Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommunikation UVEK
Département fédéral de l'environnement, des transports, de l'énergie et de la communication DETEC
Dipartimento federale dell'ambiente, dei trasporti, dell'energia e delle comunicazioni DATEC

Bundesamt für Strassen
Office fédéral des routes
Ufficio federale delle Strade

Neue Erkenntnisse zum Mobilitätsverhalten dank Data Mining?

De nouvelles découvertes sur le comportement de mobilité par Data Mining?

New findings on the mobility behavior through Data Mining?

büro widmer ag

Paul Widmer

Thomas Buhl

Institut für Datenanalyse und Prozessdesign (idp), zhaw

Andreas Ruckstuhl

Markus Dettling

Sina Rüeger

Forschungsauftrag SVI 2004/014 auf Antrag der Schweizerischen Vereinigung der Verkehrsingenieure und Verkehrsexperten (SVI)

Mai 2011 **1365**

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommunikation UVEK
Département fédéral de l'environnement, des transports, de l'énergie et de la communication DETEC
Dipartimento federale dell'ambiente, dei trasporti, dell'energia e delle comunicazioni DATEC

Bundesamt für Strassen
Office fédéral des routes
Ufficio federale delle Strade

Neue Erkenntnisse zum Mobilitätsverhalten dank Data Mining?

De nouvelles découvertes sur le comportement de mobilité par Data Mining?

New findings on the mobility behavior through Data Mining?

büro widmer ag

Paul Widmer

Thomas Buhl

Institut für Datenanalyse und Prozessdesign (idp), zhaw

Andreas Ruckstuhl

Markus Dettling

Sina Rüeger

Forschungsauftrag SVI 2004/014 auf Antrag der Schweizerischen Vereinigung der Verkehrsingenieure und Verkehrsexperten (SVI)

Mai 2011 **1365**

Impressum

Forschungsstelle und Projektteam

Projektleitung
Paul Widmer

Mitglieder
Thomas Buhl
Marcel Dettling
Andreas Ruckstuhl
Sina Rüeger

Begleitkommission

Präsident
Kay Axhausen

Mitglieder
Timo Ohnmacht
Guido Rindsfüser

Antragsteller

Schweizerische Vereinigung der Verkehrsingenieure und Verkehrsexperten (SVI)

Bezugsquelle

Das Dokument kann kostenlos von http://partnershop.vss.ch heruntergeladen werden.

Inhaltsverzeichnis

Zusammenfassung

Unter Data Mining versteht man im engeren Sinn das systematische (in der Regel automatisierte oder halbautomatisierte) Entdecken und Extrahieren von vorher unbekannten statistischen Informationszusammenhängen aus grossen Datenmengen. Im deutschen Sprachgebrauch steht Data Mining oft für den ganzen Analyse-Prozess, der auch die Vorbereitung der Daten sowie die Bewertung der Resultate umfasst. Data Mining wird in verschiedenen Bereichen erfolgreich eingesetzt. Anwendungsbeispiele aus schweizerischen Verkehrsplanungen fehlen aber bisher weitgehend. Ziel der Forschungsarbeit war es, den praktisch tätigen Verkehrsplaner mit dem Prozess und den Methoden von Data Mining vertraut zu machen und die Möglichkeiten von Data Mining als Hilfsmittel in der Verkehrsplanung auszuloten.

Data Mining wird als iterativer, lernender Prozess dargestellt, in welchem die Phasen von der Fragestellung über das Sammeln und Aufbereiten der Daten, die Modellierung und die Auswertung der Ergebnisse bis zu deren Umsetzung in die Praxis mehrfach durchlaufen werden. In dieser Arbeit wird dieser Prozess genauer beschrieben und ein Überblick über eine Auswahl von Methoden, die in der Modellierung verwendet werden, gegeben.

Beispiele aus der Literatur illustrieren das breite Anwendungsspektrum von Data Mining in der Verkehrsplanung (z.B. Verkehrserzeugung, Autobesitz, Verkehrsmittel- und Routenwahl oder Klassifikation von Mobilitätsmustern). Bei den beschriebenen Beispielen handelt es sich um Forschungsarbeiten. Deren Ergebnisse haben noch kaum breiten Eingang in die Praxis gefunden.

An Fallbeispielen wird demonstriert, wie Data Mining in der Praxis angewendet werden kann. Als Datensätze werden der Mikrozensus Verkehr 2005 und Raumstrukturdaten des Bundesamtes für Statistik verwendet. Die Fallbeispiele behandeln die Analyse der Häufigkeit von Wegeketten, die Vorhersage der Anzahl Wegeketten pro Person und Tag, die Klassifikation nach Mobilitätstypen sowie die Vorhersage des Mobilitätstyps einer Person aufgrund sozio-demographischer Merkmale und Raumstrukturinformationen zu den Wohn- und Zielorten.

Aus der grossen Vielfalt von Software-Lösungen für Data Mining wird eine Auswahl proprietärer und frei verfügbarer Pakete, welche für den Einsatz in der Verkehrsplanung als grundsätzlich geeignet beurteilt werden, grob und ohne Wertung beschrieben.

Die Studie kommt zum Schluss, dass Data Mining in der Verkehrsplanung sicher nutzbringend anwendbar ist, dass aber nicht – wie vielleicht erhofft – automatisch auf alle Fragen gute Antworten erwartet oder ohne Dazutun des Anwenders aus vorhandenen Datensätzen neue Erkenntnisse gewonnen werden können.

Empfehlenswerte Einsatzgebiete für Data Mining in der Verkehrsplanung sind beispielsweise:

- Klassifikation, z.B. des Mobilitätsverhaltens
- Visualisierung komplexer mehrdimensionaler Datensätze zum raschen Erkennen von Mustern resp. Clustern
- Rasche und automatische Erkennung der (aus statistischer Sicht) wichtigsten Prädikatoren des Mobilitätsverhaltens
- Analyse der Entscheidungsprozesse, z.B. bei der Verkehrsteilnahme.

Zusammenhänge, die mit Data Mining Methoden extrahiert werden, sind grundsätzlich Daten-getrieben und müssen keine Kausalitäten widerspiegeln. Deshalb wird empfohlen, aus Kausalitätsüberlegungen abgeleitete Modelle weiterhin mit statistischen Methoden an die Daten anzupassen. Konventionelle Modellansätze und Data Mining sollen als sich ergänzende und gegenseitig unterstützende Methoden eingesetzt werden.

Um Data Mining zukünftig auch in der Verkehrsplanung nutzbringend einsetzen zu können, bedarf es keiner weiteren Forschung. Vielmehr sind möglichst viele praktische Anwendungen erwünscht, mit denen Verkehrsplaner und Data Mining Experten in interdisziplinärer Zusammenarbeit Erfahrungen sammeln und weitergeben können.

Résumé

Le "Data Mining" (aussi appelé exploration ou fouille de données, extraction de connaissances) est un procédé utilisé avec succès dans différents domaines afin de découvrir des types caractérisés à partir de très nombreuses données. Des exemples d'application provenant de planifications des transports en Suisse manquent jusqu'à présent largement. Le but de la recherche était de familiariser le planificateur en transports dans sa pratique avec le procédé et les méthodes de "Data Mining" et de sonder les possibilités du "Data Mining" comme aide à la planification des transports.

Le "Data Mining" se présente comme un procédé d'apprentissage itératif dans lequel les phases sont parcourues plusieurs fois, du questionnement à la récolte et la préparation des données, de la modélisation et du dépouillement des résultats à leur mise en œuvre pratique. Pour donner une vue d'ensemble, une sélection des méthodes souvent appliquées est décrite en distinguant entre celles de l'apprentissage "surveillé" et "non surveillé".

Des exemples issus de la bibliographie illustrent le large éventail d'applications du "Data Mining" dans la planification des transports (par exemple génération de trafic, possession d'auto, choix du moyen de transport et de l'itinéraire ou classification de types de mobilité). Les exemples décrits proviennent de travaux de recherche. Leurs résultats n'ont pas encore vraiment trouvé leur place dans la pratique.

A l'aide d'exemples caractéristiques, il est démontré comment le "Data Mining" peut être mis en oeuvre dans la pratique. Les données utilisées proviennent du micro-recensement sur le comportement en matière de transports de 2005 et des données structurelles territoriales de l'Office fédéral du développement territorial. Les exemples typiques traitent de l'analyse de la fréquence des chaînes de déplacement, de la prévision du nombre de chaînes de déplacement par personne et jour, de la classification selon les types de mobilité ainsi que des prévisions du type de mobilité d'une personne sur la base de ses caractéristiques sociodémographiques et des informations structurelles territoriales des lieux de domicile et de destination.

Parmi la grande variété de solutions logicielles pour le "Data Mining", un choix d'ensembles propriétaires et disponibles gratuitement, qui sont en principe jugés appropriés pour une utilisation dans la planification des transports, est présenté sommairement et sans évaluation.

L'étude arrive à la conclusion que le "Data Mining" est certainement applicable utilement dans la planification des transports mais qu'il ne faut pas s'attendre – comme peut-être espéré – à de bonnes réponses ou que de nouvelles connaissances puissent émerger des données existantes sans intervention de l'utilisateur. Les domaines recommandés de recours au "Data Mining" dans la planification des transports sont par exemple:

- Classifier, par exemple les comportements en matière de mobilité
- Visualiser des données complexes et pluridimensionnelles pour rapidement reconnaître des types respectivement des groupes
- Reconnaître rapidement et automatiquement les variables déterminantes importantes (du point de vue statistique) du comportement en matière de mobilité
- Analyser les procédures de décision, par exemple en matière de déplacements.

Comme ni les règles de classification ni les méthodes pour reconnaître automatiquement les variables déterminantes statistiquement significatives ne tiennent compte des causalités entre ces dernières et les variables visées, les méthodes de "Data Mining" ne peuvent pas remplacer les méthodes conventionnelles pour l'estimation de modèles. Il est bien plutôt recommandé de recourir à des approches conventionnelles de modélisation et au "Data Mining" en les considérants comme méthodes se complétant et se soutenant réciproquement.

Afin de pouvoir recourir utilement au "Data Mining" à l'avenir aussi dans la planification des transports, il n'est pas nécessaire d'entreprendre de nouvelles recherches. Sont bien plutôt souhaitées des applications pratiques les plus nombreuses possibles par lesquelles les planificateurs en transports et les experts en "Data Mining" pourront rassembler des expériences en collaboration interdisciplinaire et les transmettre plus loin.

Summary

Data mining as a process for revealing yet unknown patterns from large amounts of data has been employed successfully in various fields. Application examples from Swiss transport planning however are largely missing so far. The aim of this research paper is to familiarize the practicing transportation planner with both process and methods of data mining and to explore the possibilities of data mining as an aid in transport planning.

Data mining is commonly perceived as an iterative learning process, in which all actions such as problem formulation, data collection, data preparation, modeling and evaluation are run through several times and finally completed by the deployment step. In this work, this process is described in some detail and an overview of a selection of methods used in the modeling is given.

A literature review illustrates the wide range of applications data mining is used for in transport planning. These include traffic generation, car ownership, mode and route choice or classification of mobility patterns. Because they mostly result from research activities, their outcomes have not yet found wide acceptance in practice.

With case studies we demonstrate how data mining can be successfully applied in a transportation planner's daily routine. The "Micro Census on Travel Behavior 2005" and the data on spatial structure of the Federal Office for Spatial Development (ARE) are used. The case studies encompass a frequency analysis of trip chaining, predicting the number of trip chains per person per day, the segmentation of people into different mobility behavior types and predicting these for a person based on their socio-demographic characteristics and spatial structure information of home and destination locations.

We continue with a review of a selection of data mining software, including both proprietary and freely available tools. All the reviewed packages are basically regarded as suitable for the use in transport planning but are not evaluated further.

As a conclusion of our research we believe that data mining can be beneficial to and yield a substantial contribution in quantitative transportation research. However, it does not – as might be hoped for by non-professionals – automatically generate sensible answers to some vague questions floating around. It still needs a clear aim and the input of an expert user to extract new insights from existing data sets.

Recommended application areas for data mining in transportation planning include:

- Classification, e.g. of the mobility behavior
- Visualization of complex multidimensional data sets for a simple identification of patterns and clusters, respectively.
- Quick and automated detection of (from a statistical point of view) most important predictors of mobility behavior
- Analysis of decision processes, e.g. in the context of traffic participation.

It is also important to keep in mind that any association that is extracted with data mining methods is inherently data driven and might not reflect causalities. We therefore recommend continuing fitting models that arise from causality considerations by traditional statistical methods. Conventional modeling approaches and data mining may be used complementary or as mutually supportive methods.

To exploit the potential of data mining in transportation planning, we think that no further research is needed. But rather a variety of practical application cases should be carried out, where transportation engineers and data mining experts can collect and share experiences in such interdisciplinary collaborations.

1 Einleitung

1.1 Ausgangslage

Wie in vielen anderen Fachgebieten werden auch im Verkehrswesen umfangreiche Datenmengen erhoben und in Datenbanken abgelegt. Das Design der Datenerhebungen erfolgt in der Regel vor dem Hintergrund der zu beantwortenden Fragestellungen sowie gestützt auf Hypothesen zu den interessierenden Zusammenhängen. Für Verkehrsplanungen sind vor allem Daten zum Mobilitätsverhalten wichtig. Solche werden beispielsweise gemeinsam vom BFS und ARE mit den in ca. 5-jährigem Zyklus durchgeführten Mikrozensen zum Verkehrsverhalten und von den SBB im Rahmen der kontinuierlichen Erhebungen des Personenverkehrs (KEP) erhoben. Im Rahmen von Forschungsarbeiten und für die Etablierung von Verkehrsmodellen werden Befragungen von Haushalten und Einzelpersonen zum tatsächlichen (Revealed Preference Erhebungen) oder zum antizipierten Verkehrsverhalten (Stated Preference Erhebungen) durchgeführt. Diese Daten bilden die Grundlage für statistische Auswertungen und Modellbildungen zur Beschreibung des Verkehrsverhaltens.

Data Mining als Prozess zur Entdeckung neuer Muster aus umfangreichen Datenmengen wird in der Schweiz in verschiedenen Branchen – vor allem zur Analyse von Kunden- und Warenkorbdaten – erfolgreich praktiziert. Anwendungsbeispiele aus schweizerischen Verkehrsplanungen fehlen aber bisher weitgehend. Mit ein Grund dürfte sein, dass der Verkehrsingenieur mit Data Mining noch wenig vertraut ist. Es fehlen Erfahrungen, ob und wie sich Data Mining nutzbringend in der Verkehrsplanung einsetzen lässt.

Ziel dieser Forschungsarbeit ist es, den praktisch tätigen Verkehrsingenieur mit dem Prozess und den Methoden von Data Mining vertraut zu machen und die Möglichkeiten und Grenzen von Data Mining als Hilfsmittel in der Verkehrsplanung auszuloten.

1.2 Aufbau des Berichtes

Das folgende Kapitel gibt einen einführenden Überblick über und den Prozess von Data Mining und die häufig zur Anwendung gelangenden Methoden.

Im dritten Kapitel wird eine Auswahl von Beispielen aus der Literatur zur Anwendung von Data Mining im Kontext verkehrsplanerischer Fragestellungen beschrieben.

Das vierte Kapitel ist der Illustration einfacher Data Mining-Anwendungen an Fallbeispielen gewidmet.

Aus der Vielfalt verfügbarer Software für Data Mining wird im fünften Kapitel eine Auswahl beschrieben. Dabei handelt es sich weder um eine abschliessende Beurteilung noch um eine Empfehlung der verschiedenen Produkte.

Im sechsten Kapitel folgen die Schlussfolgerungen aus der Sicht des Projektteams zur Anwendbarkeit von Data Mining in der Verkehrsplanung sowie Hinweise und Empfehlungen für den praktischen Einsatz durch den Verkehrsingenieur.

2 Was ist Data Mining?

2.1 Data Mining – Eine erste Umschreibung

Unter Data Mining versteht man im engeren Sinn das systematische (in der Regel automatisierte oder halbautomatisierte) Entdecken und Extrahieren von vorher unbekannten statistischen Informationszusammenhängen aus grossen Datenmengen. Die Datenbestände werden dabei nach Regelmässigkeiten, Mustern und Strukturen, Abweichungen jeglicher Art von Beziehungen und gegenseitigen Beeinflussungen untersucht. Im deutschen Sprachgebrauch steht Data Mining oft für den ganzen Analyse-Prozess, der auch die Vorbereitung der Daten sowie die Bewertung der Resultate umfasst[1]. Das folgende Zitat eines Schweizer Data Mining Anbieters[2] illustriert diese Sichtweise: *"Data Mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately comprehensible patterns or models in data to make crucial decisions. Data Mining is not a product that can be bought. Data Mining is a discipline that must be mastered."* In dieser Arbeit verwenden wir den Begriff "Data Mining" in der umfassenderen zweiten Bedeutung.

Data Mining hat grosse Ähnlichkeiten zur statistischen Datenanalyse und man kann versucht sein, irgendwelche Abgrenzungen zu finden. Wir wollen diesen Weg nicht weiter verfolgen. Stattdessen soll noch auf ein weiteres typisches Merkmal hingewiesen werden. Die Daten, die mit dem Data Mining Prozess analysiert werden, wurden im Gegensatz zu klassischen statistischen Analysen oft nicht im Wissen um die jetzt interessierende Fragestellung (Ziele der Wissensfindung) gesammelt, sondern aus verschiedenen anderen guten Gründen. Wie weit sich die Resultate auf eine ins Auge gefasste Zielpopulation verallgemeinern lassen, muss deshalb sorgfältig abgeklärt werden. Auf jeden Fall kann der Data Mining Prozess im wissenschaftlichen Umfeld erfolgreich zur Generierung von Arbeitshypothesen eingesetzt werden. In Unternehmen müssen die Resultate nutzbringend eingesetzt werden können, ansonsten Data Mining hier keine Existenzberechtigung hat.

