden Fortschritt bei der Erfassung und Speicherung von Daten. Diesen Engpass versucht KDD zu überwinden, der durch den vorwiegend zeitaufwendigen und manuellen Analysestil der klassischen Statistik verursacht wird. Die ursprüngliche Herangehensweise in der Statistik bestand meist im Aufstellen von Hypothesen, die sich beim vorliegenden Datenmaterial als richtig oder falsch bestätigten. In den späten Achtzigern begangen Forscher, sich mit neuen Herangehensweisen zu beschäftigen.

*„To deal with the data glut, a new generation of intelligent tools for automated data mining and knowledge discovery is needed."* (Fayyad, Preface, 1996)

Forscher der künstlichen Intelligenz fingen an, Algorithmen zu entwickeln, mit denen umgekehrt vorgegangen werden konnte. Aus vorhandenen Daten sollten automatisiert Hypothesen generiert werden, die neue und interessante Muster enthalten. Eine Unterstützung durch computergestützte Methoden wurde daher notwendig, wodurch die Motivation von KDD begründet ist. Data Mining wird oft im Zusammenhang mit Knowledge Discovery genannt, oder beide Begriffe werden synonym verwendet (Piatetsky-Shapiro 2007, S. 100). Dabei stellt Knowledge Discovery den gesamten Analyseprozess von der Frageformulierung bis zur Ergebnisinterpretation dar und das Data Mining lediglich im engeren Sinne die Suche nach auffälligen Mustern, Trends etc.

*„Data Mining is a step in the KKD process consisting of applying computational techniques that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data."* (Fayyad 1996: S.34)

Dabei ist eines der Hauptprobleme bei der angestrebten Automatisierung der Datenexploration die sehr große Anzahl der theoretisch möglichen Muster. Es werden intelligente Suchverfahren sowie komplexe Bewertungsstrategien benötigt, um den Suchraum vernünftig einzuschränken. Die gefundenen Muster sollen die Eigenschaften besitzen, dass sie für einen großen Teil des Datenbestandes gültig sind und bislang unbekannte, aber nützliche und verständliche Zusammenhänge beschreiben. Die gefundenen Muster können z.B. dazu dienen, auf Entscheidungsebene Maßnahmen zur Erhöhung des Produktabsatzes abzuleiten oder Produktionsfehler bei der Herstellung von Produkten zu finden. Ursprünglich wurde die Mustersuche nur auf strukturierte Daten angewendet. Inzwischen werden auch Muster auf unstrukturierten Daten, z.B. Texten und anderen

Datenquellen gesucht, so dass sich neue Forschungsrichtungen entlang unterschiedlicher Datenquellen entwickelten. So beschäftigt sich das Text Mining mit der Extraktion von Mustern aus unstrukturierten Daten (Zeitungsartikel, Patente, Versicherungsakten). Das Web Mining beschäftigt sich speziell mit der Extraktion von Mustern und Nutzerverhalten aus dem World Wide Web. Das Mobility Mining bezeichnet die Erschließung und Anwendung des KDD auf Mobilitätsdaten. Mit Mobility Mining werden Mobilitätsdaten aufbereitet, analysiert, Muster aufgedeckt und im Anschluss visualisiert. Für den Anwendungshintergrund kommen Verfahren des Data Mining zum Einsatz, da es sich bei den gesammelten GPS-Daten um sehr große Datenmengen handelt, die in dieser Form nur schwer mit den klassischen Methoden der Statistik analysiert werden können.

Bei allen speziellen Datenformaten und Anwendungen ist das typische Vorgehen im KDD-Prozess nach bestimmten Phasen geordnet. Nach Fayyad (Fayyad et al. 1996, S. 9 ff.) erfolgt in der ersten Phase des KDD-Prozesses die Auswahl des relevanten Datenbestandes für die Fragestellung. Dieser sogenannte Zieldatenbestand wird anschließend in einer Vorverarbeitungsphase bereinigt. Hierzu zählt das Beheben von etwaigen Datenqualitätsmängeln. Fehlerhafte Daten werden ersetzt, interpoliert oder entfernt. Es werden externe Datenbestände mit dem Zieldatenbestand verknüpft und angereichert. In der zweiten Phase wird der Datenbestand in die notwendige Form zur Analyse gebracht. Dazu zählen in der Regel Diskretisierungsschritte numerischer Werte, Normalisierungen oder die Umwandlung von nominalen Werten in numerische Werte. Im Anschluss folgt die Anwendung von Data Mining Verfahren zur Mustererkennung. Darauf folgen die Interpretation der Ergebnisse und die Evaluierung der Muster. Sind die Aufgaben oder Zielsetzungen mithilfe der gefundenen Muster nicht zufriedenstellend gelöst worden, ist ein Rücksprung in eine der vorherigen Phasen zur Ergebnisverbesserung möglich (vgl. Abbildung 3.12).

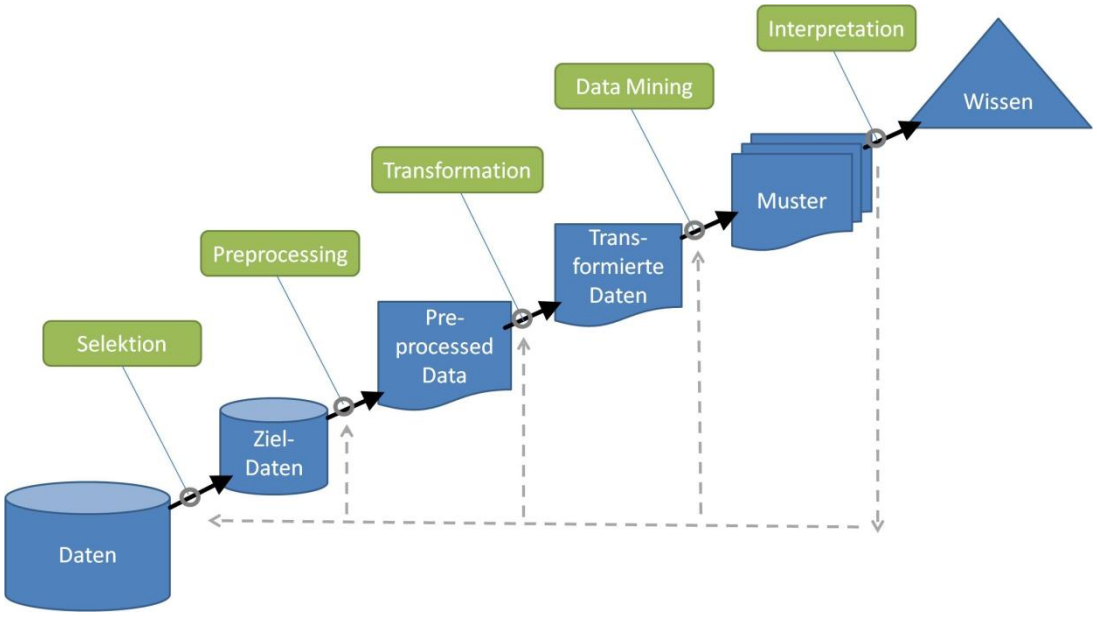



Abbildung 3.12: KKD-Prozessablauf (nach Fayyad et al. 1996)

### Aufgabenbereiche des KDD

Die Aufgabenbereiche des KDD lassen sich in verschiedene Gruppen einteilen. Nachfolgend wird die Einteilung in Regressions-, Klassifikations-, Segmentierungs/Cluster- und Abhängigkeits- Assoziationsanalyse vorgestellt:

Regressionsanalyse

Mit den Regressionsmodellen des KDD ist es das Ziel, anhand von unabhängigen Variablen die abhängige vorherzusagen wird. Mithilfe der Prognose anhand von Ausprägungen der zugehörigen unabhängigen Variablen können abhängige Variablen eingesetzt und ersetzt werden kann. Die Regressionsanalysen gehören zu den meist eingesetzten Verfahren im KDD.

Klassifikationsanalyse

Klassifikationsmodelle haben das Ziel, Elemente einer vorgegebenen Klasse zuzuordnen. Die Grundlage der Klassifikation bildet ein Lerndatenbestand, dessen Objekte einer vorgegebenen Klasse zugeordnet sind. Hierbei ergibt sich die Klassenausprägung bei einer diskreten Klassenvariablen aus den Ausprägungen der Attribute. Das Klassifikationsmodell dient dazu, die Klassenzugehörigkeit von allen Datenbankobjekten, deren Klassenzugehörigkeit bisher unbekannt ist, zu prognostizieren. Im Unterschied zum Regressionsmodell mit stetigen Werten wird im Klassifikationsansatz mit rein diskreten Werten gearbeitet.

Segmentierungs- bzw. Clusteranalyse

Bei der Segmentierungsanalyse werden Gruppen von Objekten so gebildet, dass die Objekte innerhalb einer Gruppe möglichst homogen sind. Umgekehrt sollten Objekte in unterschiedlichen Gruppen möglichst heterogen sein. Die gefundenen Gruppen von „ähnlichen“ Objekten werden als Cluster bezeichnet, das Verfahren der Gruppenzuordnung als Clustering. Im Unterschied zur Klassifikationsanalyse, bei der Daten bestehenden Klassen zugeordnet werden, ist es das Ziel der Clusteranalyse neue Gruppen/Klassen in den Daten zu identifizieren. Objekte, die keinem Cluster zugeordnet werden können, werden als Ausreißer interpretiert.

Abhängigkeits- bzw. Assoziationsanalyse

Bei der Analyse von Abhängigkeiten werden Beziehungen zwischen Attributen oder einzelnen Ausprägungen erkannt, die innerhalb einer bestimmten Teilmenge des Datenbestandes bestehen. Die gefundenen Regeln beschreiben also Korrelationen zwischen gemeinsam auftretenden Ereignissen. Assoziationsregeln werden häufig bei Warenkorbanalysen eingesetzt und helfen dabei, Werbemaßnahmen und Produktplatzierungen im Geschäft zu optimieren. Ob bei den gefundenen Mustern ein kausaler Zusammenhang besteht, muss im Anschluss mit Expertenwissen überprüft werden.

### KDD-Verfahren

Zum Erreichen der einzelnen KDD Aufgaben und Analyseziele stehen verschiedene Verfahren zur Verfügung, die z.T. aufgabenübergreifend eingesetzt werden. Im Folgenden werden einzelne Verfahren vorgestellt, dabei werden die Verfahren, die für die spätere Modellierung wichtig sind, ausführlicher beschrieben. Hierzu zählen die Nächste-Nachbar-Klassifikation und die Subgruppensuche.

**Entscheidungsbäume**

Oft eingesetzte Verfahren zur Klassifikationsanalyse sind Entscheidungsbaumverfahren. Dabei stellen Entscheidungsbäume sogenannte geordnete und gerichtete Graphen dar. Sie präsentieren in einer hierarchischen Struktur das Aufeinanderfolgen von Entscheidungen. Entscheidungsbäume bestehen immer aus einem ausgehenden Wurzelknoten und beliebig vielen weiteren inneren Knoten. Jeder Knoten stellt eine Regel dar und jede Verknüpfung zwischen den Knoten eine Antwort auf das Entscheidungsproblem. Bei jedem einzelnen Knoten wird ein Attribut abgefragt, und es wird eine Entscheidung über das weitere Vorgehen getroffen. Zur Entwicklung und Überprüfung des Modells wird beim Entscheidungsbaumverfahren die Datenmenge in eine Trainings- und Testdatenmenge aufgeteilt. Die Trainingsdatenmenge wird sukzessive nach ihren Attributwerten in einzelne Mengen aufgeteilt, die disjunkte Teilmengen darstellen. Die Qualität des Modells wird im Anschluss an einer klassifizierten Testdatenmenge überprüft. Als Maß der Güte des Modells kann die Fehlklassifikationsquote der Testmenge herangezogen werden (Quinlan 1993, Murthy 1998).

An einem praktischen Beispiel wird nun im Folgenden das Konzept der Entscheidungsbäume vorgestellt. Es beschreibt die Entscheidung, an einem Tag Tennis spielen zu können oder nicht. Dabei wird das spezifische Wetter des Tages durch vier nominale Attribute beschrieben. Das binäre Klassenattribut *„Spielen"* beschreibt, ob bei den spezifischen Wettereigenschaften Tennis gespielt werden kann oder nicht. In der beschriebenen Notation haben wir die Variablen Y und X mit den Ausprägungen Y = {ja, nein} und X = {Aussichten, Temperatur, Luftfeuchte, Wind}. Ein Entscheidungsbaum, der diese Informationen aus der Tabelle 3.2 ableitet, ist in der Abbildung 3.13 dargestellt.

| | X | | | | Y |
|---|---|---|---|---|---|
| **Tag** | **Aussichten** | **Temperatur** | **Luftfeuchte** | **Wind** | **Spielen** |
| 1 | sonnig | heiß | hoch | schwach | nein |
| 2 | sonnig | heiß | hoch | stark | nein |
| 3 | bewölkt | heiß | hoch | schwach | ja |
| 4 | Regen | mild | hoch | schwach | ja |
| 5 | Regen | kalt | normal | schwach | nein |
| 6 | Regen | kalt | normal | stark | ja |
| 7 | bewölkt | kalt | normal | stark | nein |
| 8 | sonnig | mild | normal | schwach | ja |
| 9 | sonnig | kalt | normal | schwach | ja |
| 10 | Regen | mild | normal | schwach | ja |
| 11 | sonnig | mild | normal | stark | ja |
| 12 | bewölkt | mild | hoch | stark | ja |
| 13 | bewölkt | heiß | normal | schwach | ja |
| 14 | Regen | mild | hoch | stark | nein |

Tabelle 3.2: Wetterinformationen Entscheidungsbaum (Witten et al. 2001)

Der Entscheidungsbaum kann als eine Funktion $h: X \rightarrow Y$ wie folgt interpretiert werden. Gegeben ist eine Instanz $x \in X$. Beginnend mit dem Wurzelknoten wird das Attribut in dem Knoten nach seinem Wert für $X$ überprüft und der entsprechende Zweig für den weiteren Durchlauf ausgewählt. Sobald ein Blattknoten erreicht ist, stellt der bekannte Wert das Ergebnis h(x) dar. Dabei ist zu beachten, dass es bei jeder erreichten Ebene eines Entscheidungsbaumes zu einer weiteren Instanziierung des Datensatzes kommt.

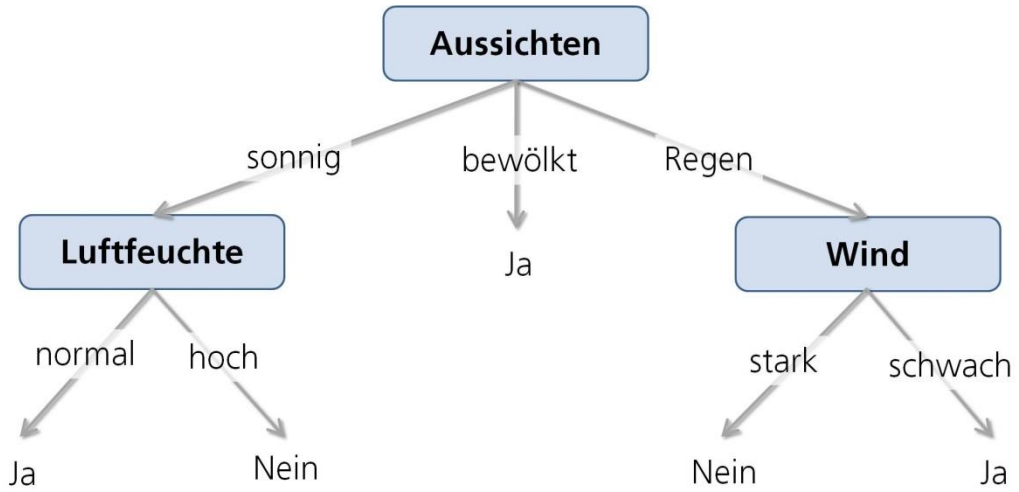



Abbildung 3.13: Entscheidungsbaumvorgehen (nach Witten et al. 2001)

Eine der wichtigen Regeln bei der Erstellung von Entscheidungsbäumen und auch von Data Mining Verfahren im Allgemeinen ist es, sie möglichst effektiv bzw. generisch zu halten. *„Entscheidungsbäume neigen ohnehin zur Übergröße, da auch zufällige Elemente in der Trainingsmenge das Erkennen von scharfen Regeln erschweren. Insbesondere in den tiefen Verzweigungen, d.h. in der Nähe der Blätter, wird der Einfluss des Zufalls größer und zwingt den Entscheidungsbaum zur Übermodellierung"* (Krahl et al. 1998, S. 73). Deswegen ist es in den meisten Fällen notwendig, den Entscheidungsbaum zu generalisieren, d.h. unnötige und nicht aussagekräftige Verästelungen zu beschneiden. Diese Verfahren zur Reduktion des Suchraums werden Pruning (stutzen) genannt. Das Pruning bewirkt, dass einzelne Entscheidungsknoten, die nur einen geringen Anteil an der Klassifikationsgüte auf unbekannten Daten haben, entfernt werden. Eine bekannte Beispielanwendung für das Pruning stellen Schachprogramme dar. Schachprogramme betrachten in der Regel nicht nur den nächsten Zug, sondern wollen möglichst eine große Anzahl der möglichen kommenden Züge durchspielen. Bei 32 Figuren, die auf eine bestimmte Art gezogen werden dürfen, und 64 Feldern ergeben sich schon nach dem ersten Zug von beiden Spielern 400 neue Kombinationsmöglichkeiten. Nach zwei Zügen von beiden Spielern ergibt sich rechnerisch schon die Möglichkeit von 197.742 neuen Varianten. In der Regel endet eine Schachpartie nach 100 Zügen. An dieser Stelle wird schnell klar, dass es nicht möglich ist, alle Situationen zu betrachten. Aus diesem Grund wird Pruning verwendet, indem Spielzüge bzw. mögliche Verästelungen bewertet werden. Der Entscheidungsbaum sucht beim Pruning zunächst nur bis zu einer gewissen Tiefe. Findet der Entscheidungsbaum Verzweigungen, die keine oder nur eine geringe Änderung bewirken, wird keine weitere Suche vorgenommen.

Entscheidungsbäume zählen zu den populärsten Verfahren des Data Mining, da die erzeugten Modelle, falls sie nicht zu groß werden, noch sehr gut lesbar, nachvollziehbar und umsetzbar sind.

**Nächste-Nachbar-Klassifikation (kNN)**

Das k Nächste-Nachbar-Verfahren (kNN) ist ein Klassifikationsverfahren, das versucht, unbekannte Objekte anhand von Analogien bekannten Objekten einer Klasse zuzuordnen. Dabei berechnet das kNN Verfahren zur Klassifikation eines Objektes die k ähnlichsten Objekte aus einem Datensatz und ordnet das gegebene Objekt in die Klasse, die am häufigsten unter den k ähnlichsten Objekten vorkommt (Belur 1991, Shakhnarovish 2005). Es handelt sich hierbei um einen einfachen Mehrheitsentscheid, denn das Objekt wird schlussendlich der Klasse zugeordnet, die am häufigsten unter den k nächsten Objekten vertreten ist (siehe Abbildung 3.14). Dabei wird die Ähnlichkeit bzw. Unähnlichkeit von Objekten z.B. über den euklidischen Abstand definiert. Zur Bestimmung der euklidischen Distanz (vgl. Kap. 3.1.2) werden die Attribute in der Datenaufbereitung normalisiert. Es seien $a$ und $b$ Datensätze, $n$ die Länge der Vektoren und die Notation $a_i$ bezeichnet die Ausprägung der $i-ten$ Dimension des Vektors $a$. Im Attributraum sind jene Vektoren $a$ und $b$ am ähnlichsten, für die die gewichtete Distanz $dist(a,b)$ am geringsten ist. Damit entscheiden die jeweils gefundenen k nächsten Nachbarn durch das Mehrheitsvotum, welcher Klasse das Objekt zugeordnet wird. Wenn die gewählte Klasse $C_j$ diejenige ist, die die Anzahl $k_{cj}$ an Beispielen aus den k-ähnlichsten Objekten hat, so wird das Objekt mit einer Konfidenz $k_{cj}/k$ der Klasse $C_j$ zugeordnet. Dabei kommt der Gewichtung der Attribute eine große Bedeutung zu. In der Regel bestimmt man die Gewichtung über eine Korrelationsanalyse. Das Attribut, welches die stärkste Korrelation mit dem Klassenattribut aufweist, wird am höchsten gewichtet. Attribute, die nur schwach oder negativ korreliert sind, werden aus dem Verfahren herausgenommen. Eine Neugewichtung nur eines Attributes führt jeweils zu einem anderen Ergebnis. Weiterhin ist die Festlegung des k von entscheidender Bedeutung. Hiermit wird bestimmt, wie viele nächste Nachbarn in die Berechnung mit einfließen. Ist das k sehr gering, so reagiert das Modell sehr anfällig auf Ausreißer und Fehler in den Daten. Bei einem sehr groß gewählten k wird sehr stark gemittelt und lokale Effekte könnten verschwinden. In der Abbildung 3.15 wird der Zusammenhang zwischen dem gewählten k und der Klassenzuordnung (blaue und weiße Fläche) von 2 unterschiedlichen Objekten (rote und blaue Punkte) deutlich gemacht. Zusätzlich wird im Vergleich eine lineare Regression dargestellt, um die Vorteile des kNN deutlich zu machen.

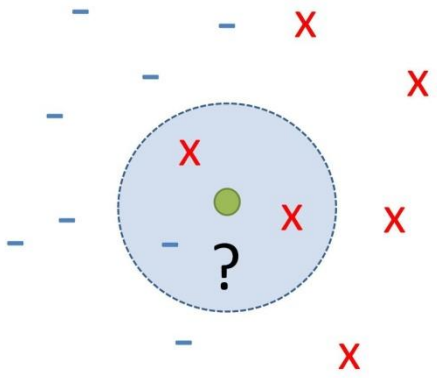



Abbildung 3.14: kNN-Verfahren k=3

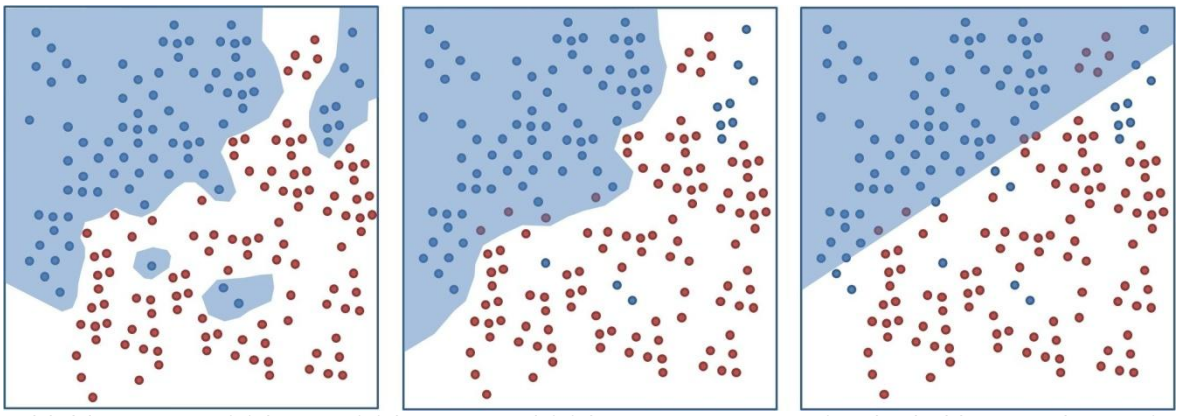


Abbildung 3.15: (1) knn 1, (2) knn 15 und (3) lineare Regression (nach Shakhnarovish 2005)

Zur Bestimmung des k wird häufig eine Kreuzvalidierung durchgeführt. Hierzu werden alle Objekte des Datensatzes in einer Trainingsphase mit bereits gelabelten Daten bei einem unterschiedlichen k-Wert mit allen anderen Objekten des Datensatzes verglichen. Für jedes gewählte k ergibt sich nach der Trainingsphase eine Klassifikationsgüte der richtig zugeordneten Objekte. Der am besten geeignete k-Wert wird dann im Anschluss der Kreuzvalidierung zur Bestimmung der unbekannten Objekte eingesetzt.

Ein Nachteil des kNN-Verfahrens ist die zum Teil sehr hohe Anzahl von irrelevanten Attributen und einer gleichmäßigen Gewichtung dieser Attribute. Ein weiterer Nachteil des Verfahrens ist, dass die Klassifikation sehr ineffizient ist, da für Ähnlichkeitsberechnungen alle Trainingsinstanzen betrachtet werden und bei hochdimensionalen Problemen oft irrelevante Attribute die Ergebnisse der Klassifikation beinträchtigen. Dieses Problem ist auch unter der Bezeichnung „Fluch der Dimensionen" bekannt und wurde von Richard Bellmann formuliert (Bellmann 1961, Mitchell 1997). Der „Fluch der Dimensionalität" besagt, dass zur Beibehaltung einer gewissen Güte der Klassifizierungsleistung die Anzahl der zur Verfügung stehenden Trainingsdaten exponentiell in der Anzahl der Attribute, wachsen muss.

**Subgruppensuche**

Die Subgruppensuche analysiert Abhängigkeiten zwischen einer diskreten Zielvariablen und mehreren erklärenden Variablen. Ziel ist es, Gruppen von Objekten zu erkennen, die eine signifikante Abweichung zum Zielwert in Bezug auf den gesamten Datenbestand besitzen. Es geht bei der Subgruppensuche nicht notwendigerweise um eine Vorhersage, sondern um eine Beschreibung des Zielattributes (Görz et al. 2003). So ist eine Abweichung dann erfüllt, wenn das Zielattribut in einer bestimmten Subgruppe einen überproportional hohen oder niedrigen Anteil am spezifischen Zielwert besitzt.

Beispiele für unterschiedliche Subgruppen im Versicherungsbereich:

- Unter den jungen Single Männern in ländlichen Gebieten ohne Studienabschluss ist der Anteil an Lebensversicherungen signifikant niedriger als im gesamten Kundenbestand.
- Verheiratete Männer im ländlichen Gebiet erzeugen 5% der Lebensversicherungssumme. Die Subgruppe der verheirateten Männer im ländlichen Gebiet mit Studienabschluss und wohnhaft in Einfamilienhäusern machen nur 5% der Versicherungskunden aus, erzeugen aber 21% der Lebensversicherungssumme.

Wir können die Subgruppe als einen Teilbereich einer Population bezeichnen. Es sei $X$ der Instanzraum mit einer Wahrscheinlichkeitsverteilung von $D$ und $L_h$ ein Hypothesenraum, in dem jede Hypothese als Extension eine Teilmenge von $X$ hat:

$$ext(h) \subseteq X \text{ für alle } h \in L_h$$

Es sei weiterhin

$$S \subseteq X$$

eine gegebene, gemäß D gezogene Subgruppe der Population. Es sei schließlich $q$ eine Funktion

$$q := L_h \to \mathbb{R}$$

Die Auffälligkeit einer Subgruppe wird durch eine sogenannte „Qualitätsfunktion" gemessen, die auffälligeren Subgruppen einen höheren Wert zuordnet als weniger auffälligen Subgruppen (Wrobel 1997). Um die statistische Auffälligkeit einer durch eine Hypothese $h$ beschriebenen Subgruppe der Größe $n := |ext(h)|$ zu bewerten, wird die Wahrscheinlichkeit $p_h$ betrachtet, dass ein Objekt aus der Subgruppe die gesuchten Eigenschaften aufweist. Es werden die Subgruppen betrachtet, die sich hinsichtlich einer zufällig ausgewählten Stichprobe der Gesamtpopulation unterscheiden. Die Qualität q der Subgruppe h beschreibt den Differenzanteil des Zielattributes in der Subgruppe p und dem gesamten Datenbestand $p_0$ sowie die Größe n der Subgruppe.

$$q(h) = \frac{|p - p_0|}{\sqrt{p_0(1 - p_0)}} \sqrt{n}$$

Der Qualitätsfunktion kommt in der Subgruppensuche die zentrale Bedeutung zu, sie bewertet eine Hypothese bzw. eine repräsentative Subgruppe $ext(h)$. Je größer die Stichprobe und je größer die Abweichung der Verteilung, desto signifikanter unterscheidet sich die Subgruppe von der Gesamtpopulation. Es erfolgt die Suche also unter zwei Gesichtspunkten. Einmal für die Subgruppen, die eine statistische Auffälligkeit aufweisen, weil sich beispielsweise die Verteilung eines bestimmten Merkmals (Anteil Lebensversicherungen) signifikant von den Verteilungen der Gesamtstichrobe unterscheidet. Der zweite Gesichtspunkt ist die Größe der gefundenen Subgruppe. Diese ist nur interessant, wenn eine bemerkenswerte statistische Auffälligkeit hinsichtlich der Gesamtpopulation aufweist und nicht nur aus einigen wenigen Fällen besteht.

Die Ergebnisse der Subgruppensuche müssen jeweils noch einmal intensiv interpretiert werden. So besteht die Gefahr, dass viele Kombinationen ähnlicher Subgruppen mit in der Ergebnisliste enthalten sind, diese müssen im Anschluss noch einmal manuell überprüft werden.

### Assoziationsregeln

Die Assoziationsregeln im Data Mining zählen zu den populärsten Techniken bei der Suche in Datenbanken. Im Gegensatz zu den vorher vorgestellten Verfahren handelt es sich hierbei nicht um eine Technik des Lernens aus bekannten Beispielen (Bollinger 1996). Ein typisches Anwendungsfeld der Assoziationsregel ist die Warenkorbanalyse. Inzwischen hat fast jeder Supermarkt ein großes Warensortiment mit Produkten unterschiedlicher Preisstufen, Größe und Hersteller. An der Kasse werden diese Artikel mithilfe von Scannerkassen und dem individuellen Strichcode für jeden individuellen Kundeneinkauf erfasst und abgespeichert. So erhält man nicht nur die Informationen, wann und was ein Kunde einkauft, sondern auch die Kombination von Produkten, für die sich der Kunde entschieden hat. Es entstehen in relativ kurzer Zeit bei großen Einkaufsmärkten Datenmengen, die das Kaufverhalten der Kunden beschreiben. Es entsteht schnell die Frage, welche Produkte besonders häufig gekauft

werden, aber auch in welcher Kombination. Wenn es solche Zusammenhänge gibt, dann ist es für den Filialleiter von großem Interesse, die häufig gemeinsam gekauften Produkte auch im Markt in direkter Nähe zu platzieren. Die Assoziationsregeln beschreiben Korrelationen zwischen gemeinsam auftretenden Ereignissen. Es werden einzelne Elemente einer Menge (Items) gesucht, die das Auftreten anderer Items implizit zur Folge haben. Assoziationsregeln lassen sich wie folgt definieren: Sei $I = \{I_1,\ I_2\ ....\ I_n\}$ eine Menge von Items, die Artikel repräsentieren. Eine Transaktion $T$ besteht aus einer Menge von Artikeln, somit gilt $T \subseteq I$. Eine Transaktion $T$ unterstützt $X$ (support $s$), eine Teilmenge $X$ von $I$, wenn $X \subseteq T$ gilt. Mit $D$ wird die Menge der Transaktionen definiert. Eine Regel $X \Longrightarrow Y$ hat einen Support von $s$, wenn ein Anteil der Transaktionen $D$ die Artikel $X \cup Y$ enthalten. Der Support einer gefundenen Regel kann wie folgt ausgedrückt werden:

$$s\ (X \Longrightarrow Y) = \frac{|\{T\ \in D|(X\ \cup Y)\ \subseteq T\}|}{|D|}$$

Die Konfidenz $c$ beschreibt den Prozentanteil der Transaktion $D$ für die gilt: wenn sie $X$ enthält, dann enthält sie auch $Y$.

$$c = \frac{|\{T\ \in D|(X\ \cup Y)\ \subseteq T\}|}{|\{T\ \in D|(X \in T\}|}$$

Weiterhin sei $s_{\min}\ \in\ [0{,}1]$ eine benutzerdefinierte Minimalhäufigkeit. Und $c_{\min}\ \in\ [0{,}1]$ eine benutzerdefinierte Minimalkonfidenz. Die Assoziationsregel hat zwei Bedingungen, die erfüllt werden müssen. Bedingung 1 ist, dass eine bestimmte Minimalhäufigkeit $s_{\min}$ der gefundenen Zusammenhänge bestehen muss. Die zweite Bedingung lautet, dass Artikel mit einer bestimmten Wahrscheinlichkeit zusammen gekauft werden. Dies verlangt, dass von den Transaktionen, die die Prämisse $X$ beinhalten, mindestens ein Anteil $c_{\min}$ auch die Konklusion $Y$ beinhalten soll. Überträgt man dies auf ein praktisches Beispiel einer Warenkorbanalyse mit den Artikeln $I = \{Cola, Chips, Schokolade\}$, und der Assoziationsregel $\{Cola, Chips\}$ $\Longrightarrow \{Schokolade\}$ haben bei einem $s_{min} = 0{,}01$ und einem $c_{\min}\ = 0{,}5$ mindestens 1% der Einkäufer die Artikel Cola, Chips und Schokolade gekauft, und mindestens 50% der Käufer haben bei einem Kauf von Cola und Chips auch Schokolade gekauft (Görz et al. 2003).

**Zusammenfassung**

Zusammenfassend kann gesagt werden, dass KDD-Verfahren auf große und komplexe Datenbestände zurückgreifen, die mit händischen Verfahren kaum noch analysierbar sind. Eine wichtige Aufgabe des KDD ist es, Muster zu erkennen, diese Datenmengen für den Nutzer in eine verständliche und handhabbare Form zu bringen und schlussendlich einen evtl. Wettbewerbsvorteil daraus zu generieren. Dabei ist es wichtig, für die jeweilige Fragestellung das geeignete Verfahren zu wählen und die Ergebnisse in ausreichender Form zu interpretieren. So liefert das Data Mining keine endgültigen Antworten, sondern nützliche Einblicke und Anhaltspunkte. Diese müssen in weiteren Prozessen getestet und überprüft werden.

## 3.5 Datengrundlagen der Modellierung in Deutschland und der Schweiz

Im folgenden Abschnitt werden die für die Arbeit zur Verfügung stehenden Datensätze beschrieben. Hierzu zählen die Straßennetzdaten, der Frequenzatlas und die mit GPS erfassten Mobilitätsdaten. Die Straßennetzdaten dienen dazu, die GPS-Daten in einem später vorgestellten Schritt aufzubereiten (siehe Kapitel 4). Die GPS-Daten stellen die Grundlage für die Leistungswertbestimmung dar und werden hinsichtlich ihrer Erfassungstechnik, der Stichprobengröße, dem Erfassungszeitraum, ihrer räumlichen Verteilung und ihren soziodemographischen Variablen vorgestellt. Der Frequenzatlas bekommt in Kapitel 5 Bedeutung, er dient dazu, mit der räumlichen Unvollständigkeit von GPS-Daten umzugehen.

### 3.5.1 Straßendaten – Vector25 und Navteq

Gerichtete Graphen $G = (K, S)$ bilden den Kern von geometrischen Modellen für Straßennetze im euklidischen Raum $\mathbb{R}^2$. Dabei besteht ein Straßennetz bzw. eine Graphenstruktur aus Knoten $K$ und Kanten $S$, die in diesem Zusammenhang als Segmente bezeichnet werden. Ein Knoten entspricht einem Punkt in der Ebene mit $x, y$ Koordinaten, und ein Segment wird als verknüpfende Verbindung dargestellt. Mithilfe von Stützpunkten können Segmente als Polylinie an individuelle Straßenverläufe angepasst werden. Segmente und Knoten geben in der Regel die realen Straßenbedingungen gut wieder, jedoch bilden sie in erster Linie ein geometrisches Modell, welches die Realität z.T. nur approximieren kann. So wird die Breite einer Straße oder die Ausdehnung einer Kreuzung nicht durch einen Knoten oder ein Segment repräsentiert. Zusätzlich sind viele Knoten im geometrischen Modell an Stellen, wo keine wirkliche Kreuzung existiert, sich jedoch eine Verkehrsrestriktion ändert (z.B. Geschwindigkeit, Gemeindegrenze). Kreuzungen weisen in der geometrischen Repräsentation in der Regel auch eine sehr komplexe Struktur auf. Insbesondere bei Autobahnkreuzungen und Auffahrten werden diese idealisiert, und es werden insbesondere getrennte Fahrrichtungen als einzelne Segmente ausgewiesen. Die genauen Details der geometrischen Repräsentation sind von Anbieter zu Anbieter unterschiedlich und können sogar innerhalb eines Straßennetzes variieren. Zu den größten Anbietern von Straßennetzen gehören weltweit die Firmen Teleatlas und NavTeq (NavTeq 2012, Teleatlas/TomTom 2012).

Abgesehen von den Informationen auf geometrischer Ebene sind noch eine Vielzahl von weiteren Attributen für Wegeberechnungen wichtig. Zu den wichtigsten Informationen gehören Fahrtrichtungsangaben, die Klassifikationen des Straßensegmentes und die zulässige Geschwindigkeit:

**Fahrtrichtungsangabe**

Die Reihenfolge von zwei Knoten eines Segmentes gibt keinen Hinweis auf die zulässige befahrbare Richtung des Autoverkehrs. Diese Information wird typischerweise als zusätzliches Attribut an ein Segment geschrieben. Es zeigt an, von welchem Knoten welche Fahrtrichtung zulässig ist.

**Straßenkategorie**

Die Straßenkategorie oder Funktionsklasse ist eine Klassifikation von Straßensegmenten anhand ihrer Wichtigkeit. Sie bilden eine Hierarchie in dem Sinne, dass das Netzwerk verbunden bleibt, wenn die weniger wichtigen Kategorien entfernt werden.

**Geschwindigkeit**

Diese hängt eng mit der jeweiligen Straßenkategorie zusammen. Je wichtiger die Straßenkategorie für den regionalen Verkehr ist, desto höher ist auch in der Regel die zulässige Geschwindigkeit. Sie dient neben der Verkehrsrestriktion der Fahrtrichtungsangabe als wichtiger Input für die Analyse der Erreichbarkeit von bestimmten Punkten.

Nachfolgend werden die zwei verwendeten Straßennetze für Deutschland und die Schweiz vorgestellt. Beide Straßennetzdaten unterscheiden sich in der Art ihrer Erstellung, der Genauigkeit und ihrer Attribute.

**Vector25**

Das bei der Modellierung in der Schweiz eingesetzte digitale Straßennetz wird vom Schweizer Bundesamt für Landestopographie erstellt. Das Bundesamt ist für die amtliche Vermessung und Kartographie zuständig (Swisstopo 2012). Das Vector25 Straßennetz umfasst das gesamte Straßen- und Wegenetz der Schweizer Landeskarte 1:25000. Die Genauigkeit des Straßennetzes wird damit durch die Qualität der zugrunde liegenden Landeskarte bestimmt, da auf dieser Basis die Digitalisierung der Straßensegmente vorgenommen wird. In der Regel weisen die Punkte und Linien eine Genauigkeit von 3-8 Metern auf. Die Klassifizierung der Straßenkategorien entspricht der Kartenlegende und wurde nicht eigens erhoben. Straßen, die sich niveaugleich kreuzen, besitzen keinen gemeinsamen Knoten. Damit ist ein Routing nur bedingt möglich, da Verkehrsrestriktionen wie Einbahnstraßen, Fahrverbote und Abbiegeverbote in Vector25 nicht enthalten sind. Straßenabschnitte, die Tunnel und Brücken darstellen, sind als solche attribuiert. Insgesamt umfasst das Vector25 Straßennetz knapp 1 Mio. Segmente.

**NavTeq**

Bei der Modellierung in Deutschland wird das digitale Straßennetz der Firma Navteq eingesetzt. Die amerikanische Firma Navteq ist neben der belgisch/niederländischen Firma Teleatlas der weltweit größte Anbieter von digitalen Straßendaten. Mit speziell ausgerüsteten Fahrzeugen fährt Navteq in regelmäßigen Abständen das komplette Straßennetz der Bundesrepublik ab. Neben der mit DGPS aufgezeichneten Wegstrecke werden über eine sprachgesteuerte Aufnahme navigationsrelevante Straßeninformationen mit erfasst. Zusätzlich wird der gefahrene Weg mit einer Digitalkamera aufgezeichnet und im Anschluss mit den Tonaufnahmen abgeglichen. Neben der reinen Geometrieaufzeichnung über das DGPS werden durch die kombinierte Methodik inzwischen insgesamt fast 120 Attribute erfasst. Im Unterschied zum Vector25 existiert an jeder Kreuzung ein Knotenpunkt. Außerdem werden Knotenpunkte eingefügt, wenn es zu einer Attributänderung kommt, wie z.B. die Veränderung der erlaubten Geschwindigkeit. Restriktionen des Straßenverkehrs sind enthalten sowie eine umfangreiche Straßentypologie. Aktuell umfasst das deutsche Straßennetz insgesamt 6,87 Millionen Straßensegmente.

In der folgenden Abbildung 3.16 sind für die Stadt Zürich die beiden Straßennetze Vector25 (gelb) und das NavTeq Straßennetz (schwarz) dargestellt. Neben den Unterschieden in der räumlichen Lage der Segmente erkennt man auch Unterschiede bei der Anzahl der Segmente. So wird in Navteq eine ÖPNV-Straßeninsel exakter dargestellt als bei Vector25. Auch bei der Darstellung von Kreuzungen gibt es im Detaillierungsgrad starke Unterschiede.

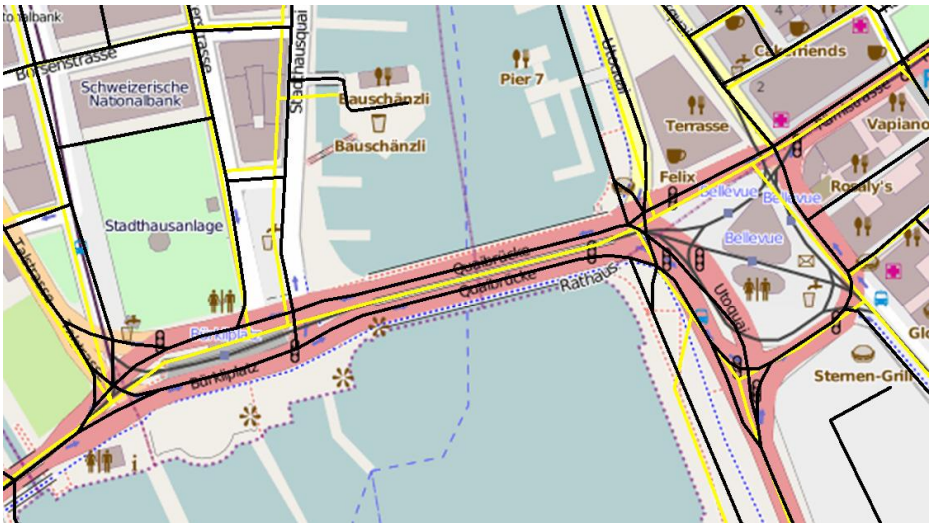



Abbildung 3.16: Vergleich Vector25 und NavTeq (topographische Hintergrundinformation OSM 2012)

### 3.5.2 FREQUENZATLAS

Der Frequenzatlas ist ein Datenprodukt, das im Auftrag des Fachverbandes für Aussenwerbung e. V. (FAW) seit 2003 vom Fraunhofer IAIS erstellt wird. Er ist die Lösung der Aufgabe, aus punktuellen Verkehrszählungen einer Stadt Verkehrsfrequenzen für das komplette Straßennetz der Stadt zu schätzen. Der Frequenzatlas ist Grundlage für die quantitative Kontaktwertberechnung (G-Wert) von Plakatflächen (vgl. Abschnitt 2.1), er liefert die geschätzte Gesamtzahl von passierenden PKW, Fußgängern und ÖPNV-Nutzern pro durchschnittlicher Stunde.

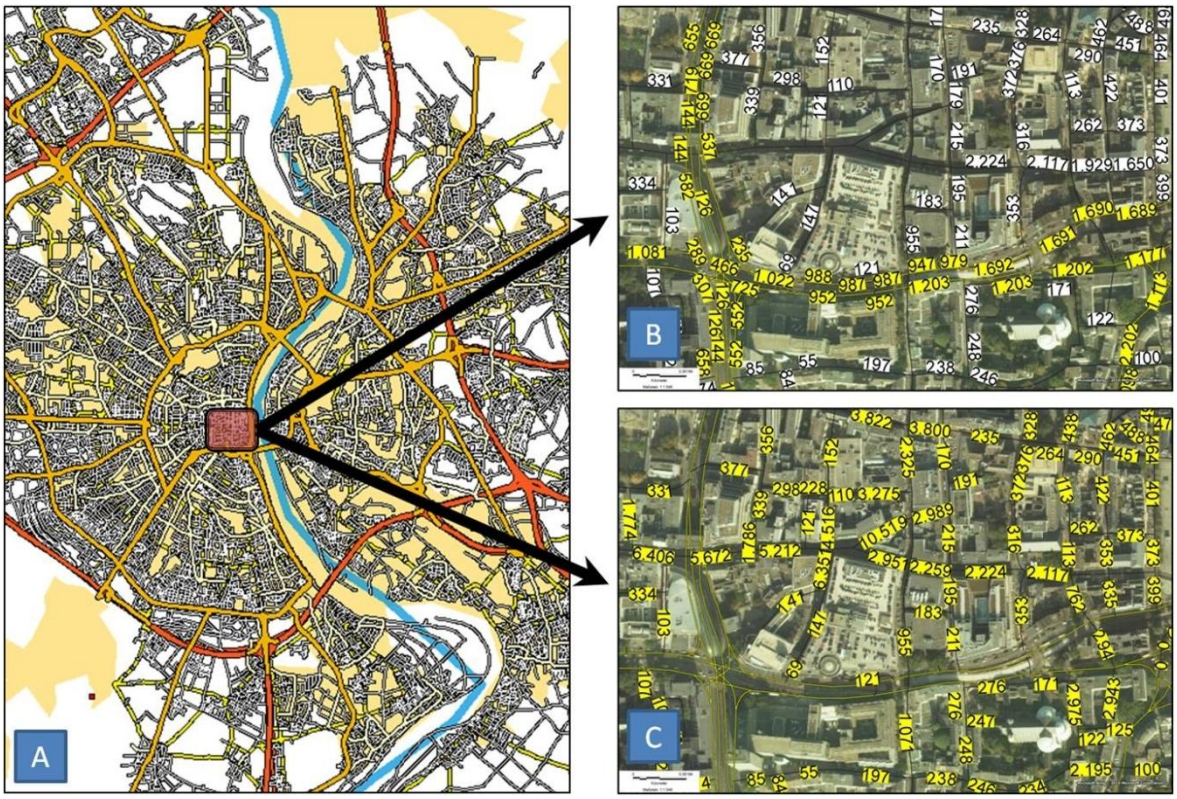



Abbildung 3.17: Frequenzatlas Köln PKW (A) Navteq Segmente Köln (B) PKW-Frequenzen (C) Fußgängerfrequenzen (topographische Hintergrundinformation Google 2012)

Hauptdatengrundlage für den Frequenzatlas sind Verkehrszählungen öffentlicher Ämter sowie eine große Anzahl von Videozählungen an Plakatstandorten (~100.000), die im Auftrag des Fachverbandes für Außenwerbung e. V. erhoben wurden. Weitere wichtige Bestandteile für die Modellierung des Frequenzatlas sind u.a. die Attributierung des NavTeq Straßennetzes (Straßenklassifikation, erlaubte Geschwindigkeit, Anzahl Fahrspuren, etc..), Points of Interest (Parkplätze, Bahnhöfe, Restaurants, etc.) und Gemeindedaten. Fraunhofer IAIS hat für die Entwicklung des Frequenzatlas ein Nächste-Nachbar-Verfahren entwickelt s-kNN, das auf Spatial Data Mining Algorithmen beruht (May et al. 2008a, May et al. 2008b). Grundsätzlich wird dabei die Frequenzschätzung als attribut-basierte Lernaufgabe verstanden, bei der die Instanzen Straßenabschnitte darstellen und die Klassenattribute die gemessene Straßenfrequenz. Die Attribute werden über die NavTeq Informationen des Straßensegmentes und die POI Informationen gebildet, die sich im Einzugsgebiet des Straßensegmentes befinden. Nächste-Nachbar-Verfahren sind in der Lage, mit räumlichen und nicht-räumlichen Informationen umzugehen, wenn angemessene Distanzfunktionen definiert sind. So kann insbesondere die räumliche Autokorrelation bei diesem Verfahren ausgenutzt werden. Der s-kNN wurde so angepasst, dass er mit Vektor-Geometrien und komplexen räumlichen Datenstrukturen umgehen kann, anstatt mit Punkt Messungen, wie es üblicherweise ein kNN macht (vgl. Abschnitt 3.3). Der s-kNN modelliert den geographischen Raum als Teilkomponente des allgemeinen Attribut Raumes. Hierzu wird die Distanz zwischen zwei Straßensegmenten $x_a$ und $x_b$ als die normierte Summe der absoluten Abstände ihrer Attribute definiert.

$$d(x_a, x_b) = \sum_{i=1}^{m} |x_{ai} - x_{bi}|$$

Zur Feinabstimmung des s-kNN werden den Attributen Gewichte zugeordnet. Die Frequenz $y_0$ eines Straßensegmentes wird berechnet als die gewichtete Summe der Frequenzen der k nächsten Nachbarn. Jedes Gewicht ist dabei indirekt proportional zum Abstand zwischen zwei Segmenten.

$$y_0 = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i} \; gegeben \; w_i = \frac{1}{d(y_0, x_i)}$$

Die Auswahl der Attribute und deren Gewichtung hängt im Speziellen von der Dichte der Verkehrsmessungen in der Region und der Art des zu prognostizierenden Verkehrs ab (PKW, Fußgänger, öffentlicher Verkehr). Dabei kommt der Dichte der Verkehrsmessungen, der Anzahl an Straßensegmenten und der Menge an Attributen eine besondere Bedeutung zu. Der kNN Algorithmus ist dafür bekannt, ein sehr rechen-, bzw. zeitintensives Verfahren zu sein. Aus diesem Grund muss ein geeignetes Optimierungsverfahren entwickelt werden. So ergibt sich ohne Optimierung für eine Stadt wie Frankfurt eine Anzahl von 43 Millionen Berechnungen bei 21.500 Segmenten und ~2.000 Messungen. Um die zeitintensiven räumlichen Berechnungen zu reduzieren, ist im s-kNN ein partielles Bewertungsschema implementiert. Während numerische Attribute in der Regel sehr schnell berechnet werden können, sind räumliche Entfernungen zwischen Liniensegmenten sehr rechenintensiv. Aus diesen Gründen wurde der s-kNN so implementiert, dass er selektiv und dynamisch die Berechnung jeder Distanz eines Straßensegmentes zu den verschiedenen gemessenen Straßensegmenten durchführt. Und zwar werden zu jeder Zeit nur die besten k Nachbarn gespeichert und gegebenenfalls dynamisch während der Iteration über die Messstellen ersetzt. Die Berechnung der Distanz wird stufenweise durchgeführt. Wenn die zusammengefassten Entfernungen aller nicht räumlichen Attribute bereits über der maximalen Distanz der aktuellen k Nachbarn liegen, kann der Kandidat als Nachbar sicher verworfen werden, und es muss keine räumliche Berechnung vorgenommen werden. Falls

die Distanz jedoch geringer ist, wird zunächst eine Abschätzung mit Hilfe eines festgelegten räumlichen Begrenzungsrechtecks der Segmente vorgenommen. Die Ausdehnung durch das Begrenzungsrechteck ist eine untere Schranke für die tatsächliche Entfernung zwischen den möglichen Segmenten, gleichzeitig wird damit auch die Rechenzeit minimiert. Wenn der Abstand der nicht räumlichen Attribute sowie der Abstand des Begrenzungsrechtecks größer oder gleich der maximalen Distanz der besten k Nachbarn liegt, kann das Segment verworfen werden. Nur wenn beide Tests bestanden sind, wird die tatsächliche räumliche Distanz bestimmt. Dadurch wird die Berechnungszeit reduziert und der erforderliche Speicherplatz erheblich verkleinert. Die Ergebnisse des Verfahrens sind dann Gegenstand weiterer Nachbearbeitungen. So kennt das s-kNN Verfahren weder Verkehrsflüsse noch deren Richtungen. Daher kann z.B. bei einer Kreuzung mit vier einfließenden Straßensegmenten die Verkehrssumme unplausibel sein. In einem zusätzlichen Schritt wird diesem Umstand Rechnung getragen, indem über ein zusätzliches Optimierungsverfahren die Kirchhoffschen Regeln erzwungen werden. Diese besagen, dass die Summe der einfließenden Verkehrsströme auch die Summe der ausfließenden Verkehrsströme sein muss. Auf der linken Seite in der Abbildung 3.18 sehen wir eine komplexe Kreuzung nach der kNN Berechnung. Wir sehen, dass das Gesamtbild der Knoten nicht plausibel aussehen. So werden insbesondere die Frequenzen im inneren Bereich der Kreuzung mit der Summe 815 durch die identische Attributierung vom kNN nicht unterschieden. Damit ein plausibleres Kreuzungsbild entsteht, wird durch ein Optimierungsverfahren mit mehrfachen Iterationen die Kreuzungsfrequenz überarbeitet. Im Ergebnis werden z.B. die Frequenzen im Inneren der Kreuzung angepasst, und die Bedingungen von Kirchhoff werden erfüllt, wie es im rechten Bild zu erkennen ist.

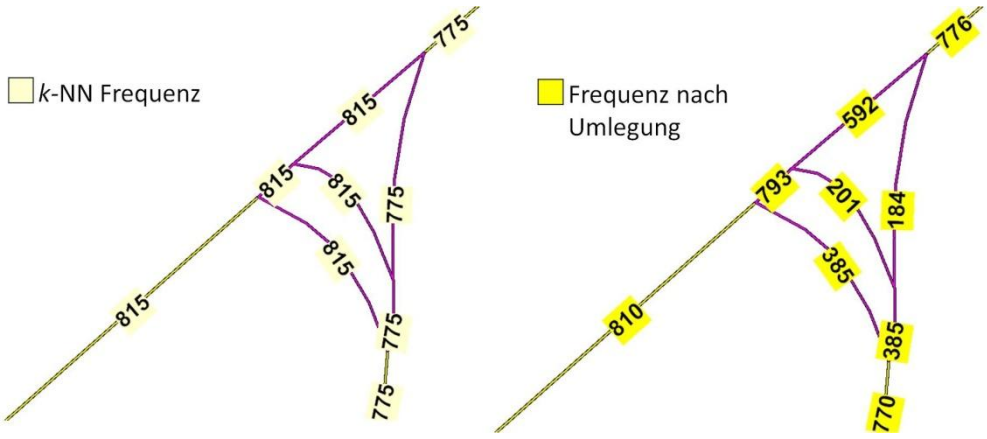



Abbildung 3.18: Kreuzungsumlegung nach kNN

Die Vorteile des entwickelten s-kNN Verfahrens sind, dass die Vorgabewerte an den Zählstellen stets eingehalten bleiben und plausible Mittelwerte bei räumlicher Aggregation an den zu prognostizierenden Straßensegmenten modelliert werden. Beim s-kNN Verfahren handelt es sich um eine *„gutmütige"* Interpolation. Es neigt nicht dazu, Extremwerte zu erzeugen, da mehrere Nachbarn zur Bildung der Frequenz herangezogen werden. Als Resultat des Frequenzatlas werden inzwischen für alle knapp 7 Millionen Segmente des Navteq Straßennetzes in Deutschland Frequenzen für die Mobilitätstypen PKW, Fußgänger und ÖPNV ausgewiesen.

### 3.5.3 GPS UND CATI FELDSTUDIE DEUTSCHLAND

Die Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma)[3] ist eine Vereinigung der Werbetreibenden (z. B. Radio, Fernsehen, Außenwerbung, usw.), Werbeagenturen und Verlage, die Interesse an der Analyse des Mediennutzungsverhaltens von verschiedenen Zielgruppen haben. Ziel der ag.ma ist es, als übergeordnete Medieninstanz objektive Medialeistungswerte für alle Mediengattungen zu publizieren. Durch die von der ag.ma in Auftrag gegebenen Studien und Analysen können Werbeleistungen innerhalb eines Mediums oder zwischen verschiedenen Medien verglichen werden. In den Jahren 2006/2007 führte die ag.ma eine bundesweite Mobilitätsstudie im Auftrag der Außenwerbung durch (ag.ma 2012a). Ziel war es, für jeden Plakatstandort in Deutschland einen objektiven Wert hinsichtlich der Leistungswerte (vgl. Abschnitt 2.1) ausweisen zu können. Die Studie wurde unter Verwendung von zwei unterschiedlichen Erhebungsmethoden durchgeführt:

1. CATI = **C**omputer **A**ssisted **T**elephone **I**nterview
2. GPS = **G**lobal **P**ositioning **S**ystem

In der folgenden Tabelle ist der Stichprobenumfang der Studie aufgeführt.

| Erfassungsmethode | Welle 2007 | Welle 2009 | Welle 2010 | Summe |
|---|---|---|---|---|
| CATI | 21.125 | 9.885 | 10.096 | 41.106 |
| GPS | 8.595 | 3.175 | - | 11.770 |
| Summe | 29.720 | 13.060 | 10.096 | 52.876 |

Tabelle 3.3: Stichprobenumfang Deutschland

Beide Erhebungen sind als rollierende Erhebungen konzipiert. So wird in den kommenden Jahren die Stichprobe permanent erweitert und erneuert. Aktuell wird diskutiert, Mobilitätsdaten, die älter als 6 Jahre sind, aus der Stichprobe herauszunehmen und durch Neuerhebungen zu ersetzen.

| | Welle 2007 | Welle 2009 | Welle 2010 |
|---|---|---|---|
| CATI | 06.08.2006-18.02.2007 | 16.02.2009-09.08.2009 | 12.04.2010-30.01.2011 |
| GPS | 06.08.2006-18.02.2007 | 16.02.2009-09.08.2009 | - |

Tabelle 3.4: Erfassungszeiträume GPS & CATI

Die jeweiligen Erfassungswellen bei GPS und CATI wurden so angelegt, dass sie über die Jahre 2006-2009 ein komplettes Erfassungsjahr abbilden, siehe Tabelle 3.4.

**CATI-Erhebung**

Bei der Erfassungsmethode über CATI wurden inzwischen insgesamt 41.106 Personen deutschlandweit über ihre Mobilität am Vortag interviewt. Die hohe Fallzahl des CATI Ansatzes beruht darauf, dass man eine flächendeckende Befragung des Bundesgebietes aus Kostengründen nicht alleine mit GPS erreichen konnte. Die Zielpersonen der Telefonbefragung wurden über einen Zufallsprozess ausgewählt. Dabei handelt es sich nur um Haushalte, die über einen Festnetzanschluss verfügen. Die mit telefonischer Befragung über Festnetz erreichbare Grundgesamtheit ist in keinem amtlichen Telefonverzeichnis vollständig verzeichnet, so dass weder individuelle noch statistische Angaben hierzu existieren (ag.ma 2012c). Um trotzdem eine bevölkerungsproportionale Verteilung im Bundesgebiet zu gewährleisten, werden die Telefonnummern nach den Vorwahlen ihrer Regionen sortiert. Je 20 Nummern aus einer Region werden zu einem Block zusammengefasst und nur eine

[3] http://www.agma-mmc.de

Nummer aus diesem Block bzw. ein Haushalt aus diesem Block angerufen. Die Auswahl der Zielperson im Haushalt erfolgt nach einer Auflistung aller im Haushalt befindlichen Personen. Ist die Person ausgewählt, wird sie am Telefon zu ihrer gestrigen Mobilität befragt. Sollte eine Person am gestrigen Tag nicht außer Haus gewesen sein, wird der Tag davor abgefragt. Unterstützt von einer kartenbasierten Routingsoftware (TripTracer) fragt dann der Interviewer jeweils nach Start- und Zielpunkt, dem Verkehrsmittel und der Dauer eines Weges (DDS 2012). Er gleicht jeweils die Start- und Zielpunkte mit seinen POI-Informationen ab und überprüft, ob evtl. Unstimmigkeiten bei der Befragung auftreten oder bestimmte Punkte nicht gefunden werden. Zusätzlich zu den Mobilitätsinformationen werden noch Merkmale zu Einkommen, Haushaltsstruktur, Mediennutzung und Freizeitaktivitäten erfasst.

**GPS-Erhebung**

Im zweiten Teil der Studie wurden insgesamt 11.770 Personen aus 42 Gemeinden in Deutschland mit einem GPS-Gerät ausgestattet. Über einen Zeitraum von 7 Tagen wurden ihre Wege aufgezeichnet. Die Personen der Stichprobe wurden in einem zweistufigen Verfahren ausgewählt. In einem ersten Schritt wird über die Auswahl von Rasterzellen innerhalb einer Gemeinde, sogenannter Sample Points, ausgewählt, wo Probanden rekrutiert werden. Die Sample Points dienen dazu, eine räumliche Streuung der Probanden innerhalb der Gemeinden zu gewährleisten. Im zweiten Schritt werden über ein Quotenverfahren die Probanden in den jeweiligen Gemeinden ausgewählt. Merkmale für das Quotenverfahren sind: Alter, Geschlecht, Anzahl der Personen im Haushalt (Single, Familienhaushalte), Ausbildung und Berufstätigkeit. Ziel ist es, über das Quotenverfahren ein verkleinertes Abbild der deutschsprachigen Bevölkerung ab 14 Jahren in Privathaushalten am Ort ihrer Hauptwohnung zu schaffen (ag.ma 2012c). Nach Festlegung der Quoten wurden von der ag.ma unterschiedliche Marktforschungsinstitute mit der Rekrutierung der Probanden beauftragt. Diese haben entweder mittels vorheriger telefonischer Befragung oder direkt per Hausbesuch die Probanden für die Teilnahme an der Studie gewonnen. Wurde ein Proband rekrutiert, wurde zu Beginn ein umfangreiches Interview durchgeführt. Neben Mobilitätsfragen wurden analag zu CATI Fragen zu Freizeitaktivitäten, Haushaltsausstattung, Einkaufsverhalten und Mediennutzung gestellt. Am Ende des Interviews wurde den Probanden ein GPS-Gerät übergeben. Dieses sollte der Proband über einen Zeitraum von 7 Tagen bei allen seinen außerhäuslichen Wegen mit sich tragen. Das Gerät ist nicht größer als ein Mobilfunktelefon (Abb. 3.19), und die Akkulaufzeit beträgt knapp 28 Stunden. Das heißt, dass das Gerät in regelmäßigen Abständen aufgeladen werden muss. Im Sekundentakt werden GPS-Koordinaten durch das Gerät gesammelt. Im Anschluss an die Tragewoche wird das GPS-Gerät vom Marktforschungsinstitut wieder abgeholt, und es wird in einem kurzen Interview nach Unregelmäßigkeiten während der Tragewoche gefragt. Danach werden die Daten vom GPS-Gerät ausgelesen und für die weiteren Nachbearbeitungsschritte zur Verfügung gestellt.

Abb. 3.19: Mobitest (MGE 2012)

Schaut man sich die Stichprobenverteilung von CATI und GPS über Deutschland hinweg an, so fällt schnell die räumliche Verteilung auf. In Abbildung 3.20 ist auf der rechten Seite die fast flächendeckende Verteilung der CATI Gemeinden über Deutschland zu erkennen. Die Erfassung mittels CATI stellt die kostengünstigere Variante der beiden Erhebungsmethoden dar und wird aus diesem Grund auch in den kommenden Jahren noch eine wichtige Grundlage in der Mobilitätserfassung der Außenwerbung spielen.

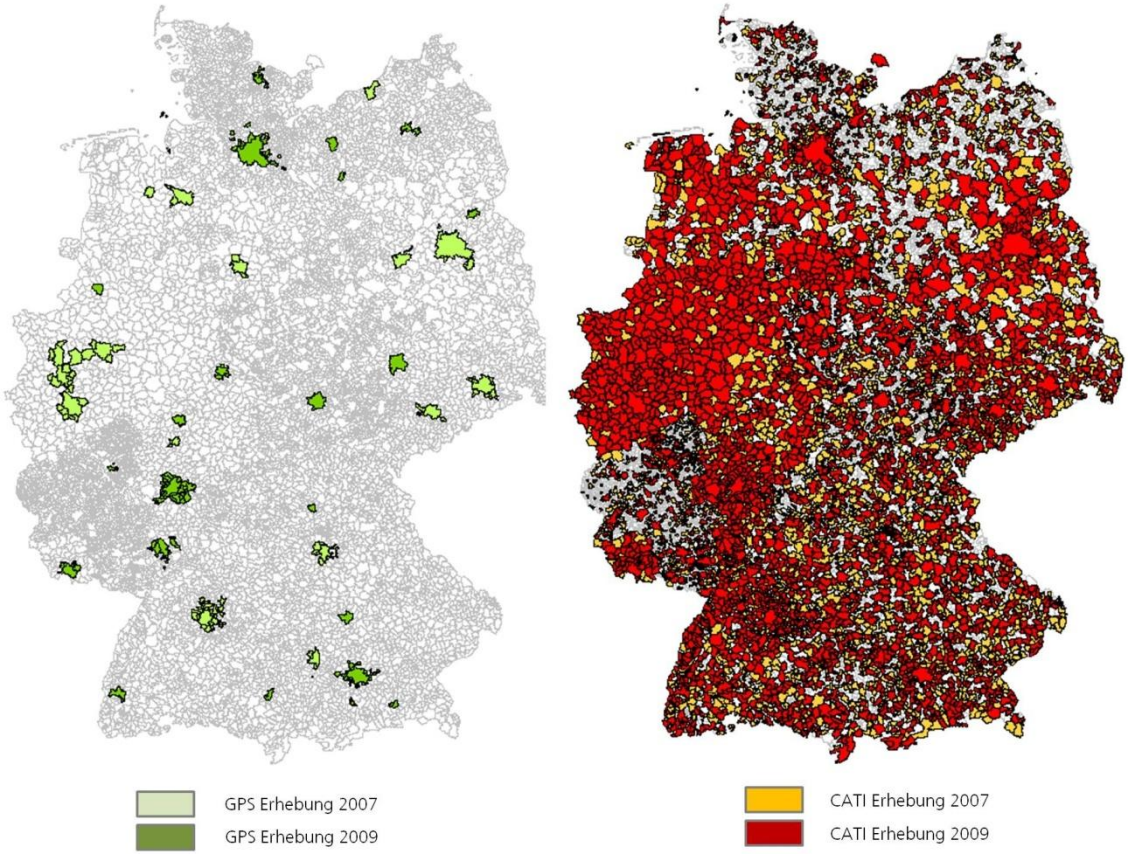



Abbildung 3.20: ag.ma Stichprobenverteilung Deutschland

Die GPS-Gemeinden sind auf der linken Seite dargestellt. Sie sind wie bereits beschrieben im Wesentlichen auf die bevölkerungsstarken Gemeinden in Deutschland konzentriert. Jedoch sind auch kleinere Gemeinden in der GPS-Erhebung enthalten, wie Brandenburg an der Havel, Lahnstein, Eberswalde, etc. Nimmt man die CATI und GPS-Stichprobe mit ihren knapp 53.000 Befragten zusammen, so handelt es sich um eine der größten Mobilitätsstichproben in Deutschland. Nur die MID 2008 des Bundesministeriums für Verkehr, Bau und Stadtentwicklung kommt mit ihren knapp 25.000 Haushalten in eine ähnliche Stichprobendimension. Die Befragungen fanden hier in einem Methodenmix aus schriftlichen und telefonischen Interviews statt. GPS wurde in dieser Studie nicht eingesetzt. Auch wurde bei der MID 2008 keine exakte Geokodierung der Wohn- und Arbeitsorte vorgenommen, sondern es wurde nur auf der relativ groben Ebene von Gemeinden gearbeitet. Aus diesem Grund existieren an dieser Stelle auch keine Wege auf Straßenvektoren zwischen den einzelnen Aufenthaltsorten der interviewten Personen. In der vorliegenden feinen, räumlichen Auflösung stellen die ag.ma Mobilitätsdaten, insbesondere die GPS-Daten, eine Datengrundlage dar, die auch jenseits des Anwendungskontextes von großem Interesse ist.

