

Masterthesis  
Fachhochschul-Studiengang  
Master Informatik

# **Implementierung eines Verfahrens zur automatisierten Verkehrsmittelerkennung auf Basis von Entscheidungsbäumen**

ausgeführt von

**Michael Zangerle, BSc**  
**1310249004**

zur Erlangung des akademischen Grades  
Master of Science in Engineering, MSc

Dornbirn, im Juli 2015

**Betreuer: Prof. (FH) Dipl.-Inform. Thomas Feilhauer**



# **Eidesstattliche Erklärung**

Ich erkläre hiermit an Eides statt, dass ich vorliegende Masterthesis selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Stellen sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder in gleicher noch in ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Dornbirn, am 29. Juli 2015

Michael Zangerle, BSc



## Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Suche nach einer Methode, welche GPS-Tracks hinsichtlich der verwendeten Verkehrsmittel analysiert. Dafür verfügen die Tracks selbst über keine zusätzlichen Informationen und es werden auch keine zusätzlichen Sensoren verwendet.

Für die Analyse von GPS-Tracks wurde eine grundlegende Filterlogik implementiert, um fehlerhafte Ausreißer aus den Rohdaten zu entfernen. Anschließend werden die verbleibenden Wegpunkte geteilt und zu Segmenten zusammengefasst, die mit nur einem Verkehrsmittel bewältigt worden sind. Für diese Segmente werden nun Zusatzwerte (z.B. Stopprate, Geschwindigkeit, ...) zu Bestimmung des Verkehrsmittels berechnet. Für die Bestimmung wurde auf Entscheidungsbäume als Schlussfolgerungsmodell gesetzt, da mit diesen in verschiedenen ähnlichen Publikationen vielversprechende Ergebnisse erzielt worden sind. Im letzten Schritt werden die Resultate ein letztes Mal mit Hilfe des Kontexts auf Plausibilität und Sinnhaftigkeit untersucht und gegebenenfalls angepasst.

Das Resultat dieser Arbeit ist der Prototyp einer Webapplikation, welche GPS-Spuren bezüglich der verwendeten Verkehrsmittel untersucht und diese anhand von verschiedenen Analysemethode bestimmt. Für diesen Prototyp wurden zwei Analysemethoden implementiert. Diese unterscheiden sich durch die Verwendung von GIS-Daten. Es konnte eine Erkennungsrate von bis zu 66% ohne bzw. 75% mit GIS-Daten erreicht werden. Jedoch muss hinzugefügt werden, dass nur wenig Daten für das Training der Schlussfolgerungsmodelle zur Verfügung standen. Test- und Trainingsdaten sollten aber in ausreichender Quantität und auch Qualität vorhanden sein, um eine bestmögliche Erkennungsrate zu erreichen.



## **Abstract**

The main subject of this thesis is the search for a method to identify transport modes from GPS tracks. Therefore the tracks itself should not contain any additional information and no additional sensor should be used.

For the analysis process a basic filter to remove the noise on recorded GPS tracks was implemented and the remaining track points divided into segments with one transport mode only. For these segments, additional values like a stop rate or velocity have been calculated. With these values the type of transport should be identified by using a decision tree as decision support tool. The decision tree was chosen because of it's promising results in other publications. Finally all segments will be analysed for the last time but with taking context into consideration to correct segments with a transport type that does not make sense or is not plausible.

The result of this thesis is a web application prototype which is able to process and identify a GPS track with different analysis methods. For the prototype two analyse methods have been implemented. They differ in their usage of GIS data. The accuracy of the classification achieved by these methods is 66% without and 75% with GIS data. But it has to be mentioned that very little data for training of the decision trees was available. Test and training data should be available in adequate quantity and quality to train a decision tree properly and reach a satisfying result when identifying GPS tracks.



# Inhaltsverzeichnis

<b>1 Einleitung</b>	<b>1</b>
1.1 Ziele dieser Arbeit . . . . .	2
1.2 Motivation und Nutzen . . . . .	4
1.3 Weiterer Aufbau der Arbeit . . . . .	5
<b>2 Stand der Technik</b>	<b>7</b>
2.1 Verwendete Daten . . . . .	8
2.2 Betrachtete Verkehrsmittel . . . . .	8
2.3 Analyse der Publikationen von Yu Zheng . . . . .	8
2.3.1 Geolife . . . . .	9
2.3.2 Segmentierung . . . . .	10
2.3.3 Schlussfolgerungsmodelle . . . . .	11
2.4 Analyse der Publikationen von Leon Stenneth . . . . .	12
2.4.1 GIS-Informationen . . . . .	12
2.4.2 Schlussfolgerungsmodelle . . . . .	13
2.5 Analyse der Publikationen von Filip Biljecki . . . . .	14
2.5.1 Segmentierung . . . . .	14
2.5.2 Schlussfolgerung . . . . .	15
2.6 Analyse der Publikationen von Sasank Reddy . . . . .	16
2.6.1 Sensoren . . . . .	17

## *Inhaltsverzeichnis*

2.6.2	Schlussfolgerungsmodelle . . . . .	17
2.7	Zusammenfassung . . . . .	18
<b>3</b>	<b>Modellbildung und Einbindung der GPS- und GIS-Daten</b>	<b>21</b>
3.1	Entscheidungsbaum als Schlussfolgerungsmodell . . . . .	22
3.1.1	Generierung eines Entscheidungsbaumes . . . . .	23
3.1.2	Die Entscheidungsbaum-Operatoren von RapidMiner . . . . .	27
3.1.3	Einfluss der GIS-Daten auf den Entscheidungsbaum . . . . .	29
3.2	Trainingsdaten . . . . .	33
3.3	Neue Aufzeichnungen . . . . .	36
3.4	GIS-Daten . . . . .	37
3.5	Weitere Daten . . . . .	38
<b>4</b>	<b>Der Prototyp</b>	<b>39</b>
4.1	Aufbau und Architektur . . . . .	40
4.1.1	Pipes-and-Filter-Architektur für Trainingsdaten . . . . .	41
4.1.2	Pipes und Filter-Architektur für die Webapplikation . . . . .	43
4.2	Verwendung der Applikation . . . . .	45
4.2.1	Create-Seite . . . . .	45
4.2.2	Results-Seite . . . . .	46
4.3	Konfiguration des Prototyps . . . . .	47
4.3.1	Konfiguration des Filters für fehlerhafte Ausreißer . . . . .	48
4.3.2	Konfiguration des Segmentierens . . . . .	48
4.3.3	Konfiguration der Analysemethoden . . . . .	49
4.3.4	Zusammenfassung . . . . .	50
<b>5</b>	<b>Segmentierung und Klassifizierung</b>	<b>53</b>
5.1	Segmentierung eines Tracks . . . . .	54

## *Inhaltsverzeichnis*

5.2	Schlussfolgerungsvariablen . . . . .	56
5.2.1	Reihung der allgemeinen Variablen . . . . .	58
5.2.2	Reihung der GIS-Variablen . . . . .	58
5.3	Berechnung der allgemeinen Entscheidungsvariablen . . . . .	61
5.3.1	Geschwindigkeit . . . . .	61
5.3.2	Beschleunigung . . . . .	62
5.3.3	Stopprate . . . . .	62
5.4	Berechnung der GIS-Entscheidungsvariablen . . . . .	64
5.4.1	Abstand zu Bushaltestellen . . . . .	64
5.4.2	Abstand zu Gleisen und zur Autobahn . . . . .	64
5.5	Klassifizierung der Verkehrsmittel . . . . .	65
5.5.1	Erstellen der Entscheidungsbäume . . . . .	65
5.5.2	Verwendung der Entscheidungsbäume . . . . .	67
5.5.3	Nachbearbeitung . . . . .	68
<b>6</b>	<b>Auswertung</b>	<b>71</b>
6.1	Erstellung des Entscheidungsbaums . . . . .	72
6.2	Analyse mit dem Prototyp . . . . .	73
6.2.1	Gesamtresultate . . . . .	73
6.2.2	Detailresultate . . . . .	74
6.3	Zusammenfassung . . . . .	76
<b>7</b>	<b>Ausblick</b>	<b>81</b>
7.1	GPX-Daten im Überblick . . . . .	81
7.1.1	GPX-Datenqualität . . . . .	81
7.1.2	Diversität der GPX-Daten . . . . .	82
7.1.3	Quantität der GPX-Daten . . . . .	83
7.2	GIS-Daten . . . . .	83
7.3	Entscheidungsbaum . . . . .	83

*Inhaltsverzeichnis*

<b>Literaturverzeichnis</b>	<b>85</b>
<b>Quellcodeverzeichnis</b>	<b>89</b>
<b>Tabellenverzeichnis</b>	<b>91</b>
<b>Abbildungsverzeichnis</b>	<b>94</b>
<b>Anhang 1</b>	<b>95</b>
<b>Anhang 2</b>	<b>103</b>

# Einleitung

Der Mensch produziert täglich eine extrem große Menge an Daten. Allein auf Youtube werden pro Minute 300 Stunden Video-Material veröffentlicht und täglich hunderte Millionen Stunden Videos konsumiert. [Youtube, 2015] Hochgerechnet auf das gesamte Internet und die gesamte Bevölkerung ergibt dies eine unvorstellbar große Menge an Daten, die bewusst oder auch unbewusst generiert werden.

Sehr viel an Daten wird auch durch diverse Fitnessgadgets, Smartwatches, Smartphones sowie Navis und ähnliche Geräte generiert. Abseits von Fitnesswerten sind viele dieser Geräte im Regelfall in der Lage GPS-Spuren aufzuzeichnen. Dies bedeutet, dass man genau nachvollziehen kann, wann man wo unterwegs war. Mit ein wenig Rechenarbeit können auch die Geschwindigkeit und viele andere Werte berechnet werden, sofern dies die Geräte nicht schon selbst machen.

Genau auf diesen GPS-Daten basiert diese Arbeit. Dabei ist es nicht wichtig, von welcher Person diese Daten stammen, sondern dass sich mit Hilfe dieser aufgezeichneten Daten feststellen lässt, wann ein Individuum sich auf welcher Strecke mit welchem Verkehrsmittel fortbewegt hat.

Analysiert man viele dieser Daten, so lassen sich viele Erkenntnisse daraus gewinnen. Unter anderem lassen sich zum Beispiel viel frequentierte Strecken in der Infrastruktur

## *1 Einleitung*

finden und mögliche Engstellen erkennen. In Folge lassen sich damit auch mögliche Verbesserungen, Einsparungsmöglichkeiten oder auch verstecktes Potential für zukünftige Projekte entdecken.

### **1.1 Ziele dieser Arbeit**

Das Hauptziel dieser Arbeit ist es, einen Prototypen zu erstellen, welcher anhand von aufgezeichneten GPS-Spuren und in Kombination mit verschiedenen Methodiken das benutzte Verkehrsmittel mit einer möglichst hohen Wahrscheinlichkeit bestimmt. Diese aufgezeichneten GPS-Spuren enthalten dabei keinerlei Informationen über die jeweilige Person. Deshalb erfolgt die Auswertung ausschließlich über die individuelle GPS-Spur sowie über öffentlich zugängliche Daten, wie zum Beispiel Busstationen und Gleise. Jede dieser Aufzeichnung kann dabei mehrere Verkehrsmittel beinhalten und von unterschiedlicher Länge und Dauer sein.

Die in dieser Arbeit berücksichtigten Verkehrsmittel sind in folgender Auflistung ersichtlich:

- Fußgänger
- Fahrrad
- Bus
- Auto (stellvertretend für Motorrad, Taxi, LKW, PWK etc.)
- Zug

Besonders interessant ist hierbei die Unterscheidung zwischen Bus und Auto in verkehrsabhängigen Situationen bzw. in der Stadt, da diese Transporttypen in diesen Situationen sehr ähnliche Verhaltensweisen und Werte zeigen.

Es soll weiters für jede Person, die ein Gerät besitzt, das in der Lage ist, eine GPS-Spur im GPX-Format aufzuzeichnen, möglich sein, diese Spur analysieren zu lassen. Dies

## *1 Einleitung*

bedeutet, dass keine speziellen Geräte oder andere Sensoren benötigt werden und dass sich dieser Prototyp mit möglichst geringen Anpassungen (Trainingsdaten, GIS-Daten und Konfigurationsparameter) auch in anderen Regionen anwenden lässt.

Im Zuge dieser Arbeit wird auch untersucht, welcher Grad an Genauigkeit sowohl mit als auch ohne geografische Zusatzinformationen im Raum Vorarlberg erreicht werden kann. Dabei soll es nach der Analyse die Möglichkeit geben, die automatisch bestimmten Verkehrsmittel manuell zu korrigieren, sollten diese nicht mit der Realität übereinstimmen. Weiters sollen diese manuellen Änderungen in die Auswertung mit einfließen.

Schlussendlich soll mit Hilfe der Auswertungen auch eine Aussage über die erzielte Genauigkeit mit und ohne die verwendeten Zusatzinformationen gemacht werden und untersucht werden, welche Zusatzwerte die Erkennungsrate wie beeinflussen. Außerdem soll erklärt werden, ob weitere Informationen (wie z.B. GPS-Daten der öffentlichen Verkehrsmittel) für eine noch genauere Bestimmung benötigt werden und welche Informationen dies sein könnten.

### **Nichtziele**

In dieser Arbeit werden die diversen Schlussfolgerungsmodelle (neuronales Netz, Bayessches Netz, Random Forest, Support Vector Machine, ...) nicht betrachtet oder verglichen, sondern es wird auf ein Modell gesetzt, das bereits bei anderen Arbeiten wie z.B. [Stenneth et al., 2011], [Reddy et al., 2010], [Sebastian Nagel, 2011] und [Zheng et al., 2008b] vielversprechende Ergebnisse erzielt hat. Dieses Modell ist der Entscheidungsbaum. Außerdem werden die Daten zur Analyse bzw. Auswertung nicht in Echtzeit betrachtet, sondern in Form einer GPS-Spur an den Prototypen übergeben.

Die Frage, an welcher Position man das Gerät zur Aufzeichnung am besten trägt, um möglichst genaue GPS-Daten zu erhalten, wird nicht weiter verfolgt, da die Daten möglichst realistisch sein sollen. Auch bleibt die Frage nach dem Energieverbrauch der App bzw. wie eine möglichst energieschonende App und die dazugehörige Kommunikation

## *1 Einleitung*

aufgebaut sein könnten, unberücksichtigt. Eine umfassende Behandlung der Themen Sicherheit und Privatsphäre ist im Rahmen der vorliegenden Arbeit nicht möglich und daher bleiben auch diese Themen unberührt. An dieser Stelle soll allerdings erwähnt werden, dass keine Informationen über den Benutzer / die Benutzerin in den GPS-Spuren benötigt oder vom Prototypen generiert oder gespeichert werden.

## **1.2 Motivation und Nutzen**

Schon früher wurde versucht, Aufzeichnungen über die Verkehrswege von verschiedenen Menschen zu sammeln. Aber die Protokolle in Papierform sowie die Telefonbefragungen waren zu aufwändig und die teilnehmenden Personen nicht zuverlässig genug. Darum ist es von entscheidendem Vorteil, eine App oder ein Gerät zur Verfügung zu haben, welches die Vorgänge des Aufzeichnens möglichst genau übernimmt. [Zheng et al., 2010]

Wird eine Auswertung mit einer für das Zielgebiet aussagekräftigen Anzahl an Personen durchgeführt, so kann das Resultat für verschiedenste Zwecke verwendet werden. Auch ohne spezielle Analyse kann rein durch die Betrachtung der gesammelten GPS-Spuren festgestellt werden, welche Routen besonders häufig benutzt werden.

Nimmt man nun verschiedene Werte aus der Auswertung hinzu, kann auch festgestellt werden, wo sich zum Beispiel verkehrstechnische Engstellen befinden und welche Routen sehr populär sind. Oder es können Aussagen über die allgemeine Verkehrssituation gemacht werden. Durch die gesammelten Daten lassen sich auch Simulationen für anstehende Bauvorhaben machen und es können Vorhersagen für bestimmte Situationen gemacht werden. Diese Aspekte können unter anderem für das Verkehrsministerium, den öffentlichen Personennahverkehr oder auch für die Stadtplanung sehr interessant sein (Optimierung von Auslastung, Einsparungspotentiale, ...).

## *1 Einleitung*

Eine ganze Reihe von Apps lässt sich mit den Auswertungen erstellen. Unter anderem könnten diese Apps die Auswertungen in soziale Medien integrieren oder für Fitnessanalysen verwendet werden. Einen einfachen Rückblick über die eigene Fortbewegung kann man damit genauso ermöglichen wie für Umweltbewusste errechnen, wie viel CO<sub>2</sub> sie produziert oder gespart haben. Ein Reisetagebuch könnte daraus genauso Nutzen ziehen wie eine App (sollten die Daten in Echtzeit ausgewertet werden), die beim Autofahren Auskunft über die aktuell billigste Tankstelle in näherer Umgebung gibt oder eine App, die einfach nur Vorschläge für alternative, schnellere Routen zu einem bekannten Ziel anbietet.

Zusammenfassend kann man sagen, dass die Verwendungsmöglichkeiten für solche Daten umfangreich sind und sich am besten unter dem Begriff kontextorientierte, geographische Applikationen zusammenfassen lassen. Nicht zuletzt öffnen sich mit solchen Daten aber auch umfangreiche Möglichkeiten für die Werbebranche.

### **1.3 Weiterer Aufbau der Arbeit**

Der Hauptteil der vorliegenden Arbeit gliedert sich in vier große Abschnitte:

Im ersten Abschnitt wird auf die Akquirierung der GPS-Daten eingegangen. Dies umfasst sowohl die gesammelten GPS-Aufzeichnungen und deren Struktur sowie die verwendeten GIS-Daten. Dabei geht es einerseits um deren Herkunft, andererseits auch darum, wie diese extrahiert wurden und wie diese Daten aufgebaut sind. Außerdem werden auch andere Daten, wie zum Beispiel GPS-Daten von Bussen des ÖPNV in Betracht gezogen.

Der zweite Abschnitt behandelt den entwickelten Prototypen, der die übergebenen GPS-Spuren analysiert. Dabei werden dessen Funktionalitäten erklärt und es wird der grundlegende Ablauf für den Benutzer/die Benutzerin dargelegt. Weiters wird auch auf die Architektur des Prototyps sowie auf dessen Konfigurationsmöglichkeiten eingegangen.

## *1 Einleitung*

Der dritte Abschnitt befasst sich sowohl mit dem Einbinden der GIS-Daten als auch dem Aufbereiten der GPS-Daten sowie dem Bestimmen der Verkehrsmittel. Mit dem Aufbereiten der Daten ist gemeint, dass zu den GPS-Punkten zusätzliche Werte für die spätere Analyse berechnet werden. Beispiele für diese Werte sind sowohl die Geschwindigkeit, Beschleunigung und Distanz als auch der Abstand zur nächsten Bushaltestelle.

Außerdem wird im dritten Abschnitt auch der Prozess des Aufteilens von GPS-Spuren in Teile (Segmente), in denen nur ein Verkehrsmittel verwendet wird, beschrieben. Diesem Schritt vorangegangen ist das in Anhang 1 beschriebene Filtern der GPS-Daten. Filtern bedeutet in diesem Zusammenhang, dass ein Teil der fehlerhaften Ausreißer aus den GPS-Spuren entfernt werden.

Aufbauend auf den Segmenten wird schließlich der Prozess der tatsächlichen Erkennung der Verkehrsmittel mit Hilfe eines Entscheidungsbaums und der berechneten Werte erklärt. Die aus dem Entscheidungsbaum gewonnenen Erkenntnisse werden schlussendlich ein letztes Mal überprüft, um sinnfreie bzw. sehr unwahrscheinliche Wechsel zwischen Verkehrsmitteln zu verhindern.

Der vierte und letzte Abschnitt des Hauptteils befasst sich mit der Auswertung der gewonnenen Erkenntnisse aus den vorhergehenden Abschnitten sowie den Testläufen mit neuen GPS-Spuren mit und ohne zusätzliche GIS-Informationen. Dabei wird auch untersucht, wie sich die verschiedenen Zusatzinformationen wie z.B. die Geschwindigkeit oder die Stoprate auf die tatsächliche Erkennung auswirken.

# Stand der Technik

Zu den Meilensteinen auf dem Forschungsgebiet der Verkehrsmittelerkennung (Transport Mode Detection) zählt die Arbeit von Yu Zheng, in welcher er unter anderem auf die Erkennung von den Abschnitten mit nur einem Verkehrsmittel eingeht. Weiters verwendete er in seiner Arbeit auch einen mit den GPS-Spuren gesammelten geographischen Wissen aufgebauten Graphen, welcher zur abschließenden Korrektur verwendet wurde. [Zheng et al., 2010] [Zheng et al., 2008a] [Zheng et al., 2008b]

Mit der Frage, wie geographische Daten in eine solche Analyse miteinbezogen werden können hat sich unter anderem Leon Stenneth beschäftigt. Dabei hat er nicht nur fixe Daten wie Gleise und Busstationen, sondern auch aktuelle Buspositionen miteinbezogen. [Stenneth et al., 2011]

Sowohl Stenneth als auch Zheng haben in ihren Arbeiten detailliert erklärt, wieso sie welche Attribute (Geschwindigkeit, Beschleunigung, ...) für die Bestimmung des Verkehrsmittels verwendet haben und sie haben diese auch durch Versuche nach ihrer Wichtigkeit gereiht. Außerdem haben beide und auch Sasank Reddy [Reddy et al., 2008] mehrere Schlussfolgerungsmodelle (Entscheidungsbaum, Bayessches Netz, Markov Modelle, Random Forest, ...) betrachtet und miteinander verglichen.

Wie man mit Verbindungsabbrüchen umgeht und zwischen ähnlichen Verkehrsmitteln

## *2 Stand der Technik*

unterscheiden kann, hat unter anderem auch Filip Biljecki beschäftigt. Des Weiteren baut er für die unterschiedlichen Kategorien von Verkehrsmitteln ein hierarchisches Modell auf, welches ihm helfen soll, bessere Entscheidungen zu treffen. [Biljecki et al., 2013]

### **2.1 Verwendete Daten**

Ein wesentlicher Unterschied zwischen all den betrachteten Publikationen sind die verwendeten Daten. Nur die GPS-Spuren bilden eine gemeinsame Basis. Manche Autoren untersuchten die Verwendung von GSM- und Wifi-Informationen [Reddy et al., 2010] oder stützten sich auf Zusatzinformationen durch weitere Sensoren wie zum Beispiel einem Beschleunigungssensor [Reddy et al., 2010] [Nadine Schüssler et al., 2011]. Andere, wie Leon Stenneth, verwendeten Echtzeitinformationen von öffentlichen Verkehrsmitteln und kombinierten diese mit GIS-Informationen, seien es Busstationen, Bahnstrecken, das Straßennetz oder Parkplätze [Stenneth et al., 2011].

### **2.2 Betrachtete Verkehrsmittel**

Ein weiterer Unterschied zwischen den Publikationen sind die betrachteten Verkehrsmittel. Hierbei reicht die Spanne der unterschiedenen Fortbewegungsmöglichkeiten von “gehen, laufen, Fahrrad und motorisiert“ [Reddy et al., 2010] bis hin zu “gehen, Zug, U-Bahn, Rad, Auto, Straßenbahn, Bus, Fähre, Segelboot und Flugzeug“ [Biljecki et al., 2013].

### **2.3 Analyse der Publikationen von Yu Zheng**

Zu den Meilensteinen auf dem Gebiet der “Transport Mode Detection“ zählen die Publikationen von Yu Zheng und seinem Team: “Understanding Mobility Based on GPS

## *2 Stand der Technik*

Data“ [Zheng et al., 2008a], “Understanding Transportation Modes Based on GPS Data for Web Applications“ [Zheng et al., 2010] und “Learning Transportation Mode from Raw GPS Data for Geographic Application on the Web“ [Zheng et al., 2008b]. In diesen Artikeln wird unter anderem auf die Erkennung von Abschnitten mit nur einem Verkehrsmittel eingegangen. Weiters wird erklärt, wie in den Projekten ein aus analysierten GPS-Spuren erstellter Graph mit geographischem Wissen zur Verbesserung der Erkennungsrate beiträgt und mit welchen Schlussfolgerungsmodellen welche Ergebnisse erzielt werden konnten.

### **2.3.1 Geolife**

Die Applikation, welche im Zusammenhang mit den oben genannten Arbeiten entstanden ist, nennt sich “Geolife“ und ist eine Webapplikation mit den Ansätzen eines sozialen Netzwerks. Dabei ging es darum, dass Benutzer und Benutzerinnen GPS-Spuren auf die Webseite laden konnten, diese Spur vom Algorithmus analysiert und die Verkehrsmittel bestimmt wurden. Dargestellt wurden die Resultate auf einer Karte, welche Benutzer und Benutzerinnen mit anderen teilen konnten.

Mit Hilfe der so gesammelten Daten konnte unter anderem Datamining betrieben und populäre Strecken festgestellt sowie Verkehrssituationen beurteilt werden. Außerdem konnten auch aktuelle Positionsdaten mit einem GPS-fähigen Smartphone/Handy ausgewertet werden und Informationen wie z.B. Abfahrtszeiten der öffentlichen Verkehrsmittel angeboten werden. Wurde aber z.B. eine Strecke mit dem Auto in die Stadt gesucht, so konnte auf Grund der bereits analysierten Fahrten über die durchschnittliche Geschwindigkeit festgestellt werden, welche Strecke am schnellsten ans Ziel führt.