Wissenschaftliche Arbeiten zu Data Mining legen den Fokus zumeist auf Data Mining Algorithmen. Bevor die Algorithmen auf Daten angewendet werden können, müssen diese aufbereitet werden. Die Auswahl, Vorverarbeitung und Aufbereitung der Daten beanspruchen einen grossen Anteil der Gesamtanstrengungen im Data Mining Prozess und haben einen entscheidenden Einfluss auf die Qualität des Gesamtergebnisses. Damit Data Mining in Unternehmen nachhaltig integriert werden kann, muss die praktische Umsetzung der Ergebnisse in das betriebliche Umfeld sorgfältig geplant und begleitet werden. Folglich rückt in der praktischen Anwendung der ganze Data Mining Prozess in den Vordergrund.

Es gibt viele Bücher und Fachartikel zu den Themen "Data Mining" und "Knowledge Discovery in Database (KDD)". Eine Einführung in diese Themengebiete findet sich in Gabriel, Gluchowski und Pastwa (2009). Runkler (2009) gibt einen einführenden Überblick über die im klassischen Data Mining verwendeten Methoden. Han und Kamber (2006) führen umfassend in das Thema ein. Das Buch von Nisbet, Elder und Miner (2009) ist als Referenz-Buch für Data Mining Anwender konzipiert. Daneben gibt es unzählige Texte, die sich vor allem mit den methodischen Aspekten auseinandersetzen, wie z.B. Bishop (2007), Duda, Hart und Stork (2000), Hastie, Tibshirani und Friedman (2009), Ye (2003) sowie Witten und Frank (2005).

[1] Diese umfassende Sicht wird auch von Forschern der Künstlichen Intelligenz mit Knowledge Discovery in Database (KDD) bezeichnet.

[2] Statoo, “What is Data Mining”, http://www.statoo.com/en/datamining/

2.2 Phasen des Data Mining Prozesses und ihre Bedeutung

2.2.1 Überblick

Wie oben im Abschnitt 2.1 dargelegt, beschreibt Data Mining im umfassenderen Sinn einen Datenanalyse-Prozess. Dieser Prozess wurde verschiedentlich, jedoch jeweils ähnlich, formalisiert. Stellvertretend wird hier der "klassische" Data Mining Prozess von CRISP (Cross Industry Standard Process for Data Mining[1], siehe Abbildung 1) näher erläutert.

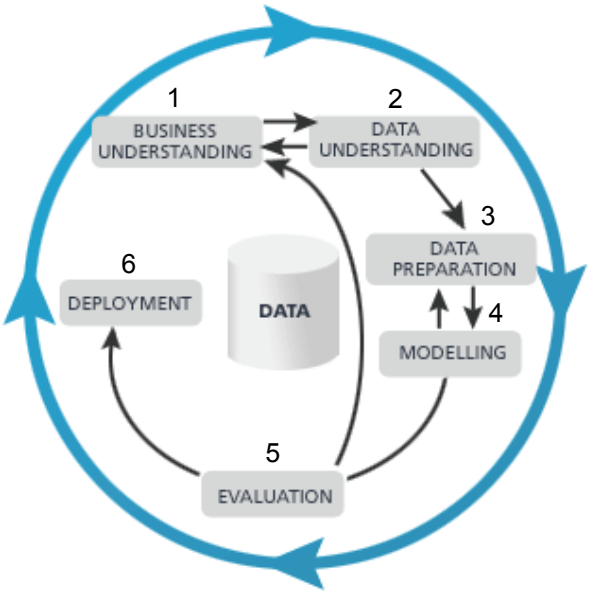

Abbildung 1: Der Data Mining Prozess nach CRISP; Quelle: http://www.crisp-dm.org/

Zentral ist das Verständnis von Data Mining als *iterativer Prozess*, in welchem die Phasen von der Idee/Fragestellung über das Sammeln/Beschaffen/Aufbereiten der Daten, die Modellierung und die Auswertung der Ergebnisse bis zu deren Umsetzung in die Praxis mehrfach durchlaufen werden. In diesem "lernenden" Prozess fliessen bei jedem Iterationsschritt die bis dahin gewonnen Erkenntnisse ein. Der Iterationsprozess endet, wenn sich die Fragestellung mit den gewonnenen Erkenntnissen beantworten lässt. Wie das Schema in Abbildung 1 deutlich zeigt, stehen die Daten im Zentrum dieses Prozesses. Mit diesem Fokus beschreibt Data Mining einen Prozess für das Daten getriebene Lernen.

2.2.2 Beschreibung der Phasen

Nach dieser Übersicht wird nun jede der sechs in Abbildung 1 dargestellten Phasen des iterativen Prozesses eines Data-Mining-Projekts genauer erläutert.

1) Business Understanding
Als erstes gilt es, die Ziele und Fragestellungen eines Data Mining Projekts genau zu definieren. Generell formulierte Absichten werden in die Welt der quantitativen Analyse übersetzt. Es wird festgelegt, womit und für wen die Resultate und Antworten dargestellt werden sollen. Überlegungen zur Verfügbarkeit von finanziellen, technischen und personellen Ressourcen sind anzustellen.

2) Data Unterstanding
Nun gilt es sicherzustellen, dass die nötigen Daten verfügbar sind. Rechtliche, finanzielle, quantitative und qualitative Aspekte sind abzuklären. Eine weitere Überlegung ist, welche Menge an zusätzlicher Arbeit erforderlich ist, um die Daten in eine Form zu bringen, in der sie mit den Methoden des Data Mining analysiert werden können.

[1] Details dazu finden sich auf http://www.crisp-dm.org/

3) Data Preparation
Die Daten müssen in "brauchbare" Form gebracht werden. Der Umgang mit fehlenden Werten ist zu klären. Mit Standardisierungen, Transformationen, Aggregierungen und Vorselektionen von Variablen sind die Daten in eine anwendbare Form zu bringen. Bei sehr umfangreichen Datensätzen kann auch eine Datenreduktion, z.B. durch Sampling, hilfreich sein.

4) Modeling
Hier werden die für die gegebene Fragestellung geeigneten analytischen Verfahren des Data Mining ausgewählt und an die Daten angepasst. Danach werden mit ihnen Vorhersagen erzeugt, Gruppierungen erstellt, aussagekräftige Grafiken kreiert usw.

5) Evaluation
Eine wichtige Frage ist nun, wie gut die angewandten Modelle resp. deren Resultate sind. Einerseits im Hinblick auf technische Betrachtungen wie Residuenanalyse oder (statistischer) Validierung, andererseits qualitativ im Rahmen der Beantwortung der gestellten Fragen und gesteckten Ziele. Zudem kann es möglicherweise weitere interessante Erkenntnisse aus dem datenanalytischen Teil geben.

6) Deployment
Die Umsetzung in die Praxis: Es ist zu überlegen, wie die Resultate kommuniziert, welche Massnahmen vorgeschlagen und wie diese umgesetzt werden. Begleitmassnahmen sind abzuklären. Weiter gilt es, Folgefragen und weitere Entwicklungen zu studieren.

2.3 Data Mining Methoden

Die Analysemethoden, die im Data Mining eingesetzt werden, stammen vielfach aus den Gebieten der (klassischen) Statistik und des maschinellen Lernens, einem Teilgebiet der theoretischen Informatik. Das zweite Teilgebiet wird oft auch als Pattern Recognition bezeichnet. In beiden Gebieten, maschinelles Lernen und Statistik, stehen für Data Mining ähnliche oder gleiche Methoden im Fokus. Die Herangehensweise und die Terminologie sind jedoch unterschiedlich. Wir nehmen hier eher einen statistischen Blickwinkel ein.

Die analytischen Methoden des Data Mining lassen sich in zwei Hauptgruppen unterteilen. Die erste Hauptgruppe kann mit "**überwachtem Lernen**" (bzw. „strukturüberprüfende[1] Verfahren“) überschrieben werden. Das Ziel dieser Methoden ist es, für jedes Objekt (oder jede Beobachtungseinheit) eines der Merkmale aus den anderen vorauszuberechnen. Die Vorausberechnung hat den Charakter einer Vorhersage. Sie muss jedoch nicht zwangsläufig im Sinne einer zeitlichen Vorhersage wie z.B. eine Wetterprognose sein, sondern es genügt, wenn ein bestehender Wert des Merkmals vorhergesagt wird, der sich zurzeit unserer Kenntnis entzieht. In der Fachsprache wird diese Hauptgruppe auch mit "supervised learning", "supervised pattern recognition" oder neuerdings mit "predictive Modeling"[2] umschrieben.

In der zweiten Hauptgruppe werden Verfahren zusammengefasst, die unbekannte Strukturen in den Daten aufdecken sollen. Beim Suchen geht man davon aus, dass man gar nichts über die zu suchende Struktur weiss, höchstens vermutet, dass es Strukturen irgendwelcher Art geben müsste. Deshalb wird diese Gruppe mit "**unüberwachtem Lernen**" (im englischen "unsupervised learning", "unsupervised pattern recognition") bzw. strukturentdeckende Verfahren bezeichnet.

[1] Strukturüberprüfende Verfahren werden oft im Zusammenhang mit konfirmatorischer bzw. induktiver Statistik gesehen. Allerdings stellen im Data Mining im Gegensatz zur induktiven Statistik nicht im Vorfeld formulierte und zu überprüfende Hypothesen den Ausgangspunkt der Betrachtungen dar.

[2] Im Fachenglisch wird das zeitliche Vorhersagen mit forecast umschrieben, wogegen predict im allgemeineren Sinn eingesetzt wird. Folglich steht forecast für eine (zeitliche) Extrapolation, wogegen predict mehr für eine Vorhersage aufgrund von (statistisch) einflussnehmenden Variablen steht (d.h. es handelt sich vorwiegend um eine Interpolation innerhalb des Merkmalsraumes).

2.3.1 Einige Algorithmen für Data Mining im Überblick

In diesem Abschnitt werden einige Algorithmen (oder besser Familien von Algorithmen) kurz vorgestellt, die Ihren Ursprung in der KI-Forschung (KI steht für künstliche Intelligenz) und im Gebiet des maschinellen Lernens haben.

a) Künstliche neuronale Netze

Künstliche neuronale Netze (engl. artificial neural networks) sind Netze aus künstlichen Neuronen. Der Ursprung der künstlichen neuronalen Netze liegt in der Forschung über künstliche Intelligenz. Man stellt sie den natürlichen neuronalen Netzen gegenüber, welche Nervenzellvernetzungen im Gehirn und im Rückenmark bilden. Bei künstlichen neuronalen Netzen geht es aber hauptsächlich um eine Abstraktion (Modellbildung) von Informationsverarbeitung und weniger um das Nachbilden biologischer neuronaler Netze.

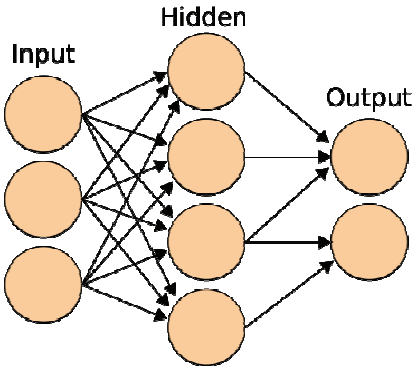

Abbildung 2: Einfaches neuronales Netz mit einem hidden Layer

Die Eingabe-Knoten (input) repräsentieren die Merkmale eines Objekts, der Ausgabe-Knoten (output) eine vorherzusagende Grösse des Objekts. Die Pfeile repräsentieren Gewichte. Das gewichtete Mittel der Werte in den Knoten der vorhergehenden Schicht (engl. layer) wird über eine Aktivierungsfunktion (meist die logistische Funktion) in den Wert des Zielknotens überführt. Die Ausgabe kann ein numerischer Wert oder eine kategorielle Grösse wie die Klassenzugehörigkeit sein. Auch die Eingabe-Grössen können sowohl numerisch als auch kategoriell sein. Die Gewichte werden mittels einer Lernstichprobe durch so genannte backpropagation für spezifische Situationen bestimmt. Wichtige Elemente, die beim Modellieren eines neuronalen Netzes festgelegt werden müssen, sind die Anzahl Knoten in den mittleren Schichten (hidden layer) und die Anzahl der mittleren Schichten. Weiter müssen für den Suchpfad die Einstellgrössen im Algorithmus geeignet gewählt werden.

Neben dem klassischen künstlichen neuronalen Netz, welches auch als **Multilayer-Perceptron** bezeichnet wird, gibt es weitere Typen von solchen Netzen:

- **Single Layer Perceptron**: Diese Netzwerke haben keine mittlere Schicht sondern nur eine Eingabe- und eine Ausgabe-Schicht. Falls als Aktivierungsfunktion die logistische Funktion gewählt wird, entspricht dieses Verfahren der **logistischen Regression**. Macht man alle Eingabe-Merkmale zu kategoriellen Grössen, wird dieses Netz (logistische Regression) zu einem einfachen, jedoch sehr mächtigen Klassifikationswerkzeug.
- Bei **selbstorganisierenden Karten** (Self-organizing maps – SOM), nach ihrem Entwickler auch **Kohonen-Karten** oder Kohonen Feature Maps genannt, handelt es sich um ein spezielles neuronales Netz, das - wie der Name schon sagt - sich selbst organisiert, also im Gegensatz zum Perceptron ohne Lernen auskommt und im unüberwachten Lernen eingesetzt wird. Es gruppiert iterativ die Objekte üblicherweise auf einem zwei-dimensionalen quadratischen Neuronengitter bis die totale Differenz zwischen den Gruppen maximiert ist. Da die euklidische Distanz verwendet wird, ist es wichtig, dass die Daten im Voraus geeignet skaliert werden.
- Ein **Bayesian Network** ist ein künstliches neuronales Netz (gerichteter Graph), bei dem die Knoten als Zufallsvariablen im Sinn von Bayes interpretiert werden. Das heisst, dass Knoten beobachtbare Grössen, latente Variablen oder unbekannte Para-

meter sein können. Die Kanten repräsentieren bedingte Wahrscheinlichkeiten. Mit diesen Vorstellungen lässt sich die Modellkomplexität besser steuern, aber auch Einstellgrössen für den Algorithmus können optimiert werden. Für jede Eingabe-Grösse erhält man zusätzlich eine Information über Ihre Wichtigkeit.

- **Probabilistic Networks** sind ebenfalls eine Form neuronaler Netze. Sie bestehen aus drei bis vier Schichten und erlauben nur kategorielle Zielgrössen, sie haben Ähnlichkeiten mit der k-nearest neighbour Methode.
- **Radial Basis Function (RBF)** Networks sind Netze mit drei Schichten. Im Gegensatz zu einem Multilayer-Perceptron benutzt ein RBF-Netz nicht die Rohdaten als Eingabegrössen sondern skalierte (euklidische) Abstände zu einem "Zentrum" der Merkmalsgrösse. Als Aktivierungsfunktion wird die Gauss'sche Glockenkurve verwendet. Der Vorteil liegt darin, dass man mit nur einer mittleren Schicht beliebige nicht lineare Funktionen approximieren kann.

b) Support Vector Machine (SVM)
Bei Support Vector Machines (SVM) handelt es sich nicht um Maschinen im herkömmlichen Sinne, sondern um ein rein mathematisches Verfahren zur Mustererkennung. Der Namensteil "Machine" weist auf das Herkunftsgebiet der Methode, das maschinelle Lernen, hin.

Ausgangsbasis für den Bau einer Support Vector Machine ist eine Menge von Trainingsobjekten, für die jeweils bekannt ist, welcher Klasse sie zugehören. Jedes Objekt wird durch einen Vektor von Objektmerkmalen in einem Vektorraum repräsentiert. Aufgabe der Support Vector Machine ist es, in diesen Raum eine Hyperebene einzupassen, die als Trennfläche fungiert und die Trainingsobjekte in zwei Klassen teilt. Der Abstand derjenigen Vektoren, die der Hyperebene am nächsten liegen, wird dabei maximiert. Dieser breite, leere Rand soll später dafür sorgen, dass auch Objekte, die nicht genau den Trainingsobjekten entsprechen, möglichst zuverlässig klassifiziert werden.

Beim Einsetzen der Hyperebene ist es nicht notwendig, alle Trainingsvektoren zu beachten. Vektoren, die weiter von der Hyperebene entfernt liegen und gewissermassen hinter einer Front anderer Vektoren "versteckt" sind, beeinflussen Lage und Position der Trennebene nicht. Die Hyperebene ist nur von den ihr am nächsten liegenden Vektoren abhängig – und auch nur diese werden benötigt, um die Ebene mathematisch exakt zu beschreiben. Diese nächstliegenden Vektoren werden nach ihrer Funktion Stützvektoren (engl. support vectors) genannt und verhalfen den Support Vector Machines zu ihrem Namen.

Eine saubere Trennung mit einer Hyperebene ist nur dann möglich, wenn durch den breiten, leeren Rand eine solche gelegt werden kann. Dies ist in realen Anwendungsfällen im Allgemeinen nicht der Fall. Support Vector Machines verwenden deshalb den so genannten "Kernel Trick" um eine nicht lineare Klassengrenze einzuziehen. Die Idee hinter dem "Kernel Trick" ist, den Vektorraum und damit auch die darin befindlichen Trainingsvektoren in einen höher dimensionalen Raum zu überführen. In einem Raum mit genügend hoher Dimensionsanzahl – im Zweifelsfall unendlich – wird auch die verschachteltste Vektormenge linear trennbar. In diesem höher dimensionalen Raum wird nun die trennende Hyperebene bestimmt. Bei der Rücktransformation in den ursprünglichen Raum wird die lineare Hyperebene zu einer nicht linearen, unter Umständen sogar nicht zusammenhängenden Hyperfläche, die die Trainingsvektoren sauber in zwei Klassen trennt. Technisch muss man allerdings dafür nicht in diese hochdimensionalen Räume gehen, sondern man ersetzt bei der Beschreibung der Trennfläche das Skalarprodukt durch geeignete so genannte Kernelfunktionen, wie z.B. die Gaussian radial basis function (RBF).