### 3.5.4 GPS- FELDSTUDIE SCHWEIZ

Swiss Poster Research Plus (SPR+)[4] ist seit dem Jahr 2000 eine Forschungseinrichtung der Schweizer Aussenwerbung. Für das Medium Plakat erforscht SPR+ neue Ansätze und Wege zur validen und transparenten Ausweisung von Leistungswerten. Die erste Pilotstudie mit GPS wurde bereits im Jahre 2002-2004 in der Agglomeration Winterthur durchgeführt. Insgesamt 630 Personen wurden in einer repräsentativen Bevölkerungsstichprobe mit einem GPS-Gerät

[4] http://www.spr-plus.ch

für 7-10 Tagen ausgerüstet. Nach dem erfolgreichen Ersteinsatz wurden in den folgenden Jahren in insgesamt 12 Schweizer Regionen GPS-Erhebungen durchgeführt (siehe Tabelle 3.5).

| Agglomeration | Einwohnerzahl der Agglomeration | Anzahl der Probanden | Erhebungsjahr |
|---|---|---|---|
| Winterthur | 103.213 | 630 | 2002 - 2004 |
| Zürich | 918.676 | 1.800 | 2004 - 2006 |
| Genf | 390.672 | 1.080 | 2005 - 2006 |
| Basel | 407.957 | 1.080 | 2005 - 2006 |
| Chur | 55.018 | 270 | 2006 |
| Sierre | 69.626 | 270 | 2006 |
| Sion | 75.644 | 360 | 2006 |
| Biel | 103.324 | 630 | 2006 - 2007 |
| St. Gallen | 122.276 | 720 | 2006 - 2007 |
| Luzern | 164.690 | 810 | 2006 - 2007 |
| Bern | 298.757 | 1.080 | 2007 |
| Lausanne | 260.206 | 1.080 | 2007 - 2008 |
| Summe | **2.970.059** | **9810** | **2002-2008** |

Tabelle 3.5: GPS-Stichproben in der Schweiz

Im Unterschied zum deutschen Erfassungsmodell wurde in der Schweiz rein auf die GPS-Technologie gesetzt. Ein weiterer Unterschied ist, dass keine schweizweite Erfassung durchgeführt wurde, sondern diese auf die wichtigsten Schweizer Agglomerationen beschränkt wurde. Die Agglomerationen werden vom Schweizer Bundesamt für Statistik (BFS) nach unterschiedlichen Kriterien definiert, damit ein räumlicher Vergleich zwischen institutionell unterschiedlich abgegrenzten städtischen Gebieten möglich wird. Zu diesen Kriterien gehören die Einwohnerzahl und die Bevölkerungsentwicklung, der bauliche Zusammenhang, das Verhältnis der Erwerbstätigen zur Wohnbevölkerung, die Wirtschaftsstruktur und die Verflechtung mit der Kernzone durch Pendler. Dabei ist das Kriterium des Wegpendleranteils in einer Kernzone eines der wichtigsten für die Zuordnung der Gemeinden zum Agglomerationsraum (Bundesamt für Statistik BFS 2011). Die erfassten Agglomerationen decken ~ 2/3 der Schweizer Bevölkerung ab. In der Abbildung 3.21 sind die 12 Agglomerationen dargestellt.

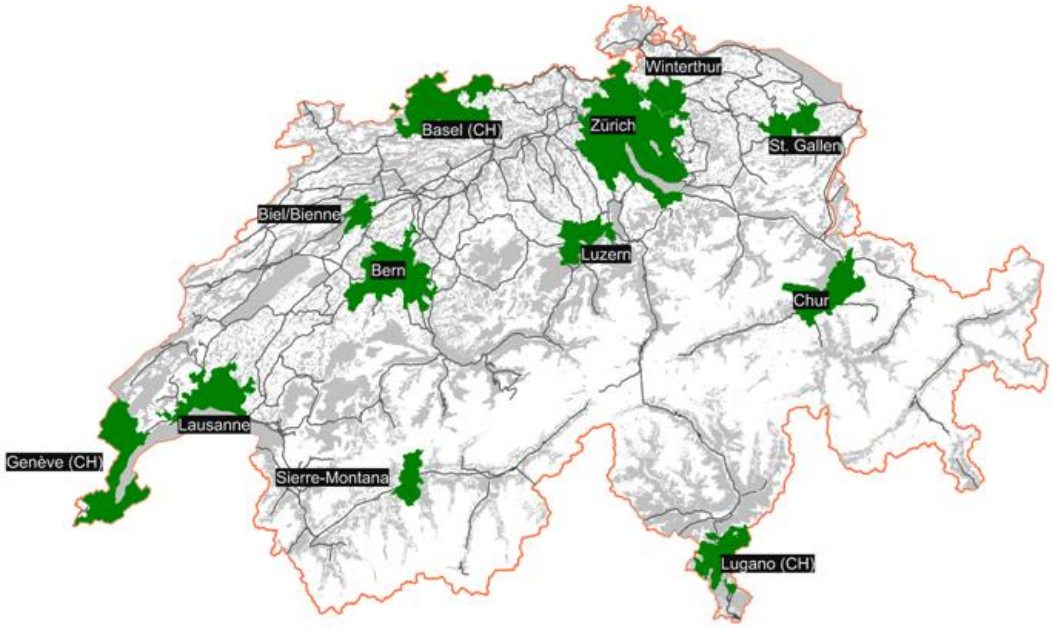



Abbildung 3.21: Agglomerationen in der Schweiz mit GPS-Stichprobe

Die Probandenauswahl erfolgte nach einem Quotenverfahren über die Merkmale Altersgruppen und Geschlecht. Um eine räumliche Verteilung der Stichprobe sicherzustellen, wurden für die Agglomerationen sogenannte Geozellen definiert. Sie stellen geographische Räume innerhalb einer Agglomeration dar, deren Bevölkerung in Bezug auf das Untersuchungsgebiet möglichst ähnliche Strukturen aufweisen. Hierzu wurden die Geozellen von SPR+ so gebildet, dass eine möglichst homogene Bebauungsstruktur, eine ähnliche Bevölkerungsstruktur und eine gemeinsame Verkehrsanbindung des übergeordneten Straßennetzes (Bundesstraßen und Autobahnen) gegeben ist. In Abbildung 3.22 sind die insgesamt 12 Geozellen für Bern und 20 Geozellen für Zürich mit den Hauptverkehrsachsen dargestellt.

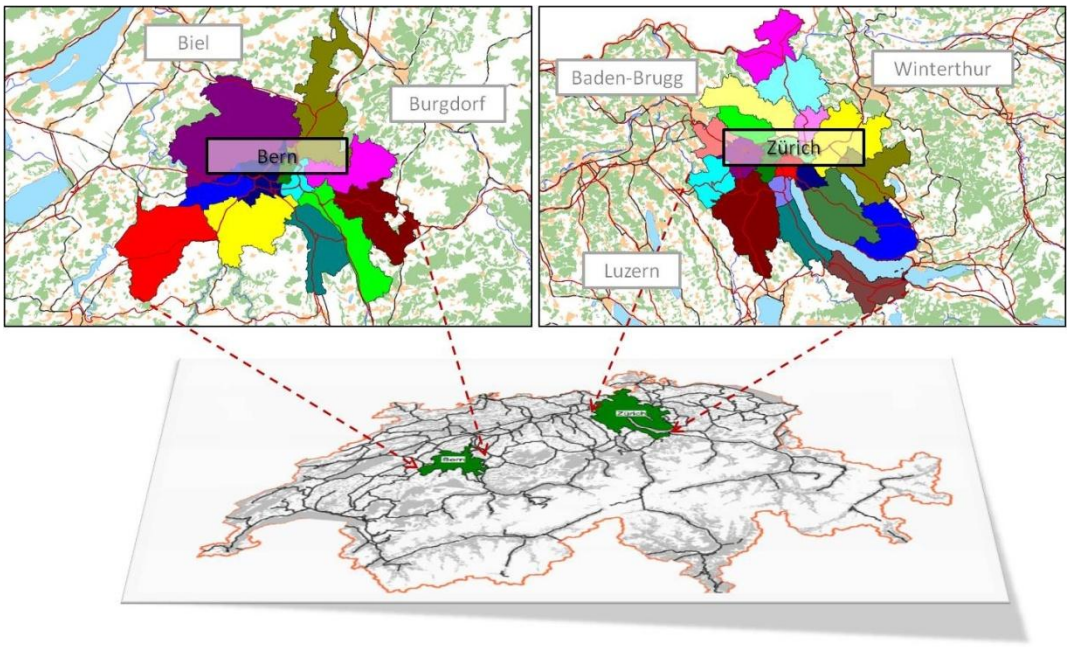



Abbildung 3.22: Geozellen (1) Bern und (2) Zürich

Pro Geozelle wurden jeweils nach drei Altersklassen (14-29 Jahre, 30-45 Jahre, >46 Jahre) und Geschlecht 90 Personen rekrutiert. Das heißt, dass in jeder Altersklasse 15 Männer und Frauen befragt worden sind. Zu den Fragen des beauftragten Marktforschungsinstitutes zählten neben dem Medienverhalten soziodemographische Variablen, das Mobilitätsverhalten, die Ausbildung und der aktuelle Berufsstand. Die Mobilität der Probanden wurde mittels des von SPR+ im Auftrag entwickelten „Mobilitymeters" aufgezeichnet. In Abbildung 3.23 ist das GPS-Gerät dargestellt. Wie das deutsche Gerät, musste auch dieses GPS-Gerät in regelmäßigen Abständen von den Probanden aufgeladen werden. Die Akkulaufzeit betrug ~ 20 Stunden bei einer sekündlichen Aufzeichnung der Position. Im Anschluss an die Erhebung, die zwischen 7 und 10 Tagen gedauert hat, wurde das GPS-Gerät von dem Marktforschungsinstitut wieder abgeholt. Das Ergebnis der SPR+ Erhebung sind im Durchschnitt 200.000 aufgezeichnete Wegpunkte pro Person und insgesamt 2,3 Millionen aufgezeichnete Wegekilometer für alle Probanden. In Abbildung 3.24 sind die aufgezeichneten GPS- Trajektorien der Luzerner (links) und der Züricher Probanden (rechts) für die komplette Schweiz abgebildet. Eine vergleichbar große Mobilitätsstudie mit GPS liegt in der Schweiz nicht vor.


Abbildung 3.23: Mobilitymeter (SPR+ 2012)

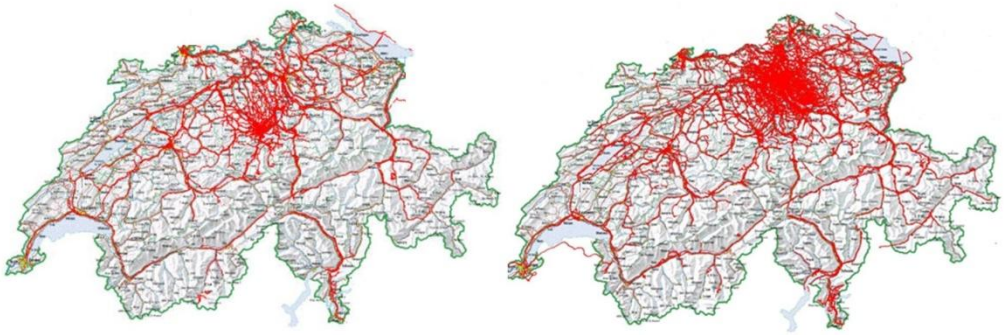


Abbildung 3.24: GPS-Aufzeichnungen (1) der Luzerner und (2) Züricher Probanden in der Schweiz (SPR+ 2012)

Die Berechnung von der reinen Passage zum Plakatkontakt ist in der Schweiz relativ einfach strukturiert. Wie in Abbildung 3.25 dargestellt, gibt es nur insgesamt 4 Gewichtungskriterien für die Kontaktberechnung:

- Durchgangswinkel
- Passagengeschwindigkeit
- Häufung (Anzahl Plakatflächen in der direkten Nachbarschaft)
- Tageszeit

Eine frontale und parallele Passage wird mit 1 gewichtet. Falls der Proband jedoch mit mehr als 10 km/h parallel an der Plakatstelle vorbeikommt, wird der Kontakt um 0,7 auf 0,3 reduziert. Eine weitere Reduktion wird über die Anzahl von Plakatflächenhäufungen vorgenommen. Befinden sich bis zu vier Flächen in direkter Nähe einer Plakatstelle, wird der Kontaktwert mit 0,5 multipliziert und weiter abgewichtet.

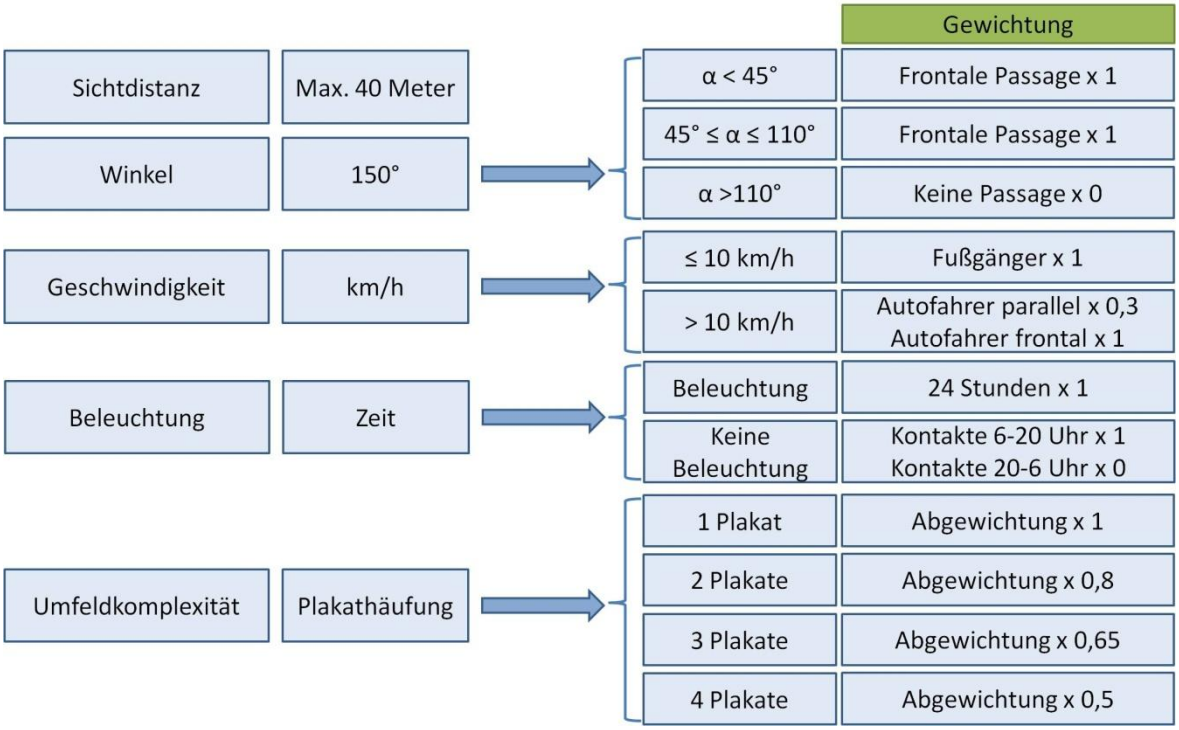



Abbildung 3.25: Gewichtungskriterien in der Schweiz (SPR+ 2011)

**Berechnungsbeispiel:**

Eine parallele Passage eines Autofahrers bei Tag mit einer Geschwindigkeit von mehr als 10 km/h durch den Beachtungsraum einer Fläche, die mit drei anderen Flächen an einer Plakatstelle steht, ergibt einen Plakatkontakt von 0,15 = 1 Passage x 0,3 (für parallel mit >10 km/h) x 0,5 (für Häufung mit 3 Nachbarbarflächen).

Wie die einzelnen Parameter in den späteren Berechnungsprozess einfließen, wird in Kapitel 5 erläutert.

## 3.6 Zusammenfassung

In diesem Kapitel wurde eine Übersicht zur räumlichen und Mobilitätsdatenanalyse gegeben. Zusätzlich wurden die einfließenden Datenquellen der Dissertation vorgestellt. Im Detail hatte das Kapitel folgenden Inhalt:

Im ersten Teil wurde ein Überblick über Datenmodelle, Strukturen und Charakteristiken von Geo- und Mobilitätsdaten gegeben. Dabei wurden insbesondere die topologischen Relationen von Polygonen und Trajektorien vorgestellt, da diese im späteren Anwendungskontext hinsichtlich der Erstellung von Plakatsichtbarkeitsräumen und der Passagenberechnung eine wichtige Rolle in der Kontaktwertberechnung spielen. Die geometrische Modellierung ist für die Ermittlung des Abstandes, der relativen Lage von Objekten zueinander und für ein späteres Routing zwischen einzelnen GPS-Koordinaten wichtig. Für die weitere Arbeit wichtig sind, die Definition von Trajektorien sowie die Charakteristik der menschlichen Mobilität, die eine starke Periodizität und eine Beschränkung auf wenige besuchte Orte zeigt. Zusätzlich wurden in diesem Abschnitt die Annotation und verschiedene Analysemethoden von Trajektorien präsentiert. Abgeschlossen wurde die Beschreibung von Mobilitätsdaten mit einer Auflistung, Gruppierung und Bewertung von Einsatzmöglichkeiten unterschiedlicher Mobilitätserfassungssysteme.

Im zweiten Teil des Kapitels wurden die Grundlagen des Knowledge Discovery in Databases und vier ausgewählte KDD-Verfahren vorgestellt. Hintergrund für die Entwicklung und den Einsatz dieser Methoden sind die enorm gestiegenen Datenmengen, die manuell nicht mehr zu bearbeiten sind. Das KDD wird beim vorgestellten Frequenzatlas und bei der später verwendeten Suche nach Stichprobenverzerrungen eingesetzt. Abgeschlossen wurde das Kapitel mit der Beschreibung der einfließenden Datengrundlagen der Dissertation. Hierzu zählen die Straßendaten NavTeq und Vector25, die aufgrund ihrer unterschiedlichen Erfassungsmethodik eine unterschiedliche Qualität und Detaillierung besitzen. Dies hat Auswirkungen auf die spätere Verknüpfung mit den beschriebenen GPS-Daten. Beide GPS-Studien haben neben der reinen Aufzeichnung der GPS-Koordinaten auch soziodemographische und weitere Variablen der Probanden erfasst. Die Schweizer und die Deutsche GPS-Stichprobe stellen zwei der größten Mobilitätsstudien in Europa dar.

Mit diesen beschriebenen Grundlagen der Methodik und den eingesetzten Daten ist man für die kommenden anwendungsorientieren Kapitel vorbereitet.

# KAPITEL 4

## 4. CHARAKTERISTIKEN, DATENAUFBEREITUNG UND VALIDIERUNG VON GPS-ERHEBUNGEN

"*Es ist nötig, alles zu messen, was messbar ist, und zu versuchen, messbar zu machen, was noch nicht messbar ist*".

Galileo Galilei (1564-1642)

Grundgedanke der neuen Datenerfassung in der Außenwerbung ist Messen statt Befragen. Wurde in der Vergangenheit noch auf die Erinnerung von Probanden gesetzt, ermöglicht die Erfassung mittels GPS-Geräten eine objektive Beobachtung des Mobilitätsverhaltens über einen längeren Zeitraum als einen einzelnen per Telefon befragten Tag. Unschärfen der subjektiven Erinnerung können dadurch ausgeschlossen werden. Erinnerungslücken, wie z. B. die Suche nach einem Parkplatz im Wohngebiet oder der kurze Weg zum Bäcker, werden aufgezeichnet. Die Erfassung der Mobilität durch GPS-Messung ist somit ein Verfahren auf einem hohen Genauigkeitsniveau. Allerdings sind derzeit noch einige Herausforderungen zu bewältigen.

In diesem Kapitel werden die Herausforderungen beschrieben, die mit der Datenaufbereitung und Validierung der GPS-Daten einher gehen. Zunächst werden in Abschnitt 4.1 grundlegende Charakteristiken bei GPS-Studien beschrieben, zu denen mehrere Arten der Unvollständigkeit des Datensatzes gehören. Der Abschnitt 4.2 befasst sich dann mit der Aufbereitung von GPS-Daten und der Verknüpfung von GPS-Rohpunkten mit dem Straßennetz. Dies ist notwendig, da die Rohdaten z.T. extreme Ausreißer und Oszillationen aufweisen. Abschnitt 4.3 stellt eine wichtige Grundlage im Anwendungskontext her, die automatisierte Erstellung von Sichtbarkeitsräumen und der Passagenberechnung für die Außenwerbung. Hier wird auch der Frage nachgegangen, welche Auswirkungen unterschiedliche Sichtdistanzen bei der Erstellung des Sichtbarkeitsraumes auf die Menge von Probandenpassagen haben. Im nachfolgenden Abschnitt 4.4 wird eine Methodik vorgestellt, die auf Basis von KDD-Verfahren Stichprobenverzerrungen in der GPS-Erfassung erkennt. Hierzu wird eine Terminologie eingeführt, die die Ursache von fehlenden Werten in Datensätzen klassifiziert. Der Abschnitt 4.5 fasst die Ergebnisse des Kapitels zusammen.

### 4.1 Messlücken bei GPS-Studien

In Mobilitätsstudien mit GPS gibt es eine Reihe von individuellen oder technischen Gründen, die zum Fehlen von Daten führen können. Es kann zwischen drei Varianten des Fehlens bei Mobilitätsdaten unterschieden werden:

1. Fehlen einer Teilstrecke eines Weges
2. Vollständiges Fehlen eines Weges
3. Fehlen eines Tages

Die Unvollständigkeit bei Teilstrecken bezieht sich auf kurze Erfassungslücken innerhalb einer Sequenz von GPS-Signalen. Mehrere Gründe können zur Entstehung solcher Datenlücken

führen, z.B. Störungen der GPS-Erfassungsqualität durch Gebäude, Bäume, Relief oder der Signalverlust innerhalb eines Tunnels oder bei der Startphase eines GPS-Gerätes. Das vollständige Fehlen eines Weges bezieht sich auf das vollständige Fehlen einer oder mehrerer Fahrten an einem Tag. Das einmalige Fehlen eines Weges ergibt sich z.B., wenn ein Proband vergisst, das GPS-Gerät für einen kurzen Einkaufsweg mitzunehmen. Das Fehlen eines Tages verweist auf das Fehlen vollständiger weiterer Messtage. Diese können z.B. durch leere Gerätebatterien oder technische Mängel des Gerätes entstehen. Fehlende Tage sind im Anwendungskontext ein schwerwiegendes Problem. Wie in Abbildung 4.1 treten fehlende Tage sehr häufig auf, wie hier dargestellt für die Anzahl der deutschen Probanden. Denn die Erfassung lief über einen Zeitraum von insgesamt 7 Tagen. Die optimistische Erwartung zu Beginn der Erhebung war, dass alle Probanden 7 valide Messtage aufweisen. Wie zu erkennen, trifft dies aber nur auf 2915 Personen zu, was einen Anteil von ~33% ausmacht.

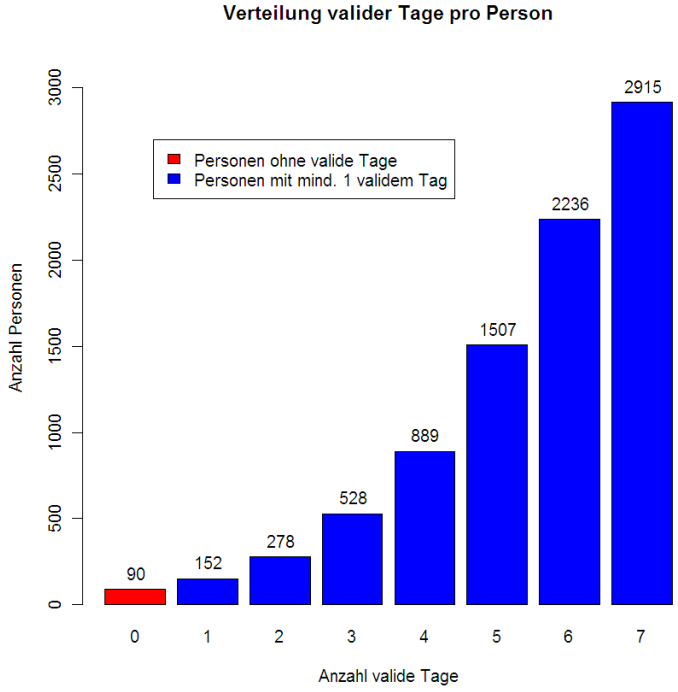



Abbildung 4.1: Fehlende Tage in Messdaten (Deutschland)

Je nach Art der Unvollständigkeit können unterschiedliche Aufbereitungsalternativen ergriffen werden. Kurze Messlücken können durch Routingverfahren geschlossen werden (Abschnitt 4.2). Die übrigen Arten von Messlücken können nur schwer identifiziert werden und können ohne weitere Informationen nicht durch ein Routing geschlossen werden. Das Fehlen einzelner Wege ist nur schwer zu identifizieren, höchstens wenn einzelne Wege nicht wieder zurück zum Startpunkt führen. Da in diesem Falle nur eine zeitliche Lücke aufgezeichnet wird, diese aber nicht bedeutet, dass auch eine räumliche Lücke in der Datenaufzeichnung entstanden ist, ist unklar, ob eine Person evtl. sogar am gleichen Ort geblieben ist oder unterwegs war. Für Personen mit fehlenden Tagen hingegen muss eine geeignete Methode für den Umgang mit Datenlücken gefunden werden, da dieses Problem, wie oben dargestellt, zu massiv auftritt. Diese folgt in Kapitel 5.

Ignoriert werden können die Messlücken nicht, da sonst im Anwendungskontext die Leistungswerte für die Außenwerbung unterschätzt würden. Fehlende mobile Tage bedeuten einen Verlust von potenziellen Plakatkontakten. Das Entfernen aller Personen mit weniger als 7 Tagen stellt auch keine Option dar, da sonst mehr als 66% der Studie entfernt würden.

## 4.2 Datenaufbereitung – von Rohdaten zu Trajektorien

In diesem Abschnitt wird die erste Variante des Fehlens bei Mobilitätsdaten angegangen (vgl. Abschnitt 4.1 - 1. Fehlen einer Teilstrecke eine Weges) sowie die Datenaufbereitung von GPS-Trajektorien vorgestellt. Datenaufbereitung bedeutet in diesem Zusammenhang das Eliminieren von GPS-Fehlverortungen, das Schließen von Teilstrecken eines Weges sowie die Zusammenführung von GPS-Rohdaten und einem digitalen Straßennetz. Die Aufbereitung wird notwendig, da unterschiedliche Fehler bei der GPS-Erfassung auftreten können. Erstens liegt die Messgenauigkeit von GPS bei rund 15 Metern und damit wird das GPS-Signal selten exakt auf dem befahrenen Straßensegment positioniert. Zweitens entstehen bei längerem Stillstand eines Probanden, z.B. an einer Ampel, sogenannte Oszillationen. Die dadurch entstehenden Punktwolken können im Extremfall sogenannte Pseudotrajektorien erzeugen, die in der Realität nicht gefahren wurden. Drittens kann der Empfang der Satellitensignale auf verschiedene Art gestört werden. Dies führt zu einem Verlust der Position und zu extremen Ausreißern von GPS-Signalen, die Kilometer vom eigentlichen Standort entfernt liegen können. Der Verlust der Position führt zu „Lücken" innerhalb einer Trajektorie. In der Praxis entstehen Lücken häufig dann, wenn ein Proband z.B. durch einen Tunnel oder eine innerstädtische Häuserschlucht fährt. Und viertens kommt es zum Fehlen von Mobilitätsaufzeichnungen infolge der Anlaufphase eines GPS-Gerätes. Mindestens 3 Satelliten muss das GPS-Gerät empfangen, damit eine Positionierung stattfinden kann, und dies kann bis zu 2-3 Minuten dauern. In Abbildung 4.2 sind die GPS-Aufzeichnungsfehler 1-3 nochmals graphisch dargestellt. Probleme für die Berechnung der Leistungswerte in der Außenwerbung können insbesondere dann auftreten, wenn eine Passage an einem Plakat real stattgefunden hat, jedoch durch Messungenauigkeiten verloren gehen.


Abbildung 4.2: Mögliche Fehler bei der GPS-Erfassung (1) Versatz der GPS-Signale (2) Oszillationen bei Stillstand (3) Lücken in der Aufzeichnung (topographische Hintergrundinformation OSM 2012)

Die geschilderten Probleme 1-4 in der GPS-Erhebung werden im Folgenden über ein Map Matching Verfahren angegangen. Hierzu werden die GPS-Koordinaten mit einem digitalen Straßennetz verknüpft und evtl. auftretende Lücken werden über ein Routing geschlossen. Zu Beginn wird in Abschnitt 4.2.1 das Map Matching eingeführt. Der Abschnitt 4.2.2 stellt anhand eines Beispiels die Durchführung eines Map Matchings vor. Abschnitt 4.2.3 fasst die Ergebnisse zusammen.

### 4.2.1 Einführung in das Map Matching

Der Erfolg einer Zusammenführung von GPS-Koordinaten und einem Straßennetz hängt im Wesentlichen von der Identifizierung homologer Objekte in beiden Datensätzen ab. Dies ist aufgrund der unterschiedlich erfassten und z.T. heterogenen Geodaten (GPS, digitales Straßennetz) eine Herausforderung. Ein wesentliches Problem ist, dass erfasste Objekte bei einer Zusammenführung unterschiedliche Lagebeziehungen, Repräsentationen oder

Klassifikationen aufweisen. Es ist möglich, dass einzelne Objekte gänzlich fehlen, oder sich die Objekte in Geometrie oder Topologie grundlegend voneinander unterscheiden, so dass eine Zusammenführung unter Umständen nicht möglich ist. Beim Map Matching wird nach (Bernstein und Kornhauser 1998) zwischen zwei Kategorien unterschieden:

- Geometrisches Matching nutzt den direkten räumlichen Bezug von Geodaten aus und stellt so eine Verknüpfung her. Die Grundlage für das geometrische Matching bildet die Berechnung der räumlichen Distanz. So wird bei diesem Verfahren die GPS-Position dem nächstliegenden Knoten oder der nächstliegenden Kante zugewiesen. Dies kann z.B. über die Euklidische Distanz zwischen zwei Koordinaten (vgl. Abschnitt 3.1.3) durchgeführt werden. In der Berechnung werden in einem Punkt-zu-Punkt Matching alle Zuordnungspaare sequentiell abgearbeitet.

- Topologisches Matching kombiniert geometrisches Matching und topologische Informationen aus zusätzlichem Wissen über den Straßengraphen und seine Verknüpfungen. So wird bei diesem Verfahren berücksichtigt, wie die Straßen miteinander verbunden sind, um zu verhindern, dass es zu Sprüngen beim Map Matching kommt. Es wird zusätzlich untersucht, auf welchen Straßen ein Fahrzeug unterwegs war und ob das nächste zugeordnete Straßensegment in einer plausiblen Art und Weise erreicht werden kann. Mithilfe von topologischen Beziehungen des Straßennetzes können Heuristiken gebildet werden, die den Matchingprozess bei der Zuordnung unterstützen.

Im Abschnitt 4.2 wird ein topologisches Matching vorgestellt, da beim geometrischen Matching insbesondere bei einem engen Straßennetz und ungenauen Koordinaten eine fehlerhafte Zuordnung durchgeführt würde. In Abbildung 4.3 ist ein Fall dargestellt, der beim rein geometrischen Matching falsch zugeordnet würde. Die GPS-Koordinaten weisen einen Versatz von bis zu 15 Metern auf und springen auf eine benachbarte parallele Straßen. Dies führt dazu, dass bei diesen zwei Straßensegmenten, die parallel zueinander laufen, die GPS-Punkte eines Autofahrers mal dem einen und mal dem anderen Segment zugeordnet werden.

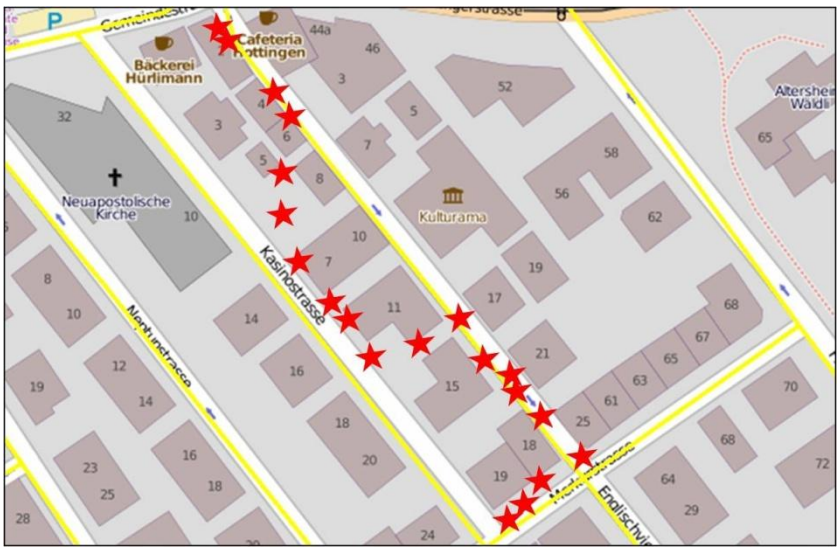



Abbildung 4.3: Geometrisches Matching bei engem Straßennetz (topographische Hintergrundinformation OSM 2012)

Aus dem Verlauf der GPS-Koordinaten kann jedoch durch ein topologisches Matching und unter Annahme gewisser Heuristiken eine exaktere und eindeutige Zuordnung vorgenommen werden (Curtin 2007). Ein rein auf geometrische Zuordnung basierendes Verfahren ordnet den Probanden evtl. mehrere Straßensegmente des Wohngebietes zu, obwohl zu vermuten ist, dass der Proband *„nur"* von seinem Startort das Wohnviertel verlassen hat. Das geometrische Matching würde zu viele besuchte Straßensegmente zuordnen. Die Passagen an Plakatstellen würden somit überschätzt. Aus diesem Grund wird zur Zusammenführung von GPS-Daten und einem digitalen Straßennetz das topologische Matching gewählt. Für ein topologisches Matching sollten folgende Voraussetzungen erfüllt sein:

1. Ein vollständiges und möglichst aktuelles Straßennetz
2. Die Positionen der Probanden, bzw. beim Routing die erste und die letzte Koordinate des Probanden
3. Ein einheitliches geographisches Bezugssystem

Im folgenden Abschnitt wird das Vorgehen eines speziell entwickelten topologischen Map Matching erläutert. Dabei handelt es sich um eine Modifikation und Erweiterung des vorgestellten Algorithmus von (Lou 2009 und Marchal et al. 2006)).

### 4.2.2 TOPOLOGISCHES MAP MATCHING

Ziel des Map Matchings ist es, die realen Trajektorien der Probanden möglichst gut zu rekonstruieren, um zu berechnen, wie und wann sie an einem bestimmten Straßensegment vorbeigekommen sind. Hierzu wurde jedem der Probanden ein GPS-Gerät ausgehändigt und in einem Log deren alltägliche Bewegungen festgehalten. Der GPS-Log ist eine Sammlung von GPS-Punkten $L = \{g_1, g_2, \dots g_n\}$. Jeder GPS-Punkt $g_i \in L$ enthält den Breitengrad $g_i\, lat$, den Längengrad $g_i\, long$ und einen Zeitstempel $p_i\, t$. Noch mal zur Erinnerung: Eine Trajektorie $T$ ist eine Sequenz von GPS-Punkten in einem bestimmten Zeitintervall (vgl. Abschnitt 3.2). Die eingesetzten GPS-Geräte in Deutschland und der Schweiz speichern in der Regel sekündlich die Ortsinformationen ab.

Das Map Matching Verfahren besteht aus vier Phasen: 1. Datenvorverarbeitung 2. Erzeugung von Optionsgraphen, 3. Einfügen von Kanten in den Optionsgraphen 4. das eigentliche Map Matching. Im nachfolgendem werden die einzelnen Phasen vorgestellt.

**Datenvorverarbeitung**

Das Map Matching beginnt mit einer Vorverarbeitung der GPS-Rohpunkte. Jeder einzelne GPS-Punkt wird attributiert, mit dem Ziel, verschiedene Filterungen und Klassifikationen durchzuführen. Dabei sollen z.B. einerseits bereits frühzeitig die gröbsten offensichtlich fehlerhaften GPS-Punkte im Vorfeld des eigentlichen Map Matchings herausgefiltert werden und andererseits die wahrscheinlichste Fortbewegungsart (PKW, Fußgänger) klassifiziert werden. Das bedeutet also, bei der Filterung werden einzelne GPS-Punkte eliminiert, bei der Klassifikation wird die Menge an GPS-Punkten mit einem zusätzlichen Attribut ausgestattet. Zu den einzelnen Schritten gehören:

**Filterung**

1. Filterung Ausreißererkennung: Analysiert die eingelesen Punkte und bestimmt die Entfernung und die Beschleunigung des vorangegangenen und des nächstliegenden GPS-Punktes. Ist die Entfernung > 15 Meter und die Beschleunigung > 3 Meter pro Sekunde, wird dieser Punkt eliminiert.
2. Filterung Stillstandserkennung: Dabei handelt es sich um die bereits erwähnten Oszillationen. Diese entstehen typischerweise an Kreuzungen und führen z.T. zu sehr großen Punktwolken. Diese Punktwolken sind häufig innerhalb eines 15 Meter

Toleranzpuffers und zeichnen sich über die Dauer des Stillstandes durch eine geringe Durchschnittsgeschwindigkeit aus. Weiterhin werden zur Erkennung der Oszillation die Abstände eines Punktes zu den 5 folgenden GPS-Punkten einbezogen. Oszilliert der Punktabstand an dieser Stelle, dann wird ein Stillstand identifiziert. Diese Punkte werden entfernt.

**Klassifikation**

1. Klassifikation Wegeanfang und –ende: Wenn der zeitliche Abstand zwischen zwei aufeinander folgenden Punkten 30 Minuten überschreitet, werden sie als Routinggrenze klassifiziert. An dieser Stelle erfolgt keine Lückenschließung mehr.
2. Klassifikation Verkehrsmittelzuordnung (PKW, Fußgänger): Hier werden Punkte unter Berücksichtigung der Geschwindigkeit klassifiziert (>10 k/mh = PKW, < 10 k/mh Fußgänger).

**Map Matching**

Das Ziel des Map Matching ist es, jedem GPS-Punkt genau das Straßensegment zuzuordnen, auf dem sich der Proband höchstwahrscheinlich aufgehalten hat. Zusätzlich sollen durch das Map Matching die ggf. auftretenden Lücken zwischen den Segmenten von aufeinander folgenden GPS-Punkten geschlossen werden.

Hätte man an dieser Stelle bereits für jeden GPS-Punkt das zugehörige Straßensegment, so könnte man die vorhandenen Lücken zwischen den Segmenten plausibel durch kürzeste Wege im Straßennetz schießen. Leider lässt sich das passende Straßensegment für einen einzelnen GPS-Punkt i.d.R. nicht eindeutig bestimmen.

Die Idee für das hier verwendete Map Matching ist es, für jeden GPS-Punkt eine Menge an möglichen Straßensegmenten zu bestimmen, die als Kandidaten bezeichnet werden. Das Ziel des Verfahrens ist es, einen kürzesten Weg durch das Straßennetz zu bestimmen, der einen Kandidaten von jedem GPS-Punkt besucht.

Dazu wird aus den gegebenen GPS-Punkten und dem Straßennetz ein weiterer Graph gebildet, der hier als Optionsgraph bezeichnet wird. Die Knoten des Optionsgraphen sind in Schichten geordnet, wobei jede Schicht zu einem GPS-Punkt mit seinen Kandidaten korrespondiert. In jeder Schicht gibt es mehrere Kandidaten/Optionen für einen GPS-Punkt *g*, die angeben, auf welchem Straßensegment *g* Mobilität aufgezeichnet worden sein könnte. Die Kanten des Optionsgraphen sind gerichtet und so angeordnet, dass die Schichten des Graphen nur in der durch die GPS-Punkte induzierten Reihenfolge durchlaufen werden können. Das letztendliche Map Matching wird sich aus einem kürzesten Weg im Optionsgraphen ergeben. Dies kann man als Anwendung von *Ockhams Razor*[5] betrachten, bei der vorausgesetzt wird, dass ein Kandidat von jedem GPS-Punkt besucht wird, und der kürzeste Weg durch diese Kandidaten als wahrscheinlichste Lösung betrachtet wird.

Der Map Matching Algorithmus verläuft in den drei folgenden Phasen:

[5] Ockhams Razor ist ein heuristisches Forschungsprinzip aus der Scholastik, das bei der Bildung von erklärenden Hypothesen und Theorien Sparsamkeit gebietet. Steht man vor der Wahl mehrerer möglicher Erklärungen für ein und dasselbe Phänomen, soll man diejenige bevorzugen, die mit der geringsten Anzahl an Hypothesen auskommt und somit die „einfachste" Theorie darstellt. Es enthält ebenso die Forderung, für jeden Untersuchungsgegenstand nur eine einzige hinreichende Erklärung anzuerkennen (Wikipedia 2012).

1. **Erzeugung der Knoten im Optionsgraphen:** Dabei wird für jeden GPS-Punkt *g* eine Schicht von Knoten eingefügt, in der für jeden Kandidaten von *g* ein Knoten im Optionsgraphen entsteht. Diese Kandidaten sind Kanten im Straßennetz, welches selber auch ein Graph ist. Formal ist ein Knoten im Optionsgraphen ein Paar (*g*,*s*) aus einem GPS-Punkt *g* und einem Straßensegment *s*.

2. **Erzeugen der Kanten im Optionsgraphen:** Es seien *g* und *g'* zwei aufeinander folgende GPS-Punkte. Nun werden alle Paare (*s*,*s'*) der Kandidaten *s* von *g* und Kandidaten *s'* von *g'* durchlaufen, um (gerichtete) Kanten im Optionsgraphen zu erzeugen. Dabei wird jeweils der kürzeste Weg *w* im Straßennetz von *s* zu *s'* gesucht. Für jeden solchen Weg *w* von *s* zu *s'* im Straßennetz wird eine Kante ((*g*,*s*),(*g'*,*s'*)) von (g,s) zu (g',s') im Optionsgraphen eingefügt und mit den Kosten *w* annotiert.

3. **Map Matching als kürzester Weg im Optionsgraphen:** Das Map Matching für die gegebenen GPS-Punkte ergibt sich als ein kürzester Weg im Optionsgraphen. Dazu werden noch zusätzlich ein Start- und ein Endknoten im Optionsgraphen eingefügt. Der Startknoten kommt vor die Schicht des ersten GPS-Punktes und erhält eine ausgehende Kante zu allen enthaltenen Knoten. Analog kommt der Endknoten hinter die Schicht des letzten GPS-Punktes und enthält eine eingehende Kante von allen enthaltenen Knoten. Alle Kanten, die dabei noch eingefügt werden, erhalten Kosten 0 im Optionsgraphen. Anschließend wird der kürzeste Weg vom Start- zum Endknoten gesucht. Dieser Weg im Optionsgraphen entspricht einer Folge von Wegen im Straßengraphen, deren Konkatenation das Map Matching ergibt.

Soweit die wesentlichen Eckpunkte der Verfahrens. Diese und noch einige weitere Details werden anhand eines Beispiels erläutert. Dabei wird ein kleines Straßennetz mit 12 Segmenten ($s_1, \dots, s_{12}$) und 4 GPS-Punkten ($g_1, \dots, g_4$) verwendet.

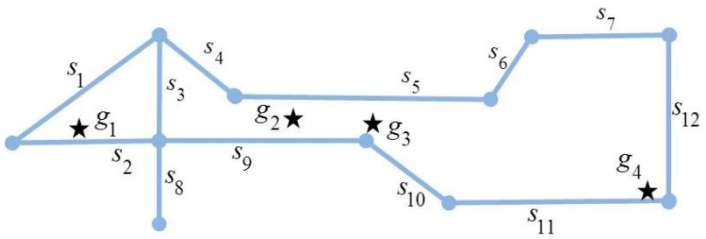



Abbildung 4.4: Beispiel Map Matching

Zunächst werden die Knoten des Optionsgraphen gebildet, wobei die Kandidaten für einen GPS-Punkt benötigt werden. Dabei wird ein Puffer um die GPS-Punkte erstellt, und alle Straßensegmente, die diesen Puffer schneiden, sind Kandidaten für einen GPS-Punkt. Diese Puffer sind hier durch blaue Kreise dargestellt.

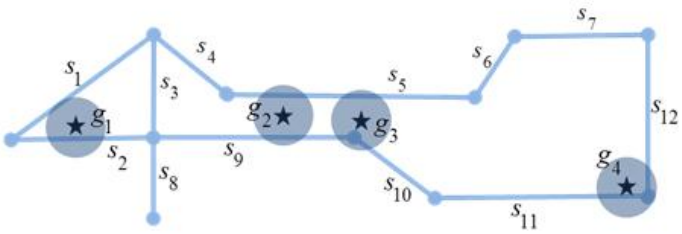



Abbildung 4.5: Kandidatenbildung Optionsgraph

Damit erhält man zunächst 4 Schichten mit 9 Knoten im Optionsgraphen.

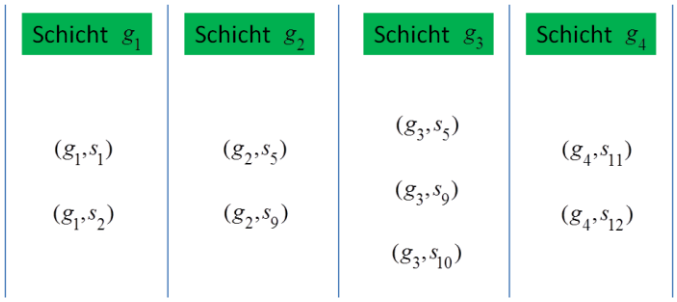



Abbildung 4.6: Optionsgraphen

Nun folgend die Kanten des Optionsgraphen. Dabei wird von jedem Knoten *x* in einer Schicht *i* zu allen Knoten *y* in Schicht *i+1* eine Kante eingefügt. Seien dabei $s^x$ und $s^y$ Straßensegmente von *x* und *y*. Gesucht wird nun ein kürzester Weg *w* im Straßengraphen von $s^x$ zu $s^y$. Dabei ist die Interpretation so, dass man sich mit *x* bereits auf dem Segment $s^x$ befindet und zum Segment $s^y$ gelangen muss. Daher endet der Weg mit $s^y$, aber startet nicht mit $s^x$, weil es im vorhergehenden Weg bereits enthalten ist. Eine Ausnahme ist der Fall, dass $s^y = s^x$ ist. In diesem Fall ist *w* leer und enthält kein Segment. Eine weitere Ausnahme gibt es in dem Fall, dass $s^y$ im Straßennetz nicht von $s^x$ aus erreichbar ist. In diesem Fall wird keine Kante von x zu y in den Optionsgraphen aufgenommen. Ansonsten sind die Kosten der Kante (*x*,*y*) im Optionsgraphen gleich den Kosten von *w* im Straßennetz. Dabei sind verschiedene Kostenmodelle im Straßennetz möglich. In der einfachsten Variante sind die Kosten einer Kante gleich der Länge in Metern. Eine andere Möglichkeit ist eine vom Verkehrsmittel und Straßenkategorie abhängige geschätzte Zeit für die Passage eines Straßensegments.

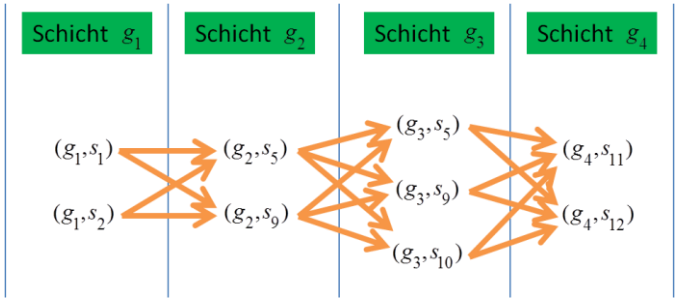



Abbildung 4.7: Optionskanten

Als nächstes werden noch ein Start- und ein Endknoten in den Optionsgraphen eingefügt. Dabei kommt der Startknoten vor die Schicht des ersten GPS-Punktes und erhält eine ausgehende Kante zu allen enthaltenen Knoten. Da die Segmente der GPS-Punkte in der ersten Schicht noch nicht besucht werden, erhalten diese Kante einen Pseudo-Weg, der aus dem jeweiligen Straßensegment-Kandidaten besteht. Die Kosten der Kanten vom Startknoten sind 0. Der Endknoten kommt hinter die Schicht des letzten GPS-Punktes und erhält eine eingehende Kante von allen enthaltenen Knoten. Der jeweils assoziierte Weg im Straßennetz ist leer und die Kosten der Kanten sind 0.

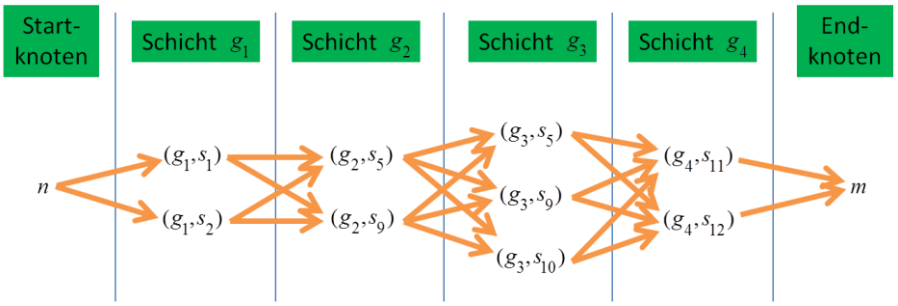



Abbildung 4.8: End- und Startknoten im Optionsgraphen

Zuletzt wird im Optionsgraphen ein kürzester Weg vom Startknoten *n* zum Endknoten *m* gesucht. Die Wege im Straßennetz, die mit den dabei besuchten Kanten im Optionsgraphen assoziiert sind, ergeben zusammen das Map Matching. In der folgenden Abbildung 4.9 sind die Kanten im Optionsgraph auf dem kürzesten Weg zwischen n und m blau gefärbt, und unter den Trennlinien der Schichten stehen die jeweils zugehörigen Wege im Straßennetz. Dabei steht () für einen leeren Weg aus 0 Straßensegmenten. Insgesamt ergibt sich der Weg $(s_2, s_9, s_{10}, s_{11})$.

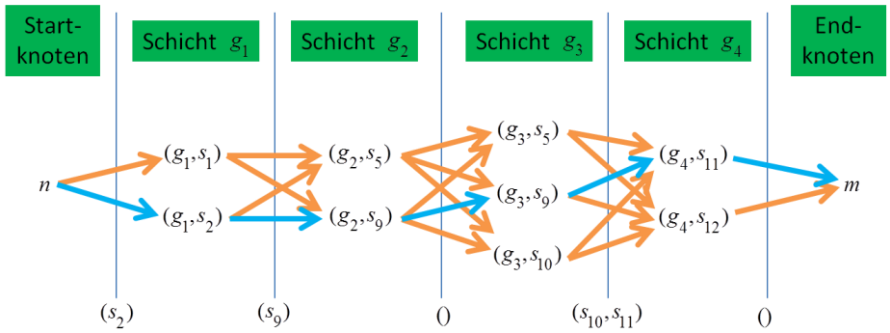



Abbildung 4.9: Map Matching

Eine Schicht steht für die „möglichen Welten" für einen einzelnen GPS-Punkt, d.h. die verschiedenen Segmente, die während der Aufzeichnung durchschritten worden sein könnten. Ein Pfad durch den Optionsgraphen, der einen Knoten aus jeder Schicht besucht, steht für eine mögliche „Gesamtwelt", in der für jeden GPS-Punkt ein Straßensegment ausgewählt wurde. Das Kriterium dafür, welche Gesamtwelt gewählt wird, ist der kürzeste Weg durch den Optionsgraphen.

**Routing Anlaufphase**

Bei den ersten in der Schweiz durchgeführten GPS-Aufzeichnungen ist schnell aufgefallen, dass z.T. der erste Streckenabschnitt der Probanden fehlte. Dies ist dadurch zu begründen, dass die GPS-Geräte eine gewisse Anlaufphase brauchten, um mindestens 3 Satelliten zu empfangen. Diese Anlaufphase kann bis zu 2-3 Minuten dauern. Um diese Lücke zu schließen, hat man die Möglichkeit, ein Routing durchzuführen. Das Routing verbindet die Wohnadresse der Probanden mit dem ersten gemessenen Punkt der GPS-Aufzeichnung. In Abbildung 4.10 ist das Vorgehen dargestellt. Bei G* liegt die Wohnadresse des Probanden. Über die Segmente $(s_{13}, s_{14})$ wird die Lücke bis zum ersten aufgezeichneten GPS-Punkt $(g_1)$ geschlossen.

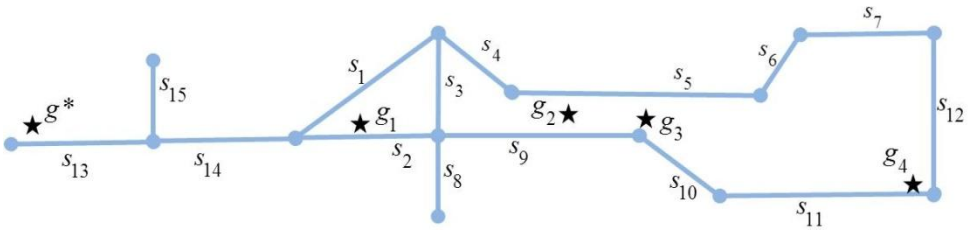



Abbildung 4.10: Routing Anlaufphase, einfügen eines künstlichen GPS-Punktes

### 4.2.3 Ergebnis Map Matching und Routing

Nach Durchführung des Map Matching Verfahrens liegen nun alle aufgezeichneten GPS-Informationen auf Straßennetzebene vor. In der Schweiz ist dies das Vector25 Straßennetz, in Deutschland das NavTeq Straßennetz. Abbildung 4.11 stellt dies nochmals graphisch dar. Auf der linken Seite der Abbildung sind für einen Tag die roh GPS-Koordinaten von 10 selektierten Probanden zu erkennen. Nach dem Map Matching liegen alle Informationen auf Straßenabschnittniveau vor (rechte Seite der Abbildung). Durch Selektion einzelner Segmente können über Abfragen für einzelne Tage die Summe der Passagen, die Anzahl der Probanden, die Probanden IDs sowie weitere Informationen abgerufen werden.

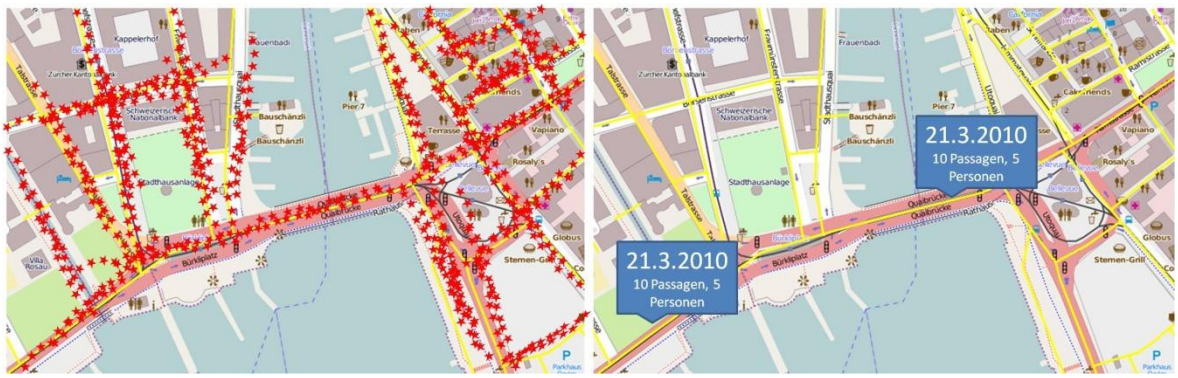



Abbildung 4.11: GPS-Daten nach dem Map Matching (topographische Hintergrundinformation OSM 2012)

Im Ergebnis liegen für Deutschland insgesamt 1.352.472 km an zurückgelegter Strecke über alle Probanden vor. Die männlichen Probanden legen dabei insgesamt 72.336 Kilometer mehr zurück als die weiblichen Probanden.

| Deutschland | Gesamt | Männlich | Weiblich |
|---|---|---|---|
| **Kilometerleistung** | 1.352.472 | 712.404 | 640.068 |
| **Probanden** | 11.458 | 5.388 | 6.070 |
| **Anzahl Valider Tage** | 56.959 | 26714 | 30.245 |
| **Kilometer pro Tag** | 23,7 | 26,7 | 21,2 |

Tabelle 4.1: Gesamtkilometerleistung in Deutschland

In der Schweiz liegt die Kilometerleistung insgesamt etwas höher. Allerdings hatten die Schweizer Probanden das GPS-Gerät auch bis zu 10 Tage im Einsatz. Wie in Deutschland liegt

auch in der Schweiz die Gesamtkilometerleistung bei den männlichen Probanden höher als bei den weiblichen. Dieses Ergebnis war zu erwarten, da auch in der MID (Mobilität in Deutschland 2008a) und dem Microzensus Schweiz (2005) eine unterschiedliche Mobilität bei den Geschlechtern festgestellt worden ist.

| Schweiz | Gesamt | Männlich | Weiblich |
|---|---|---|---|
| **Kilometerleistung** | 1.886.613 | 1.032.791 | 8.53.822 |
| **Probanden** | 11.858 | 5.706 | 6.152 |
| **Anzahl Valider Tage** | 67.085 | 32.261 | 34.824 |
| **Kilometer pro Tag** | 28,1 | 32,0 | 24,5 |

Tabelle 4.2: Gesamtkilometerleistung in Deutschland

## 4.3 Erstellung von Plakatpassagen und Plakatsichtbarkeitsräumen zur Leistungswertberechnung

Nachdem im vorherigen Abschnitt die GPS-Daten mit dem Straßennetz verknüpft worden sind, wird in diesem Abschnitt der Frage nachgegangen, wie und wann ein Proband potenziell die Möglichkeit hat, ein Plakat zu sehen. Hierzu müssen die Informationen auf der Straßensegmentebene mit Plakatinformationen in Beziehung gesetzt werden. Hierzu wird an dieser Stelle das Wort *„Plakatpassage"* eingeführt. Eine Plakatpassage bedeutet, dass ein Proband durch den Sichtbarkeitsraum eines Plakates passiert. Dabei stellt die Erfassung der Passage an der Plakatstelle und auch die Erstellung von Sichtbarkeitsräumen eine wichtige Komponente in der Leistungswertbestimmung dar. Eine exakte Verortung der Plakatstellen und der daraus abgeleitete Sichtbarkeitsraum sind ein entscheidendes und wichtiges Kriterium für eine spätere Bewertung von Plakatstellen. In diesem Abschnitt wird vorgestellt, mit welchen Regeln eine Passage eines Probanden an einer Plakatstelle gewertet wird (Abschnitt 4.3.1) und wie ein Sichtbarkeitsraum eines Plakates definiert ist (Abschnitt 4.3.2). Zusätzlich wird untersucht welchen Einfluss die maximale Sichtdistanz eines Plakates und eine evtl. Individualisierung der Sichtbarkeitsräume durch Gebäudedaten hat[6].

### 4.3.1 Passagenberechnung an Plakatstellen

Eine Passage ist eine Trajektorie, die den Sichtbarkeitsraum einer Plakatstelle schneidet und theoretisch zu einem Werbekontakt führen kann. Dabei ist ein Sichtbarkeitsraum ein geometrischer Bereich aus, dem Passanten das Plakat theoretisch sehen können. Nicht jede Passage führt automatisch zu einem Werbekontakt mit der Plakatstelle. Die definierten Sichtbarkeitsräume haben eine wichtige Rolle in der Leistungswertbestimmung, denn sie werden nach ihrer Konstruktion dazu verwendet, sie mit den GPS-Trajektorien zu verschneiden. Das bedeutet in der Regel, wenn eine GPS-Person den Sichtbarkeitsraum durchschritten hat, dann hatte sie auch die Möglichkeit eines Werbekontaktes. Diese Möglichkeit des Kontaktes ist eine der Basisgrößen in der Leistungswertbestimmung in der Außenwerbung, es ist der sogenannte opportunity of contact (ooc) (Esomar 2009). In der Schweiz tritt z.B. ein ooc auf, wenn eine Person mit einem Durchgangswinkel von kleiner oder gleich 110° den Sichtbarkeitsraum passiert. Der Durchgangswinkel ist hierbei der Winkel zwischen der Bewegungsrichtung des Passanten und der Orientierung des Plakates. Diese Einschränkung verhindert, dass eine Person, die mit dem Rücken zum Plakat den Sichtbarkeitsraum passiert, keinen Kontakt generiert. In der Schweiz gibt es drei unterschiedliche Möglichkeiten des ooc. Es wird unterschieden zwischen einer frontalen Passage <45° (a), einer parallelen Plakat Passage 45°-110° (b) und keiner Passage >110° (c) (vgl. Abbildung 4.12).

[6] Die Analyse mit Gebäudedaten wird nur in der Schweiz durchgeführt, da kein Gebäudedatensatz in Deutschland vorlag.

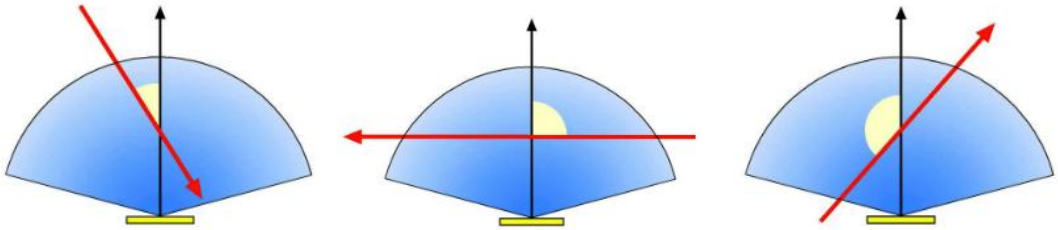


Abbildung 4.12: ooc für unterschiedliche Passagedurchgangswinkel (1) frontale Plakat Passage, (2) parallele Plakat Passage, (3) rückwärtige Plakat Passage – kein potenzieller Kontakt (SPR+ 2011)

Die Unterscheidung zwischen einer frontalen und einer parallelen Passage ist wichtig, da eine frontale Passage in der Schweiz und in Deutschland eine höhere Kontaktchancenmöglichkeit besitzt. Dies wird aufgrund der direkten Sichtlinie zum Plakat begründet, während bei einer parallelen Passage der Passant die Blickrichtung zur Seite ändern muss (vgl. Abschnitt 2.1).

