Möglichkeiten der algorithmischen Auswertung von Standortdaten

Facharbeit Informatik

Eingereicht am 7. Dez. 2011 von

Timo Grossenbacher

aus Solothurn, Schweiz

Matrikelnummer 07-707-821

Betreuer:
Prof. Dr. Lorenz Hilty

Informatik und Nachhaltigkeit
Institut für Informatik
Universität Zürich

Zusammenfassung

Aktuell ist eine starke Zunahme der Verbreitung von Smartphones und anderen Geräten, mit denen Standortdaten erfasst werden können, zu verzeichnen. Gleichermassen stark steigt das Angebot an Diensten und Anwendungen, die Standortdaten verwenden, um dem/der BenutzerIn ortsbasierte Informationen anbieten zu können. Im Zusammenhang mit der Inanspruchnahme eines Dienstes, sei es zur Selbstortung oder anderen Zwecken, werden diese Daten häufig bei einem Anbieter zwischen- oder dauerhaft gespeichert. Somit ergibt sich die Ausgangslage, dass immer mehr Anbieter über eine stetig wachsende Menge von möglicherweise identifizierbaren Standortdaten verfügen können.

In der Literatur wird die Verwendung von (vermeintlich) anonymen und pseudonymen Standortrohdaten zur Gewinnung von Informationen über das raumzeitliche Verhalten eines/einer BenutzerIn lebhaft diskutiert, und es werden verschiedene Ansätze zur automatischen Auswertung dieser Daten präsentiert. Diese Ansätze unterscheiden sich in der Form der Daten, die sie benötigen, in der Art und Weise, wie die Daten ausgewertet werden, und in den Informationen, die sie gewinnen können. Es zeigt sich zum Beispiel, dass sowohl simple Heuristiken als auch komplexe Methoden des maschinellen Lernens verwendet werden können, um aus raumzeitlichen Pfaden signifikante Orte zu gewinnen, Routen und Ziele vorauszusagen und Transportmodi zu erkennen.

Diese Facharbeit hat dementsprechend zum Ziel, den aktuellen Forschungsstand der algorithmischen Auswertung von Standortdaten aufzuzeigen, indem sie eine Klassifikation nach Art der zu gewinnenden Information und nach Art der verwendeten Methoden vornimmt. Verschiedene Ansätze werden einander gegenübergestellt und ihre Vor- und Nachteile diskutiert. Abschliessend wird eine Beurteilung der Möglichkeiten der Auswertung präsentiert und auf deren Gefahren in Bezug auf den Datenschutz hingewiesen.

Abstract

Currently, one can observe a steady rise in the dissemination of smartphones and other devices which are suited to record location data. At the same time, services and applications that offer location based information are skyrocketing. In the context of such services and applications, location data is often saved and processed on the providers' servers. Thus, more and more providers have potentially identifiable location data at their command.
In the literature, the usage of (allegedly) anonymous and pseudonymous raw location data to infer information about the spatiotemporal movement patterns of a user is actively discussed. Thus, various approaches to the automatic analysis of this kind of data exist. These approaches differ in the form of the data they require, the form of the methods applied and the information that can be gained by using them. For example, simple heuristics as well as complex techniques from machine learning can be used to gain significant places, to predict routes and destinations, and to infer information about transport modes from simple spatiotemporal trajectories.
By classifying possible approaches in regard to the inference that can be made and methods used, this work has the goal to present the current state of research concerning algorithmic analysis of location data. Various approaches are compared to each other, and advantages as well as shortcomings are discussed. Eventually, the approaches are synthesized and assessed, and possible dangers to location privacy are pointed out.

Inhaltsverzeichnis

1 Einführung **2**
1.1 Standortdaten und ihr Personenbezug 2
1.2 Auswertung von Standortdaten 3
1.3 Gegenmassnahmen 4
1.4 Inhalt dieser Arbeit 4

2 Algorithmische Auswertung **5**
2.1 Extraktion von signifikanten Orten 5
2.1.1 Mittels Heuristiken und Clustering 5
2.1.2 Mittels *Fingerprinting* 7
2.1.3 Mittels Einbezug von Kontextinformationen 8
2.2 Vorhersage von Routen und Zielen 10
2.2.1 Mittels Heuristiken 10
2.2.2 Mittels maschinellem Lernen 14
2.3 Extraktion von Transportmodi 16
2.4 Reidentifikation 17

3 Beurteilung **19**

Abbildungsverzeichnis **21**

Literaturverzeichnis **22**

Kapitel 1

Einführung

Als die Firma Apple im Jahr 2007 das erste Modell des iPhones vorstellte, trat sie damit eine Welle von relativ grundlegenden Entwicklungen los. Das sogenannte Smartphone, ein eigentlicher Computer, den der Benutzer ständig mit sich führt und der über zahlreiche Sensoren sowie Kommunikationsschnittstellen verfügt, wurde massentauglich. Bei der Benutzung dieser Geräte werden ständig in irgendeiner Form Daten akquiriert und gespeichert. Dank der grossen Vielfalt dieser Daten und der direkten Kopplung zum Datenbezüger und Datenerzeuger, dem/der BenutzerIn, eröffnen sich neue Möglichkeiten der Analyse und Informationsgewinnung (Raento et al., 2009).

1.1 Standortdaten und ihr Personenbezug

Dies ist auch der Fall bei Standortdaten, die eine Unterkategorie von räumlichen Daten, den Geodaten, darstellen. Die Möglichkeiten zur Erhebung jener sind vielfältig: Während GPS das bekannteste und zuweilen das genaueste Mittel darstellt, können Standortdaten auch via Mobilfunkzellen, WLAN-Netze und andere Systeme erhoben werden. Die Erhebung von Standortdaten dient mehreren Zwecken. Häufig dienen sie der Navigation (Selbstortung), immer mehr aber auch der Verknüpfung von Sachdaten mit einem räumlichen Kontext. Beispiele dafür sind Location Based Services (LBS), Location Sharing Systems (LSS) und andere Anwendungen wie zum Beispiel Verkehrsmonitoring. LBS bieten dem Benutzer ortsabhängige Informationen, während LSS zum Zweck haben, den Standort des Benutzers/der Benutzerin anderen BenutzerInnen mitzuteilen. Gerade auf sozialen Netzwerken wird diese Möglichkeit in letzter Zeit vermehrt eingesetzt (Beispiele dafür sind Facebook[1], Foursquare[2] und der Dienst Google Latitude[3]).

[1] https://www.facebook.com
[2] https://www.foursquare.com
[3] https://www.google.com/latitude

Selbstverständlich können nicht nur mit Smartphones Standortdaten erhoben und verwendet werden. Diese stellen jedoch eine Besonderheit dar, da sie über einen ständigen und inhärenten Bezug zum/zur EigentümerIn des Smartphones verfügen[4]. Gerade dieser Personenbezug bedingt eine genauere Betrachtung aus Sicht des Datenschutzes. Personenbezogene Geodaten werden in vielen — zumindest in europäischen — Gesetzgebungen als besonders schützenswert eingestuft (vgl. z.B. Weichert (2007) und Forgo & Krügel (2010)). Im Zusammenhang mit der Inanspruchnahme eines Dienstes, sei es zur Selbstortung oder anderen Zwecken, werden diese Daten jedoch häufig bei einem Anbieter zwischen- oder dauerhaft gespeichert. Somit ergibt sich die Ausgangslage, dass immer mehr Anbieter über eine stetig wachsende Menge von möglicherweise identifizierbaren Standortdaten verfügen können.

1.2 Auswertung von Standortdaten

Wie oben bereits angetönt, eröffnen sich damit unzählige Möglichkeiten der Auswertung. Zum Beispiel können identifizierbare Standortdaten der Erstellung von Bewegungsprofilen dienen. Andererseits können auch pseudonyme[5] oder vermeintlich anonyme Standortdaten bestimmten Bewegungsprofilen zugeordnet werden. So ist es möglich, Heimadressen und andere Points of Interest (POIs) aus rohen Standortdaten zu extrahieren. Dies kann unter anderem unter ausschliesslicher Verwendung vollautomatischer Techniken geschehen. Vollautomatische Techniken basieren ausschliesslich auf computergestützten Algorithmen, während die semiautomatische Auswertung mit Hilfe visueller Analysetechniken durch einen/eine menschliche/n AnalystIn erfolgt. Dabei werden häufig Algorithmen zur Vorprozessierung der grossen Datenmenge eingesetzt. Während der erste Ansatz im Bereich der Informatik ausgiebig behandelt wurde, stammt der zweite vor allem aus dem Gebiet der geografischen Informationsvisualisierung und der Geovisualisierung. Dieses relativ junge Forschungsfeld, das sich aus einigen wenigen AutorInnen zusammensetzt, argumentiert unter anderem, dass anonymisierten Rohdaten die nötige Semantik fehlt, um ausschliesslich mit Methoden des Data Minings analysiert zu werden (Andrienko et al., 2007: 38). Andererseits ist der Mensch nicht dazu fähig, grosse Mengen an Daten zu perzipieren und zu ordnen — dabei sollen ihn wiederum Algorithmen unterstützen.