Dieser Ansatz kann erweitert werden, so dass SVM sowohl zur Lösung von Klassifizierungsproblemen als auch für regressionsartige Fragestellungen eingesetzt werden kann.

c) Klassifikations- und Regressionsbäume

Ein Klassifikationsbaum ist die Realisierung einer Entscheidungsfindung in Baumform. Zur Klassifikation "durchläuft" ein Objekt mit seinen Merkmalen einen (umgedrehten) Baum (vgl. Abbildung unten) von oben nach unten. An jeder Verzweigung steuern Bedingungen an die Merkmale eines Objekts in welchen Ast es geleitet wird. Die geschätzte Klassenzugehörigkeit des Objekts wird durch das Blatt – den Endpunkt im Baum –, in dem das Objekt schliesslich ankommt, bestimmt.

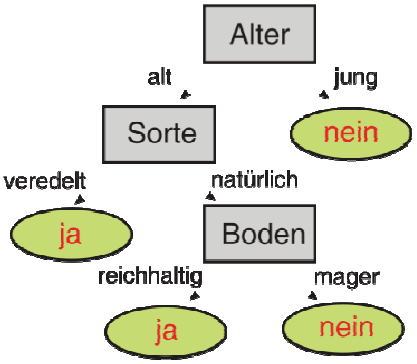

Abbildung 3: Einfacher Entscheidungsbaum zur Vorhersage, ob ein Apfelbaum Früchte tragen wird.[1]

Die Praxis unterscheidet verschiedene Baumtypen. Am bekanntesten sind **C&RT**[2] (Classification And Regression Trees) und **CHAID** (Chi-square Automatic Interaction Detectors). Ein bedeutendes Merkmal des C&RT-Algorithmus ist, dass nur Binärbäume erzeugt werden können, das heisst, dass an jeder Verzweigung immer genau zwei Äste vorhanden sind. Das zentrale Element dieses Algorithmus ist also das Finden einer optimalen binären Trennung. Die Merkmalsauswahl wird durch die Maximierung des Informationsgehalts gesteuert, welcher durch Masse wie Entropie oder Gini-Index gemessen wird.

Im Gegensatz zu C&RT erlaubt CHAID nicht nur eine binäre sondern eine Verzweigung mit beliebig vielen Ästen. Der Baum wird breiter. Er ist meist weniger tief als der korrespondierende C&RT-Baum. Dafür werden nach der ersten Aufteilung die nachfolgenden Aufsplittungen weniger bedeutungsvoll. Neben CHAID gewinnt **C4.5**[3] immer mehr an Bedeutung, welches ebenfalls eine beliebige Anzahl von Verzweigungen nach einer Entscheidung erlaubt. Im Gegensatz zu CHAID, bei dem das Teilungskriterium auf der Chi-Quadrat-Statistik beruht, wird im C4.5 wie im C&RT ein informationstheoretischer Ansatz zur Teilung eingesetzt. Ein weiterer Unterschied von CHAID zu C&RT und C4.5 besteht darin, dass der CHAID-Algorithmus das Wachsen des Baumes stoppt, bevor der Baum zu gross geworden ist. Der Baum wird also nicht beliebig wachsen gelassen, um ihn hinterher mit einer sogenannten Pruning-Methode wieder zu stutzen. C&RT erzeugt bei diesem Stutzen einige Unterbäume und testet diese mit neuen, vorher noch nicht klassifizierten Daten. C4.5 hingegen beschneidet den Baum ohne Beachtung der gegebenen Datenbasis. Ein weiterer Unterschied besteht darin, dass CHAID mit kategorial skalierten Variablen wie Farbe (rot, gelb, grün) oder Bewertung (gut, mittel, schlecht) arbeitet anstatt mit metrisch skalierten Variablen wie zum Beispiel Körpergewicht in kg.

Die Klassifikationsbäume sind sehr beliebt, weil sie im Vergleich z.B. zu künstlichen neuronalen Netzen für die Interpretation einfacher zugänglich sind. Leider sind sie aber recht sensitiv auf kleine Änderungen in den Merkmalen. Neuere Verfahren wie Random Forest und TreeNet (mehr dazu unter Ensemble-Techniken) haben diesbezüglich Abhilfe geschaffen.

[1] Quelle: http://de.wikipedia.org/wiki/Entscheidungsbaum

[2] C&RT bezeichnet die generische Version. Das Akronym CART steht für einen proprietären Algorithmus.

[3] C4.5 gilt als Nachfolger des ID3-(Iterative Dichotmiser 3)Algorithmus.

d) Ensemble-Techniken
Eine Methode, um die Klassifikationsgüte beim Einsatz eines Klassifikationsalgorithmus – vor allem eines einfachen, jedoch schwachen Algorithmus – zu steigern, ist der Einsatz einer Vielzahl von solchen Klassifikationsalgorithmen, eines so genannten Ensembles. Handelt es sich beim Klassifikationsalgorithmus um Klassifikations- und Regressionsbäume, werden diese Mengen von Bäumen auch als Wälder bezeichnet.

Die Idee hinter der Ensemble-Technik ist, dass ein einzelner Klassifikationsalgorithmus zwar nicht immer eine optimale Klassifikation liefern muss, die Mehrheitsentscheidung einer Vielzahl von solchen Klassifikationsalgorithmen dies aber sehr wohl leisten kann. Verbreitete Methoden zur Erzeugung geeigneter Ensembles sind Boosting und Bagging. Neben der Klassifikation kann diese Technik auch in der Regression eingesetzt werden. Nachteil z.B. bei Entscheidungswäldern ist, dass es nicht mehr so einfach ist, die in allen Bäumen enthaltenen Regeln in ihrer Gesamtheit in einfacher Weise zu interpretieren.

Bei **Boosting** wird nach einer ersten Klassifizierung mit einer vorgegebenen Methode eine zweite, dritte, … mit derselben Methode durchgeführt, wobei jedes Mal die fehlklassifizierten Objekte ein zusätzliches Gewicht bekommen. Am Schluss wird eine gewichtete Abstimmung unter den Klassifizierungen, die aus den verschiedenen Schritten kommen, durchgeführt. Der bekannteste Boosting-Algorithmus ist **AdaBoost** (hergeleitet aus adaptive boosting). Handelt es sich um einen einfachen Klassifikationsbaum mit nur einer Entscheidung, wird die Methode auch **TreeNet** genannt.

Das andere Verfahren, **Bagging** (das Wort leitet sich aus "bootstrap aggregation" ab), benutzt M zufällig (mit Zurücklegen) gezogene Teilstichproben von Objekten aus der Lernstichprobe und erstellt für jede Teilstichprobe einen Klassifizierer. Der Gesamtklassifizierer besteht dann aus einer Mehrheitsabstimmung unter den M Klassifizierern. Um bei Klassifikations- und Regressionsbäumen das Bagging noch effektiver zu machen, wird z.B. beim **Random Forest** eine weitere Stufe von Variabilität in den Klassifizierer hineingebracht. Beim Erstellen des jeweiligen Baumes stehen bei der Suche nach geeigneten Kassifikationen wiederum nicht alle Merkmale zur Verfügung, sondern – als zweite Zufallskomponente – nur eine zufällige Auswahl. Diese beiden Zufallskomponenten sorgen dafür, dass der Random Forest stabiler als ein einzelner Klassifikationsbaum und auch viel weniger zu Overfitting neigt. Der Einsatz des Random Forest ist ziemlich unproblematisch, da die Grundeinstellungen so robust gesetzt sind, dass er in vielen Fällen ohne weitere Feineinstellungen auf ein Problem angewandt werden kann. Das führt dazu, dass diese Methode zumindest halb, wenn nicht sogar ganz automatisch eingesetzt werden kann. In den Fallbeispielen (Kapitel 4) kommen wir auf diese Methode zurück.

e) K-Means und verwandte Methoden
Ein ***k*-Means-Algorithmus** ist ein Verfahren zur Cluster-Bildung. Dabei wird aus einer Menge von Objekten eine vorher festgelegte Anzahl von *k* Gruppen mit ähnlichen Objekten gebildet. Der Algorithmus ist eine der am häufigsten verwendeten Techniken zur Gruppierung von Objekten, da er schnell ist und sich skalieren lässt, d.h. für grosse Datensätze eingesetzt werden kann.

Die Grundversion des Algorithmus arbeitet mit numerischen Merkmalen der Objekte. Die (Un-)Ähnlichkeit zweier Objekte wird mit der euklidischen Distanz gemessen. Deshalb sucht der k-Means-Algorithmus stets (möglichst) sphärische Cluster. Ein weiterer beachtenswerter Punkt ist, dass die Lösung, die der Algorithmus liefert, von den Startpunkten (für die Clusterzentren) abhängt und demzufolge nicht zwingend die beste mögliche Lösung sein muss.

Beim **Fuzzy-Clustering** werden Objekte unscharf (also mit einem bestimmten Zugehörigkeitsgrad, oft als Wahrscheinlichkeiten ausgedrückt) auf Cluster verteilt. Setzt man diese Idee mit dem k-Means-Ansatz um, erhält man den **Fuzzy C-Means** (FCM) Algorithmus. Die Cluster werden dabei (implizit) durch eine sphärische multivariate Normalverteilungsdichte beschrieben. Es ist nun naheliegend, die multivariate Normalverteilungsdichten durch andere Dichten zu ersetzen, womit man bei der Idee des **EM-Algorithmus** angelangt ist. Grundvoraussetzung ist, dass sich alle Objekte als (Merkmals-)Vektoren darstellen lassen. Weiterhin muss eine Funktion bekannt sein, nach welcher der Mittelwert

zweier solcher Vektoren berechnet werden kann. Wie bei k-Means wird zu Beginn des Clustervorgangs eine beliebige, problemspezifische Anzahl von Clustern gewählt. Da jedes Objekt mit einer gewissen Wahrscheinlichkeit zu jedem Cluster gehört, beeinflusst auch jedes Objekt die Parameter der Wahrscheinlichkeitsverteilungen, die die entsprechenden Cluster beschreiben. Der Erfolg des Algorithmus hängt dementsprechend stark von der angenommenen Wahrscheinlichkeitsverteilung ab.

f) Assoziationsanalyse
In der Assoziationsanalyse wird nach (verwertbaren) Assoziationsregeln gesucht, die Korrelationen zwischen gemeinsam auftretenden Dingen beschreiben. Ein typisches Anwendungsfeld sind die Zusammenhänge beim Einkauf, die sogenannte Warenkorbanalyse, um gezielt Werbemassnahmen einzuleiten. Beispielsweise kann eine Assoziationsregel lauten: "Bei 80 Prozent der Einkäufe, in denen Bier gekauft wird, werden auch Pommes Chips gekauft." [1] Häufig werden diese Erkenntnisse im Crossmarketing genutzt.

Die Idee hinter der Assoziationsanalyse ist einfach. Die Suche nach verwertbaren Regeln kann aber sehr aufwendig sein, denn es gibt sehr viele Kombinationen von Produkten, zwischen denen keine starken Beziehungen bestehen. Der erste Algorithmus zur Assoziationsanalyse ist der AIS-Algorithmus (benannt nach seinen Entwicklern Agrawal, Imielinski und Swami), aus dem der wohlbekannte Apriori-Algorithmus entwickelt wurde. Eine neuere Alternative ist der FP-Growth-Algorithmus.

g) Schlussbemerkung
Obwohl es zu jeder Methode verschiedene Algorithmen gibt, ist in den meisten Data Mining Software-Paketen pro Methode nur ein bestimmter Algorithmus implementiert, sodass die Anwenderin oder der Anwender sich diesbezüglich kaum weitere Gedanken machen wird. Mehr zu den Methoden und ihren Anwendungen findet sich in Nisbet, Elder und Miner (2009) oder in einem der anderen Texte, die am Ende von Kapitel 2.1 aufgeführt sind.

Für den Anwender und die Anwenderin ist die Struktur der Fragestellung hingegen viel entscheidender, nämlich wie eingangs angesprochen, ob unbekannte Strukturen aufgedeckt (d.h. unüberwachtes Lernen) werden sollen oder ob eine Objekt-(Kenn-)Grösse aus ihren Merkmalen vorhergesagt werden soll. Deshalb werden in den beiden folgenden Abschnitten die verschiedenen Verfahren, einschliesslich Methoden aus der Statistik, unter dem Gesichtspunkt dieser Strukturen nochmals diskutiert.

2.3.2 Überwachtes Lernen

Beim überwachten Lernen steht eine Zielgrösse oder ein Zielmerkmal im Zentrum, welches aus den anderen Merkmalen vorausberechnet werden soll. Die dazu angewandten Methoden lassen sich weiter je nach Typ der Zielgrösse in zwei Untergruppen unterteilen.

Handelt es sich bei der Zielgrösse um eine kontinuierliche Grösse wie z.B. Nachfrage (in Geld oder in Menge), Lebensdauer, Fahrzeit, usw., kommen regressionsartige Verfahren zum Einsatz. Viele davon sind schon aus der "klassischen" Statistik bekannt wie gewöhnliche lineare und nichtlineare Regression, Gamma- und Weibull-Regression[2]. Bei diesen Modellen wirken die erklärenden Variablen als Linearkombination auf die Zielvariable ein, ausser bei der nicht linearen Regression, bei der aber der funktionale Zusammenhang vorgegeben ist. Zu dieser Gruppe zählen wir auch die Poisson-Regression[3], welche eine Zählung als Zielgrösse enthält. Ist der funktionale Einfluss der erklärenden Variablen auf die Zielvariable eher unklar, setzt man Methoden wie nichtparametrische Regression, generalisierte additive Modelle (GAM) oder multivariate adaptive regression splines (MARS) ein. Vielfach kommen aber Methoden zum Einsatz, die ihren Ursprung im maschinellen Lernen haben, wie z.B. Neuronale Netzwerke, Regressionsbäume, Support Vector Machines (SVM) und Random Forest.

[1] Diese Regeln werden oft schematisch in einheitlicher Form präsentiert: „Wenn Item A vorkommt, dann tritt auch Item B auf“ (Wenn-dann-Regel).

[2] Teil der generalisierten linearen Modelle (GLM)

[3] Auch Teil der generalisierten linearen Modelle (GLM)

Falls es sich bei der Zielgrösse um eine kategorielle Grösse handelt, die z.B. eine Gruppenzugehörigkeit, einen Objekttyp oder dergleichen bezeichnet, spricht man von Klassifikation oder Diskriminanzanalyse. Methoden aus der Statistik sind die verschiedenen Typen von Diskriminanzanlyse-Methoden (z.B. lineare, quadratische, flexible, mixture, ...), logistische und multinomiale logistische Klassifikation sowie discrete choice models. Die bereits zuvor erwähnten künstlichen neuronalen Netze, Klassifikationsbäume und Support Vector Machines (SVM) sowie die Nächste-Nachbar-(knn)Methode und "ensemble-learning" Ansätze (boosting, bagging und Random Forest) sind verbreitet angewendete Verfahren in Data-Mining-Anwendungen.

In der praktischen Anwendung von Klassifikationsmethoden ist die logistische Regression, die identisch zu einem einschichtigen Perzeptron, d.h. einem vereinfachten neuronalen Netz ist, die am häufigsten verwendete klassische Methode. Unter den neueren Ansätzen dominieren zurzeit die "ensemble-learning" Ansätze die Szene.

Welcher dieser Ansätze ist nun der beste? Es gibt verschiedenste Performance-Vergleiche[1], wobei oft jener Ansatz am erfolgreichsten ist, den die Autoren am besten beherrschen. Auch Diskussionen mit verschiedenen Anwendungsgruppen zeigen, dass es in erfolgreichen Projekten oft von untergeordneter Bedeutung ist, mit welchem Ansatz man arbeitet. Entscheidender ist, wie gut man die Methode beherrscht.

Im Data Mining werden also sowohl Methoden aus der Statistik als auch aus dem maschinellen Lernen eingesetzt. Statistik und Data Mining lassen sich deshalb kaum auf der methodischen Ebene unterscheiden. Im Gegensatz zur traditionellen Statistik ist jedoch die Grundhaltung im überwachten Lernen darauf ausgerichtet, die Zielgrösse möglichst gut aus den vorhandenen Merkmalen vorherzusagen. Dabei wird kaum Rücksicht genommen, ob das Merkmal einen ursächlichen Einfluss auf die Zielgrösse ausübt oder für diese eher die Funktion eines Indikators einnimmt. Der zweite Fall enthält auch die Situation, bei welcher zwischen dem Merkmal und der Zielgrösse eine Scheinkorrelation besteht. Wenn man das verhindern möchte, muss bei der Auswahl der Merkmale sehr sorgfältig auf kausale Zusammenhänge geachtet werden. In den meisten Anwendungsgebieten des Data Mining ist das aber vielfach schwierig bis unmöglich, weil die kausalen Zusammenhänge im Dunklen liegen oder sehr komplex sind. Im Weiteren können die Ausprägungen eines verwendeten Merkmals schon selber eine kausale Folge der anderen Merkmale sein. Dann stellt sich die Frage, ob es trotzdem als Merkmal in der Modellierung zugelassen oder doch besser ausgeschlossen werden soll. Bei einem Entscheid ist zu beachten, dass, sofern der kausale Zusammenhang nicht mit einer (einfachen) mathematischen Formel ausgedrückt werden kann, durch Einbezug des Merkmals die Prognosefähigkeit des Modells verbessert werden kann. Andererseits verbessert der Einschluss von Merkmalen, die bekanntermassen einen (kausalen) Einfluss auf die Zielgrösse ausüben, nicht notwendigerweise die Prognosefähigkeit. Meistens handelt es sich dabei um Merkmale, die keinen grossen Einfluss im konkreten Fall ausüben. Der Grund liegt darin, dass die Anpassung des Modells an die Daten nur mit einer gewissen Unsicherheit vorgenommen werden kann. Ist diese Unsicherheit grösser als der Gewinn in der Prognosefähigkeit, der sich durch den Einschluss eines solchen Merkmals ergibt, resultiert dann eben insgesamt eine Verschlechterung der Prognosefähigkeit.