Interessant sind neben dem Winkel der Passage auch noch folgende, weitere Informationen, die für eine spätere Kontaktwertberechnung relevant werden:

1. Uhrzeit
2. Geschwindigkeit der Passage

Die Uhrzeit wird benötigt, um tageszeitlich sehr frühe oder sehr späte Passagen bei nicht beleuchteten Plakatflächen herauszufiltern und diese nicht in die Passagenwertberechnung mit einzubeziehen. Die Geschwindigkeit ist ein weiterer wichtiger Faktor in der Berechnung von der Passage zu einem späteren Kontakt. Mit einer geringeren Geschwindigkeit hat man auch eine höhere Chance, die Werbung wahrzunehmen. In der Außenwerbung wird an dieser Stelle nur unterschieden zwischen Fußgängerpassagen und PKW-Passagen. Passagen, die als Fahrrad oder ÖPNV-Nutzer generiert werden, werden in Deutschland und der Schweiz nach Konvention wie eine PKW-Passage gewichtet (vgl. Abschnitt 2.2.2). Für jede einzelne Probandenpassage werden pro Plakatstelle Informationen gespeichert. Die Tabelle 4.3 zeigt dies beispielhaft für eine Plakatstelle.

| Plakat | Proband | Uhrzeit | Datum | Winkel | Geschwindigkeit |
|---|---|---|---|---|---|
| 190342 | 2002 | 12:45 | 1.07.2010 | 48° | 8 |
| 190342 | 2002 | 13:32 | 2.07.2010 | 49° | 5 |
| 190342 | 1231 | 22:30 | 3.07.2010 | 110° | 48 |
| 190342 | 3212 | 17:12 | 4.07.2010 | 100° | 50 |

Tabelle 4.3: Passagentabelle für ausgewählte Plakatstellen

Diese Informationen sind Kriterien, die einen maßgeblichen Einfluss auf die Qualität eines Kontaktes haben und aus den aufgezeichneten GPS-Daten ausgelesen werden können. Wie in den Abschnitten 2.2.1 und 2.2.2 dargestellt gibt es jedoch noch eine Menge von weiteren Gewichtungskriterien in Deutschland und der Schweiz, die für die Entscheidung herangezogen werden, ob aus einem ooc ein Plakatkontakt wird. Diese Faktoren sind im Wesentlichen durch unterschiedliche Plakatformate und Konventionen/Erfahrungen zu begründen. Sie sind nicht aus den aufgezeichneten GPS-Daten ableitbar und gehören zum

qualitativen Bereich der Medienforschung, welcher nicht Teil dieser Arbeit ist (vgl. Abschnitt 2.1).

### 4.3.2 Erstellung von Sichtbarkeitsräumen

Um die Konstruktion von Sichtbarkeitsräumen und ihren Auswirkungen auf die Leistungswerte zu verstehen, stelle man sich den alltäglichen Weg zur Arbeit vor, und versuche sich an die Plakate zu erinnern, die man passiert. Einige Plakate haben eher große Dimensionen und sind beleuchtet, wieder andere Plakate sind eher klein dimensioniert und leicht verdeckt, so dass das Plakat nur aus der Nähe erkannt werden kann. Daher ist auch die Sichtbarkeit jedes einzelnen Plakates einzigartig und hängt von der Größe des Plakates und seines Standortes bzw. Umfeldes ab. Es stellt sich die Frage: Wie kann man einen Sichtbarkeitsraum geographisch definieren und welche Rolle spielt die räumliche Dimensionierung des Sichtbarkeitsraumes (Hecker et al. 2010c)? Früh wird auch zusätzlich die Frage gestellt: Wie können automatisiert Sichtbarkeitsräume erstellt werden? Denn pro Jahr werden in der Schweiz und in Deutschland viele Plakatstellen umgebaut, neu aufgestellt oder abgebaut. Jeden Sichtbarkeitsraum einer Plakatstelle manuell zu erfassen/vermessen wird aus Kostengründen in der Außenwerbung derzeit nicht durchgeführt.

Zwingend notwendig für die Erstellung eines Sichtbarkeitsraumes sind die X/Y Koordinate und die Ausrichtung (Himmelsrichtung) des Plakates. Bei der Erstellung von Sichtbarkeitsräumen und der Zuordnung von GPS-Trajektorien kommen zwei Geodatensätzen eine besondere Bedeutung zu:

1. Straßennetz mit zugeordneten GPS-Trajektorien
2. Gebäudedaten

Das Straßennetz wird dazu verwendet, die vorher per Map Matching zugeordneten GPS-Daten den Plakatstellen zuzuordnen. Die Gebäudedaten spielen bei der Bestimmung von Sichtbarkeitsräumen insofern eine Rolle, da Gebäude und Anbauten die Sicht auf eine Plakatstelle beeinträchtigen können. In der Schweiz werden derzeit die Sichtbarkeitsräume mithilfe von Gebäudedaten verschnitten. Somit können Teile oder sogar komplette Segmente einer Straße, die nicht in einer direkten Sichtlinie des Plakates liegen, entfernt werden. Abbildung 4.13 zeigt ein Beispiel für die Erstellung eines Sichtbarkeitsraumes in der Schweiz. Auf der linken Seite sind Plakatsichtbarkeitsräume mit einer maximalen Sichtbarkeitsentfernung von 40 Metern und einem Öffnungswinkel von 150° aufgespannt. Im Bild auf der rechten Seite werden diese Sichtbarkeitsräume mit dem Gebäudelayer verschnitten und somit um diese Sichthindernisse reduziert.

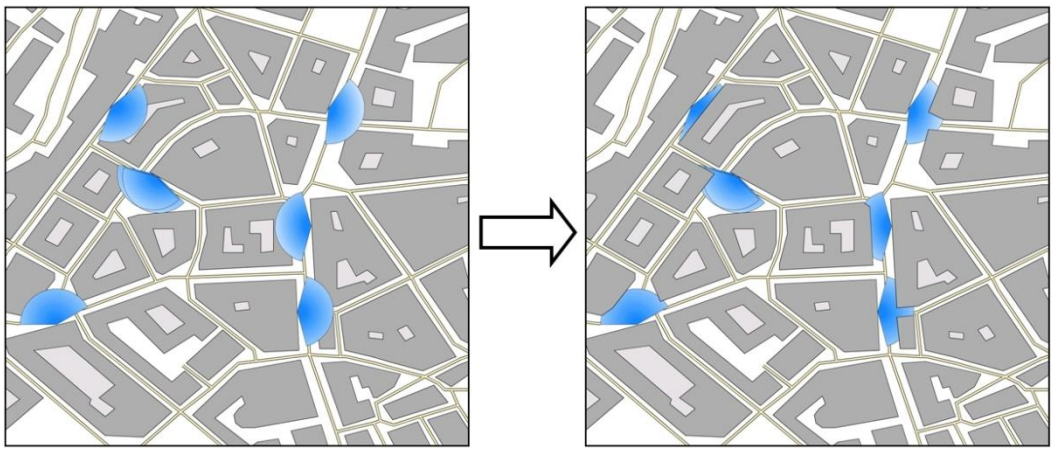


Abbildung 4.13: Sichtbarkeitsräume von Plakatstellen vor (links) und nach der Verschneidung mit Gebäuden (rechts) (Hecker et al. 2010c)

Eine der häufig zu Beginn diskutierten Fragen bei der Erstellung von Sichtbarkeitsräumen ist die Justierung der maximalen Sichtbarkeitsdistanz. Aus welcher Distanz kann man ein bestimmtes Plakatformat erkennen? Im dem oben illustrierten Beispiel wurde eine Sichtbarkeitsdistanz von 40 Metern angenommen, doch gibt es eine Reihe von unterschiedlichen Meinungen, aus welcher Distanz ein Plakat noch zu sehen ist. Dabei spielt auch die konkrete Werbung auf der Plakatstelle eine wichtige Rolle, denn die graphische Aufbereitung und Illustration hat ebenfalls Einfluss auf die Sichtdistanz. Ist viel textueller Anteil in der Werbeansprache, dann minimiert sich auch die Distanz. Derzeit wird der Einfluss des Werbeinhaltes in der Außenwerbung noch nicht berücksichtigt. Lediglich den unterschiedlichen Plakatformaten wird je nach Größe eine maximale Sichtbarkeitsentfernung zugeordnet. In Deutschland schwanken je nach Plakatformat die Sichtdistanzen zwischen 30 und 80 Metern, in der Schweiz zwischen 40 und 60 Metern. Eine der interessanten Fragen an dieser Stelle ist, wie stark die Anzahl der Passagen mit größer werdender Sichtbarkeitsdistanz zunimmt. Ist ein massiver Anstieg des ooc über die Distanz zu verzeichnen, oder ist der Anstieg eher marginal? Wie sensitiv reagiert der Passagenwert auf die Distanz?

Um dies zu testen, wird der ooc für die Sichtdistanzen zwischen 20, 40, 60, 80 und 100 Metern und erst mal ohne Individualisierung durch Gebäudedaten berechnet. Teststädte für die Berechnungen des ooc sind die Agglomerationen Basel, St. Gallen und Genf mit jeweils 2.319, 911 und 3.090 Plakatstellen. Tabelle 4.5 zeigt die Ergebnisse für die unterschiedlichen Distanzen. Die Werte sind indexiert über die geringste Distanz von 20 Metern. Das heißt, ein Wert von 100% bedeutet, dass die Ergebnisse völlig identisch sind, und ein Wert über 100% bedeutet, dass das die Berechnung mit zunehmender Entfernung zu einem Anstieg des ooc führt. Je höher die Differenz zu 100% ist, desto stärker ist der Einfluss der Entfernung.

| | **100 Meter** | **80 Meter** | **60 Meter** | **40 Meter** | **20 Meter** |
|---|---|---|---|---|---|
| **Basel** | 105,6% | 103,6% | 101,9% | 100,7% | 100,0% |
| **St. Gallen** | 109,5% | 105,9% | 102,8% | 100,9% | 100,0% |
| **Genf** | 106,0% | 103,8% | 102,0% | 100,6% | 100,0% |
| Durchschnitt | 106,2% | 104,0% | 102,1% | 100,7% | 100,0% |

Tabelle 4.4: ooc Berechnung für unterschiedliche Distanzen ohne Gebäudelayer (Hecker et al. 2010c)

Es ist zu erkennen, dass die Auswirkungen der Sichtdistanz einen Einfluss auf den ooc bis zu knapp 9,5% bei 100 Metern besitzen. Die Auswirkungen bis zu 60 Metern sind jedoch relativ gering. Dies ist darin begründet, dass der ooc erst dann ansteigt, wenn ein zusätzliches Straßensegment in den Sichtbarkeitsraum fällt. Straßensegmente, die bereits bei 20 Metern im Sichtbarkeitsraum liegen, tragen voll zum ooc bei, auch wenn es sich nur um einen kleinen Anteil des Straßensegmentes handelt.

**Die Notwendigkeit von Gebäudedaten bei der Erstellung von Sichtbarkeitsräumen**

Eine weitere Frage bei der Erstellung von Sichtbarkeitsräumen ist die Frage nach der Notwendigkeit einer Verschneidung mit Gebäudedaten, die den Sichtbarkeitsraum um etwaige Sichthindernisse reduzieren. Wie in vielen anderen Projekten, wird zu Beginn aus Kostengründen häufig die Frage gestellt: „Wie genau müssen die geographischen Daten sein, und welche Daten werden gebraucht?“. Die spontane Antwort ist in der Regel: „So

präzise und so viele Daten wie möglich!". Allerdings sind insbesondere flächendeckende und präzise Geodaten in der Regel sehr teuer. Daher ist es zweckmäßig, den Trade-off zwischen den verschiedenen Geodaten zu vergleichen. Insbesondere bei hohen Investitionskosten lohnen sich die zusätzlichen Implementierungskosten, für eine Abschätzung mit einer begrenzten Anzahl von Testdaten, um Rückschlüsse auf die Auswirkungen bei einer flächendeckenden Umsetzung zu erhalten. Gerade bei den Gebäudedaten handelt es sich um eine sehr teure Datenquelle. Bereits Burrough et al. (1996) und Agumya et al., (1999) berücksichtigt die Auswirkungen der Datenqualität auf die Modellierung und untersuchten, welche Kombinationen von Modell und Daten ein gewisses Maß an Qualität erreichen. Im Folgenden wird eine Sensitivitätsanalyse durchgeführt, um die Auswirkungen von Gebäudedaten auf die Leistungswertbestimmung in der Außenwerbung abschätzen zu können.

Um den Nutzen eines Gebäudelayer zu evaluieren, wurde jeweils für die Agglomerationen St. Gallen, Basel und Bern der ooc mit Gebäudeverschnitt berechnet. Tabelle 4.5 zeigt die Ergebnisse mit Gebäudelayer als indexierte Angaben. Das heißt ein Wert von 100% bedeutet, dass beide Ergebnisse völlig identisch sind und ein Wert über 100% bedeutet, dass das die Berechnung ohne Gebäudelayer höher liegt. Je höher die Differenz zu 100% ist, desto stärker ist der Einfluss des Gebäudelayer.

| | **100 Meter** | **80 Meter** | **60 Meter** | **40 Meter** | **20 Meter** |
|---|---|---|---|---|---|
| **Basel** | 106,3% | 103,9% | 102,1% | 100,9% | 100,0% |
| **St. Gallen** | 109,9% | 105,4% | 102,3% | 100,8% | 100,1% |
| **Genf** | 108,0% | 103,7% | 102,4% | 100,9% | 100,2% |
| Durchschnitt | 108,0% | 104,3% | 102,2% | 100,8% | 100,1% |

Tabelle 4.5: ooc Berechnung mit Gebäudelayer (Hecker et al. 2010c)

Alle Agglomerationen zeigen nur eine geringe Abweichung bis zu der Distanz von 60 Metern. Wenn die Distanz dann erhöht wird, steigt auch der Einfluss des Gebäudelayers. Dies kann durch die Tatsache erklärt werden, dass mit größer werdender Sichtbarkeit Gebäude einzelne Straßensegmente von der Sicht zum Plakat abblocken. Allerdings zeigt die Sensitivitätsanalyse auch, dass die Gebäudeeinschränkung keinen massiven Einfluss auf den ooc hat. Im Bereich der typischen Plakatsichtbarkeitsräume von bis zu 60 Metern liegt die Abweichung der Werte nur bei 2%.

Um die Ergebnisse zu verifizieren, wird ein weiterer Test durchgeführt. Da Gebäude und Straßensegmente in der Regel nicht gleichmäßig verteilt sind, sondern z.B. in der Innenstadt Gebäude und Straßen dichter und engmaschiger stehen als außerhalb der Innenstadt, wird eine zweite Sensitivitätsanalyse durchgeführt. Die Plakatstellen werden diesmal separat für innerstädtische Bereiche und Vororte berechnet. Die Ergebnisse sind in Tabelle 4.6 dargestellt. Sie zeigt, dass die Auswirkungen in der Innenstadt größer sind als in den Vororten. Es bestätigt sich die Vermutung, dass die dichtere Bebauung in der Innenstadt und das feinmaschigere Straßennetz die Bedeutung des Gebäudelayer erhöhen. Jedoch zeigt das Experiment, dass bis zu einem Abstand von 40 Metern die Differenz mit und ohne Gebäude auch gering ist.

| | 100 Meter | 80 Meter | 60 Meter | 40 Meter | 20 Meter |
|---|---|---|---|---|---|
| **Basel Innenstadt** | 106,5% | 104,2% | 102,4% | 100,8% | 100,0% |
| **Basel Vorort** | 104,7% | 103,0% | 101,5% | 100,5% | 100,0% |
| **St. Gallen Innenstadt** | 110,9% | 107,1% | 103,6% | 101,3% | 100,0% |
| **St. Gallen Vorort** | 106,7% | 103,7% | 101,2% | 100,3% | 100,0% |
| **Genf Innenstadt** | 110,8% | 107,2% | 103,6% | 100,9% | 100,0% |
| **Genf Vorort** | 102,9% | 101,6% | 100,9% | 100,4% | 100,0% |
| Durchschnitt Innenstadt | 108,8% | 105,7% | 103,0% | 100,9% | 100,0% |
| Durchschnitt Vorort | 104,0% | 102,4% | 101,2% | 100,5% | 100,0% |

Tabelle 4.6: Anzahl der Passagen mit Verschneidung des Gebäudelayer aufgeteilt nach Plakaten in Innenstadt und Vorstädten (Hecker et al. 2010c)

Erst ab einer Sichtdistanz von 60 Metern steigt der Unterschied im Durchschnitt für die Innenstadt auf 3% und in den Vororten auf 1,2%. Die Analyse zeigt, dass für eine Erstellung von Sichtbarkeitsräumen bis 40 Metern die kostenintensive Modellintegration eines Gebäudelayer nicht zwingend notwendig ist. Sind die Sichtdistanzen jenseits von 40 Metern und gehen sogar bis 100 Metern, ist eine Integration ratsam, insbesondere für innerstädtische Gebiete.

## 4.4 Analyse von Stichprobenverzerrungen in Mobilitätsdaten

Dieser Abschnitt stellt sich der Herausforderung, die eingehende GPS-Empirie nach etwaigen Stichprobenverzerrungen zu untersuchen. Die vorgestellten GPS-Versuchspersonen stellen die Basis für die Hochrechnung auf die Grundgesamtheit in Deutschland und der Schweiz dar. Sollten Verzerrungen in den erhobenen Daten existieren, so wird dieser Fehler bei Nicht-Berücksichtigung vervielfältigt. Es stellt sich die Frage, ob Verzerrungen in den Daten existieren und wie diese behandelt werden müssen.

Zu Beginn des Abschnitts wird in 4.4.1 eine Systematik über die Unvollständigkeit von Daten vorgestellt. Diese Systematik unterscheidet zwischen drei Varianten der Unvollständigkeit, wobei nur eine Variante als ignorierbares Fehlmuster klassifiziert wird. In den nachfolgenden Abschnitten wird mittels KDD-Verfahren eine Suche nach auffälligen Mustern fehlender Daten vorgestellt. Dabei sind etwaige Stichprobenverzerrungen bei folgenden Auswertungen problematisch:

1. bei Auswertungen der Mobilität,
2. bei Auswertungen über die Zeit, sowie
3. bei soziodemographischen Auswertungen.

Diese Auswertungen bzw. Analysen spielen in der späteren Leistungswertausweisung eine wichtige Rolle. Abschnitt 4.5 schließt mit einer zusammenfassenden Betrachtung der Ergebnisse.

### 4.4.1 Systematik über die Unvollständigkeit von Daten

Bereits im Jahre 1976 stellte Rubin (Rubin 1976) eine Systematik über die Mechanismen bei unvollständigen Daten vor und diskutierte den Einfluss der Mechanismen auf den Inferenzprozess. Der Mechanismus bezeichnet dabei die Beziehung zwischen dem Fehlen von Merkmalswerten und den übrigen Merkmalswerten eines Datensatzes. Auf Grund von systematischen Zusammenhängen zwischen einem Merkmalswert und der Ursache für das Fehlen von Daten können Analysen zu Falschaussagen führen. Aus diesem Grund muss überprüft werden, ob „zufällige“ oder „systematische“ Zusammenhänge in den Daten bestehen. Nach der Terminologie von Rubin können drei Varianten voneinander unterschieden werden:

- **MCAR**: „missing completely at random“
- **MAR**: „missing at random“
- **MNAR**: „missing not at random“

Die Abbildung 4.14 zeigt die Zusammenhänge von zufälligen und systematischen Zusammenhängen. Im nachfolgenden werden diese erläutert.

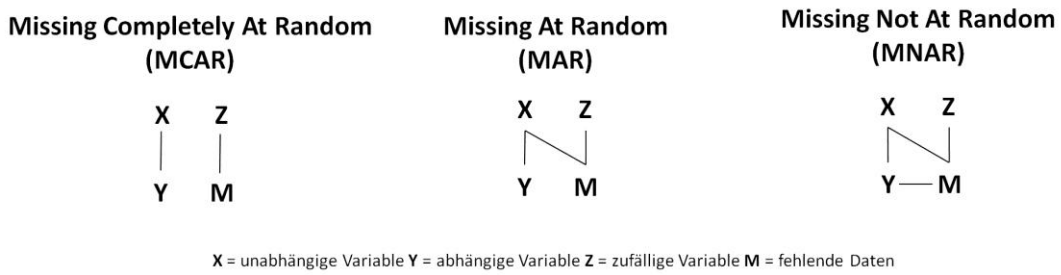



Abbildung 4.14: Systematik fehlender Daten (MCAR, MAR, MNAR) und Beziehungen zwischen einzelnen Variablen

Nehmen wir einen Datensatz mit einer erklärenden Variable X und einer abhängigen Variable Y für n Objekte. Während X vollständig erfasst worden ist, kann Y fehlende Daten enthalten. Die fehlenden Daten von Y können mit Hilfe der Variable M kodiert werden. Hat M den Wert 1, sind Werte für Y enthalten, liegt der Wert bei 0, fehlen Daten für Y. Ferner definieren wir eine Variable Z mit zufälligem Rauschen, welche keine Beziehung zu X und Y aufweist. MCAR tritt auf, wenn das Fehlen von Daten (M) unabhängig von den Daten (X,Y) ist, das heißt: P(M|X,Y) = P(M) (vgl. Abbildung 4.14). Wenn eine Beziehung zwischen M und X existiert, aber M nach wie vor unabhängig von Y ist, spricht die Terminologie von MAR. MAR bezeichnet eine bedingte Unabhängigkeit bei fehlenden Daten bei einem gegebenen Wert von X (vgl. Abbildung 4.14). Man muss allerdings beachten, dass bei MAR eine Beziehung zwischen M und Y existieren kann aufgrund ihrer gegenseitigen Abhängigkeit von X. Wenn die Verteilung von fehlenden Daten auf Y zurückzuführen ist, wird dies als MNAR bezeichnet. MCAR und MAR werden auch als vernachlässigbare Fehlmuster bezeichnet (oder auch nichtinformative Fehlmuster), während MNAR als nicht ignorierbares Fehlmuster (oder informatives Fehlmuster) bezeichnet wird.

Üblicherweise enthalten Datensätze mehrere Variablen, von denen auch mehrere Variablen fehlende Werte aufweisen können. Die oben beschriebene Terminologie von fehlenden Daten kann wie folgt formal beschrieben werden: Es sei $Y = (Y_{1,} Y_{1,} \dots, Y_p)$ ein Menge von Variablen mit Beobachtungen von n Objekten. M bezeichne eine (n x p) Matrix die das Fehlen von Werten in Y kodiert. Unter der Annahme vollständiger Information kann die Datenmenge $Y_{voll} = (Y_{beob}, Y_{fehl})$ in zwei Teilmengen zerlegt werden, die jeweils die beobachteten und fehlenden Werte enthalten. Es handelt sich um MCAR, wenn das Fehlen von Werten unabhängig von den Werten selbst ist, d.h. $P(M|Y_{voll}) = P(M)$. Falls das Fehlen von Werten unabhängig von den beobachteten Werten ist, d.h. $P(M|Y_{voll}) = P(M|Y_{beob})$, liegt MAR vor. Andernfalls sind die fehlenden Daten MNAR.

Je nach Art der Fehlmuster und der Methode der Inferenz kann der geschätzte Parameter eine Verzerrung aufweisen. Existiert MCAR, so stellt dies in der Regel kein Problem für die Schätzung von Parametern aus der Stichprobe dar. Fehlende Werte die dem MAR unterliegen, müssen methodisch behandelt werden, jedoch kann eine unverzerrte Parameterschätzung auf den beobachteten Daten über eine Konditionierung[7] vorgenommen werden. Im Falle von MNAR ist die Parameterschätzung ein schwieriges Problem und erfordert eine explizite Spezifikation der Verteilung der Fehlmuster. Jedoch ist in vielen Fällen der Mechanismus MNAR in den Daten nicht erkennbar, und daher auch nicht behandelbar.

[7] Konditionierung im statistischen Sinn meint die Zerlegung einer gemeinsamen Wahrscheinlichkeitsverteilung mehrerer Variablen in die unbedingte Wahrscheinlichkeit einer Variable und die entsprechenden bedingten Wahrscheinlichkeiten der übrigen Variablen (Bethlehem 2002).

Im folgenden Abschnitt wird ein systematischer Ansatz zur Erkennung von MAR Abhängigkeiten zwischen vollständig beobachteten soziodemographischen Variablen, beobachtetem Mobilitätsverhalten, Trageverhalten und der Messaktivität der Probanden vorgestellt (Hecker et al. 2010a). Dabei wird davon ausgegangen, dass kein MNAR Mechanismus vorliegt. Alle drei Variablen sind wechselseitig miteinander verbunden (siehe Abbildung 4.15). Die soziodemographischen Merkmale enthalten alle erfassten, beschreibenden Informationen über die Probanden und dienen als erklärende Variablen für gefundene Verzerrungen der Stichprobe. Die Messaktivität stellt die zeitliche Information der Mobilitätsstudie dar und ist definiert über die Anzahl der validen mobilen und immobilen Tage. Die Validität wird für jeden einzelnen Probanden und Tag bestimmt.

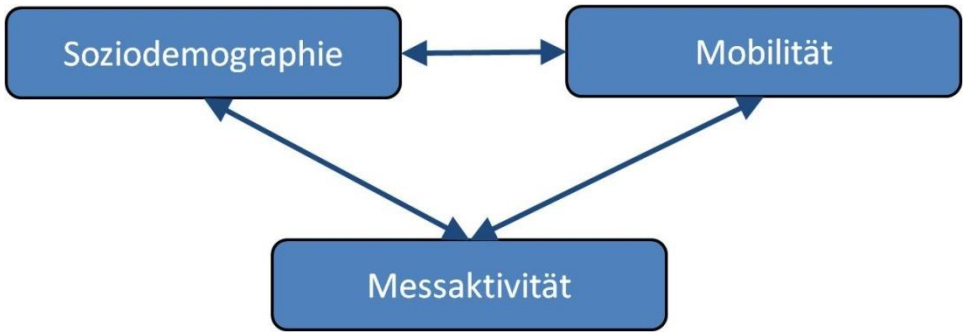



Abbildung 4.15: Zusammenhänge bei Stichprobenverzerrungen

Zuletzt stellt die Mobilität die räumliche Information der GPS-Erhebung dar. Sie ist als jede einzelne Veränderung der Position im Raum bzw. Bewegung definiert.

Um Stichprobenverzerrungen in den Ergebnissen auszuschließen, ist es wichtig, dass Mobilitätsverhalten und Messaktivität unabhängig von allen soziodemographischen Schichtungen sind. In einem ersten Schritt wird aus diesem Grund untersucht, ob eine Abhängigkeit zwischen dem Mobilitätsverhalten und der Messaktivität besteht. Wenn beide Variablen unabhängig sind, dann liefert eine Analyse des gesamten Datensatzes in Bezug auf die Population unverfälschte Ergebnisse. Falls ein Zusammenhang erkannt wird, dann müssen weitere Abhängigkeitsanalysen zwischen dem Mobilitätsverhalten, den soziodemographischen Gruppen und der Messaktivität durchgeführt werden. Auch wenn Aussagen über spezifische soziodemographische Gruppen von Interesse sind, müssen diese Gruppen analysiert werden. Denn abhängig von soziodemographischen Variablen können sich etwaige Zusammenhänge zwischen Mobilitätsverhalten und der Messaktivität ergeben. Das Ziel in beiden Fällen ist es, Faktoren zu bestimmen, bei denen das Mobilitätsverhalten und die Messaktivität in allen möglichen Untergruppen unabhängig sind (Hecker et al. 2010a).

Mit Verwendung der Subgruppenanalyse werden diese Konstellationen untersucht. Wenn gefundene Subgruppen in den Daten darauf hinweisen, dass es einen Zusammenhang zwischen Trageverhalten und Mobilität gibt, dann handelt es sich mindestens um MAR. Eine solche Abhängigkeit kann durch eine Konditionierung der gefundenen soziodemographischen Gruppen ausgeglichen werden. Zusätzlich wird der gesamte Datenbestand, der nicht durch eine Subgruppe beschrieben wird, nach weiteren Abhängigkeiten untersucht. Wenn keine Abhängigkeit vorhanden ist, können alle Abhängigkeiten durch die beobachteten Variablen erklärt werden. Alle Analysen werden unter der Annahme durchgeführt, dass das Mobilitätsverhalten von Personen über die Tragetage korreliert ist, und außerdem das MNAR ausgeschlossen werden kann.

### 4.4.2 Untersuchung von Mobilitätsdaten auf Abhängigkeiten

Die Untersuchung von Stichprobendaten aus Mobilitätsstudien auf Verzerrungen erfordert eine systematische und vollständige Analyse auf Abhängigkeiten zwischen den in 4.2.1 genannten Variablenklassen. Der entwickelte Algorithmus basiert in seinem Kern auf der Subgruppensuche. In einem vierstufigen Prozess werden Subgruppen gesucht, die signifikant in ihrer Verteilung der Messaktivität gegenüber der Grundverteilung abweichen. Der Algorithmus SAAS beschreibt das Vorgehen im Detail.

---

Algorithmus: Subgruppenbasierter Algorithmus zur Analyse von Stichproben-abhängigkeiten (SAAS) (Hecker et al. 2010a)

---

Gegeben

= Datensatz oder Datenbank $X = [x_{ij}]_{n \times m}$ mit $m$ Attributen und $n$ Testpersonen, wobei $x_{ij}$ die Wertausprägung des j-ten Attributes der i-ten Testperson repräsentiert

Gesucht

= subgruppenbasierte Abhängigkeitsbeschreibungen (Regeln) $SD = \{sd_1, \dots, sd_z\}$ als eine Menge von Termen $\{t_1, \dots, t_k, t_y\}$ mit $t_y$ als Repräsentant der Mobilität. Alle anderen Terme $t_f$, $f = 1 \dots k$, haben die Form $(a_j = v_j), v_j \in D(a_j), D$ ist die Domäne des soziodemographischen Attributes $a_j$

Vorgehen:

1: Diskretisiere jedes Attribut in eine sinnvolle Anzahl an Kategorien und begrenze die Dimensionalität der Domäne

2: Analysiere die Abhängigkeiten zwischen der Mobilität und der Messaktivität auf dem gesamten Datensatz

3: Finde alle Subgruppen $SD$ zwischen den soziodemographischen Attributen und der Mobilität als unabhängige, erklärende Attribute und mit dem Attribut zur Messaktivität als Zielattribut

4: Berechne auf der Negativmenge der Subgruppen $\widetilde{SD}$ alle Abhängigkeiten zwischen der Mobilität und der Messaktivität als Klassenattribut

---

Zu Beginn der Untersuchung werden erklärende Attribute und Klassenattribute diskretisiert. Diskretisierung ist der Prozess der Gewinnung einer diskreten Teilmenge aus zuvor kontinuierlich verteilten Datenmengen. Oftmals werden soziodemographische Attribute bereits diskretisiert erfasst oder sind natürlich kategorisiert (z. B. Geschlecht). Einige Attribute werden aus Datenschutzgründen diskret erfasst, wie z. B. Alter, Einkommen, Beruf. Auf der anderen Seite ist es erforderlich, dass Klassenattribute wie bspw. Messaktivität und räumliche Mobilität diskret verteilt sind. Die Anzahl von Kategorien korrespondiert zum Detaillevel der Analyse der Mobilitätsstichprobe. Eine kritische Mindestzahl an Instanzen pro Kategorie muss garantiert werden. Für die Diskretisierung der räumlichen Mobilität bedeutet dies, dass

ausreichend Probanden mit vergleichbarer Mobilität in einer Kategorie zusammengefasst werden müssen.

Die zweite Stufe beinhaltet die Analyse der Stichprobendaten auf Abhängigkeiten zwischen dem Mobilitätsverhalten und der Messaktivität. Mit Hilfe eines Chi-Quadrat-Tests wird geprüft, ob eine gegebene Verteilung zweier Variablen dem Produkt beider Randverteilung übereinstimmt. Sind beide Variablen unabhängig voneinander, sind Analysen auf dem Mobilitätsdatensatz unverzerrt. Andernfalls müssen die Daten in einem Vorverarbeitungsschritt behandelt werden. Am Fallbeispiel soll der beschriebene Zusammenhang erläutert werden. Nehmen wir an, dass in einer Mobilitätsstichprobe Probanden mit einer hohen Mobilität ebenfalls eine hohe Messaktivität haben. Wenn man die mittlere Reisedistanz als durchschnittliche Anzahl absolvierter Tageskilometer bestimmen möchte und vereinfacht den Durchschnitt aller gereisten Kilometer aller Probanden über alle gemessenen Tage bestimmt, würde man die mittlere Reisedistanz der weniger mobilen Probanden überschätzen. Die Ursache liegt in der fehlenden Mobilität der nicht gemessenen Messtage und der daraus resultierenden Verzerrung zwischen Messtagen mit hoher und niedriger Mobilität. Als Resultat von Stufe 2 des Algorithmus liegt eine Indikation vor, ob und in wie weit bereits eine allgemeine Stichprobenverzerrung im Datensatz existiert.

Stufe 3 ist erforderlich, wenn in der vorangegangen Stufe Abhängigkeiten entdeckt worden sind oder wenn Subanalysen für einzelne soziodemographische Gruppen aus der Gesamtstichprobe durchgeführt werden sollen. Für jede Teilmenge oder die gesamte Datenbank wird erneut eine Subgruppensuche durchgeführt, um Abhängigkeiten aufzudecken. Soziodemographische Teilmengen werden durch Splitten eines Datensatzes entlang eines soziodemographischen Terms $t_k = (a_j = v_j)$ gebildet, wobei $v_j$ die Attributausprägung für ein Attribut $a_j$ repräsentiert. Die Stärke $st(sd_z)$ und der Support $su(sd_z)$ einer beliebigen Subgruppe $sd_z = \{t_1, \dots, t_k, t_y\}$ deutet die Höhe der Abhängigkeit zwischen der soziodemographischen Eigenschaft oder der Mobilität und der Messaktivität an. Der Support beschreibt den Anteil der Stichprobe bzw. die Menge aller Instanzen, die durch die Subgruppe beschrieben sind. Als Stärke bezeichnet man den Wahrscheinlichkeitsanteil der Zielvariablen innerhalb der Subgruppe.

Wenn die Stärke $\mathrm{st}(\mathrm{sd}_z)$ einer Subgruppe mit Mobilitätsattributen einen bestimmten Schwellwert überschreitet, deutet dies auf eine problematische Beziehung zwischen der Mobilität und der Anzahl valider Messtage. Hier liegt auch eine der zentralen Vorteile der Subgruppensuche. Sie führt mit entsprechenden Heuristiken eine vollständige Durchsuchung des Suchraums durch. Damit werden alle möglichen Schichtungen der Grundgesamtheit in einem Schritt durchgeführt (Großkreutz et al. 2008).

Alle Instanzen, die nicht durch eine oder mehrere Subgruppen beschrieben sind, bilden die Negativmenge. In Schritt 4 wird für diese Negativmenge die Abhängigkeit zwischen der Mobilität und der Messaktivität berechnet. Sofern eine signifikante Abhängigkeit erkannt wird, sind die soziodemographischen Variablen nicht ausreichend, um die Verzerrung vollständig zu kompensieren.

Bei Analysen mit räumlichem Kontext ist es angebracht, die einzelnen Analysen individuell für jede Regionen durchzuführen. Um Seiteneffekte zu vermeiden, z.B. kann das Mobilitätsverhalten einer Region anders sein, als in einer anderen Region.

### 4.4.3 Motivation der Subgruppensuche

Nicht zufällige Effekte im Verlauf einer mobilitätsbezogenen Umfrage können leicht zu fehlenden Daten führen, die die Stichprobe verzerren und das Analyseergebnis beeinflussen. Geeignete algorithmische Verfahren, um derartige Einflüsse in einer Mobilitätsstichprobe aufzudecken, verfolgen vier konkrete Ziele:

(1) Unterstützen beim Datenverständnis

(2) Entdecken signifikanter Abhängigkeiten zwischen Variablen der Erhebung

(3) Liefern von Erkenntnissen über die Art und Intensität der Abhängigkeit

(4) Erkennen *aller* Abhängigkeiten über einem Qualitätslevel

Ein Verfahren, das all diese Anforderungen erfüllt, ist die Subgruppensuche. Die Subgruppensuche sucht Mustern und bzw. statistischen Abhängigkeiten zwischen Variablen einer Stichprobe. Konkret wird nach Subgruppen gesucht, die eine signifikante Abweichung in der Wahrscheinlichkeitsverteilung in Bezug auf ein Klassenattribut (Zielparameter) gegenüber der gesamten Stichprobenpopulation aufweisen (Wrobel 1997, Klösgen 2002). Folglich ist eine Subgruppe eine Teilmenge der Stichprobe, die durch eine Regel beschrieben ist. Eine Regel ist eine Kombination von Attribut-Merkmalsausprägungen. Das Ziel der Subgruppensuche ist die Identifikation von Regeln zur Beschreibung von Teilmengen der Stichprobe, die:

i. eine hohe Generalität (Support) haben, sowie

ii. eine positive Tendenz in Richtung der Zielklasse (Stärke).

Mit algorithmischen Weiterentwicklungen der Subgruppensuche ist nunmehr auch eine erschöpfende Suche nach Subgruppen im vollständigen Suchraum möglich. Wenn Zielattribute (multi-class) zusammengesetzt sind. Damit unterscheidet sich das Subgruppenverfahren von anderen Verfahren, wie dem Entscheidungsbaumlernen (vgl. Abschnitt 3.4). Ein weiterer Vorteil liegt darin, dass alle Subgruppen in einem Schritt entdeckt werden. Andere Verfahren wie die Regressionsanalyse liefern nur die signifikantesten Ergebnisse. Wenn es um die Identifikation von Abhängigkeiten in Stichproben geht, eignet sich auch die Varianzanalyse ANOVA (Fahrmeir 1999). Allerdings liefert diese keine Information zum Ursprung der Abhängigkeit oder andere erklärenden Zusammenhänge. Subgruppenbeschreibungen als Regeln sind hingegen „lesbar" und fördern das Verständnis der Stichprobendaten in Bezug auf die Verteilung, Verzerrungen und Abhängigkeiten.

Systematische Fehler können mit Hilfe statistischer Tests identifiziert werden. Hypothesen über die Verteilung der Grundgesamtheit können mittels Signifikanztests verifiziert werden. Die Subgruppensuche verwendet für die Bewertung der Hypothesen (Regel) eine Qualitätsfunktion wie $q(h) = \frac{|p-p_0|}{\sqrt{p_0(1-p_0)}}\sqrt{n}$ mit $p$ als Wahrscheinlichkeit über die Klassen in der Subgruppe, $p_0$ als Wahrscheinlichkeit über die Klassen in der gesamten Stichprobe und $n$ als Anzahl Instanzen in der Subgruppe.

### 4.4.4 Aufbereitung der Daten zur Mustererkennung

In diesem Abschnitt wird das unter Abschnitt 4.4.2 vorgeschlagene Vorgehen zum Aufdecken verzerrender Einflüsse in Mobilitätstichproben auf einen realen Datensatz angewendet. Der Datensatz umfasst das Mobilitätsverhalten von 11.000 Testpersonen. Innerhalb der Stichprobendaten sollen verzerrende Einflüsse und Abhängigkeiten aufgedeckt werden, die durch das Fehlen kompletter Erhebungstage entstehen. Das Vorgehen wird anhand dreier Beispielregionen demonstriert, darunter eine kleine, eine mittlere und eine große Agglomeration gewichtet nach den Einwohnern. Im Einzelnen sind dies:

- Chur mit 55.000 Einwohnern,
- St. Gallen mit 122.000 Einwohnern und
- Genf mit 390.000 Einwohnern

Im weiteren Verlauf werden die Subgruppen mit der höchsten Qualität präsentiert.

Diese empirische Fallstudie ist wie folgt aufgebaut. In den nächsten drei Teilabschnitten werden die Attributklassen und die Diskretisierungsprozedur beschrieben. Anschließend werden zentrale Ergebnisse der Untersuchung der Mobilitätsdaten auf Abhängigkeiten und Einflüsse durch fehlende Daten in Form identifizierter Subgruppen dargestellt und interpretiert.

**Soziodemographie**

Jeder Teilnehmer der Schweizer Mobilitätsstudie hatte zwei standardisierte Fragebögen zu beantworten. Bevor eine Person ausgewählt und eine Teilnahme angeboten wurde, musste jeder potenzielle Teilnehmer Auskunft über seine Person, seine Lebensverhältnisse und sein alltägliches Mobilitätsverhalten geben. Insgesamt 33 soziodemographische Merkmale wurden so erfasst. Darunter sind Informationen zu Alter, Geschlecht, Wohn- und Arbeitsadresse, Bildung, Beruf, Einkommen und Haushalt. Als konzipierte Mobilitätserhebung sind auch Daten zum persönlichen Mobilitätsverhalten gesammelt worden, wie Affinität zur Nutzung des öffentlichen Nahverkehrs und täglichen Aktivitäten (z.B. Hobbies).

**Messaktivität**

Im Rahmen der Studie ist jede Testperson über einen Zeitraum von 7 Tagen sensorgestützt begleitet worden. Hierfür kam ein batteriebetriebenes Erfassungsgerät mit GPS-Modul zum Einsatz („MobilityMeter“). Die Messaktivität $\mathrm{ma}_p$ einer Testperson $p \in P$ bezeichnet die individuelle Anzahl valider Erhebungstage einer Person. Während der siebentägigen Erhebungsphase gibt es unterschiedliche Einflüsse, die zu fehlenden Daten führen können (siehe Abschnitt 4.1). Relevant für die Messaktivität ist hier das Fehlen ganzer Erhebungstage.

Obwohl die Aggregation in Tagen bereits eine Diskretisierung der Messzeit darstellt, wird diese gemäß Stufe 1 des Algorithmus auf einer noch höheren Ebene zusammengefasst. Auf diese Weise lässt sich die Komplexität des Suchproblems reduzieren. Subgruppensuche und andere maschinelle Lernverfahren hängen direkt von der Dimensionalität des Suchraums ab, einerseits in Bezug auf die Laufzeit und andererseits in Bezug auf die Entdeckung von Abhängigkeiten. Hierbei spielen die Qualitätsfunktion, sowie die Stärke und der Support eine

wichtige Rolle. Fallen zu wenige Instanzen in eine Gruppe, wird diese als wenig signifikant möglicherweise verworfen. Bei kontinuierlichen Attributen führt dies dazu, dass nicht alle Subgruppen gefunden werden können. In diesem konkreten Fall werden die validen Messtage in drei sogenannte Teilnahmetypen $pt$ klassifiziert: niedrig (0-3 Tage), normal (4-5) und hoch (6-7 Tage). Für jeden Teilnehmer ergibt sich:

$$p_t = \begin{cases} niedrig \;\; wenn\, ma_p < 4 \\ hoch \;\; wenn\, ma_p > 5 \\ normal \quad sonst \end{cases}$$

Die Verteilung der Teilnahmetypen veranschaulicht Abbildung 4.16. Die übrigen erhobenen Agglomerationen (darunter Zürich) zeigen ähnliche Verteilungsmuster. Obgleich die Verteilung der Anteile auf den ersten Blick sehr homogen erscheint, zeigen einzelne Gebiete signifikante Abweichungen. Dies stützt die Empfehlung einer raumbezogenen Analyse pro Testregion zur Berücksichtigung lokaler Phänomene. Auf diese Weise wird das Entdecken räumlicher Subgruppen unterstützt.

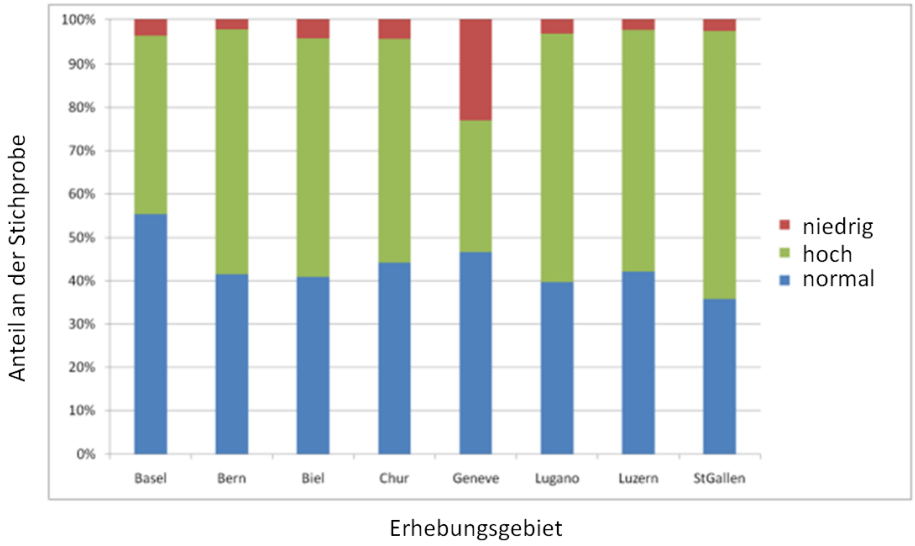



Abbildung 4.16: Verteilung der Teilnahmetypen in der Stichprobenauswahl in 8 ausgewählten Schweizer Agglomerationen (Hecker et al. 2010a)

Die Teilnehmer der Teilstudie in Genf zeigen im Vergleich zu anderen Teilstudien einen signifikant höheren Anteil vom Teilnahmetyp $pt = "niedrig"$ (vgl. Abbildung 4.16). In Basel ist der Anteil vom Teilnahmetyp $pt = "normal"$ im Vergleich am höchsten. Keine dieser Feststellungen erlaubt eine belastbare Beurteilung der Stichprobenverzerrung. Letztendlich könnte es sich um eine Eigenschaft des Gebietes bzw. der dortigen Bevölkerung handeln.

### Mobilität

Im Folgenden wird die Reisedistanz als Indikator verwendet, um mögliche Abhängigkeiten zwischen dem Mobilitätsverhalten und anderen Attributen zu erkennen. Die Mobilität $m_p$ einer Testperson $\mathrm{p} \in \mathrm{P}$ ergibt sich aus der mittleren Reisedistanz pro Messtag. Tage der Immobilität gehen mit einer Reisedistanz von 0 (Null) in die Berechnung ein. Für jede Testperson errechnet sich die mittlere individuelle Mobilität $m_p$ als:

$$m_p = \frac{1}{ma_p} \cdot \sum_{i=1}^{ma_p} len(p, d_i)$$

wobei $p$ die Testperson mit einer Messaktivität von $ma_p$ Tagen und der Reiseleistung $len(p, d_i)$ am i-ten validen Erhebungstag $d_i$ ist. $m_p$ wird in Kilometern pro Tag angegeben. Entsprechend der Mobilität kann die mittlere alltägliche Mobilität individuell ausgeprägt sein. Das metrische Attribut wird im Anschlussschritt in ein Merkmal mit nominaler Skalierung transformiert. Wie bereits erwähnt, ist es für wahrscheinlichkeitsbasierte Algorithmen leichter, Zusammenhänge und Abhängigkeiten im niedrig dimensionalen Suchraum zu finden.

Für jedes Testgebiet $r \in R$ wird die mittlere, tägliche Mobilität $m_r$ in Kilometern berechnet:

$$m_r = \frac{1}{N_r} \cdot \sum\nolimits_{p \in P_r} m_p$$

wobei $P_r$ die Menge aller Personen eines Gebiets $r$ und $N_r = |P_r|$ die Anzahl aller Testpersonen eines Gebietes ist. Die Variabilität in der Mobilität eines Gebietes sich ergibt ferner aus:

$$\sigma_r = \sqrt{\frac{1}{N_r} \cdot \sum\nolimits_{p \in P} (m_p - m_r)^2}$$

Um die Mobilität der einzelnen Personen untereinander vergleichbar zu machen, wird die persönliche mittlere Mobilität $m_p$ einer Testperson $p \in P_r$ aus einem Testgebiet $r \in R$ transformiert. Aufgrund von möglichen Messfehlern und unbekannten Ausreißern wird an dieser Stelle die z-Transformation gegenüber anderen Transformationen (z.B. Min-Max) bevorzugt. Sie errechnet sich wie folgt:

$$w(p, r) = \frac{m_p - m_r}{\sigma_r}$$

Die z-transformierten Mobilitätsindikatoren werden in Standardabweichungen angegeben und stellen die Abweichung vom Mittelwert unter Berücksichtigung der territorialen Variabilität der Mobilität dar.

Auch dieses Merkmal wird diskretisiert, um es für die Subgruppensuche aufzubereiten. Dazu wird der normalisierte Mobilitätsindikator $w$ über folgende Zuordnungsfunktion einem von drei Mobilitätstypen zugewiesen:

$$mt_p = \begin{cases} niedrig \;\; wenn\, w(p,r) < -1 \\ hoch \;\; wenn\, w(p,r) > 1 \\ normal \quad sonst \end{cases}$$

Abbildung 4.17 zeigt ein Histogramm der Verteilung der Mobilitätstypen in der untersuchten Stichprobe. Für die beispielhaft gewählten 8 Testgebiete wird deutlich, dass die Mehrheit der Testpersonen mit obiger Zuordnungsfunktion dem Mobilitätstyp „normal“ zugeordnet worden sind. Allgemein fällt auf, dass der Mobilitätstyp „hoch“ relativ gleichmäßig über alle Testgebiete verteilt ist. Dies deutet darauf hin, dass es in jedem Gebiet einen gleichen Anteil von Testpersonen gibt, deren Kilometerleistung über der mittleren Gebietsmobilität liegt. St. Gallen weist von allen Testregionen die größten Abweichungen auf. Insbesondere der Mobilitätstyp „niedrig“ ist nahezu nicht vorhanden. Genf zeigt im Besonderen im Bereich des Mobilitätstypen „niedrig“ einen höheren Anteil auf, was auf eine größere Anzahl Testpersonen mit niedrigerer Kilometerleistung schließen lässt. Grundsätzlich zeigt sich über

alle Testgebiete eine sehr gleichmäßige Verteilung der Mobilitätstypen von 2/3 vom Typ „normal“, 1/6 vom Typ „niedrig“ und 1/6 „hoch“.

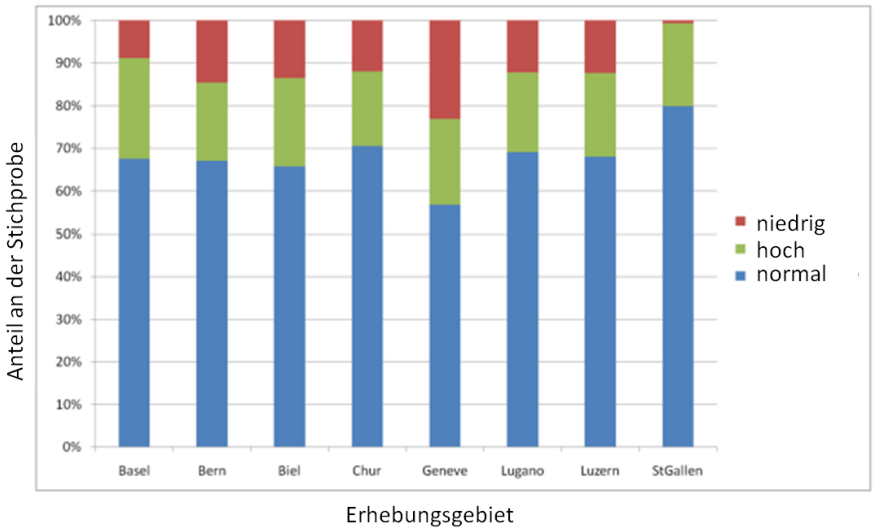



Abbildung 4.17: Verteilung der Mobilitätstypen in den Stichprobendaten (Hecker et al. 2010a)

### 4.4.5 Muster- und Abhängigkeitsentdeckung

In GPS-basierten Mobilitätsstudien weist jeder Proband ein individuelles Teilnahmeverhalten bis hierher auf. Dieses Individualverhalten ist problematisch, wenn eine Abhängigkeit zwischen der Mobilität und Messaktivität existiert. Der zweite Schritt des vorgeschlagenen Vorgehens zielt daher darauf ab, derartige Abhängigkeiten in den Stichprobendaten zu entdecken. Um die stochastische Unabhängigkeit beider kategorischer Variablen – Messaktivität und Mobilität – zu prüfen, eignet sich ein Chi-Quadrat-Test. Die Tabelle 4.7 visualisiert das Ergebnis für die untersuchte Mobilitätsstichprobe. Daraus geht hervor, dass beide Variablen voneinander abhängig sind ($x^2 = 767$ bei einem Freiheitsgrad von 4 und $\alpha = 0{,}05$ mit $x^2(0{,}95; 4) = 9{,}488$). Mit $x^2 > 9{,}488$ wird die Nullhypothese abgelehnt. In der Subgruppenanalyse wird dieses Ergebnis bestätigt. Es existiert die Subgruppe mit dem Merkmal $Mobilität = "hoch"$ für das Zielattribut $Messaktivität = "hoch"$.

| | | **Mobilität** | | | |
|---|---|---|---|---|---|
| | | niedrig | normal | hoch | Total |
| **Messaktivität** | Niedrig | 304 | 73 | 4 | 381 |
| | Normal | 813 | 666 | 11 | 1.490 |
| | hoch | 1172 | 3238 | 160 | 4.570 |
| | total | 2.289 | 3.977 | 175 | 6.441 |

Tabelle 4.7: Kreuztabelle der Mobilität und Messaktivität (Hecker et al. 2010a)

In der dritten Stufe des algorithmischen Vorgehens wird die Subgruppensuche systematisch für jede Agglomeration (Testgebiet) angewendet, um signifikante und potenziell kritische Abhängigkeiten zwischen der Mobilität der Testpersonen und Lücken in den Messdaten aufzudecken. Genaugenommen werden Regeln gesucht, die Gruppen von Testpersonen mit

signifikant niedrigerer oder höherer Messaktivität anhand ihrer Mobilitätklasse (niedrig, normal, hoch) und ihren soziodemographischen Merkmalen beschreiben. Dieses Wissen kann dazu genutzt werden, um auf die erkannten Abhängigkeiten zu konditionieren und so die Abhängigkeiten und verzerrenden Einflüsse zu kompensieren.

Die aufgedeckten Muster in den Experimenten sind in Tabelle 4.8 abgebildet. *Support* ist die Anzahl der Instanzen, die durch die Subgruppenregel beschrieben wird. *Stärke* ist der Anteil der Teilmenge der Instanzen, die zur Zielklasse gehören. Die Klassenverteilung ist in Spalte $p_0$ angegeben. Abhängig von der verwendeten Suchstrategie und der Qualitätsfunktion werden Subgruppen von $p_0$ abweichender Stärke oder höherem Support bevorzugt.

| **Subgruppen für Zielattribut: niedrige Messaktivität ($pt = \text{niedrig} \equiv$ 0-3 validen Messtagen)** | | | | | | |
|---|---|---|---|---|---|---|
| **Index Gebiet** | **Support** | **Stärke $p$** | **$p_0$** | **Größe** | **Qualität** | **Subgruppenregel (Beschreibung)** |
| 1.1 St. Gallen | 0,16 | 0,07 | 0,03 | 122 | 0,17 | $[arbeitslos] = \text{"Ja"} \wedge [Abbo\_\text{ÖPNV}] = \text{"Ja"}$ |
| 1.2 St. Gallen | 0,14 | 0,14 | 0,03 | 101 | 0,17 | $[Transportmittel_{Arbeit}] = \text{"Bus, Tram"} \wedge [Distanz\_Home2Work] = \text{"Ja"}$ |
| 1.3 Genf | 0,21 | 0,40 | 0,23 | 302 | 0,42 | $[Transportmittel_{Arbeit}] = \text{"Bus, Tram"}$ |
| 1.4 Genf | 0,13 | 0,48 | 0,23 | 190 | 0,42 | $[ANZ\_Autos\_HH] = 0$ |

| **Subgruppen für Zielattribut: hohe Messaktivität ($pt = \text{hoch} \equiv$ 6-7 validen Messtagen)** | | | | | | |
|---|---|---|---|---|---|---|
| **Gebiet** | **Support** | **Stärke $p$** | **$p_0$** | **Größe** | **Qualität** | **Subgruppenregel (Beschreibung)** |
| 2.1 St. Gallen | 0,04 | 1,00 | 0,62 | 29 | 0,48 | $[Abschluss] = \text{"Prof. Schule"} \wedge [Distanz] = \text{"Ja"} \wedge [single] = \text{"Ja"} \wedge [Zielgruppe] = \text{"Weiblich\_45+"}$ |
| 2.2 St. Gallen | 0,10 | 0,83 | 0,62 | 75 | 0,48 | $[Beruf] = \text{"Manager, Sen. Beamter"}$ |
| 2.3 Genf | 0,09 | 0,51 | 0,3 | 132 | 0,45 | $[Beruf] = \text{"Repräsent., Beamter"} \wedge [Transportmittel_{Arbeit}] = \text{"Auto"} \wedge [ANZ\_Auto\_HH] = \text{">1"} \wedge [Distanz] = \text{"Ja"}$ |
| 2.4 Genf | 0,46 | 0,39 | 0,3 | 673 | 0,45 | $[Transportmittel_{Arbeit}] = \text{"Auto"}$ |

Tabelle 4.8: Subgruppenbeschreibung mit Regeln hoher Qualität in Bezug auf eine hohe oder niedrige Mobilität gegenüber dem Zielattribut (nach Hecker et al. 2010a)

Im Allgemeinen haben die Experimente eine Vielzahl der Attribute mit Datenlücken in der Stichprobe in Verbindung gebracht. Der Einfluss von individueller Mobilität und den persönlichen Eigenschaften der Testpersonen ist anhand von identifizierten Subgruppen belegt worden. Dass der Beschäftigungsstatus oder das verwendete Verkehrsmittel die Messaktivität an einer Mobilitätsstudie beeinflussen, kann wohl erwartet werden. Auch einige Berufsgruppen oder bestimmte Lebensumstände lassen ein unterschiedliches Teilnahmeverhalten eher erwarten. Die Beziehungen zwischen soziodemographischen Attributen und der Teilnahmebereitschaft (Messaktivität) zu kennen hilft nicht nur beim Verständnis der Gruppen, sondern kann auch beim Design der Studie einbezogen werden.

Ein zentraler Vorteil der Subgruppensuche ist die Verständlichkeit der Ergebnisse. Eine Subgruppe wird durch eine konkrete Regel als Attribut-Wert-Paar beschrieben. Zur Veranschaulichung werden im Folgenden einige solcher Subgruppen im Detail erklärt. So übersetzt sich die Subgruppe 1.4 in Tabelle 4.8 als $[Anzahl\ Fahrzeuge\ im\ Haushalt] = 0$. Daraus wird ersichtlich, dass es eine Abhängigkeit zwischen einer niedrigen Messaktivität und dem Fehlen eines Fahrzeuges gibt. Andere Subgruppen lassen darauf schließen, dass die Nutzung öffentlicher Verkehrsmittel ein Indikator für eine Mehrzahl fehlender Daten darstellt (vgl. Subgruppen 1.1 und 1.3 in Tabelle 4.8). Anders in St. Gallen, dort weisen Manager und höhere Beamte eine bessere Beteiligung und signifikant weniger Fehltage auf, wie aus Subgruppe 2.2 hervorgeht: $[Beruf] = "Manager, höherer\ Beamter"$. Diese Aussage wird ebenfalls durch die Subgruppe 2.3 im Erhebungsgebiet Genf unterstützt. Subgruppe 2.3 besagt, dass Vertreter und Beamte, die mit dem Auto zur Arbeit fahren, eine signifikant höhere Messaktivität aufweisen im Vergleich zu anderen soziodemographischen Gruppen.

Bei der Analyse der Agglomeration Chur (Schweiz) sind keine signifikanten Subgruppen gefunden worden. Natürlich sind Subgruppen entdeckt worden, die einzelne soziodemographische Attributausprägungen mit einer bestimmten Messaktivität in Bezug setzen, allerdings, und dies ist entscheidend für das Aufdecken verzerrender Stichprobeneffekte, weisen diese Subgruppen von Personen kein besonderes Mobilitätsverhalten auf. Schlussfolgernd führen Analysen auf den Stichprobendaten aus Chur zu repräsentativen Ergebnissen. Diese Feststellung belegt ein zusätzlich durchgeführter Chi-Quadrat-Test auf Unabhängigkeit zwischen Messaktivität und Mobilität $(x^2 = 2{,}9\ mit\ x^2(0{,}95; 4) = 9{,}488)$. Dieser zeigt, dass keine stochastische Abhängigkeit besteht und die Nullhypothese der Unabhängigkeit bestätigt wird.

Zusammenfassend werden in Schritt drei des SAAS interessante Muster im Zusammenhang mit der Messaktivität entdeckt. Alle entdeckten Subgruppen helfen durch ihre Interpretierbarkeit dabei, die Psychologie hinter der Beteiligung besser zu verstehen. Personengruppen wurden identifiziert, die signifikant mehr oder weniger Messtage aufweisen. Für übertragbare oder repräsentative Ergebnisse in Mobilitätsstudien sind fehlende Mobilitätsdaten problematisch. Wenn diese nicht zufällig verteilt sind (MCAR), sondern in Zusammenhang mit der Mobilität der Probanden stehen (MAR). Diese müssen dann entsprechend der Abhängigkeiten behandelt werden.

Ein Ergebnis für die Agglomeration Genf ergibt sich aus den dort gefundenen Subgruppen und weist auf eine Abhängigkeit zwischen dem aufgezeichneten Mobilitätsverhalten und der Anzahl valider Messtage im Erhebungszeitraum hin (vgl. Subgruppen 1.3, 1.4, 2.3 und 2.4 aus Tabelle 4.8). Diese Verzerrungen der Stichprobe können unter Ausnutzung der Subgruppenbeschreibung behandelt werden, indem auf die gefundenen Abhängigkeiten konditioniert wird. In obigem Beispiel würde man auf das Attribut-Werte-Paar $[Beruf] = "höherer\ Beamter"$ konditionieren.

Auch die Subgruppe 2.2 in St. Gallen mit einer Stärke von 0,83 und $p_0$ von 0,62 deutet auf eine signifikante Beziehung zwischen Beruf und der Messaktivität hin. Manager und höhere

Beamte haben weniger Fehltage und zusätzliche eine höhere Mobilität. Die Analysen sollten also nach dem Beruf geschichtet werden, um diese Entdeckung zu berücksichtigen. Ansonsten wird diese berufliche Zielgruppe andere Gruppen überlagern. Derartige Stichprobenverzerrungen führen zu falschen Eindrücken oder Aussagen über die Mobilität und das Mobilitätsverhalten der studierten Zielgruppe. Im konkreten Fall würde die Mobilität stark überbewertet werden.

Im Gegensatz zur Übermobilität einzelner Gruppen in der Stichprobe existieren weitere Faktoren wie die Affinität zu gewissen Verkehrsmitteln, die die gemessene Mobilität negativ beeinflusst. Subgruppen 1.3 und 1.4 setzen eine niedrige Mobilität mit dem Verkehrsmodus in Beziehung. Allerdings erlauben sie noch keine Ergründung der Ursachen. Diese fehlen im Attributvektor der Erhebung. Es darf spekuliert werden, dass die GPS-Geräte in Bahnen oder Bussen nicht zuverlässig funktionieren.

In der vierten Stufe des SAAS wird die Negativmenge aller Subgruppendaten $\widetilde{SD}$ genauer untersucht. Die Negativmenge ergibt sich als $\widetilde{SD} = \{sd \mid \text{att(x)} \neq \text{sd} \; \forall \; sd \in SD\}$ mit $att(x)$ als Attribut-Wert-Kombination der Instanz $X$. Um zu prüfen, ob es unentdeckte Abhängigkeiten in der Restmenge der Stichprobe existieren, die durch keine Subgruppe beschrieben wird, wird die Korrelation zwischen der Anzahl valider Messtage im Erhebungszeitraum (Messaktivität) und der Mobilität der Testpersonen berechnet. Die Korrelation schätzt die Stärke der Abhängigkeit zwischen beiden Variablen. Im günstigsten Fall wäre die Korrelation sehr niedrig. Dies impliziert, dass die fehlenden Daten keinen Einfluss auf die Mobilitätsverteilung in der restlichen Stichprobe haben, d.h. die fehlenden Daten sind zufällig verteilt.

Die in St. Gallen existierenden Abhängigkeiten (ursprünglich $r_X = 0{,}21$) im gesamten Datensatz können durch geeignetes Vorverarbeiten reduziert werden auf $r_{\widetilde{SD}} = 0{,}08$. Genf hingegen zeigt die höchste Abhängigkeit in der Stichprobe. Obwohl eine Reihe von Abhängigkeiten mithilfe der Subgruppensuche aufgedeckt werden konnten, bleibt die Korrelation vergleichsweise hoch mit $r_{\widetilde{SD}} = 0{,}47$, ein Indikator für versteckte Abhängigkeiten, die durch keine erhobene Variable oder deren Kombination erklärt werden kann.

**Zusammenfassung Subgruppensuche**

Die erfassten GPS-Daten umfassen neben der aufgezeichneten Mobilität auch umfangreiche soziodemographische Daten. Nicht immer ist bekannt, welche Faktoren zu Datenlücken führen. Ist evtl. eine besondere soziodemographische Gruppe dafür verantwortlich? Dieser Frage muss nachgegangen werden, um keine Verzerrungen in den späteren Ergebnissen zu erhalten. Aus diesem Grund wurde in diesem Abschnitt ein mehrstufiges Verfahren vorgestellt, das signifikant auftretende Muster in den Daten sucht. Der Kern des Ansatzes besteht aus einer Subgruppen-Analyse, die Abhängigkeiten zwischen soziodemographischen Gruppen, fehlenden Tagen und dem Mobilitätsverhalten liefert. Die Ergebnisse einer solchen Untersuchung können dazu verwendet werden, Verzerrungen aufgrund des Trageverhaltens der Probanden mittels Konditionierung zu kompensieren (MAR).

## 4.5 Zusammenfassung

In diesem Kapitel wurden die Problematik bei der Aufbereitung und Validierung von GPS-Daten beschrieben. Zu Beginn des Kapitels wurden typische Charakteristiken bei der Erhebung mittels GPS-Studien erläutert (Abschnitt 4.1). Zu diesen zählen das Fehlen einzelner Wegteilstrecken, das Fehlen vollständiger Wege sowie das Fehlen kompletter Tage. Es wurde festgestellt, dass der Anteil fehlender Tage einen signifikanten Anteil ausmacht. Ignorieren kann man an dieser Stelle das Fehlen nicht, da sonst die Mobilität und damit die Leistungswerte unterschätzt werden.