[4] vgl. dazu auch den Begriff „Pervasive Computing“ (z.B. Satyanarayanan (2001) oder Hilty et al. (2004)).

[5] Pseudonyme Daten sind solche, die einem eindeutigen Pseudonym, z.B. einer ID oder einem Fantasienamen, zugeordnet werden können.

1.3 Gegenmassnahmen

Die Informationsgewinnung aus vermeintlich anonymen oder anonymisierten Datensätzen ist aus der Sicht des Schutzes der Privatsphäre äusserst heikel, insbesondere wenn durch die gewonnenen Informationen eine sogenannte Reidentifikation von Personen möglich ist. So sind in der Literatur denn auch nicht nur Methoden zur Informationsgewinnung, sondern auch zur Informationsverschleierung und Anonymisierung zu finden (z.B. Kalnis et al. (2007), Hoh et al. (2007)). Das Feld der Informationsvisualisierung hat dieses Problem ebenfalls erkannt und arbeitet an Lösungsansätzen (Andrienko & Andrienko, 2011).

1.4 Inhalt dieser Arbeit

Basierend auf den vorgestellten technologischen und gesellschaftlichen Entwicklungen soll in dieser Arbeit vorderhand eine Übersicht und Klassifikation der in der Forschung diskutierten Möglichkeiten zur algorithmischen Auswertung von Standortdaten erstellt werden. Schwerpunkt ist nicht zwingend, wie diese Methoden im Detail aussehen, sondern vielmehr, welche Auswertungen damit möglich sind. Trotzdem soll bewusst und in vergleichender Weise auf verschiedene Techniken zur Gewinnung einer bestimmten Information eingegangen werden. Die Arbeit verzichtet auf eine Behandlung von semiautomatischen und visuellen Techniken, wie unter Abschnitt 1.2 angedeutet. Diese könnten in einer nachfolgenden Studie tiefer betrachtet werden. Sie geht auch nicht auf spezifische Gegenmassnahmen, wie unter Abschnitt 1.3 vorgestellt, ein.
Diese Arbeit hat schliesslich den übergeordneten Zweck, den aktuellen Forschungsstand im Bereich der Auswertung von Standortdaten zusammenzufassen und somit auch auf mögliche Gefahren der Akquirierung jener Daten hinzuweisen.
Aus Gründen der einfacheren Lesbarkeit wird im Folgenden die männliche Form stellvertretend für beide Geschlechter verwendet.

Kapitel 2

Algorithmische Auswertung von Standortdaten

Die folgenden Ausführungen geben einen Überblick über die Möglichkeiten der algorithmischen Auswertung von Standortrohdaten. Anhand der aufgezählten Möglichkeiten werden verschiedene Ansätze, zum Beispiel Heuristiken oder Techniken des maschinellen Lernens, vorgestellt.

2.1 Extraktion von signifikanten Orten

Mit dem Aufkommen des öffentlich zugänglichen GPS-Systems und der zunehmenden Verbreitung von Mobiltelefonen zu Beginn des neuen Jahrtausends wurden auch die ersten Studien über die Auswertung von Standortdaten veröffentlicht. Einer der ersten Prozessierungsschritte beim Auswerten von Standortdaten ist die Extraktion von signifikanten Orten, das heisst Orten, an denen sich der Benutzer wiederkehrend und über längere Zeit aufhält. So kümmern sich denn auch die frühesten Arbeiten um dieses Problem.

2.1.1 Mittels Heuristiken und Clustering

Marmasse & Schmandt (2000) entwickeln für ein LBS-Tool eine simple Heuristik für das Vorschlagen von häufig besuchten Orten. Diese nützt die Empfangsschwäche von GPS-Signalen in Gebäuden aus. Verschwindet ein Signal innerhalb eines vordefinierten räumlichen Radius' mehr als drei mal für eine vordefinierte Dauer, schlägt das System dem Benutzer diesen Ort als „häufig besuchten" vor. Die Nachteile dieser Methode liegen auf der Hand: Einerseits können GPS-Signale teilweise auch in Gebäuden empfangen werden, andererseits dauert es je nach dem unterschiedlich lange, bis nach dem Verlassen des Gebäudes wieder ein GPS-Signal empfangen wird. Zudem müssen die

Parameter für die Erkennung von Orten an lokale Gegebenheiten angepasst werden. In Häuserschluchten reagiert die Heuristik zum Beispiel anders als in offenen Landschaften mit vereinzelten Gebäuden. Alle diese Faktoren können zu erheblichen Ungenauheiten führen.
Ashbrook & Starner (2003) verwenden den gleichen Ansatz, wobei sie anstatt eines fest definierten räumlichen Radius' einen *k-means* Clusteringalgorithmus[1] anwenden, um mehrere häufig besucht" Orte, sogenannte *places*, zu *locations* zusammenzufassen. Dies hat den Vorteil, dass Sampling-Ungenauigkeiten des GPS-Signals ausgemerzt werden. Dennoch muss auch hier für den Clusteringalgorithmus ein fest definierter Parameter *Radius* gewählt werden, der je nach beabsichtigtem Zweck völlig unterschiedliche Resultate liefern kann. Als zweiter Indikator für signifikante Orte verwenden die Autoren die Heuristik, dass ein Benutzer, der eine Geschwindigkeit von unter einer Meile pro Stunde hat, als stationär gilt. Somit können auch Gegenden, in denen nur ein schwaches Signal vorhanden ist, sinnvoll klassifiziert werden.
Ebenfalls eine Heuristik verwenden Hariharan & Toyama (2004). Sogenannte *stays* sind GPS-Signale, die innerhalb eines bestimmten Radius' und eines bestimmten Zeitraums aufgezeichnet wurden. Die im Voraus definierbaren Parameter *Radius* und *Zeitraum* können dabei unterschiedliche raumzeitliche Skalen berücksichtigen. *Stays* werden in einem zweiten Schritt zu *destinations* zusammengefasst, wobei agglomeratives Clustering angewendet wird. *Destinations* können dann mit (digitalen) Karten abgeglichen werden, um Ortsinformationen zu erhalten. Dieser Algorithmus verlässt sich somit nicht auf das Fehlen von GPS-Signalen infolge Aufenthalten in Gebäuden, sondern kann auch Aufenthalte im Freien berücksichtigen. Ein weiterer Vorteil dieser Methode gegenüber derer von Marmasse & Schmandt (2000) und derer von Ashbrook & Starner (2003) ist die einbezogenene Berücksichtigung von unterschiedlichen Skalen. So kann der Algorithmus sowohl für die Extraktion von beliebten Reisezielen als auch für die von häufigen Einkaufsorten verwendet werden. Dabei muss die Skala jedoch vor der Auswertung festgelegt werden und kann nicht automatisch aus den Daten „gelernt" werden.
Die Arbeit von Kang et al. (2004) unterscheidet sich dadurch von den anderen, als dass erstens nicht zwingend GPS verwendet werden muss, sondern ein *Stream* von Koordinaten aus einem beliebigen Ortungssystem ausreicht, und zweitens signifikante Orte in Echtzeit durch ein zeitbasiertes Clusteringverfahren erhoben werden. Damit ist gemeint, dass effektiv ein Clustering von Raum-Zeit-Punkten stattfindet (siehe Abbildung 2.1 für eine anschauliche Erklärung). Der Vorteil dieser Methode ist, dass sie rechnerisch nicht sehr intensiv ist und in Echtzeit angewendet werden kann. Des weiteren werden irrelevante Punkte verworfen, die bei purem räumlichen Clustering

[1] Für eine Übersicht über gängige Clusteringverfahren siehe z.B. Xu & Wunsch (2005).

zwangsweise einem Cluster hinzugefügt würden. Trotzdem können auch hier keine Skaleneffekte aus den Daten gelesen und somit miteinbezogen werden. Die Autoren geben jedoch an, dies in nächsten Arbeiten berücksichtigen zu wollen (Kang et al., 2004: 67).