2.3.3 Unüberwachtes Lernen

Um unbekannte Strukturen in den Daten aufdecken zu können, werden Methoden eingesetzt, die sich in drei Gruppen unterteilen lassen. Eine erste Gruppe enthält Methoden, mit denen hochdimensionale Daten (d.h. Objekte, die mit vielen Merkmalen beschrieben sind) möglichst niedrig dimensional visualisiert werden können. Konkret strebt man üblicherweise eine zweidimensionale Darstellung an. Bekannte, klassische Methoden sind die Hauptkomponenten-Analyse sowie die metrische und ordinale multidimensionale Skalierung. Eine weitere, jedoch sehr rechenintensive Methode ist projection pursuit. Oft

[1] Meyer, Leisch und Hornik (2003) kommen zu folgendem Schluss: *"Support vector machines yielded good performance, but were not top ranked on all data sets. For classification, simple statistical procedures and ensemble methods proved very competitive, mostly producing good results "out of the box" without the inconvenience of delicate and computationally expensive hyperparameter tuning. For regression tasks, neural networks, projection pursuit regression and random forests often yielded better results than SVMs."*

werden aber auch nur die zweidimensionalen Streudiagramme je zweier Merkmale betrachtet. Ein wichtiges Ziel dieser Visualisierungen ist es, Strukturen und sonstige Auffälligkeiten in diesen Projektionen zu finden. Möglicherweise lassen sich dadurch auch die Objekte in Gruppen einteilen (informelle Cluster-Analyse).

Cluster-Analysen, die auf algorithmischer Basis die Aufteilung der Objekte in Gruppen (oder eben Cluster) vornehmen, bilden die zweite Methodengruppe. Die bekanntesten Verfahren sind die k-Means Verfahren, eine so genannte Partitionierungsmethode, und die verschiedenen hierarchischen Verfahren (single, complete, average, Ward-Linkage, usw.). Auch Fuzzy Methoden und model-based clustering Verfahren (EM-Algorithmus) werden eingesetzt. Ihren Ursprung hat die Cluster-Analyse in der Taxonomie (Teilgebiet der Biologie), wo über eine Clusterung von verwandten Arten mittels Körpermerkmalen oder neuerdings mittels Gensequenzen eine Ordnung der Lebewesen ermittelt wird.

Um Objekte einer gemeinsamen Gruppe oder eben verschiedenen Gruppen zuteilen zu können, muss eine Distanz oder, allgemeiner, eine (Un-)Ähnlichkeit aufgrund der Merkmale zwischen den Objekten ermittelt werden. Die bekanntesten Distanzmasse sind die euklidische Distanzfunktion und die City-Block-Distanzfunktion (auch L1-Metrik genannt). Beide können aber nur dann eingesetzt werden, wenn alle Merkmale mit quantitativen (z.B. metrischen) Merkmalen beschrieben wurden. Da dies im Data Mining eher selten der Fall ist, werden viele weitere, teilweise sehr problemspezifische Distanzmasse verwendet[1]. Es gibt kaum wegweisende Regeln, welches Mass in welchem Fall einzusetzen ist, ausser die Unterscheidung nach Merkmalstypen, d.h. die Unterscheidung, ob es sich um quantitative Grössen wie Intervall-, Betrags- oder Zähldaten oder um kategorielle Grössen wie dichotome, nominale oder ordinale Grössen handelt. Sind die Objektmerkmale durch verschiedene Datentypen beschrieben, wird in einem der am meisten verbreiteten Verfahren, Gower's Dissimilarity genannt, für jedes Merkmal separat die Unähnlichkeit berechnet. Das Mittel dieser Unähnlichkeit über alle Objektmerkmale liefert dann die (Gesamt-)Unähnlichkeit zwischen zwei Objekten. Damit kein Merkmal die Gesamtunähnlichkeit dominieren kann, wird verlangt, dass die einzelnen Unähnlichkeiten zwischen 0 und 1 liegen. Das ist einfach ein naheliegender Ansatz ohne tiefere Begründung. Die Wahl des Unähnlichkeitsmasses ist und bleibt vielfach willkürlich. Allerdings hat die Wahl des Unähnlichkeitsmasses oftmals einen entscheidenden Einfluss[2] auf das Ergebnis der Clusteranalyse. Aus all dem folgt, dass die Resultate einer Clusteranalyse von diversen subjektiven Entscheidungen beeinflusst werden, die durchaus kritisch hinterfragt werden können. In einem wissenschaftlichen Zusammenhang können demzufolge die Resultate einer Clusteranalyse kaum für mehr als zur Generierung von (neuen) Fachhypothesen eingesetzt werden.

Auch die im 2.3.1. f) vorgestellte Assoziationsanalyse kann als ein Teil des unüberwachten Lernens angesehen werden.

[1] Wie z.B. die Levenshtein Distanz, die in Kapitel 3.6 beschrieben ist.

[2] Dieser Sachverhalt wird z.B. in Schlich (2004, Kap. 6) ausführlich im Zusammenhang mit Verkehrsverhaltensmustern diskutiert und es werden einige spezifisch in der Verkehrsverhaltensforschung verwendete Unähnlichkeitsmasse vorgestellt. – Data Mining Software bietet üblicherweise keine "fachspezifischen" Unähnlichkeitsmasse a priori an. Solche können jedoch meistens ohne Probleme hinzugefügt werden.

3 Data Mining in der Verkehrsplanung

Die folgenden der Literatur entnommenen Beispiele sollen zeigen, wie Data Mining zur Behandlung bestimmter Fragestellungen aus der Verkehrsplanung eingesetzt wurde und welche Erkenntnisse dabei gewonnen wurden.

3.1 Berechnung von Nachfrageelastizitäten

Dennerlein et al. (1990) haben im Auftrag des damaligen Stabes für Gesamtverkehrsfragen die Verkehrsverhaltensreaktionen in Bezug auf Einkommens-, Preis- und Geschwindigkeitsänderungen untersucht und die entsprechenden Elastizitäten berechnet. Auch wenn der Begriff "Data Mining" in der Studie nicht auftaucht, entspricht das Vorgehen, welches sich in die folgenden Teilschritte gliedert, weitgehend einem Data Mining Prozess:

- Aufbereitung der theoretische Grundlagen für die Analyse des Verkehrsverhaltens (Business Understanding)
- Bestandsaufnahme der verfügbaren Daten (Data Understanding)
- Aufbereitung und Bereinigung der umfangreichen Datengrundlagen sowie Zusammenführung zu einer Untersuchungsdatei (Data Preparation)
- Datenanalyse mit statistischen Methoden wie Logit-, Probit-, Tobit- und Regressionsverfahren (Modeling)
- Evaluation der Ergebnisse und Validierung der Elastizitätsansätze (Evaluation)
- Erstellen eines Strategiemodells, Bericht mit Empfehlungen (Deployment).

Die Ergebnisse der Studie bildeten einen wichtigen Beitrag zu den Arbeiten an der damaligen Gesamtverkehrskonzeption Schweiz (GVK-CH).

3.2 Zuordnung von Verkehrserzeugungsraten

In den 4-Schritt-Modellen werden die Verkehrserzeugungsraten von Personengruppen oder Gebietstypen (trip generation) üblicherweise mit Regressions- oder mit Cross-Klassifikations-Methoden, teilweise auch mit Logit-Modellen ermittelt. Die Schwierigkeit bei diesen Methoden besteht in der Wahl der "richtigen" erklärenden Variablen. Als neue Methode testeten Strambi und van de Bilt (1998) den Einsatz von CHAID (Chi-Squared Automatic Interaction Detection[1]) zur Analyse der Verkehrserzeugungsraten von Haushalten. Das Ergebnis von CHAID ist eine Kriterien-basierte Segmentierung der Haushalte in Untergruppen, welche sich bezüglich der Verkehrserzeugungsraten unterscheiden. Im Gegensatz zu einer Cluster-Analyse ist dabei ersichtlich, welche unabhängigen Variablen (predictors) die Zugehörigkeit zu einer Untergruppe bestimmen. Am Beispiel des in der Studie verwendeten Datensatzes einer in São Paulo durchgeführten Haushaltbefragung erwiesen sich für die Anzahl Arbeitswege/Haushalt die Zahl der Beschäftigten pro Haushalt als wichtigste erklärende Variable, gefolgt von Werkstudenten/Haushalt, Haushaltseinkommen, Schüler/Haushalt und Kinder[2]/Haushalt. Mit CHAID war es möglich, dieses Ergebnis "automatisch" zu erhalten, ohne Notwendigkeit, verschiedene Modellansätze zu testen. Es könnte direkt als Verkehrserzeugungsmodell verwendet werden. Die mit CHAID aufgezeigten Interaktionen zwischen den erklärenden Variablen und der Zielvariable können helfen, die Zusammenhänge besser zu verstehen oder bisher nicht beachtete Zusammenhänge zu erkennen. Die Autoren empfehlen denn auch, CHAID als exploratives Hilfsmittel für die Spezifikation von konventionellen Modellen einzusetzen.

[1] Siehe Kapitel 2.3.1 c)

[2] Jünger als 10 Jahre

3.3 Autobesitz

Am Beispiel der Modellierung der Wahl des Autos (Wahl des Autotyps und Entscheid, ob ein Neu- oder Gebrauchtwagen angeschafft werden soll) von Haushalten vergleichen Mohammadian und Miller (2002) die Performance eines künstlichen neuronalen Netzwerkes (multilayer perceptron artificial neural network[1], MLP ANN) mit jener eines Nested Logit Modells (NLM). Die Daten wurden aus einer in Toronto durchgeführten Autobesitz-Studie übernommen. Diese wurden für die Erstellung des neuronalen Netzwerkes in einen Trainings-, einen Validations- und einen Testdatensatz unterteilt[2]. Während mit dem MLP ANN die Einzelentscheide der Haushalte besser nachgebildet werden konnten als mit dem NLM (gut 80% korrekt gegenüber knapp 50%). schnitt das NLM bezüglich der aggregierten Vorhersage (Marktanteil) gemäss einer Validierung auf dem Testdatensatz besser ab. Die neuronalen Netzwerke haben nach Ansicht der Autoren den Vorteil, dass sie in kürzerer Zeit bessere Ergebnisse liefern als traditionelle Discrete Choice Models, aber den Nachteil, dass es sich um "Black Boxes" handelt mit fehlender statistischer Transparenz und Interpretierbarkeit (Elastizitäten, Goodness of fit usw.). Sie sehen das Potential der neuronalen Netzwerke daher primär darin, rasch aufzeigen, welches die wichtigen bestimmenden Variablen sind und so Hinweise für die zu wählenden Inputvariablen für andere, konventionelle, Modelltypen zu liefern.

3.4 Verkehrsmittelwahl

Die Modellierung der Wahl des Verkehrsmittels erfolgt standardmässig mit auf dem Theorem der Utility-Maximierung beruhenden Logit-Ansätzen (z.B. Multinomial Logit, MNL). Wets et al. (2000) prüften als Alternativen zu einem MNL-Modell den Einsatz der Data Mining-Methoden C4.5 und CHAID[3]. Die für die Testanwendungen verwendeten Daten stammen von einer im Raum Rotterdam durchgeführten Aktivitäten-Tagebuch-Erhebung. Wie für Data Mining typisch, mussten diese zuerst bereinigt werden, da sie Inkonsistenzen und fehlende Werte aufwiesen. Dazu gelangte eine eigens zum Zweck der Bereinigung von Tagebuch-Erhebungen entwickelte Software[4] zum Einsatz. Für den Einsatz von CHAID mussten kontinuierliche Daten diskreten Werten zugeordnet werden. Dazu wurde die equal-frequency-interval-Methode verwendet. Betrachtet wurde das jeweilige für eine Tour benutzte Haupt-Verkehrsmittel. Die Arbeit zeigte, dass sich C4.5 und CHAID sehr gut für die Abbildung der Verkehrsmittelwahl eignen. Der Vergleich der Misklassifikationsmatrizen mittels eines Testdatensatzes ergab, dass zwischen den 3 Methoden (C4.5, CHAID und MNL) kaum Unterschiede bestehen. Als Vorteil von C4.5 und CHAID gegenüber MNL erwähnen die Autoren deren grössere Robustheit resp. geringere Sensitivität bezüglich kleinen Änderungen in den unabhängigen Variablen, Diskontinuitäten oder Ausreissern und Mulitkollinearität.

Xie, Lu und Parkany (2003) verglichen für die Modellierung der Wahl des Verkehrsmittels für Arbeitswege am Beispiel eines Tagebuch-Datensatzes von San Francisco den Einsatz eines traditionellen MNL, eines Klassifikationsbaum (CT)-Algorithmus (C4.5)[5] und eines neuronalen Netzwerkes[6]. Als Vorteil der beiden Data Mining Methoden weisen sie darauf hin, dass keine spezifische Modellstruktur angenommen werden muss und das Problem der IIA-Eigenschaftentfällt. Die Ergebnisse der Studie lassen sich wie folgt zusammenfassen: Mit den beiden Data Mining-Methoden werden geringfügig bessere Ergebnisse in Bezug auf die Vorhersage-Genauigkeit (Hit-Ratio) erzielt als mit dem MNL-Modell. Die Autoren weisen darauf hin, dass bei neuronalen Netzwerken die Interpretierbarkeit der Ergebnisse ein Problem darstellen kann.

Biagioni et al. (2009) testeten verschiedene Klassifikationsmethoden zur Modellierung der Wegeketten-basierten Verkehrsmittelwahl und verglichen die Ergebnisse mit jenen eines MNL. Es wurde angenommen, die Verkehrsmittelwahl für den ersten Weg der

[1] Mehr dazu in 2.3.1 a)

[2] Diese Unterteilung dient der Vermeidung von Overfitting des neuronalen Netzwerkes

[3] Mehr zu den Klassifikationsbaum-Algorithmen C4.5 und CHAID in 2.3.1 c)

[4] SYLVIA (Arentze et al., 1999)

[5] Einen Überblick über die verschiedenen Klassifikationsbaum-Algorithmen gibt Kapitel 2.3.1 c)

[6] Mehr zum Neuronalen Netz(-werk) in 2.3.1 a)

Wegekette erfolge unabhängig von den nachfolgenden Wegen. Das für den ersten Weg gewählte Verkehrsmittel wird als Anchor-Mode bezeichnet. Die Wahl der Verkehrsmittel für die folgenden Wege wurde einmal mit und einmal ohne Berücksichtigung dieses Anchor-Modes modelliert. Als Data Mining Methoden wurden C4.5, Naïve Bayes (NB), Simple Logistic, Support Vector Machines (SVM) sowie die Boostmethode AdaBoost (AB) verwendet[1]. Die Boostmethode wurde einmal mit C4.5 (AB-C4.5) und einmal mit NB (AB-NB) kombiniert getestet. Zusätzlich wurden für die beiden Fälle mit/ohne Anchor-Mode Discrete Choice Modelle (MNL) geschätzt. Als Datenbasis dienten die Ergebnisse einer in Chicago durchgeführten Aktivitäten-basierten Erhebung (ca. 32'000 Personen). In einem ersten Schritt mussten die Daten bereinigt werden. Datensätze mit fehlenden Werten oder unvollständigen Wegeketten wurden gelöscht. Von den ursprünglich im Datensatz vorhandenen rund 35'000 Wegeketten blieben nach der Bereinigung ca. 20'000 übrig. Abhängige Variable war jeweils das gewählte Verkehrsmittel, als unabhängige Variable dienten Wege-, Transportmittel- und Haushalt-Attribute. Die Beurteilung der Performance der Data Mining-Methoden erfolgte anhand der drei folgenden Kenngrössen:

- Accuracy: Anzahl der korrekt klassifizierten Fälle dividiert durch die Gesamtzahl der Fälle
- Precision: Anzahl der korrekt klassifizierten Fälle einer Klasse A dividiert durch die Gesamtzahl der als Klasse A klassifizierten Fälle
- Recall: Anzahl der korrekt als der Klasse A angehörig klassifizierten Fälle dividiert durch die Anzahl der im Datensatz effektiv der Klasse A angehörigen Fälle.

Die Ergebnisse lassen sich wie folgt zusammenfassen:

- Die Berücksichtigung des Anchor-Modes führt generell zu besseren Ergebnissen der Data Mining-Klassifikationsmethoden
- Von den Data Mining Methoden lieferte im vorliegenden Fall die Methodenkombination AB-NB/AB-C4.5 die besten Ergebnisse für die Nachbildung der Verkehrsmittelwahl
- Diese Kombination von Data Mining Methoden ergab für alle 3 Kenngrössen (Accuracy, Precision und Recall) die besseren Resultate als das MNL-Modell

Aufgrund dieser Ergebnisse empfehlen die Autoren, zur Modellierung der Verkehrsmittelwahl MNL-Modelle und Data Mining als sich ergänzende und sich gegenseitig unterstützende Methoden einzusetzen.