Im folgenden Abschnitt 4.2 wurde die Datenaufbereitung von GPS-Trajektorien vorgestellt. Dies diente dazu, Aufzeichnungslücken bei Teilstrecken zu schließen und Fehlverortungen zu eliminieren. Über ein topologisches Map Matching Verfahren wurden nach einer ersten Filterung die Roh-GPS-Punkte mit dem Straßennetz zusammengeführt. Lücken, die bei der Anlaufphase eines GPS-Gerätes entstehen können, wurden durch ein Routing geschlossen. Weiterhin wurden die Trajektorien nach den Verkehrsarten PKW und Fußgänger klassifiziert. Im Ergebnis liegen alle gesammelten GPS-Punkte nun auf Basis von Straßensegmenten vor.

Abschnitt 4.3 befasste sich mit einem sehr wichtigen anwendungsorientierten Verfahren im Kontext der Arbeit. Es wurde vorgestellt, wie und wann eine Passage mit einer Plakatstelle zu einem opportunity of contact (ooc) führt. Hierzu sind zwei Datensätze von entscheidender Bedeutung. Erstens die Trajektorie und zweitens der individuelle Sichtbarkeitsraum eines Plakates. Es wurden Regeln aufgestellt, wann eine Trajektorie zu einem potenziellen Kontakt führt. Weiterhin wurde untersucht, welchen Einfluss die Größe und die Individualisierung von Sichtbarkeitsräumen auf Gebäuden haben. Es konnte festgestellt werden, dass der Einfluss erst ab einer Sichtdistanz von 60 Metern zunimmt, und es dann zu einem größeren Unterschied in der Anzahl der Passagen kommt.

Nach der Aufbereitung der GPS-Daten wurde untersucht, ob Stichprobenverzerrungen in den Daten enthalten sind (Abschnitt 4.4). In diesem Zusammenhang wurde eine Terminologie vorgestellt, die Zusammenhänge zwischen fehlenden Daten und den Variablenwerten eines Datensatzes beschreibt. Systematische Zusammenhänge könnten zu späteren Falschaussagen führen. Drei Varianten existieren in der Literatur, wobei nur die Variante MNAR zu einem schwerwiegenden Problem führt. In diesem Abschnitt wurde ein systematischer Ansatz zur Erkennung von MAR Abhängigkeiten zwischen vollständig beobachteten soziodemographischen Variablen, beobachteten Mobilitätsverhalten und dem Trageverhalten der Probanden vorgestellt. Es ist zu beachten, dass alle drei Variablen wechselseitig miteinander verbunden sind. Mit der Subgruppensuche wurde ein KDD Verfahren ausgewählt, welches nach Gruppen mit signifikant abweichenden Trageverhalten in einem Datensatz sucht.

Mit diesem Kapitel wurden die Grundlagen gelegt, um in den folgenden beiden Kapiteln die Reichweitenberechnung und die Ausweisung der Leistungswerte vorzustellen.

# KAPITEL 5

## 5. MODELLIERUNG VON LEISTUNGSWERTEN IN DER AUßENWERBUNG MIT GPS-MOBILITÄTSSTUDIEN

Im vorherigen Kapitel wurden die Aufbereitung, die Charakteristiken und die Validierung der zur Verfügung stehenden GPS-Stichprobe vorgestellt. Dabei wurden die Daten so aufbereitet, dass sie auf Straßenabschnittsebene vorliegen und mit Sichtbarkeitsräumen von Plakatstellen verschnitten wurden. Diese Daten dienen nun als Basis für die weiteren Modellierungsschritte.

In diesem Kapitel wird der Frage nachgegangen, wie die aufbereiteten Daten in eine valide Leistungswertberechnung einfließen können. Dabei müssen zuvor zwei wesentliche Problemstellungen der Dissertation gelöst werden, nämlich erstens, wie man mit einer GPS-Stichprobe umgeht, die eine zeitliche Unvollständigkeit in der Erhebung besitzt. Und zweitens, wie man mit einer Stichprobe umgeht, die eine unvollständige räumliche Abdeckung des Untersuchungsgebietes aufweist. Denn obwohl z.B. die GPS-Feldstudie in Deutschland eine sehr umfangreiche Stichprobe darstellt, können eine Vielzahl von Plakatstellen nicht direkt bewertet werden, da sie keine GPS-Passagen aufweisen. In Abschnitt 5.2 wird die Problemstellung nochmals klar definiert und eine Übersicht zur besseren Darstellung der einzelnen Modellierungsschritte gegeben.

In Abschnitt 5.1 wird die Problematik der komplett fehlenden Erfassungstage von GPS-Probanden dargestellt und mögliche Varianten der Modellierung vorgestellt. Fehlende Tage, an denen Mobilität ohne GPS-Aufzeichnung stattgefunden haben, müssen beachtet werden. Ignoriert man die fehlenden Messdaten, unterschätzt man die Mobilität und damit die Leistungswerte von Plakatstellen. Daher ist ein zentraler Punkt dieser Arbeit eine Methodik anzuwenden und zu adaptieren, die mit dieser zeitlichen Unvollständigkeit umgehen kann. Im Anschluss wird in Abschnitt 5.1.1 eine der vorgeschlagenen Varianten der Modellierung auf den Anwendungskontext adaptiert. In Abschnitt 5.1.3 wird dann beispielhaft für die Stadt Hamburg die Berechnung der Kontaktklassen beschrieben.

Im Anschluss wird in Abschnitt 5.2 in drei Schritten ein Weg zur Lösung der räumlichen Unvollständigkeit vorgestellt. Hierzu wird in Abschnitt 5.2.2 ein Aggregationssystem eingeführt, das den Raum in funktionale Räume einteilt und GPS-Trajektorien aggregiert. Im anschließenden Abschnitt 5.2.4 wird auf Basis des Aggregationssystems und des Frequenzatlas die Disaggregation von Trajektorien vorgestellt, die dafür zuständig ist, dass jedes Plakat eine positive Passagewahrscheinlichkeit erhält. Abschließend wird die Modellierung der mikrographischen Mobilität zur Lösung der geringen räumlichen Variabilität, die zu einer unvollständigen Abdeckung führt, am Beispiel der Stadt Köln demonstriert (Abschnitt 5.2.5). Der Abschnitt 5.2.7 befasst sich im Anschluss mit dem Zusammenspiel der räumlichen und zeitlichen Unvollständigkeit. Dabei wird einerseits die Variante der Berechnung von Plakatkontakten auf Basis von Kontaktwahrscheinlichkeiten und andererseits auf Basis von Kontaktdosen vorgestellt (vgl. Abschnitt 2.1). Mit einem Vergleich der Berechnungsmethoden über die aggregationsbasierte und rein GPS-basierte Methode befasst sich der Abschnitt 5.2.9. Abschnitt 5.3 befasst sich dann mit der

Berechnung von Durchschnittsnetzen, die insbesondere im Bereich der deutschen Mediaplanung eine wichtige Rolle spielen. Abgeschlossen wird das Kapitel mit einer Zusammenfassung der wichtigsten Erkenntnisse aus diesem Kapitel.

## 5.1 Umgang mit zeitlicher Unvollständigkeit in Daten

Der angestrebte Erhebungszeitraum der GPS-Daten erstreckte sich in der Schweiz über 7-10 Tage und in Deutschland über 7 Tage. Die Messreihen der Probanden sind aus verschiedenen Gründen unvollständig (vgl. Abschnitt 4.1). Über den Messzeitraum nimmt die Anzahl der teilnehmenden Probanden beständig ab (vgl. Hecker et al. 2010b). Dies liegt zum Einen an der Ermüdung von Probanden, an der Studie teilzunehmen, zum Anderen aber auch an technischen Problemen (Gerät defekt). Dies bedeutet, dass Tage, an denen Mobilität stattgefunden hat, zum Teil nicht aufgezeichnet wurden. Es gibt nun vier Möglichkeiten mit diesem unvollständigen Datensatz umzugehen:

**1.** Alle Personen mit fehlenden Tagen werden entfernt.

**2.** Unvollständigkeit wird ignoriert.

**3.** Fehlende Daten werden ergänzt und aufgefüllt.

**4.** Fehlende Daten werden durch eine geeignete Methodik behandelt.

Die Möglichkeit 1 kommt nur dann in Betracht, wenn die Stichprobe weiterhin eine ausreichende Größe zur Modellierung der Leistungswerte darstellt. Diese Möglichkeit scheidet allerdings aus, da die Anzahl der Personen mit fehlenden Tagen signifikant ist. Es bliebe nur die Option einer empirischen Nacherhebung. Die ist aufgrund der immens steigenden Kosten für die Außenwerbung nicht realisierbar. Möglichkeit 2 kann auch ausgeschlossen werden. Wertet man die fehlenden Informationen als nicht vorhandene Mobilität, würde aufgrund der signifikanten Menge unvollständiger Tage die Mobilität erheblich unterschätzt werden. Möglichkeit 3 ist ein Vorgehen, das nicht ohne Konventionen und weitere Vorgaben auskommt. Aus diesem Grund muss mit Möglichkeit 4 ein alternativer Weg gefunden werden, der die Unvollständigkeit explizit mit modelliert. Ein Problem der Ausgangsdaten ist, dass die Probandenzahl über den Zeitverlauf variiert. Daher kann die Reichweite nicht einfach als das Verhältnis der Summe aller Plakaterstkontakte bis zu einem bestimmten Tag in Bezug auf die Gesamtanzahl der Probanden berechnet werden, sondern muss an die variierende Stichprobengröße angepasst werden. Ein Verfahren das sich mit dieser Problematik auseinandersetzen, ist die sogenannte Ereignisanalyse (Aalen & Borgan & Gjessing 2008; Kaplan & Meier 1958, Little & Rubin 1987). Generell finden Verfahren der Ereignisanalyse einen breiten Einsatz, beispielsweise werden sie in den Anwendungsfeldern der Medizin und der Qualitätsanalyse für die Berechnung von Überlebenszeiten von Patienten mit bestimmten Krankheiten oder Ausfallwahrscheinlichkeiten von technischen Geräten benutzt (Kleinbaum & Klein 2005). Im Gegensatz zur Überlebenszeit in den obigen Beispielen interessiert jedoch im Anwendungskontext der Außenwerbung nicht, für welchen Anteil der Stichprobe kein Ereignis eintritt, sondern gerade wie viele Personen innerhalb des Betrachtungszeitraumes mindestens ein Ereignis (Plakatkontakt) aufweisen. Da dies komplementäre Ereignisse im Wahrscheinlichkeitsraum sind, kann die Reichweite als

$$\begin{array}{ccc} P(Plakatkontakt) & = 1 - & P(kein\ Plakatkontakt) \\ \hat{=} & & \hat{=} \\ Reichweite & = 1 - & Überlebenswahrscheinlichkeit \end{array}$$

berechnet werden. Bei der Reichweitenberechnung wird das Ereignis durch den Kontakt einer Person mit einem Plakat bzw. einem Plakat einer Kampagne definiert. Zur Erinnerung: Die Reichweite ist der prozentuale Anteil einer Zielgruppe, der mindestens ein Plakat einer Kampagne gesehen hat. Nimmt man die oben genannte Umformung, können Methoden der Ereignisanalyse auf den Anwendungskontext übertragen werden.

Im Detail wird nun im nächsten Abschnitt das Vorgehen bei der Ereignisanalyse im Anwendungskontext vorgestellt. Dabei wird die Kaplan-Meier Methode, die zu den bekanntesten Verfahren der Ereignisanalyse zählt, vorgestellt und später für Deutschland und die Schweiz angewendet.

### 5.1.1 Anwendung der Ereignisanalyse auf die Außenwerbung

Die Ereignisanalyse und speziell das Kaplan-Meier Verfahren erlaubt es, Personen mit fehlenden Daten (sogenannte „zensierte" Personen) in der Analyse zu verwenden, ohne die fehlenden Daten explizit zu ergänzen. Die Grundidee besteht darin, als Stichprobe für einen Tag nur die Menge der Personen zu verwenden, für die bis zu diesem Tag Daten vorliegen (d.h. vorhandene Trajektorien oder die Person hat angegeben, zuhause gewesen zu sein). Personen, für die an diesem Tag keine Informationen existieren (Gerät defekt, vergessen, …), zählen nicht länger zu der Stichprobe. Die Wahrscheinlichkeit eines Plakatkontaktes über sieben Tage wird dann als Gegenwahrscheinlichkeit zum Ergebnis einer sukzessiven Berechnung aus den Wahrscheinlichkeiten für „keinen Kontakt" über die einzelnen Tage berechnet. Anstatt die Reichweite einfach als das Verhältnis der Summe aller Plakaterstkontakte bis zu einem bestimmten Tag in Bezug auf die Gesamtanzahl der Probanden zu berechnen, passt das Kaplan-Meier Verfahren die Berechnung an die variierende Stichprobengröße an (Hecker et al. 2010b). Das Kaplan-Meier Verfahren wird formal wie folgt ausgedrückt:

Die Variable $S(t)$ bezeichnet die Wahrscheinlichkeit, mindestens bis zu Zeitpunkt $t$ kein kritisches Ereignis (Erstkontakt) zu generieren. Ein Zeitpunkt $t_i$ wird durch das Eintreten des kritischen Ereignisses (Plakatkontakt) für eine oder mehrere Personen der verbleibenden Stichprobe definiert. Dadurch wird der Zeitraum in mehrere Intervalle (hier per Konvention: ein oder mehrere Tage) zerlegt. Für jedes Intervall wird die bedingte Wahrscheinlichkeit für „keinen Kontakt" berechnet und nach dem Multiplikationssatz zur Gesamtwahrscheinlichkeit zusammengefügt. Betrachtet man nun $n$ Intervalle bis zum Zeitpunkt $t_n$, so gilt:

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1) \cdot P(A_2 \mid A_1) \cdot \ldots \cdot P(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1})$$

Das Ereignis $A_1$ entspricht bei Kaplan-Meier dem Überleben (kein Plakatkontakt) des ersten Zeitintervalls vom Zeitpunkt $t_0$ bis $t_1$. Die Zufallsvariable $T$ stellt die Zeit bis zum ersten Plakatkontakt dar.

$$P(A_1) \hat{=} P(T > t_1)$$

Die bedingte Wahrscheinlichkeit $P(A_2 \mid A_1) \triangleq P(T > t_2 \mid T > t_1)$ wiederum ist die Wahrscheinlichkeit im zweiten Zeitintervall (zwischen den Zeitpunkten $t_1$ und $t_2$) keinen Plakatkontakt zu haben, falls im ersten Zeitintervall ebenfalls kein Kontakt stattgefunden hat.

Die bedingte Wahrscheinlichkeit $P(A_n \mid \bigcap_{i=1}^{n-1} A_i)$ ist letztlich die Wahrscheinlichkeit, im $n$-ten Intervall (zwischen den Zeitpunkten $t_{n-1}$ und $t_n$) keinen Plakatkontakt zu haben, unter der Voraussetzung, dass in allen vorangegangenen Zeitintervallen bereits kein Kontakt stattgefunden hat.

Die Wahrscheinlichkeit über $n$ Zeitintervalle keinen einzigen Plakatkontakt zu erhalten $P(A_1 \cap A_2 \cap ... \cap A_n)$ ergibt sich aus der Multiplikation der bedingten Wahrscheinlichkeiten:

$$P(T > t_n) = S(t_n) = \prod_{i=1}^{n} P(T > t_i \mid T > t_{i-1}) \quad \text{mit} \quad P(T > t_0) = 1 .$$

Die bedingte Wahrscheinlichkeit „kein Plakatkontakt" eines Intervalls $i$ ergibt sich als Verhältnis der Anzahl Personen, die zum Intervallende $t_i$ noch nie einen Plakatkontakt hatten, zu der Anzahl der Personen $r(t_{i-1})$, die am Ende des vorherigen Intervalls noch keinen Kontakt hatten. Dabei bezeichnet $d(t_i)$ die Anzahl der Personen, die innerhalb des $i$-ten Intervalls (d.h. zwischen den Zeitpunkten $t_{i-1}$ und $t_i$) zum ersten Mal mindestens einen Plakatkontakt verzeichnen. Die Anzahl der Personen ohne Kontakt zum Zeitpunkt $t_i$ berechnet sich als Differenz der Personen ohne Plakatkontakte des vorangegangenen Intervalls abzüglich der Personen mit einem Erstkontakt im $i$-ten Intervall sowie den Personen $c(t_i)$, die nun aus der Stichprobe aussteigen:

$$r(t_i) = r(t_{i-1}) - d(t_i) - c(t_i) .$$

Die ausscheidenden Personen $c(t_i)$ werden jeweils zum Ende eines Intervalls $i$ zensiert. Kaplan-Meier nimmt also an, dass alle Personen die innerhalb eines Intervalls aus der Studie ausscheiden, bis zum Ende des jeweiligen Intervalls überleben. Durch die Zensur erfolgt die schrittweise Anpassung der Daten an die Stichprobengröße. Die Überlebenswahrscheinlichkeit $p_i$ im Zeitintervall $i$ ergibt sich damit als:

$$p_i = \frac{r(t_{i-1}) - d(t_i)}{r(t_{i-1})}$$

Die Wahrscheinlichkeit länger als $T$ Tage keinen Kontakt mit einer Kampagne zu erzeugen (also das Komplement der Reichweite) ergibt sich dann als

$$S(t_n) = P(T > t_n) = \prod_{i=1}^{n} p_i = \prod_{i=1}^{n} \frac{r(t_{i-1}) - d(t_i)}{r(t_{i-1})} .$$

Die Abbildung 5.1 stellt ein Beispiel zur Reichweitenberechnung dar. Sie zeigt beispielhaft für 5 Probanden Mobilitätsdaten und Kontakte mit einer gegebenen Kampagne. Im unteren Teil der Abbildung sind die Ergebnisse nach Anwendung der Zensur dargestellt. Die vorletzte Zeile enthält dabei die Personen „at Risk", also die Anzahl an Probanden, die bis zum jeweiligen Tag Mobilitätsdaten besitzen und noch nicht durch einen Plakatkontakt zensiert wurden. Die letzte Zeile enthält die Anzahl an Probanden, die an dem jeweiligen Tag ihren Erstkontakt mit der Kampagne haben. Da Messtage nicht nur am Ende, sondern auch in der Mitte des Untersuchungszeitraumes können, müssen die Daten vor Anwendung von Kaplan-Meier permutiert werden. Eine Permutation ist generell möglich, da eine Ausweisung der Reichweite stets nach dem 7 Tagen erfolgt und nicht nach speziellen Wochentagen. Konkret bedeutet dies, dass bei einer unvollständigen Messreihe, durch die Permutation alle fehlenden Tage ans Ende der Messreihe verschoben werden.

Das Kaplan-Meier Verfahren berechnet jetzt für jeden Tag eine bedingte Wahrscheinlichkeit, die angibt, welcher Anteil der Probanden „at Risk" ohne einen Kontakt bleibt. Multipliziert man diese Wahrscheinlichkeiten über alle Tage, so ergibt sich die Wahrscheinlichkeit, dass Probanden innerhalb von 7 Tagen kein Plakat der Kampagne sehen. Die Gegenwahrscheinlichkeit dieses Ereignisses ist die Reichweite. Durch die Zerlegung in bedingte Wahrscheinlichkeiten passt Kaplan-Meier die Berechnung an die abnehmende Stichprobengröße an.

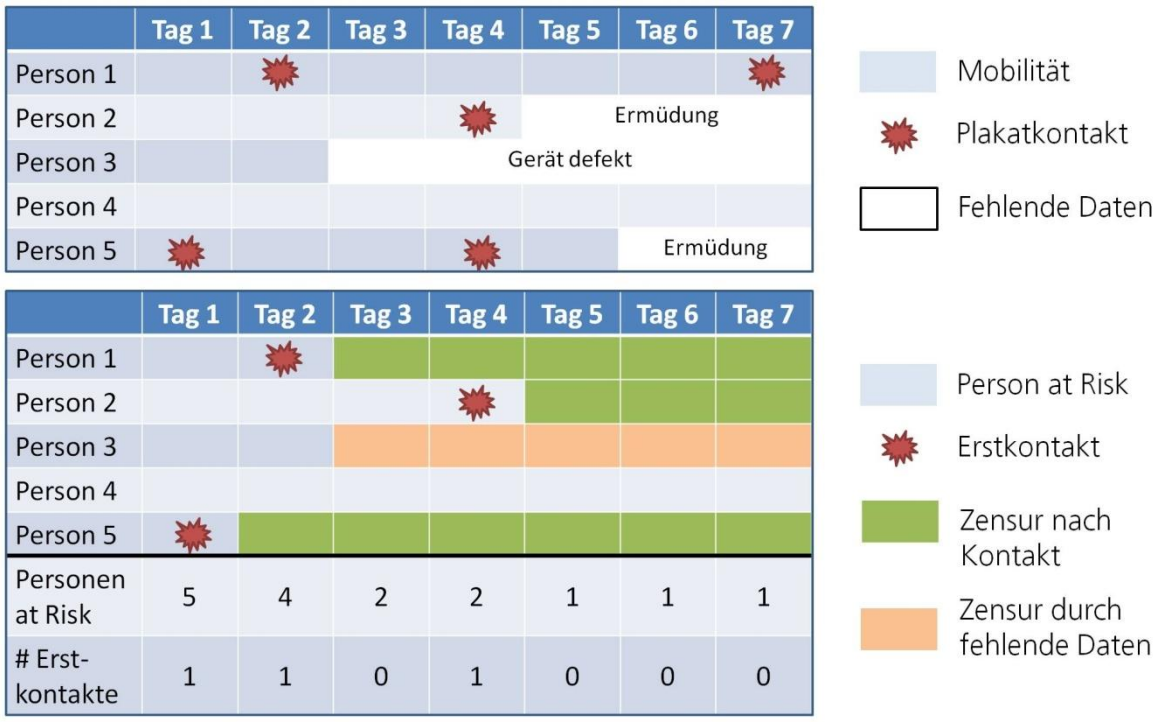



Abbildung 5.1: Zensurtechnik Kaplan-Meier (nach Hecker et al. 2010b)

Vergleicht man das Vorgehen der Ereignisanalyse mit den bisherigen Verfahren (CATI), so fällt als erstes die deutlich höhere Komplexität auf (vgl. Abschnitt 2.3). Dies liegt zum einen an der längeren zeitlichen Auflösung der Daten, zum anderen aber auch an der Kompensation von fehlenden Messdaten. Die Abbildung 5.2 stellt noch einmal die Zusammenhänge zwischen der Ereignisanalyse, der Anwendung in der Medizin/Technik und der Reichweitenberechnung in der Außenwerbung dar.

| Kaplan Meier | Technik/Medizin | Außenwerbung |
|---|---|---|
| • Objekte | • technisches Gerät/Patient | • GPS Probanden |
| • kritisches Ereignis | • Defekt/Tod | • Erstkontakt mit Plakat |
| • Zensur | • aus der Studie ausgeschlossene Geräte/Patienten | • Personen mit ungültigen Messtagen |

| Kaplan Meier | Formel | Interpretation Anwendung |
|---|---|---|
| • Zufallsvariable T | $T \in (0,\infty)$ | • Zeit bis zum ersten Plakatkontakt |
| • kum. Verteilung F(t) | $F(t)=P(T \leq t)$ | • WS einer Person, bis zu Zeitpunkt t ein bestimmtes Plakat zu sehen |
| • Survivor Funktion S(t) | $S(t)=P(T>t)=1-F(t)$ | • WS einer Person, Plakat erst nach t zu sehen |

Abbildung 5.2: Zusammenhänge Kaplan-Meier, Technik/Medizin und Außenwerbung

### 5.1.2 Validierung der Reichweitenberechnung mit Kaplan-Meier

In diesem Abschnitt wird überprüft, wie valide die Reichweitenergebnisse des beschriebenen Modellierungsansatzes mit Kaplan-Meier sind. Um eine Validierung durchführen zu können, werden nur Probanden ausgewählt, die eine Anzahl von 7 validen Messtagen aufweisen. Für diese Personengruppe kann ohne den Einsatz von Survival Techniken oder sonstigen Modellierungsschritten die Reichweite aus den Passagen an Plakatstellen ausgezählt werden. Diese Werte dienen als Bezugspunkt für die folgenden Experimente.

Als Testregionen für die Validierung wurden die Agglomerationen Bern und Zürich ausgewählt, da hier eine große GPS-Probandenmenge vorliegt. Die Abbildung 5.3 zeigt die Verteilung der validen Tage in Bern und Zürich.

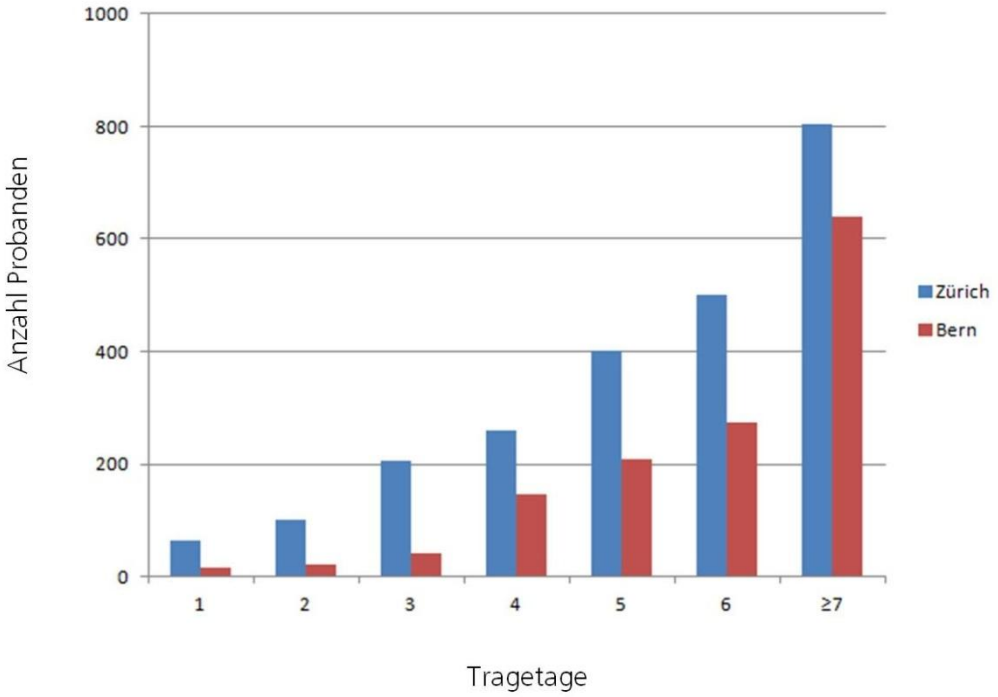



Abbildung 5.3: Anzahl valider Tage in Bern und Zürich

Diese ergibt für die Anzahl von 7 validen Messtagen eine Gesamtpersonenanzahl von 635 in Bern und 807 in Zürich. Für diese selektierten Personen wird die Reichweite nach 7 Tagen bestimmt. Als Testkampagne werden jeweils in Bern und in Zürich unterschiedlich große Kampagnen gewählt (20 und 50 Plakatstellen in Bern sowie 50 und 100 Plakatstellen in Zürich). Um die Validität der Kaplan-Meier Berechnung zu überprüfen, wird über das zufällige Löschen von Messtagen eine Stichprobe simuliert, wie sie in Deutschland und der Schweiz vorliegt. Die Ausfallrate der Messtage wird zum Vergleich mit der vollständigen Messreihe mit 7 Tagen sukzessive gesteigert. Die simulierte Fehlrate der Tage steigert sich von 0,1 auf 0,9. 0,1 bedeutet, dass bei 10% der Personen zwischen 1 und 6 Tagen zufällige eliminiert werden. Für die Agglomeration Zürich betrifft dies bei 0,1 beispielsweise 80 Personen. Die Rate von 0,0 entspricht dem vollständigen 7 Tage Datensatz. Die Ergebnisse der Reichweitenberechnungen stellen jeweils Durchschnittswerte von 100 zufällig gezogen Kampagnen der fixen Größe (20 und 50 In Bern, 50 und 100 in Zürich) dar. Für jede Kampagne wurden jeweils 10 Simulationen der Fehltage durchgeführt, um evtl. Zufälligkeiten in der Selektion zu vermeiden. Die Tabelle 5.1 enthält die Ergebnisse für Bern, die Tabelle 5.2 die Ergebnisse für Zürich.

| Fehltage | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Reichweite 20 Plakate** | 0,710 % | 0,723 % | 0,745 % | 0,716 % | 0,789 % | 0,737 % | 0,789 % | 0,719 % | 0,728 % | 0,774 % |
| **Reichweite 50 Plakate** | 0,874 % | 0,873 % | 0,862 % | 0,860 % | 0,854 % | 0,878 % | 0,811 % | 0,823 % | 0,804 % | 0,799 % |

Tabelle 5.1: Validierung Kaplan-Meier - Bern

| Fehltage | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Reichweite 50 Plakate** | 0,696 % | 0,645 % | 0,632 % | 0,699 % | 0,716 % | 0,726 % | 0,706 % | 0,666 % | 0,685 % | 0,675 % |
| **Reichweite 100 Plakate** | 0,832 % | 0,852 % | 0,867 % | 0,843 % | 0,829 % | 0,811 % | 0,862 % | 0,845 % | 0,872 % | 0,882 % |

Tabelle 5.2: Validierung Kaplan-Meier - Zürich

Die Ergebnisse der Experimente in Bern und in Zürich bestätigen die Anwendbarkeit der Kaplan-Meier Methode im Umgang mit zeitlich unvollständigen Daten. Die Unterschiede in den Reichweitenwerten sind minimal im Vergleich zur direkt ausgezählten Reichweite. Selbst bei einer sehr hohen Anzahl von Personen mit eliminierten Tagen sind die Reichweitenergebnisse noch eng an den Zahlen des vollständigen Datensatzes.

### 5.1.3 Berechnung von Kontaktklassen

Eine weitere Besonderheit in der Reichweitenberechnung stellen die sogenannten Kontaktklassen dar. Die Kontaktklasse spielt bei der Planung von Kampagnen eine sehr wichtige Rolle. Über sie kann ein Mediaplaner entscheiden, welche Kontaktsumme eine Person aufgebaut haben muss, um als erreicht zu gelten. Wenn der Proband diesen festgelegten Schwellwert des Kontaktes erzielt hat, wird erst ab diesem Zeitpunkt der Proband zur Reichweite und OTS dazu gezählt. Das bedeutet, ein Mediaplaner kann in Abhängigkeit des Plakatmotives bestimmen, wann es wahrgenommen wird. Bei Einführung eines neuen Produktes ist die Anzahl der Mindestkontakte in der Regel höher als bei einem bereits bekannten Produkt (vgl. Abschnitt 2.1). Vor diesem Hintergrund möchte der Mediaplaner gerne mit den Kontaktklassen variieren und so individuell die Leistungswerte an das Plakatmotiv anpassen können.


Abbildung 5.4:Werbemotive für unterschiedliche Plakatmotive (Ströer 2012)

In Abbildung 5.4 ist ein Beispiel für ein neues und ein bereits eingeführtes Produkt dargestellt. Beide Werbekampagnen kommen aus der Telekommunikationsbranche, besitzen jedoch eine stark unterschiedliche Markenbekanntheit. Die Farbe Magenta wird bereits bei einem ersten Blickkontakt mit dem Provider Deutsche Telekom in Verbindung gebracht, während die Firma Deutschlandsim eine solche Verbindung nicht aufweisen kann. Aus diesem Grund würden Mediaplaner die Kontaktklasse von der linken Plakatkampagne höher wählen.

Im Folgendem wird für Hamburg anhand einer Plakatkampagne die Kontaktklassenberechnung vorgestellt und ihr typisches Charakteristikum erläutert. Ausgewählt wurden 400 Plakatstellen für die Formatgattungen Citylightposter (CLP) und Großflächen (GF). Berechnet wurden jeweils die Kontaktklassen 1, 2 und 5. In der Abbildung 5.5 sind beide Plakatkampagnen (CLP und GF) jeweils dargestellt. Die X-Achse repräsentiert die Tage über die Zeit und die Y-Achse die erzielte Reichweite für die Stadt Hamburg. In jeweils 3 unterschiedlichen Farben sind die Kontaktklassen 1, 2 und 5 abgetragen. Wie zu erwarten, sinkt die Reichweite mit größer werdender Kontaktklasse. Bei der CLP Plakatkampagne erreichen wir eine Reichweite bei Kontaktklasse 1 von 84% und bei der Kontaktklasse 5 von 52%. Bei der Großflächenkampagne sind die Unterschiede noch gravierender. Während wir bei der Kontaktklasse 1 eine Reichweite von 73% erzielen, sinkt sie bei Kontaktklasse 5 auf gerade einmal 29%. Der hohe Einstieg bei der Kontaktklasse 1 ist durch die Größe der Kampagne zu erklären. Bei 400 verteilten Plakaten über die Stadt Hamburg ist es nur schwer möglich, innerhalb von 7 Tagen nicht an einem der betreffenden Plakate vorbeizukommen und einen Kontakt zu erzielen. Wie jedoch zu sehen ist, reduziert sich die Reichweite bei beiden Kampagnen, wenn man mehr als nur einen Kontakt erzielen möchte.

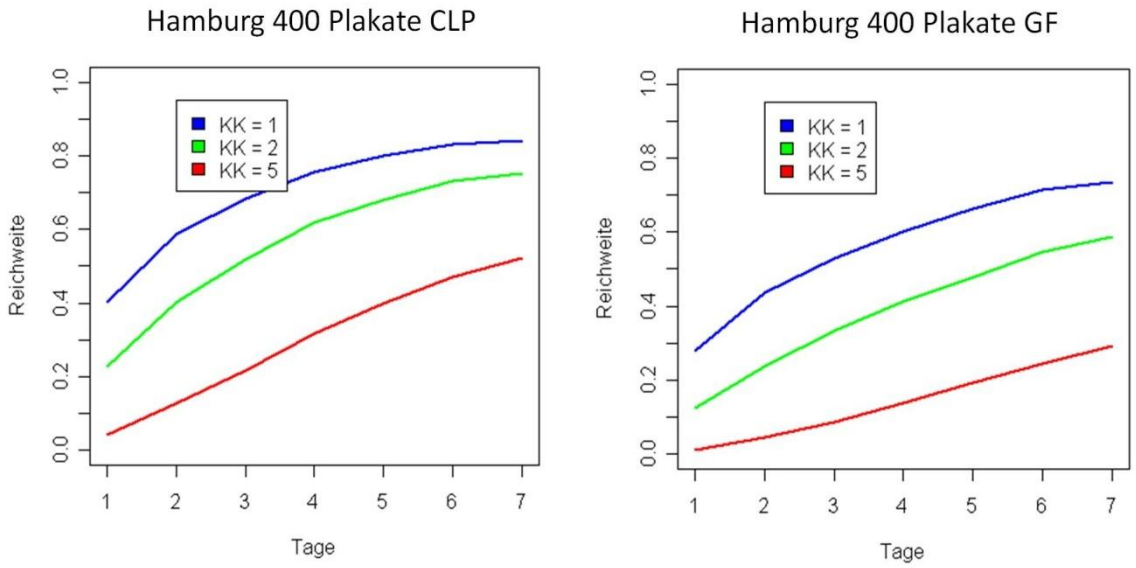



Abbildung 5.5:Kontaktklassenberechnung für Hamburg

In Tabelle 5.3 sind die Ergebnisse für die einzelnen Kontaktklassen und Kampagnentypen abgetragen. Es fällt auf, dass mit steigender Kontaktklasse wie beschrieben die Reichweite sinkt, jedoch der Opportunities to See (OTS) Wert steigt. Der OTS beschreibt die durchschnittliche Anzahl der Kontakte aller Personen in einem bestimmten Zeitraum (vgl. Abschnitt 2.1). Nimmt die Kontaktklasse zu, dann steigt auch der OTS Wert, denn es verbleiben dann nur noch die Probanden in der erreichten Menge von Personen, die mindestens die Kontakte der Kontaktklasse erzielt haben. Das heißt, der OTS Wert muss immer >= der Kontaktklasse sein.

| Hamburg | KK | Reichweite | OTS | GRP |
|---|---|---|---|---|
| Citylightposter | 1 | 84% | 8 | 698 |
| Citylightposter | 2 | 75% | 9 | 683 |
| Citylightposter | 5 | 52% | 12 | 588 |
| **Hamburg** | **KK** | **Reichweite** | **OTS** | **GRP** |
| Großfläche | 1 | 73% | 5 | 367 |
| Großfläche | 2 | 58% | 6 | 346 |
| Großfläche | 5 | 29% | 10 | 244 |

Tabelle 5.3: Ergebnisse zur Kontaktklassenberechnung für Hamburg (1) Citylightposter (2) Großfläche

Einen weiteren Effekt der Kontaktklassen kann man in Tabelle 5.3 bei der Berechnung des Gross Rating Points (GRP) erkennen. Der GRP wird auch gerne als Kontaktdruck bezeichnet und beschreibt die durchschnittliche Menge der Kontakte, die 100 Personen der anvisierten Zielgruppe produzieren (vgl. Abschnitt 2.1). Der GRP sinkt an dieser Stelle mit der kleiner werdenden Anzahl an Bruttokontakten. Dies ist wie beim OTS ein typisches Bild bei steigender Kontaktklasse.

Zusätzlich zu den beschriebenen Charakteristiken ist noch ein Unterschied zwischen den beiden Plakatarten zu erkennen. Obwohl diese die gleiche Stellenanzahl aufweisen, ist ein Unterschied in den Leistungswerten zu erkennen. Die CLP-Poster liegen bei der KK 1 mit knapp 11% über der Reichweite der Großflächen. Dieser Effekt lässt sich dadurch erklären, dass das CLP Format eine höhere Kontaktgewichtung besitzt als das Großflächen Format. Zudem sind die CLP Plakatstellen häufig in Fußgängerzonen aufgestellt, die im Vergleich zu den eher PKW-lastigen Straßenstellen der Großflächen auch eine höhere Kontaktchance besitzen.

Wie dieses Beispiel für Hamburg zeigt, sind die Kontaktklassen ein sehr probates Mittel, um die Werbewirkung von Kampagnen als Mediaplaner individuell steuern zu können. So kann z.B. über einen sogenannten Werbepretest mit einer Probandenstichprobe festgestellt werden, wie gut ein Plakatmotiv wirkt. Die Ergebnisse eines Pretests können im Anschluss für die Kontaktklassensteuerung eingesetzt werden.

## 5.2 Modellierung von Mikromobilität in Mobilitätsstudien

Im Unterschied zu soziodemographischen Variablen mit wenigen kategorischen Werten ist der geographische Wertebereich und der Raum aller möglichen Bewegungen durch ihn sehr groß. Selbst in diskretisierter Form kann der geographische Raum von ein paar hundert Regierungsbezirken, zu einigen tausenden Gemeinden, bis hin zu ein paar Millionen Straßenabschnitten reichen. Um die Mobilität und die Bewegungen auf der Ebene von Straßenabschnitten zu analysieren, müssten extrem große Datensätze erhoben werden. Genauer gesagt würde dies für Deutschland bedeuten, dass man eine repräsentative Gruppe von Testpersonen erfassen müsste, die mit ihren Bewegungen die rund 6,9 Millionen Straßenabschnitte im Erhebungszeitraum abdecken. Selbst der kombinierte deutsche Datensatz aus GPS und CATI mit insgesamt 42.780 Personen und deckt mit den gesammelten Trajektorien das deutsche NavTeq Straßennetz nur zu 26,7% ab. Eine Auswertung auf Straßenniveau erscheint auf den ersten Blick auf dieser Basis nicht möglich. In Abbildung 5.6 wird dieses Problem graphisch am Beispiel eines Berliner Wohnblocks dargestellt. Auf der linken Seite der Abbildung erkennt man die Wohnadresse des GPS-Probanden und insgesamt 4 Plakatstellen. Die linke Seite der Abbildung visualisiert die zurückgelegten Wege des Probanden nach 7 Tagen (gelb markiert). In der Summe hat der Proband 4 von 9 Straßen seines *„Quartiers"* befahren und hatte dabei die potentielle Möglichkeit 2 von 4 Plakaten zu sehen.


Abbildung 5.6: (1) Wohnadresse eines Probanden und Plakatstandorte (2) Wege des GPS-Probanden (topographische Hintergrundinformation Google 2012)

Wird auf Basis dieser Daten die Reichweite berechnet, würde für zwei der vier Plakate keine Reichweite berechnet werden können. In Abbildung 5.7 demonstriert dies noch mal anhand zweiter Plakate aus dem beschriebenen Wohnblock. Das Plakat auf der linken Seite erzielt eine Reichweite von 0%, dass auf der rechten Seite von 8%. Durch externe Datenquellen weiß man an dieser Stelle jedoch, dass die Verkehrsmenge (Frequenzatlas) und die Bevölkerungszahl in beiden Wohnstraßen identisch sind.


Abbildung 5.7: Auswertungen auf Straßensegmentniveau (topographische Hintergrundinformation Google 2012)

Damit eine Auswertung für alle Plakatstellen möglich ist, muss eine geeignete Methodik entwickelt werden. Eine Mobilitätsstudie auf Basis von GPS, die eine vollständige Abdeckung auf Straßensegmentniveau darstellt, ist aus Kostengründen nicht realisierbar und aus diesem Grund keine Option.

Bisher haben Forscher und Praktiker aus dem Bereich der Mobilitätsforschung die Bewegungen der menschlichen Mobilität entweder ohne konkrete räumliche Bezüge analysiert, oder die Mobilitätsdaten wurden auf einer sehr groben räumlichen Ebene ausgewertet. Zum Beispiel gibt die Studie Mobilität in Deutschland (BMVBS 2010) Variablen über die durchschnittliche Anzahl der Fahrten oder gefahrene Kilometer pro Tag und dem gewählten Transportmittel an. Diese Aussagen beziehen sich jedoch nicht auf das Straßenniveau, sondern werden auf Kreisebene ausgegeben. Ein zweites Beispiel ist die bundesdeutsche Arbeitswegematrix von DDS (2011), die auf kommunaler Ebene Statistiken auswertet. Das heißt, für die Frequenz auf einem bestimmten Straßensegment kann bei beiden Studien kein Wert bestimmt werden.

Für den vorgestellten Anwendungskontext sind jedoch Auswertungen auf Straßenabschnittsniveau zwingend notwendig, da Plakate an Straßenabschnitten verortet sind und gerade der Unterschied zwischen einem Plakat in einer Nebenstraße und beispielsweise einer Hauptstraße große Bedeutung für die Preisgestaltung besitzt. Aus diesem Grund wird hier ein Ansatz entwickelt, der die räumliche Variabilität der Mobilität erhöht, damit Aussagen auf Straßenabschnitten zulässt und damit wichtige Mobilitätseigenschaften beibehält (Hecker et al. 2011b).

**Forschungsfrage**

Wie kann, gegeben eine Mobilitätsstichprobe, ein Modell gefunden werden, dass die räumliche Variabilität erhöht, um eine flächendeckende Abdeckung des gewählten Untersuchungsgebietes zu gewährleisten?

**Modellierungsübersicht**

Die folgende Modellierungsübersicht gibt einen Überblick über die kommenden Modellierungsschritte. Hierzu werden der Arbeitsablauf und alle notwendigen Komponenten der Modellierung vorgestellt.

Die Modellierung zielt darauf ab, eine realistische Darstellung der räumlichen Variabilität in einem Mobilitätsdatensatz zu erzeugen. Dabei konzentriert sich der erzeugte Datensatz rein auf die Mobilität, die auf Straßensegmenten stattfindet und auf deren räumliche Verteilung. Eigenschaften der ursprünglichen Trajektorien wie Geschwindigkeit, Konnektivität und Richtung werden nicht beibehalten.

Die grundlegende Idee der Modellierung zur Erhöhung der räumlichen Variabilität von Trajektorien ist die Separierung von Makro- und Mikromobilität. Während die Bewegung auf makroskopischer Ebene beibehalten wird, wird die Bewegung auf mikroskopischer Ebene räumlich gestreut. In einem ersten Schritt werden alle Trajektorien auf eine gröbere Ebene der räumlichen Granularität abgebildet und anschließend auf Basis externer Frequenzinformationen auf das Straßennetz disaggregiert. Auf diese Weise wird die grundlegende Form der Bewegung erhalten, aber auf mikroskopischer Ebene durch die Einbindung von Hintergrundwissen stärker variiert. Das „Streuen" der Mobilität zur Erhöhung der räumlichen Variabilität wird durch wiederholte Simulation erreicht. Jede simulierte Welt enthält im Ergebnis wieder Werte auf Straßenabschnittsniveau. Vorteil dieses Vorgehens ist, dass nach der Kombination aller simulierten Welten eine Mobilitätsverteilung vorliegt, die für alle Straßensegmente im Simulationsraum eine positive Wahrscheinlichkeit des Besuchs beinhaltet.

Die Abbildung 5.8 zeigt die Komponenten und den Arbeitsablauf des Ansatzes. Die erste Komponente stellt die Trajektorien der Stichprobe auf Straßenabschnittsebene dar. Zweitens wird ein räumliches System gebraucht, welches zur Aggregation der Trajektorien genutzt werden kann. Diese Komponente fasst eine Menge von Straßensegmenten und den damit zugeordneten GPS-Trajektorien zu einer höheren Ebene zusammen. Die letzte Komponente des Ansatzes ist eine Verkehrsverteilung, die angibt, wie viele Menschen auf einem Straßensegment innerhalb einer bestimmten Zeitspanne unterwegs sind.

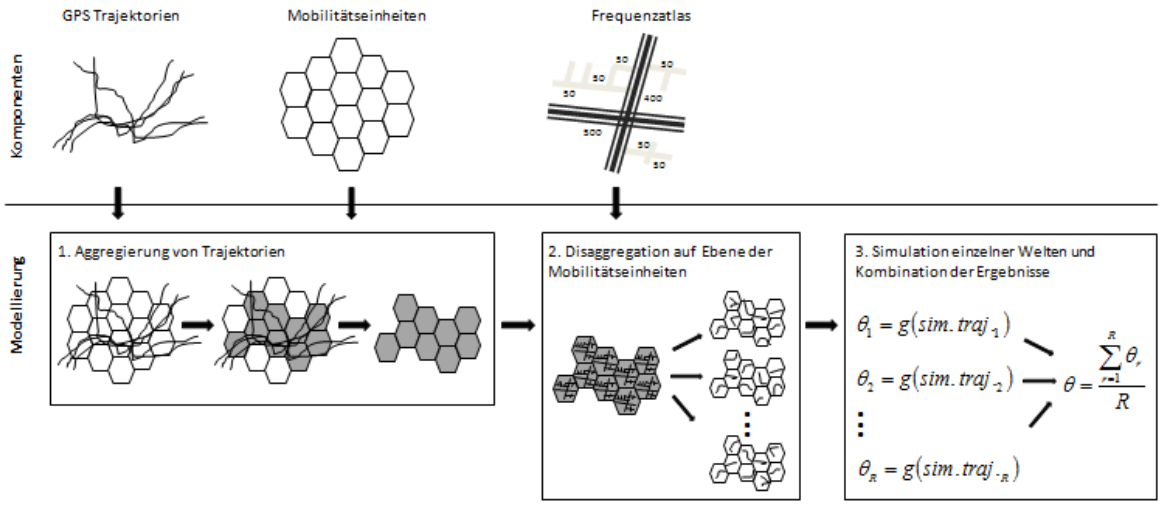



Abbildung 5.8: Modellierungsübersicht Mikromobilität (Hecker et al. 2011b)

Die Modellierung nutzt die verschiedenen Komponenten wie folgt: Im ersten Schritt (1) wird die räumliche Granularität der Trajektorien vergröbert, indem die Sequenzen der Straßensegmente in Sequenzen von räumlichen Einheiten überführt und somit vergröbert werden. Zu diesem Zweck werden die Trajektorien mit den räumlichen Einheiten verschnitten. Im zweiten Schritt (2) werden die daraus resultierenden Trajektorien in einer wiederholten Simulation disaggregiert. In jeder Simulation wird die Trajektorieninformation auf Ebene der Mobilitätseinheiten durch eine Menge von Straßensegmenten gemäß der Verkehrsverteilung innerhalb der Mobilitätseinheit ersetzt. Jede einzelne Simulation ist für

sich erst einmal so gering in ihrer räumlichen Abdeckung, wie die ursprünglichen Trajektorien, doch ergibt sich in der Kombination der einzeln simulierten Welten (3) eine vollständige Verteilung der Mobilität auf allen Segmenten. Im Folgenden werden die einzelnen Modellierungsschritte näher vorgestellt.

### 5.2.1 Anforderungen und Rahmenbedingungen der Modellierung

In diesem Abschnitt wird der Schritt der Aggregation von Trajektorien unter Verwendung eines Mobilitätseinheitssystems beschrieben. Hierzu werden zunächst Bedingungen gestellt, die die Aggregation erfüllen soll. Diese Bedingungen haben direkten Einfluss auf die Form der Mobilitätseinheiten und den Algorithmus zur Erstellung des Aggregationssystems.

**Überlegungen zum Bau des Aggregationssystems**

Je nach Form und Größe des räumlichen Aggregationssystems bleiben verschiedene Charakteristika von Mobilitätsdaten erhalten. Wichtigstes Ziel ist es, die Mikromobilität zu erhöhen, allerdings unter der Berücksichtigung, dass räumliche Gebiete mit homogener Mobilität gebildet werden. Folgende Vorgaben sollen erfüllt werden:

- **a.** Erhaltung der Homogenität und der Vielfalt der zugrundeliegenden Mobilität,
- **b.** Erhaltung der Makro-Mobilität,
- **c.** die Größe der zu bestimmenden räumlichen Aggregationseinheiten müssen zu den Informationen passen, die man später ableiten will,
- **d.** Erhaltung des lokalen (Anwohner) und Pendelverkehrs,
- **e.** Beachtung von natürlichen und künstlichen Mobilitätsbarrieren,
- **f.** Sicherstellung, dass das Aggregationssystem keine räumlichen Lücken oder Überlappungen besitzt.

Die erste Überlegung (a) bezieht sich auf ein ähnliches Mobilitätsverhalten innerhalb einer Mobilitätseinheit. Zum Beispiel wird ein Einkaufsviertel in der Innenstadt durch eine große Menge von Fußgängerbewegungen charakterisiert, während die umgebenden Ringstraßen durch den Fahrzeugverkehr gekennzeichnet sind. Um diese Mobilitätsstrukturen zu erhalten, müssen die Aggregationseinheiten möglichst homogene Einheiten bilden, z.B. Universitätsviertel, Einkaufsviertel, Wohngebiete, etc.

Die Erhaltung der Makro-Mobilität (b) bedeutet, dass die aufgezeichneten GPS-Trajektorien ihre Charakteristiken auf der makroskopischen Ebene behalten. Das heißt, dass z.B. eine einpendelnde Person aus dem Süden einer Stadt nicht plötzlich in den Norden versetzt wird. Die gewählte Aggregationseinheit (c) sollte nicht zu klein sein, da man sonst Gefahr läuft eine zu geringe Messdichte an Personen zu besitzen. Allerdings dürfen auch keine überdimensionierten Aggregationseinheiten erstellt werden, da sonst zu viele individuelle Informationen der Mobilität verloren gehen. In der Regel kann die Aggregationseinheit kleiner gewählt werden, je höher die Verkehrslast in ihr ist. Denn an diesen Stellen sind auch viele aufgezeichnete Trajektorien zu finden. Zusätzlich zu den genannten Vorgaben sollen die Aggregationseinheiten den Transit- bzw. Pendlerverkehr und die lokalen Bewegungen in Wohnvierteln bewahren (d). Das heißt, eine Person, die sich z.B. auf einer Hauptstraße in Richtung der Innenstadt bewegt, soll nicht in die anliegenden Wohngebiete verteilt werden. Die Berücksichtigung von Barrieren (e) spiegelt die Tatsache wieder, dass die Mobilität komplett straßengebunden stattfindet und Sprünge über Flüsse oder sonstige Barrieren nicht möglich sind. Die letzte Überlegung (f) ist eine Modellierungsanforderung an die Erstellung

der Aggregationseinheiten und besagt, dass die Aggregation flächendeckend sein muss und alle Straßensegmente (NavTeq) beinhaltet.

In der Literatur wird eine Reihe von Methoden angeboten, die den Raum systematisch in kleinere Teilbereiche untergliedern (vgl. Abschnitt 3.1.5). So ist z.B. eine Voronoi Partitionierung ein sehr häufig verwendetes Verfahren, das auf Grundlage von fixen Punkten und einer Distanz-Funktion den Raum in Mosaike einteilt. Für Mobilitätsstudien würde das bedeuteten, dass man die Zentroide der Mobilität bilden muss, um ein Voronoi Mosaik anlegen zu können. Ein Nachteil dieses Verfahrens ist die Nichtberücksichtigung von künstlichen und natürlichen Barrieren. Das gleiche gilt für administrative Grenzen, wie Bundesländer, Kreise und Gemeinden. Beides führt zu einer Durchmischung von lokalen und überregionalen Besonderheiten. Das nun vorgestellte Aggregationssystem soll diese Mängel des Voronoi Systems und der Aggregation über administrative Grenzen überwinden.

### 5.2.2 Erstellung von räumlichen Aggregationseinheiten (Mobilitätseinheiten)

Die Aggregationseinheiten werden über eine räumliche Aufteilung, bzw. eine Aufspaltung des Raumes erzeugt. Dabei geschieht diese Aufteilung auf Basis des Straßennetzwerkes. Das Straßennetz hat allerdings Restriktionen für die Modellierung: So können Bewegungen, z.B. über Plätze und durch einen Park nur sehr schlecht oder gar nicht abgebildet werden (vgl. Abschnitt 3.5.1). Allerdings stellt das Straßennetzwerk, in diesem Falle das NavTeq Netz, eine weit verbreitete (länderübergreifend), häufig aktualisierte und geographisch sehr exakte Datenquelle dar. Die grundlegende Idee ist, Aggregationseinheiten über den überregionalen Verkehr aufzuteilen. Diese Straßen des überregionalen Verkehrs bilden sogenannte Grenzlinien und partitionieren zusammen mit den natürlichen Barrieren die geographischen Aggregationseinheiten. Das Straßennetz kanalisiert den Verkehr und erhält weitere wertvolle Informationen über die Mobilität. Aus diesem Grund bildet das Straßennetz eine gute Grundlage für die Unterteilung von Charakteristiken des Verkehrs.

Es sei $A = \{x_1, \dots, x_z\}$ ein geographischer Raum oder der Bereich, in dem das Mobilitätsverhalten untersucht wird, und $x_i$ sei eine Koordinate im euklidischen Raum $\mathbb{R}^2$. Die aufgezeichnete GPS-Mobilität in diesem Raum ist verknüpft mit dem Straßennetz. Das Straßennetz wird durch einen Graphen $G = (N, S)$ bestehend aus einer Menge an Knoten $N \subseteq A$ und einer Menge an Kanten $S \subseteq N \times N$. (Abbildung 5.9 (1)) repräsentiert.

Der Untersuchungsraum $A$ wird in Aggregationseinheiten $u$ anhand von Merkmalen des Straßennetzes aufgeteilt. Es wird zwischen zwei Aggregationseinheiten unterschieden: Grenzeinheiten $u^{(b)}$ bestehen aus Straßensegmenten, die in erster Linie Pendlerverkehr tragen, und geschlossene Mobilitätsbereiche $u^{(i)}$, die bestimmte Räume einer Stadt wie Wohngebiete, Universitätsgebiete und Gewerbegebiete abbilden. Das Ergebnis dieser Raumaufteilung ist ein Aggregationssystem $U$ mit $U = \{\{u^{(b)}\} \cup \{u^{(i)}\}\}$. Der Prozess der Erstellung des Aggregationssystems beinhaltet drei aufeinander folgende Arbeitsschritte (Abbildung 5.9):

1. Identifizierung der übergeordneten Straßensegmente (Grenzeinheiten)
2. Unterteilung des Untersuchungsraumes in geschlossene Bereiche
3. Konstruktion der Aggregationseinheiten

Im ersten Schritt werden zwei Klassen $C = (C_1, C_2)$ von Straßenabschnitten definiert, die als übergeordnete $C_1$ und untergeordnete $C_2$ Straßenklasse bezeichnet werden. Elemente der übergeordneten Klasse $G_{C_1} \subseteq G$ enthalten in erster Linie Pendel- und Transitverkehr, während Segmente der untergeordneten Klasse $G_{C_2} \subseteq G$ hauptsächlich lokale Mobilitätsmuster enthalten. Jedem Straßensegment im Untersuchungsraum wird anhand seiner Charakteristik eine dieser Klassen zugeordnet, welche auf Basis des Straßennetzwerkes gemäß der Funktion

$h: S \rightarrow C$ zuwiesen werden. Dieser Ansatz hat den Vorteil, dass die notwendigen Informationen flächendeckend in der gleichen Qualität vorliegen und verfügbar sind. Ein Beispiel für diese Klassifizierung ist in Abbildung 5.9 (2) zu sehen. Die Segmente der Klasse $C_1$ (doppelte Strichstärke in Abbildung 5.9 (2)) werden, auch Grenzeinheiten genannt, da sie in Schritt 2 zur räumlichen Trennung der geschlossenen Mobilitätsbereiche genutzt werden.

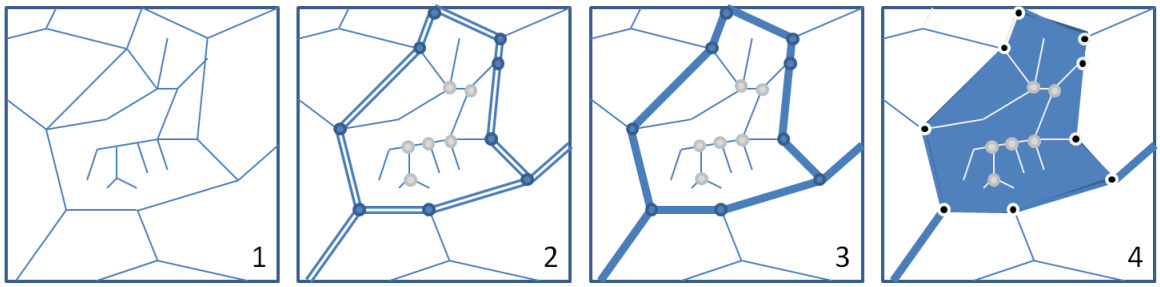



Abbildung 5.9: Prozess zur Erstellung der Aggregationseinheiten – 3 stufiger Prozess zur Konstruktion der Mobilitätseinheiten. **(1)** Straßennetzwerk, **(2)** Identifizierung der übergeordneten Straßeneinheiten, **(3)** Bildung von Grenzeinheiten und Unterteilung des Raumes in Grenz- und geschlossene Bereiche, **(4)** Konstruktion der Aggregationseinheiten (Hecker et al. 2011b)

Im zweiten Schritt werden eine Reihe von geschlossenen Wegen (Kreise), im übergeordneten Graphen $G_{C_1} = (N_{C_1}, S_{C_1})$ (siehe Abbildung 5.9 (3)) erstellt. Ein solcher Kreis, oder auch Grenzlinie $l$ genannt, ist definiert als eine Folge von Knoten und Segmenten, die als $C_1$ klassifiziert worden sind, formal ausgedrückt: $l = (n_1, n_2, \dots, n_k) | n_1 = n_k \text{und } n_i \neq n_j \text{ für } i \neq j \; sonst, \forall n_j, n_{j+1} \; \exists \; \dot{s} \; \in \; S_{c_1} \text{mit } \dot{s} = (n_j, n_{j+1})$. Überschneidungen oder selbst Kreuzungen sind nicht gestattet. Der gesamte Datensatz der Grenzlinien wird mit $L$ bezeichnet. Die Grenzlinien werden dazu verwendet, um das Straßennetz bei jedem Grenzknoten $n \in N_{C_1}$ zu trennen. Die so umschlossene Fläche eines Kreises bildet die Aggregationseinheit der inneren Mobilitätseinheiten $u^{(i)}$, während der Kreis selbst die Basis für die Grenzeinheiten $u^{(b)}$ bildet. Beide Aggregationseinheiten können entweder als Punktmenge im $\mathbb{R}^2$ oder als Menge von Straßensegmenten repräsentiert werden. Formal ausgedrückt, wird die Punktmenge einer Aggregationseinheit $u$ definiert als $A_u = \{ x \mid x \in A, h_A(u, x) = \text{true}\}$, wo die Funktion $h_A: (U, A) \rightarrow \{\text{true}, \text{false}\}$ testet, ob ein Punkt $x \in A$ zu einer Aggregationseinheit $u$ gehört oder nicht (siehe Abbildung 5.9 (4)). Die letzte Darstellung (Abbildung 5.9 (4)). beschreibt eine Aggregationseinheit $u$ als eine Menge von Straßenabschnitten und wird als $S_u = \{ s \mid s \in S, h_S(u, s) = \text{true}\}$ definiert und mit $h_S: (U, S) \rightarrow \{\text{true}, \text{false}\}$ getestet wird, ob zum Beispiel der Schwerpunkt einer Straße innerhalb $A_u$ liegt. Zusätzlich zu den oben genannten Grenzlinien werden noch die Geometrien von natürlichen Barrieren, wie z.B. Flüsse oder Eisenbahnlinien genutzt, um Grenzlinien zu bilden.

Das daraus resultierende Aggregationssystem ist die Menge aller Grenz- und inneren Einheiten im Untersuchungsraum, $\cup A_u = A, \; \cup S_u = S \mid u \in U$. Jedes Straßensegment gehört zu exakt einer Einheit.

Zur Bildung der Aggregationseinheiten wird das Straßennetzwerk von NavTeq verwendet und jedes Straßensegment der Klassen $C_1$ und $C_2$ wird gemäß seiner Funktionsklasse zugeordnet. Die Funktionsklasse unterscheidet Straßensegmente nach ihrer Geschwindigkeit, der möglichen Verkehrslast und der offiziellen administrativen Bezeichnung. Die Klasse $C_1$ wird der funktionalen Klasse {1, 2, 3} zugeordnet und $C_2$ den Funktionsklassen {4, 5}.

### 5.2.3 AGGREGATION VON TRAJEKTORIEN

Die erstellten Aggregationseinheiten dienen zur Zusammenfassung der gesammelten GPS-Trajektorien für diese Bereiche. Hierzu werden die Trajektorien von ihrer sequentiellen Abfolge der Straßensegmente in eine sequentielle Abfolge der Aggregationseinheiten umgewandelt. Eine Trajektorie mit ihrer Abfolge der Straßensegmente wird definiert als $\mathrm{t^S} = (\mathrm{s_1}, \mathrm{s_2}, \dots, \mathrm{s_k})$ mit $\mathrm{s_i} \in \mathrm{S}, \mathrm{i} = (1, 2, \dots, \mathrm{k})$ und $\mathrm{k}$ der Länge der Trajektorie. Ebenso wird die Abfolge einer Trajektorie auf der Ebene der Aggregationseinheiten definiert, $\mathrm{t^U} = (\mathrm{u_1}, \mathrm{u_2}, \dots, \mathrm{u_m})$ mit $\mathrm{u_i} \in \mathrm{U}, \mathrm{i} = (1, 2, \dots, \mathrm{m})$ und $\mathrm{m}$ der Länge der Trajektorie. Der komplette Datensatz wird mit $\mathrm{T^S}$ und $\mathrm{T^U}$ bezeichnet. Die Transformation besitzt somit folgende Formalisierung:

$$t^S = (s_1, \dots, s_k) \xrightarrow{\mathbf{GenMobility()}} t^U = (u_1, \dots, u_m), \;\; s_i \in S, u_i \in U, m \leq k.$$

Der Algorithmus 1 zeigt das Vorgehen bei der Transformation. Für jedes einzelne Segment wird nach der Aggregationseinheit gesucht, zu der es gehört. Im Ergebnis erhält man eine geordnete Liste von aufeinanderfolgenden Aggregationseinheiten, die von einer Person besucht worden sind. Jeder Kontakt mit einer Aggregationseinheit wird nur einmal gespeichert, außer, die Sequenz wird durch das Betreten einer weiteren Aggregationseinheit unterbrochen: $(\mathrm{u_1}, \mathrm{u_2}, \mathrm{u_2}, \mathrm{u_1}) \Longrightarrow (\mathrm{u_1}, \mathrm{u_2}, \mathrm{u_1})$ oder $(u_1, u_1, u_2, u_3) \Longrightarrow (u_1, u_2, u_3)$. Das Ergebnis dieses Algorithmus ist eine Menge an Trajektorien, die einer Sequenz von Aggregationseinheiten $\mathrm{T^U}$ zugeordnet worden sind. Die generalisierten Trajektorien enthalten jetzt nur noch Informationen über Bewegungen auf der Makroebene. Mit der Aggregation wird der geringen räumlichen Abdeckung der GPS-Stichprobe auf Mikroebene entgegengewirkt. Die Bewegungen der ursprünglichen GPS-Trajektorien innerhalb der Aggregationseinheiten werden bei diesem Vorgehen nicht berücksichtigt. Im folgenden Abschnitt 5.1.4. wird das System der Mobilitätseinheiten für die Stadt Köln vorgestellt. In Abschnitt 5.2.4 wird vorgestellt, wie auf Mikroebene die Modellierung weitergeführt wird.

Algorithmus 1: GenMobility – Generalisierung der Mobilitätsinformationen (Hecker et al. 2011b)

Input:
= aggregation unit system $U$,
= $S_{u_j}$ $\forall$ $u_j \in U$ are the sets of street segments of aggregation units $u_j$,
= street network S,
= trajectory sample data $T^S$on street network
Output:
= $T^U$, set of trajectories where each trajectory is a list of aggregation units $t_i^U = (u_1, \ldots, u_m)$
Method:

```
T^U = {}
t^U = ( )
previousUnit ← ∅
for each trajectory t^S ∈ T^S do
    for each street segment s_i of trajectory t^S = (s_1, s_2, .., s_k) do
        for each aggregation unit u_j ∈ U do
            # if trajectory segment inside aggregation unit
            if s_i ∈ S_uj then
                # checks whether the person is already
                #  inside the unit, re-entries allowed
                if previousUnit = ∅ or previousUnit ≠ u_j then
                    t^U = append(t^U, u_j)
                    previousUnit = u_j
                end if
            end if
        end loop
    end loop
    T^U = insert(T^U, t^U)
end loop
```

### 5.2.4 Disaggregation von Trajektorien

In diesem Abschnitt wird die Disaggregation von Mobilitätsinformationen vorgestellt. Eine wichtige Komponente, die bei dieser Modellierung herangezogen wird, ist der Frequenzatlas (vgl. Abschnitt 3.5.2). May et al. (2008a, 2008b) entwickelten für Deutschland eine Frequenzkarte, die eine durchschnittliche Anzahl von Personen auf einem Straßenabschnitt für PKW, Fußgänger und ÖPNV pro Stunde ausweist. Abbildung 5.10 zeigt einen Auszug des Frequenzatlas für die Stadt Köln. Die jeweiligen Zahlen geben Frequenzwerte pro durchschnittlicher Tagesstunde für PKW an.