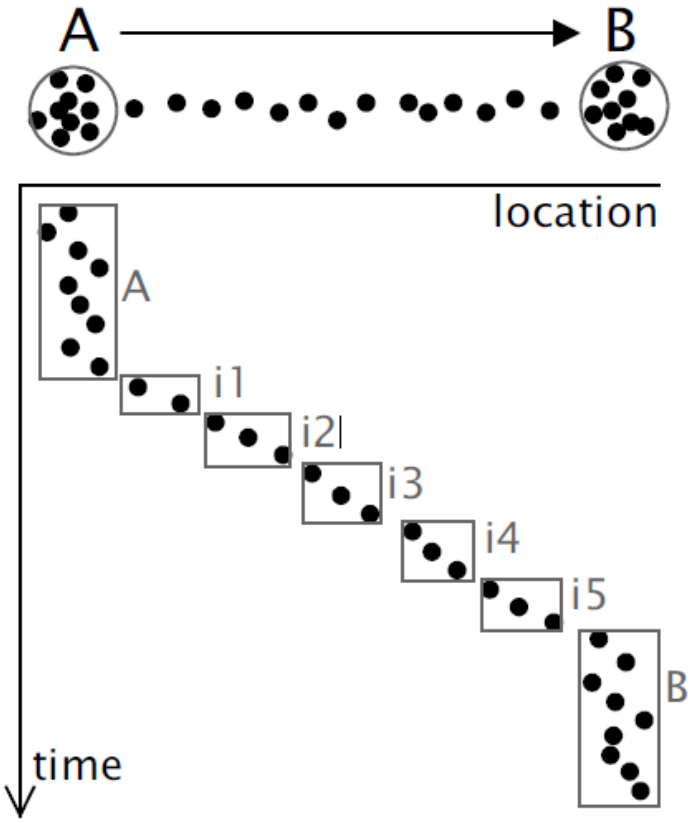


Abbildung 2.1: Eine Person bewegt sich von A nach B. Auf dem Weg dahin werden die Cluster i1–i5 und schlussendlich B erhoben. Da Cluster/Punkt B länger als eine bestimmte Zeitdauer besteht, wird er als signifikanter Ort bestimmt, während die Cluster i1–i5 verworfen werden (aus Kang et al. (2004: 61))

2.1.2 Mittels *Fingerprinting*

Hightower et al. (2005) schliesslich fassen die Arbeiten von Marmasse & Schmandt (2000), Ashbrook & Starner (2003) und Kang et al. (2004) zusammen und präsentieren einen neuen Algorithmus, der anders konzipiert ist als die oben vorgestellten Ansätze. Während diese stets einen geometrischen, das heisst geographischen, Bezug haben, gehört jener zu der Kategorie

Fingerprinting. *Fingerprinting* heisst, dass die Signalstärke von Sendestationen mit unbekannter räumlicher Position (z.B. WiFi und GSM) verwendet wird, um elektromagnetische Signaturen von bestimmten Orten zu erhalten. Verändert sich diese Signatur über eine vorher definierte Zeitspanne nicht signifikant, kann davon ausgegangen werden, dass sich der Benutzer stationär an einem Ort aufhält. Das Wiedererkennen von bereits besuchten Orten wird über statistisches Vergleichen von solchen *Fingerprints* erreicht. Um die Vorteile ihres Algorithmus' unter Beweis zu stellen, evaluieren Hightower et al. (2005) ihn und die vorgestellten Arbeiten. Dafür sammeln sie gleichzeitig GPS-, WiFi- und GSM-Daten und verwenden diese dann zum Lernen beziehungsweise zum Erkennen von Orten. Die Resultate zeigen, dass ihr Ansatz signifikant besser abschneidet als die bisherigen. Die Algorithmen von Marmasse & Schmandt (2000) und von Ashbrook & Starner (2003) leiden vor allem unter der schlechten Empfangsqualität von GPS in engen Häuserschluchten, zudem wird das Warten auf dreimaliges Aussetzen des GPS-Signals bei Marmasse & Schmandt (2000) als zu restriktiv erachtet. Zuverlässige Resultate erreichen die geometrischen Algorithmen erst, wenn mehr als drei mal ein signifikanter Ort besucht wird oder wenn man sich dort länger als drei Stunden aufhält. In dieser Hinsicht ist der Algorithmus von Hightower et al. (2005) vielfältig einsetzbar: Er erkennt signifikante Orte unabhängig von der Besuchsfrequenz und der Aufenthaltsdauer, obwohl hier natürlich auch eine minimale Aufenthaltsdauer im Vornherein festgelegt werden muss.
Ein weiteres Beispiel für die Verwendung von GSM-Daten und *Fingerprinting* präsentieren Laasonen et al. (2004) mit einem Set aus verschiedenen Algorithmen. Dabei basieren diese ausschliesslich auf Übergängen zwischen verschiedenen GSM-Zellen, die von Mobiltelefonen registriert werden. Mittels Verwendung der Graphentheorie gelingt es ihnen demnach, geographische Ortschaften zu erkennen, obwohl dass Zellinformationen von GSM-Netzen keine expliziten topologischen oder geographischen Informationen beinhalten. Das Problem bei der ausschliesslichen Verwendung von GSM-Zellen ist jedoch ihre unterschiedliche Grösse. Je nach Signalstärke und physikalischer Umwelt können diese zwischen wenigen hundert Metern und einigen Kilometern gross sein. Dementsprechend unscharf sind die Ergebnisse der Algorithmen von Laasonen et al. (2004), vor allem in ländlichen Gebieten.

2.1.3 Mittels Einbezug von Kontextinformationen

Wie oben bereits erwähnt wurde, wird das Ergebnis von vielen Algorithmen durch vordefinierte Parameter eingeschränkt. So wird zum Beispiel in fast allen bisher vorgestellten Methoden eine bestimmte minimale Zeitdauer vorausgesetzt, die der Benutzer an einem Ort verbringen muss, damit dieser als signifikant eingestuft wird. Ist diese Schwelle zu klein, können

unwichtige Ereignisse wie Stops an Verkehrsampeln fälschlicherweise als signifikant klassifiziert werden (falsch positive Orte). Ist sie zu gross, werden wichtige Orte, die jedoch nur einen kurzen Aufenthalt bedingen, vernachlässigt (falsch negative Orte). Liao et al. (2007a) gehen diese Schwäche an, indem sie ein probabilistisches Modell[2] von *Aktivitäten* und *signifikanten Orten* entwickeln. *Aktivitäten* basieren auf GPS-Signalen und werden mittels folgenden Kontextinformationen gewonnen: Zeitliche Informationen wie Tageszeit, Wochentag und Aufenthaltsdauer, Fortbewegungsgeschwindigkeit und Informationen, die aus geographischen Datenbanken extrahiert werden. Diese Informationen können zum Beispiel Busrouten oder Punktdaten wie Restaurants, Läden, etc., sein. Zusätzlich werden einzelne *Aktivitäten* mit