3.5 Routenwahl

An zwei Fallbeispielen mit jeweils zwei Alternativrouten untersuchten Yamamoto, Kitamura und Fujii (2002) den Einsatz von C4.5 zur Analyse des Routenwahlverhaltens. C4.5 ist, wie oben beschrieben, ein Algorithmus zur Bildung von Decision Trees und Production Rules[2]. Zur Charakterisierung der betrachteten Routenalternativen wurden die erwartete minimale Reisezeit, die erwartete maximale Reisezeit und die erwartete mittlere Reisezeit verwendet. Als weitere unabhängige Variable wurden sozioökonomische Daten der Verkehrsteilnehmer sowie zusätzliche Attribute der Fahrt verwendet. Der Vergleich mit einem binären Logit Modell zeigte, dass dieses einen kleineren Anteil der korrekt vorhergesagten Routenwahlentscheide (hit ratio) aufwies als die Data Mining Methoden. Die Autoren kamen zum Schluss, dass sich sowohl Decision Trees als auch Production Rules sehr gut zur Modellierung der Routenwahl eignen und dass Data Mining Methoden generell ein grosses Potential zur Analyse von Daten zum Mobilitätsverhalten aufweisen.

3.6 Identifikation und Vergleich von Mobilitätsmustern

Die mit Aktivitäten- und Mobilitätstagebüchern erhobenen Daten beinhalten eine grosse Informationsmenge, insbesondere auch zur Abfolge und Dauer der Aktivitäten. Das Erkennen von Mustern in diesen Daten und der Vergleich erkannter Muster dienen dem besseren Verständnis des Aktivitäten- und Mobilitätsverhaltens der Menschen. Auch

[1] Alle erwähnten Algorithmen ausser Naïve Bayes sind im Kapitel 2.3.1 genauer beschrieben.

[2] Ein Ast von der Wurzel bis zum Endblatt in einem Klassifikationsbaum wird in der Form „Wenn die und die Bedingungen erfüllt sind, dann gehört das Objekt zur Klasse X“ neu formuliert.

werden so wertvolle Hinweise für die Bildung von Modellen gewonnen, mit denen das Aktivitäten- und Mobilitätsverhalten abgebildet und prognostiziert werden kann. Eine oft angewandte Methode für die Zuordnung von Mustern zu Gruppen ist die Cluster Analyse (z.B. Ma und Goulias, 1996). Für den Vergleich von Aktivitäten-Sequenzen schlägt Wilson (1998) die aus der Bioinformatik bekannte Sequence Alignment Methode (SAM) vor. Bei dieser Methode wird der erforderliche Aufwand (ausgedrückt als Levenshtein Distanz resp. L-distance), welcher nötig ist um zwei Sequenzen durch Löschen, Einfügen und Ersetzen von Elementen zur Übereinstimmung zu bringen, berechnet und als Mass für die Distanz resp. die Ähnlichkeit zwischen zwei Sequenzen von Elementen (z.B. Aktivitäten) verwendet. Damit lassen sich Fragestellungen wie z.B. die folgenden behandeln (Wilson, 1998):

- Wie ähnlich sind die Mobilitätsmuster von Frauen und Männern?
- Sind die Mobilitätsmuster von Arbeitslosen ähnlicher jenen von Pensionierten oder jenen von Beschäftigten?
- Haben sich die Mobilitätsmuster während der letzten 10 Jahre signifikant verändert?

Nach Joh, Arentze und Timmermans (2001) werden in den bisher verwendeten Methoden zur Klassifizierung von Aktivitätenmustern die Informationen über die Abhängigkeiten zwischen den erklärenden Attributen und jene über die sequentiellen Abhängigkeiten zwischen den Aktivitäten nicht berücksichtigt. Sie schlagen vor, die Sequence Alignment Methode (SAM) zur multidimensionalen Sequence Alignment Methode (MDSAM) zu erweitern und mit dieser ein neues integriertes Distanzmass zu bestimmen, welches beide Arten von Informationen berücksichtigt. Neben den Aktivitätstypen werden als weitere Dimensionen wie Ort der Aktivität, benutztes Verkehrsmittel, mit/ohne Begleitperson usw. berücksichtigt. Die Autoren haben dazu eine spezielle Software (DANA[1]) geschrieben, welche die Anwendung von MDSAM unterstützt. Die Anwendung an einem Beispieldatensatz ergab, dass MDSAM potentiell eine wertvolle und sensitive Methode zur Gewinnung eines Distanzmasses für die Klassifikation von Aktivitäten- und Mobilitätsverhalten ist, vor allem dann, wenn die Interdependenzen zwischen den verschiedenen Variablen von Bedeutung sind. Es zeigte sich aber auch, dass weiterer Forschungsbedarf besteht. In einer späteren Studie wendeten die Autoren (Joh, Arentze und Timmermans, 2007) MDSAM an, um Aktivitätenketten von Frauen und Männern hinsichtlich häufig auftretender gemeinsamer Teilsequenzen (skeletal information) zu untersuchen. Es konnte festgestellt werden, dass sich im verwendeten Datensatz[2] die Aktivitätenmuster von Frauen und Männern deutlich unterscheiden. Während sich die wichtigsten Ähnlichkeiten der Aktivitätenmuster von Männern mit wenigen gemeinsamen Teilsequenzen beschreiben lassen, wurden bei den Frauen sehr viel unterschiedlichere und komplexere Sequenzen festgestellt. Generell beurteilen die Autoren MDSAM als geeignete Methode, mit skeletal information die Ähnlichkeiten von mehrdimensionalen Aktivitätensequenzen zu beschreiben. Die Anzahl der mit dieser Methode bearbeitbaren Fälle ist aber beschränkt und es fehlt vorläufig eine Methode, die statistische Signifikanz der Ergebnisse zu beurteilen. Auch diesbezüglich sehen die Autoren weiteren Forschungsbedarf.

Ausgehend von der fundamentalen Annahme, dass Haushalte und Einzelpersonen bei der Planung ihrer Aktivitäten in Raum und Zeit heuristische Entscheidungs-Regeln anwenden, müsste es nach Keuleers et al. (2001) möglich sein, z.B. mit Assoziations-Regeln raum-zeitliche Muster in den Daten von Aktivitäten-Tagebüchern zu finden. Um dies zu testen, verwendeten sie den gleichen Datensatz wie Wets et al. (2000) (siehe Abschnitt 3.4). Für die Anwendung des Algorithmus zur Entdeckung von Assoziations-Regeln zwischen Haushalt- und Personendaten auf der einen und Aktivitätenmustern auf der anderen Seite war ein Preprocessing des Datensatzes (z.B. Diskretisierung kontinuierlicher Variablen und Umwandlung von nominalen Variablen in binäre Variablen) nötig. Die Anwendung des Algorithmus mit festgesetzten Werten für Support und Confidence resultierte in einer Unmenge von Assoziation-Regeln, welche einem Postprocessing unterzogen wurden, bei dem die Gruppe der nicht-trivialen Assoziations-Regeln identifiziert wurde. Von den verbleibenden Regeln waren nur jene von Interesse, welche als Antecedent ausschliesslich Haushalts- und Personendaten und als Consequent ausschliesslich

[1] Dissimilarity Analysis of Activity Patterns (Joh et al., 2006)

[2] Aus 6'950 Fällen wurden je 100 Aktivitätenketten für Frauen und Männer zufällig ausgewählt.

nur Aktivitätendaten enthalten. Ca. 6'500 Regeln erfüllten diese Bedingung. Von diesen wurden schliesslich Lift und Coverage ermittelt und so festgestellt, welche Haushalt- und Personendaten den grössten Einfluss auf die Wahl der Aktivitäten haben. Als Schlussfolgerung stellten die Autoren fest, dass auch nach dem Postprocessing die Zahl der mit dem Algorithmus gefundenen Assoziations-Regeln noch viel zu gross und das Ergebnis deshalb nicht sehr hilfreich war, auch wenn verschiedene interessante Assoziationen gefunden wurden.

Statt die Ähnlichkeiten von Aktivitätenketten zu analysieren versuchten Joh, Ettema und Timmermans (2009) mit der SAM-Methode den mentalen Aufwand bei der kurzfristigen Anpassung von Aktivitätenplänen (Streichen einer Aktivität, Einfügen einer Aktivität, Ersetzen einer Aktivität) abzuschätzen. Gegenstand der Studie war also nicht das Ergebnis, sondern der Prozess der Entscheidungsfindung. Es wurden zwei Datensätze verwendet: einer mit den ursprünglich geplanten Aktivitäten und einer mit den tatsächlich durchgeführten Aktivitäten der gleichen Personen. Die beiden Datensätze unterscheiden sich also durch die fallengelassenen, die hinzugefügten und die ersetzten Aktivitäten. Mit SAM kann der Aufwand geschätzt werden, welcher hinter diesen Änderungen steckt. Es gibt jedoch keine empirischen Erkenntnisse zur Gewichtung der einzelnen Operationen (Löschen, Hinzufügen und Ersetzen). Unter der Annahme, dass die Leute den Aufwand für die Anpassung der Tagespläne möglichst minimieren, kann davon ausgegangen werden, dass die Verteilung der Levenshtein Distanzen zwischen geplanten und ausgeführten Aktivitätenketten mehr kleine als grosse Werte aufweist. Auf dieser Basis wurde ein genetischer Algorithmus eingesetzt, welcher Gewichte der mentalen Aufwände berechnet, für welche dann mit SAM unter Verwendung der Software DANA (siehe oben) die Levenshtein Distanzen berechnet wurden. Dieser Vorgang wurde solange wiederholt, bis die Levenshtein Distanzen die gewünschte Verteilung aufwiesen. Das Ergebnis deutete darauf hin, dass bei kurzfristigen Anpassungen der Aktivitätenpläne das Löschen einer Aktivität kaum einen mentalen Aufwand erfordert während das Einfügen einer Aktivität mit einem grösseren Aufwand verbunden ist als der Ersatz einer Aktivität. Zur Gewinnung eines besseren Verständnisses des Entscheidungsprozesses bei der Aktivitätenplanung ist die Methode ein vielversprechender Ansatz, welcher nach Ansicht der Autoren in weiteren Forschungsarbeiten vertieft untersucht werden sollte.

3.7 Fazit der Literaturrecherche

In der akademischen Welt wurden die Möglichkeiten von Data Mining zur Behandlung verschiedener Fragestellungen aus der Verkehrsplanung, namentlich auch zum Mobilitätsverhalten, ausgelotet. Dabei konnte in der Regel die Eignung dieser Methoden zur Klassifikation und Vorhersage von Mobilitätsverhalten sowie zur Erkennung von Mustern auf der Grundlage vorhandener Daten (welche nicht speziell zu diesem Zweck erhoben worden waren) gezeigt werden.

Für die Vorhersage des Mobilitätsverhaltens (z.B. Verkehrsmittel- oder Routenwahl) konnten mit den Data Mining Methoden teilweise bessere Ergebnisse erzielt werden als mit klassischen MNL-Ansätzen. Trotzdem wird kaum empfohlen, Data Mining anstelle klassischer Modellansätze zu verwenden. Vielmehr wird vorgeschlagen, Data Mining und konventionelle Modellansätze als sich ergänzende und gegenseitig unterstützenden Methoden einzusetzen, wobei Data Mining schwergewichtig als exploratives Hilfsmittel zur Erkennung der massgebenden Einflussvariablen zum Einsatz gelangen soll.

Für die Erkennung von Mustern im Mobilitätsverhalten haben sich die Sequence Alignment Methoden (SAM) als vielversprechend erwiesen. Aber auch dieses Forschungsergebnis hat den Eingang in die Praxis der Verkehrsplanung noch kaum gefunden.

4 Praktische Anwendung an Fallbeispielen

4.1 Einleitung

Wie der Data Mining Prozess (vgl. Abschnitt 2.2) in der Verkehrsplanung umgesetzt und wie einige der in Abschnitt 2.3 erwähnten Methoden eingesetzt werden können, soll an den drei Themen

- Wegeketten
- Mobilitätstypen aufgrund des täglichen Mobilitätsverhaltens
- Klassifikationsregeln für die Mobilitätstypen aufgrund von Personenmerkmalen und Raumstrukturinformationen.

illustriert werden.

Die Phase "Business Understanding" im Data-Mining Prozess werden wir im Folgenden kaum weiter ansprechen. Auch die letzte Phase kann hier nicht relevant abgedeckt werden, weil die Ergebnisse nicht in die Praxis umgesetzt werden. Zur Bearbeitung der Themen greifen wir auf die Daten des Mikrozensus Verkehr 2005 und für das dritte Thema zusätzlich auf die Raumstrukturdaten des BFS zurück.

4.2 Verwendete Datensätze

4.2.1 Mikrozensus zum Verkehrsverhalten 2005

Die Daten aus dem Mikrozensus Verkehr sind in insgesamt 9 Dateien abgelegt, welche über Schlüsselvariablen einander zugeordnet sind. Von diesen verwenden wir im Rahmen der Fallbeispiele die folgenden drei:

- Zielpersonen-Datei: Enthält die Daten zur Soziodemografie, Mobilität und zu verkehrspolitischen Einstelllungen jeder Zielperson, von welcher das Verkehrsverhalten am Stichtag erhoben worden ist. Jede Zielperson ist durch eine eindeutige Schlüsselvariable identifiziert.
- Wege-Datei: Enthält die Daten (z.B. Start- und Zielort, Zweck, Start- und Ankunftszeit, verwendetes Hauptverkehrsmittel usw.) für jeden Weg, der von den Zielpersonen am Stichtag durchgeführt wurde. Jeder Weg ist durch eine eindeutige Schlüsselvariable identifiziert und mit der Zielpersonendatei verknüpft.
- Etappen-Datei: Enthält die Daten für alle am Stichtag zurückgelegten Etappen1 (Start, Ziel, Verkehrsmittel usw.) Jede Etappe ist mit einer Schlüsselvariable identifiziert, welche eindeutigen eine Zuordnung zum zugehörigen Weg und zur zugehörigen Zielperson gestattet.

Alle drei Dateien sind, wie oben erwähnt, durch Schlüsselvariable miteinander verknüpft.

4.2.2 Raumstrukturdaten

Für die Behandlung des oben erwähnten dritten Themas wurden jedem im Mikrozensus Verkehr rapportierten Weg die Raumstrukturdaten des Quell-(Wohn-) und des Zielortes zugespielt. Dabei wurden zwei Typen von Raumstrukturdaten unterschieden. Der erste erfasst die Struktur auf Stufe Gemeinde. Der zweite Typ enthält die lokale Situation am Quell- oder Zielort. Einerseits wurden Gegebenheiten innerhalb von 300 bis 600 Meter erfasst und andererseits z.B. Grössen wie die Distanz zum Gemeindezentrum, zum nächsten Autobahnanschluss oder zur nächsten ÖV-Haltestelle. Die verwendeten Daten sind im Anhang I zusammengestellt und beschrieben.

[1] Eine Etappe ist der Teil eines Weges, der mit ein und demselben Verkehrsmittel zurückgelegt wird. Etappen sind als Unterebene der Wege die kleinste Erfassungseinheit im Mikrozensus zum Verkehrsverhalten 2005.

4.3 Häufigkeit von Wegeketten-Typen

4.3.1 Business Understanding, Data Understanding, Data Preparation

Eine Wegekette (Tour, Ausgang) ist eine zeitliche Abfolge von Wegen, wobei der erste Weg von zu Hause startet und der letzte Weg zu Hause endet. Die Wege werden durch ihren Zweck charakterisiert. Der letzte Weg hat in der Regel den Zweck "Rückkehr".

Die benötigten Informationen sind in der Wege-Datei abgelegt. Da für jeden Weg die Einträge auf einer separaten Zeile erfasst sind, haben die Einträge für eine Wegekette insgesamt den Charakter einer Liste. Klassische Data Mining Werkzeuge verlangen aber die Daten in Matrixform, d.h. pro Beobachtungseinheit (hier pro Wegekette) müssen alle Informationen auf einer Zeile mit fester Anzahl von Einträgen liegen. Um diese Dateistruktur zu erhalten werden die Wegeketten als Zeichenfolge kodiert, wobei jedes Zeichen (Buchstaben) den jeweiligen Zweck des einzelnen Weges identifiziert. Die Zeichenkette kann beliebig lang sein und wird als "String" (d.h. als ein Eintrag) gespeichert.

4.3.2 Modeling

Für die Modellierung greift man hier auf aus der Statistik bekannte Werkzeuge zurück. Es wird ausgezählt, wie oft jede Zeichenkette vorkommt. Aus fachlichen Überlegungen hat man entschieden, dass die Reihenfolge der Zeichen in der Kette relevant ist. Das heisst, dass z.B. die Wegeketten "Freizeit – Einkaufen – Rückweg" und "Einkaufen – Freizeit – Rückweg" als ungleich betrachtet werden.

Die Auszählung wurde für die ganze Schweiz, nur für den Kanton Thurgau und nur für den Kanton Zürich vorgenommen und jeweils in einem Balkendiagramm dargestellt (vgl. Abbildung 4, wo jeweils die 17[1] häufigsten Wegeketten aufgeführt sind). Es zeigt sich, dass fünf Wegeketten in allen 3 Stichproben deutlich gehäufter vorkommen und es sich in allen 3 Fällen jeweils um die gleichen handelt: "Freizeit – Rückkehr" (FR), "Arbeit – Rückkehr" (AR), "Einkauf – Rückkehr" (ER), "nur Freizeit" (F)[2] und "Schule – Rückkehr" (SR).