In diesem von Microsoft Research geführten Projekt standen umfangreiche Test- und Trainingsdaten (1,2 Millionen Kilometer und 48.000 Stunden) zur Verfügung. [Microsoft Research, 2015] Ein Bildschirmfoto der Applikation ist in Abbildung 2.1 zu sehen.

## 2 Stand der Technik

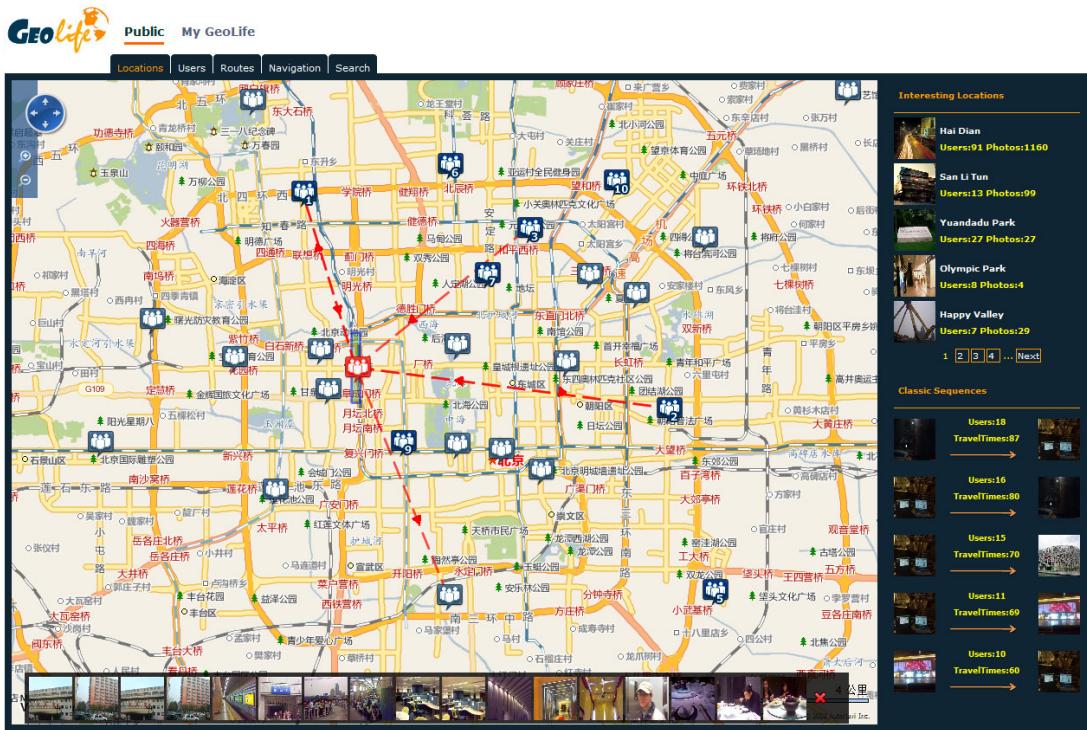


Abbildung 2.1: Geolife (Quelle: research.microsoft.com)

### 2.3.2 Segmentierung

Die Publikationen von Zheng beinhalten auch eine der detailliertesten Beschreibungen des Segmentierungsvorgangs. Dabei werden GPS-Spuren in Abschnitte, in welchen nur ein Verkehrsmittel verwendet wird, unterteilt. Diese Abschnitte können dann in weiterer Folge genauer analysiert und das benutzte Verkehrsmittel bestimmt werden.

Beim Segmentieren stützt sich Zheng darauf, dass Personen bei einem Wechsel des Verkehrsmittels stehen bleiben und sich ein Stück zu Fuß bewegen. Dies bedeutet, dass es einige GPS-Punkte mit einer Geschwindigkeit von genau oder beinahe 0 km/h gibt. Diese aufeinanderfolgenden Punkte können dann in einem "Gehen"-Segment zusammengefasst werden. Alle anderen Segmente werden vorläufig als "Nicht-Geh"-Segmente klassifiziert. Zheng sagt außerdem, dass der Beginn und das Ende eines Segments mit

## *2 Stand der Technik*

dem Typ “Gehen“ ein wichtiger Indikator für einen Wechsel des Fortbewegungsmittels ist. Diese Aussage stützt sich auf die Erkenntnisse aus GPS-Daten, die von 65 Personen über 10 Monate gesammelt wurden.

### **2.3.3 Schlussfolgerungsmodelle**

Um die Typen der “Nicht-Geh-Segmente“ bestimmen zu können, hat sich Zheng auf verschiedene Zusatzinformationen basierend auf den Segmenten gestützt, darunter:

- Distanz
- maximale Geschwindigkeit
- maximale Beschleunigung
- durchschnittliche Beschleunigung
- Richtungswechsel
- Stopprate
- erwartete Geschwindigkeit
- Geschwindigkeitsänderungsrate

In weiterer Folge stellte er fest, dass Stopp-, Richtungswechsel- und Geschwindigkeitsänderungsrate am effektivsten und stabilsten gegenüber den verschiedenen Verkehrssituationen sind. Diese drei Eigenschaften können auch mit ein paar der anderen kombiniert werden, um noch weitere Verbesserungen zu erhalten. Wurden allerdings zu viele Eigenschaften miteinbezogen, so konnte er eine Verringerung der Genauigkeit beim Bestimmen des Typs feststellen.

Für die tatsächliche Bestimmung des Typs führte Zheng verschiedene Experimente mit dem Entscheidungsbaum, der Support Vector Machine, dem bayesschen Netz und dem Conditional Random Field durch. Dabei stellte er fest, dass der Entscheidungsbaum die besten Ergebnisse im Zusammenspiel mit der in Abschnitt “2.3.3 Schlussfolgerungsmodelle“ beschriebenen Segmentierungsmethode liefert.

## *2 Stand der Technik*

Während des Analysierens der Trainingsdaten wurde im Hintergrund ein Graph aufgebaut, welcher Informationen über die jeweils betrachtete geografische Region widerspiegelt. In diesen geografischen Regionen waren 28 große Städte in China sowie mehrere Städte in den USA, Südkorea und Japan enthalten. Mit diesem Graph wurden die Schlussfolgerungen nochmals überprüft, und es konnte eine weitere Verbesserung erzielt werden. Schlussendlich konnte der von Zheng entwickelte Algorithmus 76,2% der Verkehrsmittel ohne jegliche Zusatzinformationen im Sinne von GIS-Daten oder anderen Sensordaten korrekt identifizieren.

## **2.4 Analyse der Publikationen von Leon Stenneth**

Leon Stenneth vergleicht in der Publikation “Transportation Mode Detection using Mobile Phones and GIS Information” [Stenneth et al., 2011], ähnlich wie Zheng auch, verschiedene Schlussfolgerungsmodelle. Allerdings verwendete er auch zusätzliche GIS-Informationen, um eine genauere Bestimmung zu ermöglichen. Außerdem arbeitet die entwickelte Applikation nicht mit ganzen GPS-Spuren, sondern analysiert immer zwei Punkte innerhalb eines 30-Sekunden-Intervalls, wodurch auch das eigentliche Segmente entfällt.

### **2.4.1 GIS-Informationen**

Die von Stenneth verwendeten GIS-Informationen beinhalten sowohl die Gleise von Zügen als auch Bushaltestellen sowie die aktuelle Position von allen Bussen in Chicago. Daraus berechnete er verschiedene Zusatzinformationen für die Bestimmung des Transporttyps:

- durchschnittliche Nähe zu den Gleisen
- durchschnittliche Nähe zu Bushaltestellen
- durchschnittliche Nähe zum nächsten Bus

## *2 Stand der Technik*

### **2.4.2 Schlussfolgerungsmodelle**

Um den Typ des jeweils betrachteten Abschnitts zu bestimmen, zieht auch Stenneth mehrere Zusatzwerte als Kriterien zu Hilfe:

- durchschnittliche Geschwindigkeit
- durchschnittliche Nähe zu den Gleisen
- durchschnittlicher Abstand zu einem Bus
- durchschnittliche Beschleunigung
- durchschnittlicher Richtungswechsel
- durchschnittliche Bushaltestellennähe
- durchschnittliche Genauigkeit der Koordinaten
- Distanz zum nächsten Bus

Wie Zheng stellt auch Stenneth fest, dass nur ein paar der Zusatzwerte wirklich effektiv sind. In seinem Fall waren das die durchschnittliche Geschwindigkeit und Beschleunigung, die Nähe zu den Gleisen, der Abstand zu anderen Bussen sowie die Distanz zum nächsten Bus.

Die von Stenneth betrachteten Modelle sind Naive Bayes, bayessches Netz, Entscheidungsbaum, Random Forest und Multilayer Perceptron. In den Experimenten stellte er fest, dass der Random Forest mit den GIS-Informationen das vielversprechendste Modell mit mehr als 93% richtig erkannten Verkehrsmitteln ist. Ohne GIS-Informationen schneidet auch der Random Forest ähnlich wie bei Zheng der Entscheidungsbaum mit 76% ab. Mit GIS-Informationen erreicht auch der Entscheidungsbaum eine Erkennungsrate von mehr als 92%.

## **2.5 Analyse der Publikationen von Filip Biljecki**

Filip Biljecki veröffentlichte eine Arbeit mit dem Titel “Transportation mode-based segmentation and classification of movement trajectories“ [Biljecki et al., 2013]. Darin vergleicht er nicht nur eine Vielzahl von Publikationen zu dem Thema Verkehrsmittelerkennung (unterer anderem [Schuessler and Axhausen, 2009], [Zheng et al., 2010], [Reddy et al., 2010], [Gonzalez et al., 2010]) anhand der Fortbewegungsmittel sowie der Zusatzwerte, ob GIS-Daten verwendet wurden und welches Resultat erzielt worden ist, sondern führt auch ein hierarchisches Modell für die Erkennung der Transportmittel ein. Weiters verwendet er zur Bestimmung des Transportmittels auch GIS-Daten wie Bus- und Straßenbahnhaltestellen. Ein Kapitel in dieser Arbeit ist auch der Behandlung von Störungen oder Unterbrechungen des GPS-Signals gewidmet. Darin wird aufgeführt, welche Fälle weiter behandelt werden können und welche sich mit diesem Ansatz nicht beheben lassen. In dieser Arbeit konnte eine Genauigkeit von 95% erreicht werden.

### **2.5.1 Segmentierung**

Für die Segmentierung verwendet Biljecki dieselbe Methodik wie Zheng, aber erweitert diese dahingehend, dass auch dann segmentiert wird, wenn ein Signalverlust (30 Sekunden keine neuen Werte) feststellt werden kann. Diese Entscheidung wird damit begründet, dass es bei einem Verkehrsmittelwechsel oft zu einem Signalverlust kommt. Überschreitet ein Stopp eine bestimmte Dauer (12 Sekunden) so wird auch segmentiert, denn dies könnte auch auf einen Wechsel hindeuten. Da die Track-Punkte nicht 100% genau sind, wird alles unter 2km/h als Stopp eingestuft. Weil nach der Erkennung alle aufeinanderfolgenden Segmente vom selben Typ zusammengefasst werden, ist eine Übersegmentierung kein Problem.

### 2.5.2 Schlussfolgerung

Um das Fortbewegungsmittel feststellen zu können, verwendet Biljecki ein Expertensystem. Dieses System basiert auf einer Fuzzy-Logic und klassifiziert die Segmente mit Wahrscheinlichkeitswerten. Dazu verwendet dieses System auch die hierarchische Gliederung der verschiedenen Verkehrsmittel, welche in Abbildung 2.2 ersichtlich ist. Diese Gliederung soll verhindern, dass sich das System zu früh auf einen Typ festlegen muss - das System kann den Typ dadurch schrittweise bestimmen.

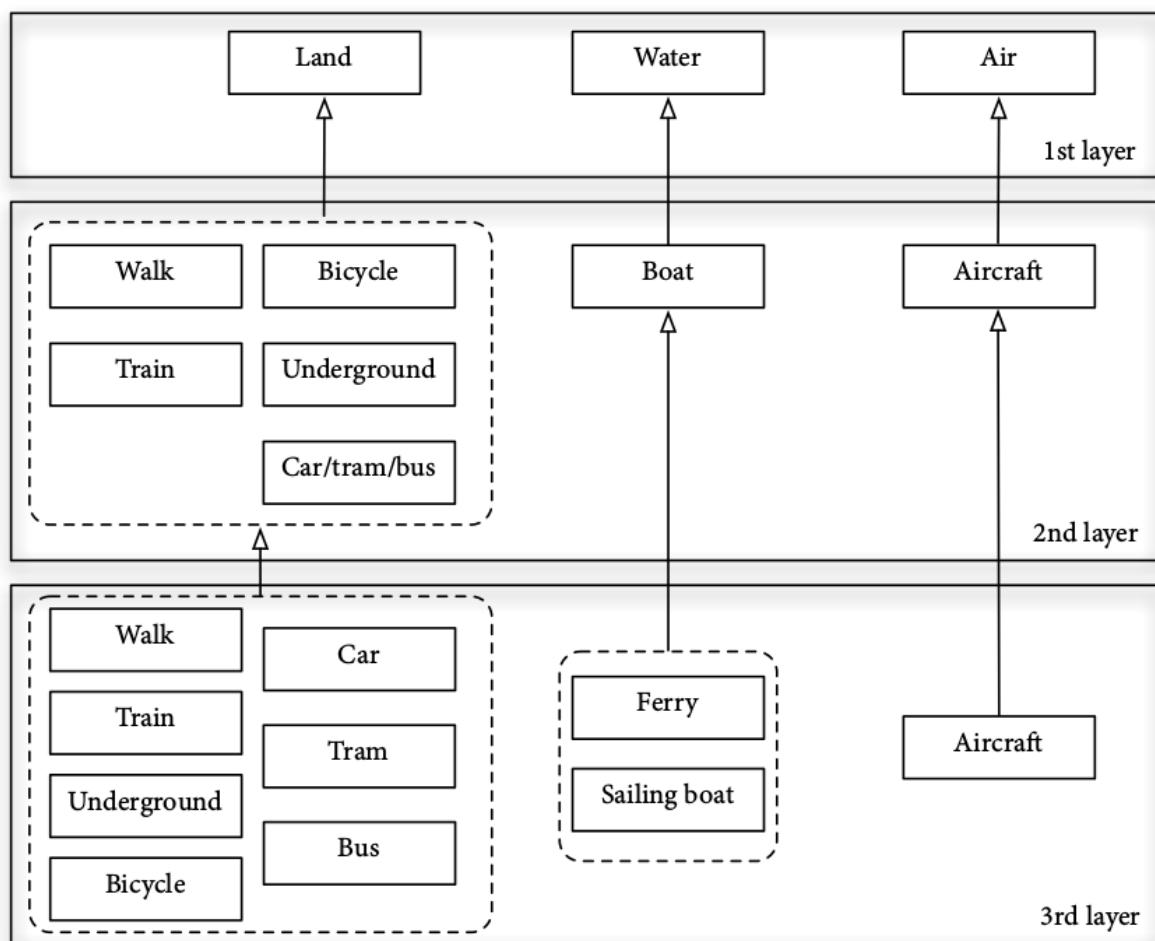


Abbildung 2.2: Fortbewegungsmittel-Hierarchie (Quelle: [Biljecki et al., 2013])

## *2 Stand der Technik*

Zusammenfassend sagt Biljecki, dass es sehr schwierig ist, alle möglichen Fälle der Realität abzubilden und wirklich zufriedenstellende Resultate zu erhalten. Dies betrifft vor allem den Segmentierungs- und Bestimmungsprozess von Daten mit sehr vielen fehlerhaften Ausreißern (Rauschen) oder jene Fällen, in denen wenige Beispieldaten vorhanden sind. Weiters meint er, dass Fehler nicht zwangsläufig auf das System zurückzuführen sind, sondern dass es sich bei diesen Fehlern oft um spezielle Situationen handelt, welche sehr kompliziert zu modellieren sind oder drastische Auswirkungen auf die allgemeine Performanz haben.

Alles in allem kann aber gesagt werden, dass das von Biljecki vorgestellte Modell zehn verschiedene Verkehrsmittel unterscheiden kann. Dies sind mehr als in allen anderen vom Autoren betrachteten Publikationen. Außerdem konnte durch die Verwendung von GIS-Daten bessere Resultate (95% bis 100% je nach Daten) als in vielen anderen Publikationen im Bereich der Verkehrsmittelerkennung erzielt werden.

## **2.6 Analyse der Publikationen von Sasank Reddy**

Sasank Reddy hat sich in den Publikationen “Using Mobile Phones to Determine Transportation Modes” [Reddy et al., 2010] und “Determining Transportation Mode on Mobile Phones” [Reddy et al., 2008] wie auch Stenneth und Zheng mit verschiedenen Schlussfolgerungsmodellen befasst. Bevor diese Modelle aber zum Einsatz kamen, wurde evaluiert, mit welchen weiteren Sensoren man sinnvolle Daten aufzeichnen könnte. Im Zuge dessen wurde auch überprüft, an welcher Stelle am Körper das Smartphone die genauesten Daten liefert und wie möglichst energieeffizient Daten aufgezeichnet und übermittelt werden können. Im Gegensatz zu anderen Publikationen wurde in diesen Arbeiten nicht genauer zwischen den motorisierten Verkehrsmitteln unterschieden.

## *2 Stand der Technik*

### **2.6.1 Sensoren**

Da die meisten Smartphones nicht nur über ein GPS-Modul sondern auch über WIFI, einen Beschleunigungssensor sowie Bluetooth und natürlich über ein GSM-Modul verfügen, wurde in diesen Arbeiten auch evaluiert, ob es möglich und sinnvoll ist, Daten von diesen Geräten und Sensoren miteinzubeziehen.

Bluetooth wäre zwar interessant, aber es kommt hauptsächlich innerhalb von Gebäuden zum Einsatz (TV, Radio, Computer, ...). Darum konnte dieser Sensor nicht für die Bestimmung des Verkehrsmittels eingesetzt werden.

Wifi und GSM in den Erkennungsprozess miteinzubeziehen konnte nach einer Reihe von Versuchen ausgeschlossen werden, da der Grad der Verbesserung nur 0,6% betrug und ein wesentlich höherer Energiebedarf gegeben war. Außerdem war die Abhängigkeit von Wifi und GSM in ländlichen Gegenden mit schlechtem Empfang und / oder wenig Wifis ein weiterer Grund, der gegen diese Sensoren sprach.

Die vielversprechendste Kombination war jene aus GPS-Daten und den Daten des Beschleunigungssensors, welche auch für die weiteren Experimente verwendet wurde.

### **2.6.2 Schlussfolgerungsmodelle**

Um das Fortbewegungsmittel des jeweiligen Abschnittes zu erkennen, zieht Reddy folgende Zusatzinformationen heran:

- Geschwindigkeit
- Varianz des Beschleunigungssensorsignals
- 3 Beschleunigungssensorwerte (3Hz, 2Hz, 1Hz)

Die in dieser Publikation betrachteten Schlussfolgerungsmodelle sind der Entscheidungsbaum, K-Means Clustering, Naive Bayes, Nearest Neighbor, Support Vector Machine sowie ein Continuous Hidden Markov Model und ein Entscheidungsbaum in Kombination

## *2 Stand der Technik*

mit einem Discrete Hidden Markov Model. Dabei konnte festgestellt werden, dass die letzte Variante die besten Resultate mit 93,6% erzielte. Wie aber auch in den anderen Publikationen war der Entscheidungsbaum mit 91,3% nicht weit abgeschlagen hinter dem kombinierten Ansatz.

## **2.7 Zusammenfassung**

Neben den vorgestellten Publikationen zum Thema Verkehrsmittelerkennung gibt es noch weitere interessante Publikationen, deren Erfahrungen zum Teil in diese Arbeit eingeflossen sind. Unter diesen Publikationen sind unter anderem “Processing raw data from global positioning systems without additional information“ [Schuessler and Axhausen, 2009] und “Improving post-processing routines for gps oversavations using propted-recall data“ [Nadine Schüssler et al., 2011] von Nadine Schüssler sowie “Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks“ [Gonzalez et al., 2010] von Paola Gonzalez.

Für diese Arbeit wurde für den Segmentierungsprozess die Vorgehensweise von Yu Zheng [Zheng et al., 2010] verwendet. Darin sind aber auch die von Filip Biljecki [Biljecki et al., 2013] vorgeschlagenen Änderungen eingeflossen, welche eine weitere Segmentierungsregel bei Empfangsverlust beinhalten.

Für die weitere Verarbeitung der GPS-Daten wurde der Entscheidungsbaum aufgrund des guten Abschneidens in den erwähnten Publikationen verwendet. Der Entscheidungsbaum wurde zur nachfolgenden Analyse und Auswertung der verbesserten Verkehrsmittelerkennung in zwei Varianten implementiert. In der einen Variante verwendet er nur aus den GPS-Daten berechnete Werte (Geschwindigkeit, Beschleunigung, ...), wie es auch in vielen anderen Arbeiten gemacht wurde. In der zweiten Variante greift der Entscheidungsbaum aber auch auf GIS-Informationen zurück, wie es zum Beispiel Leon

## *2 Stand der Technik*

Stenneth in seiner Publikation [Stenneth et al., 2011] beschrieben hat.

Ein weiterer Grund für die Verwendung von GIS-Daten war neben dem guten Abschneiden der GIS-Daten-gestützten Verkehrsmittelerkennung, die eigene Zielsetzung, dass keine zusätzlichen Sensoren oder speziellen Geräten neben dem Gerät für die Aufzeichnung der GPS-Daten verwendet werden sollen.



# Modellbildung und Einbindung der GPS- und GIS-Daten

Dieser Abschnitt behandelt die Akquirierung und die Struktur der verwendeten GPS-Daten für das Training des Entscheidungsbaums. Außerdem wird erläutert, wieso der Entscheidungsbaum als Schlussfolgerungsmodell ausgewählt und wie er erstellt wurde.

Sowohl für die Trainingsdaten für den Entscheidungsbaum als auch für die Testdaten wird erläutert, wie diese aufgezeichnet wurden. Dies inkludiert sowohl die Geräte als auch die Software, welche dazu verwendet wurde. Weiters wird auf die Struktur der GPS-Spuren und der GIS-Daten eingegangen.

Außerdem wird dargelegt, woher die verwendeten GIS-Daten stammen, wie diese extrahiert wurden und welche Rolle sie im weiteren Prozess spielen. Abschließend wird auch auf die Positionsdaten der verschiedenen Verkehrsmittel des ÖPNV eingegangen und in welcher Weise diese hätten eingesetzt werden können.

### **3.1 Entscheidungsbaum als Schlussfolgerungsmodell**

Aufgrund des guten Abschneidens des Entscheidungsbaums in verschiedenen Arbeiten - [Stenneth et al., 2011], [Reddy et al., 2010], [Sebastian Nagel, 2011] und [Zheng et al., 2008b] - wurde auch in dieser Arbeit ein Entscheidungsbaum als Schlussfolgerungsmodell verwendet. Um einen Entscheidungsbaum erstellen zu können, werden Trainingsdaten benötigt. Daraus werden dann die Regeln für den Entscheidungsbaum bzw. die jeweiligen Entscheidungen des Baums abgeleitet.

Deshalb wurde für diese Trainingsdaten ein Teil der gesammelten GPS-Daten manuell segmentiert und mit dem benutzten Verkehrsmittel ergänzt. Mit Hilfe des Prototyps wurden die Trainingsdaten eingelesen, gefiltert und mit zusätzlichen Informationen angereichert. Diese zusätzlichen Informationen sind bei der Variante ohne GIS-Daten zum Beispiel Geschwindigkeit und Beschleunigung. Bei der Variante mit GIS-Daten wird unter anderem die Nähe zu Bushaltestellen oder zu Schienen ergänzt. Danach wurden für alle eingelesenen Segmente die berechneten Werte und das dazugehörige Verkehrsmittel in einer Datei im CSV-Format abgelegt.

Für die konkrete Generierung der Entscheidungsbäume wurde das Tool RapidMiner verwendet. Dabei handelt es sich um eine Datamining Software, welche sowohl verschiedene Datenquellen unterstützt (Datenbanken und auch einzelne Dateien in unterschiedlichen Dateiformaten) als auch die Generierung von Entscheidungsbäumen über eine komfortable grafische BenutzerInnenoberfläche erlaubt. In der freien Version ist jedoch nur der CSV-Import verfügbar. Deshalb wurde vom Prototypen auch eine CSV-Datei generiert, welche in der Folge einfach in RapidMiner eingebunden und ausgewertet werden konnte.

Schlussendlich wurden aus den eingebundenen Daten Entscheidungsbäume für die Variante mit bzw. ohne GIS-Daten generiert. Diese können sowohl als Bilder als auch in Textform exportiert werden. In weiterer Folge wurden aus diesen Entscheidungsbäumen

### *3 Modellbildung und Einbindung der GPS- und GIS-Daten*

in Textform PHP-Klassen generiert (siehe Abschnitt “5.5.1 Erstellen der Entscheidungsbäume“).