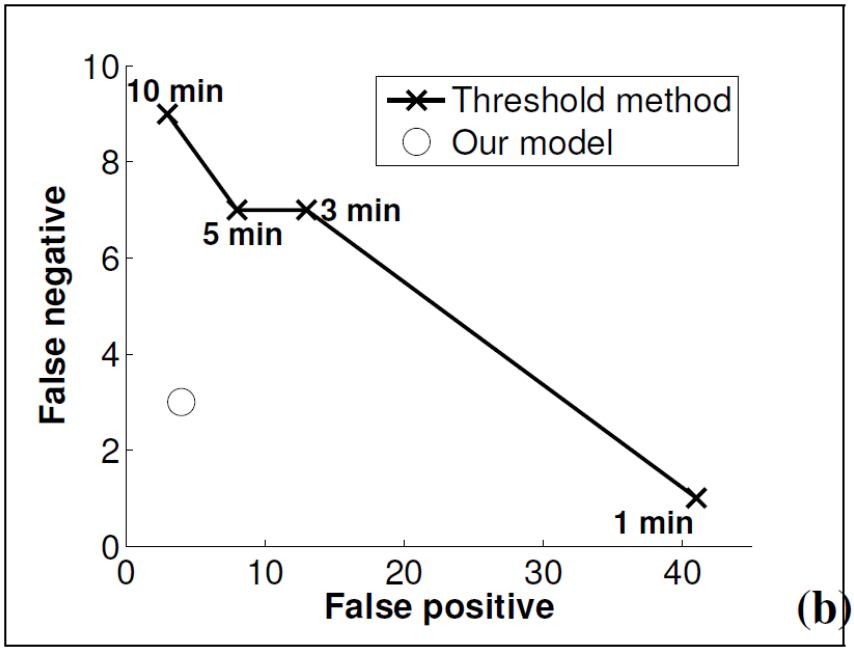


Abbildung 2.2: Eine zu hohe oder zu kleine Schwelle der Aufenthaltsdauer an einem Ort produziert falsch negative beziehungsweise falsch positive Orte. Die Methode von Liao et al. (2007a), gekennzeichnet durch einen Punkt, ist von dieser Einschränkung nicht betroffen (aus Liao et al. (2007a: 16))

anderen, nachbarschaftlichen *Aktivitäten* verglichen, um ihre Plausibilität zu überprüfen. So kann es zum Beispiel nicht sein, dass jemand in einer *Aktivität* auf den Bus steigt und in der nächsten einen Zug verlässt. *Aktivitäten* werden dann in einem iterativen Verfahren zu *signifikanten Orten* zusammengefasst, wobei schliesslich auch eine *Aktivität* an diesem Ort vorliegt (z.B. Schlafen, Transfer von Bus zu Zug, etc.). Um die Wirksamkeit

[2]Die meisten Ansätze können unter dem Überbegriff „Maschinelles Lernen“ zusammengefasst werden (siehe Punkt 2.2.2 für eine kurze Einführung).

ihres komplexen Ansatzes unter Beweis zu stellen, vergleichen die Autoren ihn mit solchen Algorithmen, die auf einer fixen zeitlichen Aufenthaltsdauer beruhen. Sie kommen zum Schluss, dass aus 51 signifikanten Orten nur 4 falsch positiv und 3 falsch negativ klassifiziert wurden, während der andere Ansatz deutlich schlechter abschneidet (Abbildung 2.2).

2.2 Vorhersage von Routen und Zielen

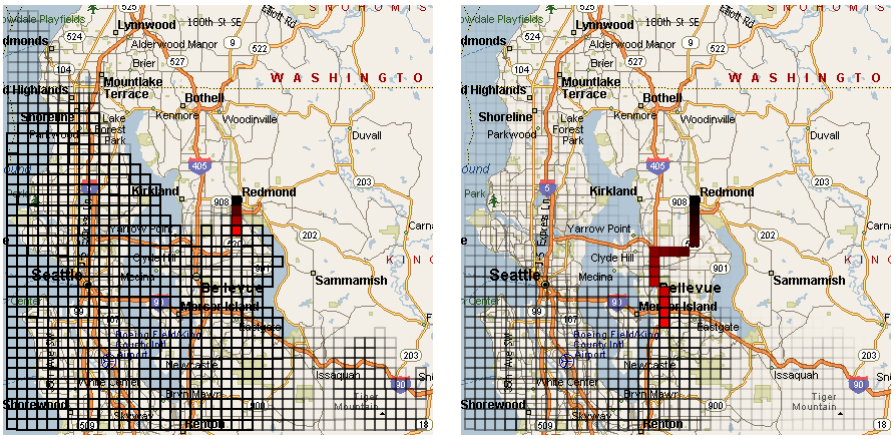
2.2.1 Mittels Heuristiken

Krumm (2006) präsentiert eine einfache Methode, die es ermöglicht, Ziele einer Autoreise in Echtzeit vorherzusagen. Sie basiert auf der simplen Heuristik, dass Autofahrer grundsätzlich bestrebt sind, effizient zu fahren. Formal bedeutet dies, dass die restliche Zeit, die benötigt wird, um ein Ziel zu erreichen, mit dem Fortschreiten einer Reise monoton sinken sollte. Um dies zu überprüfen, teilt der Autor ein Testgebiet von $41km^2$ in ein gleichmässiges Raster aus Zellen von $1km$ Seitenlänge ein. Zuerst wird ein konventioneller Routenplaner genommen, um die geschätzte Fahrzeit von jeder Zelle zu allen restlichen Zellen zu berechnen. Danach werden über 4000 gesammelte GPS-Trips in diesem Testgebiet auf das Raster diskretisiert, das heisst, es entstehen Sequenzen aus Zellen. Nun wird jede Sequenz auf ihre Effizienz analysiert. Es wird für jede neue Zelle in der Sequenz überprüft, ob die geschätzte Zeit zum Ziel (basierend auf den Daten aus dem Routenplaner) kleiner oder grösser ist als die bisher minimale Zeit zum Ziel. Ergibt sich in jeder neuen Zelle eine neue minimale Zeit zum Ziel, ist die Route zu 100% effizient. Durch die Auswertung der gesammelten GPS-Trips stellt der Autor eine durschnittliche Effizienz e von 0.625, also 62.5%, fest. Ein einfaches bayessches Modell

$$p(c_i|S) = \frac{p(S|c_i)p(c_i)}{\epsilon}$$

das die Wahrscheinlichkeit $p(S|c_i)$ einer Sequenz S, gegeben eine Zielzelle c_i, und die *a priori* Wahrscheinlichkeit $p(c_i)$ dieser Zielzelle, multipliziert, gibt dann Auskunft über die Wahrscheinlichkeit $p(c_i|S)$ dieser Zielzelle, gegeben eine Sequenz (die gewünschte Information). ϵ ist ein Normalisierungsfaktor. In diesem Fall wird jeder Zielzelle die gleiche *a priori* Wahrscheinlichkeit $p(c_i) = 1/N_c$ zugeordnet — Verbesserungen des Modells setzen in erster Linie an dieser Schätzung an (siehe unten). $p(S|c_i)$ wird folgendermassen berechnet: Multipliziere $p(S|c_i)$ für jede neue Zelle in einer bisherig befahrenen Sequenz mit der durchschnittlichen Effizienz e, wenn diese Zelle (zeitlich) näher beim (zu schätzenden) Ziel ist als jede vorherhige Zelle. Ist dies nicht der Fall, multipliziere mit der Gegenwahrscheinlichkeit $e - 1$. Vereinfacht gesagt erhöht sich die Wahrscheinlichkeit einer (zu schätzenden) Zielzelle, je effizienter eine Sequenz, die dahin führt, durchfahren wird (solange $e > 0.5$).

Dieser Algorithmus hat intuitiv zur Folge, dass, je mehr Zellen bereits durchfahren worden sind, desto genauer die Vorhersage der Zielzelle wird. Dies bestätigt Abbildung 2.3.

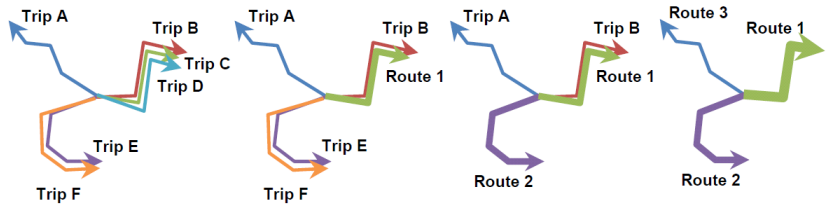


(a) Zu Beginn des Trips (b) Nach einiger Zeit

Abbildung 2.3: Die Wahrscheinlichkeit einer Zielzelle wird durch ihre Helligkeit angezeigt (je dunkler, desto wahrscheinlicher). In (b) reduzieren sich im Vergleich zu (a) die möglichen Zielzellen signifikant.