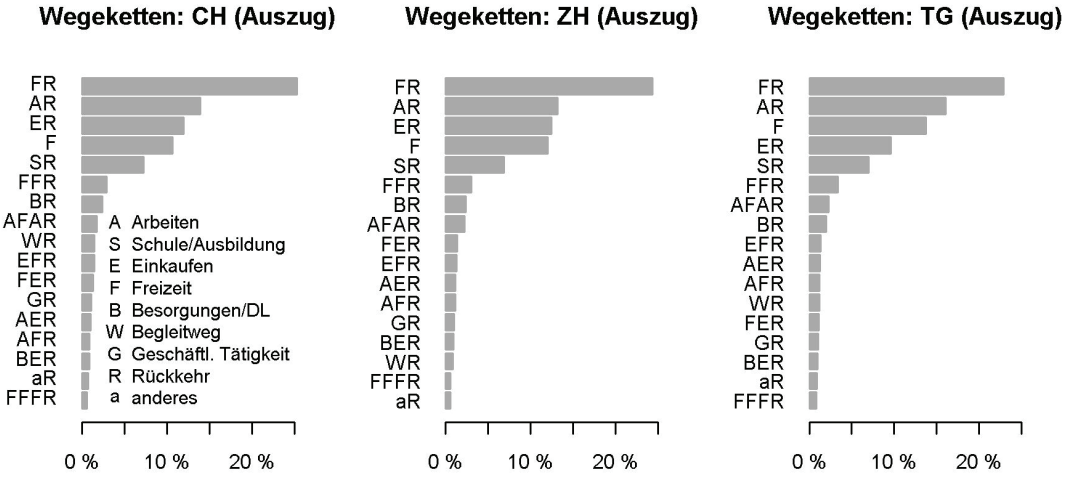

Abbildung 4: Häufigkeiten der am häufigsten vorkommenden Wegeketten in der Schweiz (links), im Kanton Zürich (Mitte) und im Kanton Thurgau (rechts). Der Zweck wird durch einen Buchstaben identifiziert (vgl. Legende links).

Als nächstes wurde untersucht, welche Personen welche Wegeketten zurücklegen. Dazu wurden die Personengruppen mit den fünf häufigsten Wegeketten (FR, AR, ER, F und SR) einzeln betrachtet und jene mit den restlichen Wegeketten in einer Gruppe zusammengefasst (a). Anschliessend wurden diese sechs Personengruppen (mit je identischen

[1] Die Zahl 17 ist willkürlich gewählt. Sie ist gross genug, um aufzuzeigen, wie Rang stabil die Wegeketten für TG, ZH und CH sind.

[2] Das sind Rundwege, z.B. Spaziergänge von zu Hause und zurück nach Hause, ohne weitere Aktivität dazwischen.

Wegeketten) hinsichtlich ihrer (soziodemographischen) Zusammensetzung bezüglich Altersklassen und Geschlecht analysiert. Das Resultat kann in einer mehrdimensionalen Kreuztabelle (vgl. Tabelle 1), in einem Hypercube (nicht gezeigt) oder in einem Mosaik-Plot (Abbildung 5) dargestellt werden.

Tabelle 1: Kreuztabelle der Personengruppen nach Wegeketten-Typen, Altersklassen und Geschlecht (Legende: siehe Abbildung 4)

Mann:						
Alter	a	FR	AR	ER	F	SR
(0,16]	472	1016	59	111	270	1455
(16,20]	260	303	149	41	32	121
(20,35]	1417	1044	950	325	227	84
(35,64]	3429	1988	2451	812	820	21
(64,100]	839	1084	95	595	647	3
Frau:						
Alter	a	FR	AR	ER	F	SR
(0,16]	480	805	32	135	282	1317
(16,20]	257	230	122	56	45	114
(20,35]	1610	1047	683	500	334	90
(35,64]	3829	2478	1662	1704	1405	41
(64,100]	1238	1345	44	1081	723	6

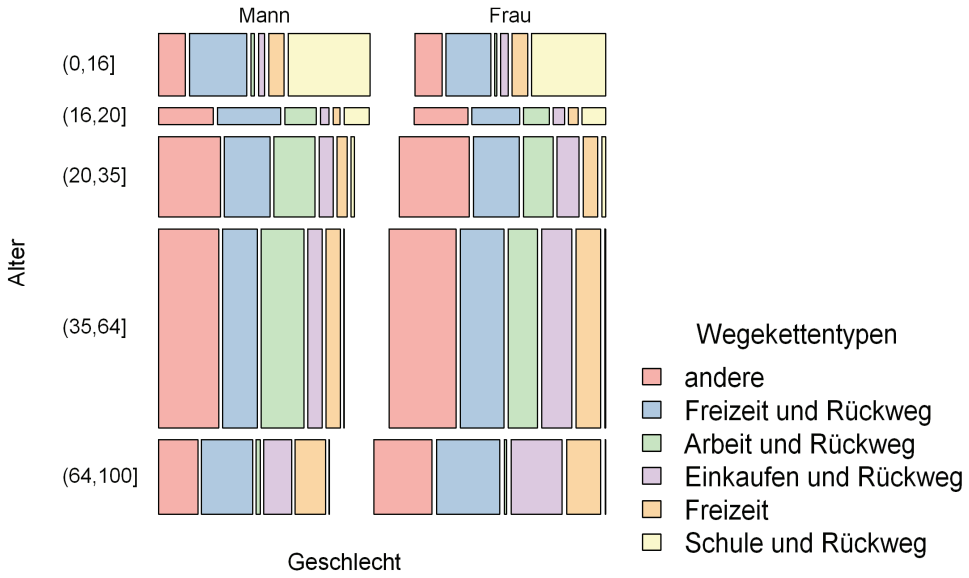

Abbildung 5: Mosaikplot der Auszählung der Wegeketten nach Typ und Personengruppen (Altersklassen und Geschlecht) für die Schweiz

4.3.3 Evaluation

Auf den ersten Blick ist der Mosaik-Plot eine recht anspruchsvolle Darstellung. Die "Zeilenhöhe" der fünf Altersklassen entspricht dem Anteil der Wegketten, die durch Personen der jeweiligen Altersklasse zurückgelegt wurden. Die Zeile ist in zwei Blöcke von Tafeln aufgeteilt. Der linke Block fasst die Wegketten, die Männer zurücklegen, und der rechte die Wegketten, die Frauen zurücklegen, zusammen. Die Breite der Blöcke ist proportional der Anzahl Wegketten, die Männer, respektive Frauen in der jeweiligen Altersklasse zurücklegen. Die Tafelbreiten sind dann proportional zur Häufigkeit der entsprechenden Wegeketten innerhalb einer Altersklasse und des Geschlechts. Falls Altersklasse und Geschlecht keinen Einfluss auf die Wegeketten hätten, müsste die Verteilung innerhalb eines Blocks (d.h. das Breitenverhältnis der Tafeln) gleich sein. Dies ist aber klar nicht der Fall: Die Wegekette "Schule – Rückkehr" (SR) kommt bei den tiefen Altersklassen häufiger vor als bei den oberen. Die Wegekette "Arbeit – Rückkehr" (AR) kommt bei

Männern verhältnismässig häufiger vor als bei Frauen. Bei der Wegekette "Einkauf – Rückkehr" (ER) ist es gerade umgekehrt. Ein weiterer auffälliger Punkt ist, dass die Tafelbreiten bei Männern systematisch von den tiefen Altersklassen zu den hohen abnimmt. Das muss aber nicht notwendigerweise ein Hinweis für eine nicht repräsentative Erhebung sein, da die Häufigkeiten in Bezug auf Wegeketten gerechnet wurden und nicht in Bezug auf die befragten Personen. Der Grund für dieses Phänomen kann ein unterschiedliches Verhalten der Geschlechter in den verschiedenen Altersklassen bezüglich der Rückkehr nach Hause zwischen verschiedenen Tätigkeiten sein, welches jeweils in der Folge zu neuen Wegeketten führt.

4.4 Anzahl Wegeketten pro Person und Tag

4.4.1 Business Understanding, Data Understanding, Data Preparation

Oben wurde darauf hingewiesen, dass Personen ganz unterschiedlich viele Wegeketten pro Tag durchführen, abhängig von ihrem Alter, ihrem Geschlecht und vielen weiteren sozio-demografischen Merkmalen. Um diesen Aspekt weiter zu erkunden, wurde versucht, die (erwartete) Anzahl Wegeketten pro Person und Tag möglichst gut durch eine Auswahl von sozio-demografischen Merkmalen vorherzusagen. Ob die gewählten soziodemografischen Merkmale einen kausalen Einfluss auf die Anzahl Wegeketten haben oder nicht, bleibt vorerst dahingestellt. Dies kann gegebenenfalls im Nachhinein weiter diskutiert werden.

Entsprechend unserem Vorhaben waren die Daten aus den verschiedenen in Kapitel 4.2.1 erwähnten Datensätzen des Mikrozensus zum Verkehrsverhalten zusammenzuführen und geeignet aufzubereiten.

4.4.2 Modeling

Das geläufigste Modell, um Anzahlen (hier die Anzahl Wegeketten) zu modellieren, ist die Poisson-Regression. Dabei geht man davon aus, dass die Anzahl Wegeketten poissonverteilt und deren logarithmierter Erwartungswert eine Linearkombination von ausgewählten soziodemografischen Merkmalen ist. Folglich wirken die Merkmale multiplikativ auf die erwartete Anzahl Wegeketten. Die Anpassung solcher Modelle erfolgt z.B. mit dem Scoring-Algorithmus, einem iterativen Algorithmus, welcher üblicherweise für die Anpassung generalisierter linearer Modelle (GLM) eingesetzt wird.

Der Modellierungsprozess wurde gestartet, indem subjektiv einige wichtige soziodemografischen Merkmale ausgewählt wurden. Anschliessend wurde ein schrittweises Variablenselektionsverfahren eingesetzt, um unter allen zur Verfügung stehenden soziodemografischen Merkmalen jene auszuwählen, welche die Zielgrösse "Anzahl Wegeketten" empirisch am besten beschreiben[1]. Das Variablenselektionsverfahren beruhte in diesem Fall auf dem so genannten Bayes'schen Informationskriterium (BIC), welches in der Lage ist, verschiedene Modelle bezüglich ihrer Möglichkeiten zu bewerten, die Zielvariable zu beschreiben. Die statistischen Details des Modellierungsprozesses sind im Anhang II zusammengestellt.

Als wichtige Merkmale zur Vorhersage der erwarteten Anzahl Wegeketten haben sich die in Tabelle 2 aufgeführten sozio-demografischen Merkmale erwiesen. Die "Verankerung" des Modells liegt in einer (Referenz-) Person, die zwischen 35 und 64 Jahre alt und Vollzeit (mit gratis Parkplatz am Arbeitsort) angestellt ist, immer ein Auto oder ein Fahrrad zur Verfügung hat und in einem Haushalt mit 2 Personen wohnt. Sie legt im Durchschnitt 1.64 Wegeketten pro Tag zurück. Um die durchschnittliche Anzahl Wegeketten pro Tag für eine Person mit anderen Merkmalen zu bestimmen, können die entsprechenden Faktoren (rechte Spalte) aus der Tabelle 2 miteinander multipliziert werden. (Die Referenz-

[1] Falls vor allem die Prognosefähigkeit der Beschreibung im Zentrum steht, kann ein automatisches Variablenselektionsverfahren eingesetzt werden. Wären wir vor allem an einer Erklärung der Ziel-Variablen interessiert, könnte ein rein automatischer Einsatz des Variablenselektionsverfahrens zu unbefriedigenden Resultaten führen.

person hat einen Faktor von 1.00). Die so erhaltene Zahl muss noch mit 0.64 multipliziert und anschliessend zu 1 addiert werden[1].

Tabelle 2: Multiplikativer Einfluss (jeweils in Klammern nach dem Merkmalstufennamen) der wichtigsten sozio-demografischen Merkmale auf die erwartete Anzahl Wegeketten pro Person und Tag

Personen im HH	1 (1.01), 2 (1.00), 3 (1.04), 4 (1.15), 5 (1.22), 6 und mehr (1.19)
Velo vorhanden	Immer (1.00), nach Absprache (0.86), nicht verfügbar (0.86), weiss nicht (0.50)
MIV vorhanden	Immer (1.00), nach Absprache (0.84), nicht verfügbar (0.78), kein Führerausweis / unter 18 (0.89)
Wochentag	Mo-Fr (1.00) , Sa (1.07), So (0.59)
Parkplatz am Arbeitsort	Ja - gratis (1.00), ja - bezahlt (0.84), nein (0.94), weiss nicht (0.87)
Arbeitsstatus	Anderes (1.50), arbeitslos (1.34), Ausbildung (1.07), Hausarbeit (1.43), weiss nicht (1.21), Rente (1.24), Teilzeit anderes (1.14), Teilzeit angestellt (1.25), Teilzeit Kader (1.27), Selbständig (1.34), Vollzeit anderes (1.11), Vollzeit angestellt (1.00), Vollzeit Kader (0.96), Vollzeit Selbständig (1.11)
Alter falls Arbeitstag	(0,16] (1.79), (16,20] (1.10), (20,35] (0.93), (35,64] (1.00), (64,100] (1.24)
Alter falls kein Arbeitstag[2]	(0,16] (0.81), (16,20] (0.80), (20,35] (0.93), (35,64] (0.99), (64,100] (0.89)

Legende: (0,16] ist gleichbedeutend mit (≥0 und <16)

Wichtig ist hier der Hinweis, dass die Modellierung so durchgeführt wurde, dass die Anzahl Wege, die eine Person an einem Tag zurücklegt, möglichst gut vorhergesagt werden kann. Ob diese Merkmale und nur diese einen kausalen Einfluss auf die Zielgrösse haben, kann jedoch aufgrund des Vorgehens nicht zwingend gefolgert werden.

4.4.3 Evaluation

Das Ergebnis der Poisson-Regression ist intuitiv im Grossen und Ganzen interpretierbar. Im Einzelnen kann die Interpretation von Effekten gewisser Faktorstufen schwierig sein, insbesondere wenn es sich um Stufen wie "weiss nicht"/"keine Angaben" oder "anderes" handelt. Insgesamt zeigt aber das vorliegende Beispiel, dass die Poisson-Regression ebenfalls Hinweise auf mögliche "kausale" Einflussvariablen zu liefern vermag.

4.5 Mobilitätstypen

4.5.1 Business Understanding, Data Understanding, Data Preparation

In diesem Abschnitt wird mit Data Mining Methoden geprüft, welche Mobilitätstypen aufgrund des Verkehrsmittelwahl-Verhaltens unterschieden werden können. Dazu wird die von einer Person gewählte Benutzung der Verkehrsmittel für die am Erhebungstag durchgeführten Wege betrachtet[3]. Die Wichtigkeit eines von einer Person verwendeten Verkehrsmittels kann mit der damit zurückgelegten Distanz, mit der damit verbrachten Zeit oder der Häufigkeit der Nutzung am Erhebungstag beschrieben werden. Jede dieser Betrachtungsweisen kann das eine oder andere Verkehrsmittel bevorzugen. Mit dem Au-

[1] Das Addieren von 1 ist notwendig, da wir die Anzahl Wegeketten minus 1 modelliert haben.

[2] Die Referenzperson hier ist „Alter (35,64] falls Arbeitstag“ aus der vorhergehenden Gruppe.

[3] Für die Charakterisierung von Mobilitätstypen ist uns kein zwingend „richtiges“ Vorgehen bekannt. Die Charakterisierung, die wir in dieser Arbeit verwenden, basiert auf der Verkehrsmittelwahl. Es wäre aber auch denkbar, die Abfolge der gewählten Verkehrsmittel mit zu berücksichtigen, wie dies z.B. in Schlich (2004, Kap. 6) gemacht wurde.

to legt man sicher meistens grössere Distanzen zurück als zu Fuss. Andererseits legt man vielleicht eher mehr Etappen zu Fuss zurück als mit dem Auto.

Wir haben uns entschlossen, für jede Person die am Beobachtungstag zurückgelegte totale Distanz, die für die Wege total aufgewendete Zeit sowie die Anzahl Etappen zu betrachten. Diese Merkmale bezeichnen wir als "absolute Merkmale". Zusätzlich wurde jeweils pro verwendetes Verkehrsmittel der Anteil an der Distanz, an der Zeit und an der Anzahl Etappen erfasst. Somit liegt eine zweite Gruppe von Merkmalen vor, die als relative Merkmale bezeichnet werden. Die Verkehrsmittel wurden in "öffentliche Verkehrsmittel (ÖV)", "motorisierte individuelle Verkehrsmittel (MIV)", "langsame Verkehrsmittel (LV)", welche das Velo und das zu Fuss gehen umfassen, sowie "andere", unterteilt.

Um nun mit Cluster-Analyse-Methoden Gruppen von Personen mit einem ähnlichen Mobilitätsverhalten zu suchen, müssen aufgrund der gerade entwickelten Merkmale Ähnlichkeiten oder Unähnlichkeiten zwischen allen Personen berechnet werden. Auch hierfür gibt es viele unterschiedliche Ansätze. Im vorliegenden Fall gab es keine Gründe, einen bestimmten Ansatz zu bevorzugen. Deshalb haben wir den einfachsten, die L1-Metrik (es werden die Beträge der Unterschiede in jedem einzelnen Merkmal aufsummiert), gewählt[1]. In den meisten Ansätzen ist es sinnvoll, die Merkmale in vergleichbaren Skalen zu erfassen. Bei den relativen Merkmalen ist das eher der Fall als bei den absoluten. Bei den absoluten Merkmalen (Dauer, Distanz) wird im Allgemeinen vorzugsweise der Unterschied relativ gemessen (z.B. Person A legt eine um 20% grössere Distanz zurück als Person B). Den gleichen Effekt kann man erreichen, indem die absoluten Merkmale logarithmiert und anschliessend davon die absoluten Unterschiede betrachtet. Dieses Vorgehen hat den Vorteil, dass alle Merkmale gleich behandelt werden können. Das Logarithmieren reicht aber noch nicht, da der Wertebereich im Vergleich zu den relativen Merkmalen immer noch zu unterschiedlich ist. Deshalb werden die logarithmierten absoluten Merkmale noch so skaliert, dass sie sich mit grosser Wahrscheinlichkeit wie die relativen in einem Wertebereich der Länge 1 ausbreiten. Dazu wird zusätzlich durch die dreifache Streuung dividiert. Diese Art der Bestimmung der Unähnlichkeiten mag komplex und teilweise willkürlich erscheinen. Aber es führt kaum ein Weg an solchen Überlegungen vorbei und in den aller meisten Fällen müssen Entscheide gefällt werden, die im Nachhinein höchstens durch eine erfolgreiche Clusteranalyse zu rechtfertigen sind.