#### **Überanpassung (Overfitting)**

Beim Erstellen von Entscheidungsbäumen wie auch bei anderen von Trainingsdaten lernenden Algorithmen muss beachtet werden, dass das Ergebnis nicht zu sehr auf die Trainingsdaten zugeschnitten ist. Man möchte also, dass mit Hilfe der Trainingsdaten ein Modell generiert wird, welches auch für Nichttrainingsdaten ein akzeptables Resultat liefert und nicht zu sehr auf die Gegebenheiten / Ungenauigkeiten in den Trainingsdaten spezialisiert ist. Ist dies jedoch der Fall, so spricht man von einer Überanpassung (Overfitting) des Modells an die Daten. [Tom Dietterich, 1995]

#### **Zurückschneiden (Pruning)**

Ist ein Entscheidungsbaum zu sehr auf die Trainingsdaten angepasst, so muss dieser wieder zurückgeschnitten werden. Dies bedeutet, dass einzelne Blätter oder auch Teile des Baums entfernt werden, um ein bestmögliches Resultat (geringe Anzahl an Fehlern) für Nichttrainingsdaten zu erhalten. Für diese Aufgabe wurden verschiedene Algorithmen entwickelt, welche unter anderem in der Publikation “Pruning Decision Trees with Misclassification Costs“ von Jeffrey Bradford beschrieben werden. [Jeffrey P. Bradford et al., 1998]

#### **3.1.1 Generierung eines Entscheidungsbaumes**

Um einen Entscheidungsbaum generieren zu können, gibt es verschiedene Algorithmen und Ansätze, die im folgenden Abschnitt kurz erklärt werden. Die betrachteten Algorithmen sind:

- ID3

### *3 Modellbildung und Einbindung der GPS- und GIS-Daten*

- C4.5
- CART
- CHAID

#### **ID3**

Der ID3 Algorithmus (Iterative Dichotomiser 3) wurde von John Ross Quinlan 1986 entwickelt und ist der Vorgänger von C4.5. ID3 wurde entworfen, um nicht alle möglichen Entscheidungsbäume generieren zu müssen, wenn man nur einen möglichst guten Entscheidungsbaum für die gegebenen Daten benötigt. Im Allgemeinen generiert ID3 einfache Entscheidungsbäume, aber es kann nicht garantiert werden, dass es keine besseren Entscheidungsbäume gibt. Der allgemeine Algorithmus ist iterativ und geht dabei wie folgt vor: [Quinlan, 1986]

- Es wird eine Teilmenge der Trainingsdaten per Zufall ausgewählt (Window) und für diese ein Entscheidungsbaum generiert. Dieser Entscheidungsbaum ist für alle Elemente innerhalb der Teilmenge gültig und kann diese korrekt klassifizieren.
- Nachfolgend werden alle anderen Elemente mit diesem Entscheidungsbaum klassifiziert. Kann der Baum alle Elemente korrekt klassifizieren, so ist ein ausreichend guter Entscheidungsbaum gefunden und das Ziel erreicht. Wenn nicht alle Elemente korrekt klassifiziert werden können, wird eine Auswahl der falsch klassifizierten Elemente in die Teilmenge (Window) übernommen und erneut ein Entscheidungsbaum gesucht.

Für das Erstellen des Entscheidungsbaumes selbst wird dabei auf das Finden des Attributs / der Entscheidungsvariable mit der niedrigen Entropie gesetzt. Diese Attribute oder Entscheidungsvariablen sind in diesem Fall die zusätzlich berechneten Werte pro Segment (Beschleunigung, Stopprate, ...). Entropie ist in diesem Zusammenhang ein Maß für die Reinheit bzw. die Verunreinigung. [Howard Hamilton, 2009] Ein hoher Reinheitswert bedeutet zugleich einen Gewinn an Information (information-gain),

### *3 Modellbildung und Einbindung der GPS- und GIS-Daten*

da sich der Informationsgewinn mit Hilfe der berechneten Entropie-Werte bestimmen lässt. [Thomas Mitchell, 1997]

Dies bedeutet, man sucht ein Attribut, welches für eine möglichst reine Menge an Objekten in diesem Knoten sorgt. Hat man zum Beispiel eine Menge von männlichen und weiblichen Personen, so sucht man ein Attribut, welches bewirkt, dass im nächsten Knoten eine möglichst hohe Wahrscheinlichkeit für männliche oder weibliche Personen gegeben ist und somit die Anzahl möglichst nicht ausbalanciert ist. Wäre die Anzahl genau ausbalanciert, so würde man durch diesen Knoten keinen Schritt näher an die schlussendliche Klassifizierung kommen, was bedeutet, die Verunreinigung wäre maximal und der Informationsgewinn minimal.

Mit diesem Ansatz können Entscheidungsbäume bereits nach wenigen Iterationen für ein Trainingsset von bis zu 30.000 Objekten und 50 Attributen gefunden werden. [Quinlan, 1986]

ID3 macht kein Backtracking und überdenkt sozusagen seine Auswahl an Attributen zu keinem späteren Zeitpunkt. Dies kann auch dazu führen, dass ein lokales, aber kein globales Optimum gefunden wird. [Howard Hamilton, 2009] Weiters kann es vorkommen, dass der erstellte Entscheidungsbaum überangepasst ist, wenn eine zu kleine Menge an Trainingsdaten verwendet worden ist. [RapidMiner, 2015]

## C4.5

Der C4.5 Algorithmus ist eine Erweiterung des ID3 Algorithmus, welche unter anderem versucht, folgende Probleme des ID3 Algorithmus zu adressieren: [Howard Hamilton, 2009]

- Überanpassen des Entscheidungsbaums an die Daten
- Vermeiden von fehlerhaftem Zurückschneiden
- Verbesserung der Effektivität der Berechnung

### *3 Modellbildung und Einbindung der GPS- und GIS-Daten*

- Handhabung von nicht diskreten Werten durch Umformen der Bedingung für das Attribut auf größer oder kleiner gleich. Kann ein Attribut eine unendliche Anzahl von Werten annehmen, so wird durch diese umgeformten Bedingungen entschieden, welchem Pfad gefolgt werden soll.
- Handhabung von Trainingsdaten mit fehlenden Attributen
- Nachträgliches Zurückschneiden von Teilbäumen und Ersetzen dieser durch Blätter-Knoten

#### **CART**

Der CART-Algorithmus benutzt im Gegensatz zu C4.5 und ID3 den Gini-Index als Entscheidungskriterium zur Bestimmung des nächsten Attributs. [RapidMiner, 2015] Der Gini-Index ist ein weiteres statistisches Maß für die Angabe der Reinheit.

Der Algorithmus wählt die Attribute auch nach dem Kriterium der Reinheit für die daraus folgenden Knoten. Das bedeutet, dass er die Daten so aufzuteilen versucht, dass möglichst reine Ergebnisse in den neuen Knoten entstehen. An dieser Stelle stoppt der Algorithmus jedoch nicht, sondern er wählt jenes Attribut, welches die Reinheit maximiert, und erzeugt eine Reihe weiterer Teilbäume. Diese Teilbäume werden dann bis zur Wurzel zurückgeschnitten. Dabei wird abgeschätzt, wie hoch die Kosten für eine falsche Klassifizierung sind, und jener Teilbaum mit den geringsten Kosten gewählt. [Wei-Yin Loh, 2008]

#### **CHAID**

CHAID steht für “CHi-squared Automatic Interaction Detection“ und verwendet statt der Entropie und des Informationsgewinns den Chi-Quadrat-Test für die Auswahl des nächsten Attributs. Eine Erklärung zu diesem Test liefert zum Beispiel Dr. Chandran in der Präsentation mit dem Titel “Chi-Square Test“ <http://de.slideshare.net/syamchandran3/chi-squared-test-2>. [RapidMiner, 2015]

### *3 Modellbildung und Einbindung der GPS- und GIS-Daten*

Weiters arbeitet dieser Algorithmus nicht mit numerischen Werten und stoppt das Wachstum, des Baumes, bevor der Baum zu groß wird. CHAID benötigt eine große Menge an Daten, um verlässliche Ergebnisse / Entscheidungsbäume erzeugen zu können. [Rapid-Miner, 2015]

#### **Zusammenfassung**

Die betrachteten Algorithmen unterscheiden sich nicht nur durch ihre eigenständigen Vorgehensweisen, sondern auch durch die verwendeten Entscheidungskriterien. Dabei setzen ID3 und C4.5 auf den Informationsgewinn, CART auf den Gini-Index und CHAID auf den Chi-Quadrat-Test. Der Informationsgewinn und der Gini-Index unterscheiden sich dabei zwar in der Berechnung, wählen aber laut Laura Raileanu in nur zwei Prozent der Fälle unterschiedliche Attribute aus [Raileanu and Stoffel, 2004].

#### **3.1.2 Die Entscheidungsbaum-Operatoren von RapidMiner**

Wie bereits am Beginn des Kapitels beschrieben wurde, wurde die Datamining-Software RapidMiner verwendet. Diese Software bietet verschiedene Operatoren für die verschiedenen Algorithmen an, darunter einen ID3-Operator für den ID3-Algorithmus und einen CHAID-Operator für den CHAID-Algorithmus. Weiters gibt es einen Decision-Tree-Operator, welcher abhängig vom ausgewählten Klassifizierungskriterium, den CART-Algorithmus (“gini-index”) oder C4.5-Algorithmus (“gian-ratio”, “information-gain”) für den Entscheidungsbaum auswählt. [RapidMiner, 2015]

Neben den erwähnten Algorithmen gibt es noch den C5.0- und den MARS-Algorithmus, welche aber nicht in RapidMiner verfügbar sind und deshalb auch nicht weiter betrachtet werden.

RapidMiner bietet die Möglichkeit, sowohl Einfluss auf das Zurückschneiden als auch auf das Überanpassen von Entscheidungsbäumen zu nehmen. Grundsätzlich kann ein

### *3 Modellbildung und Einbindung der GPS- und GIS-Daten*

Entscheidungsbaum durch verschiedene Parameter beeinflusst werden (siehe Abbildung 3.1). Die für diese Arbeit interessanten Parameter waren die maximale Tiefe des Baumes, der Zuversichtswert und die Option, welche das Zurückschneiden anwendet.

Die interessanteste Option ist allerdings das Feld Kriterium (criterion), da mit diesem beeinflusst werden kann, mit welchem Entscheidungskriterium entschieden (Gini-Index, Informationsgewinn, ...) wird, welches Attribut beim Aufbau eines Entscheidungsbaums als nächstes verwendet wird. Indirekt kann dadurch auch beeinflusst werden, welcher Algorithmus (CART, C4.5) verwendet wird um den Entscheidungsbaum zu generieren. RapidMiner bietet dabei mehrere Optionen an, wobei die Interessantesten folgende sind:

- **Informationsgewinn (information\_gain)** Der Informationsgewinn wird über die Entropie (die Reinheit) der entstehenden Teilresultate in diesem potentiellen neuen Knoten berechnet und wird vom ID3- und C4.5-Algorithmus verwendet.
- **Informationsgewinnrate (gain\_ratio)** Die Informationsgewinnrate berechnet für alle Attribute eine Gewichtung auf Basis der Anzahl und der Größe der entstehenden Teilbäume. Je höher das Gewicht, umso wichtiger ist das Attribut. Dadurch versucht man die Nachteile, die durch den reinen Informationsgewinn entstehen können, zu verhindern. Ein Beispiel wäre bei Kundendaten der hohe Informationsgewinn durch die Kreditkartennummer, da diese einen Kunden eindeutig identifiziert. Aber ein solches Attribut ist nicht zwangsläufig jenes, welches eine so hohe Relevanz im Entscheidungsbaum haben soll. Dies versucht man durch eine Gewichtung von Attributen durch die Informationsgewinnrate zu verhindern. Diese Rate kann vom ID3- und C4.5-Algorithmus als Entscheidungskriterium verwendet werden. [RapidMiner, 2015] [Johannes Fürnkranz, 2008]
- **Gini-Index (gini\_index)** Der Gini-Index ist ein statistisches Maß (wie der Informationsgewinn) und wird vom CART-Algorithmus verwendet. Dabei wird im Gegensatz zum Informationsgewinn nicht die Entropie sondern die Verunreinigung

### *3 Modellbildung und Einbindung der GPS- und GIS-Daten*

gemessen. Diese wird in weiterer Folge minimiert um die das beste Attribut bestimmen zu können. [Johannes Fürnkranz, 2008]

Wie sich die unterschiedlichen Methodiken / Berechnungsarten auf die Resultate auswirken, wird im Abschnitt “6.1 Erstellung des Entscheidungsbaums“ beschrieben.

Die maximale Tiefe des Baumes ist dabei ein maximaler Wert, der das Wachstum für den Baum einschränkt und dadurch die Spezialisierung und mögliche Überanpassung an die Testdaten ab einem gewissen Punkt beschränkt. Der minimale Informationsgewinn hat dabei einen ähnlichen Effekt wie die maximale Tiefe. Kann beim Aufteilen des aktuellen Knotens ein weiterer Informationsgewinn über dem angegebenen Wert erreicht werden, so wird dieser aufgeteilt. Ansonsten wird der Baum an diesem Punkt nicht mehr wachsen, was wieder eine Überanpassung verhindern kann. [RapidMiner, 2015]

Der Zuversichtswert ist ein Wert, der für das Zurückschneiden des Baumes verwendet wird und angibt, welches Level an Zuversicht gegeben sein muss, um das Zurückschneiden tatsächlich anzuwenden (er steht im Zusammenhang mit den durch das Zurückschneiden verursachten Fehlentscheidungen). [RapidMiner, 2015]

#### **3.1.3 Einfluss der GIS-Daten auf den Entscheidungsbaum**

Der Entscheidungsbaum ohne GIS-Daten ist in Abbildung 3.2 zu sehen. Bei diesem Entscheidungsbaum wurden mittlere und maximale Geschwindigkeit, Stopprate sowie mittlere und maximale Beschleunigung als Indikatoren für den Entscheidungsprozess gewählt. Wie diese Werte berechnet und ergänzt werden, wird im Abschnitt “5.2 Schlussfolgerungsvariablen“ erklärt.

Der Entscheidungsbaum mit GIS-Daten ist in Abbildung 3.3 zu sehen. Bei diesem Entscheidungsbaum wurden zusätzlich zu den Geschwindigkeits-, Beschleunigungs- und Stoppwerten auch Werte für die verwendeten GIS-Daten gesammelt. Die zusätzlichen

### 3 Modellbildung und Einbindung der GPS- und GIS-Daten

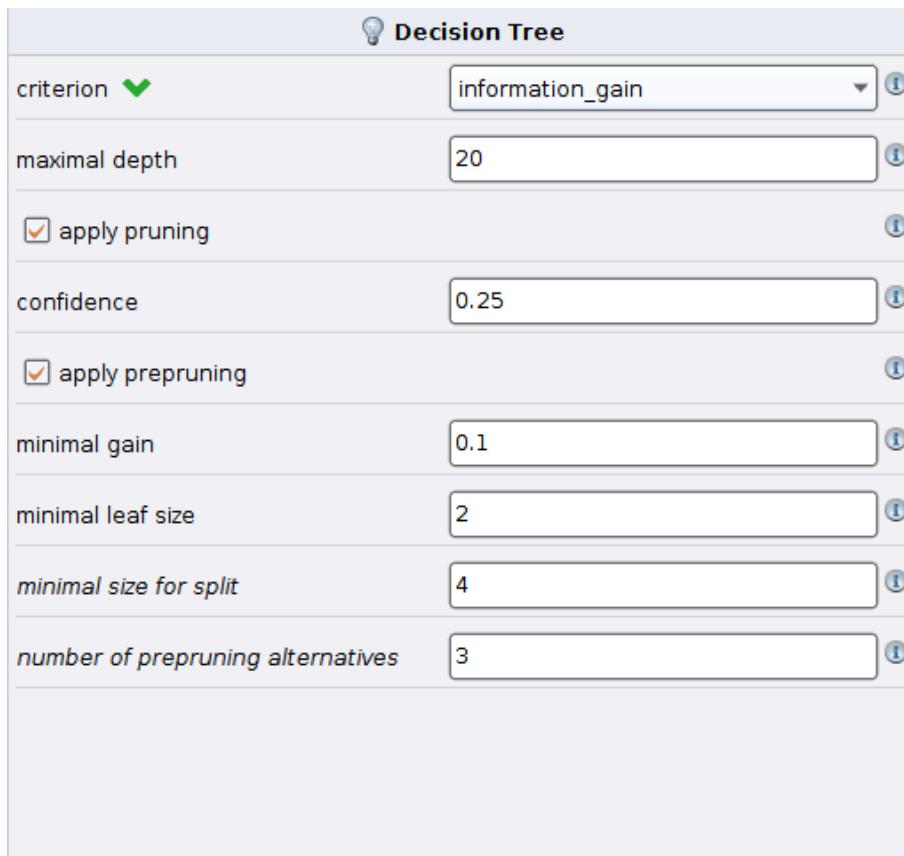


Abbildung 3.1: Konfigurationsmöglichkeiten für Entscheidungsbäume

Werte beinhalten einen Wert (pts closeness), welcher die Stopps in der Nähe von Bushaltestellen und Bahnhöfen widerspiegelt sowie zwei Werte für die Nähe zur Autobahn (highwaycloseness) und zu Gleisen (railcloseness). Wie diese Werte berechnet und ergänzt werden, wird im Abschnitt “5.2 Schlussfolgerungsvariablen“ erklärt.

### 3 Modellbildung und Einbindung der GPS- und GIS-Daten

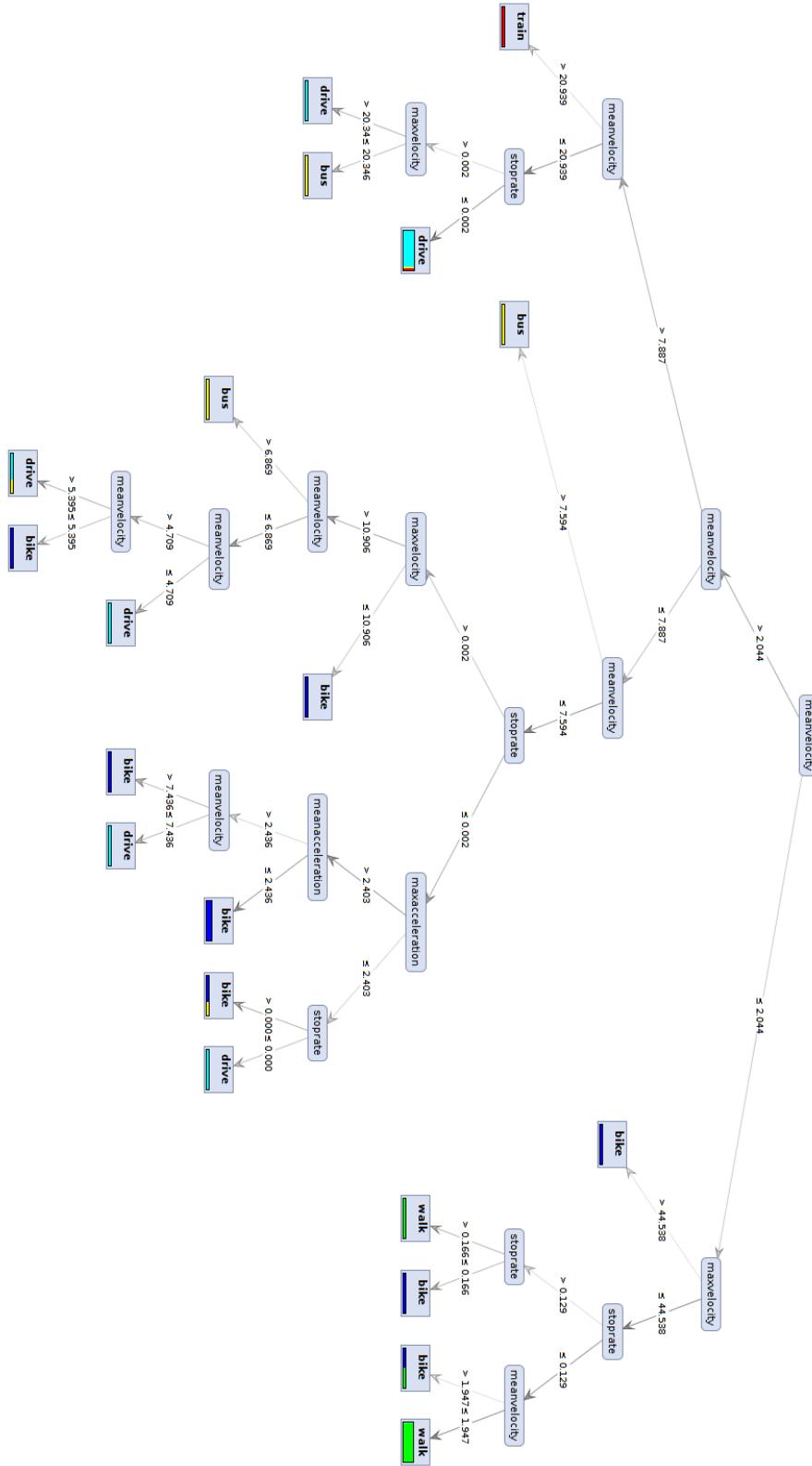


Abbildung 3.2: Entscheidungsbaum ohne GIS-Daten

### 3 Modellbildung und Einbindung der GPS- und GIS-Daten

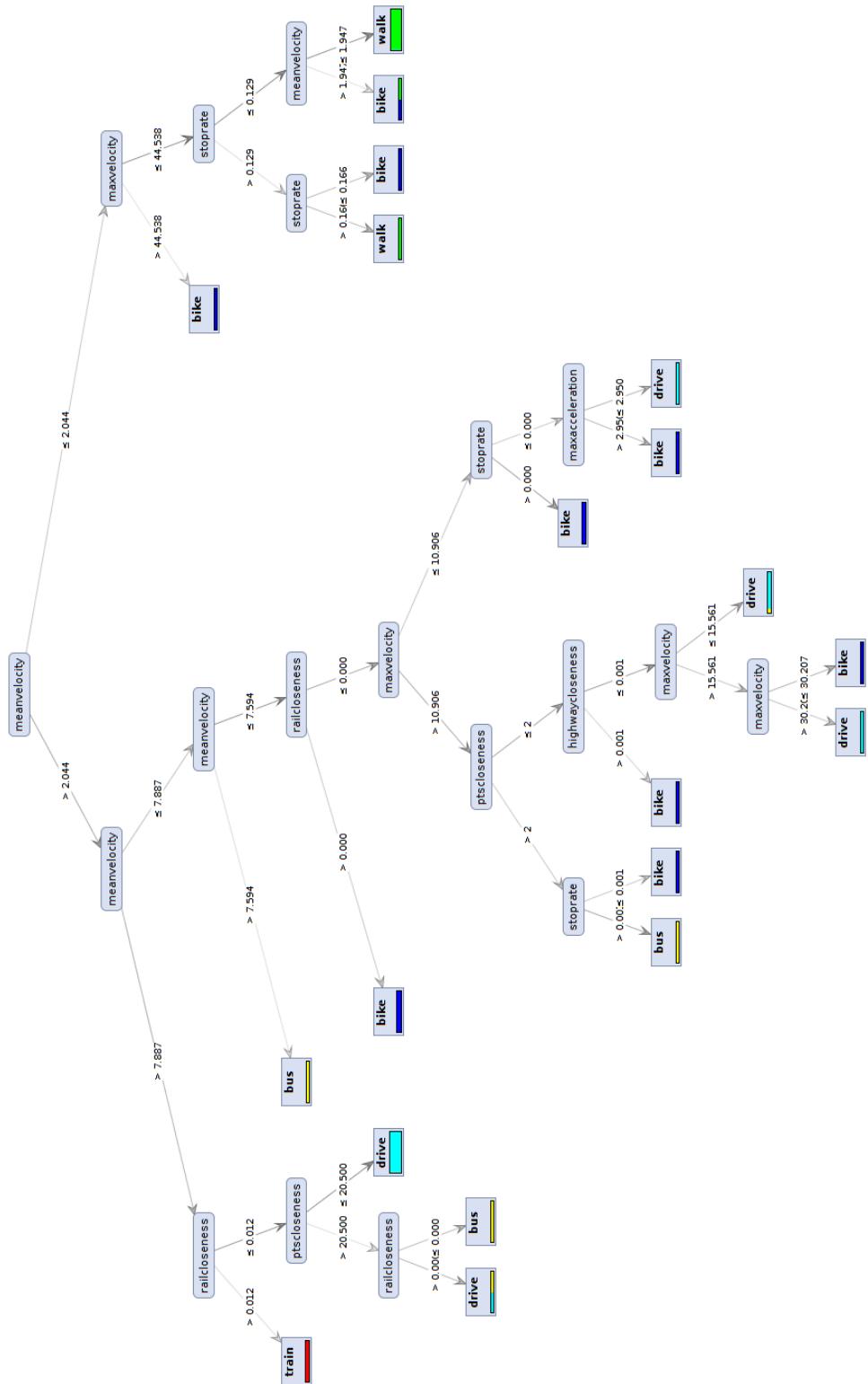


Abbildung 3.3: Entscheidungsbaum mit GIS-Daten

### 3 Modellbildung und Einbindung der GPS- und GIS-Daten

Typ	Anzahl	Distanz (km)
Auto	73	752,33
Fußgänger	53	71,73
Fahrrad	41	1.120,93
Zug	8	239,04
Bus	16	66,03
<b>Gesamt</b>	<b>191</b>	<b>2.250,06</b>

Tabelle 3.1: Trainingsdatenübersicht

## 3.2 Trainingsdaten

Für das Training der Entscheidungsbäume konnten GPS-Aufzeichnungen aus einem Projekt von Sebastian Nagel “Möglichkeitsstudie zum Projekt: Mobilitäts-Tracker“ verwendet werden. Diese Daten wurden mit verschiedenen GPS-Geräten (Wintec WTB-202, Columbus V-900, photoMate 887 Lite, qStarz BT-Q1300, xaiox) und der Hilfe von mehreren Personen aufgezeichnet. [Sebastian Nagel, 2011]

Weiters wurden zum Training auch neue Datensätze verwendet, die mit zwei verschiedenen Smartphones und mit Hilfe der App “MyTrack“ aufgezeichnet wurden. Diese App wurde ausgewählt, da sie sich sehr einfach handhaben lässt und die einzelnen Aufzeichnungen komfortabel exportiert werden können.