In einer Evaluation seiner Methode vergleicht (Krumm, 2006: 6) den Anteil der bereits durchfahreren Sequenz mit der Vorhersagegenauigkeit der Zielzelle. Für einen Anteil von 25% erhält die Schätzung einen Genauigkeit von $21km$, für 75% erhöht sich diese Genauigkeit auf $3km$. Diese Resultate zeigen, dass die vorgeschlagenene Heuristik höchstens für eine Schätzung ausreicht, wenn der Trip schon beinahe beendet ist. Ein möglicher Schwachpunkt ist die grobe Aufteilung auf ein $1km$-Raster. Eine feinere Aufteilung hätte jedoch höhere Rechenzeiten bei der Verwendung des Routenplaners zur Folge, da eine grössere Anzahl von Zellkombinationen berechnet werden müssten. Andererseits kann man der Methode anrechnen, dass sie ohne komplexe probabilistische Modelle auskommt und keine Kontextinformationen benötigt.
In einer sehr aktuellen Publikation zeigt Krumm (2011: 233-234), dass die *a priori* Wahrscheinlichkeit $p(c_i)$ einer Zielzelle besser geschätzt werden kann, indem Kontextinformationen berücksichtigt werden. Als Beispiele dafür nennt er erstens Statistiken über durchschnittliche Fahrzeiten eines Trips, die dazu benutzt werden können, weit entfernte oder sehr nahe Zellen mit einer kleinen Wahrscheinlichkeit zu behaften. Auch die Landnutzung und Form der Landschaft kann dazu genutzt werden, um zum Beispiel Zellen, die im Wasser liegen, auszuschliessen, und solche, die in Wohngebieten liegen, zu favorisieren. Der Autor nennt auch die Möglichkeit, oft besuchte Zellen zu registrieren und mit einer höheren Wahrscheinlichkeit zu belegen,

oder Zellen zu favorisieren, die über POIs wie Restaurants und dergleichen verfügen. Dies wurde in Krumm & Horvitz (2006) umgesetzt, allerdings unter Einbezug von Techniken des maschinellen Lernens (siehe Punkt 2.2.2).
Eine Sammlung von Algorithmen zur Vorhersage von Routen mit dem Auto entwickeln Froehlich & Krumm (2008), wobei ihre Methoden zur Vorhersage der gesamten Route und nicht nur der nächsten Segmente gedacht sind. Ihre Methoden basieren auf der Annahme, dass Autofahrer eine Menge von Routen haben, die sie regelmässig und wiederkehrend befahren. Die Vorgehensweise besteht im Wesentlichen aus drei Schritten: 1. Extraktion von *Trips* mit Start- und Endpunkten aus GPS-Daten, 2. Zusammenfassung und Abstraktion von *Trips* zu *Routen*, 3. Echtzeit-Vorhersage von *Trips* basierend auf bestehenden *Routen*. Diese Schritte und die dafür verwendeten Algorithmen sollen hier kurz angerissen werden.
Für die Segmentation von *Trips* aus GPS-Daten verwenden die Autoren einen Schwellenwert von einigen Minuten, in denen ein Fahrzeug stillsteht. Danach werden Ausreisser in den Daten durch eine Überprüfung der Geschwindigkeit und Beschleunigung erkannt und entfernt. Sind diese Parameter aussergewöhnlich hoch, handelt es sich mit grosser Sicherheit um einen unrealistischen Ausreisser. Zuletzt werden verschiedene Filter angewandt, um nicht-valide *Trips* zu entfernen, zum Beispiel solche, die weniger als eine Mindestmenge von Datenpunkten beinhalten, oder solche, die über ungewöhnlich viele Richtungswechsel verfügen. Solche entstehen meist bei Stillstand des Fahrzeuges wegen der Varianz des GPS-Signals.
In einem zweiten Schritt werden ähnliche *Trips* zu *Routen* zusammengefasst, wobei hierarchisches Clustering angewandt wird. Dafür muss jedoch zuerst eine Ähnlichkeitsmatrix erhoben werden, wo die gegenseitige Ähnlichkeit jeder *Trip*-Kombination erfasst wird (siehe Abbildung 2.4). Für die Festsetzung von Ähnlichkeit wird die sogenannte *Hausdorff-Metrik* angewandt, die basierend auf den Distanzen zwischen einzelnen Datenpunkten eines *Trips* und Segmenten des anderen *Trips* einen Wert für die räumliche Nähe zweier Linien angibt. Dabei wird die Richtung berücksichtigt, das heisst, dass räumlich gleiche *Trips* in entgegengesetzter Richtung nicht als ähnlich festgelegt werden. Aus den Clustern von *Trips* müssen nun noch *Routen* mit einer eigenen räumlichen Ausdehnung extrahiert werden. Dafür verwenden die Autoren einen Algorithmus, der für jeden *Trip* eines Clusters in einem festen räumlichen Abstand neue Datenpunkte erstellt und dann den räumlichen Durchschnitt dieser Punkte, gegeben eine Distanz, bildet. Das Resultat ist eine *Route*, deren Punkte aus den Durchschnitten aller *Trips* im Cluster gebildet werden.
In einem letzten Schritt müssen dann *Trips* in Echtzeit mit *Routen* abgeglichen und so vorhergesagt werden. Dabei wird wiederum die Ähnlichkeits-Metrik aus dem letzten Paragraphen verwendet. Als Resultat liefert der Matching-Algorithmus eine nach Wahrscheinlichkeit geordnete Liste von *Routen*, zu denen der aktuelle *Trip* gehören könnte. Hier soll noch einmal



	A	B	C	D	E	F
A	0	5	8	9	14	16
B	5	0	3	4	9	11
C	8	3	0	1	6	8
D	9	4	1	0	5	7
E	14	9	6	5	0	2
F	16	11	8	7	2	0

(a) The initial trip similarity matrix for the above trips.

	A	B	R1	E	F
A	0	5	8.5	14	16
B	5	0	3.5	9	11
R1	8.5	3.5	0	5.5	7.5
E	14	9	5.5	0	2
F	16	11	7.5	2	0

(b) Trip C and Trip D are merged to form Route 1.

	A	B	R1	R2
A	0	5	8.5	15
B	5	0	3.5	10
R1	8.5	3.5	0	6.5
R2	15	10	6.5	0

(c) Trip E and Trip F are merged to form Route 2.

	R3	R1	R2
R3	0	7.3	15
R1	7.3	0	7.7
R2	15	7.7	0

(d) The final trip similarity matrix.

Abbildung 2.4: *Trips* werden, basierend auf ihrer Ähnlichkeits-Metrik, hierarchisch zu Clustern zusammengefasst. Dies geschieht solange, bis ein bestimmter Schwellenwert einer minimalen Ähnlichkeit nicht mehr erreicht werden kann (aus Froehlich & Krumm (2008: 7)

auf die Annahme zurück gekommen werden, die die Autoren zu Beginn ihres Artikels treffen, nämlich dass viele *Routen* regelmässig befahren werden. Froehlich & Krumm (2008) zeigen dies, in dem sie GPS-Aufzeichnungen von ungefähr 240 Autofahrern auswerten. 40% aller *Routen* werden demnach mehr als einmal befahren, und die 10 meist befahrenen *Routen* umfassen 50% aller *Trips*. Solche Werte können jedoch erst erreicht werden, wenn *Trips* bereits mindestens für eine Woche aufgezeichnet wurden, und im Regelfall erhöhen sie sich, je länger aufgezeichnet wird.
In einer experimentellen Evaluation zeigen die Autoren dann, dass nach einer Fahrtdistanz von ungefähr 16km eine 50%-Chance besteht, die befahrene *Route* richtig vorherzusagen. Dies ist jedoch nur der Fall, wenn die Evluation nur solche *Trips* beachtet, die mehrmals vorkommen. Werden alle, auch nur einmalig befahrene, *Trips* miteinbezogen, sinkt der Wert auf 20%. Mit andauernder Fahrt werden die Vorhersagen stetig genauer, solange man in der Nähe von bereits befahrenen *Routen* bleibt. So ist es mit den vorgestellten Methoden selbstverständlich nicht möglich, etwas vorauszusagen, das noch nie befahren wurde. Die Autoren geben an, dass man durch die Berücksichtigung von zusätzlichen Informationen wie Fahreffizienz (siehe Krumm (2006)), Tageszeiten und Wochentagen die Vorhersage verbessern könnte. Ausserdem sei es auch schon wertvoll, *nur* eine Liste von möglichen *Routen* zur Hand zu haben, denn basierend darauf könnte zum Beispiel ein Hybrid-Motor seinen Ressourcenverbrauch heuristisch steuern.