Der vorhergehende Absatz legt dar, wie die Daten für die Clusteranalyse prinzipiell aufbereitet werden müssen. Beim konkreten Umsetzen wurden teilweise grobe Fehler in den Daten des Mikrozensus Verkehr festgestellt, welche das Löschen einzelner Datensätze erforderten[2].

4.5.2 Modeling

Die Clusteranalyse wurde je für die Kantone Thurgau und Zürich durchgeführt[3]. Vorgängig wurden verschiedene hierarchische Methoden getestet. Dabei hat sich gezeigt, dass das Verfahren von Ward die klarste Lösung liefert. Das Resultat lässt sich am einfachsten mit einem Dendrogramm präsentieren (vgl. Abbildung 6). Das ist ein auf den Kopf gestelltes Baumdiagramm. Jede Verzweigung zeigt an, wo sich jeweils zwei Cluster zu einem verschmelzen. Je länger die vertikalen Linien bis zur nächsten Verschmelzung sind, desto weiter sind die Cluster voneinander entfernt. Die Anzahl Cluster kann aus dem Dendrogramm herausgelesen werden, indem ein horizontaler Schnitt dort angebracht

[1] Weil die Charakterisierung nicht über das Aktivitätenmuster erfolgt, sondern rein auf den benutzten Verkehrsmitteln basiert, kommen die Unähnlichkeitsmasse, die in Schlich (2004, Kap. 6) vorgestellt werden, hier nicht in Frage.

[2] Wie allgemein üblich werden im Bericht darüber keine weiteren Worte verloren. Aber solche Plausibilisierungen brauchen viel Zeit. Da die gefundenen Unstimmigkeiten nirgends zentral festgehalten werden, müssen die Plausibilisierungen in jedem Projekt wieder neu vorgenommen werden, ohne Garantie, dass jeweils alle Unstimmigkeiten wieder gefunden werden.

[3] Ein Grund für diese Einschränkung liegt darin, dass unsere RAM-Speicherkapazitäten nicht ausreichten, um mit der Datenanalyse-Software R die hierarchische Clusteranalyse für die ganze Schweiz zu rechnen. Im Wesentlichen müssen für eine Clusteranalyse, in der die Unähnlichkeiten im Voraus gerechnet werden, n(n-1)/2 Unähnlichkeiten (n=Anzahl Personen) im RAM geladen werden, und dies bei ungefähr n=29'000 Personen. Solche Speicherplatz-Probleme können auch Gründe sein, auf andere Cluster-Analyse-Verfahren auszuweichen.

wird, wo ein Übergang von kurzen zu langen vertikalen Linien erfolgt[1]. Die jeweiligen Lösungen sind in Abbildung 6 mit einer horizontalen Linie eingezeichnet.

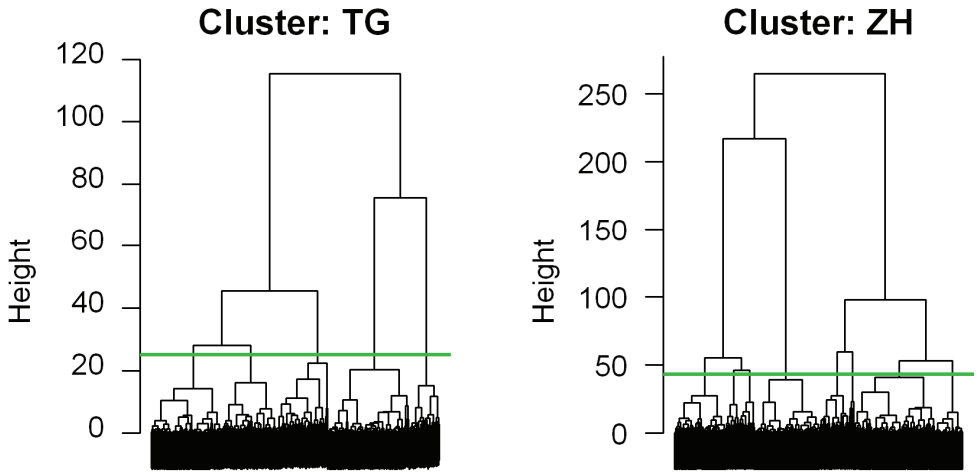

Abbildung 6: Dendrogramm der Resultate für die Kantone Thurgau (links) und Zürich (rechts). Die horizontale Linie markiert die gewählte Stelle, an welcher die ganze Stichprobe in Clusters unterteilt wird.

Wie im Abschnitt 2.3 dargelegt wurde, kann eine Clusteranalyse auch über die Visualisierung der Daten, möglichst in zwei Dimensionen, erfolgen. Eine Möglichkeit dazu besteht darin, die Methode der metrischen multidimensionalen Skalierung (MDS) auf die berechneten Unähnlichkeiten anzuwenden. Die Methode nützt aus, dass die Unähnlichkeiten zwischen z.B. 1'000 Personen im 999-dimensionalen euklidischen Raum dargestellt[2] werden können. Anschliessend sucht man jenen zweidimensionalen Unterraum, in dem die Daten am meisten streuen[3]. Das Resultat ist in Abbildung 7 dargestellt. Die Resultate der hierarchischen Clusteranalyse, durch die Einfärbung der Punkte in Abbildung 7 sichtbar gemacht, stimmen mit den Ansammlungen in dieser Darstellung recht gut überein.

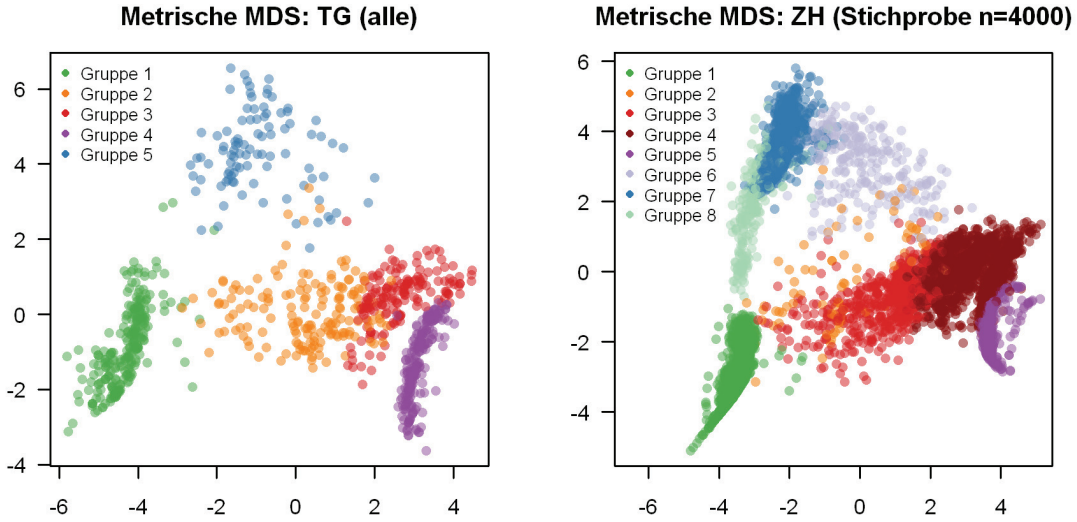

Abbildung 7: Zweidimensionale Repräsentation der Unähnlichkeiten für die Kantone Thurgau (links) und Zürich (rechts) mit Hilfe der metrischen multidimensionalen Skalierung (MDS). Die Einfärbung der Punkte widerspiegelt die Clusterzugehörigkeiten gemäss der hierarchischen Clusteranalyse mit der Methode von Ward (vgl. Abbildung 6).

[1] Oft findet sich keine eindeutige Stelle. Dann entscheidet man sich erst, nachdem man für verschiedene Varianten die Cluster zu interpretieren versucht hat.

[2] Es gibt hier noch einige Details, die im konkreten Fall genauer beachtet werden müssen.

[3] Das ist ein sinnvolles Vorgehen, falls die wichtigen Informationen dort liegen, wo die Streuung gross ist.

4.5.3 Evaluation

Eine erste Art der Evaluation erfolgt dadurch, dass man versucht, die Cluster durch die Merkmale der dazugehörigen Personen zu charakterisieren. Dazu werden für jeden Cluster jeweils die Mediane aller Merkmale bestimmt und in eine Grafik (vgl. Abbildung 8) eingetragen.

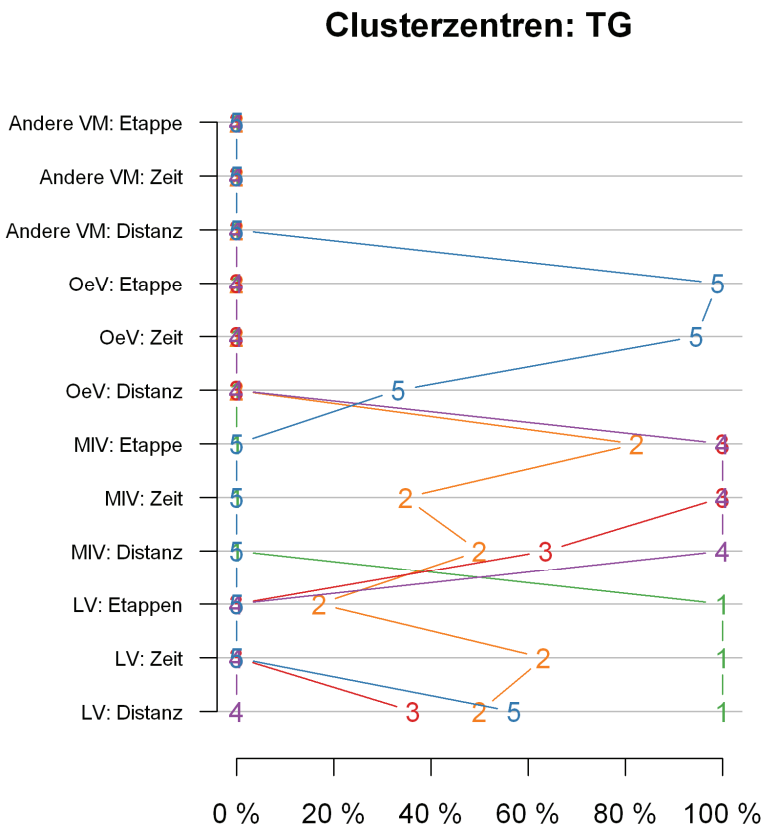

Abbildung 8: Median (x-Achse) der relativen Mobilitätsmerkmale[1] (y-Achse) für jeden Cluster aus der Analyse für den Kanton Thurgau

Wir stellten fest, dass sich die Cluster in den absoluten Merkmalen im Gegensatz zu den relativen Merkmalen fast nicht unterscheiden, weshalb die absoluten Merkmale in Abbildung 8 weggelassen wurden. Die Cluster sind mit Nummern identifiziert. Für den Thurgau lassen sie sich folgendermassen charakterisieren: Personen aus Cluster 1 benutzen langsame Verkehrsmittel (LV), Personen aus Cluster 2 benutzen etwa gleich häufig motorisierte individuelle Verkehrsmittel (MIV) wie LV, Personen aus Cluster 3 benutzen schwergewichtig MIV und wenig LV, Personen aus Cluster 4 benutzen fast ausschliesslich MIV und Personen aus Cluster 5 benutzen neben LV öffentliche Verkehrsmittel (ÖV). Vor allem in Cluster 4 ist ein sehr extremes Verhalten charakterisiert, das wahrscheinlich so über eine längere Zeitperiode gar nicht vorkommt. Der Grund liegt in der Erhebungsart: Das Mobilitätsverhalten einer Person wird nur aufgrund ihres Verhaltens am Erhebungstag erfasst. Da kann es gut möglich sein, dass eine Person z.B. den ganzen Tag mit dem Auto unterwegs ist, obwohl sie üblicherweise auch andere Verkehrsmittel benutzt. Deshalb sind diese Ergebnisse in der Praxis[2] nur bedingt nutzbar. Ähnliche Resultate zeigen sich auch für den Kanton Zürich.

[1] Die Abkürzungen auf der y-Achse haben folgende Bedeutung: VM: Verkehrsmittel; ÖV: Öffentliche Verkehrsmittel; MIV: Motorisierte individuelle Verkehrsmittel; LV: Velo und zu Fuss.

[2] Dies ist eine typische Konsequenz, die sich daraus ergibt, dass Fragen anhand von Daten untersucht werden, für die sie nicht zwingend geeignet sind. Data Mining setzt aber gerade diesbezüglich vorerst kaum Grenzen.

4.6 Klassifikationsregeln für die Mobilitätstypen

4.6.1 Business Understanding, Data Understanding, Data Preparation

In diesem Anwendungsbeispiel wird versucht, den Mobilitätstyp einer Person aus deren sozio-demografischen Merkmalen und aus Raumstrukturinformationen zu den Wohn- und Zielorten vorherzusagen. Dazu werden die Daten aus dem Mikrozensus Verkehr 2005 und die Raumstrukturdaten zusammengeführt. Dies geschieht einerseits dadurch, dass der jeweilige Wohnort der Person mit Raumstrukturinformationen beschrieben wird. Andererseits werden auch alle Zielorte, die die Person während des Erhebungstages besucht hat, adäquat beschrieben, indem die Raumstrukturinformationen für die einzelnen Zielorte geeignet zusammengefasst werden. Die Details zu allen verwendeten Variablen finden sich im Anhang III. Dieser Ansatz kann allerdings nur für die Raumstrukturdaten auf Stufe Gemeinde durchgeführt werden. Bei vielen Zielorten fehlen die Raumkoordinaten – vermutlich wegen unklaren Angaben bei der Erhebung – und entsprechend können deren lokalen Raumstrukturinformationen nicht ermittelt werden.

Einer Person ist am Morgen meistens klar, ob sie zur Arbeit oder Schule gehen oder einer anderen Tätigkeit nachgehen wird. Meist weiss sie auch schon, ob sie nur eine kleine Distanz (z.B. weniger als 2 km) oder eine grössere zurücklegen wird. Solche Informationen zum Mobilitätsverhalten sind nicht direkt in den Datensätzen vorhanden. Es ist jedoch möglich, aus der Information über den Zweck des Weges auf Mobilitätsentscheide zu schliessen, wenn davon ausgegangen wird, dass der Weg zur Arbeit oder zur Schule nicht spontan festgelegt wird. Ähnlich kann bei der Distanz vorgegangen werden, obwohl sich auch längere Wege spontan ergeben können und somit am Morgen nicht immer klar ist, ob im Laufe des Tages insgesamt nur eine kurze Distanz zurückgelegt werden soll. Trotz dieser Einschränkungen verwenden wir diese konstruierte Information und schauen, inwiefern interessante, plausible Erkenntnisse (eigentlich besser als Hypothesen zu bezeichnen) gewonnen werden können.

4.6.2 Modeling

Die Modellierung erfolgte zweistufig, je mit einem Random Forest. In einem ersten Schritt wurde der Random Forest eingesetzt, um mit dem darin automatisch berechneten Wichtigkeitsmass ("Importance") die für die Vorhersage relevanten Merkmale herauszufiltern[1]. Von den vielen Merkmalen wurden für den Kanton Thurgau 37 wichtige Merkmale selektioniert. In einer zweiten Runde wurde der Random Forest nur mit diesen selektionierten Merkmalen gerechnet. Wie wichtig die einzelnen Merkmale in der zweiten Runde sind, ist in der Abbildung 9 grafisch dargestellt. Auf der Abszisse (x-Achse) sind die Merkmale angegeben, auf der Ordinate (y-Achse) die Clusternummern. Die Wichtigkeit eines Merkmals ist durch die Fläche der Kreisscheibe angezeigt.

Auffällig in Abbildung 9 ist, dass zur Vorhersage der Clusterzugehörigkeit für die Cluster 5 (ÖV) und 1 (LV) viele Merkmale wichtig sind, hingegen für die drei Cluster 2, 3 und 4 (MIV) nur wenige. Insbesondere Raumstrukturmerkmale zum Wohnort und zum Zielort sind für diese zweite Gruppe von Clustern von geringer Wichtigkeit. Für Cluster 2 gibt es keine Merkmale, die besonders wichtig sind. Bei der Evaluation wird sich dann zeigen, dass dieser Cluster auch am schlechtesten vorhergesagt werden kann.

Um die Resultate in Abbildung 9 besser einordnen zu können, müssen folgende Aspekte berücksichtigt werden: Einerseits werden durch den Random Forest nur jene Merkmale einbezogen, die helfen, den Mobilitätstyp möglichst gut vorherzusagen. Die einbezogenen Merkmale müssen dabei nicht notwendigerweise einen kausalen Einfluss auf das Mobilitätsverhalten haben (auch müssen die nicht einbezogenen Merkmale nicht zwangsläufig keinen kausalen Einfluss ausüben). Andererseits ist zu vermuten, dass sich gewisse erwartete Effekte deshalb nicht zeigen, weil die Mobilitätstypen (d.h. Cluster) aufgrund des Mobilitätsverhaltens an einem einzigen Stichtag konstruiert worden sind. Die Cluster

[1] Es ist für das Endergebnis sehr entscheidend, welche Variablen man zur Vorhersage zulässt. Je nachdem werden andere Variablen als wichtig angesehen und selektioniert. Man kann das "Spiel" sogar so weit treiben, dass man die Modellierung wiederholt, bis am Ende jene Variablen als wichtig selektioniert werden, die man aus fachlicher Sicht gerne sehen möchte. In diesem Sinne sind die Ergebnisse im Data Mining eben explorativ und nicht abschliessend.

sind also aufgrund des kurzfristigen Verhaltens zusammengestellt worden. Grundsätzliche Einstellungen zur Mobilität resp. Wertvorstellungen, welche auch das langfristige Verhalten (z.B. in Bezug auf die Wahl des Wohnortes und des Besitzes von Mobilitätswerkzeugen) beeinflussen, können vermutlich das momentane (an einem Stichtag festgestellte) Verhalten zu wenig klar durchdringen.