Einen groben Überblick über die verwendeten Trainingsdaten bietet die Tabelle 3.1. Die erste Spalte enthält den Verkehrsmitteltyp, die zweite die Anzahl der Segmente und die dritte Spalte enthält die Gesamtdistanz in Kilometern vom jeweiligen Typ. Insgesamt beinhalten die Trainingsdaten 191 Segmente der verschiedenen Transportmittel und erstrecken sich über 2.250 Kilometer.

## Struktur der GPS-Daten

Die Training- und Testsdaten sind in einzelnen Dateien als XML abgelegt und entsprechen dem gängigen GPX-Format, wie es im Listing 3.1 ersichtlich ist. Die Spezifikation für GPX kann auf der Webseite von Topografix unter <http://www.topografix.com/GPX/1/1/> gefunden werden. Ein zugehöriges XML-Schema findet man hier: <http://www.topografix.com/GPX/1/1/gpx.xsd>. [Topografix, 2004]

**Track (trk)** Üblicherweise beginnt eine solche Datei mit einem gpx-Element, welches wiederum einen Track enthält. Ein Track repräsentiert eine Aufzeichnung oder Spur und enthält eine Folge aller aufgezeichneten Trackpoints. Diese sind aber wiederum in ein oder mehrere Tracksegmente gegliedert. [Topografix, 2004]

**Tracksegment (trkseg)** Ein Track kann aus einem oder mehreren Tracksegmenten bestehen. Die Tracksegmente enthalten wiederum beliebig viele aufeinander folgende Trackpoints. Mit diesen Segmenten kann ein Track in logische Abschnitte unterteilt werden. Außerdem kann ein neues Segment begonnen werden, wenn zum Beispiel die Verbindung verloren oder der GPS-Empfänger aus- und wieder eingeschaltet wurde. [Topografix, 2004]

**Trackpoint (trkpt)** Ein Trackpoint entspricht einem Punkt des aufgezeichneten Tracks und enthält eine Koordinate (Längen- und Breitengrad) sowie einen Zeitstempel und die Höhenmeter. Es gibt aber auch Varianten, bei denen Geschwindigkeits- und Beschleunigungswerte zu einem Trackpoint abgelegt werden. [Topografix, 2004]

### 3 Modellbildung und Einbindung der GPS- und GIS-Daten

```
1 <?xml version="1.0" encoding="UTF-8" standalone="yes"?>
2 <gpx xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.topografix.com/GPX/1/1" ...>
3   <metadata>
4     <name>Badgasse – FH</name>
5     <desc></desc>
6   </metadata>
7   <trk>
8     <name>Badgasse – FH</name>
9     <trkseg>
10       <trkpt lat="47.39786" lon="9.735109">
11         <ele>475.0</ele>
12         <time>2015-02-19T07:20:18.156Z</time>
13       </trkpt>
14       ...
15       <trkpt lat="47.405439" lon="9.744841">
16         <ele>492.0</ele>
17         <time>2015-02-19T07:24:35.160Z</time>
18       </trkpt>
19     </trkseg>
20   </trk>
21 </gpx>
```

Listing 3.1: GPX-Datei

### 3 Modellbildung und Einbindung der GPS- und GIS-Daten

## 3.3 Neue Aufzeichnungen

Wie bereits im Abschnitt “3.2 Trainingsdaten“ erwähnt, wurden die neuen Daten mit zwei Smartphones (Samsung Galaxy S und einem LG Nexus 5) und der App “MyTrack“ (siehe Abbildung 3.4) aufgezeichnet. Neben der einfachen Handhabung bietet diese App auch an, nur Punkte mit einer Mindestgenauigkeit aufzuzeichnen, was beim nachfolgenden Filtern ein wenig Arbeit abnimmt. Diese neuen Daten werden hauptsächlich zum Testen verwendet, nur ein Teil davon ist in die Trainingsdaten eingeflossen. Weiters kann man zwar ein Fortbewegungsmittel pro Aufzeichnung angeben, aber dies hat keinerlei Einfluss auf die Aufzeichnung selbst oder die Daten - es dient lediglich der visuellen Darstellung/Unterscheidung der einzelnen Aufzeichnungen.

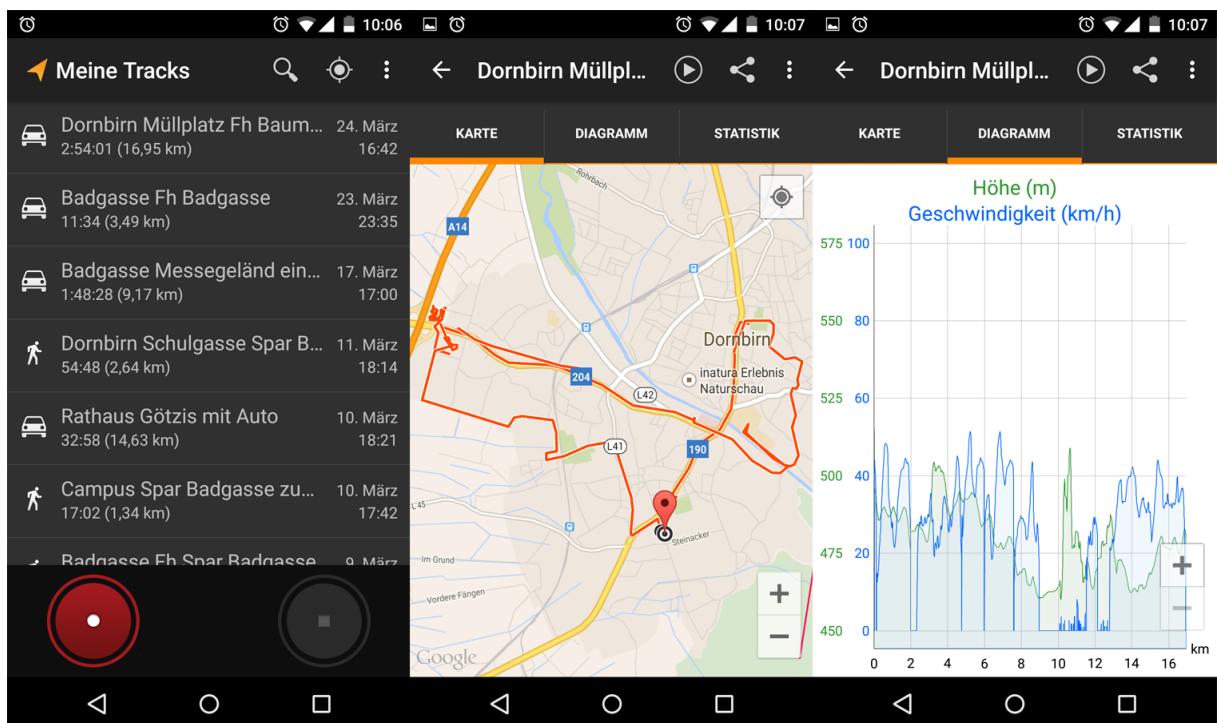


Abbildung 3.4: Die App myTrack

## **3.4 GIS-Daten**

Als relevante GIS-Daten kommen laut den erwähnten Publikationen zum Thema Verkehrsmittelerkennung einige in Frage, wie z.B. Parkplätze, Busstationen, Gleise, Bahnhöfe und das gesamte Straßennetz.

All diese Daten mögen zwar relevant sein, aber es handelt sich auch um sehr viele Daten, was wiederum bedeutet, dass die Bearbeitungszeit einer Aufzeichnung rasch ansteigt. Da der Prototyp auch für konkrete EndbenutzerInnen interessant sein soll, wird auf die Verwendung des gesamten Straßennetzes und der Parkplätze verzichtet, um die Bearbeitungszeit möglichst gering zu halten. Deshalb wird auf die Verwendung von Busstationen und der Gleise, gesetzt da dies schon in der Arbeit von Stenneth [Stenneth et al., 2011] zu guten Resultaten geführt hat. Dieser Ansatz wird ergänzt mit den GIS-Daten des Autobahnnetzes, da in diesen viel Potential vermutet wird und es sich um eine überschaubare Menge an Daten handelt.

### **Akquirierung**

Allgemein sind GIS-Daten z.B. via OpenStreetMaps oder Google-Maps verfügbar, aber in einzelnen Stichproben hat sich herausgestellt, dass die Zusatzinformationen in OpenStreetMaps wesentlich detaillierter und einfacher zu extrahieren sind. Dafür wurde in Kauf genommen, dass diese Daten nicht standardisiert eingetragen wurden.

Die Österreich-Daten von OpenStreetMaps wurden als Archiv heruntergeladen und mit Hilfen von JOSM auf den relevanten Bereich eingegrenzt. JOSM ist ein Tool mit welchem die Daten von OpenStreetMaps gepflegt werden können (zu finden unter <https://josm.openstreetmap.de/>). Nachdem der Bereich auf Vorarlberg eingegrenzt worden ist, konnte dieser mit Hilfe von osmosis (weiteres Tool der OpenStreetMaps-Community <http://wiki.openstreetmap.org/wiki/Osmosis#Downloading>) nach bestimmten Punkten und Verbindungen gefiltert werden. Dadurch war es möglich, das Schienennetz von Vorarlberg sowie die Busstationen und das Autobahnnetz von Vorarlberg als XML-Datei

### *3 Modellbildung und Einbindung der GPS- und GIS-Daten*

zu exportieren. Diese XML-Dateien für Busstationen, das Autobahnnetz und das Schienennetz wurden anschließend eingelesen und in die Datenbank importiert.

In weiterer Folge können diese Daten zur genaueren Bestimmung eines Verkehrsmitteltyps verwendet werden. Dazu werden Werte für die Nähe zu Bushaltestellen sowie Bahnhöfen bei Trackpoints ohne bzw. mit minimaler Bewegung ermittelt. Außerdem wird in regelmäßigen Abständen die Nähe zu den Schienen und der Autobahn für jedes Segment berechnet. Diese Werte fließen dann in den Bestimmungsprozess ein und werden im Abschnitt “5.4 Berechnung der GIS-Entscheidungsvariablen“ genauer beschrieben.

## **3.5 Weitere Daten**

Neben den zusätzlichen Werten, die aus den GPS-Spuren berechnet werden können, und den GIS-Daten, bestand die Möglichkeit, Daten des öffentlichen Personennahverkehrs einzubinden, da diese in Vorarlberg die GPS-Tracks eines jeden Busses enthalten. Die Verwendung dieser Daten wäre insofern vielversprechend gewesen, als dass man einen ähnlichen Ansatz wie Stenneth verfolgen hätte können. Man hätte dadurch überprüfen können, ob an der jeweiligen Stelle gerade ein Bus steht und daraus Rückschlüsse ziehen können. Da diese Daten aber zum Zeitpunkt dieser Arbeit weder für diese Arbeit noch für die Öffentlichkeit verfügbar waren, konnte dieser Ansatz nicht weiter verfolgt werden.

# Der Prototyp

Im Rahmen dieser Thesis wurde ein Prototyp zur Erkennung von Verkehrsmitteln auf Basis von GPS-Daten entwickelt. Im Prinzip sollte jede Person, die ein Gerät zur Aufzeichnung von GPS-Tracks sowie einen Computer mit Internetzugang besitzt, ihre Daten mit diesem Prototypen analysieren können. In Abstimmung mit diesem Ziel wurde eine Webapplikation entwickelt. Diese Applikation verwendet serverseitig das PHP-Framework Symfony als Basis und stellt eine REST-Schnittstelle zur Kommunikation zur Verfügung. Für die Persistierung wurde das Doctrine-ORM in Verbindung mit einer mySQL-Datenbank verwendet.

Der Prototyp bietet neben dem Filtern, Segmentieren und Analysieren der GPS-Daten auch eine Visualisierung der analysierten Daten an. Dies bedeutet, dass das Resultat der Verarbeitung auf einer Karte dargestellt wird. Zusätzlich zur Darstellung der Karte wird dem Benutzer/der Benutzerin dadurch auch die Möglichkeit geboten, die Resultate des Prozesses zu bearbeiten und somit gegebenenfalls das Verkehrsmittel zu korrigieren.

Sowohl der vom System bestimmte als auch der vom Benutzer/von der Benutzerin korrigierte Verkehrsmitteltyp werden in der Datenbank abgelegt, um Aussagen über die Genauigkeit beider Verfahren (mit und ohne GIS-Daten) treffen zu können. Diese Aus-

## *4 Der Prototyp*

wertungen werden wiederum durch verschiedene Diagramme und Statistiken visualisiert.

### **4.1 Aufbau und Architektur**

Der Kern der Webapplikation ist in einer Pipes-and-Filter Architektur aufgebaut. Diese Architektur bietet sich an, wenn man einen Datenstrom verarbeiten will. Dabei werden die einzelnen Schritte der Verarbeitung in sogenannte Filter unterteilt, und diese werden mit Kanälen (Pipes) verbunden. Die Daten fließen somit in einen Filter, werden dort bearbeitet und schlussendlich über die verbundenen Kanäle zu einem oder mehreren nachfolgenden Filtern weitergeleitet. Am Ende einer solchen Kette steht eine Senke (sink), in welcher die Daten dann abgelegt werden (z.B. eine Datei). Durch dieses Architekturmuster wird eine Codegliederung in einzelne Komponenten/Arbeitsschritte in gewisser Weise bereits vorgegeben. Aber es ergeben sich auch andere Vorteile, wie z.B. Flexibilität bezüglich der Datenquelle, Austauschbarkeit von einzelnen Filterimplementierungen oder auch der Repräsentation und Persistierung der Endergebnisse. Weiters müssen bei diesem Ansatz auch keine Zwischendateien o.Ä. geschrieben werden. [Buschmann, 1998]

Diese Flexibilität, welche es ermöglicht, Filter bis zu einem gewissen Grad zu kombinieren und in verschiedenen Reihenfolgen zu verbinden, wurde für den Prototyp benötigt, um z.B. beide Analysemethoden möglichst einfach abbilden zu können. Außerdem werden im Falle dieses Prototyps die Trainingsdaten in eine Datei geschrieben. Bei einer Anfrage über die Webapplikation werden die Daten in der Datenbank abgelegt. Hierbei unterscheiden sich nicht nur die Persistierungsarten sondern auch die Darstellung der Daten selbst, da in die Datei im CSV-Format geschrieben wird. Durch diesen Ansatz konnten viele Komponenten in anderen Benutzungsfällen wiederverwendet werden.

## 4 Der Prototyp

### 4.1.1 Pipes-and-Filter-Architektur für Trainingsdaten

Die Anordnung der einzelnen Filter für die Verarbeitung der GPS-Daten ist im Falle der Trainingsdaten in Abbildung 4.1 ersichtlich. Hierbei wird das Verzeichnis mit den Trainingsdaten definiert und dem ersten Filter (FileReader) werden nach und nach die einzelnen Dateien übergeben. Nachdem die Daten den Trackpoint- und Tracksegmentfilter (grün ohne GIS-Daten, orange mit GIS-Daten) durchlaufen haben, werden sie schließlich von der letzten Komponente (FileWriter) in eine Datei geschrieben.

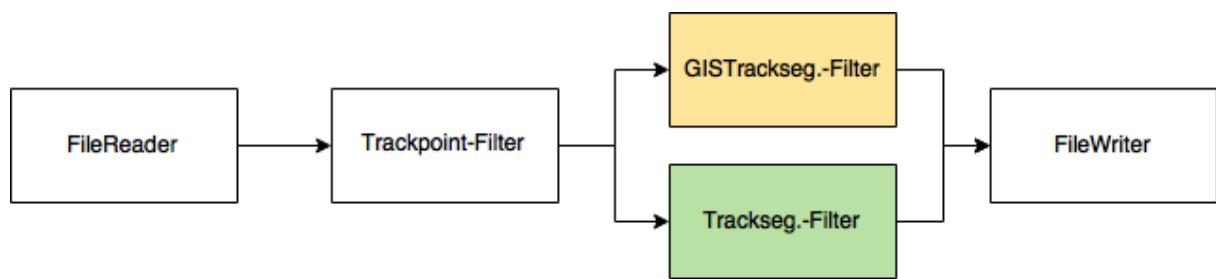


Abbildung 4.1: Pipes- und Filter- Struktur für Trainingsdaten

#### FileReader-Filter

Der FileReader-Filter liest die Datei mit dem angegebenen Dateinamen als XML-Datei ein, sucht sich alle Segmente (trkseg) aus dieser Datei und gibt diese schließlich an den nächsten Filter weiter. Als eingelesene Datei erwartet sich dieser Filter eine XML-Datei im GPX-Format.

#### Trackpoint-Filter

Der Trackpoint-Filter ist nicht nur ein Filter im Sinne des Architekturmusters sondern auch im Sinne seiner Aufgabe. Er versucht all jene Trackpunkte (trkpt) herauszufiltern, die nicht den konfigurierten Grenzwerten entsprechen. Dies bedeutet, dass all jene Trackpunkte gefiltert werden, bei welchen der Zeitabstand, die Geschwindigkeit

## *4 Der Prototyp*

oder die Änderung im Bereich der Höhenmeter unter/über den minimalen/maximalen Grenzwerten liegen. Dadurch werden die einzelnen GPX-Segmente von den gröbsten fehlerhaften Ausreißern bereinigt, bevor sie an den nächsten Filter weitergegeben werden.

### **Tracksegment-Filter**

Der Tracksegment-Filter ist einer der wichtigsten Filter im Verarbeitungsprozess. Er berechnet für jedes Segment die für die weitere Verarbeitung benötigten Zusatzwerte. Dies bedeutet, dass er die mittlere und maximale Beschleunigung, die mittlere und maximale Geschwindigkeit sowie die Stoprate berechnet. Weitere Informationen über die gewählten Zusatzwerte sowie die Auswahl selbst sind im Abschnitt “5.2 Schlussfolgerungsvariablen“ zu finden. Im Falle der Trainingsdaten ergänzt er die Segmentdaten noch mit dem in den Trainingsdaten definierten Verkehrsmitteltyp.

### **GISTracksegment-Filter**

Der GISTracksegment-Filter macht im Wesentlichen dasselbe wie der normale Tracksegment-Filter, allerdings berechnet er zusätzlich zu den Geschwindigkeits- und Beschleunigungswerten noch die Abstände zu den verschiedenen Infrastrukturen wie Busstationen, Schienen und Autobahnen. Diese Werte werden bei der Analyse mit GIS-Daten benötigt.

### **FileWriter**

Der FileWriter übernimmt das Schreiben der übergebenen Daten in eine Datei. Hierbei werden die vom Tracksegment-Filter übergebenen Segmente in einer Datei im CSV-Format abgelegt, welche in weiterer Folge im RapidMiner verwendet wird.

## *4 Der Prototyp*

### **4.1.2 Pipes und Filter-Architektur für die Webapplikation**

Die Anordnung der einzelnen Filter für die Verarbeitung der GPS-Daten, die über die Webapplikation verarbeitet werden sollen, ist in Abbildung 4.2 ersichtlich.

Der Ablauf ist hierbei ähnlich wie jener bei den Trainingsdaten, und es wird wiederum zwischen den zwei Modi mit GIS-Daten (orange) oder ohne GIS-Daten (grün) unterschieden. Ausgehend vom FileReader und Trackpoint-Filter kommen die Daten in den Segmentation-Filter. Von dort kommen sie, je nach Analyse-Art, in den Tracksegment-Filter oder in den GISTracksegment-Filter. Das Ergebnis dieser Filter kommt danach in den TravelModel-Filter, wo die eigentlichen Verkehrsmittel bestimmt werden. Anschließend werden die Ergebnisse dem Postprocess-Filter übergeben, welcher die bestimmten Verkehrsmittel ein letztes Mal überprüft. Nun werden sie an den Database-Filter übergeben, welcher die Ergebnisse für die Datenbank aufbereitet und in dieser ablegt.

#### **Segmentation-Filter**

Der Segmentation-Filter teilt die Tracksegmente der einzelnen Tracks in Teile, welche mit hoher Wahrscheinlichkeit nur mit einem Verkehrsmittel bewältigt wurden. Dazu verwendet dieser ermittelte Geschwindigkeits- sowie Beschleunigungswerte. Dieser Vorgang stützt sich auf die Ergebnisse aus den Publikationen von Zheng [Zheng et al., 2010] und Biljecki [Biljecki et al., 2013]. Der genaue Ablauf dieses Prozesses ist im Abschnitt “5.1 Segmentierung eines Tracks“ beschrieben.

#### **TravelMode-Filter**

In diesem Filter wird der eigentliche Verkehrsmitteltyp anhand der berechneten Werte bestimmt. Dies geschieht in beiden Fällen (mit und ohne GIS-Daten) mit Hilfe des Entscheidungsbaumes (siehe Abschnitt “3.1.3 Einfluss der GIS-Daten auf den Entscheidungsbaum“ und “3.1.3 Einfluss der GIS-Daten auf den Entscheidungsbaum“), welcher

#### 4 Der Prototyp

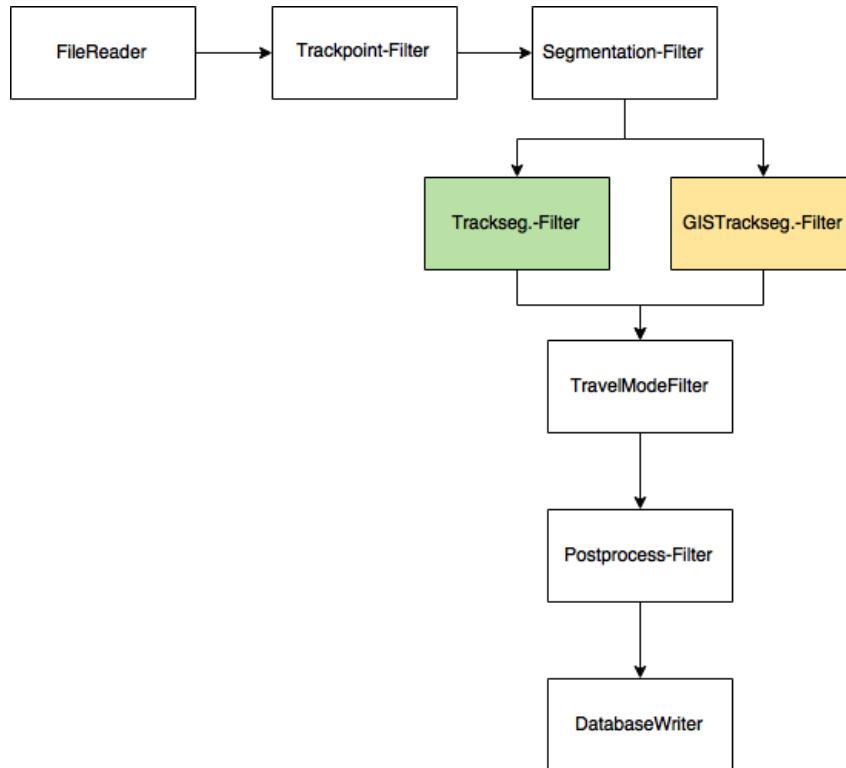


Abbildung 4.2: Pipes- und Filter- Struktur der Webapplikation

mit Hilfe der jeweiligen Trainingsdaten generiert worden ist. Der genaue Ablauf dieses Prozesses ist im Abschnitt “5.5.2 Verwendung der Entscheidungsbäume“ beschrieben.

#### Postprocessing-Filter

Bei diesem Filter geht es darum, dass nach der Bestimmung aller Verkehrsmittel der Segmente eines Tracks, die Plausibilität und Sinnhaftigkeit der Wechsel nochmals überprüft wird. Diese letzte Überprüfung der Resultate der Entscheidungsbäume basiert auf Zheng [Zheng et al., 2010] und soll verhindern, dass sehr unwahrscheinliche Verkehrsmittelwechsels wie zum Beispiel “Auto->Bus->Auto->Bus->Auto“ entstehen. Basierend auf der Aussage von Zheng [Zheng et al., 2010], dass sich zwischen allen Verkehrsmittelwechsel ein (kleines) Segment vom Typ “zu Fuß“ befinden muss, können

## *4 Der Prototyp*

Wechsel wie im obigen Beispiel verhindert werden. Somit wird das Resultat von “Auto->Bus->Auto->Bus->Auto“ zu “Auto->Auto->Auto->Auto->Auto“ korrigiert. Der genaue Ablauf dieses Prozesses ist im Abschnitt “5.5.3 Nachbearbeitung“ beschrieben.

### **DatabaseWriter**

Da das Ablegen der ermittelten Ergebnisse in der Datenbank nur eine Variante (neben dem Ablegen in einer Datei usw.) von vielen ist, kümmert sich diese Komponente darum, dass die übergebenen Daten für die Datenbank aufbereitet und schlussendlich persistiert werden.

## **4.2 Verwendung der Applikation**

Die Oberfläche der Webapplikation ist sehr einfach gehalten und besteht im Wesentlichen aus drei verschiedenen Seiten. Einer Startseite (Home), welche das Projekt kurz vorstellt, einer Seite (Create), auf welcher ein GPS-Track analysiert werden kann, und einer Seite (Results), auf welcher die Resultate aller Auswertungen angezeigt werden.