2.2.2 Mittels maschinellem Lernen

In der Literatur gibt es eine Vielzahl von Arbeiten, die Ansätze aus dem Bereich des maschinellen Lernens verwenden, um basierend auf bestehenden Standortdaten ein Ziel, eine Route oder beides vorauszusagen. Maschinelles Lernen hat häufig zum Ziel, basierend auf sogenannten Trainingsdaten, Aussagen über unbekannte Zustände eines Modells zu machen (siehe z.B. Bishop (2007)). Details zu diesen Methoden sollen hier nicht wiedergegeben werden, da sie einerseits relativ komplex sind und andererseits den Umfang dieser Arbeit übersteigen würden. Im Folgenden soll jedoch anhand eines kurzen Beispiels eine Einführung in eine mögliche Ausprägung des maschinellen Lernens gegeben werden. Danach wird eine Übersicht über die Ansätze im Bereich der Routen- und Zielvorhersage, die in der Literatur diskutiert werden, präsentiert. Diese Ansätze unterscheiden sich vor allem in den Ergebnissen, die sie liefern können, sowie in den Vorbedingungen und Annahmen, die sie treffen.

Die Arbeit von Ashbrook & Starner (2003) kann als Beispiel gesehen werden anhand dessen eine Art von maschinellem Lernen anschaulich erklärt werden kann. Die Autoren verwenden die im ersten Schritt angefertigte, chronologische geordnete Liste von *locations* (siehe Punkt 2.1.1), um ein Graphenmodell zu trainieren, dessen Knoten *locations* repräsentieren und dessen Kanten die Wahrscheinlichkeit eines Übergangs zwischen diesen *locations* bezeichnen. Diese Wahrscheinlichkeiten werden aus der Menge aller *locations* berechnet — so hat zum Beispiel die Route „von Zuhause zur Arbeit“ eine höhere Frequenz und damit grössere Wahrscheinlichkeit als die Route „vom Supermarkt zur Arbeit“ (siehe Abbildung 2.5). Ein simples Modell würde nun die Wahrscheinlichkeit der nächsten *location* („Arbeit“ in diesem Fall) basierend auf der letzten *location* („Supermarkt“ in diesem Fall) berechnen (Modell 1. Ordnung). Bezieht man jedoch die n letzten *locations* auch noch mit ein, können genauere Vorhersagen gemacht werden (Modell n. Ordnung). So bekommt die Route „vom Supermarkt zur Arbeit“ plötzlich eine höhere Wahrscheinlichkeit, wenn ihr zuerst die Route „von Zuhause zum Supermarkt“ vorangegangen ist. Zusammenfassend kann man sagen, dass das Modell in einer ersten Phase basierend auf der chronologisch geordneten Liste von *locations* trainiert wird, um Wahrscheinlichkeiten zu erhalten. In einem zweiten Schritt kann dann basierend auf den n letzten besuchten *locations* der nächste signifikante Ort mit einer bestimmten Wahrscheinlichkeit vorausgesagt werden.

Patterson et al. (2003) entwickeln ein graphenbasiertes Modell, das aus GPS-Daten gleichzeitig den Transportmodus und das wahrscheinlichste Ziel eines Benutzers lernen und voraussagen kann. Dafür werden zuerst häufig besuchte Orte gelernt, um danach Vorhersagen zu treffen. Zusätzlich demonstrieren die Autoren, dass durch das Hinzufügen von Kontextinformationen wie Standorten von Busstops und Busrouten die Genauigkeit ihrer Methode

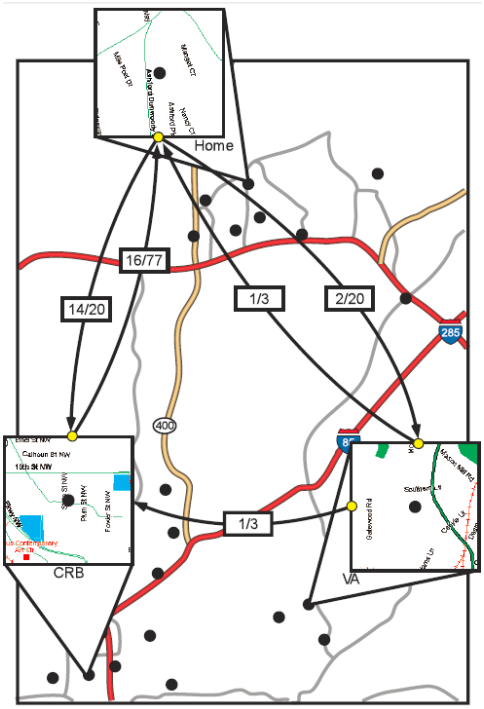


Abbildung 2.5: Die Karte zeigt Knoten und Kanten eines Graphen, der basierend auf einer Chronologie aus GPS-Daten erstellt wurde. Die Nummern an den Kanten bezeichnen die Wahrscheinlichkeit eines Übergangs von einem Ort zum nächsten (aus Ashbrook & Starner (2003: 282))

erhöht werden kann. Die Arbeit von Liao et al. (2006) geht in die gleiche Richtung. Auch hier werden verschiedene Techniken des maschinellen Lernens kombiniert, um gleichzeitig Information über Transportmodus und Ziel zu lernen und vorauszusagen.

Simmons et al. (2006) verwenden ebenfalls eine graphenbasierte Technik des maschinellen Lernens, um Routen und Ziele vorherzusagen. Dabei wird konkret Wissen über das Strassennetzwerk mit in das Modell einbezogen. Die Autoren geben eine Genauigkeit von 98% der Vorhersage des nächsten Routenabschnittes an, obwohl in den meisten Fällen das nächste Segment das einzig mögliche ist. Gleichzeit sagt ihr System ein mögliches Ziel der aktuellen Route voraus. Die Autoren evaluieren auch die Wirksamkeit des Einbezugs von Kontextinformationen wie der Tageszeit, des aktuellen Wochentags und der momentanen Geschwindigkeit des Fahrzeugs. Dabei stellt sich heraus, dass in ihrem Modell nur die Geschwindigkeit die Routenvorhersage signifikant verbessern kann.

2.3 Extraktion von Transportmodi

Die Extraktion von Transportmodi aus einem GPS-Track oder anderen Standortdaten ist eine Aufgabe, die in der Literatur weniger stark diskutiert wird als die in Abschnitt 2.2. Es finden sich vor allem Ansätze aus dem Bereich des maschinellen Lernens (siehe Punkt 2.2.2), obwohl auch einfache Heuristiken verwendet werden können. So könnte zum Beispiel eine bestimmte Mindestgeschwindigkeit verwendet werden, um die Bewegung „Laufen“ von der Bewegung „Bus“ zu unterscheiden, oder es könnten Beschleunigungen und Bremsvorgänge heuristisch analysiert werden. Dabei werden Kontextinformationen wie das vorherige und nachfolgende Transportmittel oder Verkehrszustände jedoch nicht beachtet — so würde eine Busfahrt, die in einem Stau stecken bleibt, ebenfalls als „Laufen“ klassifiziert werden.
Die Arbeit von Zheng et al. (2008) kann stellvertretend für solche, die maschinelles Lernen anwenden, genannt werden (siehe z.B. auch Patterson et al. (2003); Gogate et al. (2005); Liao et al. (2007b)). Die Autoren entwickeln eine Methode, die basierend auf puren GPS-Daten, das heisst ohne Kontextinformationen, Transportmodi erkennen kann. In einem ersten Schritt werden einfache Metriken wie Geschwindigkeit und Beschleunigung aus den GPS-Daten errechnet. Diese werden dazu benutzt, einen GPS-Trip in bestimmte *Segmente* aufzuteilen. Dabei wird eine Heuristik angewendet, die auf bestimmten Annahmen beruht, zum Beispiel dass kurze Wechsel zwischen „Laufen“ und „Autofahren“ nicht sehr wahrscheinlich sind, und es sich wohl eher um einen Stau handelt. Das Ergebnis dieses Schrittes ist eine Segmentation in *walk segments* und *non-walk segments*, wobei GPS-Signale mit einer Geschwindigkeit und Beschleunigung nahe bei Null als Übergänge zwischen diesen Segmenten dienen. Für alle Segmente werden dann Statistiken wie die Durchschnittsgeschwindigkeit, die drei höchsten Geschwindigkeitszustände und weitere erhoben. Danach erst werden verschiedene Methoden des maschinellen Lernens angewendet, um die Segmente genauer in „Laufen“, „Radfahren“, “Autofahren“ und „Busfahren“ zu unterteilen. Vereinfacht gesagt beziehen jene Methoden einerseits Informationen aus vorangegangenen und nachfolgenden *Segmenten* sowie andererseits Informationen zu Übergangswahrscheinlichkeiten zwischen einzelnen Modi in die Erkennung ein.
Die Autoren klassifizieren bei einer experimentellen Evaluation der verschiedenen Methoden fast 70% der Länge eines Trips korrekt. Dafür verwenden sie GPS-Daten mit einer Gesamtlänge von über $20'000km$, wovon 70% als Trainingsdaten, deren Transportmodi bekannt sind, dienen. Die Arbeit von Zheng et al. (2008) zeigt somit, dass auch ein kombinierter Ansatz aus Heuristiken und probabilistischen Modellen durchaus zu brauchbaren Resultaten führen kann.