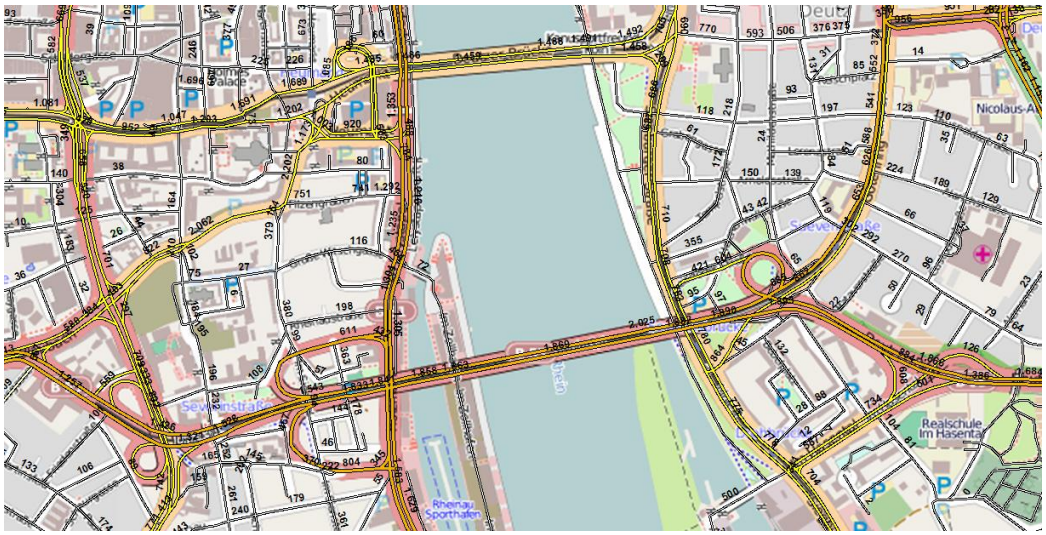


Abbildung 5.10: Frequenzatlas Köln (PKW-Frequenzen) (topographische Hintergrundinformation OSM 2012)

Für ein gegebenes Straßennetz $G$ mit Segmenten $S = \{s_1, s_2, \ldots, s_{|S|}\}$ bezeichnet der Frequenzatlas mit $F = \{f_1, f_2, \ldots, f_{|S|}\}$ mit $f_i$ die entsprechende Frequenz auf Segment $s_i$.

Im folgenden Disaggregationschritt werden die vorher generalisierten Trajektorien wieder zurück auf das Straßenniveau abgebildet. Dieser Prozess wird mehrfach wiederholt, so dass eine Reihe von Abbildungswelten auf Segmentniveau entstehen. Die Abbildungswelten jeder simulierten Welt werden jeweils separat ausgewertet und im Anschluss zusammengefasst. Die Disaggregation einer Trajektorie pro Aggregationseinheit wird gemäß der Verkehrsverteilung des Frequenzatlas vorgenommen. Aus diesem Grund ist der erste Schritt zur Disaggregation die Umwandlung des Frequenzatlas in eine Wahrscheinlichkeitsverteilung des Verkehrs nach Verkehrsart und Aggregationseinheit. Der Algorithmus für die Transformation ist in Alg. 2 dargestellt. Die Abbildung 5.11 veranschaulicht dieses Vorgehen. Im Wesentlichen gibt die resultierende Verteilung an, mit welcher Wahrscheinlichkeit eine Person, die die betreffende Aggregationseinheit betritt, an einem bestimmten Straßensegment anzutreffen ist.

---

Algorithmus 2. Transformation des Frequenzatlas in eine Verkehrsverteilung (Hecker et al. 2011b)

---

Input:

- = streetsystem with segments S = $\{s_1, s_2, \ldots . s_{|S|}\}$,
- = frequency map F = $\{f_1, f_2, \ldots, f_{|S|}\}$,
- = aggregation unit system U = $\{u_1, u_2, \ldots, u_{|u|}\}$, each aggregation unit consists of a set $S_{u_i} = \{s_{i_1}, s_{i_2}, \ldots, s_{i|S_{u_i}|}\} \subseteq S$ of street segments

Output:

- = set of traffic distributions $D = \{d_{u_1}, d_{u_2}, \ldots, d_{u_{|U|}}\}$ with a probability distribution $d_{u_i}$ for each aggregation unit $u_i$

1: for all $u_i \in U$ do
2: for all $s_{i_j} \in S_{u_i}$ do
3: calculate $pr(s_{i_j}) = f_{i_j} / \sum_{p=1}^{|S_{u_i}|} f_{i_p}$
4: end for
5: end for

---

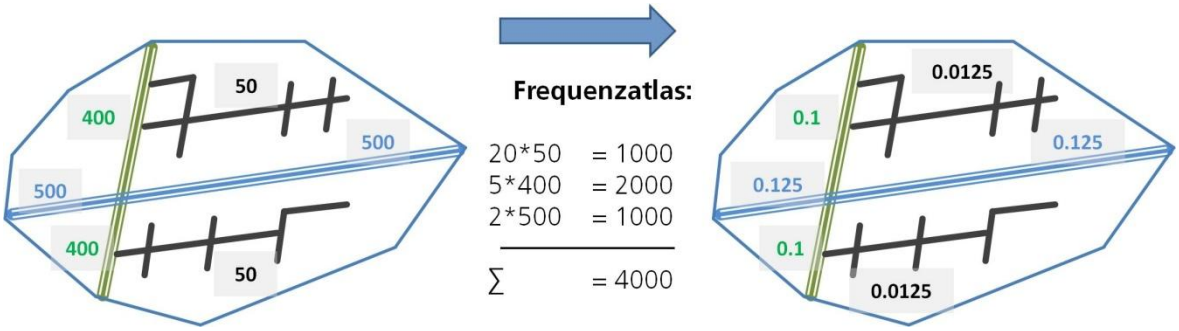



Abbildung 5.11: Transformation des Frequenzatlas in eine Verkehrsverteilung für eine Aggregationseinheit

Als nächstes muss bestimmt werden, wie viele Straßenabschnitte pro Simulationsdurchgang für eine Aggregationseinheit verwendet werden. Diese Information lässt sich aus den originalen GPS-Trajektorien der betreffenden Aggregationseinheit ableiten, indem die durchschnittliche Anzahl an berührten Straßensegmenten berechnet wird. Formal ausgedrückt wird für eine Aggregationseinheit $u_i$ mit den Straßensegmenten $\mathrm{S}_{\mathrm{u_i}} = \{\mathrm{s}_{\mathrm{i}_1}, \mathrm{s}_{\mathrm{i}_2}, \dots, \mathrm{s}_{\mathrm{i}|\mathrm{S}_{\mathrm{u_i}}|}\}$ , den Trajektorienmengen $T^s$ und $T^u$, $t_j^S = (s_1, s_2, \dots, s_k)$ und $t_j^u = (u_1, u_2, \dots, u_k)$ mit $j = 1..|T^S|$, die durchschnittliche Anzahl der passierten Straßenabschnitte pro Aggregationseinheit wie folgt berechnet:

$$n_{u_i} = \frac{\sum_{t_j^S \in T^S} \sum_{s_p \in t_j^S} I\left(s_p \in S_{u_i}\right)}{\sum_{t_j^U \in T^U} \sum_{u_q \in t_j^U} I\left(u_q = u_i\right)} .$$

In der obigen Gleichung bezeichnet $I(\cdot)$ eine Boolesche Funktion, die einen Wert von 1 besitzt, wenn ihr Argument wahr ist, oder 0 wenn ihr Argument falsch ist. An dieser Stelle wäre es auch möglich, statt dem Mittelwert eine Verteilung der passierten Straßensegmente pro Aggregationseinheit anzunehmen. Im Folgenden wird jedoch mit dem einfacheren Weg der Durchschnittsbildung gearbeitet.

Mit dem Wissen über die Verkehrsverteilung und der durchschnittlichen Anzahl der passierten Segmente für eine Aggregationseinheit ist es möglich, für eine Aggregationseinheit $u_i$ eine Disaggregation durchzuführen. Die Disaggregation aller Aggregationseinheiten resultiert in einer Menge von simulierten Trajektorien $\dot{\mathrm{T}}^{\mathrm{S}}$, die auf Basis von Straßenabschnitten ausgewertet werden kann. Die ausgewählte Anzahl der durchschnittlich berührten Straßensegmente pro Aggregationseinheit $\mathrm{n}_{\mathrm{u_i}}$ wird auch bei jeder der simulierten Trajektorien Welten eingehalten. Bei vielfacher Wiederholung der Simulation konvergiert nach dem Gesetz der großen Zahlen die Verteilung aller simulierten Straßensegmente zu der Verteilung des Verkehrs im Frequenzatlas. Algorithmus 3 zeigt den Prozess im Detail. Hierbei beschreibt die Funktion *multinominal* das zufällige Ziehen von einer polynominalen Verteilung. Die Verteilung wird im ersten Argument festgelegt und die Anzahl der Ziehungen im zweiten.

Algorithmus 3. Disaggregation der aggregierten Trajektorien durch widerholte Simulation (Hecker et al. 2011b)

Input:
- = aggregation-unit-trajectory sample $T^u$,
- = set of traffic distributions $\mathrm{D} = \{\mathrm{d}_{\mathrm{u}_1}, \mathrm{d}_{\mathrm{u}_2}, \dots, \mathrm{d}_{\mathrm{u}_{|\mathrm{U}|}}\}$,
- = set of the average number of passed street segments per mobility unit $= \{\mathrm{n}_{\mathrm{u}_1}, \mathrm{n}_{\mathrm{u}_2}, \dots, \mathrm{n}_{\mathrm{u}_{|\mathrm{U}|}}\}$,
- = number of simulations w

Output:
- = set of simulated trajectory worlds on street level $\mathrm{TW} = \{\dot{\mathrm{T}}_1^{\mathrm{S}}, \dot{\mathrm{T}}_2^{\mathrm{S}}, \dots, \dot{\mathrm{T}}_{\mathrm{w}}^{\mathrm{S}}\}$

1: TW = {} # initialize set of trajectory worlds

2: for r = 1..w do # perform w simulations
3: $\dot{\mathrm{T}}_{\mathrm{r}}^{\mathrm{S}} = \{\}$ # initialize single trajectory world

4: for all $t^u \in T^u$ do # for each aggregated trajectory
5: $\dot{\mathrm{t}}_{\mathrm{j}}^{\mathrm{S}} = ()$ # initialize simulated trajectory

6: for all $\mathrm{u}_{\mathrm{j_i}} \in \mathrm{t}_{\mathrm{j}}^{\mathrm{U}}$ do # for each aggregation unit per trajectory
# randomly draw $\mathrm{n}_{\mathrm{u}_{\mathrm{j_i}}}$ segments and append to trajectory

7: $\dot{\mathrm{t}}_{\mathrm{j}}^{\mathrm{S}} = \mathrm{append}\left(\dot{\mathrm{t}}_{\mathrm{j}}^{\mathrm{S}},\ \mathrm{multinomial}\left(\mathrm{d}_{\mathrm{u}_{\mathrm{j_i}}}, \mathrm{n}_{\mathrm{u}_{\mathrm{j_i}}}\right)\right)$

8: end for
9: $\dot{\mathrm{T}}_{\mathrm{r}}^{\mathrm{S}} = \mathrm{insert}(\dot{\mathrm{T}}_{\mathrm{r}}^{\mathrm{S}},\ \dot{\mathrm{t}}_{\mathrm{j}}^{\mathrm{S}})$ # insert trajectory to trajectory set

10: end for
11: $\mathrm{TW} = \mathrm{insert}(\mathrm{TW}, \dot{\mathrm{T}}_{\mathrm{r}}^{\mathrm{S}})$ # insert simulation to world set
12: end for

Für jede einzelne generierte Trajektorienwelt werden die Ergebnisse berechnet und im Anschluss über die Mittelung aller Trajektorienwelten kombiniert, das heißt:

$$\theta_r = g\left(\dot{T}_r^S\right) \qquad \forall r = 1...w,$$

$$\theta = \frac{\sum_{r=1}^{w} \theta_r}{w}.$$

Durch die beschriebene Simulation verlieren die Trajektorien ihre erfasste Konnektivität innerhalb der einzelnen Aggregationseinheiten. Die erfasste Sequenz der Aggregationseinheiten bleibt jedoch erhalten und damit auch die Makromobilität der aufgezeichneten GPS-Trajektorien.

### 5.2.5 Aggregationseinheiten am Beispiel von Köln

Geht man nach der beschriebenen Erstellung der Aggregationseinheiten vor, ergibt sich für Deutschland ein Gesamtdatensatz von 194.331 Einheiten. Im Folgenden werden am Beispiel von Köln Abbildungen und Statistiken zu den Aggregationseinheiten gezeigt. Köln umfasst insgesamt 2.256 Aggregationseinheiten.

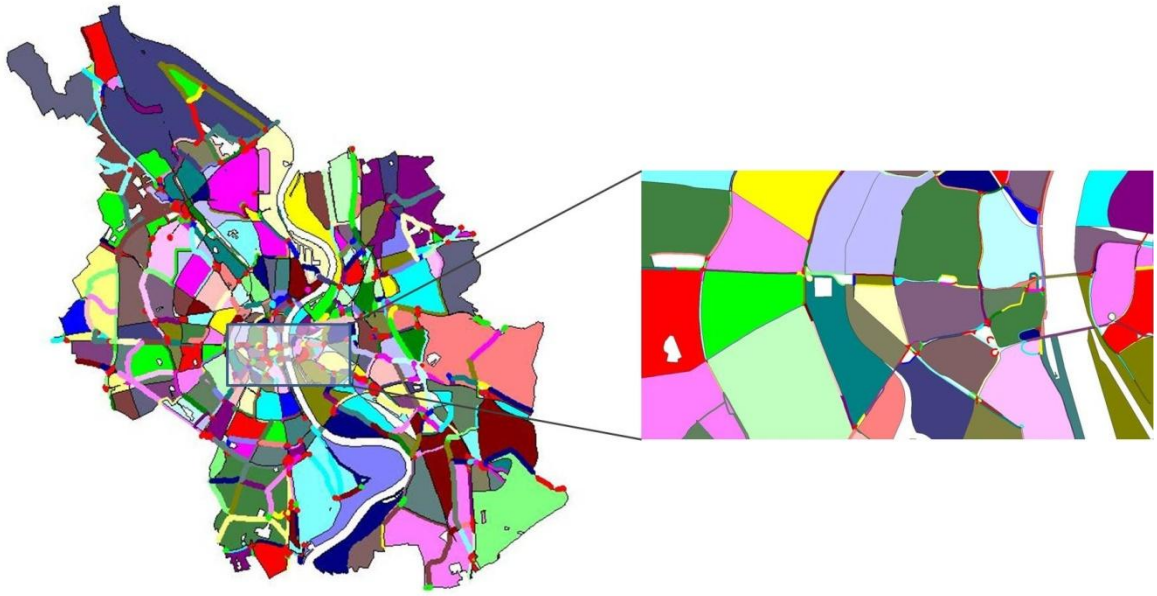


Abbildung 5.12: Aggregationseinheiten für die Gemeinde Köln und Innenstadt Köln (Hecker et al. 2011b)

Wie Tabelle 5.4 zeigt, sind die meisten Aggregationseinheiten in Köln Grenzeinheiten des Verkehrs. Sie stellen relativ kleine Einheiten dar, doch aufgrund des hohen Verkehrsaufkommens auf diesen Einheiten ist die Anzahl der zugeordneten Trajektorieninformationen in der späteren Disaggregation ausreichend. Zusätzlich ist in der Tabelle noch die durchschnittliche Anzahl der Straßensegmente pro Einheit und die durchschnittliche Anzahl der berührten Straßensegmente angegeben, welche für die Disaggregation und Erstellung der einzelnen Simulationswelten genutzt wird.

| | Anzahl der Aggregationseinheiten | Durchschnittliche Anzahl (und Median) Straßensegmente pro Einheit | Durchschnittliche Anzahl (und Median) der berührten Straßensegmente (*n*) pro Aggregationseinheit |
|---|---|---|---|
| Grenz-einheiten | 1.997 | 4,6 (1,0) | 2,4 (1,0) |
| Innere Einheiten | 259 | 137,9 (43,0) | 5,4 (4,8) |
| Summe | 2.256 | 19,9 (1,0) | 2,8 (1,1) |

Tabelle 5.4: Statistiken für die Aggregationseinheiten in Köln (Hecker et al. 2011b)

### 5.2.6 Zusammenfassung der Aggregation und Disaggregation im Anwendungskontext

Die aktuelle Anzahl der GPS und CATI Datenerhebung in Deutschland umfasst insgesamt 42.780 Probanden mit einer Erfassungsdauer bis zu 7 Tagen. Trotz dieser großen

Datenmenge stellt die räumliche Abdeckung eine Herausforderung dar. Nicht alle Plakatstandorte werden von den Probanden der Stichprobe besucht. Ohne diese Informationen ist eine Ausweisung von Leistungswerten jedoch nicht möglich. Abbildung 5.13 visualisiert dies exemplarisch. Wird ein Plakatstandort ohne GPS-Passagen ausgewertet, ist sein Leistungswert gleich Null (linke Seite). Dies ist jedoch nicht der wahre Leistungswert des Standortes, er ist zurückzuführen auf die fehlende Variabilität und damit räumlichen Abdeckung in der Stichprobe.

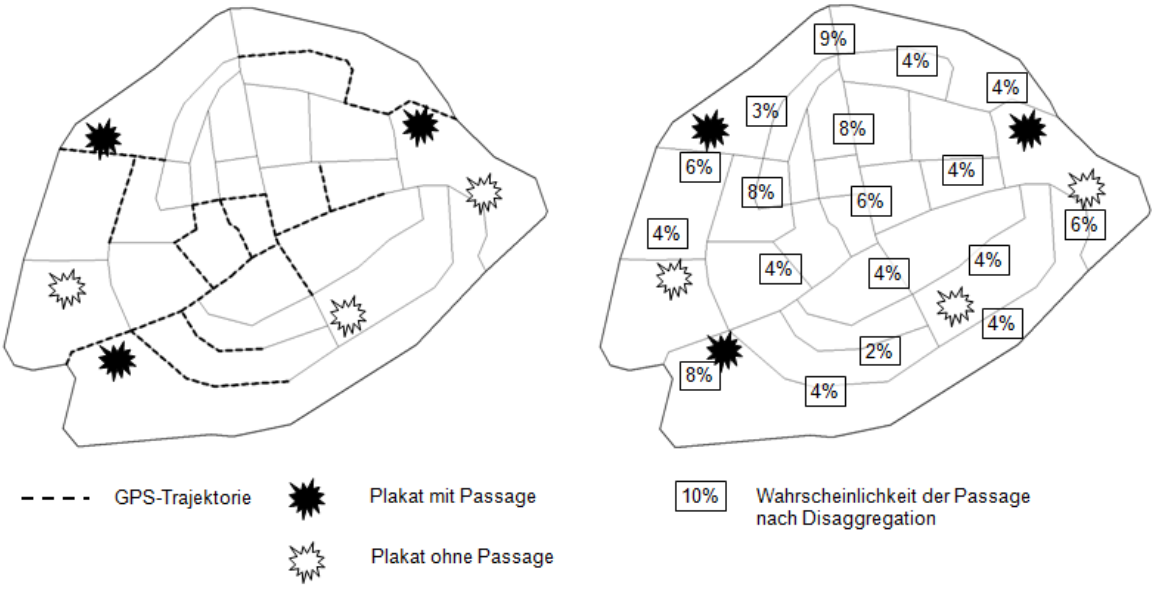



Abbildung 5.13: Beispiel für Plakatpassagen vor und nach der Modellierung (Hecker et al. 2011b)

Um valide und gerechte Leistungswerte für alle Plakatstellen zu bestimmen, muss man auf diese mangelnde Varianz in der Mobilität reagieren. Dies kann mit dem vorgestellten Ansatz zur Erhöhung der räumlichen Varianz vorgenommen werden. Wie in Abbildung 5.13 auf der rechten Seite visualisiert, erhält jedes Segment in einer Aggregationseinheit eine positive Wahrscheinlichkeit der Passage. Die Wahrscheinlichkeit spiegelt die Frequenzen des Frequenzatlas (vgl. Abschnitt 3.5.2) wieder.

Die Abbildung 5.14 zeigt die Aggregation und Disaggregation einer Trajektorie am Beispiel der Stadt Köln. Die Trajektorie startet und endet in einer inneren Aggregationseinheit. In diesen Aggregationseinheiten kann die Mobilität durch die Simulation variieren. An den Rändern verlaufen die Aggregationen der Grenzeinheiten. Hier sind die Variationen bei einer durchgeführten Simulation gering oder gar nicht vorhanden. Die Unterteilung zwischen inneren und Grenzräumen der Mobilität trägt der Verkehrslast Rechnung. Die Grenzeinheiten sind die verkehrsstärkeren Einheiten und sind aus diesem Grund kleiner, mit einer geringeren Segmentanzahl. Die Bewegungen in ihnen sind kompakt und tragen in der Regel den überregionalen Pendlerverkehr und den Transitverkehr zu Funktionsräumen einer Stadt. Mit diesem Vorgehen wird eine Vermischung von regionalem Verkehr und lokalem Verkehr vermieden. Auf makroskopischer Ebene bleibt die Mobilität erhalten.

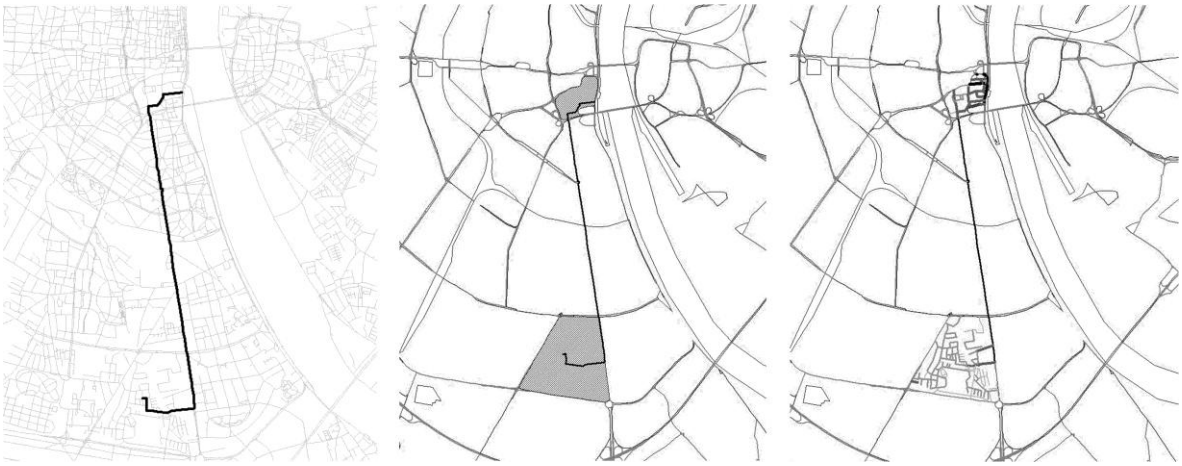


Abbildung 5.14: Original, aggregierte und simulierte Trajektorie für die Stadt Köln (Hecker et al. 2011b)

Greift man noch mal das Beispiel des Berliner Quartiers aus Abschnitt 5.2 auf, so ergibt sich nun folgendes Bild (Abbildung 5.15). Beide Plakate besitzen nach den externen Informationen des Frequenzatlas die gleiche Passagewahrscheinlichkeit.

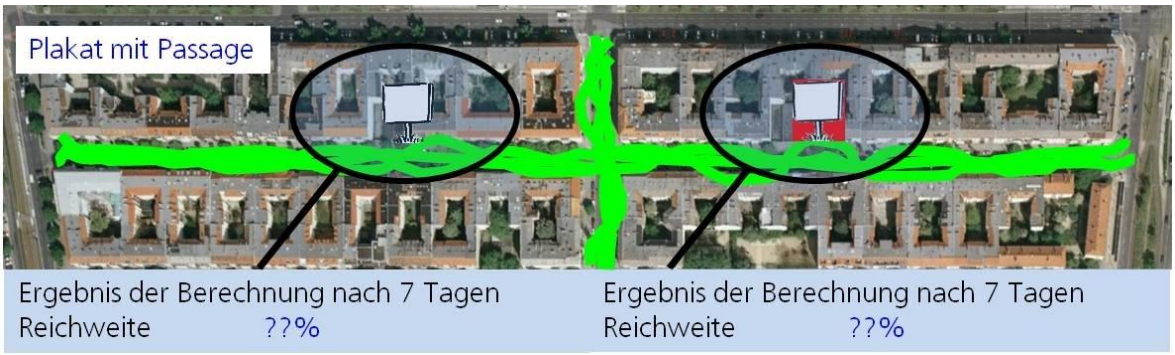



Abbildung 5.15: Auswertungen auf Straßensegmentniveau nach der Modellierung (topographische Hintergrundinformation Google 2012)

Wie nun die plakatindividuellen Kontakte berechnet werden, wird im Folgenden Abschnitt beschrieben.

### 5.2.7 Berechnung Reichweiten mit Kontaktwahrscheinlichkeiten in Deutschland

Für die Berechnung von Reichweiten auf Basis von Kaplan-Meier können individuelle Kontaktwahrscheinlichkeiten oder Kontaktdosen an Plakatstellen zugrunde gelegt werden (vgl. Abschnitt 5.1.1). Im folgenden Abschnitt wird der Berechnungsweg für Kontakwahrscheinlichkeiten vorgestellt. Die Kontaktwahrscheinlichkeit gibt die Wahrscheinlichkeit an, ob bei einer Passage ein Kontakt mit einem Plakat eintritt, oder nicht eintritt $P(Kontakt|Passage)$. Das heißt, bei einer festen Anzahl von Passagen ergibt sich eine Verteilung über die erzeugten Kontakte. Beispielsweise beträgt die Wahrscheinlichkeit, ein Plakat mit der Kontaktwahrscheinlichkeit 0,6 bei 3 Passagen genau zweimal zu sehen, $p$ = 0,6² · (1-0,6) · 3 = 0,432. Die Wahrscheinlichkeit, es überhaupt nicht zu sehen, beträgt $p$ = (1-0,6)³ = 0,064. Die Ereignisanalyse (Kaplan-Meier) geht jedoch davon aus, dass ein Ereignis entweder stattfindet oder nicht stattfindet, also entweder 0 oder 1 ist. In einem Wahrscheinlichkeitsmodell ist diese Voraussetzung jedoch nicht erfüllt. Um die Ereignisanalyse auf ein probabilistisches Modell anzuwenden, müssen daher die Kontakte über eine zweite Simulation realisiert werden. Dieses Vorgehen lässt sich am Beispiel eines

Plakatnetzes mit 2 Plakaten verdeutlichen. Die Plakate besitzen die Kontaktwahrscheinlichkeiten 0,8 und 0,4. An beiden Plakaten erfolgen 10 Passagen.

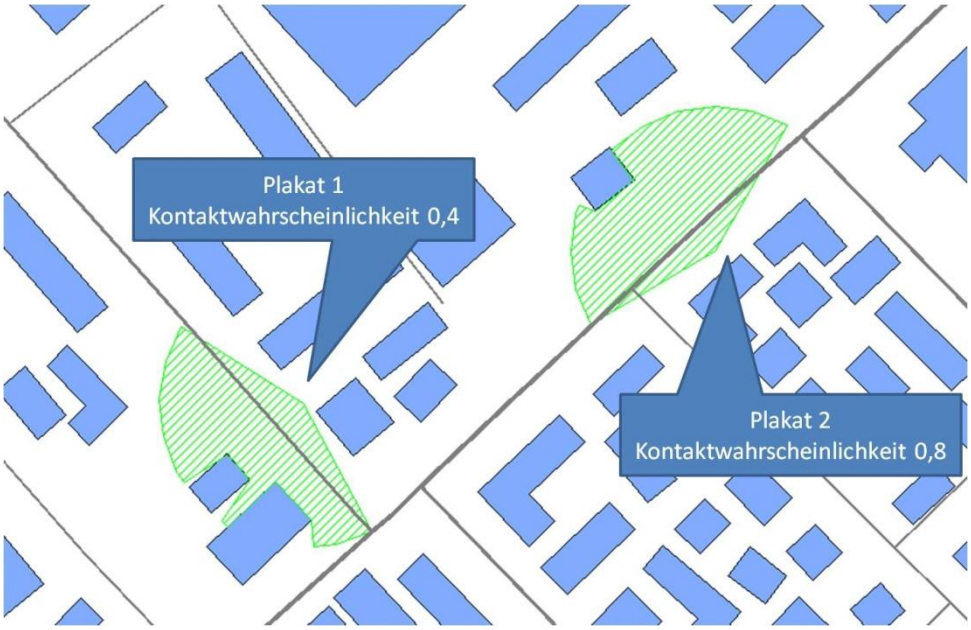



Abbildung 5.16: Beispiel Kontaktsimulation

Die möglichen Ausgänge sind wie folgt:

- P1 gesehen, P2 gesehen
- P1 nicht gesehen, P2 gesehen
- P1 gesehen, P2 nicht gesehen
- P1 nicht gesehen, P2 nicht gesehen.

Jede dieser Welten kann mit Kaplan-Meier behandelt werden, da sie feste Ereignisse enthalten. Bei $n$ Plakaten gibt es $2^n$ mögliche Ausgänge, die jedoch mit verschiedenen Wahrscheinlichkeiten eintreten. Bei mehrfacher Simulation werden im obigen Beispiel (Abbildung 5.16) im Mittel 8 Kontakte beim ersten Plakat und 4 Kontakte beim zweiten Plakat erzeugt, was jeweils der Anzahl an erwarteten Kontakten bei 10 Passagen entspricht. Das bedeutet, dass bei der Berechnung auf Basis von Kontaktwahrscheinlichkeiten immer ein durchschnittlicher Reichweitenwert berechnet wird. Zuvor muss jedoch die Berechnung, die auf Basis von echten GPS-Trajektorien erläutert worden ist, auf das in Abschnitt 5.2.2 vorgestellte System der Mobilitätseinheiten angepasst werden.

**Berechnung von Passagewahrscheinlichkeiten mit einem Plakat auf Basis der Mobilitätseinheiten**

Durch den Einsatz des Aggregationssystems werden die Passagen der Trajektorien zur Erhöhung der Variabilität in der Mobilitätseinheit verteilt. Dadurch erhält, wie in Abschnitt 5.2.4 beschrieben, jedes Straßensegment einer passierten Mobilitätseinheiten eine bestimmte Passagenwahrscheinlichkeit. Für die Bestimmung der Passagenwahrscheinlichkeit in den Mobilitätseinheiten wird der Frequenzatlas mit seinen für jedes Straßensegment vorliegenden Frequenzen herangezogen. Abbildung 5.17 zeigt nochmals schematisch einen solchen Funktionsraum mit 2 Plakatstellen und den anliegenden Frequenzen.

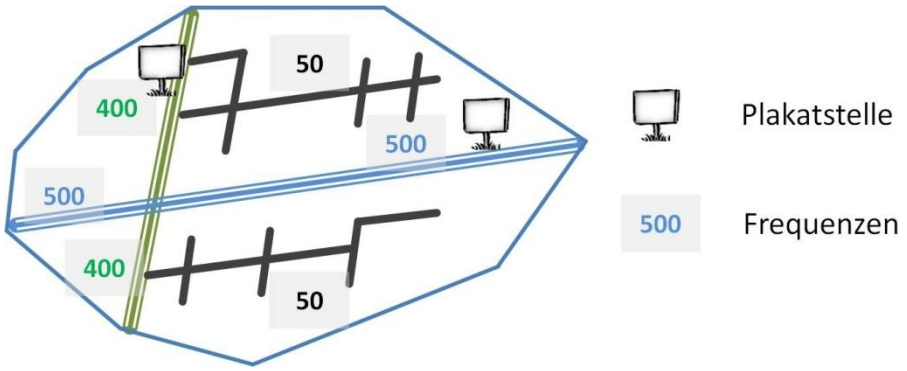



Abbildung 5.17: Beispiel innere Mobilitätseinheit

Die Zahlen in der Abbildung geben jeweils die Frequenzen des Frequenzatlas wieder. Die hochfrequenten Straßen, hier die blauen Doppellinien, tragen mehr Frequenz als die anliegenden Wohnstraßen mit nur 50 Pkws pro Stunde. Dabei tragen zur Vereinfachung alle schwarzen Straßensegmente die Frequenz 50 (20 Segmente), die grünen Straßensegmente die Frequenz 400 (5 Segmente) und die blauen Straßensegmente die Frequenz 500 (2 Segmente). Um die Passagewahrscheinlichkeit eines Straßensegmentes in einer Mobilitätseinheit zu bestimmten, wird nun die relative Frequenzmasse für jedes Straßensegment errechnet. Diese ergibt sich aus dem Quotienten der Straßensegmentfrequenz und der Summe aller Straßensegmentfrequenzen in der Mobilitätseinheit (vgl. Abschnitt 5.2.4).

Die Abbildung 5.18 zeigt das Vorgehen im Fall der idealisierten Mobilitätseinheit. Jedes Straßensegment erhält eine individuelle Passagewahrscheinlichkeit nach seinem jeweiligen Frequenzanteil des Frequenzatlas.

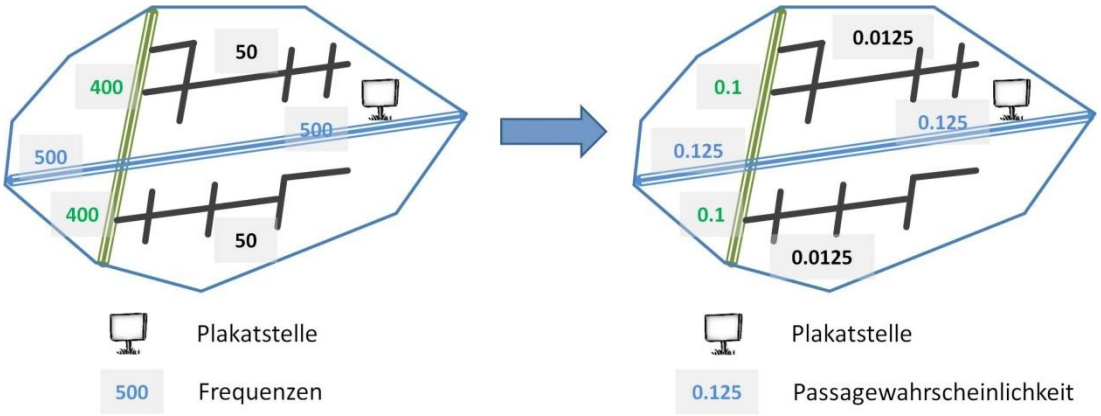



Abbildung 5.18: Berechnung der Passagewahrscheinlichkeit

Die Wahrscheinlichkeit, ein bestimmtes Plakat $P_i$ zu passieren, wenn die Mobilitätseinheit betreten wird, entspricht der Passagewahrscheinlichkeit $p_i$ des Segmentes, an dem das Plakat steht. Auch die Wahrscheinlichkeit, wenigstens eines von mehreren Plakaten zu passieren, ist einfach zu berechnen. Sie ergibt sich aus der Summe der einzelnen Passagewahrscheinlichkeiten aller betrachteten Plakate.

Bei der Durchquerung einer Mobilitätseinheit werden im Schnitt $n$ Segmentpassagen erzeugt. Diese Zahl ergibt sich aus den GPS-Originaldaten. Die Wahrscheinlichkeit, dass mindestens eine dieser Passagen Segment $i$ trifft, ist:

$$P(Segment|Mobilitätseinheit) = 1 - (1 - p_i)^n$$

**Berechnung der Kontaktwahrscheinlichkeit mit einem Plakat auf Basis von Mobilitätseinheiten**

Für die Leistungswertberechnung müssen Kontakt- und Passagewahrscheinlichkeiten kombiniert werden. Ziel ist es, zu jedem Plakat einer Kampagne die individuelle Kontaktwahrscheinlichkeit (Winkel der Plakatfläche, Umfeldkomplexität, Höhe des Werbeträger, etc.) mit einem gegebenen Track zu bestimmen. Da die Wahrscheinlichkeit eines Plakatkontaktes einer Passage unabhängig von der Passagewahrscheinlichkeit eines Segmentes ist, können für die Berechnung der gemeinsamen Wahrscheinlichkeit beide Werte miteinander multipliziert werden. Es wird als eine 2-stufige Simulation durchgeführt:

1. Wahrscheinlichkeit der Passage
2. Wahrscheinlichkeit des Kontaktes

Für die Wahrscheinlichkeit eines Plakatkontaktes beim Betreten einer Mobilitätseinheit ergibt sich somit:

$$P(Kontakt|Mobilitätseinheit) = P(Kontakt|Segment) * P(Segment|Mobilitätseinheit)$$
$$= 1 - (1 - p * k)^n$$

Für jede Trajektorien-Simulation kann nun die Kontaktwahrscheinlichkeit mit jedem Plakat einer Kampagne in den durchquerten Mobilitätseinheiten berechnet werden. Daraus ergibt sich für jede Mobilitätseinheit ein Wahrscheinlichkeitsvektor, der pro Plakat eine Kontaktwahrscheinlichkeit enthält. Diese Struktur ist in Abbildung 5.19 mit insgesamt 4 inneren Mobilitätseinheiten idealisiert dargestellt. Die berechneten Kontaktwahrscheinlichkeiten dienen als Grundlage für die Simulation. Mobilitätseinheiten, die nicht von einer Trajektorie eines Probanden (rote Linie) berührt werden, fließen nicht in den Wahrscheinlichkeitsvektor ein.

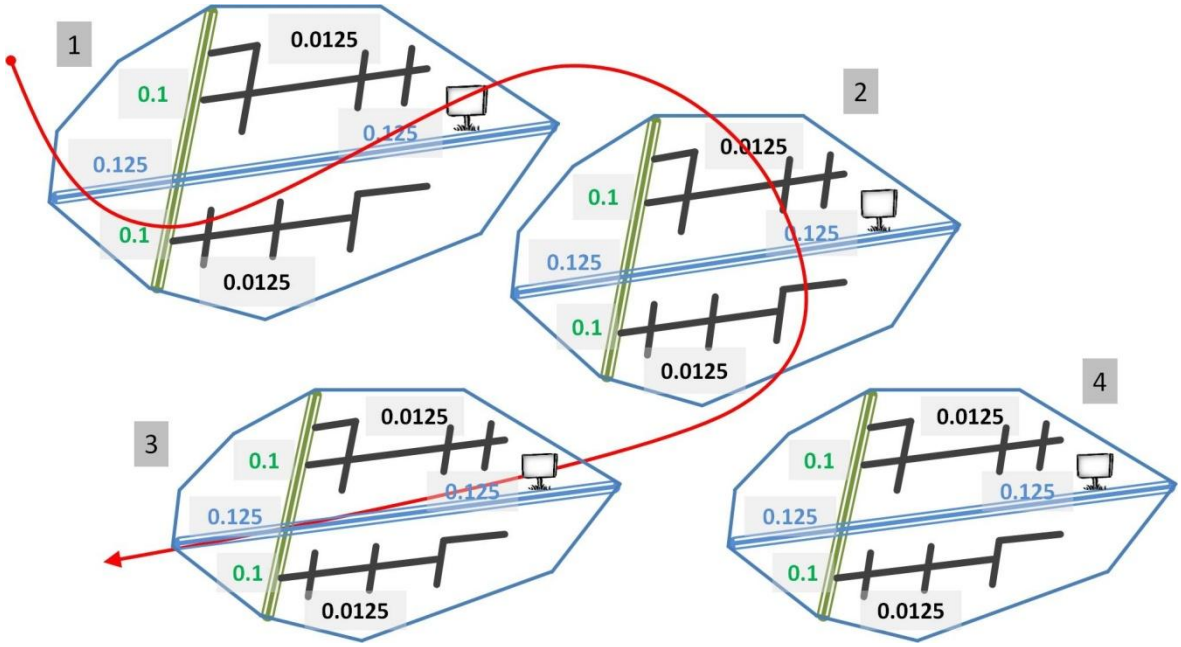



Abbildung 5.19: Simulation einer Kampagne über mehrere Mobilitätseinheiten

In jedem der 4 idealisierten Mobilitätseinheiten steht genau eine Plakatstelle. Für diese ergibt sich ein Wahrscheinlichkeitsvektor für den Beispielprobanden von:

$$< 0.125; 0.125; 0.125; 0 >$$

Im Anschluss erfolgt die Zufallsrealisierung, hier beispielhaft dargestellt bei einer 10 fachen Simulation (Tabelle 5.5):

| **10 Fache Simulation** | | | |
|---|---|---|---|
| **Mobilitätseinheit 1** | Mobilitätseinheit 2 | Mobilitätseinheit 3 | Mobilitätseinheit 4 |
| **1** | 0 | 1 | 0 |
| **0** | 0 | 0 | 0 |
| **0** | 0 | 0 | 0 |
| **0** | 0 | 0 | 0 |
| **0** | 1 | 0 | 0 |
| **0** | 0 | 0 | 0 |
| **1** | 0 | 0 | 0 |
| **0** | 0 | 0 | 0 |
| **0** | 0 | 1 | 0 |
| **0** | 0 | 0 | 0 |
| **∑ 2** | **∑ 1** | **∑ 2** | **∑ 0** |

Tabelle 5.5: Simulation einer Plakatkampagne mit 4 Plakatstellen

Es ist zu erkennen, dass für jede Simulationswelt ein Wert 0 oder 1 bestimmt wird. 1 bedeutet in diesem Fall Kontakt mit einem Plakat. Jede dieser einzelnen Welten wird mit Kaplan-Meier gerechnet. Am Ende werden die jeweiligen Ergebnisse verdurchschnittlicht und ausgewiesen.

Die Wahrscheinlichkeitsverteilung kann nach Verkehrsarten unterschieden werden. Da der Frequenzatlas die Verkehrsarten PKW, Fußgänger und ÖPNV ausweist, kann für jede Mobilitätseinheit auch eine Passagengewichtung anhand dieser Verkehrsarten vorgenommen werden. Das bedeutet, dass die Wahrscheinlichkeiten für jede Mobilitätseinheit nach dem Vorhandensein der Verkehrsarten z.T. dreimal berechnet werden müssen. Ein Segment, das keinen ÖPNV-Verkehr trägt, hat demnach eine Passagenwahrscheinlichkeit von Null, kann aber von Fußgängern und PKW Fahrern gesehen werden (analoges Beispiel für Kfz-Verkehr ist eine Fußgängerzone).

Die Abbildung 5.20 zeigt einen abschließenden Überblick über die einzelnen Modellierungsschritte. Im ersten Schritt (1) wird ein System der Mobilitätseinheiten erstellt und mit der externen Datenquelle des Frequenzatlas Wahrscheinlichkeiten für das Betreten jedes Straßensegmentes gebildet. Im zweiten Schritt (2) werden die erfassten GPS-Trajektorien den Mobilitätseinheiten zugeordnet und im Anschluss über mehrfache Simulation räumlich disaggregiert. Kommt eine Person an einer Plakatstelle vorbei, so können individuelle Plakatfaktoren wie Ausrichtung des Plakates, Beleuchtung, Größe, etc. Rechnung getragen werden (3). Auf der Grundlage der Simulation und der plakatindividuellen Faktoren wird ein Wahrscheinlichkeitsvektor aufgebaut (4). Dieser ist Input für die N-fache Reichweitenberechnung durch Kaplan-Meier (5). Im letzen Punkt (6) werden die jeweiligen Ergebnisse gemittelt und ausgewiesen.

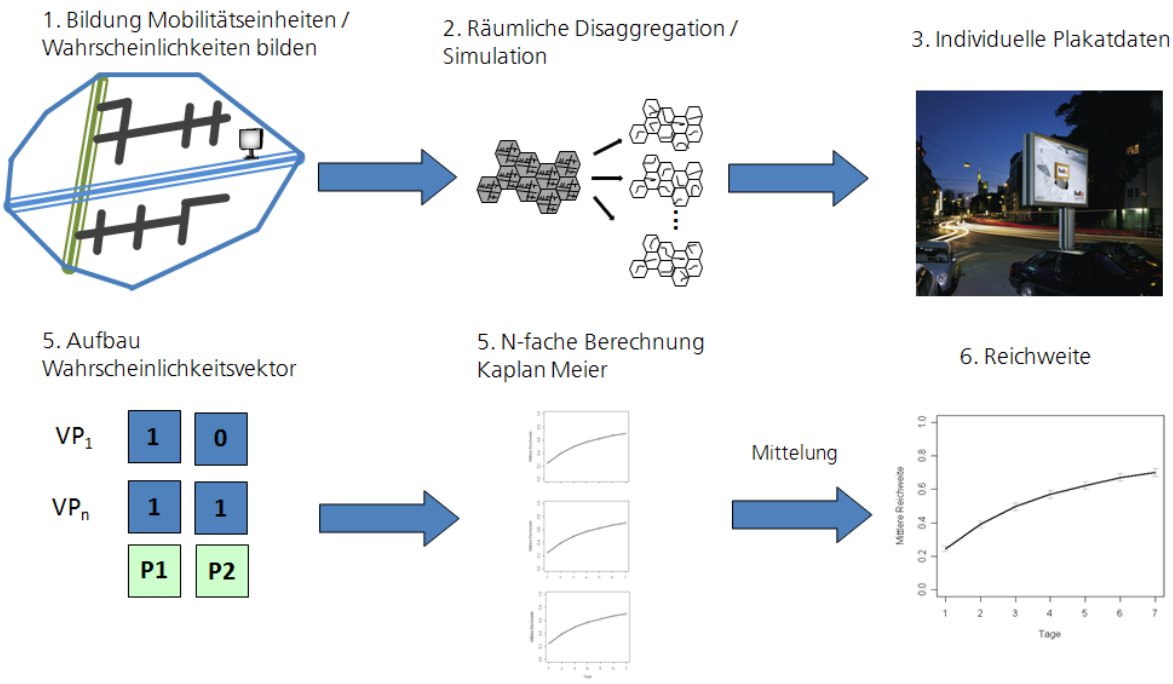



Abbildung 5.20: Modellierungsübersicht Deutschland

### 5.2.8 Berechnung der Reichweiten mit Kontaktdosen in der Schweiz

Die Reichweitenberechnung in der Schweiz unterscheidet sich in zwei wesentlichen Punkten vom Berechnungsprozess in Deutschland.

Der erste Unterschied bezieht sich auf die unterschiedliche Stichprobengröße in Deutschland und der Schweiz. So kommt in der Schweiz in den 12 erhobenen Agglomerationen von insgesamt 25.990 Plakatstellen nur bei 0.4% der Plakatstellen keine Passage vor. Das ist zum einen dadurch zu begründen, dass die Plakatstellen häufiger an verkehrsreichen Standorten stehen und zum anderen die Stichprobe im Vergleich zu Deutschland massiv größer ist (vgl. Abschnitte 3.5.3 und 3.5.4). Bei Plakatstellen, die keine Passage aufweisen, handelt es sich zudem auch häufig um Fehlverortungen der X,Y Koordinate (Plakatfläche steht im See), die Plakatfläche schaut in die Gegenrichtung des Straßensegmentes, oder es handelt sich um eine Objektfläche, die innerhalb eines Gebäudes steht, aber als Straßenfläche gekennzeichnet ist.

Den zweiten Unterschied stellt die Berechnung der Kontakte an der jeweiligen Plakatstelle dar. Während in Deutschland jede Person, die ein Plakat passiert, eine bestimmte Wahrscheinlichkeit besitzt, das Plakat zu sehen (Wert entweder 1 oder 0), wird in der Schweiz bei jeder Passage, die entweder frontal oder parallel am Plakat vorbeigeht, ein Kontakt erzeugt (zwischen 0,3 als niedrigster Wert und 1). Einzelne Plakatkontakte von Probanden werden kumuliert über die Zeit berechnet (vgl. Abschnitt 2.1).

Diese Unterschiede haben einen entscheidenden Einfluss auf den Modellierungsweg. Aufgrund der umfangreicheren Abdeckung der Stichprobe wird in der Schweiz auf den Einsatz der Moilitätseinheiten verzichtet. Das bedeutet, dass die Kontakte direkt aus den einzelnen Passagen der GPS-Stichprobe ausgelesen werden können. Hier tritt auch der zweite Unterschied hervor. Es muss aufgrund der Behandlung der Plakatkontakte mit einer Kontaktdosis keine Simulation durchgeführt werden. Da jeweils reelle positive Zahlen vorliegen, können diese in direkter Weise in die Ereignisanalyse einfließen.

Auch in der Schweiz muss aufgrund der zeitlichen Unvollständigkeit das beschriebene Kaplan-Meier Verfahren eingesetzt werden. Die Transformation von der Überlebenswahrscheinlichkeit zur Plakatreichweite geschieht wie bei der Umformulierung bei der beschriebenen Wahrscheinlichkeitsberechnung in Abschnitt 5.2.7. $S(t)$ gibt die Wahrscheinlichkeit an, dass ein Proband aus der GPS-Stichprobe bis zum Zeitpunkt $t$ kein Plakat der Kampagne gesehen hat.

$$\mathrm{S(t)} = \mathrm{P(T} > t)$$

Folglich ist die Reichweite in der Schweiz zur Berechnung einer Kampagne durch die komplementäre Wahrscheinlichkeit eines Kontaktes gegeben:

$$\mathrm{F(t)} = \mathrm{P(t} \leq \mathrm{T)} = 1 - \mathrm{S(t)}$$

Die gespiegelte Survivalkurve resultiert in der Reichweitenkurve. In der Abbildung 5.21 sind beide Kurven beispielhaft für eine Kampagne mit insgesamt 25 Plakaten in der Agglomeration Winterthur dargestellt. Auf der linken Seite wird die Survivalkurve als Treppenfunktion dargestellt. Jedes Mal, wenn ein neuer Plakatkontakt generiert wird, sinkt die Überlebenswahrscheinlichkeit. Auf der rechten Seite ist die komplementäre Reichweitenkurve mit einer linearen Interpolation über einen Zeitraum von 7 Tagen zu sehen. Im Ergebnis wurden mit dieser Kampagne 79% der GPS-Probanden erreicht.

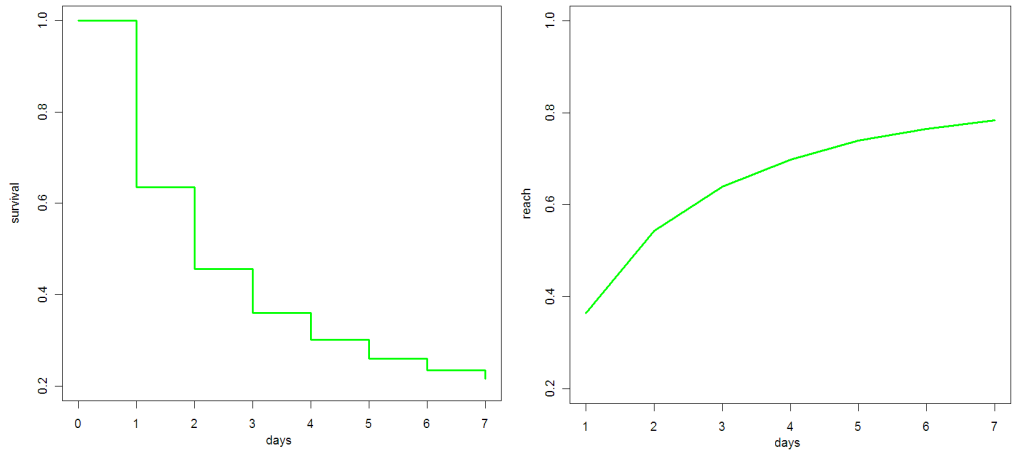



Abbildung 5.21: Reichweitenberechnung in der Schweiz

Dabei hat jede Person, die ein Plakat in Winterthur passiert, eine plakatindividuelle „Dosis", das Plakat zu sehen. So baut ein PKW Fahrer, der an einem Plakat vorbeikommt das parallel zur Straße positioniert ist, eine Dosis von 0,3 auf. Ein Fußgänger der an der gleichen Stelle vorbeikommt, hat einen Vollkontakt von 1. Jede einzelne Passage wird nun in der Schweiz über die Zeit kumulativ aufaddiert, wobei den plakatindividuellen Sichtbarkeitsfaktoren aus Abschnitt 2.2.2 Rechnung getragen wird.

Für die Berechnung der Reichweiten in der Schweiz mit Kontaktdosen wird folgt vorgegangen. Für den gegebenen GPS-Datensatz wird mit Kaplan-Meier berechnet, zu welchem Zeitpunkt $t_i$ ein Ereignis eintritt. Folgende Variablen sind notwendig:

- $r_i$ - Anzahl der Personen „at Risk" zum Zeitpunkt $t_i$
- $d_i$ - Anzahl der Ereignisse zum Zeitpunkt $t_i$ auf Basis von Kontaktdosen
- $c_i$ - Zahl der zensierten Personen zwischen $t_i$ und $t_{i+1}$
- Zufallsvariable $T$ – Die Zeit bis zum Eintreten des ersten Plakatkontaktes

Die Anzahl der Personen „at Risk" zu einem Zeitpunkt $t_{i+1}$ besteht aus allen Personen, die nicht zensiert worden sind und noch keinen Kontakt mit einer Plakatstelle hatten. Personen, die eine Passage an einem Plakat haben und einen Kontaktwert > 0 erzielen, werden als Ereignis aufgezeichnet. Überschreitet der kumulierte Wert dieser Person einen definierten Schwellwert (Kontaktklasse, vgl. Abschnitt 2.1), dann wird diese Person für $t_{i+1}$ zensiert. Bei einem Schwellwert von 1 muss eine Person bei einer Kontaktdosis 0,3 also mindestens viermal an der bestimmten Plakatkampagne vorbeikommen, um in den Zensurmechanismus von Kaplan-Meier aufgrund eines Plakatkontaktes zu fallen. Die bedingte Wahrscheinlichkeit $p_i$ zum Zeitpunkt $t_i$ gegeben das $t_{i-1}$ zu überleben wird dann wie folgt berechnet:

$$p_i = P(T > t_i \mid T > t_{i-1}) = \frac{r_{i-1} - d_i}{r_{i-1}}.$$

Die Anzahl der Ereignisse $d_i$ werden jeweils solange für jeden einzelnen Probanden addiert, bis der Schwellwerte überschritten wird.


Abbildung 5.22: Modellierungsübersicht Schweiz

Die Abbildung 5.22 zeigt einen abschließenden Überblick über die Modellierungsschritte in der Schweiz. Im ersten Schritt (1) werden die erfassten GPS-Trajektorien mit den Sichtbarkeitsräumen der Plakatstellen verschnitten. Dabei bauen Probanden, die an Plakaten vorbeikommen, über die Zeit eine Kontaktdosis auf. Diese Kontakte werden mit den individuellen Plakatfaktoren verrechnet (2). Im Anschluss erfolgt Kaplan-Meier Reichweitenberechnung.

Fasst man das Vorgehen der Modellierung zusammen, so handelt es sich um eine deutlich komplexere Lösung als das bisher eingesetzte Verfahren des A-Wertes in der Schweiz (vgl. Abschnitt 2.3), allerdings auch um eine weniger komplexe Modellierung als in Deutschland. Mit dem vorgestellten Berechnungsweg rein über das Kaplan-Meier Verfahren erfolgt die Modellierung in einer sehr direkten Weise auf den erhobenen Daten über eine modell-inhärente Kompensation von fehlenden Messdaten. Individuelle Plakatsichtbarkeitsfaktoren können passagenindividuell berücksichtigt werden. Wie realitätsnah und plausibel die Ergebnisse für die Schweiz sind, wird in Kapitel 6 vorgestellt. Die Unterschiede zum System der Mobilitätseinheiten werden in Abschnitt 5.2.5 vorgestellt.

### 5.2.9 Vergleich der Reichweitenberechnung mit und ohne Erhöhung räumlicher Variabilität

Um einen Vergleich der beiden vorgestellten Berechnungswege auf Basis der Mobilitätseinheiten und direktem Weg zu ermitteln, wird in diesem Abschnitt ein Vergleich zwischen einer direkten Berechnung aus den GPS-Trajektorien und mit dem Verfahrensweg der Erhöhung der räumlichen Variabilität (Mobilitätseinheit) eine Leistungswertberechnung durchgeführt. Ziel ist es, etwaige Unterschiede bei beiden Berechnungsvorgängen zu identifizieren. Es ist zu erwarten, dass Unterschiede insbesondere bei den

Reichweitenergebnissen existieren, da die Variabilität der Mobilität beim System der Mobilitätseinheiten angehoben wird und dadurch die Reichweite höher liegen müsste. Bei beiden durchgeführten Berechnungswegen wird das Kaplan-Meier Verfahren zur Kompensierung der fehlenden Messtage eingesetzt.

Um die beiden Wege der Berechnung miteinander zu vergleichen, werden Durchschnittsplakatkampagnen mit unterschiedlicher Größe für Deutschland berechnet. Einmal auf Basis der Mobilitätseinheiten und einmal mit dem beschriebenen Weg, der in der Schweiz eingesetzt wird. Hierzu werden jeweils 20 Netze gleicher Größe zufällig gezogen. Jede Berechnung mit der Methode der Mobilitätseinheiten wird mit 100 Simulationen ausgeführt.

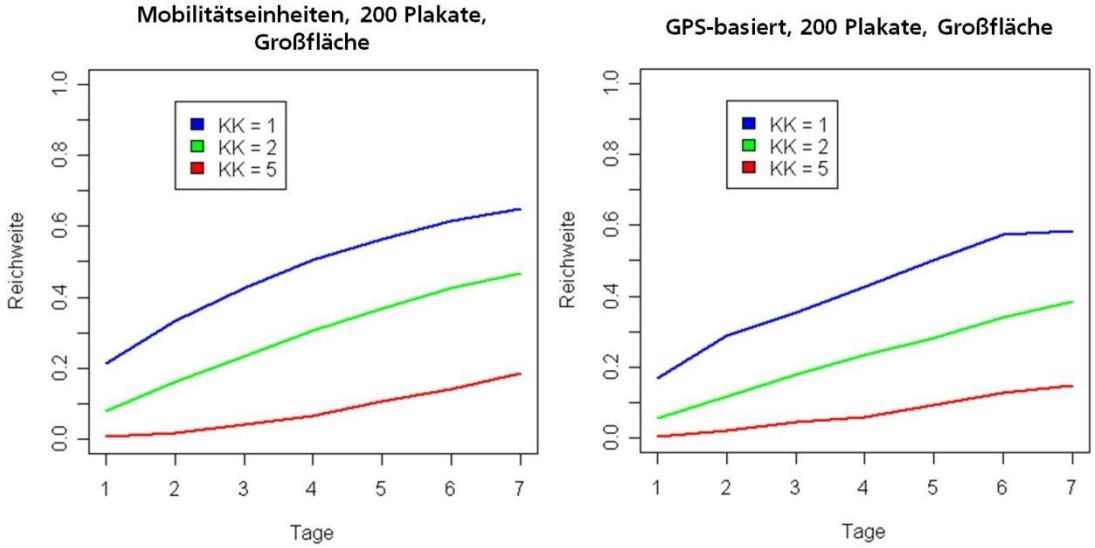



Abbildung 5.23: Vergleich Mobilitätseinheiten und rein GPS-basierte Berechnung mit 200 Plakaten in Hamburg

In Abbildung 5.23 sind die Ergebnisse für 200 Plakate in Hamburg dargestellt. Die X-Achse stellt die Tage von 1 bis 7 dar und die Y-Achse die erzielte Reichweite. Zum Vergleich wurden jeweils für beide Berechnungswege die Kontaktklassen 1, 2 und 5 berechnet. Auf der linken Seite sind die Reichweiten nach Kontaktklassen für die Mobilitätseinheit basierten Berechnungsweg dargestellt, auf der rechten Seite die Berechnung nach reiner GPS-Wegeauswertung. Man erkennt, dass der Reichweitenverlauf über die Zeit und die Kontaktklassen relativ gleichmäßig ist, jedoch bei der absoluten Höhe Unterschiede in den Leistungswerten existieren. In der Tabelle 5.6 ist dieser Unterschied noch deutlicher zu erkennen. Bei der Berechnung mit Mobilitätseinheiten liegen die Reichweitenwerte über alle Kontaktklassen hinweg über den Werten der rein auf GPS-Basis berechneten Werte.

| | | # Plakate | Kontaktklasse 1 | Kontaktklasse 2 | Kontaktklasse 5 |
|---|---|---|---|---|---|
| Reichweite | Mobilitäts-einheit | 200 | 64% | 42% | 18% |
| OTS | Mobilitäts-einheit | 200 | 4 | 5 | 8 |
| Reichweite | GPS | 200 | 58% | 38% | 14% |
| OTS | GPS | 200 | 4 | 5 | 10 |

Tabelle 5.6: Ergebnisse zur Kontaktklassenberechnung für Hamburg

In einem weiteren Beispiel wurde in Abbildung 5.24 die Berechnung für insgesamt 400 Großflächen in Hamburg durchgeführt, auch hier jeweils für die Kontaktklassen 1, 2 und 5 und für beide Varianten. Die Reichweiten der einzelnen Kontaktklassenberechnungen liegen wieder über denen der rein GPS-basierten Methode.

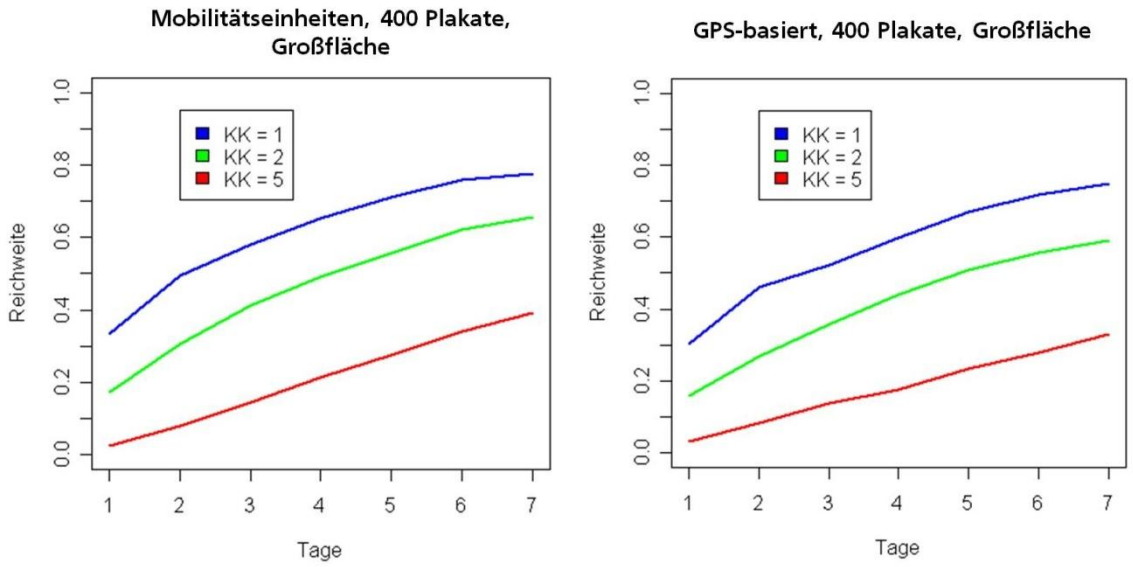



Abbildung 5.24: Vergleich Mobilitätseinheiten und rein GPS-basierte Berechnung mit 400 Plakaten in Hamburg

Die Tabelle 5.7 zeigt die Ergebnisse der Berechnung im Detail auch für den OTS. Wie in der vorangegangenen Berechnung sind hier die Unterschiede kleiner. Erst mit steigendem Kontaktklassenwert liegt der OTS weiter auseinander.

| | | # Plakate | Kontaktklasse 1 | Kontaktklasse 2 | Kontaktklasse 5 |
|---|---|---|---|---|---|
| Reichweite | Mobilitäts-einheit | 400 | 77% | 65% | 39% |
| OTS | Mobilitäts-einheit | 400 | 6 | 7 | 11 |
| Reichweite | GPS | 400 | 74% | 58% | 32% |
| OTS | GPS | 400 | 6 | 8 | 13 |

Tabelle 5.7: Ergebnisse zur Kontaktklassenberechnung für Hamburg

In der Abbildung 5.25 sind jeweils für die Kontaktklassen 1,2 und 5 die Reichweitenverläufe der Berechnungswege miteinander verglichen. Blau dargestellt ist die Methode mit Mobilitätseinheiten, grün die GPS-basierte Methode. Es ist zu erkennen, dass über alle Kontaktklassen hinweg die Methode mit Mobilitätseinheiten von Tag 1-7 über den Werten der GPS-basierten Methode liegt. Der Unterschied bleibt auch zwischen den einzelnen Tagen relativ konstant und steigt nicht etwa über die Zeit an.