### **4.2.1 Create-Seite**

Auf dieser Seite wird dem Benutzer/der Benutzerin die Möglichkeit geboten, einen Track hochzuladen und diesen mit einem von den zwei Varianten (mit und ohne GIS-Daten) analysieren zu lassen. Ist dies geschehen, so wird eine Karte angezeigt, auf welcher die ausgelesenen Segmente dargestellt werden. Diese Segmente werden, je nach Verkehrsmitteltyp, in verschiedenen Farben visualisiert. Per Klick auf eines dieser Segmente kann der Verkehrsmitteltyp des Segments korrigiert werden (siehe Abbildung 4.3). Diese Korrektur hat wiederum Einfluss auf die Auswertung der Ergebnisse, welche in Abbildung 4.4 ersichtlich sind.

## 4 Der Prototyp

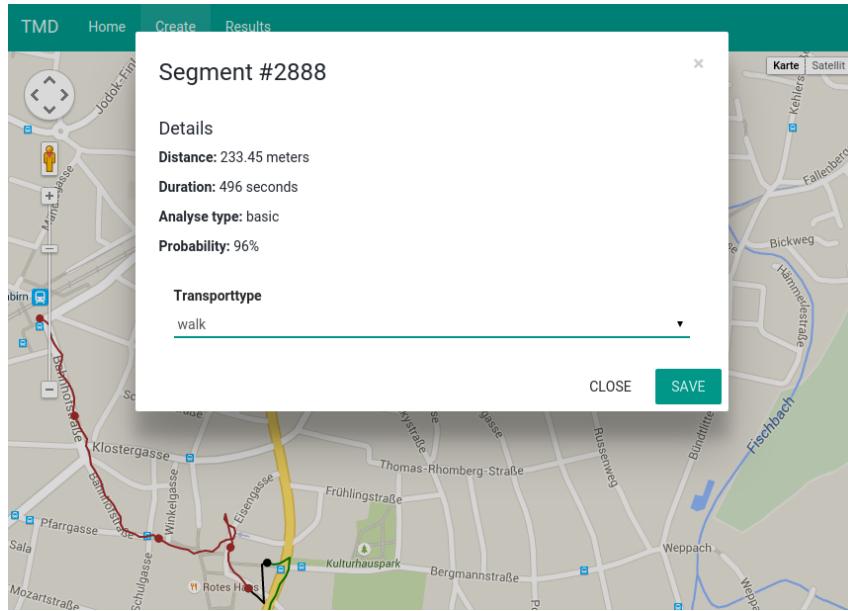


Abbildung 4.3: Korrigieren eines Verkehrsmitteltyps eines analysierten Tracks

### 4.2.2 Results-Seite

Die Resultate aller vorgenommenen Analysen und Änderungen werden auf dieser Seite als Diagramme dargestellt. Dabei zeigt das erste Diagramm das Verhältnis aller analysierten zu den korrekt erkannten Segmenten. Das zweite Diagramm zeigt die Anzahl der korrekt analysierten Segmente und die Gesamtanzahl der Segmente je nach Verkehrsmittel. Alle weiteren Diagramme betreffen ein spezifisches Verkehrsmittel und geben darüber Auskunft, wie oft das jeweilige Verkehrsmittel richtig bzw. nicht richtig erkannt worden ist. Dabei werden die nicht richtig erkannten Resultate nach Verkehrsmitteln sortiert.

Alle Diagramme zeigen dabei die Werte für sämtliche verfügbaren Analysemethoden an, um einen direkten Vergleich zu ermöglichen. Ein Ausschnitt dieser Seite ist in Abbildung 4.4 zu sehen. In dieser Abbildung sind auch die Resultate der Untersuchungen mit verschiedenen Kombinationen der Zusatzvariablen sichtbar (siehe Abschnitt 6 Auswertung).

## 4 Der Prototyp

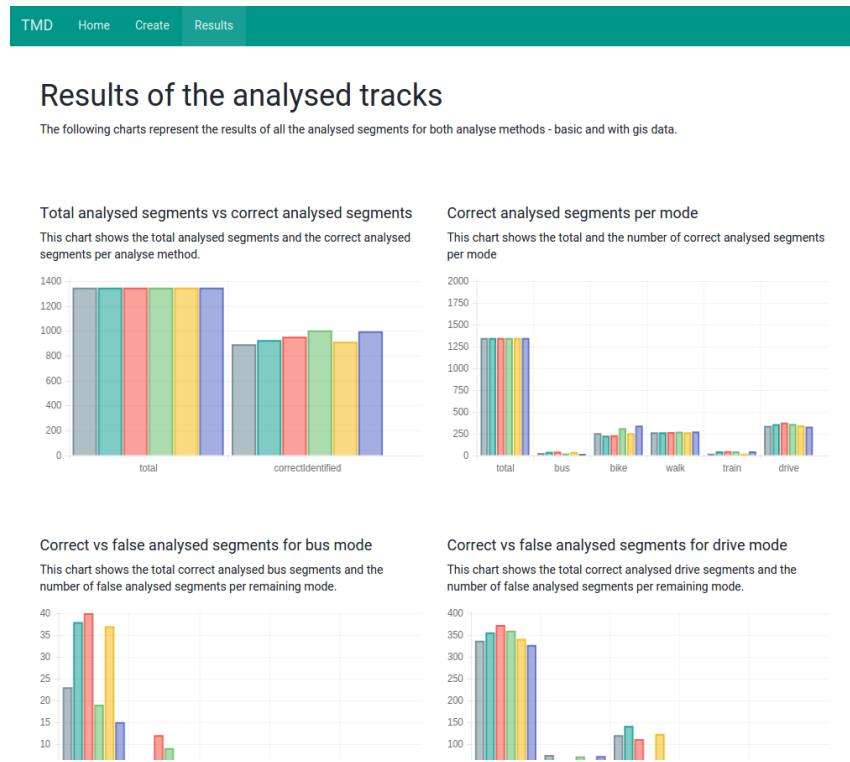


Abbildung 4.4: Ausschnitt zu den Resultaten der analysierten Tracks

## 4.3 Konfiguration des Prototyps

Der Prototyp lässt sich in den verschiedenen Bereichen weitestgehend konfigurieren. Diese Konfiguration kann in einer separaten Datei (app/config/config.yml) vorgenommen werden. Die Konfigurationsmöglichkeiten beinhalten unter anderem:

- Konfiguration des Default-Namespace für das Parsen der GPX-Dateien, wenn keiner in der Datei gefunden wurde.
- Verschiedenste Grenzwerte zum Beeinflussen des Filterns von fehlerhaften Ausreißern in den Trackpunkten.
- Diverse Grenzwerte für den Segmentierungsprozess, welcher die einzelnen Tracks in Segmente mit nur einem Verkehrsmittel unterteilt.

## 4 Der Prototyp

```
1   filter:
2     min_distance: 0.1
3     max_distance: 50
4     max_altitude_change: 25
5     min_trackpoints_per_segment: 2
6     min_time_difference: 2
7     points_to_skip_from_start: 2
8     min_valid_points_ratio: 0
```

Listing 4.1: Filterkonfiguration

- Definitionsmöglichkeiten für alle Analysemethoden sowie deren Eigenschaften.

### 4.3.1 Konfiguration des Filters für fehlerhafte Ausreißer

In diesem Konfigurationsbereich (siehe Listing 4.1) inkludiert sind sowohl eine minimale als auch eine maximale Distanz in Metern pro Zeiteinheit sowie ein Maximalwert für die Änderung der Höhenmeter. Weiters kann hier definiert werden, wie hoch die minimale Anzahl der Trackpunkte pro Segment sein muss, wie groß die minimale Zeitdifferenz (in Sekunden) zwischen zwei Trackpunkten sein soll und wie viele Punkte am Start übersprungen werden sollen. Außerdem kann ein Prozentwert für die Anzahl der validen Trackpunkte im Verhältnis zu allen Trackpunkten eingestellt werden. Eine genauere Erklärung zum Filterprozess ist im Anhang 1 zu finden.

### 4.3.2 Konfiguration des Segmentierens

Im Abschnitt der Segmentierungskonfiguration (siehe Listing 4.2) kann festgelegt werden, bis zu welcher Geschwindigkeit ( $m/s$ ) und Beschleunigung ( $m/s^2$ ) ein Wegpunkt als Geh-Punkt gilt. Weiters kann festgelegt werden, welches die minimale Zeitspanne (in Sekunden) und die minimale Distanz (in Metern) ist, die ein Segment haben muss, um nicht mit dem vorangegangenen Segment vereint zu werden. Schlussendlich gibt

## 4 Der Prototyp

```
1 segmentation:
2     max_walk_velocity: 2.78
3     max_walk_acceleration: 1.5
4
5     min_segment_time: 20
6     min_segment_distance: 50
7     max_time_difference: 30
8
9     max_time_without_movement: 10
10    max_velocity_for_nearly_stoppoints: 0.55
11
12    certain_segments_min_time: 60
13    certain_segments_min_distance: 100
```

Listing 4.2: Segmentierungskonfiguration

es drei Werte, welche zum Beenden eines Segments führen können. Darunter ist sowohl ein Stopp, welcher durch eine sehr geringe oder gar keine Bewegung über eine bestimmte Zeitspanne (in Sekunden) definiert wird als auch eine Zeitspanne (in Sekunden), in welcher keine neuen Trackpunkte gefunden werden. Wofür diese Werte benötigt werden, wird im Abschnitt “5.1 Segmentierung eines Tracks“ detailliert erklärt. Die letzten zwei Konfigurationsvariablen beschreiben die minimale Zeit und Distanz, die ein Segment haben muss, um vom Status eines ungewissen Segments in den Status eines sicheren übergehen zu können.

### 4.3.3 Konfiguration der Analysemethoden

Für die Analyse können verschiedenste Methoden definiert werden. In Listing 4.3 ist die Konfiguration für die Analysemethoden ohne GIS-Daten abgebildet. Hierbei werden unter einem Namen für die Analysemethode (basic) verschiedene Konfigurationsvariablen abgelegt, darunter zum einen Variablen für die Erstellung des Entscheidungsbaums, zu

## 4 Der Prototyp

```
1 basic:
2     class: "BasicDecisionTree"
3     cacheDir: "%kernel.root_dir%/../decisionTrees/basic"
4     txtFilePath: "%kernel.root_dir%/../decisionTrees/basic"
5     txtFileName: "basicDecisionTree.txt"
6     csv_columns:
7         - "stoprate"
8         - "meanvelocity"
9         - "meanacceleration"
10        - "maxvelocity"
11        - "maxacceleration"
```

Listing 4.3: Analysekonfiguration

anderen Variablen für die Generierung der Datei mit den Trainingsdaten.

Die Konfiguration für den Entscheidungsbaum umfasst einen Klassennamen (zugleich auch Name der generierten Datei), das Verzeichnis, in welchem die generierte Entscheidungsbaum-Datei abgelegt werden soll, sowie den Pfad, wo die Textdatei mit der Definition des Entscheidungsbaums gefunden wird. Wozu diese Werte benötigt werden, wird im Abschnitt “5.5.1 Erstellen der Entscheidungsbäume“ genauer erklärt. Hinter der Konfigurationsvariable “csv\_columns“ verstecken sich die Spalten, welche beim Erstellen der Trainingsdatendatei verwendet werden. Diese definieren somit auch, welche Entscheidungsvariablen im Entscheidungsbaum verwendet werden.

### 4.3.4 Zusammenfassung

Der gesamten Analyseprozess mit den Hauptschritten - dem Filtern, dem Segmentieren, dem Bestimmen der Verkehrsmittel und dem Nachbearbeiten - ist in den folgenden Abbildungen ersichtlich. In der Abbildung 4.5 sind die Rohdaten eines GPS-Tracks abgebildet. Am rechten Ende des Tracks sind geradlinige Spitzen zu erkennen, was darauf

#### 4 Der Prototyp

schließen lässt, dass es sich dabei um fehlerhafte Ausreißer handelt, die herausgefiltert werden sollen.



Abbildung 4.5: Rohdaten eines GPS-Tracks

In Abbildung 4.6 sind die bereits gefilterten Daten zu sehen. Zu erkennen ist auch, dass die Spitzen am rechten Ende des GPS-Tracks entfernt bzw. geglättet wurden.

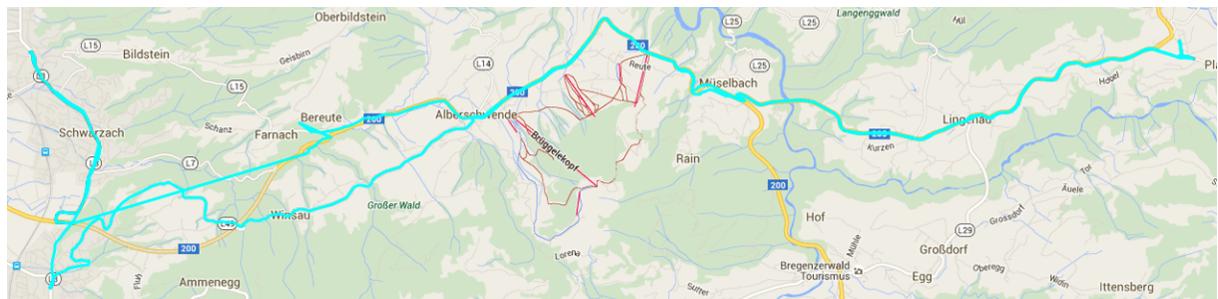


Abbildung 4.6: Gefilterte Rohdaten eines GPS-Tracks

In Abbildung 4.7 wurde der GPS-Track segmentiert und in einzelne Abschnitte aufgeteilt (jeder Kreis symbolisiert den Beginn oder das Ende eines Segments). Dabei sind die roten Abschnitte jene, die zu Fuß bewältigt wurden. Die anderen Segmente sind noch unbestimmt.

In Abbildung 4.8 wurden die Verkehrsmittel der einzelnen Segmente bereits bestimmt und in Abbildung 4.9 wurden diese zum Schluss auf Sinnhaftigkeit und Plausibilität kontrolliert (ein vereinzeltes Fahrradsegment in Grün wurde korrigiert).

#### 4 Der Prototyp



Abbildung 4.7: Segmentierter GPS-Track



Abbildung 4.8: GPS-Track mit bestimmten Verkehrsmitteln

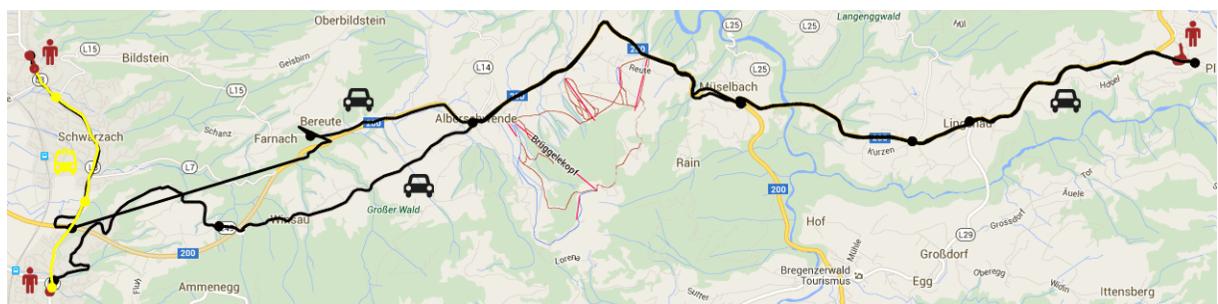


Abbildung 4.9: Nachbearbeiteter GPS-Track

# Segmentierung und Klassifizierung

In diesem Abschnitt wird der Prozess der Klassifizierung von einzelnen Segmenten erklärt. Dies inkludiert den Segmentierungsvorgang sowie die Berechnung der Zusatzinformationen, welche für die Segmentierung benötigt werden. Weiters wird dargelegt, wie die Zusatzvariablen ausgewählt worden sind.

Unter dem Segmentierungsvorgang versteht man das Aufteilen einer GPS-Spur in Abschnitte, in welchen mit hoher Wahrscheinlichkeit nur ein Verkehrsmittel benutzt worden ist. Dabei wird versucht, die Segmente zu finden, in welchen man zu Fuß unterwegs war. Diese werden dann als Geh-Segmente gekennzeichnet. Die anderen Segmente werden als Nicht-Geh-Segmente klassifiziert. Da nur zwischen zwei temporären “Verkehrsmitteln“ unterschieden wird, vereinfacht sich die Verkehrsmittelbestimmung drastisch.

Nach der Segmentierung folgt die konkrete Bestimmung des jeweiligen Verkehrsmittels für die einzelnen Abschnitte, welche aus dem Segmentierungsvorgang hervorgegangen sind. In dieser Arbeit wurde die Bestimmung mit Hilfe von Entscheidungsbäumen realisiert. Dabei gibt es einen Entscheidungsbaum für die Klassifizierung mit und einen für die Klassifizierung ohne GIS-Daten. Für diese Entscheidungsbäume wurden verschiedene Entscheidungsvariablen ausgewählt. Sowohl die Auswahl als auch die Berechnung

## *5 Segmentierung und Klassifizierung*

dieser Werte wird erklärt.

Schlussendlich werden die klassifizierten Abschnitte ein letztes Mal auf ihre Plausibilität in dieser Reihenfolge überprüft. Dabei stützt sich dieser Prozess auf die Aussage von Zheng [Zheng et al., 2010], nach welcher sich immer - wenn mitunter auch kurze - Geh-Segmente zwischen Verkehrswechseln befinden müssen und man nicht von einem in ein anderes Verkehrsmittel wechseln kann, ohne sich eine kurze Strecke zu Fuß bewegen zu müssen. Durch Miteinbeziehen des Kontexts lassen sich Abfolgen wie z.B. "Auto->Bus->Auto->Bus->Auto" verhindern.

### **5.1 Segmentierung eines Tracks**

Bei der Segmentierung geht es in erster Linie darum, den Umfang des Problems der Verkehrsmittelbestimmung zu verkleinern. Man versucht also, das angegebene GPX-Tracksegment in Geh-Segmente und Nicht-Geh-Segmente aufzuteilen. Dadurch muss man statt zwischen mehreren Verkehrsmitteln (Gehen, Bus, Zug, Auto, Fahrrad) nur mehr zwischen zwei unterscheiden. Die genauere Unterscheidung kann dann in einem weiteren Schritt folgen, in welchem jedoch schon klar ist, ob es sich um ein Segment handelt, in welchem eine Person zu Fuß unterwegs war oder nicht.

Zheng sagt in diesem Zusammenhang, dass es zwischen jedem Wechsel eines Verkehrsmittels einen Abschnitt gibt, in welchem man zu Fuß unterwegs war und ein Stopp stattgefunden hat, auch wenn dieser Abschnitt sehr klein war. Ein Wechsel erfolgt nie direkt, wie z.B. von einem Zug in den Bus ohne anzuhalten und über einen Bahnsteig gehen zu müssen. Wenn also, ein Verkehrsmittelwechsel stattgefunden hat, dann gibt es neben Geh-Punkten auch Trackpunkte, in welchen eine Geschwindigkeit und eine Beschleunigung von nahezu 0 zu erwarten ist. Diese Aussage hat sich aus seinen umfangreichen Testdaten ableiten lassen. Daraus hat er folgenden Algorithmus abgeleitet: [Zheng et al., 2010]

## *5 Segmentierung und Klassifizierung*

- Finde alle Geh-Punkte und Nicht-Geh-Punkte des betrachteten Abschnitts oder Tracks anhand von Grenzwerten für Geschwindigkeit und Beschleunigung
- Fasse die aufeinander folgenden Punkte vom selben Typ in Segmente zusammen.
- Liegt die Distanz oder die Zeit eines solchen Segments unterhalb einer definierten Grenze, so vereine dieses Segment mit dem vorherigen.
- Überschreitet ein Segment eine bestimmte Länge (200 Meter), ist es ein “sicheres Segment“. Liegt die Distanz eines Segments jedoch unterhalb dieses Werts, so ist es ein “unsicheres Segment“. Überschreitet die Anzahl der aufeinander folgenden “unsicheren Segmente“ einen bestimmten Grenzwert, so werden die aufeinander folgenden “unsicheren Segmente“ vereint und als Nicht-Geh-Segment betrachtet.

Biljecki ergänzt diesen Ansatz von Zheng mit seinen Erfahrungen, in welchen er feststellte, dass bei einem Wechsel des Verkehrsmittels oft auch das Signal verloren geht. Aus diesem Grund beendet Biljecki ein Segment, wenn die Verbindung verloren wurde und startet ein neues. Den Verbindungsverlust interpretiert Biljecki so, dass er für mindestens 30 Sekunden keinen weiteren Trackpoint findet. Weiters segmentiert Biljecki dann, wenn er für mehr als 12 Sekunden keine bzw. wenig Bewegung ( $< 2\text{km/h}$ ) feststellen konnte. Diese beiden Änderungen am Algorithmus von Zheng begründet Biljecki damit, dass eine Übersegmentierung besser ist als eine Untersegmentierung. Aufeinander folgende Segmente vom Typ kann man immer noch in einem nächsten Schritt zusammenlegen. [Biljecki et al., 2013]

Für den Prototyp wurde im Wesentlichen der durch Biljecki erweiterte Ansatz von Zheng gewählt. Vorallem das Einteilen der Segmente in “sichere“ und “unsichere“ Segmente hat sich bei der Bestimmung des Verkehrsmittels als essentiell herausgestellt. Liegt die Distanz des Segments unterhalb einer gewissen Grenze, so tendiert die Klassifizierung dazu, falsche Entscheidungen zu treffen. Speziell im Kontext mit GIS-Daten konnte man feststellen, dass diese ihre Wirkung erst ab einer bestimmten Distanz entfalten konnten (z.B. mehrere kurze Segmente mit jeweils einem Stopp bei einer Bushaltestelle im Ge-

## 5 Segmentierung und Klassifizierung

gensatz zu einem längeren Segment mit vielen Stopps bei Bushaltestellen).

Das Resultat eines Segmentierungsvorgangs kann man in Abbildung 5.1 sehen. Jeder Kreis auf der eingezeichneten Route zeigt den Start bzw. das Ende eines Segments. Auf Basis dieser ersten Aufteilung kann nun die genauere Bestimmung des Verkehrsmittels erfolgen.

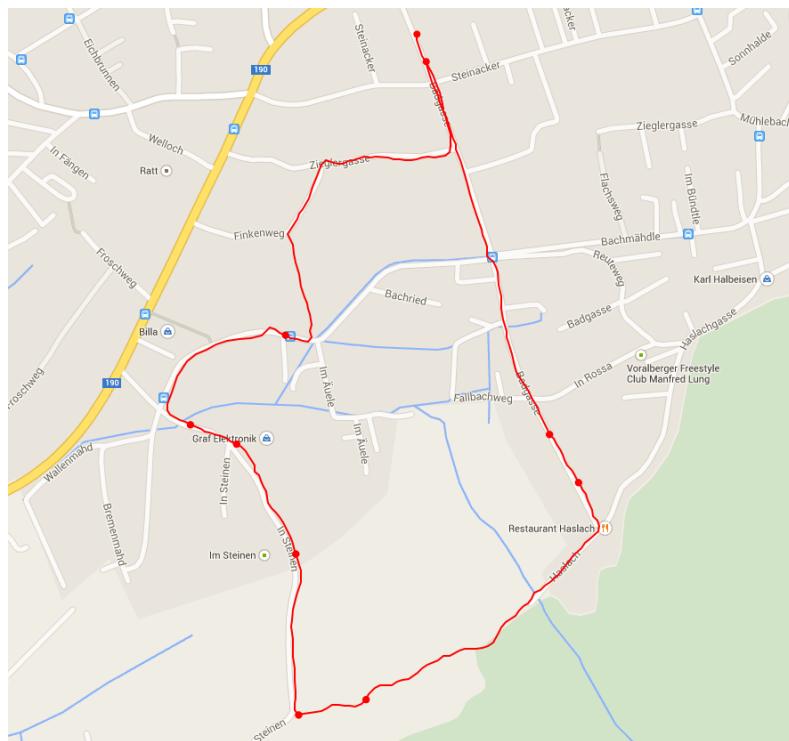


Abbildung 5.1: Grundlegende Segmentierung

## 5.2 Schlussfolgerungsvariablen

Als Grundlage für die Schlussfolgerung durch die Entscheidungsbäume müssen zuerst Variablen festgelegt werden, die möglichst ausschlaggebend für die betrachteten Verkehrsmittel sind. Diese Schlussfolgerungs- oder auch Entscheidungsvariablen sind ein

## *5 Segmentierung und Klassifizierung*

Teil einer Bedingung in einem Knoten eines Entscheidungsbaums. Eine Bedingung besteht dabei aus der Variable (z.B. Stopprate), einem Operator ( $\leq$  oder  $\geq$ ) sowie einem Wert. Wird ein Baum mit einem Segment traversiert, so muss an jedem Knoten eine Entscheidung gefällt werden, welche dann vorgibt, ob mit dem linken oder rechten Kind fortgefahrene werden soll. Dies wird gemacht, bis ein Blatt im Baum erreicht wird. Ein Blatt enthält das Resultat für die getroffenen Entscheidungen und gibt Auskunft darüber, zu wie viel Prozent Wahrscheinlichkeit ein Segment von einem bestimmten Transporttyp ist.

Diese Variablen können dabei diverse geschwindigkeitsabhängige Werte sein, wie durchschnittliche und maximale Geschwindigkeit oder auch Beschleunigungswerte und Abstände zu bestimmten Infrastrukturen. Eine Übersicht über die Schlussfolgerungsvariablen in den betrachteten Publikationen ist in der Tabelle 5.1 abgebildet. Diese Werte sind nach der Anzahl der Vorkommnisse in den Publikationen gereiht und bilden eine Grundlage für die in dieser Arbeit verwendeten Variablen.