2.4 Reidentifikation

Alle bisher vorgestellten Methoden haben zum Ziel, Informationen aus Rohdaten zu gewinnen. Viele dieser Informationen können — kombiniert mit nicht-räumlichen oder nur implizit räumlichen Daten wie Adressen — dazu dienen, anonyme oder pseudonyme Rohdaten einer bestimmten Identität zuzuordnen. In der Literatur gibt es jedoch auch Arbeiten, die explizit Methoden zur Reidentifikation von Benutzern diskutieren. Meist geschieht dies im Hinblick auf die Datenschutzproblematik, und es werden gleichzeitig Gegenmassnahmen präsentiert. Im Folgenden soll ein Auszug aus dieser Forschung präsentiert werden.

Krumm (2007) entwickelt vier verschiedene Algorithmen zur Extraktion einer Heimadresse aus mehreren GPS-Tracks. Diese Algorithmen basieren auf verschieden Heuristiken, zum Beispiel dass die Heimadresse meist der letzte signifikante Ort an einem Tag ist, oder dass man an der Heimadresse die meiste Zeit verbringt. Ebenfalls verwendet wird ein hierarchisch-agglomeratives Clusteringverfahren, um Datentupel aus GPS-Koordinaten zusammenzufassen. Danach wird der Mittelpunkt (Zentroid) des Clusters mit den meisten Tupeln als Heimadresse angenommen. In einer empirischen Evaluation zeigt Krumm (2007) dann, dass der Algorithmus, der die letzte Destination an einem Tag als Heimadresse annimmt, am besten abschneidet. Die Ergebnisse dieser Methoden liegen in geographischen Koordinaten vor — diese verwendet der Autor, um sie über *reverse geocoding* mit effektiven Adressen abzugleichen. Hier liegt denn auch der Flaschenhals der Methode: Oft ist es schwer, von einem — erst noch relativ ungenauen — Tupel aus Koordinaten mit hoher Sicherheit die richtige Adresse zu eruieren. So erreicht der Autor in einer experimentellen Evaluierung eine Erfolgsquote von 13%. Nichtsdestotrotz ist in den allermeisten Fällen zumindest die Ortung auf Quartierstufe möglich.

Eine ähnliche, aber erfolgreichere Heuristik zur Entdeckung von Heimadressen entwickeln Hoh et al. (2006). Sie unterscheidet sich dadurch von jenen von Krumm (2007), als dass sie der iterativen, manuellen Festlegung von plausiblen Gebieten (Wohnzonen, etc.) bedarf und deshalb eigentlich nicht den vollautomatischen Methoden zugeordnet werden kann. Die Autoren beziffern die Erfolgsquote der Eruierung von Adressen aus Rohdaten auf 85%. Diese Zahl ist mit Vorsicht zu geniessen, da die Autoren nicht über Angaben zu „richtigen“ Adressen verfügen, vielmehr nehmen sie einfach an, diese würden stimmen.

Während die bisherigen Methoden immer nur auf zusammenhängende Datensätze von einzelnen Benutzern anwendbar sind, demonstrieren Gruteser & Hoh (2005) eine Methode, die aus anonymen Daten einer unbekannten Menge von Benutzern einzelne Benutzer extrahieren kann. Ihre Methode basiert im Wesentlichen auf der Annahme, dass sich ein Benutzer konstant auf einem Pfad bewegt und eher selten abrupt die Richtung wechselt

(sogenanntes *trajectory-based linking* von Standortdaten (Gruteser & Hoh, 2005: 181)). Dazu entwickeln sie einen probabilistischen Algorithmus, um erstens basierend auf einem bestehenden Pfad (*trajectory*) den nächsten Zustand eben dieses vorauszusagen und zweitens Koordinaten-Tupel (*samples*) diesem nächsten Zustand eindeutig zuzuordnen. So ergibt sich eine voneinander unterscheidbare Menge an Pfaden, extrahiert aus einer zuerst zufällig anmutenden Menge von Koordinaten-Tupeln mehrerer Benutzer. In einem Experiment zeigen die Autoren, dass ihr Algorithmus auch dann mit ausreichender Qualität funktioniert, wenn angenommen wird, die Benutzer hätten sich alle zur selben Zeit bewegt. Berücksichtigt man nämlich unterschiedliche Zeitstände in realen Daten, vereinfacht sich das Problem der Extrahierung von einzelnen Pfaden erheblich. In einem weiteren Schritt könnten diese Pfade nun mit den oben vorgestellten Methoden weiter analysiert werden, um so an identitätsrelevante Informationen zu gelangen.

Kapitel 3

Beurteilung

Das vorangehende Kapitel hat verschiedene Ansätze zur Auswertung von Standortdaten aufgezählt. Es wurden Methoden zur Gewinnung von signifikanten Orten, zur Voraussage von Routen und Zielen sowie zur Erkennung von Transportmodi vorgestellt. Obwohl sich die Methoden in ihrer Art teilweise stark unterscheiden, können sie alle basierend auf nur wenigen Informationen, genau gesagt basierend auf zwei Koordinaten und einer Zeitangabe, umfassende Aussagen über den Datenerzeuger machen. Gewisse Methoden benötigen nicht einmal georeferenzierte Daten, sondern machen sich die spektrale Signatur von WiFi- und GSM-Signalen oder die Zelleninfos aus GSM-Netzen zu Nutze.
In den obigen Ausführungen wurde ebenfalls gezeigt, dass die einzelnen Methoden und Algorithmen spezifische Schwächen aufweisen, die entweder in späteren Arbeiten verbessert wurden, oder die den jeweiligen Ansätzen schlichtweg inhärent sind. So hat sich gezeigt, dass man sich nicht nur auf das Verschwinden von GPS-Signalen in Gebäuden verlassen kann, wenn man einen signifikanten Ort aus einem raumzeitlichen Pfad extrahieren will. Man kann dementsprechend folgern, dass verschiedene Methoden zu verschiedenen Einsatzzwecken gebraucht werden können, aber es so etwas wie eine universelle Technik zur Informationsgewinnung nicht geben kann. Es stellt sich denn auch die Frage, ob sich bei gewissen Fragestellungen der Aufwand lohnt, ein komplexes probabilistisches Modell zu trainieren, oder ob eine simple Heuristik nicht genügend genaue Resultate liefert. Andererseits muss gesagt werden, dass es wahrscheinlich Probleme gibt, die wegen ihrer Komplexität mit Heuristiken ungenügend verlässlich behandelt werden können, und ein Griff zu Techniken des maschinellen Lernens unabdingbar ist.
Aus diesen Erkenntnissen lässt sich schliessen, dass für eine umfassende Informationsgewinnung aus Standortrohdaten eine Kombination aus verschiedenen Ansätzen die besten Resultate liefert. Meistens lassen sich für das Vorprozessieren von Daten, zum Beispiel zur Segmentation von GPS-Daten in spezifische Trips, Heuristiken nutzen. Ein probabilistisches Modell könnte