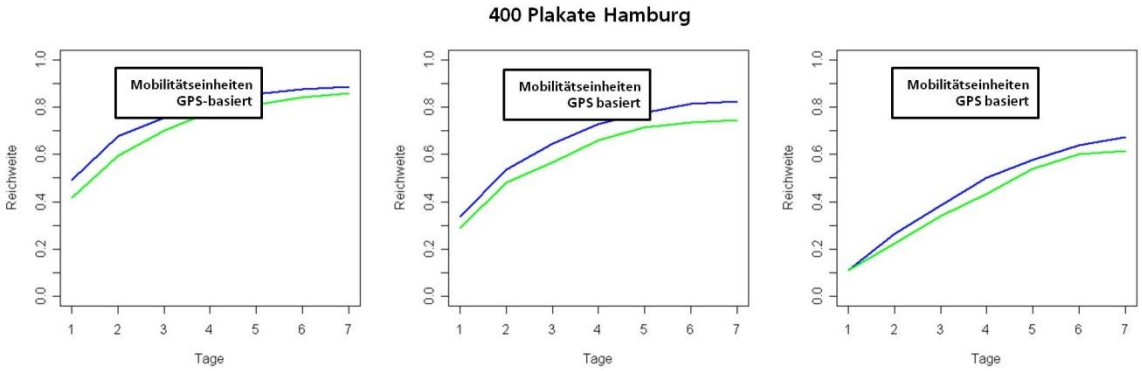



Abbildung 5.25: Vergleich Reichweitenverlauf nach Kontaktklassen 1,2 und 5

In einem abschließenden Vergleich wird in Abbildung 5.26 beispielhaft die Verteilung der Werte von 100 zufällig gezogenen Plakatkampagnen mit je 400 Plakaten in Hamburg dargestellt. Im Ergebnis dieser Berechnung liegt bei der Methode der Mobilitätseinheiten eine 3,4% höhere Reichweite bei leicht sinkendem OTS vor als bei einer rein GPS-basierten Methode. Die Standardabweichung der Reichweite beträgt dabei 0,6%.

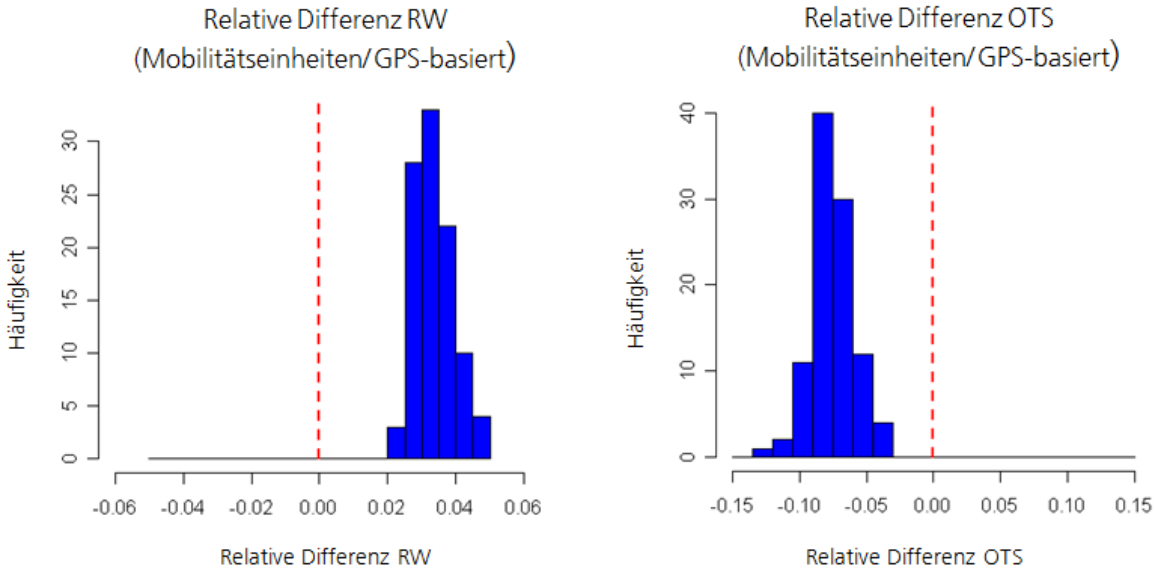



Abbildung 5.26: Relativer Vergleich Mobilitätseinheiten und rein GPS-basierte Berechnung mit 400 Plakaten in Hamburg

In der Analyse zeigen die vorgestellten Vergleiche generell leicht höhere Reichweitenwerte und niedrigere OTS-Werte als bei der Methode der Mobilitätseinheiten. Dies erklärt sich durch das Verwischen von Wegen innerhalb der Mobilitätseinheiten. Wir wissen, dass die Probandenvariabilität die Variabilität der Gesamtpopulation unterschätzt. Anstatt dass Kontakte nun durch Probanden sehr konzentriert erzeugt werden, bewirkt die Streuung der Wege über die Mobilitätseinheiten auch eine Streuung der Kontakte über die Probanden. Durch die Umlegung über den Frequenzatlas wird diese Diversität im Verhalten der Population approximiert. Es gibt mehr Personen, welche eine Passage mit einem Plakat haben, aber im Durchschnitt haben die Personen weniger Kontakte pro Plakat. Dies hat eine steigende Reichweite und einen sinkenden OTS zur Folge.

## 5.3 Berechnung von Durchschnittsnetzen

Neben der Berechnung von spezifischen Plakatnetzen wird die Reichweite von Mediaplanern auch gerne für Teilbelegungen, d.h. Plakatnetze verschiedener Größe und durchschnittlicher Reichweite, ausgewiesen. Diese dienen in der Außenwerbung einerseits als Benchmark für spezifische Netze und andererseits zur Bestimmung des Leistungswertes sollte keine individuelle Zusammenstellung des Netzes möglich sein. Da eine Teilbelegung fester Größe, je nach Auswahl der Plakate, eine andere Reichweite besitzt, wird die durchschnittliche Reichweite für Netze fester Größe per mehrfacher Berechnung bestimmt. Bei der Berechnung von Durchschnittsnetzen ergibt sich das Problem der kombinatorischen Vielfalt, da die Anzahl möglicher Plakatkombinationen exponentiell steigt. Beispielsweise gibt es in Hamburg 4179 City Light Poster. Um eine Teilbelegung mit 100 Plakaten zu bilden, gibt es

$$\binom{n}{k} = \binom{4179}{100} = \frac{4179!}{4079! * 100!} \approx 4{,}16 * 10^{203}$$

verschiedene Möglichkeiten, die für den mittleren Teilbelegungswert zu beachten sind. Dieses Problem wird durch Anwendung der Monte-Carlo-Simulation gelöst. Die Monte-Carlo-Simulation verwendet Prinzipien der Wahrscheinlichkeitsrechnung und Statistik, um komplexe Probleme näherungsweise zu lösen. Sie wird auch als Methode der statistischen Versuche bezeichnet und hat das Ziel, Kenngrößen der Verteilung zu schätzen. Mit Hilfe wiederholter, zufälliger Simulation wird die Lösung des Problems näherungsweise bestimmt. Dabei ist das Vorgehen wie folgt:

1. Übertragung des Problems auf ein stochastisches Modell,
2. Durchführung einer großen Anzahl von Zufallsexperimenten (Simulation),
3. Bestimmung von Schätzwerten für das Ausgangsproblem aus den Ergebnissen der Zufallsexperimente.

Für die Reichweitenberechnung von Teilbelegungen nach Monte-Carlo werden zufällig n Plakatnetze gleicher Größe gezogen, deren Reichweite dann gemittelt wird. Je größer die Anzahl der Simulationen ist, desto genauer wird das Ergebnis. Es stellt sich die Frage, wie viele Simulationen für eine robuste Aussage einer durchschnittlichen Anzahl berechnet werden müssen.

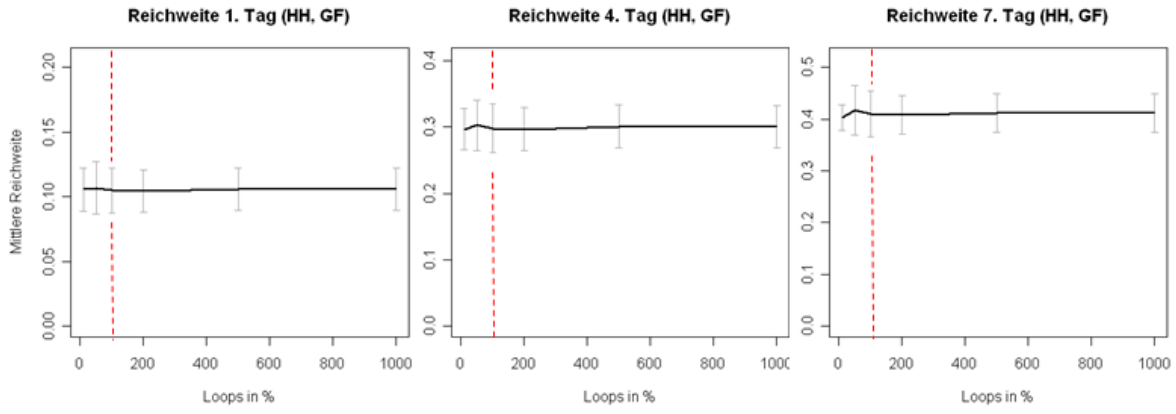



Abbildung 5.27: Monte Carlo Berechnung für Durchschnittsnetze

In Abbildung 5.27 sind für die Stadt Hamburg und die Plakatnetzgröße 100 die Ergebnisse der Monte-Carlo-Simulation für Großflächen (GF) abgetragen. Jeweils nach Tag 1, 4 und 7

wurde die mittlere Reichweite (Y-Achse) berechnet. Die Anzahl der Simulationen ist auf der X-Achse abgetragen und reicht bis zu einer Anzahl von 1000 Wiederholungen. Die Reichweite steigt über die Anzahl der Tage an. Die durchschnittliche Reichweite ist als schwarze Linie für die jeweiligen Simulationen abgetragen. Zusätzlich wird die Standardabweichung dargestellt. Man erkennt eine Stabilisierung der Mittelwerte und der Standardabweichung ab einer Simulationsanzahl von 100 Wiederholungen.

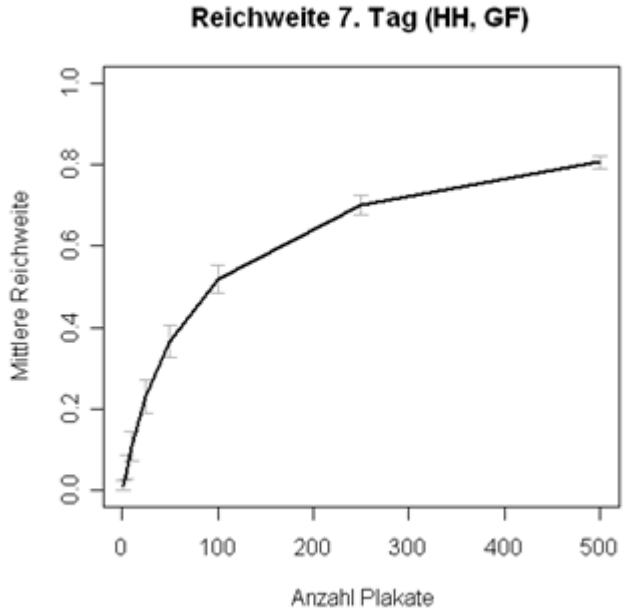



Abbildung 5.28: Berechnung der durchschnittlichen Reichweite über eine unterschiedliche Plakatanzahl

In der Abbildung 5.28 ist die Reichweite für unterschiedliche Teilbelegungsquoten nach 7 Tagen abgebildet. Auf der X-Achse sind unterschiedliche Plakatnetzgrößen von 0 bis 500 abgetragen. Die Y-Achse beschreibt die mittlere Reichweite bezogen auf die Plakatnetzgrößen. Die angegebenen Intervalle spiegeln die Standardabweichung der Reichweite der jeweils 200 zufällig gezogenen Netze wider. Die mittlere Reichweite nimmt mit steigender Netzgröße und steigender Aushangdauer zu. Die Standardabweichung nimmt mit zunehmender Plakatnetzgröße ab. Das heißt, je größer ein Netz ist, desto geringer ist auch die Reichweitenschwankung.

Die Ergebnisse der Berechnung können als Benchmark für speziell zugeschnittene Plakatkampagnen angesehen werden. Sobald der Reichweitenwert über dem Durchschnittswert liegt, kann man davon ausgehen, dass mehr kontaktstarke Plakate und eine gute Streuung des Plakatnetzes über die Stadt vorliegen. Mediaagenturen können mit dieser Hilfe die Qualität einer Plakatkampagne bewerten.

## 5.4 Zusammenfassung

In diesem Kapitel wurden Lösungen dafür vorgestellt, wie man mit einer unvollständigen räumlichen und zeitlichen GPS-Daten umgehen kann. Damit liefert dieses Kapitel auf zwei der drei eingangs gestellten Herausforderungen eine methodische Antwort. Wie die Ergebnisse der Reichweitenberechnung aussehen und wie robust die Ergebnisse sind, wird im folgenden Kapitel 6 erläutert. Im Detail haben sich die Abschnitte des Kapitels mit folgenden Themen beschäftigt.

Im ersten Abschnitt 5.1 wurde der Frage nachgegangen, wie man mit einem nicht vollständigen Messzeitraum umgeht. Die Anzahl der teilnehmenden Probanden nimmt in Deutschland und der Schweiz über den Erhebungszeitraum kontinuierlich ab. Es wurden insgesamt vier Möglichkeiten des Umgangs mit den unvollständigen Daten vorgestellt, wobei nur eine als sinnvoll und geeignet erscheint. Das gewählte Verfahren stammt aus dem Bereich der Ereignisanalyse und wurde für den Anwendungskontext der Außenwerbung adaptiert. Die generelle Idee bei dieser Technik ist, dass die Stichprobe an ihre variierende Größe tageweise angepasst wird. Probanden, die aufgrund ihres nicht vollständigen Trageverhaltens früher aus der Erhebung ausfallen, werden nach Kaplan-Meier zensiert. Personen, die genügend viele Plakatkontakte hatten, ebenfalls (Justierung über die Kontaktklasse). Durch die Zerlegung der Stichprobe in bedingte Wahrscheinlichkeiten passt das Verfahren die Stichprobe an die abnehmende Größe an und löst damit das Problem des unvollständigen Messzeitraumes. Abschnitt 5.1.2. überprüfte die Validität des vorgestellten Kaplan-Meier Verfahrens, indem bei Probanden mit einem vollständigen 7 Tage Datensatz eine künstliche Zensur der Tage durchgeführt wurde. Die Ergebnisse waren für unterschiedliche Kampagnengrößen und Zensurraten durchweg positiv. In Abschnitt 5.1.4 wurden die Berechnung und die Charakteristiken von Kontaktklassen vorgestellt. Diese sind ein probates Mittel zur Steuerung von Plakatkampagnen mit einer unterschiedlichen Werbewirkung.

Im zweiten Abschnitt 5.2 wurde der Frage nachgegangen, wie man mit der unvollständigen räumlichen Abdeckung der GPS-Empirie umgeht. Der vorgestellte Lösungsweg zielt darauf ab, eine realistischere Darstellung der räumlichen Variabilität der Mobilität zu erzeugen. Die grundlegende Idee der Modellierung ist dabei, die Makro- und Mikromobilität zu separieren und diese getrennt zu behandeln. Hierzu wird mithilfe eines Systems der Mobilitätseinheiten, dem Frequenzatlas und im Anschluss einer mehrfach durchgeführte Simulation eine erhöhte Variabilität erzeugt. Diese Herangehensweise ist auch über den Anwendungskontext hinaus eine sehr interessante Methode für andere Fragestellungen die sich mit kleinräumiger Mobilität auf Straßenniveau befassen.

Im Anschluss wurde in den Abschnitten 5.2.7 und 5.2.8 das Vorgehen zur Reichweitenberechnung mit dem System der Mobilitätseinheiten sowie Kaplan-Meier Verfahren in Deutschland und der Schweiz vorgestellt. Dabei ist ein wesentlicher Unterschied bei beiden Ländern, dass das Verfahren der Mobilitätseinheiten nur in Deutschland Anwendung findet. Darüber hinaus wurde der Modellierungsverlauf für die Berechnung mit Kontaktdosen und Kontaktwahrscheinlichkeiten vorgestellt. Der Abschnitt 5.2.9 validierte das Vorgehen der Berechnung der Reichweiten mit Mobilitätseinheiten, indem die Leistungswerte Reichweite und OTS der rein GPS-basierten Berechnung gegenübergestellt wurden. Es stellt sich heraus, dass die Methode bei der Reichweiteberechnung mit Mobilitätseinheiten generell höhere Reichweitenwerte, aber niedrigere OTS-Werte liefert. Diese Ergebnisse ließen sich methodisch durch das Verwischen von Wegen innerhalb der Mobilitätseinheiten erklären und sind in dieser Form gewollt. Mit der Umlegung durch den Frequenzatlas wird die Diversität im

Verhalten der Population besser approximiert und so entsteht ein plausibleres Bild der Reichweiten. Abgeschlossen wurde das Kapitel 5 mit Abschnitt 5.2.6 und der Berechnung von durchschnittlichen Plakatkampagnen. Diese spielen bei der Mediaplanung als sogenannter Benchmark für individuell zusammengestellte Plakatkampagnen eine wichtige Rolle.

Mit diesen Methoden werden im Kapitel 7 exemplarisch für unterschiedliche Kampagnenkonfigurationen und Städte Leistungswerte berechnet und die Vorteile der neuen Herangehensweise dargestellt. Zuvor wird noch eine Robustheitsanalyse der Reichweitenergebnisse durchgeführt.

# KAPITEL 6

## 6. ROBUSTHEITSANALYSE DER REICHWEITENERGEBNISSE

Im vorigen Kapitel wurde der Berechnungsweg zur Leistungswertbestimmung von Plakatkampagnen vorgestellt. Dieses Kapitel geht der Frage nach, wie robust die Ergebnisse der Reichweitenbestimmung sind. Hierzu werden Kampagnen bzgl. ihrer Stabilität bei einer geringeren GPS-Stichprobe analysiert. Zusätzlich wird die Frage bearbeitet, was passiert, wenn eine neue GPS-Stichprobe die bisherige Stichprobe ersetzt. Bleiben die Leistungswerte stabil, oder kommt es zu großen Veränderungen in der Ausweisung? Desweiteren wird die Frage untersucht, ob in Zukunft auch eine evtl. kleinere und damit kostengünstigere GPS-Stichprobe zur Leistungswertberechnung eine ausreichende Validität garantieren kann.

Das Kapitel 6 schließt mit einer Zusammenfassung der Ergebnisse ab (Abschnitt 6.3) und leitet dann über in die Präsentation einzelner Reichweitenergebnisse in Kapitel 7.

### 6.1 Robustheitsanalyse der Ergebnisse

Wie in den vorangegangen Kapiteln dargestellt, sind Mobilitätsinformationen eine der zentralen Datenquellen für die Außenwerbung geworden. In der Zukunft wird die Außenwerbung ältere Mobilitätsinformationen mit jüngerer Information erneuern müssen, um die Leistungswerte auf dem aktuellen Stand zu halten. In diesem Abschnitt werden Experimente vorgestellt, die die Auswirkungen veränderter GPS-Datensätze in der Außenwerbung zeigen. Dabei stellt die Außenwerbung nach wie vor ein besonderes Anwendungsgebiet dar, da sie die Evaluierung von Mobilitätsdaten in einer hoch aufgelösten räumlichen Auflösung benötigt. Im Vergleich dazu liegt der Fokus klassischer Mobilitätsstudien auf regionalen und nationalen Zusammenhängen, die weniger von individuellen Veränderungen im jeweiligen Datensatz beeinflusst werden. Gegenwärtige Mobilitätsstudien aus der Mobilitätsforschung befassen sich nicht mit dem Aspekt der Stabilität, um a) die Auswirkung einer wiederholten Mobilitätsstudie und b) die Auswirkung einer Mobilitätsstudie kleinerer Größe auf bereits vorhandene Ergebnisse zu überprüfen.

Im diesem Abschnitt wird untersucht, inwiefern neue und kleinere Datensätze von wiederholten Mobilitätsstudien die Reichweiten von Werbekampagnen beeinflussen (Hecker et al. 2011a).

#### 6.1.1 METHODISCHES VORGEHEN

Im diesem Abschnitt wird eine Übersicht über das methodische Vorgehen zur Robustheitsanalyse von Mobilitätsstudien gegeben. Dabei werden die einzelnen Aufbereitungsschritte erklärt und die verwendeten Algorithmen vorgestellt.

Dabei hat das methodische Vorgehen zur Robustheitsanalyse in dieser Arbeit zwei Ziele: Zum einen wird die Auswirkung eines geänderten GPS-Datensatzes auf die Leistungswerte untersucht. Dabei soll analysiert werden, wie sich ein Austausch eines bestehenden GPS-Datensatzes durch eine GPS-Neuerhebung auf die Reichweiten auswirkt. An dieser Stelle wird die Varianz von Leistungswerten für eine neue GPS-Stichprobe gleicher Größe untersucht.

Zum anderen wird die Variabilität des Standardfehlers, die von einem GPS-Datensatz kleineren Umfangs verursacht werden kann, untersucht, um dadurch eine angemessene, vielleicht kostengünstigere Datenmenge für zukünftige Mobilitätsstudien und Neuerhebungen bestimmen zu können. Oder anders ausgedrückt, wie hoch ist die stichprobenbedingte Variabilität des Schätzers bei einem kleineren GPS-Datensatz? In beiden Fällen liegt das Interesse der Robustheitsanalyse in statistischen Kennzahlen, mit denen Aussagen über Werbekampagnen in der Außenwerbung getroffen werden können, wie z.B. die Reichweite.

Das Ziel vieler Studien ist es, mit Hilfe eines erhobenen Datensatzes eine gewisse Kenngröße $\theta$ einer Bevölkerung zu analysieren. Neben dem Schätzwert $\hat{\theta}$ ist die Fehlergenauigkeit des Schätzwertes von großer Bedeutung. In diesem Zusammenhang ist das meist verwendete genauigkeits Maß der Standardfehler (se), welcher die Standardabweichung von $\hat{\theta}$ durch Ziehung einer Stichprobe beschreibt. Ist der Standardfehler bekannt, können beispielsweise Konfidenzintervalle für den tatsächlichen Wert von $\theta$ bestimmt werden. Obwohl der Standardfehler ein sehr probates Maß zur Berechnung der statistischen Genauigkeit darstellt, hat er den Nachteil, dass er für die meisten Kennzahlen, mit Ausnahme des Mittelwertes, nicht anhand einer Formel aus dem Datensatz berechnet werden kann (Efron 1993). Eine Lösung des Problems ermöglicht das Bootstrap-Verfahren, welches 1979 von Efron vorgestellt wurde (Efron 1979). Es bietet eine Methode, um den Standardfehler und anderweitige statistische Kenngrößen aus einer Stichprobe zu berechnen.

Wie in (Efron 1993) beschrieben, kann der Standardfehler eines Datensatzes durch das Bootstrap-Verfahren wie folgt berechnet werden:

Es bezeichne die Menge $x = (x_1, x_2, \dots x_n)$ einen Datensatz und $x^* = (x^*_1, x^*_2, \dots x^*_n)$ eine Bootstrap-Stichprobe, die durch ein n-faches Ziehen mit Zurücklegen aus der ursprünglichen Stichprobe $x$ generiert wird. Die Stichprobenziehung wird $r$-mal wiederholt, um eine Reihe von unabhängigen Bootstrap-Stichproben $x^*_1, x^*_2, \dots x^*_r$ zu erhalten. Für jede Stichprobe wird der sogenannte Bootstrap-Replikationswert $\hat{\theta}^{*i}$ mit $i = 1 \dots r$ gebildet. Die Schätzgröße des Standardfehlers ergibt sich dann aus der gemittelten Standardabweichung der einzelnen Bootstrap-Replikationswerte.

$$\hat{se} = \sqrt{\frac{\sum_{i=1}^{r} (\hat{\theta}^{*i} - \hat{\theta}^{*(\cdot)})^2}{r-1}} \text{ mit } \hat{\theta}^{*(\cdot)} = \frac{\sum_{i=1}^{r} \hat{\theta}^{*i}}{r}$$

Wenn die Anzahl der Stichprobenziehungen *r* im Bootstrap-Verfahren gegen unendlich geht, nähert sich der Schätzwert *sê* dem idealen Bootstrap Schätzer für den jeweiligen Datensatz an. Typischerweise sollte die Anzahl von Replikationen zwischen 25 und 200 liegen.

Eine Hypothese im Vorfeld der Experimente ist: Steigt die Anzahl von Plakaten einer Werbekampagne, so wirkt sich das stabilisierend auf die Leistungswerte aus, da lokale Effekte einer Kampagne verschwinden und sich gegenseitig aufheben. Folglich werden die Berechnungen in den folgenden Experimenten mit unterschiedlichen Kampagnengrößen durchgeführt, wofür mehrere zufällig gewählte Kampagnen herangezogen werden, um gemittelte Ergebnisse zu erhalten.

### Vorgehen mit Bootstrap bei gleicher Datensatzgröße

Wird in der Außenwerbung in Zukunft ein GPS-Datensatz verändert, indem neuere Daten von gleichem Umfang verwendet werden, muss überprüft werden, wie sich die statistischen Kennzahlen der neuen Daten auf die der älteren Daten auswirken. Hierzu wird im Folgendem

das Bootstrap-Verfahren verwendet. In diesem Experiment werden anhand des gesamten Datenumfangs zufällige Bootstrap-Stichproben gezogen und die Reichweite der jeweiligen Werbekampagne evaluiert. Für jede einzelne Anwendung des Bootstrap-Verfahrens wird der Standardfehler einer spezifischen Werbekampagne gebildet, welcher anschließend über Kampagnen mit der gleichen Größe gemittelt wird. Im folgenden Algorithmus werden die jeweiligen Schritte dargestellt:

Algorithmus 4. Bootstrap on Full Sample (Hecker et al. 2011a)

Input:

= set of campaign sizes $S_c = \{10,20,....,100\}$

= set of test persons *Pers*, set of poster locations *Loc* and set of poster passages *Pass*

= # bootstrap repetitions $r_b$, # campaign repetitions $r_c$

Output:

= $s\hat{e} = (s\hat{e}_{10}, s\hat{e}_{20}, \ldots, s\hat{e}_{100})$ vector with estimates of standard error for campaign sizes $s_c$

Method:

1: for $(s \in S_c)\{$ // iterate over campaign sizes
2: for $(j = 1..r_c)\{$ // iterate over $r_c$ campaigns per size
3: C = sample (*Loc*, s) // sample campaign
4: for (i = 1.. $r_b$) { // calc. bootstrap replications
5: $\hat{\theta}^{*i} = calcBootstrapReplication(Pers, C, Pass)$
6: }
7: $s\hat{e}_j = std(\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*r_b})$ // calc. stand. error
8: }
9: $s\hat{e}_s = avg(s\hat{e}_j \mid j = 1..r_c)$ // average se per camp. size
10: }
11: $s\hat{e} = (s\hat{e}_{10}, s\hat{e}_{20}, \ldots, s\hat{e}_{100})$

**Vorgehen mit Teilmengen und Bootstrap bei kleinerer Datensatzgröße**

Ist es das Ziel, die Auswirkung einer reduzierten GPS-Stichprobe auf Reichweiten zu analysieren, können zwei Fälle voneinander unterschieden werden. Ersterer bezieht sich alleinig auf die Auswirkung eines verkleinerten Datensatzes derselben Datenquelle. Im zweiten Fall wird die Auswirkung eines neuen GPS-Datensatzes kleineren Umfangs (einer anderen Datenquelle) analysiert. Aus diesem Grund wird zusätzlich zur Berechnung der Reichweite anhand unterschiedlich großer Werbekampagnen in diesen Experimenten die Größe des Datensatzes verändert, indem Teilmengen (Subsampling) des Datensatzes gebildet werden. Dabei werden ähnlich wie im vorhergehenden Beispiel für jede Teilmenge mehrere Stichproben gleicher Kampagnengröße gezogen, um eventuelle Abweichungen und Zufälligkeiten auszuschließen.

Um die eventuelle Veränderung der Leistungswerte einer Werbekampagne durch kleinere Teilmengen zu berechnen, wird die Abweichung zwischen der Reichweite im reduzierten und im vollen Datensatz berechnet. Genauer gesagt, wird die Wurzel aus dem mittleren quadratischen Fehler für verschieden große Probanden Teilmengen $S_s = \{2.5,5.0, \ldots, 97.5\}$ berechnet. In diesem Zusammenhang ist die Erwartung, dass der Fehler bei stetig kleiner werdenden Teilmengen des Datensatzes und abnehmender Kampagnengröße

zunimmt. Interessant ist hierbei, ob eine Korrelation zwischen diesen beiden Variablen existiert.

Im zweiten Fall wird zur Bestimmung die Auswirkung veränderter und kleinerer Datensätze auf die Reichweite noch zusätzlich das Bootstrap-Verfahren angewendet. Hierzu wird im ersten Schritt der GPS-Datensatz verkleinert, und dann im folgenden Schritt das Bootstrap-Verfahren durchgeführt. Im folgenden Algorithmus werden beide Methoden beschrieben:

Algorithmus 5. Bootstrap and Subsampling for Varying Numbers of GPS-Persons (Hecker et al. 2011a)

Input:

= set of campaign sizes $S_c = \{10,20,....,100\}$

= set of person subsample sizes in percent $S_s = \{2.5,5.0,\ldots,97.5\}$

= set of test persons *Pers*, set of poster locations *Loc* and set of poster passages *Pass*

= # bootstrap repetitions $r_b$, # campaign repetitions $r_c$ and # subsample repetitions $r_s$

Output:

= rmse = ($rmse_{ts}$) and sê = ($s\hat{e}_{ts}$) with s ∈ $S_c$ and t ∈ $S_s$ matrix with estimates of RMSE and standard error for campaign sizes Sc and subsample sizes $S_s$

Method:

1: for $(s \in S_c)\{$ // iterate over campaign sizes

2: for $(t \in S_s)\{$ // iterate over person subsample sizes

3: for $(j = 1..r_c)\{$ // iterate over $r_c$ campaigns per size

4: C = sample (Loc, s) // sample campaign

5: for $(k = 1..r_s)\{$ // iterate over $r_s$ persons groups per subsample size

6: D = sample (*Pers*, t) // sample pers. subsample

7: for (i = 1..rb) { // calc. bootstrap replication

8: $\hat{\theta}^{*i} = calcBootstrapReplication(D, C, Pass)$

9: }

10: $s\hat{e}_{jk} = std(\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*r_b})$ // calc. stand. error

11: $err_{jk} = calcReach(D, C, Pass) - calcReach(Pers, C, Pass)$ // calc. error

12: }

13: $rmse_j = \sqrt{\sum_{k=1}^{r_s} err_{jk}^2} \Big/ r_s$ // calc. RMSE over subsample

14: }

15: $s\hat{e}_{ts}$ = avg($s\hat{e}_{jk}$ | j = 1.. $r_c$, k = 1.. $r_s$) // average results

16: $rmse_{ts}$ = avg($rmse_j$ | j = 1.. $r_c$) // average results

17: }

18: }

19: sê = ($s\hat{e}_{ts}$)

20: rmse = ($rmse_{ts}$)

### 6.1.2 EXPERIMENTSETUP UND EXPERIMENTE

In diesem Abschnitt werden die vorgestellten Methoden auf Basis des GPS-Datensatzes der Schweiz angewandt. Die Berechnungen werden auf das Ballungsgebiet Zürich mit 1.956 Probanden und insgesamt 10.093 Plakatstandorten beschränkt. In der Summe hatten die GPS-Probanden 2.071.124 mal Kontakt mit Plakatstandorten innerhalb von 7 Tagen.

Im folgenden Experiment wird die Auswirkung eines veränderten GPS-Datensatzes für unterschiedliche Größen von Werbekampagnen evaluiert. Dabei wird das Bootstrap-Verfahren mit insgesamt 30 Wiederholungen angewendet, um den Standardfehler zu berechnen. Die Größe der Werbekampagne steigt inkrementell zwischen zehn und 100 Plakaten jeweils um 10 Plakate an. Für diese zehn Größen werden die Leistungswerte für je 15 simulierte Kampagnen berechnet und anschließend der Mittelwert gebildet. Tabelle 6.1 und Abbildung 6.1 zeigen die durchschnittliche Reichweite und den durchschnittlichen Standardfehler, die durch das Bootstrap-Verfahren für die unterschiedlichen Kampagnengrößen gebildet wurden. Bei zunehmender Plakatanzahl erhöht sich die Reichweite, während der Standardfehler bei wachsender Kampagnengröße abnimmt. Dieser Zusammenhang ist zu erwarten, da wie bereits oben beschrieben, lokale Effekte durch große Kampagnen ausgemittelt werden.

| Kampagnengröße | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Durchschnittliche Reichweite | 38,8 | 58,7 | 67,0 | 74,9 | 79,3 | 83,3 | 84,3 | 86,5 | 89,2 | 90,1 |
| Durchschnittlicher Standardfehler | 1,5 | 1,4 | 1,4 | 1,3 | 1,2 | 1,1 | 1,1 | 1,0 | 0,9 | 0,9 |

Tabelle 6.1: Durchschnittliche Reichweite und Standardfehler für unterschiedliche Kampagnengrößen

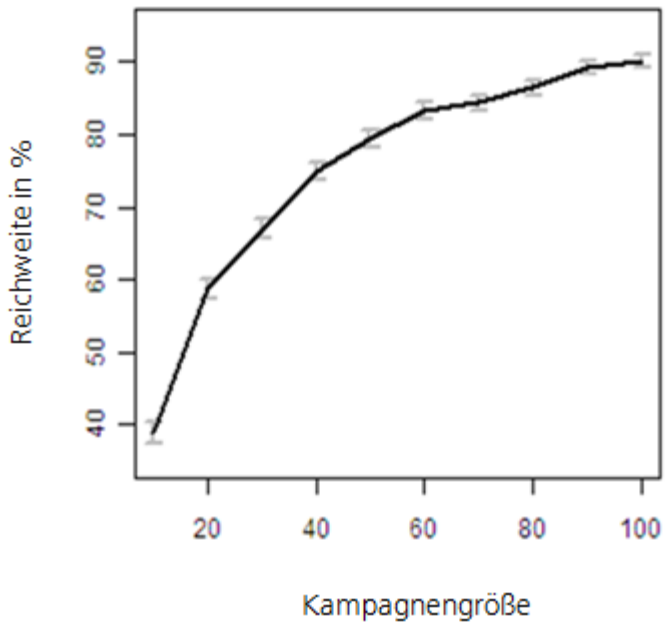



Abbildung 6.1: Durchschnittliche Reichweite und Standardfehler für unterschiedliche Kampagnengrößen (Hecker et al. 2011a)

Wie in Abbildung 6.1 zu sehen ist, ist der Standardfehler für alle Kampagnengrößen vergleichbar klein. Um Unterschiede im Detail zu erkennen, wird der im Bootstrap-Verfahren gebildete Standardfehler in vergrößerter Darstellung in Abbildung 6.2 gezeigt. Die Abbildung enthält auch eine lineare Regressionsfunktion, die aus den Ergebnissen abgeleitet wurde. Es zeigt sich eine lineare Beziehung zwischen dem Standardfehler einer geänderten GPS-Probe und der Größe von Plakatkampagnen. In der Praxis bedeutet dies, dass die Zahl der Plakate in einer Kampagne einen direkten Einfluss auf die Variabilität der Leistungswerte hat.

Eine typische Werbekampagne in Zürich umfasst etwa 50 Plakate und hätte nach dieser Berechnung einen gemittelten Standardfehler von lediglich 1.5% der durchschnittlichen Reichweite, welcher in der Praxis einer akzeptablen Abweichung entspricht. Daraus kann in einem ersten Schritt geschlossen werden, dass ein vollständiger Austausch von GPS-Daten nicht zu großen Veränderungen in den Leistungswerten für zufällig verteilte Werbekampagnen führt.

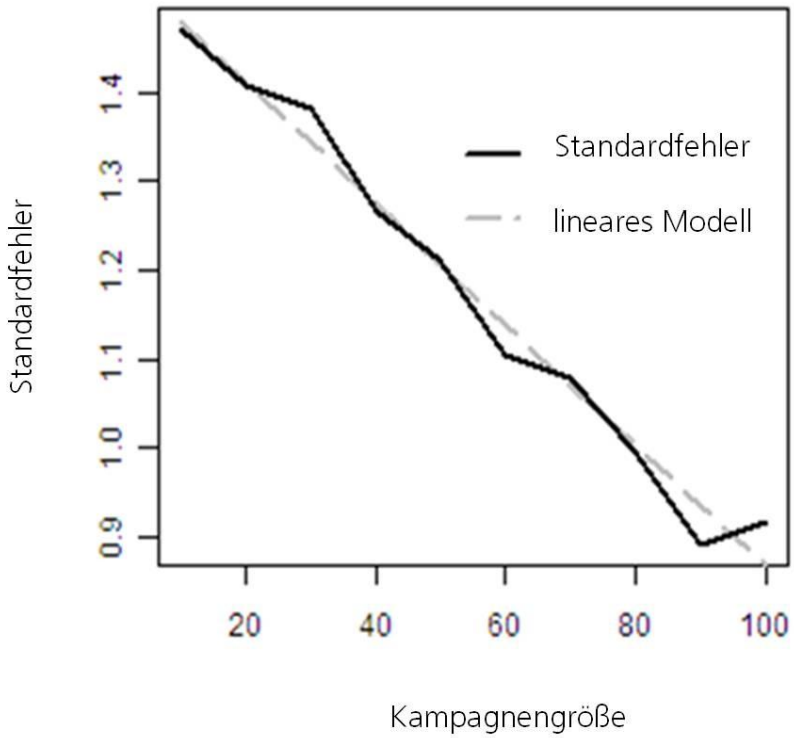



Abbildung 6.2: Durchschnittlicher Standardfehler für unterschiedliche Kampagnengrößen (Hecker et al. 2011a)

### 6.1.3 Subsampling und Bootstrap mit einer reduzierten Stichprobe

Im zweiten Experiment wird geprüft, ob ein GPS-Datensatz von kleinerem Umfang Veränderungen der Leistungswerte hervorruft. Dabei werden wiederum Kampagnengrößen zwischen 10 und 100 Plakaten gewählt (mit schrittweiser Erhöhung um 10) und die dazugehörigen Durchschnitte anhand von 15 unterschiedlichen Kampagnen pro Größe gemittelt. Zusätzlich werden Teilmengen von 2.5% bis 97.5% der GPS-Probanden gebildet, die jeweils dem entsprechenden Anteil von Probanden aus der Agglomeration Zürich entsprechen. Für jede Teilmenge wurden 15 unterschiedliche Datensätze von Probanden gewählt, um potenzielle Ausreißer zu glätten. Folglich sind die Ergebnisse Durchschnitte von 225 Werten (15*15), und zwar für jede Werbekampagne und Probandenteilmenge.

Im ersten Schritt wird der Fehler, welcher durch einen verkleinerten GPS-Datensatz verursacht wird, evaluiert und die Wurzel aus dem mittleren quadratischen Fehler berechnet. Hier sei nochmals darauf hingewiesen, dass in diesem Zusammenhang die Leistungswerte einer Probanden Teilmenge und nicht die Leistungswerte des vollständigen GPS-Datensatzes untersucht werden.

Die Ergebnisse der jeweiligen Probanden-Teilmenge sind in Tabelle 6.2 aufgelistet und in Abbildung 6.3 visualisiert. Die Wurzel aus dem mittleren quadratischen Fehler ist sowohl für kleinere Werbekampagnen als auch kleine Probanden-Teilmengen am höchsten, während sie für eine hohe Anzahl von Plakaten sowie großen Probanden-Teilmengen am kleinsten ist. Auch in diesem Beispiel erkennt man einen linearen Zusammenhang zwischen der Kampagnengröße und der Wurzel aus dem mittleren quadratischen Fehler. Bei kleineren Teilmengen, in diesem Zusammenhang bis etwa 40% der Gesamtstichprobe, steigt die Wurzel aus dem mittleren quadratischen Fehler exponentiell an. Übertragen in die Praxis bedeutet dieser Zusammenhang, dass der GPS-Datensatz auf etwa 50% seiner ursprünglichen Größe verkleinert werden kann, um weiterhin zuverlässige Werte bezüglich der Reichweite einer Kampagne zu erhalten. Zu beachten ist hierbei, dass dieses Experiment allerdings den Leistungswert einer Werbekampagne für die Gesamtheit der Bevölkerung und nicht für eine bestimmte soziodemographische Gruppe berechnet. Ist man jedoch daran interessiert, wie die Reichweite bei einer bestimmten Zielgruppe ausfällt, verkleinert sich der Datensatz zusätzlich. An dieser Stelle muss eine gesonderte Untersuchung für die kleinste Zielgruppe von Interesse durchgeführt werden. Sollten in diesem Zusammenhang ungleich verteilte soziodemographische Merkmale auftreten, kann eine geschichtete Stichprobe helfen, um mit großen Fehlerintervallen umzugehen.

| **Kampagnen Größe** | | **10** | **20** | **30** | **40** | **50** | **60** | **70** | **80** | **90** | **100** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Subsample Größe** | 2.5% | 8.7 | 8.2 | 8.7 | 7.9 | 6.9 | 6.9 | 5.4 | 6.1 | 5.7 | 4.9 |
| | 10.0% | 3.9 | 4.1 | 4.2 | 3.5 | 3.8 | 3.1 | 3.0 | 2.6 | 2.5 | 2.5 |
| | 25.0% | 2.3 | 2.3 | 2.3 | 2.1 | 2.0 | 1.8 | 1.7 | 1.8 | 1.4 | 1.5 |
| | 50.0% | 1.5 | 1.4 | 1.2 | 1.2 | 1.1 | 1.0 | 1.0 | 1.1 | 0.8 | 0.9 |
| | 75.0% | 1.0 | 1.0 | 0.9 | 0.9 | 0.9 | 0.9 | 0.8 | 0.9 | 0.7 | 0.7 |
| | 90.0% | 0.7 | 0.7 | 0.6 | 0.7 | 0.7 | 0.6 | 0.6 | 0.7 | 0.5 | 0.5 |
| | 97.5% | 0.7 | 0.6 | 0.5 | 0.6 | 0.6 | 0.5 | 0.5 | 0.7 | 0.4 | 0.4 |

Tabelle 6.2: Mittlerer quadratischer Fehler für eine reduzierte GPS-Stichprobe (Hecker et al. 2011a)

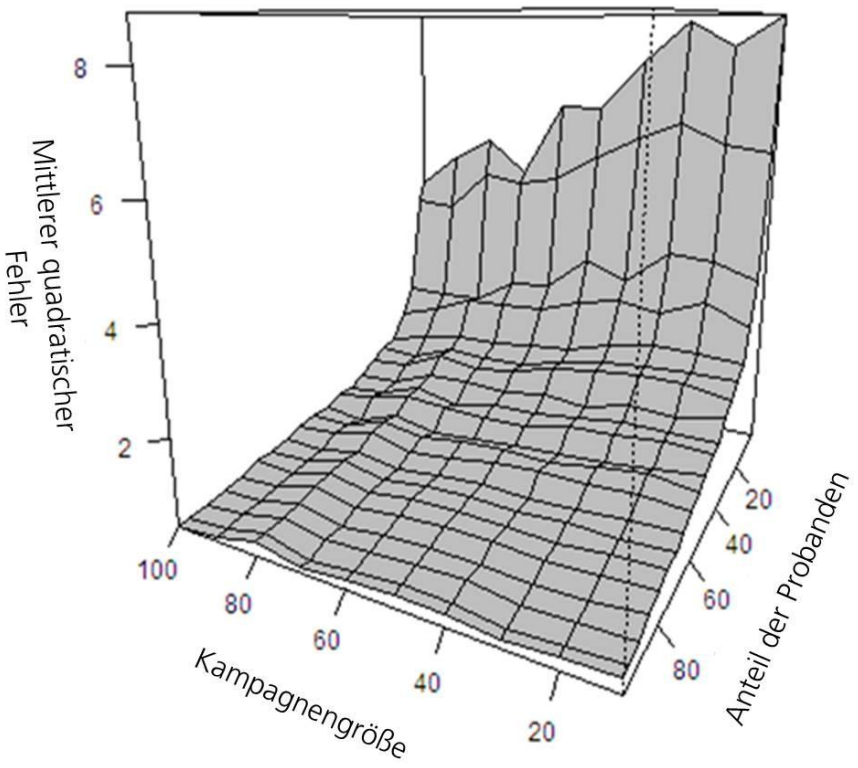



Abbildung 6.3: Mittlerer quadratischer Fehler für eine reduzierte GPS-Stichprobe (Hecker et al. 2011a)

Die durchgeführten Experimente beziehen sich bis jetzt ausschließlich auf Probanden-Teilmengen des GPS-Datensatzes und haben die Variabilität einer kleineren Stichprobengröße untersucht.

Es wurde jedoch noch nicht untersucht, welche Variation aufgrund einer anderen, neueren und kleineren GPS-Probe entsteht. Dies könnte ein evtl. zukünftiges Szenario sein, wenn in eine GPS-Stichprobe nacherhoben wird.

Für diesen Fall wird das beschriebene Bootstrap-Verfahren angewendet, welches mit 30 Wiederholungen Ergebnisse aus der Kombination von unterschiedlichen Kampagnengrößen (gemittelt über mehrere Werbekampagnen) und Probanden-Teilmengen (gemittelt über unterschiedliche Probanden Stichproben) liefert. Die Ergebnisse sind in Tabelle 6.3 aufgelistet und in Abbildung 6.4 abgebildet. Auffällig dabei ist, dass aufgrund der 30 Bootstrap Wiederholungen mit unterschiedlichen Datensätzen eine im Vergleich zu Abbildung 6.3 glattere Oberfläche hervorgerufen wird.

| Kampagnen Größe | | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Subsample Größe | 2.5% | 8.2 | 8.5 | 8.1 | 7.3 | 6.9 | 6.2 | 6.1 | 5.8 | 5.0 | 4.7 |
| | 10.0% | 4.0 | 4.2 | 4.0 | 3.7 | 3.5 | 3.2 | 3.0 | 2.8 | 2.7 | 2.5 |
| | 25.0% | 2.5 | 2.6 | 2.5 | 2.3 | 2.2 | 2.0 | 1.9 | 1.9 | 1.7 | 1.6 |
| | 50.0% | 1.8 | 1.9 | 1.7 | 1.6 | 1.6 | 1.4 | 1.4 | 1.3 | 1.2 | 1.1 |
| | 75.0% | 1.5 | 1.5 | 1.5 | 1.3 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| | 90.0% | 1.3 | 1.4 | 1.3 | 1.2 | 1.2 | 1.1 | 1.0 | 1.0 | 0.9 | 0.9 |
| | 97.5% | 1.3 | 1.3 | 1.3 | 1.2 | 1.1 | 1.0 | 1.0 | 1.0 | 0.9 | 0.8 |

Tabelle 6.3: Durchschnittlicher Standardfehler für neu generierte sowie kleinere Datensätze (Hecker et al. 2011a)

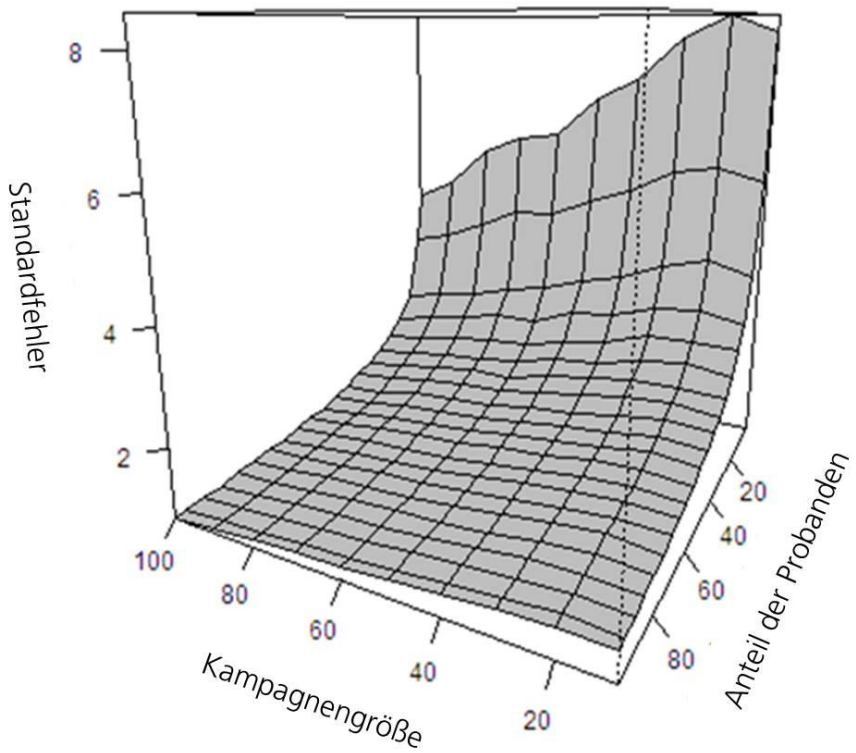



Abbildung 6.4: Durchschnittlicher Standardfehler für neu generierte sowie kleinere Datensätze (Hecker et al. 2011a)

Der Trend in den Ergebnissen ist vergleichbar mit den früheren Beobachtungen in den Experimenten. Der Standardfehler reagiert indirekt proportional zu der Größe der Probanden-Teilmenge und der Kampagnengröße. Ein Unterschied besteht lediglich in der Höhe des Fehlers. So erzielt der Standardfehler, berechnet durch das Bootstrap-Verfahren, bei einer Teilmenge von mehr als 25% einen höheren Wert als die Wurzel aus dem mittleren quadratischen Fehler im vorangehenden Experiment. Demzufolge kann man darauf schließen, dass unterschiedliche GPS-Datensätze zu einer erhöhten Varianz führen. Allerdings sind die Ergebnisse bei Probanden-Teilmengen kleiner 25% in beiden Fällen etwa gleich. Das heißt, dass beide Unterstichproben nahezu unabhängig von dem GPS-Datensatz sind. Insgesamt sind die Unterschiede der Standardfehler in beiden Beispielen aber eher gering. Daher kann geschlussfolgert werden, dass verkleinerte Stichproben und ein Austausch von GPS-Daten ohne nennenswerte Nachteile in der Berechnung von Leistungswerten von Werbekampagnen durchgeführt werden können.

Beide Experimente zeigen noch ein weiteres interessantes Ergebnis. Für eine kleinere Probandenanzahl sind die Auswirkungen einer großen Kampagne auf die Wurzel aus dem mittleren quadratischen Fehler oder Standardfehler größer als bei einer großen Anzahl von Testpersonen. Das ist ein willkommener Effekt in der Praxis, weil es bedeutet, dass im Falle von kleineren Stichproben größere Kampagnen den Fehler auf eine vernünftige Größe reduzieren könnten.

## 6.2 Zusammenfassung

Dieses Kapitel befasste sich mit der Robustheit der berechneten Ergebnisse. Während gegenwärtige Mobilitätsstudien sich nicht mit dem Aspekt der Stabilität von Ergebnissen beschäftigen, wurde in diesem Abschnitt a) die Auswirkung einer wiederholten Mobilitätsstudie und b) die Auswirkung einer Mobilitätsstudie kleinerer Größe untersucht. Die Ergebnisse zeigten zum einen, dass der berechnete Standardfehler vergleichbar klein ist und zum anderen, dass ein linearer Zusammenhang zwischen Standardfehler und der Größe der Werbekampagne existiert. Weiterhin wurde festgestellt, dass beim Verkleinern des Datensatzes der Standardfehler exponentiell ansteigt, wobei der Zuwachs des Standardfehlers bis zu einer Teilmenge von 40% linear verläuft. Daraus ergibt sich, dass ein Verkleinern der Stichprobe oder ein Austausch von GPS-Daten durchgeführt werden kann, ohne dass sich nennenswerte Veränderungen in der Leistungswertbestimmung von Werbekampagnen entwickeln. Weiterhin war zu erkennen, dass mit steigender Anzahl von Plakaten in einer Kampagne der Standardfehler und der mittlere quadratische Fehler geringer wurde. Das heißt, sollte in Zukunft eine GPS-Neuerhebung durchgeführt werden, kann dies in einem kleineren Umfang geschehen, allerdings nur, wenn die Kampagnengrößen einen bestimmten Schwellwert nicht unterschreiten. Für zukünftige empirische Erhebungen auch jenseits des Außenwerbekontextes können die Ergebnisse ein wertvoller Input sein.

# KAPITEL 7

## 7. AUSWEISUNG VON RÄUMLICH DIFFERENZIERTEN REICHWEITEN

In Kapitel 5 wurde der Berechnungsweg zur Leistungswertbestimmung von Plakatkampagnen und im anschließenden Kapitel 6 eine Robustheitsanalyse vorgestellt. In diesem Kapitel werden nun ausgewählte Ergebnisse der Reichweitenmodellierung für die Schweiz und Deutschland gezeigt (Abschnitt 7.1). Hierzu wird einerseits die Möglichkeit der räumlichen Streuung von Plakatkampagnen und andererseits die Zielgruppenausweisung nach Herkunftsgebiet präsentiert. Mit diesem Kapitel erhält der Leser einen Eindruck davon, welche neuen Möglichkeiten ein Mediaplaner in Zukunft durch die vorgestellte Methodik offeriert bekommt.

Das Kapitel 7 schließt mit einer Zusammenfassung der Ergebnisse (Abschnitt 7.2) und leitet dann über in das Diskussionskapitel.

### 7.1 Reichweiten und Passagenwertberechnung nach Zielgruppen, Herkunft und räumlicher Streuung

In diesem Abschnitt werden Beispielrechnungen für Plakatkampagnen vorgestellt, um zu zeigen, wie wichtig eine räumlich differenzierte Ausweisung von Leistungswerten für Plakatkampagnen ist. Die räumliche Differenzierung wird sowohl für die Wahl der Plakatstandorte als auch für die Auswahl der Zielgruppe vorgestellt (Hecker et al. 2010b). Als Testregionen werden die Stadt Bern und die Stadt Köln ausgewählt. Die Ergebnisse stellen einen Zugewinn in der Leistungsausweisung dar, der erst durch das GPS-mobilitätsdatenbasierte Verfahren ermöglicht wird.

#### 7.1.1 KAMPAGNENBERECHNUNG FÜR BERN

Die Anzahl der selektierten GPS-Probanden $(P)$ in den Experimenten für Bern beträgt 635. Alle erstellten Kampagnen beruhen auf real existierenden Plakaten der Schweizer Außenwerbung (Hecker et al. 2010b). Um die Auswertungen zu vereinfachen und eine bessere Vergleichsmöglichkeit zu bieten, wurden die Schweizer Wahrnehmbarkeitskriterien wie Beleuchtung, die Anzahl anderer Werbeträger im Umfeld des Plakates, etc. nicht berücksichtigt. Verglichen werden die reinen Passagen im individualisierten Sichtbarkeitsraum eines Plakates und die daraus resultierenden Reichweiten bei Kontaktklasse >0 nach Kaplan-Meier. Das Vorgehen stellt dabei das beschriebene Vorgehen aus Abschnitt 5.2.8 dar. In den Berechnungen wird die Gesamtzahl der Passagen, die durchschnittlichen Passagen pro Person sowie die Reichweite über einen Zeitraum von $t = 7$ Tagen bestimmt.

Das Experiment 1 basiert auf einer Kampagne $(K_G)$ mit 50 Plakaten, die über die gesamte Agglomeration gestreut sind. Diese Plakate werden einmal zufällig in zwei Kampagnen $(K_{R1}, K_{R2})$ mit jeweils 25 Plakaten (siehe Abbildung 7.1 links) und einmal in je eine Kampagne im Westen $(K_W)$ und Osten $(K_O)$ von Bern mit ebenfalls je 25 Plakaten aufgeteilt (siehe Abbildung 7.1 rechts). Die Leistungswerte der Kampagnen gemessen auf Basis aller Probanden in Bern sind in Tabelle 7.1 dargestellt.

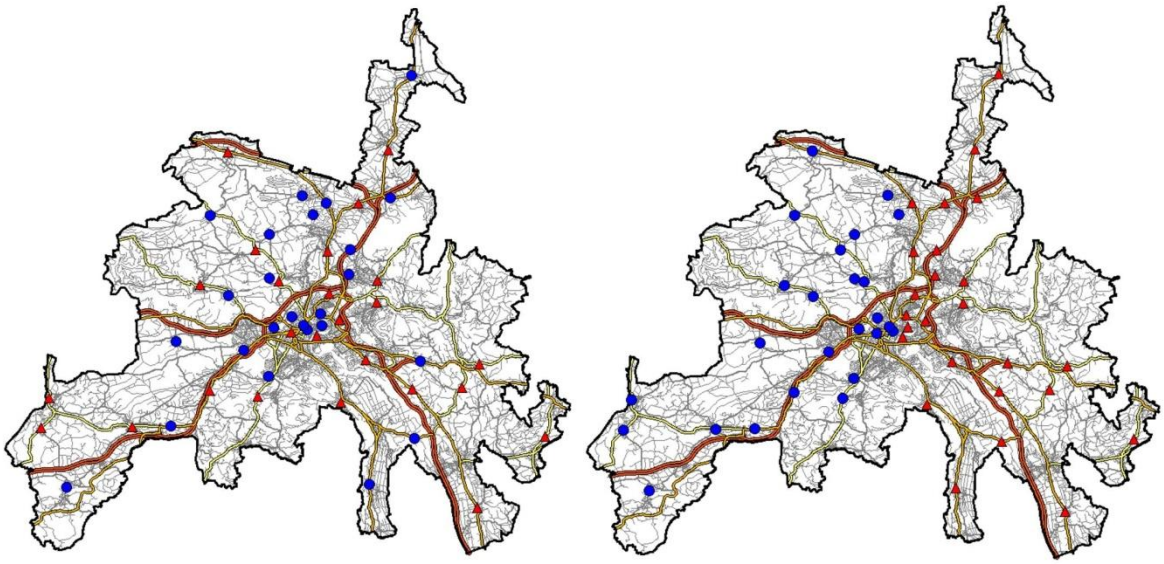


Abbildung 7.1: Zufällige (links) und räumliche orientierte (rechts) Aufteilung der Kampagne $(\mathbf{K_G})$ (Hecker et al. 2010b)

| # Probanden | # Plakate | ∑ Passagen | Ø Passagen (OTS) | Reichweite | Kampagne |
|---|---|---|---|---|---|
| \|P\|= 635 | 50 | 4730 | 7,4 | 0,861 | $(K_G)$ |
| \|P\|= 635 | 25 | 2417 | 3,8 | 0,712 | $(K_{R1})$ |
| \|P\|= 635 | 25 | 2313 | 3,6 | 0,663 | $(K_{R2})$ |
| \|P\|= 635 | 25 | 1866 | 2,9 | 0,499 | $(K_W)$ |
| \|P\|= 635 | 25 | 2864 | 4,5 | 0,676 | $(K_O)$ |

Tabelle 7.1: Leistungswerte für Berechnung 1 (Hecker et al. 2010b)

Insgesamt erzeugen die Probanden 4.730 Passagen mit allen 50 Plakaten $(K_G)$. Dies ergibt einen Durchschnitt von 7,4 Passagen pro Person. Allerdings erreicht die Kampagne nicht alle Probanden. Nur 86,1 % der Probanden kommen innerhalb von 7 Tagen durch den Sichtbarkeitsraum wenigstens eines Plakates. Vergleicht man diese Werte mit den Leistungswerten der beiden durch zufällige Aufteilung entstehenden Kampagnen $(K_{R1})$ und $(K_{R2})$, so erkennt man, dass sich die Gesamtzahl der Passagen relativ gleichmäßig auf die beiden Kampagnen verteilt. Auch die Reichweite ist bei beiden Kampagnen ähnlich und immer noch recht hoch. Dies liegt daran, dass beide Kampagnen gleichmäßig über den gesamten Raum verteilt sind und immer noch eine hohe Chance besteht, dass ein Proband ein Plakat passiert. Teilt man jedoch die Kampagne $(K_G)$ nach räumlichen Aspekten auf (hier z.B. durch eine Trennung West/Ost), kann es zu starken Unterschieden in den Leistungswerten kommen. Solche Effekte können z. B. durch die unterschiedliche Attraktivität von Stadtvierteln oder unterschiedliche Wohndichten entstehen. Anhand von Berechnung 1 kann man also sehen, dass bei gleichmäßig im Raum verteilten Kampagnen eine Mittelwertschätzung, wie z. B. durch das Copland Modell (vgl. Abschnitt 2.2.2), eine gute Annäherung an die tatsächlichen Leistungswerte ermöglicht. Erfolgt die Aufstellung einer Kampagne jedoch nach räumlichen Kriterien, so ist eine differenzierte Ausweisung der Reichweite notwendig, da es hier zu starken Schwankungen in den Leistungswerten kommen kann.

In Berechnung 2 werden zwei Kampagnen mit jeweils 10 Plakaten im Stadtzentrum von Bern auf ihre Leistungswerte untersucht. Die erste Kampagne $(K_{IG})$ ist über die gesamte Innenstadt gestreut während die zweite Kampagne um einen zentralen Platz in Bern konzentriert ist $(K_{IK})$. Beide Kampagnen sind in Abbildung 7.2 dargestellt.

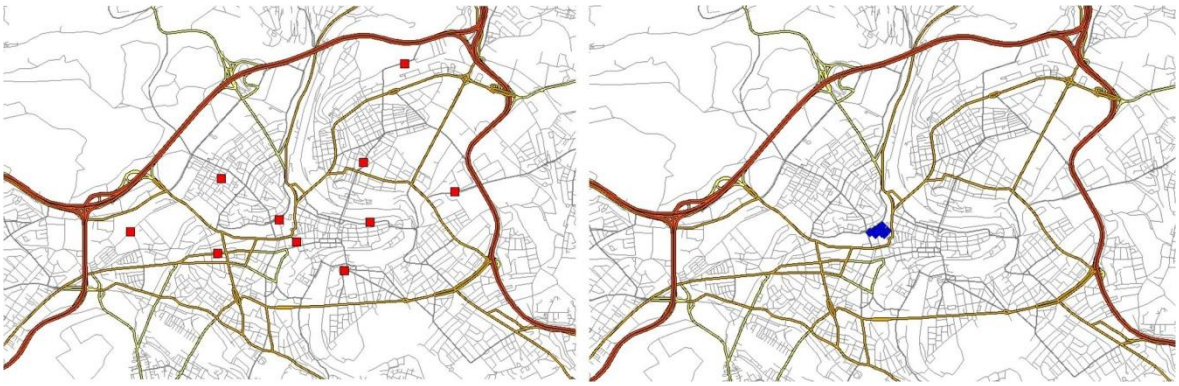


Abbildung 7.2: Gestreute Kampagne $(\mathbf{K_{IG}})$ (links) und geklumpte Kampagne $(\mathbf{K_{IK}})$ (rechts) in Bern (Hecker et al. 2010b)

Tabelle 7.2 enthält die Leistungswerte der beiden Kampagnen bezogen auf alle Probanden in Bern. Erwartungsgemäß ist die Reichweite der gestreuten Kampagne $(K_{IG})$ höher als die Reichweite der geklumpten Kampagne$(K_{IK})$. Jedoch ist die Anzahl der Plakatpassagen bei der gestreuten Kampagne $(K_{IG})$ wesentlich niedriger als bei der räumlich konzentrierten Kampagne$(K_{IK})$. Die Berechnung verdeutlicht also, dass eine hohe Reichweite nicht automatisch eine hohe Anzahl an Plakatpassagen bedeutet und dass umgekehrt eine hohe Anzahl an Passagen nicht unbedingt auf eine hohe Reichweite schließen lässt.

| # Probanden | # Plakate | ∑ Passagen | Ø Passagen (OTS) | Reichweite | Kampagne |
|---|---|---|---|---|---|
| \|P\|= 635 | 10 | 1314 | 2,1 | 0,504 | $(K_{IG})$ |
| \|P\|= 635 | 10 | 2704 | 4,3 | 0,243 | $(K_{IK})$ |

Tabelle 7.2: Leistungswerte für Berechnung 2 (Hecker et al. 2010b)

Dies steht im Gegensatz zu der in Kapitel 2 angegebenen Reichweitenformel (vgl. Copland Modell im Abschnitt 2.2.2), bei der die Reichweite mit zunehmendem AWert - unter Konstanz aller anderen Faktoren - steigt. Durch die fehlende räumliche Differenzierung des Copland Modells ergeben sich ungenaue Leistungswerte, d. h. bei räumlicher Konzentration von Plakaten zeigt sich der Vorteil von räumlich differenzierenden Verfahren zur Leistungswertbestimmung.

Im Folgenden werden die Auswirkungen der Auswahl von Zielgruppen nach geographischen Aspekten auf die Leistungswerte untersucht. Hierfür werden die Probanden aus Bern $(|P| = 635)$ anhand ihrer Wohnadresse in drei Gruppen aufgeteilt (vgl. Abbildung 7.3 links). Die Gruppe $P_C$ wohnt in der Innenstadt von Bern (blau), die Gruppe $P_W$ wohnt im Westen (grün) und die Gruppe $P_O$ wohnt im Osten von Bern (rot). Die Kampagne $(K_{SO})$ besteht aus 10 Plakaten und ist im Südosten von Bern positioniert (siehe Abbildung 7.3 rechts). Tabelle 7.3 zeigt die Ergebnisse der Berechnung.