Aus Gründen der Vollständigkeit muss noch erwähnt werden, dass aufgrund eines anderen Segmentierungsverfahrens nicht alle Variablen von Gonzales [Gonzalez et al., 2010] aufgeführt sind. Außerdem handelt es sich bei der maximalen Geschwindigkeit nicht um das absolute Maximum, sondern um 95% davon bzw. die dritthöchste Geschwindigkeit, die gemessen wurde.

Weiters muss noch erwähnt werden, dass sowohl Zheng [Zheng et al., 2010] als auch Stenneth [Stenneth et al., 2011] nicht alle Variablen schlussendlich verwendet haben. Sie haben allerdings evaluiert, welche Variablen in ihrem Projekt die größte Wirkung zeigen. Dies war bei Zheng [Zheng et al., 2010] die Kombination der Stopprate mit der Geschwindigkeitsänderungsrate und der Richtungsänderungsrate. Fügte er weitere Variablen hinzu, konnte er eine Verschlechterung der Ergebnisse beobachten. Bei Stenneth [Stenneth et al., 2011] hat die Evaluierung ergeben, dass die durchschnittliche Geschwindigkeit und Beschleunigung kombiniert mit verschiedenen GIS-Werten

## *5 Segmentierung und Klassifizierung*

das beste Ergebnis liefert.

Die betrachteten GIS-Werte waren bei Stenneth [Stenneth et al., 2011] der durchschnittliche Abstand zu Gleisen, Bussen und dem Buskandidaten (jenen Bus, in dem sich die Person am ehesten befand). Biljecki verwendete den Abstand zu Gleisen, Bushaltestellen, Buslinien, U-Bahn und Straßenbahn als Indikatoren für die Schlussfolgerung. Schüssler [Nadine Schüssler et al., 2011] inkludiert von den möglichen GIS-Werten nur den Abstand zu öffentlichen Verkehrsmitteln aller Art.

### **5.2.1 Reihung der allgemeinen Variablen**

Die in dieser Arbeit verwendeten Schlussfolgerungsvariablen basieren auf der Reihung, welche in Tabelle 5.1 ersichtlich ist. Die allgemeine Geschwindigkeit an fünfter Stelle wurde deshalb übersprungen, weil sie bereits zwei mal vertreten ist und bereits Zheng gesagt hat, dass die Geschwindigkeit allein kein aussagekräftiger Indikator für ein Verkehrsmittel ist [Zheng et al., 2010]. Weiters hat Zheng für die Auswahl der Stopprate eine interessante Erklärung, welche im Abschnitt “5.3.3 Stopprate“ genauer erklärt wird. Deshalb wurde die Stopprate auch als fünfte Variable ausgewählt.

1. durchschnittliche Geschwindigkeit
2. maximale Geschwindigkeit
3. durchschnittliche Beschleunigung
4. maximale Beschleunigung
5. Stopprate

### **5.2.2 Reihung der GIS-Variablen**

Bei der Auswahl der GIS-Variablen fallen jene mit U-Bahn und Straßenbahn weg, da es diese im Raum Vorarlberg nicht gibt. Jene Variablen, die spezifische GPS-Daten von

## *5 Segmentierung und Klassifizierung*

einzelnen Verkehrsmitteln benötigen, konnten aufgrund von fehlenden Schnittstellen zu den Daten der öffentlichen Verkehrsbetriebe nicht verwendet werden. Somit wurden folgende Variablen für die GIS-gestützte Analyse verwendet:

- durchschnittliche Nähe zu Busstationen und Bahnhöfen
- durchschnittliche Nähe zu Gleisen
- durchschnittliche Nähe zu Autobahnen

	Zheng <sup>1</sup>	Stenneth <sup>2</sup>	Reddy <sup>3</sup>	Biljecki <sup>4</sup>	Gonzales <sup>5</sup>	Schüssler <sup>6</sup>	<b>Gesamt</b>
durchschn. Geschwindigkeit	x	x		x	x	x	<b>5</b>
max. Geschwindigkeit *	x			x	x		<b>3</b>
durchschn. Beschleunigung		x	x		x		<b>3</b>
verwendet GIS Daten		x		x		x	<b>3</b>
Geschwindigkeit			x			x	<b>2</b>
max. Beschleunigung	x				x		<b>2</b>
Richtungswechselrate	x	x					<b>2</b>
Stopprate	x						<b>1</b>
Distanz des Segments	x						<b>1</b>
Distanz des Tracks					x		<b>1</b>
Geschwindigkeitswechselrate	x						<b>1</b>
durchschn. Varianz d. Beschl.						x	<b>1</b>
durchschn. beweg. Geschw.				x			<b>1</b>
durchschn. Genauigkeit		x					<b>1</b>
erwartete Geschwindigkeit	x						<b>1</b>
Varianz d. Geschwindigkeit	x						<b>1</b>

Tabelle 5.1: Entscheidungsvariablenübersicht

<sup>1</sup> [Zheng et al., 2010], <sup>2</sup> [Stenneth et al., 2011], <sup>3</sup> [Reddy et al., 2010], <sup>4</sup> [Biljecki et al., 2013], <sup>5</sup> [Gonzalez et al., 2010], <sup>6</sup> [Nadine Schüssler et al., 2011]

## 5.3 Berechnung der allgemeinen Entscheidungsvariablen

Das Hinzufügen und Berechnen der Entscheidungsvariablen ohne GIS-Daten (z.B. Geschwindigkeit aus Weg und Zeit), wird vom Tracksegment-Filter (siehe Abschnitt “4.1.1 Tracksegment-Filter“) realisiert. Dabei ist die Berechnung der grundlegenden Geschwindigkeit zwischen zwei GPS-Punkten essentiell, da alle weiteren Entscheidungsvariablen darauf aufbauen.

### 5.3.1 Geschwindigkeit

Berechnet wurde die Geschwindigkeit ( $v$ ) als Durchschnittsgeschwindigkeit, die als Verhältnis vom zurückgelegten Weg ( $s$ ) zur Zeit ( $t$ ) ausgedrückt wird, wie es z.B. im Buch “Physik“ von Douglas Giancoli definiert wird [Douglas C. Giancoli, 2010, S. 27] und in 5.1 ersichtlich ist.

$$v = \frac{s}{t} \quad (5.1)$$

Die Zeit ist dabei die Differenz zwischen den Zeitstempeln von zwei GPS-Trackpunkten und der Weg die Differenz zwischen den zwei Koordinaten. Zur Berechnung des Weges (der Luftlinienentfernung) zwischen zwei Koordinaten gibt es laut [Movable Type Ltd, 2015] mehrere Möglichkeiten. Welche man verwendet, hängt von den Längen der betrachteten Strecken, der gewünschten Genauigkeit sowie der benötigten Performanz ab (schnell, aber ungenau vs. langsam, aber genau). Da es sich bei den hier betrachteten Distanzen immer um sehr kleine Distanzen im Verhältnis zur Erde selbst handelt, aber diese aufgrund ihrer weiteren Verwendung zur Berechnung der Geschwindigkeit möglichst genau sein soll, wird in diesem Prototyp die Haversine-Formel (siehe die Gleichungen 5.2) verwendet. Dabei sind  $\varphi$  die Breitengrade,  $\lambda$  die Längengrade,  $R$  der Erdradius und  $\Delta\varphi$  bzw.  $\Delta\lambda$  die Differenz der Breiten- bzw. Längengrade. [Movable Type Ltd, 2015]

## 5 Segmentierung und Klassifizierung

$$\begin{aligned} a &= \sin^2(\Delta\varphi/2) + \cos\varphi_1 * \cos\varphi_2 * \sin^2(\Delta\lambda/2) \\ c &= 2 * \text{atan}2(\sqrt{a}, \sqrt{(1-a)}) \\ d &= R * c \end{aligned} \tag{5.2}$$

**Durchschnittliche Geschwindigkeit** Die durchschnittliche Geschwindigkeit wird aus allen Geschwindigkeitswerten eines Segments berechnet.

**Maximale Geschwindigkeit** Die maximale Geschwindigkeit wird aus allen Geschwindigkeitswerten des Segments ermittelt.

### 5.3.2 Beschleunigung

Die Beschleunigung ist als Geschwindigkeitsunterschied zwischen zwei Punkten pro Zeiteinheit definiert [Douglas C. Giancoli, 2010, S. 51]. Somit wurde diese auf Basis der ermittelten Geschwindigkeit berechnet und dadurch wurden folgende zwei Entscheidungsvariablen bestimmt:

- **Durchschnittliche Beschleunigung:** die durchschnittliche Beschleunigung für das betrachtete Segment.
- **Maximale Beschleunigung:** die maximal gemessene Beschleunigung für das betrachtete Segment.

### 5.3.3 Stoprate

Wie Zheng in der Publikation [Zheng et al., 2010] beschrieben hat, ist die Geschwindigkeit als Basis für Entscheidungsvariablen nur bedingt geeignet, da diese sehr von der aktuellen Verkehrssituation abhängig ist. Deshalb setzt Zheng auf die Kombination von Richtungswchsel, Geschwindigkeitswechsel und der Stoprate. Dieser Aussage

## 5 Segmentierung und Klassifizierung

widerspricht wiederum indirekt die Auswahl der Entscheidungsvariablen von anderen Autoren wie Biljecki [Biljecki et al., 2013], Gonzales [Gonzalez et al., 2010], Schüssler [Nadine Schüssler et al., 2011] und Reddy [Reddy et al., 2010] (siehe Tabelle 5.1).

Die Stopprate ist für Zheng insofern sehr aussagekräftig, da er an spezifischen Verkehrsmitteln auch eine spezifische Stopprate bzw. ein Stoppverhalten festmachen kann. Dies ist beispielhaft in Abbildung 5.2 ersichtlich. Dabei stoppt eine Person im Auto (a) wesentlich weniger oft wie zum Beispiel ein Bus (b), da dieser nicht nur bei Kreuzungen sondern auch bei Bushaltestellen anhalten muss. Noch öfter stoppt laut Zheng nur ein Fußgänger (c) [Zheng et al., 2010]. Im Falle des Prototyps wird die Erkennung von Stopps bereits beim Segmentieren benötigt. Da die anderen Entscheidungsvariablen auf der Geschwindigkeit basieren, wurde die Stopprate als weitere Entscheidungsvariable, die nicht abhängig von der Geschwindigkeit ist, ausgewählt. Als Stopp gilt dabei eine bestimmte Anzahl von Trackpunkten (z.B. über eine Zeit von 5 Sekunden), bei welchen die Geschwindigkeit unterhalb eines Grenzwerts (z.B. unter 2km/h) liegt. Diese Werte können wiederum in der Konfiguration eingestellt werden.

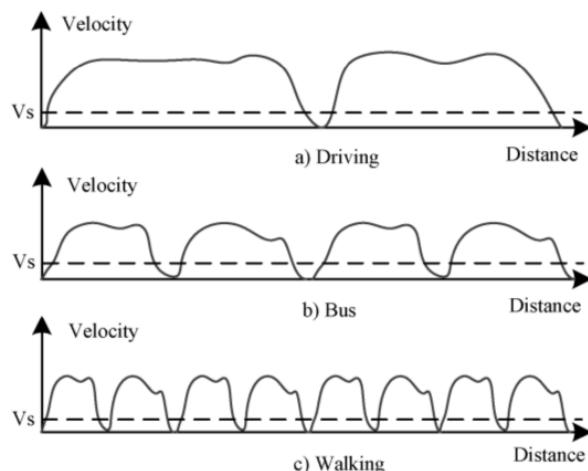


Abbildung 5.2: Stopprate laut Zheng (Quelle: [Zheng et al., 2010])

## 5.4 Berechnung der GIS-Entscheidungsvariablen

Das Hinzufügen und Berechnen der Entscheidungsvariablen mit GIS-Daten wird vom GISTracksegment-Filter (siehe Abschnitt “4.1.1 GISTracksegment-Filter“) realisiert. Dabei erweitert der GISTracksegment-Filter den Tracksegment-Filter und ergänzt diesen mit den in Abschnitt “5.4.1 Abstand zu Bushaltestellen“ und in Abschnitt “5.4.2 Abstand zu Gleisen und zur Autobahn“ beschriebenen GIS-Werten. Bei allen GIS-Werten wird eine Bounding-Box berechnet und auf die zuvor in die Datenbank importierten GIS-Daten zugegriffen. Der Radius der Boundingbox lässt sich in der Konfiguration festlegen.

### 5.4.1 Abstand zu Bushaltestellen

Da der Beginn und das Ende eines Segments immer ein Stopp ist, wird hierbei auch überprüft, ob sich eine Bushaltestelle in der Nähe befindet. Wird eine Bushaltestelle gefunden, so wird diese in den GIS-Wert für die Bestimmung des Verkehrsmittels mit-einbezogen. Ist sowohl der Beginn als auch das Ende eines Segments in der Nähe einer Bushaltestelle, so wird dies höher gewichtet. Innerhalb eines Segments können sich weitere kürzere Stopps befinden. Diese werden zusätzlich verwendet, um festzustellen, ob Bushaltestellen in der Nähe sind und es sich im Endeffekt um einen Bus handelt der von Station zu Station fährt. Im Gegensatz zu den Haltestellen am Beginn und am Ende eines Segments werden diese allerdings nicht höher gewichtet, da sich Bushaltestellen oft in der Nähe von Ampeln oder Kreuzungen befinden. Dies stellte auch Biljecki [Biljecki et al., 2013] schon fest.

### 5.4.2 Abstand zu Gleisen und zur Autobahn

Ähnlich wie bei den Bushaltestellen werden auch die GIS-Werte für die Nähe zu Gleisen und der Autobahn am Beginn und am Ende eines jeden Segments berechnet. Allerdings

## *5 Segmentierung und Klassifizierung*

wird für die Überprüfung innerhalb eines Segments ein konfigurierbarer Zeitwert verwendet. Im Prototyp wird alle 20 Sekunden überprüft ob sich ein Punkt in der Nähe eines Gleises oder der Autobahn befindet.

## **5.5 Klassifizierung der Verkehrsmittel**

Die Klassifizierung der Segmente bzw. das Bestimmen des Verkehrsmittels für ein Segment basiert bei beiden Analysearten (mit und ohne GIS-Daten) auf Entscheidungsbäumen. Wie bereits in Abbildung 3.2 und in Abbildung 3.3 ersichtlich und im Abschnitt “5.2 Schlussfolgerungsvariablen“ genau beschrieben, verwendet der Entscheidungsbau ohne GIS-Daten ausschließlich Geschwindigkeits- und Beschleunigungs- werte sowie die Stoprate als Indikatoren. Der Entscheidungsbau mit GIS-Daten fügt diesen aber noch einen Wert für die Nähe zu verschiedenen Infrastrukturen hinzu.

### **5.5.1 Erstellen der Entscheidungsbäume**

Wie bereits im Abschnitt “3.1 Entscheidungsbau als Schlussfolgerungsmodell“ beschrieben, konnte in der freien Version von RapidMiner der generierte Entscheidungsbau als Bild oder Text exportiert werden. Um jedoch beim Erstellen des Entscheidungsbau möglichst flexibel zu sein (z.B. Erweitern der Trainingsdaten), wurde ein Parser für die textuelle Darstellung des Entscheidungsbau implementiert.

Ein Ausschnitt des Entscheidungsbau als Text ist in Listing 5.1 zu sehen. Dabei stellt jede Zeile mindestens einen Knoten im Baum dar. Eine Zeile kann mit Indikatoren für die Tiefe des Knotens im Baum beginnen. Dabei steht jedes “|“ für eine Ebene tiefer im Baum. Jene zwei Zeilen am Beginn ohne Tiefenangabe bilden somit den Wurzelknoten, da sie über keine Tiefenangabe verfügen. Außerdem gibt es jeden Knoten, der kein Blatt ist, zwei mal in dieser Darstellung. Diese beiden Knoten unterscheiden sich dabei

## 5 Segmentierung und Klassifizierung

```
1 meanvelocity > 20.830: train {bike=0, walk=0, car=0, bus=0, train=5}
2 meanvelocity <= 20.830
3 |   meanvelocity > 2.041
4 |   |   meanvelocity > 7.837
5 |   |   |   meanvelocity > 8.772: car {bike=0, walk=0, car=46, bus=1,
6 |   |   |   train=2}
7 |   |   |   |   meanacceleration > 2.728: bus {bike=0, walk=0, car=0,
8 |   |   |   |   bus=2, train=0}
9 |   |   |   |   meanacceleration <= 2.728: car {bike=0, walk=0, car=3,
10 |   |   |   |   bus=0, train=0}
11 ...
12 ...
```

Listing 5.1: Entscheidungsbaum in Textform

nur durch ihren Vergleichsoperator, denn dieser ist das genaue Gegenteil des anderen. Dadurch können beide Fälle einer Bedingung abgebildet werden.

Jede Zeile besteht aus dem Namen der Entscheidungsvariable, einem Vergleichsoperator und einem Wert. Hat ein Knoten nur mehr einen Kind-Knoten mit dem Resultat, so folgt hinter dem Wert das Resultat. Das Resultat besteht dabei aus dem bestimmten Verkehrsmittel und einer Übersicht über die Vorkommnisse aller Verkehrsmittel für die getroffenen Entscheidungen aus den Trainingsdaten.

Alle Kind-Knoten sind jeweils dem letzten Knoten auf dem vorherigen Level zuzuordnen.

### Parsen und Cachen der Entscheidungsbäume

Die von RapidMiner gelieferten Entscheidungsbäume wurden im Prototyp als Teil der Konfiguration hinterlegt. Damit die Entscheidungsbäume nicht für jede Anfrage geparsst und erstellt werden müssen, bietet das Framework Symfony die Möglichkeit, die aus Konfigurationsdateien entstandenen Resultate zu cachen. Das bedeutet in diesem Fall, dass die Entscheidungsbäume eingelesen sowie geparsst werden und danach mit Hilfe

## 5 Segmentierung und Klassifizierung

```
1 ...
2 class BasicDecisionTree implements DecisionTreeInterface
3 {
4     protected $tree;
5
6     function __construct()
7     {
8         $node0 = new Node();
9         $node0->setDecision(new Decision('meanvelocity', '>', 20.83));
10        $node1 = new Node();
11        $node1->setResult(new Result(0,0,0,0,5));
12        ...
13        $node0->setRight($node2);
14        $node1->setParent($node1);
15        $node2->setParent($node0);
16        $node2->setLeft($node3);
17        ...
```

Listing 5.2: Ausschnitt des generierten Entscheidungsbaums als PHP-Klasse

des Resultats eine PHP-Datei erstellt wird. Darin wird die Initialisierung des Entscheidungsbaums als Code abgelegt (siehe Listing 5.2). Dadurch muss der Entscheidungsbaum nicht jedes Mal neu geparsst, sondern nur ein Objekt mit genau diesem Baum instanziert werden.

Damit dieser Vorgang nur gemacht wird, wenn sich die Datei mit dem Entscheidungsbaum in Textform verändert, merkt Symfony sich das Änderungsdatum und entscheidet basierend darauf, ob die PHP-Datei neu generiert werden muss.

### 5.5.2 Verwendung der Entscheidungsbäume

Die jeweiligen Entscheidungsbäume werden basierend auf der Analysemethode im Travel-Mode-Filter verwendet. Dabei wird für jedes Segment der Entscheidungsbaum traversiert.

## *5 Segmentierung und Klassifizierung*

siert und das Resultat beim jeweiligen Segment hinterlegt. Schlussendlich hat jedes Segment einen Verkehrsmitteltyp zugeordnet bekommen und wird zur Nachbearbeitung weitergegeben.

### **5.5.3 Nachbearbeitung**

Die Nachbearbeitung der Segmente mit dem bestimmten Verkehrsmittel wurde sowohl von Zheng [Zheng et al., 2010] als auch Biljecki [Biljecki et al., 2013] vorgeschlagen bzw. beschrieben. Dabei geht es darum, die klassifizierten Segmente auch im Kontext zu sehen. So kann laut Zheng [Zheng et al., 2010] zum Beispiel nicht ein “Auto“-Segment auf ein “Bus“-Segment folgen, ohne dass sich zwischen ihnen ein Segment “zu Fuß“ befindet (auch wenn dieses sehr kurz ist). Durch diese Aussage können sich Verkehrsmittelwechsel, die zum Beispiel aufgrund der Testdaten entstanden, aber nicht plausibel sind, beseitigen lassen. Dies bedeutet, dass das vorherige korrekt identifizierte Segment (im Sinne des Kontexts) als Ausgang für das nächste Segment verwendet wird und dabei auch der Verkehrsmitteltyp des vorherigen Segments übernommen wird.

Ein Beispiel dafür ist in Abbildung 5.3 zu sehen. Links sieht man dabei den analysierten Track ohne Nachbearbeitung und rechts mit Nachbearbeitung. Die Abfolge der Verkehrsmittel links ist dabei Bus->Auto->Fahrrad->Auto->Fahrrad->Bus->Auto, was schlicht nicht möglich ist aber aufgrund der Trainingsdaten und des fehlenden Kontexts entstanden ist. Am Ende der Nachbearbeitung ist das korrekte Resultat eine durchgängige Busfahrt.

Eine weitere Aufgabe der Nachbearbeitung ist das herausfiltern von einzelnen Fahrradsegmenten. Diese sind aufgrund der vielen unterschiedlichen Fahrrad-Trainingsdaten entstanden. In einigen Fällen konnte es dadurch vorkommen, dass Segmente als Fahrrad-Segment identifiziert wurden, diese aber für sich alleine standen wie z.B. zu Fuß->Fahrrad->Bus->Bus->Bus. Aus diesem Grund wurde das Nachbearbeiten dahinge-

## 5 Segmentierung und Klassifizierung

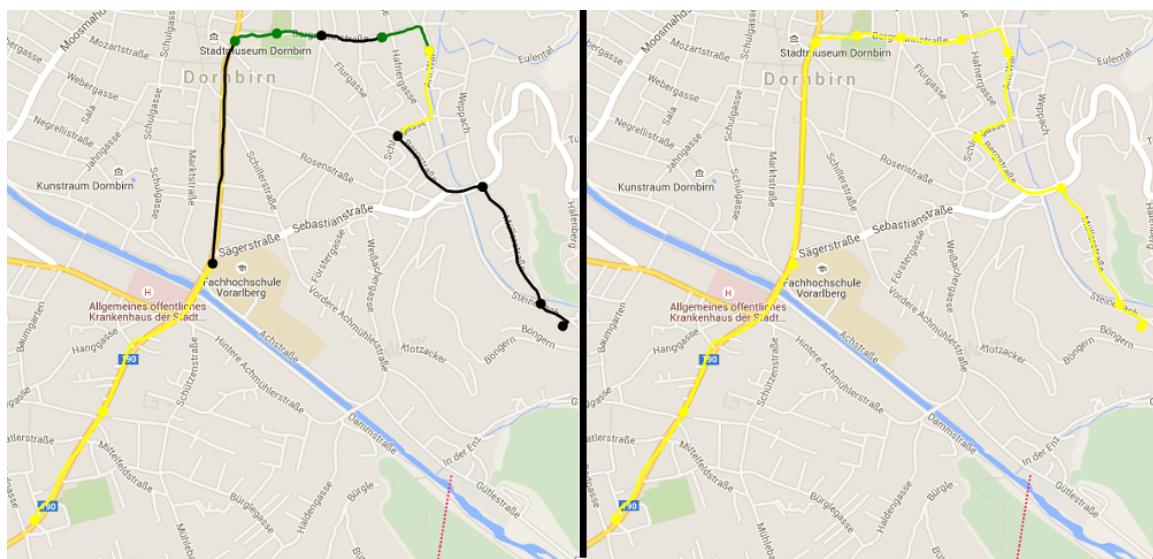


Abbildung 5.3: Kein Nachbearbeiten vs. Nachbearbeiten

hend erweitert, dass einzelne Fahrradsegmente herausgefiltert werden bzw. der Typ des Segments mit dem des vorherigen oder nächsten Segments abgestimmt wird.



## *6 Auswertung*

# Auswertung

Zur Auswertung der Ergebnisse der zwei Analysemethoden wurden, wie bereits in Abschnitt “4.2.2 Results-Seite“ erwähnt, mehrere Schnittstellen und schlussendlich auch Diagramme erstellt. Diese geben einen grundlegenden Überblick über die Resultate der verschiedenen Analysemethoden. Dabei wird zwischen den Gesamtwerten aller Verkehrsmittel sowie den einzelnen Verkehrsmitteln unterschieden. Dadurch lassen sich Tendenzen bezüglich der falschen Klassifizierungen erkennen.

Neben diesen zwei Analysemethoden wurde auch mit der Auswahl der Zusatzwerte für die GIS-Analysemethode experimentiert. Damit wurde überprüft, welche Zusatzwerte welche Auswirkungen auf die Verkehrsmittelerkennung haben.

Darüber hinaus wurde untersucht, welche Erkennungsraten mit Hilfe der unterschiedlichen Algorithmen zur Erstellung eines Entscheidungsbaums bzw. zur Findung des nächsten Attributs zur Erstellung eines Entscheidungsbaums, erreicht werden können. Dies wurde in erster Instanz in RapidMiner und in zweiter Instanz mit dem Prototyp analysiert.