dann so trainiert werden, dass es die Unsicherheiten, die aus der Anwendung der Heuristik resultieren, mit in die Schätzung und Vorhersage von Informationen einbezieht. Des weiteren könnten zur Gewinnung einer bestimmten Information, zum Beispiel des Transportmodus' in einem bestimmten Abschnitt eines Pfades, verschiedene Ansätze angewendet werden. Ein Toolkit, das verschiedene Algorithmen zur Lösung desselben Problems anbietet, würde einen Analysten dabei unterstüzen, die Resultate dieser Algorithmen qualitativ und quantitativ zu vergleichen, um zum wahrscheinlichsten Resultat zu gelangen. Hier liegt auch der Einsatz von visuellen Techniken zur Unterstützung der Auswertung durch einen Analysten nicht fern. Möglicherweise könnte zukünftige Forschung die Schaffung solcher Toolkits begleiten und/oder evaluieren. Auf der anderen Seite sind Arbeiten, die verschiedene, bereits existierende Ansätze miteinander vergleichen und auch experimentell evaluieren, faktisch nicht existent. Hier gibt es einen Nachholbedarf.
Im letzten Abschnitt des vorangehenden Kapitel wurde auf das konkrete Problem der Reidentifikation von Personen aus anonymen oder pseudonymen Standortdaten eingegangen. Es wurde gezeigt, dass bereits verschiedene Ansätze bestehen, um zum Beispiel Heimadressen zu gewinnen oder einzelne Benutzer aus einer unbekannten Menge zu extrahieren. In der Kombination dieser Methoden mit den obigen liegt denn auch die Möglichkeit der Erstellung kompletter Bewegungs-, ja Lebensprofile, zieht man die ständige Verwendung und Mitführung von Smartphones in Betracht. Verschiedene Ansätze können auch hier zur gegenseitigen Unterstützung herbeigezogen werden. Zum Beispiel könnten zuerst aus einer beliebigen Menge von Standortdaten einzelne Benutzer extrahiert werden, um dann in einem weiteren Schritt deren Pfade mit den obigen Methoden unter die Lupe zu nehmen. Basierend auf den gewonnenen signifikanten Orten könnten dann Aussagen über verschiedene Aspekte des Lebens des jeweiligen Benutzers gemacht werden, wie zum Beispiel über seine Arbeitsstelle oder religiöse Zugehörigkeit (vgl. dazu Gasson et al. (2011)). Dass dies aus datenschutzrechtlicher Sicht höchst brisant ist, ist selbstredend. In diesem Sinne soll diese Arbeit auch darlegen, dass die Auswertung von Standortdaten in vielen Fällen Informationen, die einen hohen Personenbezug aufweisen, zu Tage bringt. Dementsprechend sollten die vorgestellten Methoden auch hinterfragt werden. Die Tendenz in der Literatur ist eher die, dass auf Fragen der Privatsphäre und des Datenschutzes meistens nicht eingegangen wird. Es existieren zwar Artikel und Studien, die genau dieses Problem ansprechen, jedoch sind diese vergleichsweise rar. In diesem Gebiet müsste also noch mehr Forschung betrieben werden, besonders in Hinblick auf den stetig wachsenden Einbezug von Standortdiensten in sozialen Netzwerken und der allgemein zunehmenden Verbreitung von persönlichen Daten im Internet.

Abbildungsverzeichnis

2.1 Zeitbasiertes Clustering 7
2.2 Vergleich von verschiedenen Ansätzen zur Erkennung von signifikanten Orten 9
2.3 Vorhersagegenauigkeit von Zielzellen 11
2.4 Clustering von *Trips* 13
2.5 Wahrscheinlichkeiten von Routen 15

Literaturverzeichnis

Andrienko, G., & Andrienko, N. (2011). Position statement: Privacy issues in geospatial visual analytics. In *Advances in Location-Based Services*, Lecture Notes in Geoinformation and Cartography, (p. 239–246). Berlin: Springer-Verlag Berlin.

Andrienko, G., Andrienko, N., & Wrobel, S. (2007). Visual analytics tools for analysis of movement data. *ACM SIGKDD Explorations Newsletter*, *9*(2), 38–46.

Ashbrook, D., & Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, *7*(5), 275–286.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. New York: Springer New York.

Forgo, N., & Krügel, T. (2010). Der Personenbezug von Geodaten: Cui bono, wenn alles bestimmbar ist? *Zeitschrift für Informations-, Telekommunikations- und Medienrecht*, *1*, 17–23.

Froehlich, J., & Krumm, J. (2008). Route prediction from trip observations. In *Proceedings of SAE World Congress*. Detroit, MI, USA.

Gasson, M. N., Kosta, E., Royer, D., Meints, M., & Warwick, K. (2011). Normality mining: Privacy implications of behavioral profiles drawn from GPS enabled mobile phones. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *41*(2), 251–261.

Gogate, V., Dechter, R., Rindt, C., & Marca, J. (2005). Modeling transportation routines using hybrid dynamic mixed networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, (p. 217–224). Edingburgh, Scotland.

Gruteser, M., & Hoh, B. (2005). On the anonymity of periodic location samples. In *Security in Pervasive Computing*, Lecture Notes in Computer Science, (p. 179–192). Berlin: Springer-Verlag Berlin.

Hariharan, R., & Toyama, K. (2004). Project lachesis: Parsing and modeling location histories. In *Geographic Information Science*, Lecture Notes in Computer Science, (p. 106–124). Berlin: Springer-Verlag Berlin.

Hightower, J., Consolvo, S., LaMarca, A., Smith, I., & Hughes, J. (2005). Learning and recognizing the places we go. In *Proceedings of UbiComp 2005: Ubiquitous Computing*, (p. 159–176). Tokyo, Japan.

Hilty, L., Som, C., & Köhler, A. (2004). Assessing the human, social, and environmental risks of pervasive computing. *Human and Ecological Risk Assessment*, *10*(5), 853–874.

Hoh, B., Gruteser, M., Xiong, H., & Alrabady, A. (2006). Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, *5*(4), 38–46.

Hoh, B., Gruteser, M., Xiong, H., & Alrabady, A. (2007). Preserving privacy in GPS traces via uncertainty-aware path cloaking. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*, (p. 161–171). Alexandria, VA, USA.

Kalnis, P., Ghinita, G., Mouratidis, K., & Papadias, D. (2007). Preventing location-based identity inference in anonymous spatial queries. *IEEE Transactions on Knowledge and Data Engineering*, *19*(12), 1719–1733.

Kang, J., Welbourne, W., Stewart, B., & Borriello, G. (2004). Extracting places from traces of locations. In *Proceedings of the 2nd ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, (p. 110–118). Philadelphia, PA, USA.

Krumm, J. (2006). Real time destination prediction based on efficient routes. In *Proceedings of SAE World Congress*. Detroit, MI, USA.

Krumm, J. (2007). Inference attacks on location tracks. In *Proceedings of the Fifth International Conference on Pervasive Computing*, (p. 127–143). Toronto, Canada.

Krumm, J. (2011). Trajectory analysis for driving. In *Computing with Spatial Trajectories*, (p. 213–241). New York: Springer-Verlag New York.

Krumm, J., & Horvitz, E. (2006). Predestination: Inferring destinations from partial trajectories. In *Proceedings of UbiComp 2006: Ubiquitous Computing*, (p. 243–260). Orange County, CA, USA.

Laasonen, K., Raento, M., & Toivonen, H. (2004). Adaptive on-device location recognition. In *Pervasive Computing*, Lecture Notes in Computer Science, (p. 287–304). Berlin: Springer-Verlag Berlin.

Liao, L., Fox, D., & Kautz, H. (2007a). Extracting places and activities from gps traces using hierarchical conditional random fields. *The International Journal of Robotics Research*, *26*(1), 119–134.

Liao, L., Patterson, D., Fox, D., & Kautz, H. (2006). Building personal maps from GPS data. *Annals of the New York Academy of Sciences*, *1093*, 249–265.

Liao, L., Patterson, D., Fox, D., & Kautz, H. (2007b). Learning and inferring transportation routines. *Artificial Intelligence*, *171*(5-6), 311–331.

Marmasse, N., & Schmandt, C. (2000). Location-aware information delivery with commotion. In *Handheld and Ubiquitous Computing*, Lecture Notes in Computer Science, (p. 361–370). Berlin: Springer-Verlag Berlin.

Patterson, D., Liao, L., Fox, D., & Kautz, H. (2003). Inferring high-level behavior from low-level sensors. In *Proceedings of UbiComp 2003: Ubiquitous Computing*, (p. 73–89). Seattle, WA, USA.

Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones: An emerging tool for social scientists. *Sociological Methods & Research*, *37*(3), 426 –454.

Satyanarayanan, M. (2001). Pervasive computing: Vision and challenges. *IEEE Personal Communications*, *8*(4), 10–17.

Simmons, R., Browning, B., Zhang, Y., & Sadekar, V. (2006). Learning to predict driver route and destination intent. In *Proceedings of the Intelligent Transportation Systems Conference, ITSC'06.*, (p. 127–132). Toronto, Canada.

Weichert, T. (2007). Der Personenbezug von Geodaten. *Datenschutz und Datensicherheit-DuD*, *31*(2), 113–119.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678.

Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008). Learning transportation mode from raw gps data for geographic applications on the web. In *Proceeding of the 17th International Conference on World Wide Web*, (p. 247–256). Beijing, China.