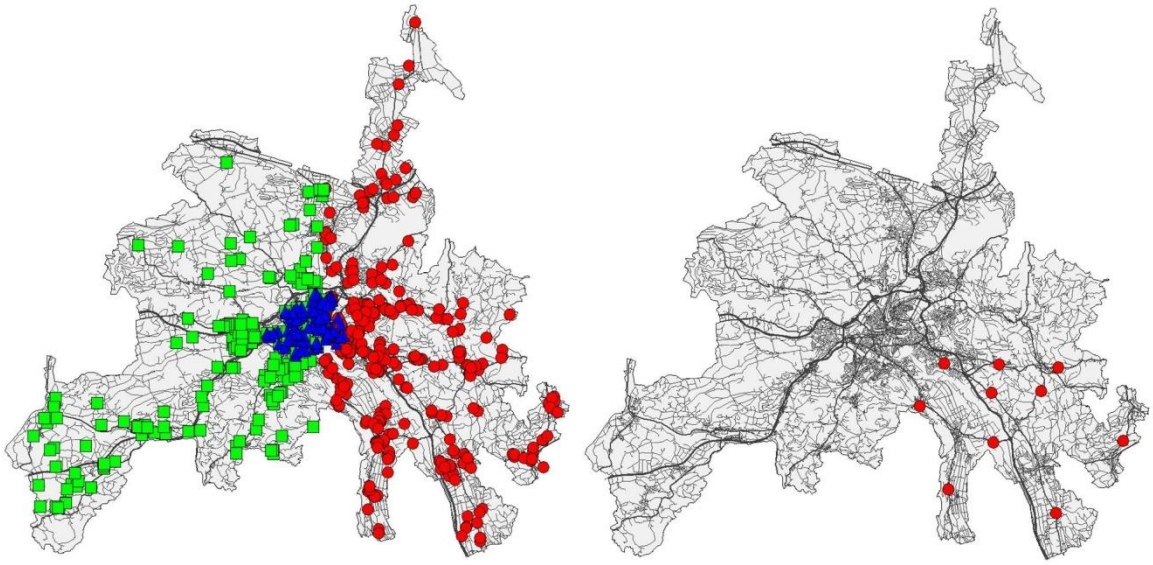


Abbildung 7.3: Gruppierung von Probanden (links) und eine Kampagne im Südosten Berns (rechts) (Hecker et al. 2010b)

| # Probanden | # Plakate | ∑ Passagen | Ø Passagen (OTS) | Reichweite | Kampagne |
|---|---|---|---|---|---|
| $\|P_C\|$ = 128 | 10 | 54 | 0,4 | 0,203 | $(K_{SO})$ |
| $\|P_W\|$ = 183 | 10 | 61 | 0,3 | 0,153 | $(K_{SO})$ |
| $\|P_O\|$ = 324 | 10 | 1388 | 4,3 | 0,562 | $(K_{SO})$ |

Tabelle 7.3: Leistungswerte für Berechnung 3 (Hecker et al. 2010b)

Die durchschnittliche Anzahl von Passagen, sowie die Reichweite von Probanden aus dem Osten von Bern sind am höchsten. Dies ist plausibel, da die Kampagne geographisch ungefähr die Hälfte des Wohngebietes dieser Zielgruppe abdeckt, sich mit den Wohngebieten der anderen Gruppen jedoch nicht überschneidet. Berechnungen mit geographisch eingeschränkten Zielgruppen sind besonders nützlich, um Leistungswerte getrennt für Bewohner einer Stadt und für Pendler zu unterscheiden oder um die Leistung einer Kampagne gezielt für einen bestimmten Raum zu bestimmen. Bei diesen Berechungen zeigt sich also, dass Verfahren zur räumlich differenzierten Ausweisung von Leistungswerten unabdingbar sind, um plausible Werte für individuell zusammengestellte Kampagnen und Zielgruppen auszuweisen.

### 7.1.2 KAMPAGNENBERECHNUNG FÜR DIE STADT KÖLN

Die Anzahl der selektierten GPS-Probanden ($|P|$) in den Kölner Berechnungen beträgt 266. Die Auswahl der Probanden wurde eingeschränkt auf die GPS-Erhebung und den angegebenen Wohnort Köln. Alle erstellten Kampagnen beruhen auch hier auf real existierenden Plakaten. Um die Auswertungen und die Interpretation der Ergebnisse zu vereinfachen, wurden Wahrnehmbarkeitskriterien nicht berücksichtigt. Das Vorgehen zur Leistungswertbestimmung folgt dabei dem im Abschnitt 5.2.7 vorgestellten Vorgehen, das heißt, zur Berechnung wird das vorgestellte System der Mobilitätseinheiten und Kaplan-Meier eingesetzt. Wie in den zuvor durchgeführten Berechnungen in der Schweiz werden für einen Zeitraum von 7 Tagen die Passagen, der OTS und die Reichweite bestimmt.

Die Berechnung 4 basiert auf 250 Plakaten $K_G$, die über das gesamte Stadtgebiet gestreut sind. Bei der Streuung wurde das Netz aus allen Werbeträgern in Köln gezogen. In der Abbildung 7.4 ist das Netz auf der linken Seite zu erkennen. Rechts daneben ist ein geklumptes Plakatnetz $K_{IK}$ zu erkennen. Dieses Plakatnetz ist rein auf die Stadtteile Kalk,

Buchheim, Mülheim und Holweide beschränkt und dient als Vergleichskampagne zum gestreuten Netz. Die Leistungswerte der Kampagnen sind in Tabelle 7.4 dargestellt.

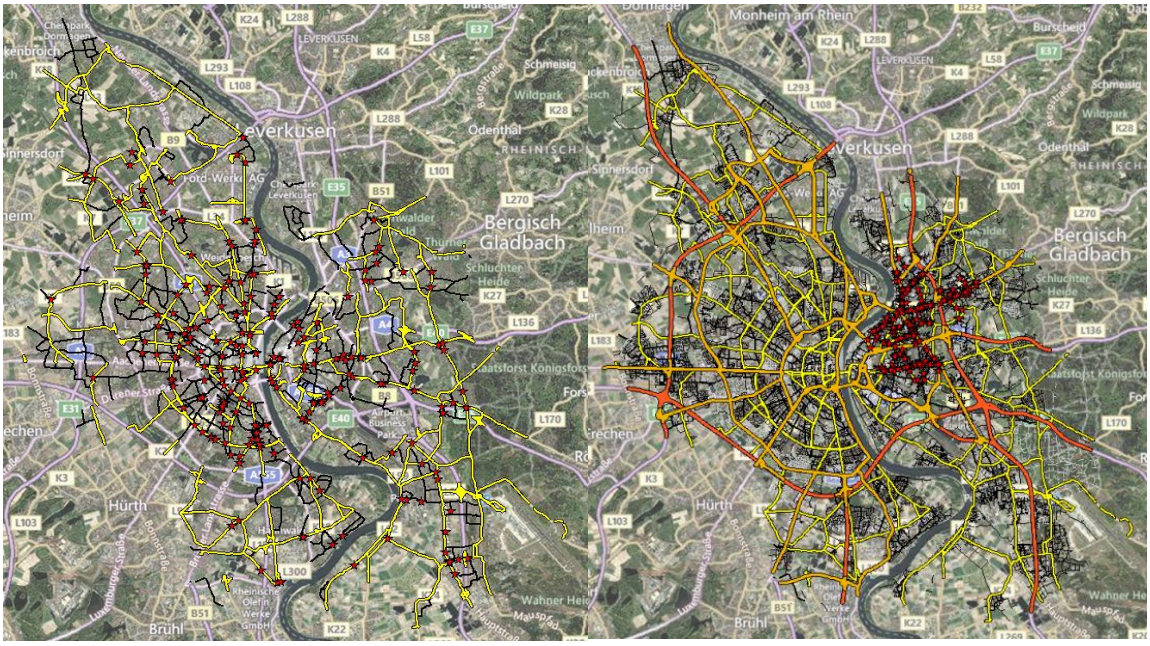



Abbildung 7.4: Gestreutes (links) und geklumptes (rechts) Plakatnetz in der Gemeinde Köln mit insgesamt 250 Plakatstellen (topographische Hintergrundinformation Google 2012)

| # Probanden | # Plakate | ∑ Passagen | Ø Passagen (OTS) | Reichweite | Kampagne |
|---|---|---|---|---|---|
| \|P\|= 266 | 250 | 1691 | 6,9 | 0,92 | $K_G$ |
| \|P\|= 266 | 250 | 1754 | 14,3 | 0,45 | $K_{IK}$ |

Tabelle 7.4: Leistungswerte für Berechnung 4

Erwartungsgemäß ist auch hier wie in der Agglomeration Bern die Reichweite der gestreuten Kampagne $K_G$ höher als die Reichweite der geklumpten Kampagne $K_{IK}$. Die Anzahl der Gesamtpassagen ist bei der konzentrierten Kampagne an dieser Stelle sogar leicht höher. Dies liegt insbesondere daran, dass die gewählten Plakatstandorte bei der geklumpten Kampagnenauswahl an wichtigen Zubringerstraßen zum Zentrum der Kölner Innenstadt platziert sind und so häufig Kontakte von den gleichen Personen über die 7 Tage hinweg einsammeln. Dies kann man auch daran erkennen, dass der OTS Wert sehr hoch im Vergleich zur gestreuten Kampagne ist.

In der nächsten Berechnung 5 werden zwei Plakatkampagnen miteinander verglichen, die über das gesamte Kölner Stadtgebiet räumlich verteilt sind. Der einzige Unterschied bei diesen beiden Kampagnen ist, dass die NavTeq Straßenkategorie zur Auswahl der Plakate mit herangezogen wird. Bei der ersten Kampagne $K_E$ wurden nur Plakate ausgewählt, die an der NavTeq Straßenkategorie 2 stehen. Diese Straßen stellen Ein- und Ausfallstraßen in das Kölner Stadtgebiet dar. Für die zweite Kampagne $K_H$ wurden Straßensegmente der Kategorie 3 und 4 ausgesucht. Dies sind in der Regel Haupt- und Geschäftsstraßen. In Abbildung 7.5 sind beide Plakatkampagnen und ihr dazugehöriges Straßennetz dargestellt.

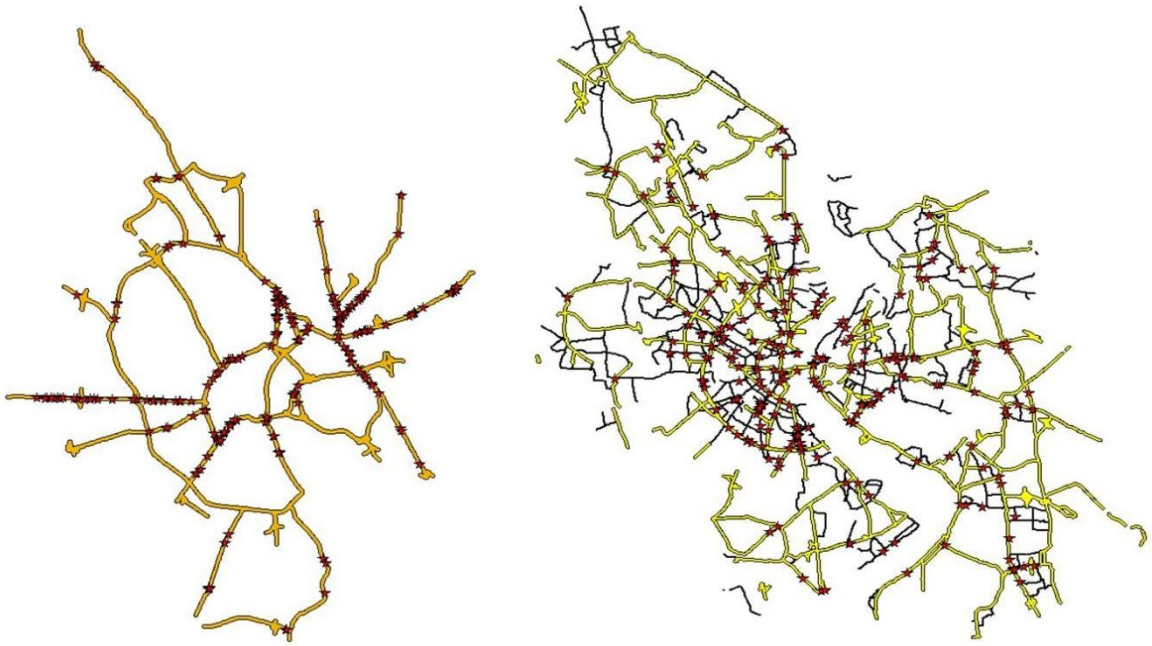


Abbildung 7.5: Plakatnetz für Ein- und Ausfallstraßen (links) und Haupt- und Geschäftsstraßen (rechts) in Köln mit insgesamt 250 Plakatstellen

| # Probanden | # Plakate | ∑ Passagen | Ø Passagen (OTS) | Reichweite | Kampagne |
|---|---|---|---|---|---|
| \|P\|= 266 | 250 | 1666 | 7,9 | 0,79 | $K_E$ |
| \|P\|= 266 | 250 | 1698 | 7,08 | 0,90 | $K_H$ |

Tabelle 7.5: Leistungswerte für Berechnung 5

Die Tabelle 7.5 enthält die Leistungswerte der beiden Kampagnen. Die Reichweite der Kampagne $K_E$ in den Ein- und Ausfallstraßen ist niedriger als die Reichweite der Kampagne $K_H$ in den Haupt- und Geschäftsstraßen. Dieser Effekt konnte erwartet werden. Sobald die Probanden ihre Wohnquartiere verlassen, nutzen sie die Haupt- und Geschäftsstraßen, um z.B. Einkäufe zu tätigen. Die Straßen der Kategorie 2 und ihrer dazugehörigen Plakatstellen werden dagegen häufiger besucht, wenn weiter entfernte Ziele angesteuert werden.

In Berechnung 6 werden die zuvor vorgestellten Plakatkampagnen $K_H$ $K_E$ hinsichtlich soziodemographischer Variablen und einer räumlichen Probandenaufteilung miteinander verglichen. Hierzu werden GPS-Personen in eine Zentrums-, in eine rechtsrheinische- und linksrheinische Population (Abbildung 7.6) aufgeteilt. Zusätzlich zu dieser räumlichen Aufteilung der Probanden werden soziodemographische Variablen untersucht. Von den 266 Probanden sind 135 männlich und 131 weiblich und teilen sich auf Altersklassen wie folgt auf: 14-29 Jahre: 48 Probanden, 30-49 Jahre: 99 Probanden und ab 50 Jahren: 119 Probanden. Diese werden nun hinsichtlich ihres Leistungswertbeitrages untersucht.

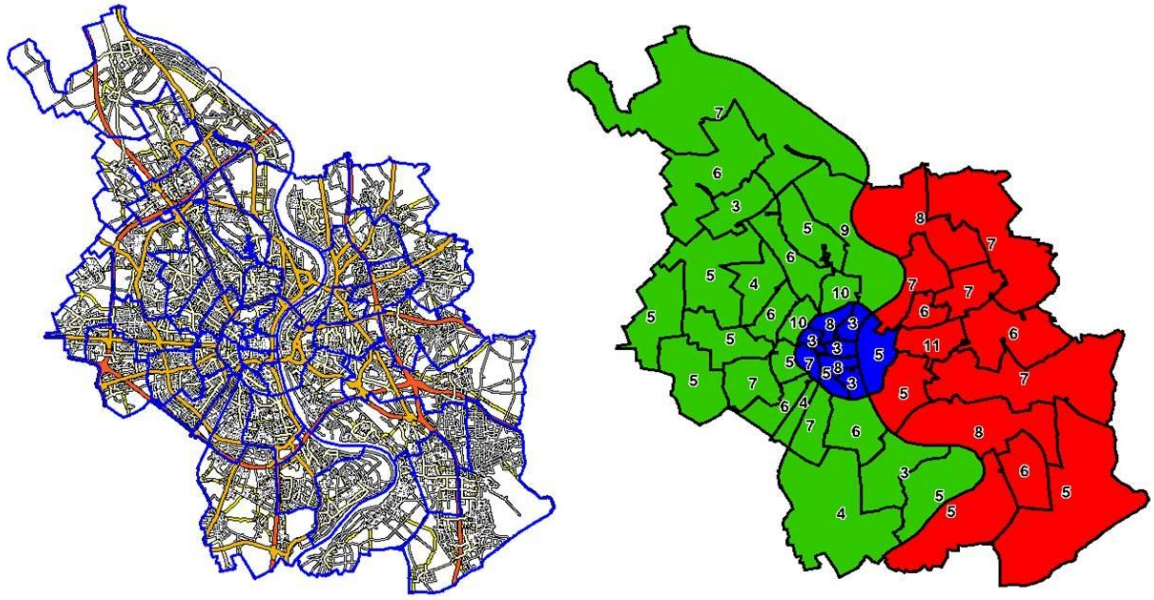


Abbildung 7.6: Probandenverteilung nach Herkunftsgebieten: (links) PLZ-Gebiete in Köln; (rechts) Aufteilung der Probanden nach PLZ Grün: linksrheinisch, Blau: Zentrum, Rot: rechtsrheinisch

In Tabelle 7.6 sind die jeweiligen Ergebnisse zu der räumlichen Verteilung und den soziodemographischen Variablen dargestellt. Auffällig ist, dass im Vergleich der Reichweiten über alle Kampagnen hinweg die Reichweite bei Männern leicht höher ist als bei Frauen. Dies ist darüber zu begründen, dass Männer in der Regel mobiler sind als Frauen. Dieser Unterschied in den Leistungswerten, bzw. in der Mobilität, wird z.B. auch durch die Studie Mobilität in Deutschland aus dem Jahre 2010 belegt (MID 2008a, vgl. Abschnitt 3.3). Ein weiterer interessanter Unterschied fällt bei der Probandenverteilung auf. Die Probanden, die aus dem Zentrum kommen, erzielen über alle soziodemographischen Gruppen hinweg die höchsten Reichweiten. Soziodemographische Unterschiede sind bei den Altersklassen zu beobachten. Zu den für die Leistungswerte kontaktstärksten Zielgruppen gehören die Altersklassen 14-29 und 30-49 Jahre. Beide zeigen in ihren Leistungswerten deutlich höhere Werte als die Klasse der ab 50 jährigen. Es scheint so, als wären diese beiden Altersklassen mobiler. Auffällige Unterschiede bei den Herkunftsgebieten der Probanden und Unterschiede bei den gewählten Netzen fallen bei den Probanden aus dem Zentrum von Köln auf. Das Netz der Haupt- und Geschäftsstraßen erzielt bei diesen Probanden über Geschlecht und Altersgruppen hinweg die höchsten Reichweitenwerte.

| # NETZ | # Gebiet | # Schicht | # Probanden | ∑ Passagen | Ø Passagen (OTS) | Reichweite | Kampagne |
|---|---|---|---|---|---|---|---|
| **Ausfall** | linksrh. | 14-29 JAHRE | 22 | 127 | 6,7 | 0,865 | $K_E$ |
| **Ausfall** | linksrh. | 30-49 JAHRE | 50 | 310 | 7,3 | 0,845 | $K_E$ |
| **Ausfall** | linksrh. | 50-.. JAHRE | 61 | 307 | 7,1 | 0,709 | $K_E$ |
| **Ausfall** | linksrh. | GESAMT | 133 | 744 | 7,1 | 0,786 | $K_E$ |
| **Ausfall** | linksrh. | männlich | 57 | 329 | 7,4 | 0,782 | $K_E$ |
| **Ausfall** | linksrh. | weiblich | 76 | 415 | 6,9 | 0,789 | $K_E$ |
| **Ausfall** | rechtsrh. | 14-29 JAHRE | 17 | 88 | 8,1 | 0,639 | $K_E$ |
| **Ausfall** | rechtsrh. | 30-49 JAHRE | 28 | 241 | 9,8 | 0,876 | $K_E$ |
| **Ausfall** | rechtsrh. | 50-.. JAHRE | 43 | 350 | 10,4 | 0,782 | $K_E$ |
| **Ausfall** | rechtsrh. | GESAMT | 88 | 680 | 9,8 | 0,784 | $K_E$ |
| **Ausfall** | rechtsrh. | männlich | 47 | 372 | 10,4 | 0,761 | $K_E$ |
| **Ausfall** | rechtsrh. | weiblich | 41 | 308 | 9,3 | 0,811 | $K_E$ |
| **Ausfall** | Zentrum | 14-29 JAHRE | 9 | 60 | 6,6 | 1,000 | $K_E$ |
| **Ausfall** | Zentrum | 30-49 JAHRE | 21 | 100 | 5,3 | 0,901 | $K_E$ |
| **Ausfall** | Zentrum | 50-.. JAHRE | 15 | 96 | 8,7 | 0,731 | $K_E$ |
| **Ausfall** | Zentrum | GESAMT | 45 | 256 | 6,6 | 0,864 | $K_E$ |
| **Ausfall** | Zentrum | männlich | 31 | 187 | 7,0 | 0,866 | $K_E$ |
| **Ausfall** | Zentrum | weiblich | 14 | 69 | 5,7 | 0,860 | $K_E$ |

| # NETZ | # Gebiet | # Schicht | # Probanden | ∑ Passagen | Ø Passagen (OTS) | Reichweite | Kampagne |
|---|---|---|---|---|---|---|---|
| **Hauptst.** | linksrh. | 14-29 JAHRE | 22 | 124 | 6,4 | 0,877 | $K_H$ |
| **Hauptst** | linksrh. | 30-49 JAHRE | 50 | 352 | 7,9 | 0,889 | $K_H$ |
| **Hauptst** | linksrh. | 50-.. JAHRE | 61 | 348 | 6,6 | 0,863 | $K_H$ |
| **Hauptst** | linksrh. | GESAMT | 133 | 823 | 7,1 | 0,875 | $K_H$ |
| **Hauptst** | linksrh. | männlich | 57 | 393 | 7,6 | 0,902 | $K_H$ |
| **Hauptst** | linksrh. | weiblich | 76 | 430 | 6,6 | 0,856 | $K_H$ |
| **Hauptst** | rechtsrh. | 14-29 JAHRE | 17 | 129 | 7,9 | 0,961 | $K_H$ |
| **Hauptst** | rechtsrh. | 30-49 JAHRE | 28 | 167 | 6,4 | 0,929 | $K_H$ |
| **Hauptst** | rechtsrh. | 50-.. JAHRE | 43 | 249 | 6,3 | 0,916 | $K_H$ |
| **Hauptst** | rechtsrh. | GESAMT | 88 | 545 | 6,7 | 0,929 | $K_H$ |
| **Hauptst** | rechtsrh. | männlich | 47 | 330 | 7,7 | 0,909 | $K_H$ |
| **Hauptst** | rechtsrh. | weiblich | 41 | 215 | 5,5 | 0,951 | $K_H$ |
| **Hauptst** | Zentrum | 14-29 JAHRE | 9 | 73 | 8,6 | 0,941 | $K_H$ |
| **Hauptst** | Zentrum | 30-49 JAHRE | 21 | 160 | 7,6 | 0,995 | $K_H$ |
| **Hauptst** | Zentrum | 50-.. JAHRE | 15 | 100 | 7,8 | 0,861 | $K_H$ |
| **Hauptst** | Zentrum | GESAMT | 45 | 332 | 7,9 | 0,940 | $K_H$ |
| **Hauptst** | Zentrum | männlich | 31 | 239 | 8,4 | 0,915 | $K_H$ |
| **Hauptst** | Zentrum | weiblich | 14 | 93 | 6,7 | 0,994 | $K_H$ |

Tabelle 7.6: Leistungswerte für Berechnung 6

In der abschließenden Berechnung 7 werden die zuvor vorgestellten Plakatkampagnen hinsichtlich ihrer Kontaktklassenentwicklung und nach Probandenherkunftsgebiet miteinander verglichen. Untersucht wird die erzielte Kontaktklasse am Tag 7. In der Abbildung 7.7 sind jeweils vier Grafiken zu erkennen. Die X-Achse gibt die erzielte Reichweite wieder und die Y-Achse die Kontaktklasse 1-5. In Grafik 1 sind für alle 266 Kölner Probanden die Leistungswerte nach 7 Tagen für die ersten 5 Kontaktklassen und für die 4 vorgestellten Kampagnen abgetragen (Streuung, Klumpung, Haupt/Geschäftsstraßen und Ausfallsstraßen). Man erkennt, dass die gestreute Plakatkampagne mit ihren 250 Plakaten über alle gewählten Kontaktklassen die reichweitenstärkste ist. Die Leistungswerte nehmen wie zu erwarten kontinuierlich mit steigender Kontaktklasse ab. Die Reichweite beginnt bei Kontaktklasse 1 bei 95% und fällt dann bei Kontaktklasse 5 bis auf 59% Reichweite. Die Reichweiten für die Haupt-/Geschäftsstraßen und Ausfallstraßen sind zu Beginn bei Kontaktklasse 1 noch in den Werten weiter auseinander, während sie bei Kontaktklasse 5 fast gleichauf liegen. Die Plakatkampagne, die geklumpt auf der rechtsrheinischen Seite ausgewählt worden ist, besitzt auch den niedrigsten Reichweitenwert.

In den drei folgenden Grafiken der Abbildung 7.7 sind, wie in Berechnung 6, die Probanden auf drei Kölner Regionen Zentrum, rechtsrheinisch und linksrheinisch aufgeteilt. Bei der Darstellung der Probanden aus dem Zentrum fällt auf, dass kaum ein Unterschied zwischen der gestreuten und der Kampagne in Haupt-/Geschäftsstraßen existiert. Die Kontaktklassenwerte liegen hier sehr eng beieinander.

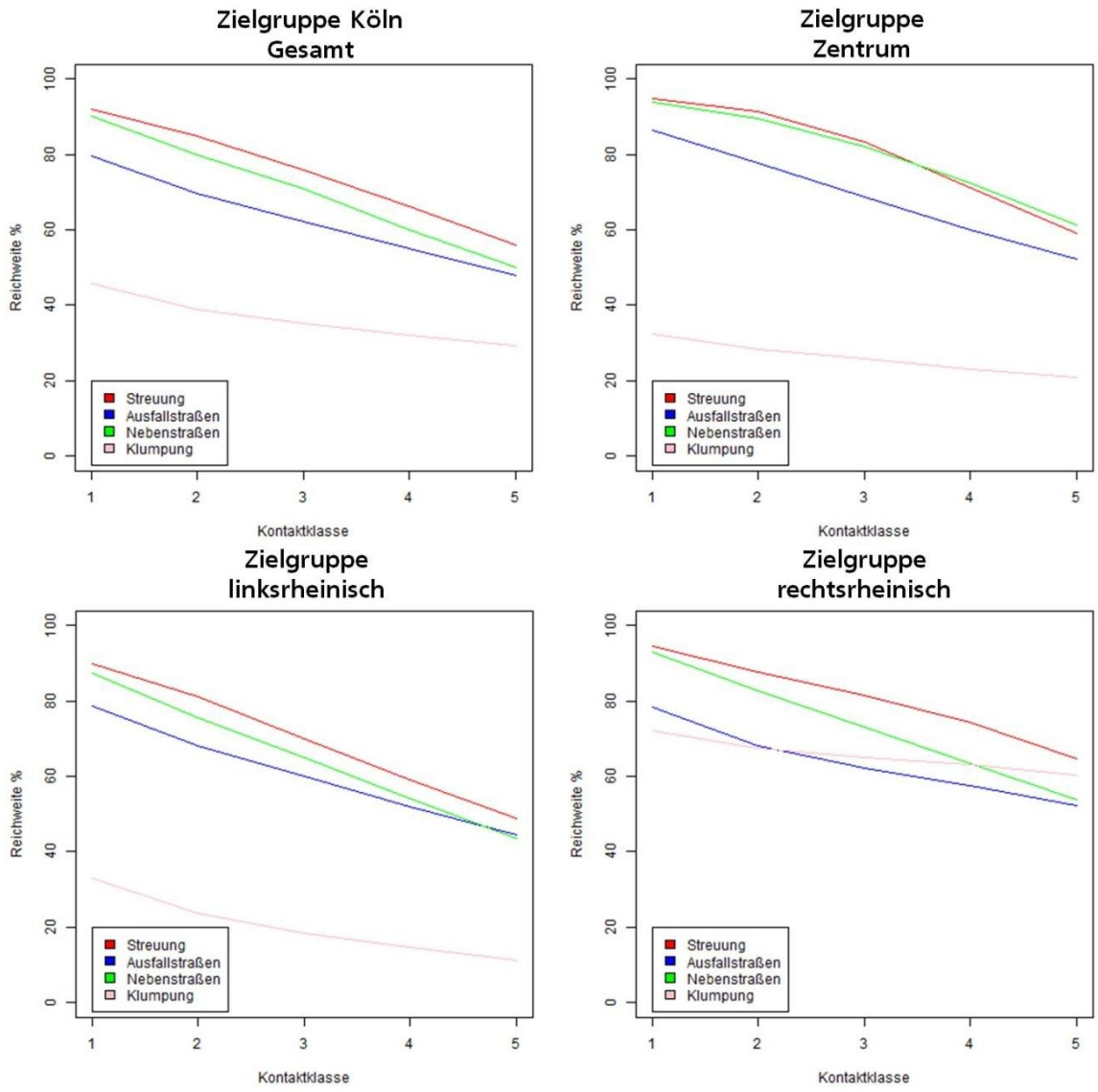



Abbildung 7.7: Kontaktklassen 1-5 für unterschiedliche Kampagnen und Probandenherkunftsgebiet

Dies lässt sich dadurch erklären, dass viele der zufällig ausgewählten Plakatstellen in Nebenstraßen im Innenstadtbezirk liegen und somit eine hohe Wahrscheinlichkeit besitzen, von den Anwohnern gesehen zu werden. Bei der Personenauswahl des linksrheinischen Köln liegen die Kampagnen Ausfall-, Haupt-/Geschäftsstraßen und Streuung sehr dicht beieinander, wobei die gestreute Kampagne immer noch die reichweitenstärkste ist. Die geklumpte Plakatkampagne auf der rechtsrheinischen Seite sackt bei Kontaktklasse 5 auf einen Wert von unter 20% ab. Das bedeutet, dass sehr wenig linksrheinische Probanden in das rechtsrheinische Gebiet kommen. Zum Abschluss dieser Auswertung werden die rechtsrheinischen Probanden noch auf ihren Reichweitenbeitrag bei den 4 ausgewählten Netzen untersucht. Der positive Sprung der geklumpten rechtsrheinischen Plakatkampagne fällt sofort auf. Die Reichweite liegt bei über 75% und nimmt auch über die Kontaktklassen weit weniger ab als die drei anderen Kampagnen. Dies kann dadurch erklärt werden, dass Probanden aus dem rechtsrheinischen Köln auch eine weit höhere Kontaktchance haben, häufige Kontakte mit ihrem sehr lokalen Netz zu erzielen, als Probanden aus dem Zentrum oder dem linksrheinischen Köln. Eine weitere Besonderheit ist bei den Haupt-/Geschäftsstraßen zu erkennen. Zu Beginn ist die Kampagne in der Reichweite noch gleichauf

mit der gestreuten Kampagne, nimmt dann aber mit steigender Kontaktklasse viel stärker ab. Dies bedeutet, dass z.B. als Werbekampagne eines neuen, bisher unbekannten Produktes, wie sie in Abschnitt 5.1.3 vorgestellt worden ist, die gestreute Kampagne die beste Wahl wäre, während man für ein bereits bekanntes Produkt eine Plakatkampagne in Nebenstraßen wählen könnte, da sie keine so hohe Kontaktklasse braucht.

Auch bei diesen Berechnungen in Deutschland zeigt sich, dass die Verfahren zur räumlich differenzierten Ausweisung von Leistungswerten, Werte für individuell zusammengestellte Kampagnen und Zielgruppen ausweisen und dies zu sehr unterschiedlichen Reichweitenwerten führen kann. Dies gilt sowohl für unterschiedliche soziodemographische Gruppen als auch für Kontaktklassen.

## 7.2 Zusammenfassung

In diesem Kapitel wurde auf Basis von realen Plakat- und Mobilitätsdaten die vorgestellte Modellierung aus Kapitel 5 für unterschiedliche Kampagnenkonfigurationen umgesetzt. In ersten Abschnitt 7.1 wurden für die Schweiz und Deutschland ausgewählte Plakatkampagnen für die Städte Bern und Köln berechnet. Dabei wurde untersucht, wie sich für beide Städte eine unterschiedliche räumliche Verteilung der Kampagnen auf die Leistungswerte auswirkt. In einer zweiten Untersuchung wurden für beide Städte noch Berechnungen nach Herkunft, soziodemographischen Variablen der Probanden und Kontaktklassen durchgeführt. Alle Berechnungen zeigten plausible und nachvollziehbare Ergebnisse. Obwohl die Modellierung in Bezug zu der bisherigen Modellierung in beiden Ländern massiv an Komplexität hinzugewinnt, zeigen die Beispielrechnungen, wie wichtig räumlich differenzierende Verfahren für eine sachgerechte Ausweisung von Leistungswerten sind. Die vorgestellten Kampagnen stellen typische Anwendungsbeispiele von Mediaplanern dar und können nur aufgrund der neuen Methodik so in die Planung einfließen.

# KAPITEL 8

## 8. ZUSAMMENFASSUNG UND DISKUSSION

In diesem Kapitel werden die Ergebnisse der Reichweitenmodellierung zusammengefasst, bewertet und die Vor- und Nachteile diskutiert. Die Diskussion erfolgt in der Reihenfolge der methodischen Teilschritte, die in dieser Arbeit vorgestellt worden sind (Abschnitt 8.1). Die verwandten Publikationen zur Außenwerbeforschung (Pasquier 1997, Engel und Hofsäss 2003) stammen noch aus der Zeit vor der Mobilitätserfassung mit GPS sowie dem Einsatz von Geographischen Informationssystemen. Diese Arbeiten betrachten keine geographischen Aspekte, die mit dem neuen Ansatz in den Fokus der Leistungsbewertung rücken. Somit ist ein direkter Vergleich nicht sinnvoll (vgl. Kapitel 2.3). Der zweite Abschnitt 8.2 stellt sich der Frage, wie zukunftsfähig die vorgestellte Systematik zur Reichweitenberechnung ist und welche weiteren Perspektiven für den Anwendungsbereich der Außenwerbung existieren.

### 8.1 Diskussion zu den vorgestellten technischen und methodischen Ansätzen

Die Außenwerbung steht europaweit vor einem Umbruch in der Leistungswertberechnung. Neue Techniken der Mobilitätserfassung und die digitale Verfügbarkeit von Plakatstandortdaten machen es möglich, Leistungswerte individuell für beliebige Kampagnen und Zielgruppen zu bestimmen. Auf Basis von real erfassten GPS-Daten in Deutschland und der Schweiz wurden einzelne Modellierungsschritte vorgestellt, diese werden nun im Folgenden diskutiert. Die Diskussion wird chronologisch nach den einzelnen methodischen Teilschritten durchgeführt. Es werden für jeden einzelnen Teilschritt nochmals die wichtigsten Fakten zusammengefasst, im Anschluss die jeweilige Limitation des Ansatzes diskutiert und etwaige Handlungsalternativen vorgestellt. Pro Teilschritt werden die wichtigsten Ergebnisse in einer abschließenden Tabelle festgehalten. In Abbildung 8.1 sind das Vorgehen und die Teilschritte der Arbeit nochmals graphisch dargestellt.

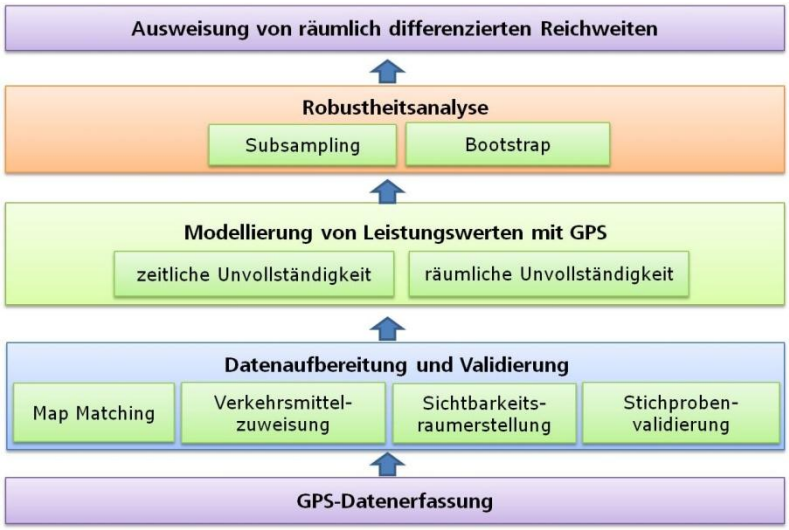



Abbildung 8.1: Darstellung der methodischen Teilschritte zur Reichweitenberechnung

**Erfassung, Datenaufbereitung und Validierung**

Eine der grundlegenden Ideen der neuen Mobilitätserfassungsmethodik ist es, nicht mehr rein auf die Erinnerung von Probanden per Telefoninterview zu setzen, sondern deren Mobilität mittels GPS längere Zeit zu erfassen und aufzuzeichnen. Unschärfen in der subjektiven Erinnerung der zurückgelegten Mobilität werden durch dieses Vorgehen ausgeschlossen. Zusätzlich wird nunmehr im Vergleich zu den bisherigen Telefonbefragungen nicht mehr nur ein einzelner Tag erfasst und mit diesem Tag hochgerechnet, sondern es werden insgesamt 7 Tage erfasst. Die reine Datenerfassung ist somit umfassender, genauer und direkter im Vergleich zu den bisher eingesetzten Telefonbefragungen. Ein Nachteil ist allerdings, dass die Erhebung teurer und die anschließende Aufbereitung und Modellierung der Daten deutlich komplexer ist. Zudem wird der Mensch bei der GPS-basierten Methode als Wahrnehmer der Werbung ausgeblendet. Wie in Abschnitt 2.1 vorgestellt, ist die Bestimmung des Werbekontaktes von vielen einzelnen Faktoren (Winkel des Plakates zum Betrachter, Umfeldkomplexität, Entfernung, etc.) abhängig und stellt den qualitativen Bereich der Außenwerbeforschung dar. Dieses gilt es jedoch in dieser Arbeit nicht zu untersuchen, stellt jedoch eine wichtige Determinante in der Leistungswerteberechnung dar.

Zu den ersten Problematiken, denen sich die Arbeit gestellt hat, gehören der Umgang mit Fehlverortungen, Oszillationen und dem Fehlen von einzelnen Teilstrecken eines aufgezeichneten GPS-Weges. Hierzu wurde in der Arbeit ein Map Matching Verfahren vorgestellt, das die GPS-Daten mit digitalen Straßennetzen (NavTeq, Vector25) zusammenführt. Die Herausforderung beim Map Matching ist es, homologe Objekte bei den z.T. heterogenen Datenquellen zu finden. So existieren unterschiedliche Lagebeziehungen, Datenlücken und Klassifikationen in den jeweiligen Datensätzen. Die Idee des vorgestellten Map Matching ist es, für jeden individuellen GPS-Punkt eine Menge an möglichen Straßensegmenten (Kandidaten) zu bestimmen. Das anschließende Ziel des Verfahrens ist es, einen kürzesten Weg durch das Straßennetz zu bestimmen, der die Kandidaten eines Weges besucht. An dieser Stelle werden für die Auswahl der Kandidaten und die Bestimmung des kürzesten Weges bestimmte Annahmen und Konventionen getroffen. So wird für die Kandidatensuche ein Puffer von 15 Metern Größe gewählt. Zukünftig wäre es vorstellbar, Puffer in Abhängigkeit von der aufgezeichneten GPS-Erfassungsqualität zu justieren. Der Vorteil wäre, dass über diesen zusätzlichen Dateninput die Kandidaten evtl. reduziert oder erweitert werden können, so der mögliche Suchraum eingeschränkt bzw. ausgedehnt und die Vollständigkeit möglicher Kandidaten innerhalb vorgegebener Wahrscheinlichkeitsschranken erhöht wird. Ein weiterer Punkt ist das Map Matching an sich. Nach Auswahl der Kandidaten wird ein Optionsgraph erstellt, der über die Erstellung des kürzesten Wegebaumes die Verknüpfung der einzelnen Straßensegmente zu einem Weg herstellt (Lou 2009 und Marchal et al. 2006). Dies kann als Anwendung von Ockhams Razor oder als Verhalten des Homo oeconomicus betrachtet werden, die voraussetzen, dass der kürzeste Weg über die selektierten Kandidaten der wahrscheinlichste ist. Der kürzeste Weg muss an dieser Stelle jedoch nicht immer der Richtige sein. Liegen GPS-Daten in einer niedrigen zeitlichen Auflösung vor, stößt das Map Matching an seine Grenzen und es müssen alternative Modellierungsschritte gefunden werden. Ein guter Weg, um die Anzahl der möglichen Optionen für die Erstellung der kürzesten Wegebäume einzugrenzen, ist eine möglichst exakte sekündliche Positionierung der GPS-Signale sowie eine ständige Aktualisierung des Satellitenalmanach (schnellere und bessere Positionierung). Eine weitere Möglichkeit besteht darin, die Verbindung zwischen dem ersten und dem letzten GPS-Punkt zu wählen, die zeitlich am besten passt.

Auf ein Map Matching kann auch in Zukunft für den Anwendungskontext der Außenwerbung nicht verzichtet werden, da sonst die Mobilität bei bestimmten Konstellationen unterschätzt werden würde (Häuserschluchten, Tunnel, etc.) und es in

bestimmten Konstellationen (Reflektion der GPS-Signale, Parallelsegmente zu einer Autobahn) zu einer Überschätzung der Leistungswerte kommen kann.

Einen großen Vorteil, den CATI Befragungen gegenüber sensorgestützten Erhebungen (GPS, Mobilfunk, Bluetooth, etc.) haben, sind semantische Informationen, die bei der Telefonbefragung mit aufgenommen werden. Informationen über das gewählte Verkehrsmittel, aber auch der jeweilige Reisezweck sind Bestandteil solcher Befragungen und werden dokumentiert. Bei GPS-Erhebungen müssen diese Informationen erst im Anschluss über eine nachträgliche Annotation der Trajektorien ergänzt werden. Aktuell geschieht die Zuordnung des Verkehrsmittels über eine Geschwindigkeitskonvention und nicht über eine Analyse der GPS-Datenpunkte. Erste Arbeiten zu diesem Thema sind bereits publiziert (Schüssler et al. 2009, Spindler 2008, Sester et. al 2012, Yan et al. 2011a, Yan et al. 2011b) und können in Zukunft in eine intelligentere Klassifikation der Verkehrsarten einfließen. Denn ob ein Proband als Fußgänger, ÖV-Nutzer, Fahrradfahrer oder per PKW an einem Plakat vorbeikommt, hat Einfluss auf die Kontaktmöglichkeit mit dem Plakat. Zusätzlich können noch weitere interessante Informationen aus den GPS-Daten extrahiert werden, z.B.: Wie lange hat sich eine Person an einem bestimmten Ort aufgehalten? Wann war die Person dort? Mit welchen anderen Personengruppen war sie zeitgleich dort? Dies sind Qualitäten, die die GPS-Daten besitzen, die jedoch aktuell noch nicht ausgewertet worden sind.

Nach der Aufbereitung und Klassifikation der GPS-Daten stellt sich die Frage, wie man Trajektorien und Plakate in Beziehung setzt. Zwingend notwendig hierfür sind die Koordinaten und die Himmelsrichtung des Plakates. Liegen diese Informationen vor, können Sichtbarkeitsräume erzeugt werden, die einen maximalen Korridor definieren, aus dem das Plakat gesehen werden kann. Über eine anschließende Verschneidung der GPS-Trajektorien und der Plakatsichtbarkeitsräume erhält man eine Plakatpassage. Plakatpassagen werden in Deutschland und der Schweiz noch zusätzlich mit qualitativen Sichtbarkeitsfaktoren des Plakates (Höhe, Beleuchtung, Größe des Plakates, etc.) zu einem werberelevanten Kontakt verrechnet. Dies ist eine Aufgabe der qualitativen Plakatforschung und war nicht Teil dieser Arbeit. Es wurde jedoch evaluiert, welchen Einfluss die Sichtdistanz auf die Entwicklung der Plakatpassagen hat (Hecker et al. 2010c). Hierzu wurde eine Sensitivitätsanalyse mit unterschiedlich großen Sichtbarkeitsräumen (20-100 Meter) durchgeführt. Es wurde untersucht, wie sich der ooc (opportunity of contact) bei größer werdenden Sichtbarkeitsräumen entwickelt. Ergebnis war, dass die Abweichung der Werte bis zu einer Entfernung von 60 Metern bei 2% liegt. Erst ab einer Entfernung von 60 Metern steigt sie stärker an. Neben der Größe des Sichtbarkeitsraumes wurde auch untersucht, welchen Einfluss eine Individualisierung des Sichtbarkeitsraumes mit Gebäudedaten hat. Gebäudedaten stellen Sichthindernisse für den Sichtbarkeitsraum eines Plakates dar und begrenzen ihn so in seiner Ausdehnung. Es konnte festgestellt werden, dass insbesondere in der Innenstadt ab einer Entfernung von 40 Metern eine Integration von Gebäudedaten ratsam ist. Allerdings sind Gebäudedaten nur eine Möglichkeit von vielen Sichthindernissen im Vorfeld eines Plakates. Hierzu gehören temporäre Sichthindernisse wie Autos, Baustellen, etc. und permanente Sichthindernisse, wie Bäume, Laternen, etc.. Verfügbare Geodaten zu diesen möglichen Hindernissen liegen aktuell nicht vor. Eine Alternative bieten für die Zukunft evtl. Daten wie Google Street View, Bing Streetsight oder 3D Stadtmodelle (HPI 2012). Aus diesen Daten können über geeignete Verfahren der automatisierten Mustererkennung (Image Recognition) Sichthindernisse extrahiert und für eine exaktere Justierung des Sichtbarkeitsraumes herangezogen werden. Allerdings stellt sich die Frage, wie man bei der Mustererkennung mobile von statischen Sichthindernissen unterscheiden kann und ob die Daten auch eine ausreichende Aktualität haben.

Bei Plakatstellen, die keine Passage im Sichtbarkeitsraum aufweisen, handelt es sich häufig um Fehlverortungen der X,Y Koordinate (Plakatfläche schaut in die Gegenrichtung des

Straßensegmentes). Aber es gibt auch Fälle, wo das digitale Straßennetz an seine Grenzen stößt. In Abbildung 8.2 ist ein solcher Fall am Beispiel des Kölner Neumarktes dargestellt. Plakatstellen, die zentral auf Plätzen, Bahnhofsvorplätzen, Markplätzen, etc. stehen, erreichen mit ihrem Sichtbarkeitsraum nicht das nächstliegende Straßensegment und erhalten damit keinen Leistungswert.


Abbildung 8.2: Repräsentation des Kölner Neumarktes am Beispiel von NavTeq (topographische Hintergrundinformation Google 2012)

Eine Option diese Problematik zu beheben ist es, manuell zusätzliche Segmente oder Knotenpunkte bei Platzsituationen einzufügen. Eine Perspektive für die Zukunft stellt die Entwicklung von NavTeq Discover Cities dar (NavTeq 2012). NAVTEQ Discover Cities ist ein Produkt außerhalb der bekannten Autonavigation. Die Lösung bietet eine Vielzahl an fußgängerrelevanten Attributen wie Brücken, Gehsteige, Fußgängerwege und Unterführungen. Virtuelle Verbindungen ermöglichen Wege über Plätze, Parks und anderen Freiflächen. Über die Einbindung von Straßenbahn- bzw., Bushaltestellen sowie von Taxiständen wird ein multimodales Routing möglich.

Die erfassten GPS-Daten umfassen neben der aufgezeichneten Mobilität auch umfangreiche soziodemographische Daten. Eine der Fragen, die in dieser Dissertation gestellt wurde ist: Gibt es systematische Zusammenhänge zwischen fehlenden GPS-Daten, soziodemographischen Variablen sowie der Mobilität von Probanden? Sollte es Verzerrungen in den Daten geben, würden diese über die Hochrechnung vervielfältigt. Aus diesem Grund wurde ein mehrstufiges Verfahren vorgestellt, das signifikant auftretende Muster in den Daten sucht (Hecker et al. 2010a). Der Kern des Ansatzes besteht aus einer Subgruppen-Analyse, die Abhängigkeiten zwischen soziodemographischen Gruppen, fehlenden Tagen und dem Mobilitätsverhalten liefert. Die Ergebnisse einer solchen Untersuchung können verwendet werden, um Verzerrungen aufgrund des Trageverhaltens der Probanden zu kompensieren (MAR – Missing at Random). Sollte es auffällige Muster geben, dann kann eine Konditionierung der Daten vorgenommen werden. Ob ein Muster auffällig oder weniger auffällig ist, wird über die Stärke der gefundenen Subgruppe entschieden. Dem Entscheider ist es an dieser Stelle überlassen, ab welcher Stärke er eine kritische Abhängigkeit detektiert. Hier fehlt es bisher an einer einheitlichen Definition für kritische Grenzwerte. Weiterhin sollte die Untersuchung in Abhängigkeit von der soziodemographischen Ausweisungstiefe betrachtet werden. Möchte ein Mediaplaner sehr detaillierte Reichweitenausweisungen nach

Geschlecht, Alter, Beruf, Ausbildung etc., wird es schwieriger, mit einzelnen Verzerrungen in Subgruppen umzugehen, da der entsprechende Suchraum größer wird.

| | |
|---|---|
| **Stärken** | 1. Datenerfassung mit GPS ist umfangreicher, genauer und direkter im Vergleich zu CATI-Befragungen.<br>2. Kurze Datenlücken können über Map Matching Verfahren geschlossen werden.<br>3. Es kann eine räumliche Verschneidung von GPS-Trajektorien und Sichtbarkeitsräumen vorgenommen werden. Direkte Erfassung der Passage, keine Erinnerungsleistung von Probanden. |
| **Schwächen** | 1. GPS-Erhebung ist sehr kostenintensiv im Vergleich zu CATI-Befragungen.<br>2. Bei größeren Datenlücken hat das Map Matching viele Wegeoptionen.<br>3. Sichtbarkeitsräume geben nicht die aktuelle Situation vor Ort wider.<br>4. Digitale Straßennetze können bisher kaum mit Plätzen, Freiflächen und Parks umgehen.<br>5. Verzerrungen in den GPS-Daten werden detektiert, es liegen aber keine Grenzwerte vor. |
| **Perspektive** | 1. Erfassung der GPS-Empfangsqualität für ein qualitätsabhängiges Matching.<br>2. Nutzung von externen Daten (Google, Bing, 3D Stadtmodelle) zur Individualisierung von Sichtbarkeitsräumen.<br>3. NavTeq Discover Cities zur Modellierung von Platzsituationen. |

Tabelle 8.1: Zusammenfassung Datenaufbereitung und Validierung

**Modellierung von Leistungswerten mit GPS**

Bei der Erhebung von Mobilitätsstudien kann es leicht dazu kommen, dass Messtage aus unterschiedlichen Gründen nicht erfasst werden. Dies können technische Gründe sein (Gerät defekt, Batterie nicht aufgeladen), oder Ermüdungsgründe der Probanden (kein Interesse mehr an der Studie, Vergessen des Gerätes). Die fehlenden Tage stellen signifikante Datenlücken in den Ausgangsdaten dar und können nicht ignoriert werden. Es stellt sich die Frage, wie fehlende Messtage in einer Modellierung behandelt werden können. Zur Lösung der Problematik wurde auf eine Methode der Survival-Analyse zurückgegriffen (Kaplan-Meier) (Hecker et al. 2010b). Die Messreihen der Personen, die aufgrund von technischen oder individuellen Gründen aus der Studie ausscheiden, werden nach Kaplan-Meier zensiert. Die generelle Idee bei dieser Technik ist, dass die Stichprobe an ihre variierende Größe schrittweise angepasst wird. Durch diese Zerlegung der Stichprobe in bedingte Wahrscheinlichkeiten passt das Verfahren die Stichprobe an die abnehmende Größe an und löst damit das Problem des unvollständigen Messzeitraumes. Die Validierungen zeigen, dass Kaplan-Meier ein guter Schätzer für den Umgang mit zeitlich unvollständigen Daten ist. Allerdings hat die vorgestellte Modellierung den Nachteil, nur mit kompletten Fehltagen

umgehen zu können. Bei technisch bedingten größeren Mobilitätslücken, die nicht durch ein Map Matching geschlossen werden können, oder beim Fehlen ganzer Teilstrecken, wie z.B. dem kurzen Weg zum Bäcker, der morgens vergessen wird, kann das Kaplan-Meier Vorgehen keine Lösung anbieten. Diese Problematik stellt weiterhin eine Herausforderung dar. Es besteht auch noch die Frage, wie man solche fehlenden Wege eines Probanden identifizieren kann? Eine Möglichkeit besteht evtl. darin, eine längere GPS-Mobilitätsstudie zu initiieren, um somit die Möglichkeit zu haben, Regelmäßigkeiten im Mobilitätsverhalten zu erkennen und die fehlenden Wege aufzufüllen. Forschergruppen von der Universität Hasselt (Bellemans et al. 2010, Han et al. 2009a, Han et al. 2009b) und der ETH Zürich und TU Berlin (Bekhor et al. 2011, Meister et al. 2010, Rieser 2010) beschäftigen sich mit aktivitätsbasierenden Simulationen des Verkehrs, die das Verkehrsverhalten von Personen tageweise komplett simulieren. Diese Option scheint aber aktuell, wie in Abschnitt 5.1 beschrieben, nicht ohne eine Menge von Konventionen und Annahmen in der Modellierung auszukommen. Der Vorteil in der Kaplan-Meier Methode liegt eindeutig in ihrer extremen Datennähe, die keine Parametrisierung verlangt. Außerdem werden keine Daten ergänzt, sondern die Lücken in der Methode behandelt.

Eine zweite Problemstellung bei der Modellierung von GPS-Stichproben ist die räumliche Abdeckung. Bei den GPS-Erfassungen in Deutschland handelt es sich zwar um eine der europaweit größten GPS-Stichproben, doch ist die Stichprobe vergleichsweise gering in Bezug auf die notwendige räumliche Abdeckung. Die erhobenen Probandenwege decken nicht alle Straßensegmente einer Stadt ab. Würde man an dieser Stelle die unvollständige räumliche Abdeckung der GPS-Daten ignorieren, so könnten eine Vielzahl von Plakatstellen nicht bewertet werden, da sie keine GPS-Passage aufweisen. Zur Lösung dieses Problems wurde ein Verfahren vorgeschlagen, das eine Stadt in Räume (Mobilitätseinheiten) aufgeteilt und im Anschluss die aufgezeichneten Mobilitätsinformationen auf der Aggregationsebene dazu nutzt, die räumliche Mikrovariabilität per Simulation zu erhöhen (Hecker et al. 2011b). Hierzu wurde ein mehrstufiges Verfahren vorgestellt:

1. Bildung von Mobilitätseinheiten für das Stadtgebiet
2. Aggregation von GPS-Tracks auf Ebene der Mobilitätseinheiten
3. Aufbau einer Mobilitätsverteilung innerhalb der Mobilitätseinheiten mit externen Daten
4. Disaggregation der GPS-Tracks in wiederholten Simulationen
5. Berechnung der Leistungswerte nach jeder Simulation

Der vorgestellte Lösungsweg zielt darauf ab, eine höhere räumliche Variabilität der Mobilität zu erzeugen. Die grundlegende Idee ist, die Mobilität auf Mikro- und Makroebene zu separieren und getrennt voneinander zu behandeln. Mobilität auf Mikroebene wird an dieser Stelle definiert als lokaler Verkehr, die Makromobilität als Pendler und überregionaler Verkehr. Mit der Sekundärdatenquelle Frequenzatlas und dem vorgeschlagenen System der Mobilitätseinheiten wird durch eine mehrfach durchgeführte Simulation die Variabilität der Mobilität erhöht. Im Ergebnis bekommt jedes Plakat eine Passagewahrscheinlichkeit, konsistent mit den Frequenzen des Frequenzatlas.

Diskutiert werden sollten an dieser Stelle zwei wichtige Komponenten der Modellierung, erstens die Erstellung des Systems der Mobilitätseinheiten und zweitens die Simulation innerhalb des Systems der Mobilitätseinheiten. Zu Beginn der Erstellung des Systems der Mobilitätseinheiten wurden folgende Vorgaben an die Modellierung gestellt:

1. Die Makromobilität soll erhalten bleiben
2. Es soll zu keiner Vermischung von Pendler- und lokaler Mobilität kommen
3. Natürliche und andere Barrieren sollen berücksichtigt werden
4. Das System soll keine Lücken und keine Überlappungen aufweisen

Grundlegende Basis für die Erstellung des Systems der Mobilitätseinheiten ist das NavTeq Straßennetz. Jedem Straßensegment wurde gemäß der NavTeq-Funktionsklasse eine von zwei Klassen zugeordnet. Die NavTeq-Funktionsklassenzuordnung basiert auf der möglichen Verkehrslast (Spuren der Straße, Geschwindigkeit, mögliche Verkehrslast etc.), der offiziellen administrativen Bezeichnung und weiteren NavTeq-spezifischen Annahmen. Die Klasse der Grenzlinien wird anhand der funktionalen NavTeq-Klasse {1, 2, 3} zugeteilt und die inneren Mobilitätseinheiten anhang der Funktionsklassen {4, 5}. Innere Mobilitätseinheiten bilden Aggregationseinheiten für den lokalen Verkehr und Grenzeinheiten für den Pendler und überregionalen Verkehr. Jedes Straßensegment ist im Anschluss eindeutig zugeordnet. Die Einteilung erzeugt eine gute und plausible Trennung des lokalen und überregionalen Verkehrs. Häufig werden durch diese Einteilung auch Funktionsräume innerhalb einer Stadt differenziert (Universitätsviertel, Wohnviertel, Gewerbegebiete), jedoch gibt es immer wieder Fälle, an denen diese Einteilung nicht zufriedenstellend ist. Das liegt daran, dass diese Einteilung nicht der originäre Zweck des NavTeq Netzes ist. Neben den vorgeschlagenen zwei Klassen besteht die Möglichkeit noch weitere Klassen einzufügen, um insbesondere bestimmte Viertel und Stadtteile noch besser voneinander abzugrenzen. Die Frage ist, ob es Alternativen zur Erstellung des Systems der Mobilitätseinheiten jenseits des vorgestellten Ansatzes aus Abschnitt 5.2 gibt? Eine der wesentlichen Voraussetzungen hierfür muss sein, dass die Systematik eine eindeutige NavTeq-Definition (oder auch eines anderen digitalen Straßennetzes) besitzen muss. Denn das NavTeq-Netz stellt die Grundlage für das Map Matching dar, und damit ist man über die Trajektorienzuordnung an dieses Netz oder ein alternatives digitales Straßennetz gebunden. Damit entfallen mögliche Ansätze mit Stadtteilen oder PLZ-Gebieten, da hier keine eindeutige Zuordnung zu einem digitalen Straßennetz besteht. Optional könnte in Zukunft eine Einteilung in Mobilitätseinheiten über andere Attribute des Netzes, externe Attribute, die mit dem Netz verbunden werden, z.B. Einwohnerdichte, oder einer Kombination von unterschiedlichen Daten durchgeführt werden.

Die zweite wichtige Komponente bei der Erzeugung einer höheren räumlichen Variabilität ist die Simulation. Dabei kommt der externen Datenquelle Frequenzatlas eine wichtige Bedeutung zu. Im ersten Schritt der Simulation wird eine räumliche Disaggregation durchgeführt. Hierzu werden die Frequenzen des Frequenzatlas in eine Wahrscheinlichkeitsverteilung des Verkehrs nach Verkehrsart und pro Aggregationseinheit umgewandelt. Mit dem Wissen über die Verkehrsverteilung und durchschnittliche Anzahl der berührten Segmente durch das Map Matching wird die Simulation pro Mobilitätseinheit durchgeführt. Bei der Simulation werden zufällig, nach der Anzahl der durchschnittlichen Berührungen, Straßensegmente in der jeweiligen Mobilitätseinheit gezogen. Das heißt, die Trajektorien verlieren bei der Simulation auf Ebene der Mobilitätseinheiten ihre Konnektivität. Zukünftig kann man sich vorstellen, an dieser Stelle die Simulation noch weiter zu verfeinern und ein wegebasiertes Routing mit einer Konnektivität der Wege zu implementieren. Unklar ist, auf welche alternative Verkehrsverteilung man an Stelle des Frequenzatlas zurückgreifen kann. Vergleichbare Sekundärdaten, die bis auf die unterste Straßenfunktionsebene Frequenzen für alle Verkehrsarten zur Verfügung stellen, sind aktuell am Geodatenmarkt nicht verfügbar.

| | |
|---|---|
| **Stärken** | 1. Fehlende Tage können über Kaplan-Meier behandelt werden.<br>2. Die unvollständige räumliche Abdeckung kann über externe Daten (Frequenzatlas), ein Aggregations- und Simulationsmodell behandelt werden.<br>3. Es wird kein Pendler- und lokaler Verkehr vermischt.<br>4. Es handelt sich um eine sehr datennahe Modellierung. |
| **Schwächen** | 1. Das Problem der fehlenden Wege ist weiterhin eine Herausforderung.<br>2. Die Konnektivität innerhalb der Aggregationseinheiten geht verloren. |
| **Perspektive** | 1. Längere Mobilitätsaufzeichnungen per Handy Applikation.<br>2. Auffüllen fehlender Wege und Tage über gefundene Muster. |

Tabelle 8.2: Zusammenfassung Modellierung von Leistungswerten mit GPS

**Robustheitsanalyse**

In der anschließenden Robustheitsanalyse wurde erstens untersucht, wie sich Reichweiten ändern, wenn in den kommenden Jahren eine kleinere Stichprobe erhoben wird. Auf Basis einer Subsampling-Methode wurde eine Verkleinerung der vorhandenen Stichprobe durchgeführt und die Ergebnisse miteinander verglichen. Zweitens wurde analysiert, was passiert, wenn sich die Personenstichprobe verändert oder ausgetauscht wird. Eine weitere Frage war, wie sich der Standardfehler in Bezug auf die Kampagnen- und Personengröße verhält. Über die Anwendung von Bootstrap wurde eine Variation der Netzgröße und Probandengröße erzeugt und die Ergebnisse miteinander verglichen (Hecker et al. 2011a).

Die Ergebnisse zeigten zum einen, dass der berechnete Standardfehler klein ist und zum anderen, dass ein linearer Zusammenhang zwischen Standardfehler und der Größe der Werbekampagne existiert. Weiterhin wurde festgestellt, dass beim Verkleinern des Datensatzes der Standardfehler exponentiell ansteigt, wobei der Zuwachs des Standardfehlers bis zu einer Teilmenge von 40% nahezu linear verläuft. Daraus ergibt sich, dass begrenzt ein Verkleinern von Datensätzen oder ein Austausch von GPS-Daten möglich ist, ohne dass sich nennenswerte Veränderungen in der Leistungswertbestimmung von Werbekampagnen entwickeln.

Zukünftig kann man sich insbesondere noch differenziertere Untersuchungen für einzelne soziodemographische Gruppen vorstellen. Die dargestellten Methoden zur Evaluierung beschäftigten sich ausschließlich mit Berechnungen von Leistungswerten einer ganzen Population. Künftig sollten Analysen auf geschichteten Datensätzen durchgeführt werden, indem die Probandenzahl auf die kleinste zu untersuchende demographische Gruppe beschränkt wird. Denn in der Regel möchte ein Mediaplaner noch eine feinere Differenzierung angezeigt bekommen als die Reichweite der Gesamtpopulation. Sind Auswertungen hinsichtlich Geschlecht, Altersgruppen und Berufsgruppen erwünscht, müssen noch weitere Analysen durchgeführt werden, so dass Mediaplaner auch klare Ausweisungsgrenzen mitgeteilt werden können.

| | |
|---|---|
| **Stärken** | 1. Ergebnisse sind sehr robust gegenüber neuen und kleineren Stichproben.<br>2. Linearer Zusammenhang zwischen Standardfehler und Kampagnengröße. |
| **Schwächen** | 1. Auswertungen wurden auf keinen soziodemographischen Gruppen durchgeführt. |
| **Perspektive** | 1. Bestimmung der Ausweisungsgrenzen für Mediaplaner in Abhängigkeit von der Stichprobengröße. |

Tabelle 8.3: Zusammenfassung Robustheitsanalyse

**Ausweisung von räumlich differenzierten Reichweiten**

Im abschließenden Kapitel 7 wurde auf Basis von realen Plakat- und Mobilitätsdaten die vorgestellte Modellierung für unterschiedliche Kampagnenkonfigurationen und Zielgruppen umgesetzt. Es wurde gezeigt, wie sich eine unterschiedliche räumliche Verteilung der Kampagnen auf die Leistungswerte auswirkt. Zusätzlich wurden noch Experimente nach Herkunft, soziodemographischen Variablen der Probanden und Kontaktklassen durchgeführt (Hecker et al. 2010b). Insbesondere Kontaktklassen stellen ein interessantes Mittel zur Steuerung von unterschiedlich wirksamen Werbekampagnen dar. Alle Experimente zeigten plausible und nachvollziehbare Ergebnisse. Die vorgestellten Kampagnen stellen typische Anwendungsbeispiele von Mediaplanern dar und können nur aufgrund der neuen Methodik räumlich differenziert in die Planung einfließen. Sie zeigten, welchen Einfluss die räumliche Verteilung von Plakatkampagnen auf die Leistungswerte hat. Allerdings hat die feinräumige Ausweisung auch ihren Preis, die Modellierung in der Schweiz und in Deutschland hat massiv an Komplexität dazugewonnen.

Weiterhin ist zu beachten, dass auch die feinräumige Ausweisung aufgrund der vorliegenden Daten ihre Grenzen hat. Es sollte mit der vorhandenen Stichprobe sehr vorsichtig bei einer Differenzierung in der Zielgruppenkombination z.B. nach Alter, Herkunft, Geschlecht und Berufsgruppe vorgegangen werden. Mit zunehmender Differenzierung bei der Zielgruppenselektion wird die GPS-Datenbasis immer kleiner, und die Ergebnisse sind an dieser Stelle nicht mehr robust. Die jeweiligen Fallzahlen sollten bei den Analysen mit herangezogen werden.