## *6 Auswertung*

### **6.1 Erstellung des Entscheidungsbaums**

Bei der Erstellung eines Entscheidungsbaums bzw. zur Bestimmung des nächsten Attributs bei der Erstellung eines Entscheidungsbaums kann in RapidMiner auf verschiedene Algorithmen bzw. Berechnungsarten zurückgegriffen werden. Darunter befinden sich der Informationsgewinn, der Gini-Index und die Informationszugewinn-Rate. Es wurden in weiterer Folge verschiedene Entscheidungsbäume mit verschiedenen Varianten der Zusatzvariablen und mit unterschiedlichen Algorithmen generiert. Neben den Entscheidungsbäumen wurde mit Hilfe von RapidMiner auch berechnet, welche Erkennungsrate mit diesen Entscheidungsbäumen zu erwarten ist. Dabei stützt sich dieser Wert nur auf die verarbeiteten Trainingsdaten. Die Resultate sind in Tabelle 6.1 zu sehen.

Die Tabelle 6.1 zeigt, dass sich die zu erwartende Erkennungsrate zwischen 61% und fast 74% befindet. Dabei wird mit dem Gini-Index das beste Ergebnis (70,68% im Schnitt) und mit der Zugewinn-Rate (im Schnitt 62,53%) das schlechteste erzielt. Der Informationsgewinn liegt mit 65,66% dazwischen.

Weiters erkennt man, dass die verschiedenen Zusatzwerte unterschiedlich viel Einfluss auf die allgemeine Erkennung haben. So fällt die Erkennungsrate mit dem Gini-Index von 73,68% auf 57,89%, wenn man die berechneten Mittelwerte für Geschwindigkeit und Beschleunigung nicht miteinbezieht. Die maximalen Werte für Beschleunigung und Geschwindigkeit bewirken sowohl zusammen als auch getrennt wenig bis keinen Unterschied für die Erkennung, wenn man den Gini-Index verwendet. Interessant ist hingegen, dass die GIS-Analysemethode ohne Stopprate mit dem Gini-Index auf 75,44% kommt.

Bei den Werten für die Berechnung mittels Informationsgewinn lässt sich noch hervorheben, dass die Testdatenvarianten ohne Beschleunigungswerte bzw. ohne mittlere Beschleunigung Erkennungsraten über von 70% erzielten.

## *6 Auswertung*

Erkennen lässt sich weiters, dass zwischen der Variante mit und der ohne GIS-Daten nur 2% Prozent liegen und diese im Falle der Zugewinn-Rate sogar für die Variante ohne GIS-Daten sprechen. Diese Werte sind jedoch nur auf Basis der Testdaten errechnet und müssen nicht mit der tatsächlichen Erkennungsrate übereinstimmen. Deshalb werden im nächsten Abschnitt die Resultate des Prototyps genauer betrachtet.

## **6.2 Analyse mit dem Prototyp**

Für die Analyse mit Hilfe des Prototyps wurden neben den zwei Analysemethoden mit und ohne GIS-Daten noch drei weitere Entscheidungsbäume ausgewählt. Die wurden auf Grund ihres Abschneidens in Abschnitt “6.1 Erstellung des Entscheidungsbaums“ ausgewählten Entscheidungsbäume und sind:

- mit GIS-Daten ohne Maximalwerte
- mit GIS-Daten ohne Mittelwerte
- mit GIS-Daten ohne Gleise
- mit GIS-Daten ohne Stoprate

### **6.2.1 Gesamtresultate**

Wie in Tabelle 6.2 ersichtlich ist, schneiden alle Analysemethoden mit GIS-Daten besser ab als jene ohne. Mit der Analysemethode ohne GIS-Daten wurden 66,30% richtig erkannt. Die Analyse mit GIS-Daten konnte im Gensatz dazu fast 69% richtig erkennen. Die Auswirkungen der verschiedenen Variationen der Entscheidungsvariablen für die GIS-Analyse werden auch durch dieses Resultat sehr deutlich. So kann die Variante ohne Gleis-Daten 67,71% richtig erkennen, aber die Varianten ohne Durchschnittswerte oder ohne Stoprate erkennen fast 75% richtig.

## *6 Auswertung*

Es wird deutlich, dass verschiedene Kombinationen von Entscheidungsvariablen sehr unterschiedliche Auswirkungen auf die tatsächliche Erkennung haben. Dies bestätigt auch die Aussagen von Zheng und Stenneth, wonach zu viele Variablen das Ergebnis verschlechtern können. [Zheng et al., 2010] [Stenneth et al., 2011]

Da basierend auf den Testdaten und den Gesamtresultaten festgestellt werden kann, dass ohne Mittelwerte für Geschwindigkeit und Beschleunigung sowie der Stopprate bessere Ergebnisse erzielt werden können, kann festgehalten werden, dass folgenden Entscheidungsvariablen die sich sehr gut für den Erkennungsprozess eignen:

- Maximale Geschwindigkeit
- Maximale Beschleunigung
- GIS-Daten für Gleise
- GIS-Daten für Autobahnen
- GIS-Daten für Busstationen

### **6.2.2 Detailresultate**

Die Detailresultate sind spezifisch auf ein Verkehrsmittel ausgerichtet und betrachten die richtig klassifizierten Segmente dieses Verkehrsmittels. Ihnen gegenüber stehen die fälschlicherweise als der jeweilige Verkehrsmitteltyp klassifizierten Segmente. Diese werden auf die eigentlich richtigen Verkehrsmittel aufgeteilt.

#### **Bus**

Für die Segmente, die laut Prototyp mit einem Bus bewältigt worden sind, kann festgestellt werden, dass ein paar wenige Segmente eigentlich Fahrrad- oder Auto-Segmente gewesen sind. Weiters wurden ohne GIS-Daten, ohne Mittelwerte und ohne Stopprate deutlich weniger Bus-Segmente erkannt wie mit den anderen Entscheidungsbäumen die GIS-Daten verwendeten. Ein Auszug der Statistik über die richtig erkannten Bus-

## *6 Auswertung*

Segmente ist in Tabelle 6.3 zu sehen und zeigt, dass Bus-Segmente mit einer Wahrscheinlichkeit von 36,19% mit der GIS-Analyse erkannt werden.

### **Auto**

Bei den Segmenten, welche als Auto-Segmente klassifiziert worden sind, kann festgestellt werden, dass die Erkennungsrate ungleich höher als jene der Bus-Segmente ist und zwischen 75 und 86% liegt. Die meisten falsch als Auto klassifizierten Segmente sind dabei Bus- und Fahrrad-Segmente. Die Ähnlichkeiten zwischen Bus und Auto waren bereits bekannt und nicht weiter überraschend. Die häufige Fehlklassifizierung von Fahrrad-Segmenten als Auto-Segmente führt jedoch zu der Erkenntnis, dass einige Fahrrad-Segmente in den Trainingsdaten sind, welche über eine hohe Geschwindigkeit verfügen und deshalb eine Fehlklassifizierung ermöglichen. Ein Auszug der Statistik über die richtig erkannten Auto-Segmente ist in Tabelle 6.4 zu sehen und zeigt, dass die Erkennung von Auto-Segmenten für die GIS-Analyse bei 82,22% liegt. Die Analyse-Methoden ohne Beschleunigung liefern für diesen Fall die beste Erkennungsrate mit 86,12%.

### **Fahrrad**

Die Resultate für die Fahrrad-Segmente zeigen, dass sich die Fehlklassifizierungen großteils in den Segmenten vom Typ “zu Fuß“ und Auto befinden. Der Grund für die falschen Klassifizierungen als Auto-Segment ist derselbe wie bei den Resultaten zum Typ Auto. Schnelle Radsegmente, die z.B. durch einen Rennradfahrer entstanden sind, können nicht gut von langsameren Autofahrten z.B. in der Stadt unterschieden werden. Weiters ist es auch schwierig, eine sehr langsame Radfahrt z.B. wenn es bergauf geht von einem Fußgänger zu unterscheiden, da Geschwindigkeit, Beschleunigung und Stopprate von beiden Fortbewegungsarten sehr ähnlich sein können. Ein Auszug der Statistik über die richtig erkannten Fahrrad-Segmente ist in Tabelle 6.5 zu sehen und zeigt, dass 46,68% der Fahrradsegmente mit Hilfe der GIS-Analyse erkannt werden. Alle anderen Analysen liefern in diesem Fall bessere Ergebnisse. Besonders die Variante ohne Stopprate liefert

## *6 Auswertung*

in diesem Fall ein wesentlich besseres Ergebnis mit 70%.

### **Zu Fuß**

Die Ergebnisse für die “zu Fuß“-Segmente sind sehr eindeutig und liegen für alle Analyse-Methoden über 93%. Aber auch hier liefert die GIS-Analyse-Methoden mit weniger Zusatzvariablen bessere Ergebnisse. Die falschen Klassifizierungen liegen hauptsächlich im Bereich der Fahrrad-Segmente und Auto-Segmente. Dies lässt in beiden Fällen auf eine sehr langsame Fortbewegungsgeschwindigkeit schließen. Ein Auszug der Statistik über die richtig erkannten “zu Fuß“-Segmente ist in Tabelle 6.6 zu sehen.

### **Zug**

Bei den Resultaten der Zug-Segmente konnten die GIS-Daten ihre Stärke voll ausspielen und im Gegensatz zu den Analyse-Methoden ohne GIS-Informationen von Gleisen, die eine ca. 35%ige Erkennungsrate haben. Mit Hilfe der GIS-Daten über Gleise konnte die Erkennung auf bis zu 93% gesteigert werden. Außerdem konnten auch die Fehlklassifizierungen als Auto-Segmente eliminiert werden. Ein Auszug der Statistik über die richtig erkannten Zug-Segmente ist in Tabelle 6.7 zu sehen.

## **6.3 Zusammenfassung**

Wird das Gesamtergebnis der GIS-Analysevarianten betrachtet, so kann festgestellt werden, dass für Segmente vom Typ Auto, Zug und “zu Fuß“, je nach Kombination der verwendeten Zusatzwerte, gute Erkennungsraten (75-86%, 91-93%, 93-97%) gefunden worden sind. Diese Verkehrsmittel können also gut erkannt werden. Wo es Verbesserungsbedarf gibt, sind eindeutig die Bus- und Fahrrad-Segmente.

Die Fahrrad-Segmente haben zum Teil schlecht abgeschnitten, weil die Trainings- und Testdaten für die Fahrrad-Segmente großteils von Sportlern stammen. Zum einen ist damit gemeint, dass im Flachen (Rennrad) und abwärts sehr hohe Geschwindigkeiten gemessen worden sind. Dies erklärt die Fehlklassifizierungen der Fahrrad-Segmente

## *6 Auswertung*

als Auto-Segmente. Zum anderen wurden bei Fahrrad-Segmenten auch sehr langsame Geschwindigkeiten gemessen. Dies ist dann vorgekommen, wenn es bergaufwärts ging. Dabei waren diese Abschnitte aufwärts meistens nicht kurz (so wie in einer Stadt) und hatten Charakteristika, die eher zu Geh-Segmenten passten als zu einem Fahrrad-Segment.

Weiters konnte beobachtet und schlussendlich auch bestätigt werden, dass zu viele Zusatzvariablen das Ergebnis verschlechtern können und mit weniger Variablen unter Umständen eine bessere Erkennungsrate erreicht werden kann. Dies bekräftigt auch die Aussagen von Zheng [Zheng et al., 2010] und Stenneth [Stenneth et al., 2011], welche ihre verwendeten Zusatzvariablen evaluiert haben. Daraus ist bei beiden hervorgegangen, dass bei einer bestimmten Anzahl von Zusatzvariablen keine Verbesserung, sondern eher eine Verschlechterung des Ergebnisses zu erwarten ist. Dies lässt sich auch in den vorherigen Resultaten beobachten. Die fehlende Stopprate bei der Erkennung von Fahrradsegmenten bewirkt, dass deutlich mehr Fahrradsegmente richtig klassifiziert werden können. Im Gegensatz dazu steht allerdings das Resultat für die Analysevariante ohne Stopprate bei Bus-Segmenten.

Bei der Betrachtung der Ergebnisse muss schlussendlich noch hervorgehoben werden, dass es sich bei den verwendeten Trainings- und Testdaten um überschaubare Datenmengen handelt und sich mit einer größeren Datenmenge wahrscheinlich bessere Erkennungsraten erzielen lassen werden. Dies kann damit begründet werden, dass bei kleinen Datenmengen wenige spezielle Testdaten, wie z.B. sehr langsame Fahrradsegmente (aufwärts) oder sehr schnelle Fahrradsegmente (abwärts, Rennradfahrer), wesentlich größeren Einfluss auf den Entscheidungsbaum und im Endeffekt auf die Erkennung haben.

## 6 Auswertung

Trainingsdatenvarianten	Informationsgew.	Gewinn-Rate	Gini-Index
Ohne GIS-Daten	61,40	66,67	71,93
Mit GIS-Daten	64,91	64,91	73,68
Mit GIS-Daten ohne max. Beschl.	64,91	59,65	73,68
Mit GIS-Daten ohne max. Geschw.	64,91	64,91	73,68
Mit GIS-Daten ohne max. Werte	68,42	61,40	73,68
Mit GIS-Daten ohne mittl. Beschl.	71,93	64,91	71,93
Mit GIS-Daten ohne mittl. Geschw.	61,40	63,16	68,42
Mit GIS-Daten ohne Mittelwerte	61,40	57,89	57,89
Mit GIS-Daten ohne Geschw.	61,40	61,40	68,42
Mit GIS-Daten ohne Beschl.	70,18	59,65	71,93
Mit GIS-Daten ohne Gleis	64,91	63,16	68,42
Mit GIS-Daten ohne Autobahn	66,67	61,40	70,18
Mit GIS-Daten ohne Bushaltest.	64,91	61,40	70,18
Mit GIS-Daten ohne Stoprate	71,93	64,91	75,44

Tabelle 6.1: Genauigkeit von Entscheidungsbäumen mit verschiedenen Entscheidungskriterien und Trainingsdatenvarianten (siehe Abschnitte “3.1.1 Generierung eines Entscheidungsbaumes“ und “3.1.2 Die Entscheidungsbaum-Operatoren von RapidMiner“)

## 6 Auswertung

	#	%
Gesamt	1.347	100,00
Ohne GIS-Daten	893	66,30
Mit GIS-Daten	926	68,75
Mit GIS-Daten ohne Maximalwerte	953	70,75
Mit GIS-Daten ohne Durchschnittswerte	1.004	74,54
Mit GIS-Daten ohne Gleise	912	67,71
Mit GIS-Daten ohne Stoprate	997	74,02

Tabelle 6.2: Gesamtresultate im Überblick

	%
Ohne GIS-Daten	21,90
Mit GIS-Daten	36,19
Mit GIS-Daten ohne Maximalwerte	38,10
Mit GIS-Daten ohne Durchschnittswerte	18,10
Mit GIS-Daten ohne Gleise	35,24
Mit GIS-Daten ohne Stoprate	14,29

Tabelle 6.3: Richtig erkannte Bus-Segmente

	%
Ohne GIS-Daten	77,83
Mit GIS-Daten	82,22
Mit GIS-Daten ohne Maximalwerte	86,14
Mit GIS-Daten ohne Durchschnittswerte	83,14
Mit GIS-Daten ohne Gleise	78,75
Mit GIS-Daten ohne Stoprate	75,52

Tabelle 6.4: Richtig erkannte Auto-Segmente

## *6 Auswertung*

	%
Ohne GIS-Daten	52,49
Mit GIS-Daten	46,68
Mit GIS-Daten ohne Maximalwerte	47,72
Mit GIS-Daten ohne Durchschnittswerte	64,52
Mit GIS-Daten ohne Gleise	52,70
Mit GIS-Daten ohne Stoprate	70,33

Tabelle 6.5: Richtig erkannte Fahrrad-Segmente

	%
Ohne GIS-Daten	93,93
Mit GIS-Daten	93,93
Mit GIS-Daten ohne Maximalwerte	94,64
Mit GIS-Daten ohne Durchschnittswerte	96,43
Mit GIS-Daten ohne Gleise	93,93
Mit GIS-Daten ohne Stoprate	97,14

Tabelle 6.6: Richtig erkannte “zu Fuß“-Segmente

	%
Ohne GIS-Daten	35,42
Mit GIS-Daten	91,67
Mit GIS-Daten ohne Maximalwerte	93,75
Mit GIS-Daten ohne Durchschnittswerte	91,67
Mit GIS-Daten ohne Gleise	35,42
Mit GIS-Daten ohne Stoprate	91,67

Tabelle 6.7: Richtig erkannte Zug-Segmente

# Ausblick

Die Ergebnisse dieser Arbeit sind bereits im Abschnitt “6 Auswertung“ detailliert beschrieben. Es kann jedoch hervorgehoben werden, dass die verwendeten Algorithmen und Methodiken die erwarteten Ergebnisse lieferten. Es ist weiters zu erwarten, dass die Trefferquote beim Klassifizieren der Segmente mit weiteren Trainingsdaten ansteigt.

## 7.1 GPX-Daten im Überblick

Wie bereits erwähnt, konnte während der Arbeit an dem Prototypen festgestellt werden, dass die Qualität und Diversität der verwendeten GPX-Daten verbesserungswürdig ist. Mit einer größerer Menge an Trainingsdaten könnte ein besseres Resultat bei der Erkennung bzw. Klassifizierung erzielt werden.

### 7.1.1 GPX-Datenqualität

Mit der Qualität der GPX-Daten sind zwei verschiedene Dinge gemeint. Zum einen ist das Filtern einfacher, wenn die GPX-Daten ein Genauigkeitswert (z.B. errechnet über die Anzahl der momentan verfügbaren Satelliten) enthalten, wie es bereits in anderen Publikationen ([Stenneth et al., 2011], [Nadine Schüssler et al., 2011]) gezeigt wurde.

## *7 Ausblick*

Dadurch könnte besser abgeschätzt werden, ob die momentanen Geschwindigkeitswerte auch im momentanen Kontext realistisch sind und nicht nur im globalen Kontext. Ist also die Geschwindigkeit von mehr als 100 km/h realistisch, wenn die aufzeichnende Person gerade zu Fuß unterwegs ist, oder handelt es sich eher um einen fehlerhaften Ausreißer, der sich über einen niederen Genauigkeitswert erkennen lässt?

Unter der Qualität der GPX-Daten ist auch zu verstehen, dass die Genauigkeit bei der händischen Segmentierung der Trainingsdaten sehr von Person zu Person variiert und es durchaus vorkommen kann, dass eine Zugfahrt laut GPX-Datei 10 GPX-Punkte länger dauert, als sie eigentlich sollte. Dies kann vor allem bei kurzen Strecken, wie sie in den aktuellen Testdaten oft vorkommen, zu einer Beeinflussung der Geschwindigkeits- und Beschleunigungswerte führen.

### **7.1.2 Diversität der GPX-Daten**

Mitunter konnte festgestellt werden, dass bei den verwendeten GPX-Daten öfter gleichen Strecken mit denselben Verkehrsmitteln aufgezeichnet wurden. Wesentlich interessanter wäre es natürlich, wenn es eine größere Anzahl von Teilnehmern beim Aufzeichnen der GPX-Tracks gäbe. Außerdem wäre es natürlich interessant, wenn diese Teilnehmer möglichst unterschiedliche Fortbewegungsgewohnheiten hätten. Damit ist gemeint, dass es durchaus öfter dieselbe Strecke in den Trainingsdaten geben darf, aber diese mit möglichst vielen verschiedenen Verkehrsmitteln bewältigt wird. Weiters wäre es auch beim Individualverkehr interessant, wie sich verschiedene Persönlichkeiten im Sinne eines aggressiven/geschwindigkeitsbetonten bzw. passiven/gemütlichen Fahrstils auswirken.

Ferner hat sich gezeigt, dass man die Testdaten möglichst auf das Hauptinteressensgebiet beschränken soll. Damit ist gemeint, dass das Aufnehmen einer MTB-Tour in die Trainingsdaten nur bedingt interessant bzw. hilfreich ist. Denn früher oder später

## *7 Ausblick*

bringen die entweder sehr hohen (abwärts) oder sehr niederen (aufwärts) Geschwindigkeiten unnötige Irritationen in den Klassifizierungsprozess. Ambitionierte Sportler liefern keine repräsentativen Daten zur Analyse für den durchschnittlichen Verkehrsteilnehmer.

### **7.1.3 Quantität der GPX-Daten**

Neben qualitativen Verbesserungen ist es auch praktisch, wenn man auf eine große Datenmenge zurückgreifen kann. Zumal abhängig vom gewählten Algorithmus für die Generierung des Entscheidungsbaums eine große Datenmenge benötigt wird, um ein repräsentatives Ergebnis zu erhalten (siehe Abschnitt 3.1.1 Generierung eines Entscheidungsbaumes).

## **7.2 GIS-Daten**

Wie bereits am Beginn der Arbeit erwähnt, gab es keine Möglichkeit auf die Daten der öffentlichen Verkehrsmittel zuzugreifen. Allerdings konnte bereits Stenneth [Stenneth et al., 2011] sehr genaue Resultate mit diesen Daten erzielen. Deshalb kann angenommen werden, dass auch für diesen Prototyp die Genauigkeit erhöht werden kann, wenn auf solche Daten zurückgegriffen wird. Dies könnte auch dann von großem Vorteil sein, wenn man zwischen verschiedenen öffentlichen Verkehrsmitteln unterscheiden möchte, wie zum Beispiel U-Bahn, Straßenbahn sowie Bus.

## **7.3 Entscheidungsbaum**

Bezüglich der Entscheidungsbäume sollte erwähnt werden, dass die Möglichkeit der automatisierten Aktualisierung bzw. die erneute Generierung der Entscheidungsbäume

## *7 Ausblick*

eine erhebliche Vereinfachung beim Einfügen von neuen Trainingsdaten darstellen würde. Die Verarbeitung der Trainingsdaten kann bereits mit einem Befehl gestartet werden. Jedoch muss weiterhin die erzeugte Trainingsdatendatei in RapidMiner importiert, der Entscheidungsbaum neu generiert und schlussendlich der Baum in Textform in das Projekt kopiert werden (siehe Anhang 2). Das Parsen des Entscheidungsbaumes wird dann wiederum automatisch vom Prototypen gemacht (siehe Abschnitt 5.5.1 Erstellen der Entscheidungsbäume).

Die Schritte im RapidMiner sind mit der Community-Version leider unumgänglich. Allerdings gibt es auch eine Serverversion von RapidMiner, bei der es laut Dokumentation über verschiedene Schnittstellen, wie z.B. via Webservices, möglich sein soll, zu kommunizieren.

# Literaturverzeichnis

[Biljecki et al., 2013] Biljecki, F., Ledoux, H., and van Oosterom, P. (2013). Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science*, 27(2).

[Buschmann, 1998] Buschmann, F. (1998). Pipes and Filters. In *Pattern-orientierte Software-Architektur: ein Pattern-System*, Professionelle Softwareentwicklung, pages 54–71. Addison-Wesley.

[Caron et al., 2006] Caron, F., Duflos, E., Pomorski, D., and Vanheeghe, P. (2006). GPS/IMU data fusion using multisensor Kalman filtering: introduction of contextual aspects. *Information Fusion*, 7(2):221–230.

[Douglas C. Giancoli, 2010] Douglas C. Giancoli (2010). *Physik: Lehr- und Übungsbuch*. Pearson Deutschland GmbH, 3 edition.

[Gonzalez et al., 2010] Gonzalez, P. A., Weinstein, J. S., Barbeau, S. J., Labrador, M. A., Winters, P. L., Georggi, N. L., and Perez, R. (2010). Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. *IET intelligent transport systems*, 4(1):37–49.

[Howard Hamilton, 2009] Howard Hamilton (2009). Machine Learning/Inductive Inference/Decision Trees/Overview. <http://www2.cs.uregina.ca/~dbd/cs831/index.html>.

## *Literaturverzeichnis*

- [Jeffrey P. Bradford et al., 1998] Jeffrey P. Bradford, Clayton Kunz, Ron Kohavi, Cliff Brunk, and Carla E. Brodley (1998). Pruning Decision Trees with Misclassification Costs. Technical Report 51, Purdue University.
- [Johannes Fürnkranz, 2008] Johannes Fürnkranz (2008). Decision-Tree Learning. <http://www.ke.tu-darmstadt.de/lehre/archiv/ws0809/mldm/dt.pdf>.
- [Jun et al., 2006] Jun, J., Guensler, R., and Ogle, J. H. (2006). Smoothing methods to minimize impact of Global Positioning System random error on travel distance, speed, and acceleration profile estimates. *Transportation Research Record: Journal of the Transportation Research Board*, 1972(1):141–150.
- [Microsoft Research, 2015] Microsoft Research (2015). GeoLife GPS Trajectories - Microsoft Research. <http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/>.
- [Movable Type Ltd, 2015] Movable Type Ltd (2015). Calculate distance and bearing between two Latitude/Longitude points using haversine formula in JavaScript. <http://www.movable-type.co.uk/scripts/latlong.html>.
- [Nadine Schüssler et al., 2011] Nadine Schüssler, Lara Montini, and Christoph Dobler (2011). Improving post-processing routines for gps oversavations using prompted-recall data. In *9th International conference on survey methods in transport*.
- [Quinlan, 1986] Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.
- [Raileanu and Stoffel, 2004] Raileanu, L. E. and Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93.
- [RapidMiner, 2015] RapidMiner (2015). RapidMiner Studio - RapidMiner Documentation. <http://docs.rapidminer.com/studio/>.

## *Literaturverzeichnis*

- [Reddy et al., 2008] Reddy, S., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2008). Determining transportation mode on mobile phones. In *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*, pages 25–28. IEEE.
- [Reddy et al., 2010] Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2).
- [Schuessler and Axhausen, 2009] Schuessler, N. and Axhausen, K. W. (2009). Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board*, 2105(1):28–36.
- [Sebastian Nagel, 2011] Sebastian Nagel (2011). Möglichkeitsstudie zum Projekt: Mobilitäts-Tracker.
- [Stenneth et al., 2011] Stenneth, L., Wolfson, O., Yu, P. S., and Xu, B. (2011). Transportation mode detection using mobile phones and GIS information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 54–63. ACM.
- [Thomas Mitchell, 1997] Thomas Mitchell (1997). Which Attribute Is the Best Classifier? In *Machine Learning*, pages 55–59. McGraw-Hill Publ.Comp., international edition edition.
- [Tom Dietterich, 1995] Tom Dietterich (1995). Overfitting and Undercomputing in Machine Learning. 27(3):326–327.
- [Topografix, 2004] Topografix (2004). GPX 1.1 Schema Documentation. <http://www.topografix.com/GPX/1/1/>.
- [Wei-Yin Loh, 2008] Wei-Yin Loh (2008). Classification and Regression Tree Methods.

### *Literaturverzeichnis*

- In *Encyclopedia of Statistics in Quality and Reliability*, pages 315–323. Department of Statistics, University of Wisconsin.
- [Youtube, 2015] Youtube (2015). Youtube statistics. <http://www.youtube.com/yt/press/statistics.html>.
- [Zheng et al., 2010] Zheng, Y., Chen, Y., Li, Q., Xie, X., and Ma, W.-Y. (2010). Understanding transportation modes based on GPS data for web applications. *ACM Transactions on the Web (TWEB)*, 4(1).
- [Zheng et al., 2008a] Zheng, Y., Li, Q., Chen, Y., Xie, X., and Ma, W.-Y. (2008a). Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM.
- [Zheng et al., 2008b] Zheng, Y., Liu, L., Wang, L., and Xie, X. (2008b). Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*, pages 247–256. ACM.

# Listings

3.1 GPX-Datei . . . . .	35
4.1 Filterkonfiguration . . . . .	48
4.2 Segmentierungskonfiguration . . . . .	49
4.3 Analysekonfiguration . . . . .	50
5.1 Entscheidungsbaum in Textform . . . . .	66
5.2 Ausschnitt des generierten Entscheidungsbaums als PHP-Klasse . . . . .	67
7.1 GPX-Datei für Trainingsdaten . . . . .	104
7.2 Kommando für das Erstellen der Trainingsdaten . . . . .	104



# Tabellenverzeichnis

3.1 Trainingsdatenübersicht . . . . .	33
5.1 Entscheidungsvariablenübersicht . . . . .	60
6.1 Genauigkeit von Entscheidungsbäumen mit verschiedenen Entscheidungskriterien und Trainingsdatenvarianten (siehe Abschnitte “3.1.1 Generierung eines Entscheidungsbaumes“ und “3.1.2 Die Entscheidungsbaum-Operatoren von RapidMiner“) . . . . .	78
6.2 Gesamtresultate im Überblick . . . . .	79
6.3 Richtig erkannte Bus-Segmente . . . . .	79
6.4 Richtig erkannte Auto-Segmente . . . . .	79
6.5 Richtig erkannte Fahrrad-Segmente . . . . .	80
6.6 Richtig erkannte “zu Fuß“-Segmente . . . . .	80
6.7 Richtig erkannte Zug-Segmente . . . . .	80



# Abbildungsverzeichnis

2.1	Geolife (Quelle: research.microsoft.com) . . . . .	10
2.2	Fortbewegungsmittel-Hierarchie (Quelle: [Biljecki et al., 2013]) . . . . .	15
3.1	Konfigurationsmöglichkeiten für Entscheidungsbäume . . . . .	30
3.2	Entscheidungsbaum ohne GIS-Daten . . . . .	31
3.3	Entscheidungsbaum mit GIS-Daten . . . . .	32
3.4	Die App myTrack . . . . .	36
4.1	Pipes- und Filter- Struktur für Trainingsdaten . . . . .	41
4.2	Pipes- und Filter- Struktur der Webapplikation . . . . .	44
4.3	Korrigieren eines Verkehrsmitteltyps eines analysierten Tracks . . . . .	46
4.4	Ausschnitt zu den Resultaten der analysierten Tracks . . . . .	47
4.5	Rohdaten eines GPS-Tracks . . . . .	51
4.6	Gefilterte Rohdaten eines GPS-Tracks . . . . .	51
4.7	Segmentierter GPS-Track . . . . .	52
4.8	GPS-Track mit bestimmten Verkehrsmitteln . . . . .	52
4.9	Nachbearbeiteter GPS-Track . . . . .	52
5.1	Grundlegende Segmentierung . . . . .	56
5.2	Stopprate laut Zheng (Quelle: [Zheng et al., 2010]) . . . . .	63
5.3	Kein Nachbearbeiten vs. Nachbearbeiten . . . . .	69

## *Abbildungsverzeichnis*

7.1	Filtrern - 1. Fall	98
7.2	Filtrern - 2. Fall	98
7.3	Filtrern - 3. Fall	99
7.4	GPS-Track ohne Filter	101
7.5	GSP-Track mit Filter	102

# **Anhang 1**

Wie auch bei vielen anderen Publikation, die sich mit GPS-Daten beschäftigen, konnte auch bei dieser Arbeit festgestellt werden, dass sich in den GPS-Spuren einige Ausreißer befanden. Dies konnte vor allem dann beobachtet werden, wenn man sich in einem Zug befand, durch einen Tunnel fuhr oder auch wenn man sich auf einem überdachten Bahnsteig befand. Die Ausreißer werden durch unrealistisch große Distanzabstände oder Sprünge in den Höhenwerten bemerkt. Da diese Werte das Ergebnis verfälschen würden, mussten verschiedene Filter implementiert werden.

In verschiedensten Publikation zum Thema GPS wird beim Filtern von fehlerhaften Ausreißern auf den Kalman-Filter gesetzt (z.B. [Caron et al., 2006] und [Jun et al., 2006]). Da das Filtern an sich aber nicht im Fokus dieser Arbeit stand und es zur Zeit dieser Arbeit keine Implementation des Kalmanfilters für die gewählte Programmiersprache PHP gab, wurde auf einfachere Filtermöglichkeiten, wie sie in [Schuessler and Axhausen, 2009] grob beschrieben werden, gesetzt.

Konkret wurden Filter für Zeit, Distanz und Höhenmeter implementiert. Die somit bereinigten Resultate hatten bereits auf den Entscheidungsbaum ohne GIS-Daten große Auswirkungen. Die Grenzwerte für diese Filter können in einer Konfigurationsdatei festgelegt und je nach geografischer Region und Testdaten angepasst werden. Diese Filter können auch als konfigurierbares Regelwerk verstanden und wie folgt beeinflusst werden:

## Anhang 1

- minimale Zeitspanne zwischen 2 Punkten
- maximale Distanz pro Sekunde
- minimale Distanz pro Sekunde
- maximale Änderung der Höhenmeter zwischen 2 Punkten pro Sekunde
- minimaler Wert für das Verhältnis zwischen gültigen und aufgezeichneten Punkten
- Anzahl der zu überspringenden Punkte am Start
- minimale Anzahl von Punkten pro Segment

GPS-Punkte, welche nicht den definierten Grenzwerten entsprechen, werden in die weiteren Verarbeitung nicht einbezogen. Abgesehen vom Zeitfilter betrachten alle anderen Filter die gemessenen Werte in Relation zur gemessenen Zeit, was bedeutet, dass der Zeitwert größer als 0 sein muss und von mehreren GPS-Punkten mit derselben Zeit nur einer betrachtet wird. In den nächsten Absätzen findet sich eine genauere Beschreibung der einzelnen Parameter und ihrer Auswirkungen.

**Minimale Zeitspanne zwischen 2 Punkten** Mit Hilfe dieses Konfigurationsparameters kann zum einen sichergestellt werden, dass sich nicht mehrere Trackpunkte mit dem selben Zeitstempel eingeschlichen haben. Zum anderen kann damit auch die Genauigkeit bzw. die Anzahl der zu verarbeitenden Punkte steuern.

**Maximale/minimale Distanz pro Sekunde** Durch diesen Parameter kann sichergestellt werden, dass man innerhalb einer bestimmten Zeitspanne nur eine gewisse Strecke zurücklegen kann. Es können auch Punkte gefiltert werden, in denen man sich nicht wirklich bewegt hat. Dadurch können die richtig großen Sprünge und Punkte ohne Bewegung herausgefiltert werden.

## Anhang 1

**Maximale Änderung der Höhenmeter zwischen 2 Punkten pro Sekunde** Hierbei wird überprüft, ob es große Sprünge im Bereich der Höhenmeter gibt.

**Minimaler Wert für das Verhältnis zwischen gültigen und aufgezeichneten Punkten**  
Durch dieses Verhältnis kann entschieden werden, ob sich überhaupt noch genug gültige GPS-Informationen für den weiteren Bestimmungsprozess in einem Track befinden.

**Anzahl der zu überspringenden Punkte am Start** Durch diesen Konfigurationsparameter kann gesteuert werden, wie viele Punkte am Beginn einer Aufzeichnung übersprungen werden. Dies röhrt daher, dass sich beim Start einer Aufzeichnung gerne Ausreißer einschleichen, bis sich der Positionsbestimmungsvorgang eingependelt hat. Durch diesen Parameter können einige dieser Trackpoints übersprungen werden.

**Minimale Anzahl von Punkten pro Segment** Über diesen Parameter kann gesteuert werden, wie viele Punkte sich in einem Tracksegment befinden, damit es überhaupt weiterverarbeitet wird. Der minimale Wert für diesen Parameter ist 2.

## Verschiedene Fälle beim Filtern

Es gibt drei Fälle, welche beim Filtern von Ausreißern abgedeckt werden sollten. Der grundlegende Algorithmus, welcher vom aktuellen Punkt ausgehend einen neuen gültigen Punkt sucht und alle ungültigen überspringt, funktioniert in den ersten zwei Fällen. Im dritten Fall muss noch eine zusätzliche Überprüfung stattfinden.

### 1. Fall

Beim ersten Fall befinden sich ein oder mehrere Ausreißer am Ende der GPS-Spur, wie es in Abbildung 7.1 beim letzten Punkt der Fall ist. Dies bedeutet, dass ab einem gewissen

## *Anhang 1*

Punkt keine weiteren validen Punkte gefunden und alle folgenden Punkte übersprungen werden.

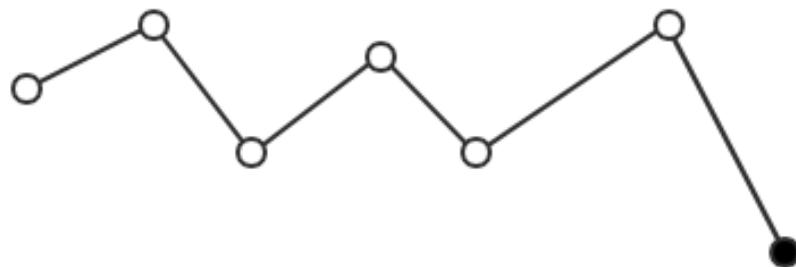


Abbildung 7.1: Filtern - 1. Fall

## **2. Fall**

Beim zweiten Fall befinden sich ein oder mehrere Ausreißer zwischen validen vorangegangenen und nachfolgenden Punkten. Ein Beispiel ist in Abbildung 7.2 mit dem vierten Punkt als Ausreißer abgebildet. Dies bedeutet, dass ein oder mehrere Punkte übersprungen werden und danach mit den gültigen Punkten weitergearbeitet werden kann.

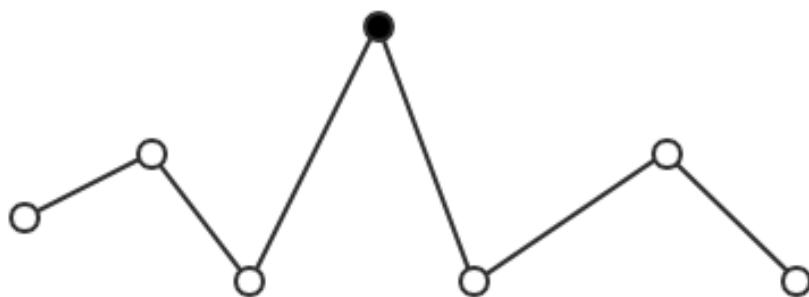


Abbildung 7.2: Filtern - 2. Fall

## **3. Fall**

Im dritten Fall befinden sich ein oder mehrere Ausreißer am Beginn der GPS-Spur. Damit

## Anhang 1

ist gemeint, dass vom Start weg keine gültigen Punkte vorhanden sind und erst im Laufe der Aufzeichnung gültige Punkte aufgezeichnet werden. Dies kann vorkommen, wenn die Aufzeichnung der GPS-Spur sofort nach dem Aktivieren des GPS-Moduls startet. Die Position konnte noch nicht mit ausreichender Genauigkeit bestimmt werden, und es wird mit einer niederen Genauigkeit gestartet. Im Laufe der Aufzeichnung steigt die Genauigkeit, wodurch es zu einem Sprung von ungenauen zu den genauen Punkten kommen kann. Ein Beispiel hierfür ist in Abbildung 7.3 mit den ersten 3 Punkten als Ausreißern ersichtlich.

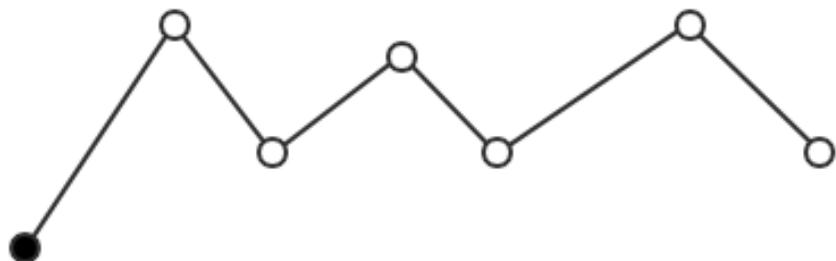


Abbildung 7.3: Filtern - 3. Fall

Da keinerlei Information über die Genauigkeit der aufgezeichneten GPS-Punkte gesammelt wurde, ist es eine komplexe Aufgabe, einen gültigen Startpunkt zu finden ohne unnötig viele GPS-Daten zu überspringen. In einzelnen Testfällen kam es vor, dass ein großer Teil der Strecke einfach weggelassen wurde, weil kein gültiger Startpunkt bzw. nicht genügend valide und aufeinanderfolgende Punkte gefunden werden konnten. Deshalb wurde ein konstanter Parameter für die Anzahl der zu überspringenden GPS-Punkte am Anfang eines Tracks festgelegt.

## *Anhang 1*

### **Zeitfilter**

Der Zeitfilter überprüft, ob der Abstand zwischen zwei GPS-Punkten größer oder gleich einem minimalen Wert (in diesem Fall 0) ist. Dadurch wird verhindert, dass zwei Punkte mit demselben Zeitstempel verarbeitet werden und bei den zeitabhängigen Berechnungen durch 0 dividiert wird. Außerdem kann man dadurch auch steuern, wie viele Punkte pro GPS-Spur überprüft werden (z.B. nur jeder 2. Punkt). Es kann festgelegt werden, welche Punkte ausgelassen werden sollen, um den Prozess zu beschleunigen oder weil sich der Grad an Genauigkeit nicht wesentlich verbessert.

### **Distanzfilter**

Der Distanzfilter kontrolliert, ob sich der Abstand zwischen zwei Punkten im Verhältnis zur Zeit in einem gewissen Bereich befindet. In dieser Arbeit wurde größer 0 m pro Zeiteinheit als minimale und kleiner 50 m pro Zeiteinheit als maximale Distanz festgelegt. Liegt ein Punkt nicht innerhalb dieser Grenzen, so wird der aktuelle Punkt mit dem Punkt nach dem Ausreißer verglichen. Dies wird solange gemacht, bis wieder ein Punkt mit valider Distanz gefunden wird oder keine GPS-Punkte mehr vorhanden sind.

### **Höhenfilter**

Der Höhenfilter entfernt, ähnlich wie der Geschwindigkeitsfilter, jene GPS-Punkte, bei welchen die Differenz der Höhenwerte zu groß ist. Im Fall der hier verwendeten Trainingsdaten wurde 25 m/s für diesen Filter festgelegt, alle Punkte mit einer größeren Differenz werden herausgefiltert.

## Anhang 1

# Resultat der Filter

Das Resultat der angewandten Filter auf die GPS Daten ist in Abbildung 7.4 (vor Filterung) und Abbildung 7.5 (nach Filterung) zu sehen. Es konnte jedoch festgestellt werden, dass zum Beispiel eine zu hohe Mindestdistanz bei Testdaten einer Mountainbike-Tour zu einem Verlust von sehr vielen Trackpunkten führen würde. Darum sollte die Mindestdistanz eher kleiner gewählt werden, um auch bei solchen Tracks gute Resultate und eine gewisse Detailtreue zu erhalten oder gänzlich auf solche Testdaten verzichtet werden.

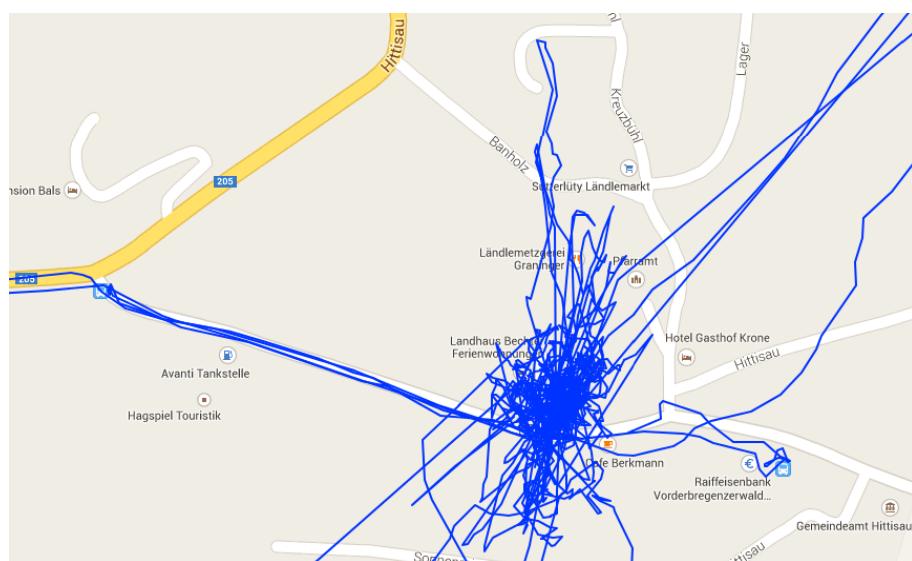


Abbildung 7.4: GPS-Track ohne Filter

## Anhang 1

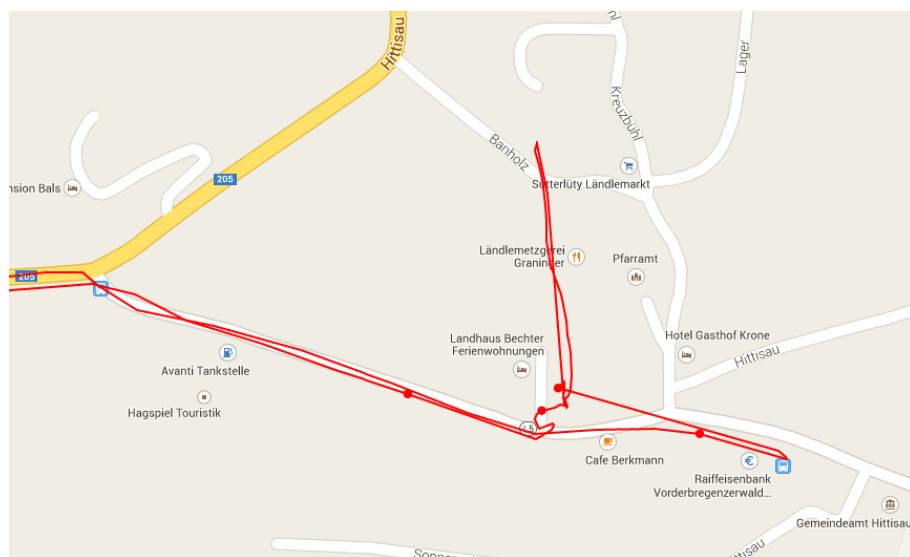


Abbildung 7.5: GSP-Track mit Filter

# **Anhang 2**

Im folgenden Abschnitt werden zwei Vorgänge beschrieben, die für das Hinzufügen von weiteren Analysemethoden (neben der Basis- und GIS-Variante) oder das Erstellen von Entscheidungsbäumen anhand neuer Daten essentiell sind. Der erste Abschnitt beschreibt dabei, welche Schritte auszuführen sind, um mit Hilfe neuer/geänderter Rohdaten neue Trainingsdaten zu erstellen. Der zweite Abschnitt geht dann auf die Erstellung eines Entscheidungsbaums mit RapidMiner auf Basis der Trainingsdaten ein.

## **Ermittlung der Trainingsdaten**

Grundsätzlich unterscheiden sich Trainingsdaten im Sinne von GPX-Dateien nicht wesentlich von den Rohdaten, die man vom GPX-Gerät bekommt. Da die anschließende Erstellung eines Entscheidungsbaums auf Basis dieser Daten erfolgt, wird eine zusätzliche Information über das tatsächlich verwendete Verkehrsmittel für das jeweilige Segment benötigt. Dies bedeutet, dass jede GPX-Trainingsdatei denselben Aufbau wie eine normale Datei hat, aber bei jedem Tracksegment (trkseg) ein Verkehrsmitteltyp mit angegeben werden muss (siehe Listing 7.1).

Bei jedem Verkehrsmittelwechsel muss ein GPX-Tracksegment in die Trainingsdaten eingepflegt werden. Dieses Segment muss zusätzlich über das Attribut “type“ verfügen, welches den eigentlichen Verkehrsmitteltyp angibt.

## Anhang 2

```
1 ...
2     <trk>
3         ...
4             <trkseg type = "bike">
5                 <trkpt lat = "47.39786" lon = "9.735109">
6                     <ele>475.0</ele>
7                     <time>2015-02-19T07:20:18.156Z</time>
8                 </trkpt>
9             ...
10            </trkseg>
11        ...
12    </trk>
13 ...
```

Listing 7.1: GPX-Datei für Trainingsdaten

```
1 app/console tmd:generate:trainingdata
```

Listing 7.2: Kommando für das Erstellen der Trainingsdaten

Auf Basis dieser Trainingsdaten kann dann mit folgendem Kommando (siehe Listing 7.2) eine CSV-Datei generiert werden. Dieses Kommando muss im Wurzelverzeichnis des Projekts ausgeführt werden und akzeptiert drei Argumente in folgender Reihenfolge:

- der **Verzeichnisname** bzw. Pfad zum Verzeichnis mit den Trainingsdaten
- der **Dateiname** für die generierte Datei
- die **Analysemethode**, die zum Generieren verwendet werden soll

## Generierung des Entscheidungsbaums

Die im vorherigen Abschnitt generierte CSV-Datei mit den Trainingsdaten kann nun in RapidMiner importiert werden. Aus dieser kann dann schlussendlich mit folgenden Schritten ein Entscheidungsbaum generiert werden:

1. Nachdem das Programm geöffnet worden ist, kann über das Hauptmenü File -> Import Data -> Import CSV File die generierte Trainingsdaten-Datei als Datenresource importiert werden.
2. Dafür wählt man nun die entsprechende Datei im Dialog aus und folgt anschließend dem Dialog bis Schritt 4 (bei den Schritten 2 und 3 mussten für den Prototypen keine Änderungen vorgenommen werden).
3. In Schritt 4 ist es wichtig, für die Resultat-Spalte (jene Spalte in, welcher die tatsächlichen Verkehrsmittel stehen) als Datentyp “text“ und als allgemeinen Typ “label“ auszuwählen.
4. Im nächsten Schritt kann man diese importierten Dateien zur Verwendung in RapidMiner unter dem gewünschten Namen ablegen.
5. Für die Generierung des Entscheidungsbaums wird als Erstes ein neuer Prozess benötigt. Diesen kann man über das Hauptmenü anlegen (File -> New Process).
6. Als Nächstes kann man die zuvor importieren Daten aus dem Repository-Bereich (links unten) zum neuen Prozess per Drag-and-Drop hinzufügen.
7. Im nächsten Schritt zieht man den “Decision-Tree“-Operator aus dem Operatoren-Bereich (links oben) in den Prozess. Diesen verknüpft man nun mit den Daten (out/tra) sowie mit dem Ende des Prozesses (mod/res).

Jetzt kann man mit Hilfe des “Run“-Buttons (blaues Dreick) unterhalb des Hauptmenüs den Prozess ausführen und bekommt den Entscheidungsbaum sowohl in grafischer als

## *Anhang 2*

auch textueller Darstellung zu sehen. Eine etwas ausführlichere Anleitung ist auch auf Youtube unter <https://www.youtube.com/watch?v=Vf6G1HNdBoI> zu finden.

Der Entscheidungsbaum in Textform kann nunmehr in einer Datei im Projekt abgelegt werden (definiert in der Konfiguration). Beim nächsten Analyseprozess wird erkannt, dass es sich um eine geänderte/neue Datei handelt und automatisch eine neue PHP-Datei mit dem Baum als PHP-Code generiert (siehe Abschnitt 5.5.1 Erstellen der Entscheidungsbäume).