Eine Frage, die in dieser Arbeit bei der Ausweisung von Leistungswerten nicht adressiert worden ist: Wie kommt man von der Passage zu einem potenziellen Kontakt mit der betreffenden Werbung? In den 90er Jahren hat die Gesellschaft für Konsumforschung (GfK) Tests im Umfeld von Tankstellen durchgeführt. Hierzu wurden mobile Werbeflächen in unterschiedlichen Winkeln, Formaten, etc. im Umfeld der betreffenden Tankstelle positioniert und die Kunden der Tankstelle im Anschluss per Fragebogen befragt. Über die Ergebnisse der Befragung wurden für die unterschiedlichen Konstellationen Faktoren ermittelt, mit der eine Abgewichtung der Passage zu einem Kontakt vorgenommen worden ist. Aktuelle Ideen zur Neuauflage einer solchen Untersuchung gehen in die Richtung des Einsatzes von Fahrsimulatoren. Das Untersuchungsdesign wäre vergleichbar mit den durchgeführten Tests der GfK. Mit einem Fahrsimulator bewegt sich eine Testperson durch eine virtuelle Stadt, und pro Testdurchlauf werden Plakatstellen in unterschiedlichen Winkeln, Höhen usw. positioniert. Im Anschluss folgt eine Auswertung der Untersuchung über die

Plakatwahrnehmung der Probanden. Der Erfolg des Einsatzes solcher Simulatoren wird davon abhängen, wie realistisch das Stadtbild im Fahrsimulator wiedergegeben werden kann, insbesondere in der Tiefendarstellung (Ab wann kann ich ein Plakat erkennen?). Auch stellt sich die Frage, ab wann der Gewöhnungsprozess im Fahrsimulator einsetzt und der Proband sich in einer alltäglichen Fahrsituation fühlt.

Eine weitere Möglichkeit zur Erfassung potenzieller Kontakte sind Kamerasysteme. Es wurden bereits unterschiedlichste Tests mit Helmkameras und Kamerasystemen an Plakatstellwänden durchgeführt. Ziel war es festzustellen, ob ein Blickkontakt mit dem Plakat hergestellt worden ist und wie lange dieser gedauert hat. Bei diesen Untersuchungen wurden keine unterschiedlichen Plakatkonstellationen, wie Entfernung, Format oder ähnliches getestet, sondern die Wirksamkeit des Werbeinhaltes. Welche Werbegestaltung zieht mehr oder weniger Blickkontakte auf sich? An dieser Stelle zieht die Außenwerbebranche allerdings auch eine Ausweisungsgrenze. Wie wirksam ein bestimmter Werbeinhalt einer individuellen Kampagne ist und wie gut dieser angenommen und dann später in Umsatz umgemünzt wird, ist aus Sicht der Außenwerbung Aufgabe der jeweiligen Mediaagenturen.

Fazit ist, dass im Bereich der Kontaktgewichtung aktuell noch mit vielen Konventionen gearbeitet wird und noch kein Verfahren/Vorgehen existiert, das diese ersetzen könnte.

| | |
|---|---|
| **Stärken** | 1. Reichweiten können räumlich differenziert ausgewiesen werden.<br>2. Zur Werbekampagnensteuerung kann in Zukunft mit unterschiedlichen Kontaktklassen gearbeitet werden. |
| **Schwächen** | 1. Es ist nicht klar definiert, wann ein Kontakt mit einer Kampagne entsteht (Passage versus Kontakt). |
| **Perspektive** | 1. Kontaktstudie über Fahrsimulatoren und Kamerasysteme. |

Tabelle 8.4: Zusammenfassung Ausweisung von räumlich differenzierten Reichweiten

## 8.2 Zukunftsfähigkeit der Modellierung und Perspektiven für die Zukunft

Diese Dissertation ist die erstmalige und vollständige Darstellung zum Vorgehen der Leistungswertermittlung mittels GPS in der Außenwerbung. Die Analyse beinhaltet mehrere Probleme: 1. Wann findet ein Kontakt mit einem Plakat statt? 2. Fast alle Bewegungsdaten (Mobilfunk, Bluetooth, etc.) besitzen fehlende zeitliche und räumliche Daten, wie kann man mit diesem Problem umgehen? 3. Wie robust sind die Ergebnisse in der Zukunft? Diese Probleme wurden in dieser Arbeit im Anwendungskontext der Außenwerbung adressiert. Die einzelnen Bausteine der Modellierung stellen zum Teil neue oder adaptierte Verfahren im Anwendungskontext dar. Die Fragen in diesem Abschnitt sind, wie zukunftsfähig ist die Modellierung und welche Perspektiven gibt es für die Zukunft? Wie relevant sind die Ergebnisse für eine zukünftige Forschung im Anwendungskontext?

Bereits heute existieren in der Mediaplanung Wünsche für weitere Ausweisemöglichkeiten von Plakatkampagnen. Neben den bereits erwähnten soziodemographischen Variablen ist es ein Wunsch, Leistungswerte über einen Zeitraum von 7 Tagen hinaus ausweisen zu können. Die Frage ist, auf welcher Datenbasis soll und kann dies geschehen, ohne dass die Kosten für die empirische Erhebung immens steigen? Als im Jahre 2003 in der Schweiz die ersten Versuche mit GPS durchgeführt wurden, war die GPS Technologie nur in speziellen GPS

Trackern verfügbar. Inzwischen kann man sich Smart Phones ohne GPS nicht mehr vorstellen. Dies birgt für die Zukunft neue Möglichkeiten der Mobilitätserfassung. Der Vorteil würde darin bestehen, dass Probanden kein zusätzliches Gerät mehr zur Erfassung mitnehmen müssen, sondern über eine spezielle App die Mobilität über ihr eigenes Mobilfunkgerät erfassen können. Probanden hätten zusätzlich die Möglichkeit, bestimmte Wege aus Gründen der Privatsphäre nicht mit aufzeichnen zu lassen. Da es sich um ein eigenes Gerät handelt, würde das Vergessen der Gerätemitnahme wahrscheinlich stark reduziert werden, und die benötigten Hardwarekosten für eine Erhebung würden entfallen. Erste Tests mit Mobilfunkgeräten und einer speziellen App wurden im Jahre 2010/2011 in der Schweiz durchgeführt. Leider war das Ergebnis ernüchternd: Die geringe Batteriekapazität, bzw. der hohe Energiebedarf des GPS-Empfängers erlaubten keine ausreichende Betriebsdauer. Um den Energiebedarf zu reduzieren wurde nur in einem zeitlichen Intervall von 30 Sekunden ein GPS-Punkt aufgenommen. Die Laufzeit der Geräte betrug bei diesen Tests maximal 4-5 Stunden, was einer aktuellen Mobilitätserfassung noch entgegen spricht.

Eine weitere Perspektive stellen zukünftig die Auswertungen von Mobilfunkproviderdaten dar. Über diese Datenquellen kann passiv die Mobilität ausgewertet werden. Wie bereits in Abschnitt 3.3.1 vorgestellt, sind hier noch insbesondere Fragen zum Schutz der Privatsphäre zu lösen und die Problematik zu untersuchen, wie aus einem aufgezeichneten Telefonieverhalten auf ein Mobilitätsverhalten geschlossen werden kann. Einer der immensen Vorteile bei diesem Datensatz wäre, dass man eine massive Steigerung der Stichprobe hätte, einen evtl. langen Zeitraum der Aufzeichnung (Monate, Jahre) zur Verfügung stehen hat, und die Daten zukünftig vielleicht auch in Echtzeit vorliegen können. Abhängig vom jeweiligen Provider muss darauf geachtet werden, ob es Verzerrungen in den Daten gibt. So sprechen viele Provider über spezielle Angebote besondere Zielgruppen an. Dies muss in der evtl. Nutzung der Daten berücksichtigt werden. Bei einer evtl. Nutzung von Mobilfunkdaten ist allerdings auch noch unklar, wie man eine ausreichend feine räumliche Auflösung erreicht, damit eine Plakatpassage identifiziert werden kann. Basierend auf Funkzellendaten, der empfangenen Signalstärke der aktuellen sowie der benachbarten Funkzellen, hat Zimmermann et al. (2004) eine Positionierungsgenauigkeit von weniger als 80m in 67% und 200 m in 95% in einem städtischen Szenario erreicht. Mit einem ähnlichen Verfahren hat Haeb-Umbach et al. (2007) eine Positionierungsgenauigkeit von 124m zu 67% erzielt. Diese Ergebnisse wurden jeweils bei einem fast idealen Setup und im Zentrumsbereich einer Stadt erreicht. Für einen flächendeckenden, landesweiten Einsatz bestehen noch weitere Herausforderungen.

Eine weitere Frage ist, wann eine GPS-Erhebung ersetzt werden muss. Wie kann man feststellen, dass sich das Mobilitätsverhalten innerhalb einer Stadt geändert hat? Eine Möglichkeit besteht darin, Verkehrsveränderungen über die Analyse des Straßennetzes zu identifizieren. Sind neue wichtige Straßen in einer Stadt dazu gekommen oder haben sich die Straßenrestriktionen in einer Stadt stark geändert. Gibt es neue Verkehrsangebote, z.B. beim ÖPNV, sind neue wichtige POI´s, Arbeitsplätze oder Geschäfte dazugekommen? In Deutschland wird aktuell eine rollierende GPS-Stichprobenerhebung etabliert. Nach 6 Jahren werden alle alten GPS-Daten durch neue Erhebungen vollständig ersetzt. Ob dies notwendig ist, ist noch zu klären. Hier besteht aktuell noch Forschungsbedarf.

Das vorgestellte Vorgehen bietet einen sehr offenen Rahmen für weitere Ideen und Forschungsvorhaben mit weiteren Datenquellen der Mobilitätserfassung. Jeden Tag besuchen oder passieren wir bei unserer täglichen Mobilität geographische Orte im Raum. Das Wissen über solche Interaktionen ist jenseits der Außenwerbung auch für viele andere Branchen interessant.

# KAPITEL 9

## 9. RESÜMEE UND AUSBLICK

Die Außenwerbung steht in Europa vor einem Umbruch in der Leistungswertberechnung. Neue Techniken der Mobilitätserfassung und die digitale Erfassung von Plakatstandortdaten machen es möglich, Leistungswerte individuell für beliebige Kampagnen und Zielgruppen zu bestimmen. Mobilitätsanalysen und Auswertungen sind in der Außenwerbung die entscheidende Basis zur individuellen Leistungswertberechnung einzelner Werbeträger und ganzer Kampagnen. Der konkurrierende Markt in der Mediabranche fordert eine exakte und möglichst genaue Bewertung des Mediaerfolges. Liegen intransparente, zu unspezifische oder gar keine exakten Kennziffern vor, werden Mediabudgets in andere Mediasparten transferiert. Aus diesem Grund ist die Außenwerbung eine Branche, die in den letzten Jahren intensiv in die Mobilitätsforschung investiert hat.

Als Pioniere im Einsatz innovativer Verfahren zur Leistungswertbestimmung haben in den letzten Jahren Deutschland und die Schweiz umfangreiche GPS-Mobilitätsstudien durchgeführt. Die vorliegende Dissertation behandelt den Umgang mit einer umfangreichen GPS-Stichprobe im Anwendungskontext der Außenwerbung für Deutschland und die Schweiz. Dabei lautete eine der zentralen Fragen: *"Wie können GPS-Daten dazu eingesetzt werden, fundierte Leistungswerte für die Außenwerbung zu bestimmen?".* Vor dem Hintergrund dieser Frage wurden Tests definiert, die Selektionseffekte in der Rekrutierung von Probanden offenlegen und Verzerrungen in den Daten erkennen. Es wurden geeignete Lösungswege im Umgang mit zeitlichen und räumlichen Datenlücken diskutiert und erprobt, sowie Validitätsanalysen zu den einzelnen Methoden durchgeführt. Es wurde die Problematik behandelt, wie Leistungswerte auf eine Neuerhebung der Mobilität reagieren und wo die Grenzen einer Leistungswertbestimmung bei kleineren und damit kostengünstigeren GPS-Erhebungen liegen. Zum Abschluss wurden die Vorteile und neuen Ausweisungsmöglichkeiten von räumlich differenzierten Plakatkampagnen für die Beispielstädte Köln und Bern vorgestellt. Eine grundlegende Motivation der Dissertation war es, die zentralen Fragen und Herausforderungen, die während des Erfahrungsaufbaus und der ersten Analyse der GPS-Daten entstanden sind, in einer systematischen Form niederzuschreiben und so ein Vorgehen zu skizzieren, welches auf weitere Länder übertragen werden kann. Die vorliegende Arbeit zeigt, dass ein GPS-Mobilitätsdatensatz mithilfe der vorgestellten Modellierungsschritte eine geeignete Datenquelle zur Leistungswertbestimmung in der Außenwerbung darstellt. Im Ergebnis liefert der neue Ansatz eine differenziertere Ausweisung von Leistungswerten im Vergleich zu dem vorher am Markt bestehenden Copland Modell, das z. T. in anderen Ländern noch im Einsatz ist. In England, den Niederlanden und Österreich sind aktuell GPS-Stichproben im Feld, um in Zukunft eine vergleichbare Leistungsausweisung wie in Deutschland und der Schweiz dem Werbemarkt anbieten zu können.

Bei allen Vorteilen, die der Einsatz der GPS-Technologie mit sich bringt, hat er insbesondere einen Nachteil. GPS kann aufgrund des Signalverlustes nicht in geschlossenen Räumen zur Datenerfassung verwendet werden. In der Schweiz und Deutschland stehen jedoch viele Plakate in öffentlichen Gebäuden, wie Bahnhöfen oder Einkaufszentren. Aktuell ist deren Leistungsbewertung bei vielen Außenwerbefirmen von hohem Interesse. Erste Studien liefern hier bereits Möglichkeiten der Integration von Bahnhöfen in die Leistungsbewertung, doch

müssen für diesen Zweck zusätzliche Daten erfasst werden (Liebig et al. 2010). Für die Modellierung bedeutet dies zusätzliche Kosten und weitere Komplexität. Zukünftig wäre es erstrebenswert, eine Erfassungstechnologie einzusetzen, die auch in Gebäuden wirksam ist. W-LAN und Bluetooth kommen für diese Erfassung beispielsweise in Frage.

Eine weitere Herausforderung stellt die sogenannte Verkehrsmittelwerbung dar. Bei dieser weiteren Gattung der Außenwerbung handelt es sich um öffentliche und private Bus- und Straßenbahnunternehmen, die Werbung außen an den Fahrzeugen anbringen. Zielgruppe bei diesen Werbeformen sind neben den Fahrgästen auch Passanten. Die besondere Herausforderung für eine Leistungsbewertung bei dieser Werbeform ist, diese *„mobilen Plakatstellen"* zu bewerten, da sie sich dynamisch durch den Verkehrsraum bewegen und sozusagen ihre möglichen Kontakte über die Zeit einsammeln.

Ein Trend, der sich in den beiden letzten Jahren in der Außenwerbung abzeichnet, ist die Digitalisierung der Plakatformate. Digitale Werbetafeln können sowohl statische als auch animierte Inhalte liefern. Inhalte können dynamisch während des Tages geändert werden, so dass der Mediaplaner zu unterschiedlichen Tageszeiten die Werbung auf die passierenden Zielgruppen an der Plakatstelle anpassen kann. In Zukunft wird es notwendig sein, nicht nur zu berechnen, wie die Reichweite pro Woche, sondern auch pro individuellem Tag der Woche und sogar pro Tageszeit ist, damit eine tageszeitlich und kundenspezifische Werbung ermöglicht werden kann.

Jenseits der Themen aus dem Bereich der Außenwerbebranche haben die vorgestellten Methoden und erfassten GPS-Datensätze in Deutschland und der Schweiz das Potenzial für weitere interessante Forschungsaktivitäten. Zu diesen zählen erstens standortbezogene Auswertungen für die Immobilienbranche, z.B. für die Bewertung von Einzelhandelsfilialen (Supermärkte, Drogerieketten, etc.). So kann auf Grundlage der vorliegenden Daten ermittelt werden, welches Potenzial ein Filialist auf dem anliegenden Straßensegment seines Geschäftes hat, und dies Kassendaten oder Frequenzzählungen im Geschäft gegenüberstellen. Diese Information stellt die sogenannte Konversions Rate dar, die ein wichtiger Indikator dafür ist, wie attraktiv ein Ladengeschäft für die Passanten ist. Zusätzlich kann noch untersucht werden, wie viele unterschiedliche Personen pro Woche ein Geschäft passieren. Insbesondere Einzelhändler, die keine Produkte für den alltäglichen Bedarf anbieten, sondern Produkte für Spontankäufe, sind daran interessiert, viele unterschiedliche Personen pro Woche oder Monat zu erreichen.

Eine zweite interessante Forschungsaktivität stellen mobilitätsbezogene Auswertungen auf den GPS-Daten dar. So können auf Basis der aufbereiteten GPS-Daten Analysen hinsichtlich der zurückgelegten Fahrtstrecken, z.B. nach unterschiedlichen Gruppen (Alter, Geschlecht, Einkommen), nach Zeit (Montag-Freitag versus Samstag und Sonntag), nach besuchten POI´s sowie nach individuellen Aktivitätsräumen der Probanden untersucht werden. Diese Informationen können wiederum dazu genutzt werden, Einzugsgebiete für standortbezogene Analysen zu überprüfen oder gegebenenfalls anzupassen.

Ein offener Punkt für zukünftige Erhebungen z.B. mittels Applikationen auf Mobilfunkgeräten, die GPS-Koordinaten aufzeichnen, ist das Thema Datenschutz. Bei Mobilitätsdaten handelt es sich um sehr private, sensitive und schützenswerte Daten. Bei diesen Daten reicht es nicht nur, die persönliche Identifikation aus dem Datensatz zu löschen, denn über die Start- und Endkoordinate eines Probanden lässt sich relativ einfach seine Wohnkoordinate ableiten. Für die Zukunft müssen Forscher für Erhebungen, die vielleicht auf einer noch größeren und umfangreicheren Basis beruhen, geeignete Verfahren und Methoden entwickeln, die die Privatsphäre von teilnehmenden Personen in geeigneter Weise schützen.

## 10. LITERATURVERZEICHNIS

AALEN, O. O. BORGAN, Ø. AND GJESSING, H. K.: SURVIVAL AND EVENT HISTORY ANALYSIS. STATISTICS FOR BIOLOGY AND HEALTH. SPRINGER, 2008

AGUMYA, A. AND HUNTER, G.J.: A RISK-BASED APPROACH TO ASSESSING THE 'FITNESS OF USE' OF SPATIAL DATA, URISA JOURNAL, 11, 1999

ALLEN, J. F.: TOWARDS A GENERAL THEORY OF ACTION AND TIME. ARTIFICIAL INTELLIGENCE, 23(2): 123-154. 1984

ANDRIENKO, G., ANDRIENKO, N., KOPANAKIS, I., LIGTENBERG, A., WROBEL, S.: VISUAL ANALYTICS METHODS FOR MOVEMENT DATA. IN: F. GIANNOTTI UND D. PEDRESCHI, EDS. MOBILITY, DATA MINING AND PRIVACY. HEIDELBERG, BERLIN: SPRINGER-VERLAG, 375-410. 2008A

ANDRIENKO, N., ANDRIENKO, G., PELEKIS, N., SPACCAPIETRA, S.: BASIC CONCEPTS OF MOVEMENT DATA. IN: F. GIANNOTTI UND D. PEDRESCHI, EDS. MOBILITY, DATA MINING AND PRIVACY. HEIDELBERG, BERLIN: SPRINGER-VERLAG, 15-38. 2008B

ANDRIENKO, N. ANDRIENKO, G. STANGE, H. LIEBIG, T. HECKER, D.: VISUAL ANALYTICS FOR UNDERSTANDING SPATIAL SITUATIONS FROM EPISODIC MOVEMENT DATA JOURNAL KI KÜNSTLICHE INTELLIGENZ, THEMENHEFT SPATIOTEMPORAL MODELING AND ANALYSIS, IM ERSCHEINEN, 2012

ANKERST, M., BREUNIG M. M., KRIEGEL H-P. AND SANDER J.: OPTICS: ORDERING POINTS TO IDENTIFY THE CLUSTERING STRUCTURE. IN: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA. ACM PRESS. PP. 49-60. 1999

ARBEITSGEMEINSCHAFT MEDIA-ANALYSE E.V. (AG.MA), 2012A. METHODENSTECKBRIEF ZUR BERICHTERSTATTUNG. HTTP://WWW.AGMA-MMC.DE/03_FORSCHUNG/PLAKAT/BERICHTERSTATTUNG/THEMEN.ASP?TOPNAV=10&SUBNAV=199, 2012-05-11

ARBEITSGEMEINSCHAFT MEDIA-ANALYSE E.V. (AG.MA), 2012B. DIE MEDIA-ANALYSE PLAKAT. HTTP://WWW.AGMA-MMC.DE/03_FORSCHUNG/PLAKAT.ASP?TOPNAV=10&SUBNAV=199, 2012-05-11.

ARBEITSGEMEINSCHAFT MEDIA-ANALYSE E.V. (AG.MA), 2012C. DIE MEDIA-ANALYSE PLAKAT, STICHPROBE. HTTP://WWW.AGMA-MMC.DE/03_FORSCHUNG/PLAKAT/ERHEBUNG_METHODE/STICHPROBE.ASP?TOPNAV=10&SUBNAV=199, 2012-05-11.

ARBEITSGEMEINSCHAFT MEDIA-ANALYSE E.V. (AG.MA), 2012D. DIE PMA HTTP://WWW.AGMA-MMC.DE/04_PRESSE/DETAIL.ASP?ID=1&TOPNAV=12&SUBNAV=379&JAHR=2004, 2012-05-11.

ARBEITSGEMEINSCHAFT MEDIA-ANALYSE E.V. (AG.MA), 2012E. MEDIALEISTUNGSWERTE HTTP://WWW.AGMA-MMC.DE/03_FORSCHUNG/DIE_MEDIA_ANALYSE.ASP?SUBNAV=73&TOPNAV=10, 2012-05-11.

AXHAUSEN, K.W., S. SCHÖNFELDER, J. WOLF, M. OLIVEIRA AND U. SAMAGA: 80 WEEKS OF GPS-TRACES: APPROACHES TO ENRICHING THE TRIP INFORMATION, ARBEITSBERICHT VERKEHRS- UND RAUMPLANUNG, 178, INSTITUT FÜR VERKEHRSPLANUNG UND TRANSPORTSYSTEME, ETH ZÜRICH, ZÜRICH, 2003

BARABASI, A. BURSTS: THE HIDDEN PATTERN BEHIND EVERYTHING WE DO, NEW YORK: DUTTON BOOKS, 2010

BARTELME, N.: GEOINFORMATIK: MODELLE, STRUKTUREN, FUNKTIONEN. SPRINGER, BERLIN. 2005.

BEKHOR, S. DOBLER, C. AND AXHAUSEN, K.W.: INTEGRATION OF ACTIVITY-BASED WITH AGENT-BASED MODELS: AN EXAMPLE FROM THE TEL AVIV MODEL AND MATSIM PAPER PRESENTED AT THE 90TH ANNUAL MEETING OF THE TRANSPORTATION RESEARCH BOARD, WASHINGTON, D.C., 2011

BELLMAN, R.E. : ADAPTIVE CONTROL PROCESSES. PRINCETON UNIVERSITY PRESS, PRINCETON, NEW YORK. 1961

BELLEMANS, T., KOCHAN, B., JANSSENS, D., WETS, G., ARENTZE, T., TIMMERMANS, H.: IMPLEMENTATION FRAMEWORK AND DEVELOPMENT TRAJECTORY OF THE FEATHERS ACTIVITY-BASED SIMULATION PLATFORM. TRANSPORTATION RESEARCH RECORD: JOURNAL OF THE TRANSPORTATION RESEARCH BOARD 2175, TRANSPORTATION RESEARCH BOARD, WASHINGTON D.C. 2010

BELUR V. DASARATHY, ED, NEAREST NEIGHBOR (NN) NORMS: NN PATTERN CLASSIFICATION TECHNIQUES, 1991

BERNSTEIN, D. KORNHAUSER, A.: AN INTRODUCTION TO MAP MATCHING FOR PERSONAL NAVIGATION ASSISTANTS. NEW JERSEY TIDE CENTER, 1998

BETHLEHEM, J.G.: WEIGHTING NONRESPONSE ADJUSTMENTS BASED ON AUXILIARY INFORMATION. IN: R.M. GROVES, D.A. DILLMAN, J.L. ELTINGE, AND R.J.A. LITTLE (EDS.), SURVEY NONRESPONSE. WILEY, NEW YORK. 2002

BOLLINGER, TONI: ASSOZIATIONSREGELN – ANALYSE EINES DATA MINING VERFAHRENS. INFORMATIK-SPEKTRUM, 19, 257-261, 1996

BUNDESAMT FÜR STATISTIK (BFS), 2011. AGGLOMERATIONEN UND METROPOLRÄUME: HTTP://WWW.BFS.ADMIN.CH/BFS/PORTAL/DE/INDEX/REGIONEN/11/GEO/ANALYSE_REGIONEN/04.HTML, 2011

BURROUGH P.A., VAN RIJN R. AND RIKKEN M,. SPATIAL DATA QUALITY AND ERROR ANALYSIS ISSUES: GIS FUNCTIONS AND ENVIRONMENTAL MODELING. IN: GOODCHILD MF AND STEYAERT LT (EDS), GIS AND ENVIRONMENTAL MODELING: PROGRESS AND RESEARCH ISSUES, WILEY & SONS, 1996

CAO, H., MAMOULIS, N. AND HEUNG, C.D.W.K.: MINING FREQUENT SPATIO-TEMPORAL SEQUENTIAL PATTERNS. IN ICDM (2005) PAGES 82-89. IEEE. 2005

CORMEN, T. H. LEISERSON, C. RIVEST, R. STEIN, C,: ALGORITHMEN – EINE EINFÜHRUNG. OLDENBOURG, MÜNCHEN, S. 598–604. WIEN 2004

CURTIN, K.M.: NETWORK ANALYSIS IN GEOGRAPHIC INFORMATION SCIENCE: REVIEW, ASSESSMENT, AND PROJECTIONS. CARTOGRAPHY AND GEOGRAPHIC INFORMATION SYSTEMS, 34(2), 103-111. 2007

DDS, 2011. HTTP://WWW.DDSGEO.DE/PRODUKTE/ARBEITSWEGEMATRIX.HTML, 2011-03-12

DDS, 2012. HTTP://WWW.DDSGEO.DE/PRODUKTE/TRIP-TRACER.HTML, 2012-05-11

DIJKSTRA, E. W.: A NOTE ON TWO PROBLEMS IN CONNEXION WITH GRAPHS. IN: NUMERISCHE MATHEMATIK, S. 269–271. 1959

DODGE, S., WEIBEL, R., LAUTENSCHÜTZ, A.-K.: TOWARDS A TAXONOMY OF MOVEMENT PATTERNS. INFORMATION VISUALIZATION, 7, 240-252. 2008

EFRON, B. BOOTSTRAP METHODS: ANOTHER LOOK AT THE JACKKNIFE, IN: THE ANNALS OF STATISTICS 7 (1), 1–26, 1979

EFRON, B. AND R. J. TIBSHIRANI, AN INTRODUCTION TO THE BOOTSTRAP, CHAPMAN & HALL, 1993

EGENHOFER, M.: REASONING ABOUT BINARY TOPOLOGICAL RELATIONS, IN: GÜNTHER, O. (EDT.), ADVANCES IN SPATIAL DATABASES - PROCEEDINGS OF THE 2ND SYMPOSIUM ON LARGE SPATIAL DATABASES SSD '91, LECTURE NOTES IN COMPUTER SCIENCE 525, PP. 143-160. 1991

ELLERSIEK, T., LIEBIG, T. HECKER D. AND KÖRNER. C.: ANALYSE VON RAUM-ZEITLICHEN BEWEGUNGSMUSTERN AUF BASIS VON BLUETOOTH-SENSOREN. IN ANGEWANDTE GEOINFORMATIK 2012, BEITRÄGE ZUM 24. AGIT SYMPOSIUM SALZBURG, IM ERSCHEINEN, 2012

ENGEL, D. AND HOFSÄSS, M.: PRAXISBUCH MEDIAPLANUNG. BERLIN, 2003.

ESOMAR, 2009. GLOBAL GUIDELINES ON OUT-OF-HOME AUDIENCE MEASUREMENT VERSION 1.0, ISBN: 92-831-0234-7, PAGES 28, 2009

FAHRMEIR, L. HAMERLE, A. UND TUTZ, G.: STATISTIK. DER WEG ZUR DATENANALYSE. 2. AUFLAGE. SPRINGER, BERLIN, 1999

FAW 2011 (FACHVERBAND AUSSENWERBUNG E.V.): OUT-OF-HOME-MEDIEN. HTTP://WWW.FAW-EV.DE/32C63818FD88DD23A2E68EAD86B8EC87/DE/FAW/OUT-OF-HOME-MEDIEN/INDEX.HTML, 2011-03-11

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. AND UTHURUSAMY, R., "PREFACE", IN ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING. CAMBRIDGE, USA, AAAI PRESS, PP. 560. 1996

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P.: FROM DATA MINING TO KNOWLEDGE DISCOVERY IN DATABASES. IN: FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. (EDS.), ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING. AMERICAN ASSOCIATION FOR ARTIFICAL INTELLIGENCE, MENLO PARK, 37-54. 1996

GIANNOTTI F., NANNI M. AND PEDRESCHI D.: EFFICIENT MINING OF TEMPORALLY ANNOTATED SEQUENCES. IN: PROC. SPATIAL DATA MINING 2006 (SDM'06). SDM, PP 346-357. 2006

GIANNOTTI F., NANNI M., PEDRESCHI D. AND PINELLI F.: TRAJECTORY PATTERN MINING. IN: PROC. OF THE 13TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD'07). ACM, PP 330-339. 2007

GIANOTTI, F., PEDRESCHI, D.: MOBILITY, DATA MINING AND PRIVACY: A VISION OF CONVERGENCE. IN: GIANOTTI, F., PEDRESCHI, D. (EDS.), 2008. MOBILITY, DATA MINING AND PRIVACY. GEOGRAPHIC KNOWLEDGE DISCOVERY. SPRINGER, BERLIN, 1-11. 2008

GOLLEDGE, R.G. AND R.J. STIMSON SPATIAL BEHAVIOR, THE GUILFORD PRESS, NEW YORK. LNCS, VOL. 5211. SPRINGER, PP. 440–456. 1997

GONZALEZ, M.C. HIDALGO, C.A., UND BARABASI, A.: UNDERSTANDING INDIVIDUAL HUMAN MOBILITY PATTERNS. NATURE, 453(7196): 779-782, 2008

GÖRZ G., ROLLINGER, C.-R., SCHNEEBERGER, J. (ED.): HANDBUCH DER KÜNSTLICHEN INTELLIGENZ. MÜNCHEN: OLDENBOURG, 4. AUFLAGE, 2003

GOOGLE (HRSG.): GOOGLE MAPS 2012. HTTP://MAPS.GOOGLE/MAPS. 2012-06-15

GROSSKREUTZ, H. RÜPING, S. AND WROBEL, S.: "TIGHT OPTIMISTIC ESTIMATES FOR FAST SUBGROUP DISCOVERY," IN PROC. OF THE 2008 EUROPEAN CONFERENCE ON MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES - PART I (ECML/PKDD'08). 2008

GUDMUNDSSON J, KREVELD M, SPECKMANN B.: EFFICIENT DETECTION OF PATTERNS IN 2D TRAJECTORIES OF MOVING POINTS. IN: GEOINFORMATICA 11(2):195-215. 2007

HAEB-UMBACH, R. AND PESCHKE, S.: "A NOVEL SIMILARITY MEASURE FOR POSITIONING CELLULAR PHONES BY A COMPARISON WITH A DATABASE OF SIGNAL POWER LEVELS," VOL. 56, NO. 1, PAGES 368–372, 2007

HAEGERSTRAND, T.: WHAT ABOUT PEOPLE IN REGIONAL SCIENCE? PAPERS OF THE REGIONAL SCIENCE ASSOCIATION 24, 7-21. 1970

HAN, Q. ARENTZE, T. TIMMERMANS, H. JANSSENS, D. WETS, G.: A MULTI-AGENT MODELLING APPROACH TO SIMULATE DYNAMIC ACTIVTY-TRAVEL PATTERNS. IN: BAZZAN, A.L.C., KLÜGL, F. (EDS.). MULTI-AGENT SYSTEMS FOR TRAFFIC AND TRANSPORTATION ENGINEERING. IDEA GROUP REFERENCE, HERSHEY PAGES, 36-56. 2009A

HAN, Q., ARENTZE, T., TIMMERMANS, H., JANSSENS, D., WETS, G.: DEVELOPING DYNAMIC MODELS OF ACTIVITY-TRAVEL BEHAVIOR: PRINCIPLES, MECHANISMS, CHALLENGES IN DATA COLLECTION AND METHODOLOGICAL ISSUES. PROCEEDINGS OF THE 88TH ANNUAL MEETING OF THE TRANSPORTATION RESEARCH, 2009B

HART, P. E. NILSSON, N. J. RAPHAEL, B.: A FORMAL BASIS FOR THE HEURISTIC DETERMINATION OF MINIMUM COST PATHS, IEEE TRANSACTIONS ON SYSTEMS SCIENCE AND CYBERNETICS SSC4 (2), PP. 100–107, 1968

HECKER, D. MAY, M. SCHEIDER, S. UND VOSS, A.: PUNKTWERBUNG. IN GEOBIT, NO.8, PAGES 30-31, 2004

HECKER, D. WARMELINK, F.: DIE EINSATZMÖGLICHKEITEN VON GEOINFORMATIONSSYSTEMEN (GIS) IM VERLAGSWESEN. EINE EINFÜHRUNG ZUM THEMA GIS UND GEOMARKETING. BVDA, PAGES 1-74, 2005

HECKER, D. STANGE, H. KÖRNER, C. MAY, M.: SAMPLE BIAS DUE TO MISSING DATA IN MOBILITY SURVEYS. IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS - ICDM 2010: PAGES 241-248, 2010A

HECKER, D. KÖRNER C. UND MAY, M.: RÄUMLICH DIFFERENZIERTE REICHWEITEN FÜR DIE AUßENWERBUNG. IN ANGEWANDTE GEOINFORMATIK 2010, BEITRÄGE ZUM 22. AGIT SYMPOSIUM SALZBURG, PAGES 194-203, 2010B

HECKER, D. KÖRNER, C. STREICH, H. UND HOFMANN, U.: A SENSITIVITY ANALYSIS FOR THE SELECTION OF BUSINESS CRITICAL GEODATA IN SWISS OUTDOOR ADVERTISEMENT. IN GISCIENCE 2010, EXTENDED ABSTRACTS VOLUME, 2010C

HECKER, D. KÖRNER C. UND MAY, M.: ROBUSTNESS ANALYSES FOR REPEATED MOBILITY SURVEYS IN OUTDOOR ADVERTISING. IN PROC. OF THE 1TH INTERNATIONAL CONFERENCE ON SPATIAL DATA MINING AND GEOGRAPHICAL KNOWLEDGE SERVICES (ICSDM'11), PAGES 148-153. 2011A

HECKER, D. KÖRNER, C. STANGE, H. SCHULZ D. UND MAY, M: MODELING MICRO-MOVEMENT VARIABILITY IN MOBILITY STUDIES. IN LECTURE NOTES IN GEOINFORMATION AND CARTOGRAPHY, VOLUME 1, PART 2, PAGES 121-140, 2011B

HECKER, D. KÖRNER C. UND MAY, M.: CHALLENGES AND ADVANTAGES OF USING GPS DATA IN OUTDOOR ADVERTISEMENT. IN PROC. OF THE 3TH CONFERENCE ON GEOINFORMATIK - GEOCHANGE, PAGES 257-260. AKADEMISCHE VERLAGSGESELLSCHAFT. 2011C

HERNANDEZ, D.: QUANTITIVE REPRESENTATION OF SPATIAL KNOWLEDGE. SPRINGER, BERLIN. 1994.

HPI 2012. ANSÄTZE ZUR KARTOGRAPHISCHEN GESTALTUNG VON 3D-STADMODELLEN. HTTP://WWW.HPI.UNI-POTSDAM.DE/DOELLNER/PUBLICATIONS/YEAR/2011/1425/STD11.HTML, 2012-05-11

HUDEC, M.: DATA MINING. EIN NEUES PARADIGMA DER ANGEWANDTEN STATISTIK. IN: A JOURNAL OF STATISTICS 31, NR. 1, 55-65. 2002

HWANG, S.Y. LIU, Y.H. CHIU, J.K. LIM, E.P.: MINING MOBILE GROUP PATTERS: A TRAJECTORY-BASED APPROACH. IN: PROC. OF THE 9TH PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (PAKDD'05). SPRINGER, PP 713-718. 2005

KANG, J. AND YONG, H-S.: MINING SPATIO-TEMPORAL PATTERNS IN TRAJECTORY DATA. JIPS 6(4) THE JOURNAL OF INFORMATION PROCESSING SYSTEMS, VOLUME 6: 521-536. 2010

KAPLAN, E. L. UND MEIER, P.: NON-PARAMETRIC ESTIMATION FROM INCOMPLETE OBSERVATIONS. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, 53: SEITEN 457–481, 1958

KIM, J. AND BOHACEK, S.: EXPLOITING MULTIHOP DIVERSITY THROUGH EFFICIENT LOCALIZED SEARCHING WITH CDMA AND ROUTE METRIC-BASED POWER CONTROL THE 9-TH INTERNATIONAL SYMPOSIUM ON MODELING, ANALYSIS AND SIMULATION OF WIRELESS AND MOBILE SYSTEMS (MSWIM),TORREMOLINOS, MALAGA, SPAIN, OCTOBER 2006

KLEIN, R.: ALGORITHMISCHE GEOMETRIE. SPRINGER, HEIDELBERG, 2005

KLEINBAUM, D. G. UND KLEIN, M.: SURVIVAL ANALYSIS. STATISTICS FOR BIOLOGY AND HEALTH. SPRINGER, 2005

KLÖSGEN, W. AND MAY, M.: "SPATIAL SUBGROUP MINING INTEGRATED IN AN OBJECT-RELATIONAL SPATIAL DATABASE," IN PROC. OF THE 6TH EUROPEAN CONFERENCE ON DATA MINING AND KNOWLEDGE DISCOVERY (PKDD), 2002, PP. 275–286. 2002

KOSCHNIK, W. J.: DIE ZUKUNFT DER KLASSISCHEN ELEKTRONISCHEN MEDIEN. IN: FOCUS-JAHRBUCH 2011.

KÖRNER, C. HECKER, D. MAY, M. UND WROBEL, S.: VISIT POTENTIAL: A COMMON VOCABULARY FOR THE ANALYSIS OF ENTITY-LOCATION INTERACTIONS IN MOBILITY APPLICATIONS. IN PROC. OF THE 13TH INTERNATIONAL CONFERENCE ON GEOGRAPHIC INFORMATION SCIENCE (AGILE'10), 2010

KÖRNER, C. HECKER, D. KRAUSE-TRAUDES, M. MAY, M. SCHEIDER S., SCHULZ, D. STANGE, H. UND WROBEL, S.: SPATIAL DATA MINING IN PRACTICE: PRINCIPLES AND CASE STUDIES. IN C. SOARES UND R. GHANI, EDITORS, DATA MINING FOR BUSINESS APPLICATIONS. IOS PRESS, 2010

KÖRNER, C.: MODELING VISIT POTENTIAL OF GEOGRAPHIC LOCATIONS BASED ON MOBILITY DATA, DISSERTATION UNIVERSITÄT BONN, 2012

KRAHL, D. WINDHEUSER, U., ZICK, K.: DATA MINING – EINSATZ IN DER PRAXIS. ADDISON WESLEY-LONGMAN VERLAG, BONN. 1998

LAUBE, P. IMFELD, S.: ANALYZING RELATIVE MOTION WITHIN GROUPS OF TRACKABLE MOVING POINT OBJECTS. IN: PROC. OF THE 2ND INTERNATIONAL CONFERENCE ON GEOGRAPHIC INFORMATION SCIENCE (GISCIENCE'02). SPRINGER, PP 132–144. 2002

LEVINSON, S. C.: SPACE IN LANGUAGE AND COGNITION: EXPLORATIONS IN COGNITIVE DIVERSITY. CAMBRIDGE: CAMBRIDGE UNIVERSITY PRESS. 2003

LIEBIG, T. STANGE, H. HECKER, D. MAY, M. KÖRNER C. AND HOFMANN. U.: A GENERAL PEDESTRIAN MOVEMENT MODEL FOR THE EVALUATION OF MIXED INDOOR-OUTDOOR POSTER CAMPAIGNS. IN PROC. OF THE THIRD WORKSHOP ON PERVASIVE ADVERTISING AND SHOPPING, 2010

LITTLE, R. J. A. AND RUBIN, D.B.: STATISTICAL ANALYSIS WITH MISSING DATA. WILEY SERIES IN PROBABILITY & MATHEMATICAL STATISTICS. JOHN WILEY & SONS, 1987

LOU, Y., ZHANG, C., ZHENG, Y., XIE, X., WANG, W., HUANG, Y.: MAP-MATCHING FOR LOW-SAMPLING-RATE GPS TRAJECTORIES. PROCEEDINGS OF THE 17TH ACM SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, NOVEMBER 4-6, 2009, SEATTLE, USA. 2009

MARCHAL,F., J.K. HACKNEY, AND K.W. AXHAUSEN: EFFICIENT MAP-MATCHING OF LARGE GPS DATA SETS - TESTS ON A SPEED MONITORING EXPERIMENT IN ZURICH, TRANSPORTATION RESEARCH RECORD, 1935, 93-100, 2006

MARCHAL, P. YUAN, S. FLAVIGNY, P.O.:PERSON-BASED GPS SURVEYS IN FRANCE: „LILLE EXPERIMENT" BY ISL, AND GPS SUBSET IN THE FRENCH NATIONAL TRAVEL SURVEY (ENTD 2007-2008). COST 355 PROJECT MEETING. HTTP://COST355.INRETS.FR/IMG/PPT/WG3-TORINO-051007-MARCHAL-YUAN-FLAVIGNY-GPS_V2DU05100700.PPT. 2010

MAY, M. HECKER, D. KÖRNER, C. SCHEIDER S. UND SCHULZ, D.: A VECTOR-GEOMETRY BASED SPATIAL KNN-ALGORITHM FOR TRAFFIC FREQUENCY PREDICTIONS. IN PROC. OF THE 2008 IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS (ICDMW '08), PAGES 442-447. IEEE COMPUTER SOCIETY, 2008A

MAY, M. SCHEIDER, S. RÖSLER, R. SCHULZ, D. HECKER, D.: PEDESTRIAN FLOW PREDICTION IN EXTENSIVE ROAD NETWORKS USING BIASED OBSERVATIONAL DATA. PROCEEDINGS OF THE 16TH ACM SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, NOVEMBER 5-7, 2008, IRVINE, USA, 1-4. 2008B

MAY, M. KÖRNER, C. HECKER, D. PASQUIER, M. HOFMANN, U. UND MENDE, F.: HANDLING MISSING VALUES IN GPS SURVEYS USING SURVIVAL ANALYSIS: A GPS CASE STUDY OF OUTDOOR ADVERTISING. IN ADKDD '09: PROCEEDINGS OF THE THIRD INTERNATIONAL WORKSHOP ON DATA MINING UND AUDIENCE INTELLIGENCE FOR ADVERTISING, PAGES 78-84, NEW YORK, NY, USA, 2009

MAY, M. KÖRNER, C. HECKER, D. PASQUIER, M. HOFMANN, U. UND MENDE. F.: MODELLING MISSING VALUES FOR AUDIENCE MEASUREMENT IN OUTDOOR ADVERTISING USING GPS DATA. IN GI JAHRESTAGUNG, VOLUME 154 OF LNI, PAGES 3993-4006. GI, 2009

MCNETT, M. AND VOELKER, G. M.: ACCESS AND MOBILITY OF WIRELESS PDA USERS. IN: ACM SIGMOBILE MOBILE COMPUTING AND COMMUNICATIONS REVIEW. VOLUME 9 ISSUE 2, APRIL 2005. PAGES 40-55.

MEISTER, K., BALMER, M., CIARI, F., HORNI, A., RIESER, M., WARAICH, R.A., AXHAUSEN, K.W.: LARGE-SCALE AGENT-BASED TRAVEL DEMAND OPTIMIZATION APPLIED TO SWITZERLAND, INCLUDING MODE CHOICE. PROCEEDINGS OF THE 12TH WORLD CONFERENCE ON TRANSPORTATION RESEARCH, JULY 11-15, 2010, LISBON, PORTUGAL. 2010

MGE (MARKETING GEOGRAPHICS ENVIRONMENT DATA), 2012. MOBITEST. HTTP://WWW.MGEDATA.COM/DE/HW-UND-SW-PRODUKTE/CUSTOM-PRODUKTE/MOBITEST/MOBITEST-SL, 2012-05-11

MIKROZENSUS 2005, BUNDESAMT FÜR RAUMENTWICKLUNG (BFS), STATISTIK DER SCHWEIZ. HTTP://WWW.BFS.ADMIN.CH/BFS/PORTAL/DE/INDEX/NEWS/PUBLIKATIONEN.HTML?PUBLICATIONID=2700, 2012-05-11

MIKROZENSUS 2010, BUNDESAMT FÜR RAUMENTWICKLUNG (BFS), STATISTIK DER SCHWEIZ. HTTP://WWW.BFS.ADMIN.CH/BFS/PORTAL/DE/INDEX/NEWS/PUBLIKATIONEN.DOCUMENT.155067.PDF, 2012-05-11

MITCHELL, T. M. MACHINE LEARNING, MCGRAW-HILL, 1997

MOBILITÄT IN DEUTSCHLAND (MID), 2008A. ERGEBNISBERICHT. BONN, BERLIN.

MOBILITÄT IN DEUTSCHLAND (MID), 2008B. KURZBERICHT. BONN, BERLIN.

MONREALE, A., PINELLI, F., TRASARTI, R. AND GIANNOTTI F.: WHERENEXT: A LOCATION PREDICTOR ON TRAJECTORY PATTERN MINING. KDD '09. PROCEEDINGS OF THE 15TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. PAGES 637-646. 2009

MURTHY, S.K.: AUTOMATIC CONSTRUCTION OF DECISION TREES FROM DATA: A MULTIDISCIPLINARY SURVEY. DATA MINING AN KNOWLEDGE DISCOVERY, VOL. 2, PAGES 345-389, 1998

NANNI, M. AND PEDRESCHI, D.: TIME-FOCUSED CLUSTERING OF TRAJECTORIES OF MOVING OBJECTS. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS. NOVEMBER 2006, VOLUME 27, ISSUE 3, PP 267-289.

NAVTEQ, 2012. NAVTEQ DATEN. HTTP://CORPORATE.NAVTEQ.COM/DEUTSCH/PRODUCTS_DATA.HTM, 2012-05-11

OSM (HRSG.): OPEN STREET MAPS 2012. HTTP://OPENSTREETMAP.ORG. 2012-06-15

OPFER, G.: REICHWEITENMESSUNG FÜR PLAKATANSCHLAGSTELLEN IN DEUTSCHLAND. IN: FOCUS-JAHRBUCH 2005. BEITRÄGE ZU WERBE- UND MEDIAPLANUNG, MARKT-, KOMMUNIKATIONS- UND MEDIAFORSCHUNG. MÜNCHEN, S. 291-314. 2005

PARKES, D. AND THRIFT, N. J.: TIMES, SPACES, AND PLACES: A CHRONOGEOGRAPHIC PERSPECTIVE. J. WILEY PUBLISHERS. CHICHESTER ENGLAND, 1980.

PASQUIER, M.: PLAKATWIRKUNGSFORSCHUNG. THEORETISCHE GRUNDLAGEN UND PRAKTISCHE ANSÄTZE, UNIVERSITÄTSVERLAG FREIBURG SCHWEIZ. 1997

PASQUIER, M. HOFMANN, U., MENDE, F. H., MAY, M. HECKER, D. KÖRNER, C.: MODELLING AND PROSPECTS OF THE AUDIENCE MEASUREMENT FOR OUTDOOR ADVERTISING BASED ON DATA. PROCEEDINGS OF THE 8TH INTERNATIONAL CONFERENCE ON SURVEY METHODS IN TRANSPORT, MAY 25-31, 2008, ANNECY, FRANCE. 2008

PELEKIS, N. KOPANAKIS, I. NTOUTSI, I. MARKETOS, G. ANDRIENKO, G. THEODORIDIS, Y.: SIMILARITY SEARCH IN TRAJECTORY DATABASES, IN: PROC. OF THE 14TH IEEE INTERNATIONAL SYMPOSIUM ON TEMPORAL REPRESENTATION AND REASONING (TIME 2007). IEEE COMPUTER SOCIETY PRESS, PP 129-140. 2007

PIATETSKY-SHAPIRO, G.: DATA MINING AND KNOWLEDGE DISCOVERY 1996 TO 2005: OVERCOMING THE HYPE AND MOVING FROM "UNIVERSITY" TO "BUSINESS" AND "ANALYTICS". IN: DATA MINING AND KNOWLEDGE DISCOVERYM 15 NR. 1, S. 99-105. 2007

QUINLAN, J. R.: C4.5: PROGRAMS FOR MACHINE LEARNING. MORGAN KAUFMANN PUBLISHERS, 1993

RIESER, M.: ADDING TRANSIT TO AN AGENT-BASED TRANSPORTATION SIMULATION: CONCEPTS AND IMPLEMENTATION PHD THESIS, VSP, TU BERLIN, GERMANY, 2010

RINZIVILLO, S. PEDRESCHI, D. NANNI, M. GIANNOTTI, F. ANDRIENKO, N. ANDRIENKO, G.: VISUALLY DRIVEN ANALYSIS OF MOVEMENT DATA BY PROGRESSIVE CLUSTERING. IN: INFORMATION VISUALIZATION 7(3):225-239. 2008

RUBIN, D.B.: INFERENCE AND MISSSING DATA. BIOMETRIKA, 63(3): 581-592, 1976

SALTENIS, S. JENSEN, C.S. LEUTENEGGER, S. T. AND LOPEZ. M. A.: INDEXING THE POSITIONS OF CONTINUOUSLY MOVING OBJECTS. IN SIGMOD, 2000

SCHEIER, C.: WIE WIRKEN PLAKATE? IN: FOCUS-JAHRBUCH 2005. BEITRÄGE ZU WERBE- UND MEDIAPLANUNG, MARKT-, KOMMUNIKATIONS- UND MEDIAFORSCHUNG. MÜNCHEN 2005, S. 265-290. 2005

SCHLICH, R. UND AXHAUSEN, K.W.: HABITUAL TRAVEL BEHAVIOUR: EVIDENCE FROM A SIX-WEEK TRAVEL DIARY. TRANSPORTATION, 30: 13-36, 2003

SCHLOSSBAUER, S.: HANDBUCH DER AUSSENWERBUNG VERLAG MD MEDIEN DIENSTE; AUFLAGE: 1. AUFL. 1997

SCHREINEMACHER, J. KÖRNER, C. HECKER, D. BARETH, G.: ANALYZING TEMPORAL USAGE PATTERNS OF STREET SEGMENTS BASED ON GPS- DATA – A CASE STUDY IN SWITZERLAND. IN PROC. OF THE 15TH INTERNATIONAL CONFERENCE ON GEOGRAPHIC INFORMATION SCIENCE (AGILE'10), IM ERSCHEINEN, 2012

SCHUESSLER, N., AXHAUSEN, K.W.: MAP-MATCHING OF GPS TRACES ON HIGH-RESOLUTION NAVIGATION NETWORKS USING THE MULTIPLE HYPOTHESIS TECHNIQUE (MHT). WORKING PAPER, 568, INSTITUTE FOR TRANSPORT PLANNING AND SYSTEM (IVT), ETH ZÜRICH. 2009

SESTER, M. FEUERHAKE, U. KUNTZSCH, C. UND ZHANG, L.: REVEALING UNDERLYING STRUCTURE AND BEHAVIOUR FROM MOVEMENT DATA. JOURNAL KI KÜNSTLICHE INTELLIGENZ, THEMENHEFT SPATIOTEMPORAL MODELING AND ANALYSIS, IM ERSCHEINEN, 2012

SHAKHNAROVISH, D. AND INDYK, ED.: NEAREST-NEIGHBOR METHODS IN LEARNING AND VISION. MIT PRESS. 2005

SISSORS, J.Z. BARON, R.B.: ADVERTISING MEDIA PLANNING. MCGRAW-HILL, CHAPTER 4-5. 2002

SONG, C., QU, Z.,BLUMM, N., BARABÁSI, A.-L.: LIMITS OF PREDICTABILITY IN HUMAN MOBILITY, SCIENCE 327, 1018-1021. 2010

SPACCAPIETRA, S., PARENT, C., DAMIANI, M. L., DE MACEDO, J. A., PORTO, F., VANGENOT, C.: A CONCEPTUAL VIEW ON TRAJECTORIES. DATA & KNOWLEDGE ENGINEERING, 65, 126-146. 2008

SPIEGEL 2011. WIE HANDYDATEN-SCHNÜFFELEI UNS HELFEN KANN. HTTP://WWW.SPIEGEL.DE/WISSENSCHAFT/MENSCH/SPIONAGE-IN-GUTER-MISSION-WIE-HANDYDATEN-SCHNUEFFELEI-UNS-HELFEN-KANN-A-771085.HTML, 2012-04-03

SPINDLER, R., 2008. KLASSIFIKATION VON GPS-DATEN NACH FORTBEWEGUNGSMITTELN. DIPLOMARBEIT (UNVERÖFFENTLICHT), RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN, DEUTSCHLAND. 2008

SPR+ 2011 (SWISS POSTER RESEARCH PLUS 2011): GEWICHTUNGSKRITERIEN. HTTP://WWW.SPR-PLUS.CH/MAIN.ASPX?TABID=297, 2011-03-09

SPR+ 2012 (SWISS POSTER RESEARCH PLUS 2012): STRAßENSTUDIE METHODENSTECKBRIEF. HTTP://WWW.SPR-PLUS.CH/MAIN.ASPX?TABID=289, 2012-05-11

STANGE, H. LIEBIG, T. HECKER, D. ANDRIENKO, G. UND ANDRIENKO, N.: ANALYTICAL WORKFLOW OF MONITORING HUMAN MOBILITY IN BIG EVENT SETTINGS USING BLUETOOTH. IN PROCEEDINGS OF THE 3RD ACM SIGSPATIAL INTERNATIONAL WORKSHOP ON INDOOR SPATIAL AWARENESS.PAGES 51-58, 2011

STIFTUNG WERBESTATISTIK SCHWEIZ, 2011. HTTP://WWW.WERBESTATISTIK.CH/DOWNLOAD.PHP?ID=26_684C49CC, 2012-03-02

STOPHER, P., XU, M. AND FITZGERALD, C.: ASSESSING THE ACCURACY OF THE SYDNEY HOUSEHOLD TRAVEL SURVEY WITH GPS, TRANSPORTATION, 34 (6), 723-741, 2007

STRÖER, 2012. PRODUKTE. HTTP://WWW.STROEER.DE/OUT-OF-HOME-MEDIEN.MEDIENUNDANGEBOTE.0.HTML, 2012-05-11

SWISSTOPO, 2012. PRODUKTE. HTTP://WWW.SWISSTOPO.ADMIN.CH/INTERNET/SWISSTOPO/DE/HOME/PRODUCTS/LANDSCAPE/VECTOR25.HTML, 2012-05-11

TAO, Y., PAPADIAS, D., SUN, J.: THE TPR*-TREE: AN OPTIMIZED SPATIO-TEMPORAL ACCESS METHOD FOR PREDICTIVE QUERIES.IN: VLDB '03 PROCEEDINGS OF THE 29TH INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES . VOLUME 29. PAGES 790-801. 2003

TELEATLAS/TOMTOM, 2012. PRODUCTS. HTTP://WWW.TOMTOM.COM/EN_GB/LICENSING/PRODUCTS/, 2012-05-11

VOSS, H. SCHEIDER, S. HECKER, D. UND VOSS, A.: GIS-VISIONEN: FAST ALLE DATEN WERDEN VERORTET SEIN. IN GEOBIT, NO.11, PAGE 18, 2004

WANG, Y., LIM, E-P. AND HWANG, S-Y.: ON MINING GROUP PATTERNS OF MOBILE USERS, 14TH INTERNATIONAL CONFERENCE ON DATABASE AND EXPERT SYSTEMS APPLICATIONS (DEXA2003), PRAGUE, CZECH REPUBLIC, SEPTEMBER, 2003

WEGENER, D. HECKER, D. KÖRNER, C. MAY, M. UND MOCK, M.: PARALLELIZATION OF R-PROGRAMS WITH GRIDR IN A GPS TRAJECTORY MINING APPLICATION. IN PROC. OF THE 1ST UBIQUITOUS KNOWLEDGE DISCOVERY WORKSHOP (UKD'08), 2008

WIKIPEDIA 2012. OCKHAMS RASIERMESSER. HTTP://DE.WIKIPEDIA.ORG/WIKI/OCKHAMS_RASIERMESSER, 2012-05-11

WITTEN, I.H., FRANK, E.: DATA MINING. PRAKTISCHE WERKZEUGE UND TECHNIKEN FÜR DAS MASCHINELLE LERNEN. CARL HANSER VERLAG, MÜNCHEN. 2001

WOLF, J.: USING GPS DATA LOGGERS TO REPLACE TRAVEL DIARIES IN THE COLLECTION OF TRAVEL DATA, DISSERTATION, GEORGIA INSTITUTE OF TECHNOLOGY, SCHOOL OF CIVIL AND ENVIRONMENTAL ENGINEERING, ATLANTA, GEORGIA, 2000

WOLF, J., GUENSLER, R. AND BACHMAN, W.: ELIMINATION OF THE TRAVEL DIARY: AN EXPERIMENT TO DERIVE TRIP PURPOSE FROM GPS TRAVEL DATA, TRANSPORTATION RESEARCH RECORD, 1768, 125-134, 2001

WROBEL, S.: "AN ALGORITHM FOR MULTI-RELATIONAL DISCOVERY OF SUBGROUPS," IN PROC. OF THE PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES (PKDD'97). SPRINGER, PP. 78–87. 1997

YAN, Z. CHAKRABORTY, D. PARENT, C. SPACCAPIETRA, S. UND ABERER, K.: SEMITRI: A FRAMEWORK FOR SEMANTIC ANNOTATION OF HETEROGENEOUS TRAJECTORIES. IN PROCEEDINGS OF THE 14TH INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY (EDBT´11), SEITE 259-270. ACM, 2011A

YAN, Z. GIATRAKOS, N. KATSIKAROS, V. PELEKIS, N. UND THEODORIDIS, Y. SETRASTREAM: SEMANTIC-AWARE TRAJECTORY CONSTRUCTION OVER STREAMING MOVEMENT DATA. IN PROCEEDINGS OF THE 12TH INTERNATIONAL SYMPOSIUM ON ADVANCES IN SPATIAL AND TEMPORAL DATABASES (SSTD´11), SEITEN 367-385, 2011B

YANG, Y. HU, M.: TRAJPATTERN: MINING SEQUENTIAL PATTERNS FROM IMPRECISE TRAJECTORIES OF MOBILE OBJECTS. IN: PROC. OF 10TH INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY. SPRINGER, PP 664-681. 2006

ZHENG, Y. ZHANG, L. XIE, X. MA, W.Y.: MINING INTERESTING LOCATIONS AND TRAVEL SEQUENCES FROM GPS TRAJECTORIES. IN: PROC. OF THE 18TH INTERNATIONAL WORLD WIDE WEB CONFERENCE (WWW'09). ACM, PP 791-800. 2009

ZIMMERMANN, D. BAUMANN, J. LAYH, A. LANDSTORFER, F. HOPPE, R. AND WOLFLE, G.: DATABASE CORRELATION FOR POSITIONING OF MOBILE TERMINALS IN CELLULAR NETWORKS USING WAVE PROPAGATION MODELS. IN VEHICULAR TECHNOLOGY CONFERENCE, VTC2004-FALL. 2004 IEEE 60TH, VOL. 7, , PAGES. 4682– 4686, 2004

ZMUD, J. AND WOLF, J.: IDENTIFYING THE CORRELATIONS OF TRIP MISREPORTING – RESULTS FOR THE CALIFORNIA STATEWIDE HOUSEHOLD TRAVEL SURVEY GPS STUDY. IN: PROC. OF THE 10TH INTERNATIONAL CONFERENCE ON TRAVEL BEHAVIOUR RESEARCH, 2003

## 11. TEILPUBLIKATIONEN

HECKER, D. STANGE, H. KÖRNER, C. MAY, M.: SAMPLE BIAS DUE TO MISSING DATA IN MOBILITY SURVEYS. IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS - ICDM 2010: PAGES 241-248, 2010A

HECKER, D. KÖRNER C. UND MAY, M.: RÄUMLICH DIFFERENZIERTE REICHWEITEN FÜR DIE AUßENWERBUNG. IN ANGEWANDTE GEOINFORMATIK 2010, BEITRÄGE ZUM 22. AGIT SYMPOSIUM SALZBURG, PAGES 194-203, 2010B

HECKER, D. KÖRNER, C. STREICH, H. UND HOFMANN, U.: A SENSITIVITY ANALYSIS FOR THE SELECTION OF BUSINESS CRITICAL GEODATA IN SWISS OUTDOOR ADVERTISEMENT. IN GISCIENCE 2010, EXTENDED ABSTRACTS VOLUME, 2010C

HECKER, D. KÖRNER C. UND MAY, M.: ROBUSTNESS ANALYSES FOR REPEATED MOBILITY SURVEYS IN OUTDOOR ADVERTISING. IN PROC. OF THE 1TH INTERNATIONAL CONFERENCE ON SPATIAL DATA MINING AND GEOGRAPHICAL KNOWLEDGE SERVICES (ICSDM'11), PAGES 148-153. 2011A

HECKER, D. KÖRNER, C. STANGE, H. SCHULZ D. UND MAY, M: MODELING MICRO-MOVEMENT VARIABILITY IN MOBILITY STUDIES. IN LECTURE NOTES IN GEOINFORMATION AND CARTOGRAPHY, VOLUME 1, PART 2, PAGES 121-140, 2011B

HECKER, D. KÖRNER C. UND MAY, M.: CHALLENGES AND ADVANTAGES OF USING GPS DATA IN OUTDOOR ADVERTISEMENT. IN PROC. OF THE 3TH CONFERENCE ON GEOINFORMATIK - GEOCHANGE, PAGES 257-260. AKADEMISCHE VERLAGSGESELLSCHAFT. 2011C

KÖRNER, C. HECKER, D. MAY, M. UND WROBEL, S.: VISIT POTENTIAL: A COMMON VOCABULARY FOR THE ANALYSIS OF ENTITY-LOCATION INTERACTIONS IN MOBILITY APPLICATIONS. IN PROC. OF THE 13TH INTERNATIONAL CONFERENCE ON GEOGRAPHIC INFORMATION SCIENCE (AGILE'10), 2010A

KÖRNER, C. HECKER, D. KRAUSE-TRAUDES, M. MAY, M. SCHEIDER S., SCHULZ, D. STANGE, H. UND WROBEL, S.: SPATIAL DATA MINING IN PRACTICE: PRINCIPLES AND CASE STUDIES. IN C. SOARES UND R. GHANI, EDITORS, DATA MINING FOR BUSINESS APPLICATIONS. IOS PRESS, 2010B

MAY, M. HECKER, D. KÖRNER, C. SCHEIDER S. UND SCHULZ, D.: A VECTOR-GEOMETRY BASED SPATIAL KNN-ALGORITHM FOR TRAFFIC FREQUENCY PREDICTIONS. IN PROC. OF THE 2008 IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS (ICDMW '08), PAGES 442-447. IEEE COMPUTER SOCIETY, 2008A

MAY, M. SCHEIDER, S. RÖSLER, R. SCHULZ, D. HECKER, D.: PEDESTRIAN FLOW PREDICTION IN EXTENSIVE ROAD NETWORKS USING BIASED OBSERVATIONAL DATA. PROCEEDINGS OF THE 16TH ACM SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, NOVEMBER 5-7, 2008, IRVINE, USA, 1-4. 2008B

MAY, M. KÖRNER, C. HECKER, D. PASQUIER, M. HOFMANN, U. UND MENDE, F.: HANDLING MISSING VALUES IN GPS SURVEYS USING SURVIVAL ANALYSIS: A GPS CASE STUDY OF OUTDOOR ADVERTISING. IN ADKDD '09: PROCEEDINGS OF THE THIRD INTERNATIONAL WORKSHOP ON DATA MINING UND AUDIENCE INTELLIGENCE FOR ADVERTISING, PAGES 78-84, NEW YORK, NY, USA, 2009A

MAY, M. KÖRNER, C. HECKER, D. PASQUIER, M. HOFMANN, U. UND MENDE. F.: MODELLING MISSING VALUES FOR AUDIENCE MEASUREMENT IN OUTDOOR ADVERTISING USING GPS DATA. IN GI JAHRESTAGUNG, VOLUME 154 OF LNI, PAGES 3993-4006. GI, 2009B

PASQUIER, M. HOFMANN, U. MENDE, F.H. MAY, M. HECKER, D. AND KÖRNER, C.: MODELLING AND PROSPECTS OF THE AUDIENCE MEASUREMENT FOR OUTDOOR ADVERTISING BASED ON DATA COLLECTION USING GPS-DEVICES. PROC. OF THE 8TH INTERNATIONAL CONFERENCE ON SURVEY METHODS IN TRANSPORT. 2008

WEGENER, D. HECKER, D. KÖRNER, C. MAY, M. UND MOCK, M.: PARALLELIZATION OF R-PROGRAMS WITH GRIDR IN A GPS TRAJECTORY MINING APPLICATION. IN PROC. OF THE 1ST UBIQUITOUS KNOWLEDGE DISCOVERY WORKSHOP (UKD'08), 2008

## ERKLÄRUNG

**Erklärung gem. § 4, Abs. (1) Nr. 9 der Promotionsordnung**

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Georg Bareth und Prof. Dr. apl. Klaus Zehner betreut worden.

Köln, den 25. Februar 2013 Dirk Hecker