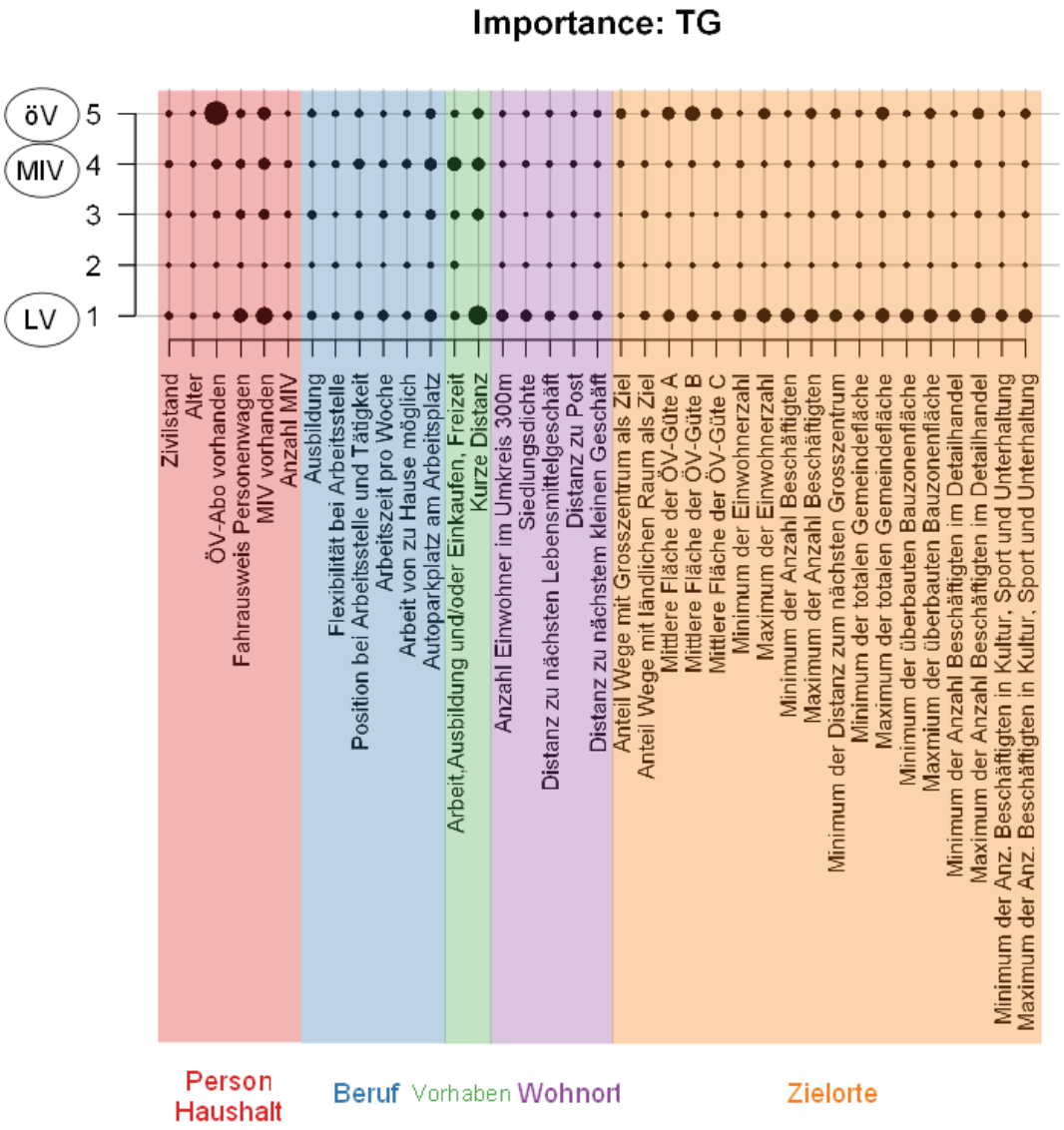

Abbildung 9: Wichtigkeit der Merkmale basierend auf der zweiten Runde des Random Forest. Je wichtiger die Variable für die Vorhersage der Klassenzugehörigkeit ist, desto grösser ist die entsprechende Kreisscheibe.

4.6.3 Evaluation

Wie gut sich mit der Random Forest Methode die Zugehörigkeit einer Person zu einem Mobilitätstyp aufgrund der sozio-demografischen und der Raumstruktur-Merkmale vorhersagen lässt, kann mit der Fehlerrate, die sich aus der Konfusionsmatrix berechnen lässt, beurteilt werden. Die Konfusionsmatrix ist in Tabelle 3 dargestellt. Die Bestimmung der Fehlerrate erfolgte mit Beobachtungen, die nicht bei der Anpassung verwendet wurden. Beim Random Forest wird diese Bestimmung "out-of bag" (OOB)[1] genannt. Die Fehlerrate ist mit 49% recht hoch. Man muss aber dabei berücksichtigen, dass reines Raten bei fünf Klassen zu einer Fehlerrate von 80% führt. Wird die Konfusionsmatrix im Detail

[1] Bei OOB handelt es sich um eine ganz spezielle Art der Kreuzvalidierung, die ausnützt, dass bei den einzelnen Bäumen nicht alle Objekte zur Erzeugung des Baumes verwendet werden, sondern nur eine Zufallsauswahl davon.

studiert, kann festgestellt werden, dass vor allem die Cluster 2, 3 und 4 schlecht auseinander gehalten werden können. Legt man diese drei Cluster zusammen, halbiert sich die totale Fehlerrate auf 23%.

Tabelle 3: Konfusionsmatrix und Fehlerrate für den Random Forest

	1	2	3	4	5	Cluster spez. Fehlerrate		Cluster spez. Fehlerrate (zus'gelegt)
1	192	25	4	22	4	22.27%		22.27%
2	51	32	33	40	3	79.87%	}	
3	16	18	62	61	6	61.96%		20.26%
4	28	27	44	108	4	48.82%	}	
5	15	8	6	10	50	43.82%		43.82%
Totale Fehlerrate (OOB)						48.91%		23.25%

5 Data Mining Software

Es existiert eine fast unüberblickbare Vielfalt von Software-Lösungen zum Data Mining. Entsprechende Übersichtsseiten[1] listen Dutzende von sogenannten "Suites" auf. Diese lassen sich grob in proprietäre und in frei verfügbare, oft mit einsehbarem Quellcode versehene (d.h. Open Source), Pakete, einteilen. In beiden Gruppen gibt es sowohl schlanke, einfache, auf einzelne spezielle Analyseschritte und Spezialanwendungen zugeschnittene Tools, wie auch sehr umfangreiche, mächtige Pakete, welche eine Menge an Optionen bieten.

Eine wesentliche Trennlinie verläuft zwischen den eigentlichen Data Mining Suites, bei denen der Anwender durch den Data Mining Prozess[2] geführt wird und solchen, die mehr auf einzelne Tasks fokussieren oder zwar umfangreich sind, jedoch keine Prozessstruktur bieten. Diese letzte Gruppe ohne Struktur umfasst auch sämtliche auf der Kommandozeile basierenden Methoden.

Viele der Suites verwenden grafische User Interfaces (GUIs). Sind diese knotenbasiert, lassen sie das einfache, bequeme und schnelle Aneinanderreihen von Modulen für Datenvorverarbeitung, Modellierung, Analyse, Validierung und Visualisierung zu. Meist gehen solche GUIs mit einer Gruppierung der Module einher, d.h. man wird gleichzeitig durch den Data Mining Prozess geführt. Einige einfachere auf GUIs basierende Pakete sind nicht knotenbasiert, sondern erlauben nur das Ansteuern verschiedener Routinen via "point and click".

Die (teilweise kostspieligen) kommerziellen Tools erlauben in der Regel die Bearbeitung deutlich grösserer Datenmengen und bieten eine grössere Anzahl an Methoden, teilweise interaktive Routinen zur Datenvorbereitung, Grafiken mit Punktidentifikation, (semi-)automatische Modellvalidierungen sowie die Möglichkeit zum Export von Scoring Codes für das effiziente Deployment im betrieblichen Alltag.

Es ist nicht möglich, unter der Vielfalt von Möglichkeiten ein "bestes" Tool zu identifizieren, zu unterschiedlich sind die Bedürfnisse und die Vorbildung der Benutzer sowie die Anwendungszwecke. Es gilt, den Spagat zwischen dem Angebot von möglichst vielen Möglichkeiten und der Übersichtlichkeit sowie einer einfachen, intuitiven Bedienbarkeit zu berücksichtigen. Gerade in den kommerziellen Tools werden häufig breit geeignete Default-Einstellungen gesetzt, die vom Laien unangetastet bleiben können und dem Profi dennoch volle Flexibilität bieten.

Wir stellen in der Folge eine Auswahl von Software-Tools vor, die wir für die Anwendung im Verkehrsbereich als geeignet erachten, und welche für uns verfügbar waren.

a) SAS Enterprise Miner
Es handelt sich um eine proprietäre Suite mit beträchtlichen Lizenzkosten. Sie basiert auf einem knotengesteuerten GUI, welches den Anwender gemäss der dem CRISP verwandten Data Mining Strategie SEMMA (Sample, Explore, Modify, Model, Assess) durch die Analyse führt. Es stehen nicht nur eine Vielzahl an statistischen Methoden zur Verfügung, sondern auch benutzerfreundliche Möglichkeiten zur (interaktiven) Datenaufbereitung und Support für Batch-Prozesse sowie Scoring Code. Die Vorteile des Enterprise Miners liegen auch im effizienten Umgang mit grossen Datenmengen.

b) IBM SPSS Modeler (ehemals Clementine)
Auch hier handelt es sich um eine proprietäre Software mit beträchtlichen Lizenzkosten. Sie verfügt über ein knotengesteuertes GUI, das den Anwender gemäss CRISP durch den Data Mining Prozess führt. Die Anzahl der zur Verfügung stehenden Methoden ist etwas geringer als beim Enterprise Miner, dafür stufen wir die Bedienung als etwas intuitiver ein. Als Stärke des Modelers ist die Verfügbarkeit von weiterer Software zu sehen,

[1] Siehe z.B. http://www.kdnuggets.com/software/suites.html

[2] Im Sinne von CRISP (siehe Kapitel 2.2), oder ähnlichen, verwandten Formulierungen.

die eine sehr effiziente Anbindung an Datenbanken und eine Automatisierung von Tasks ermöglicht.

c) KNIME
Der KNIME ist eine frei verfügbare Open Source Software, welche ebenfalls auf einem knotengesteuerten GUI basiert. Auch hier wird man durch den Data Mining Prozess geführt, wenn auch ein bisschen weniger klar und deutlich als bei a) und b). Es ist eine ausreichende Anzahl an Data Mining Methoden vorhanden. Ausserdem besteht die Möglichkeit, weitere Methoden und ganze Knoten hinzuzufügen. Zudem kann das gesamte Erscheinungsbild umprogrammiert werden. Der Umgang mit grossen Datenmengen und die Möglichkeiten zur Operationalisierung von Tasks sind jedoch gegenüber den Tools a) und b) deutlich eingeschränkt. Insgesamt handelt es sich um eine Software, welche in Design und Auftreten ähnlich zu a) und b) ist, jedoch auf allen Ebenen weniger Möglichkeiten bietet.

d) Rattle
Der Name "Rattle" steht für "R Analytical Tool To Learn Easily". Es handelt sich um ein für die Kommandozeilensprache R (siehe e)) programmiertes, nicht knoten- sondern menübasiertes GUI, welches ein beschränktes Data Mining Toolkit anbietet. Man wird auch hier grob durch den Data Mining Prozess geführt, wenn auch mässig intuitiv. Die zur Verfügung stehenden Funktionalitäten und Methoden sind eingeschränkt, so dass der erfahrene Anwender schnell an die Grenzen stösst. Es handelt sich um ein einfaches Tool, welches einem wenig erfahrenen Anwender den Start ins Data Mining ohne Kostenfolge und lange Einarbeitungszeit erleichtern kann.

e) R
R ist eine frei verfügbare Open Source Software für alle Arten von statistischen Analysen und grafischen Darstellungen. Es handelt sich um eine Kommandozeilensprache, welche nicht in erster Linie für Data Mining ausgelegt ist. Vom erfahrenen, mit dem Data Mining Prozess vertrauten Anwender lassen sich aber auch solche Analysen bestens durchführen. Es steht eine sehr umfangreiche Bibliothek an Zusatzpaketen zur Verfügung, so dass fast jede denkbare Analysemethode möglich ist. Allerdings muss man aufgrund des Charakters von R mit einer erheblichen Einarbeitungszeit rechnen und etwas Programmiererfahrung mitbringen. Weiter ist R eher für sogenanntes "Prototyping" geeignet. Die Operationalisierung von Tasks lässt sich weniger einfach bewerkstelligen als bei a) und b). Zudem sind für die Bearbeitung von sehr grossen Datensätzen gewisse Limiten vorhanden.

f) Zusammenfassung
Zusammenfassend kann festgestellt werden, dass eine grosse Anzahl von Software-Paketen zum Data Mining existiert. Die meisten verfügen über Module zum Einlesen und Aufbereiten von Daten, für supervised und unsupervised Learning, für die Validierung und Visualisierung von Daten usw. Die Wahl eines spezifischen Produkts hängt vom Einsatzzweck (z.B. einfache explorative Analyse oder professionell operationalisierter Analysetask) und von den Kenntnissen des Anwenders (verfügt er lediglich über Kenntnisse der Grundzüge der statistischen Datenanalyse oder handelt es sich um einen Experten mit ausführlichen Programmierkenntnissen?) ab.

6 Schlussfolgerungen

6.1 Zur Anwendbarkeit der Methode in der Verkehrsplanung

Als Fazit der Studie kann festgestellt werden, dass Data Mining in der Verkehrsplanung grundsätzlich anwendbar ist. Dies bestätigen sowohl die in der Literatur berichteten Erfahrungen (siehe Kapitel 3) als auch die in dieser Studie durchgeführten Anwendungsbeispiele. Allerdings hat sich auch gezeigt, das Data Mining nicht – wie vielleicht erhofft – automatisch auf alle Fragen gute Antworten liefert oder aus vorhandenen Datensätzen ohne Dazutun des Anwenders neue Erkenntnisse generiert.

Wie in anderen Sachgebieten hängt auch in der Verkehrsplanung der Nutzen von Data Mining in erster Linie davon ab, wie geeignet die vorhandenen Daten zur Beantwortung der interessierenden Fragen sind. Dieser Aspekt bestätigt sich auch in dieser Arbeit. Obwohl die Werkzeuge des Data Mining erfolgreich eingesetzt werden konnten, bleiben bei der Interpretation der konkreten Ergebnisse gewisse Unsicherheiten bezüglich Brauchbarkeit. Ein Hauptgrund liegt sicher darin, dass jeweils Datensätze verwendet wurden, die nicht im Hinblick auf die Beantwortung der hier gestellten Fragen erhoben wurden. Beim Mikrozensus Verkehr handelt es sich um eine Querschnittserhebung. Zur Beantwortung der hier gestellten Fragen wären aber Zeitreihendaten, z.B. aus einer Panelstudie, besser geeignet,a sich das Mobilitätsverhalten einer Person mit einer Befragung, welche die Verkehrsteilnahme an nur an einem einzigen Tag erfasst, nicht vollständig erheben lässt.

Zur Anwendung von Data Mining in der Verkehrsplanung haben wir generell folgende Eindrücke gewonnen:

- Die Datenaufbereitung kann sehr aufwändig sein, weil sie sich nicht automatisieren lässt. Sie muss auch sehr sorgfältig durchgeführt werden, da die Qualität der Daten zentral für die Brauchbarkeit der Resultate ist (auch hier gilt: "garbage in – garbage out").
- Es stehen verschiedene "automatisierte" Methoden zum Modellieren der Daten bereit; für die Anpassung der Modelle benötigt man kaum vertiefte technische Kenntnisse, weil die Algorithmen der neusten Generation fast ohne Projekt-spezifische Feinanpassungen auskommen. Dieser Aspekt ist für die Anwendung durch Verkehrsplaner nur vorteilhaft. Trotzdem muss man wissen, wie mit den Methoden umzugehen ist, wo ihre Grenzen sind und was die Algorithmen grundsätzlich als Resultat liefern.
- Weil man in der Verkehrsplanung gut mit der klassischen statistischen Modellierung vertraut ist, sei hier nochmals darauf hingewiesen, dass Data Mining zwar oft auch aus der Statistik bekannte Methoden einsetzt, jedoch den Fokus mehr auf die Vorhersage von Zielgrössen richtet oder auf das Auffinden von Gruppen von Objekten mit ähnlichem Verhalten. Im ersten Fall werden regressionsartige Modelle, im zweiten Fall oft auch Verfahren aus der multivariaten explorativen Statistik eingesetzt.
- Bei den regressionsartigen Modellen ist weiter zu beachten, dass im Data Mining die für die Vorhersage notwendigen Merkmale datenbasiert selektioniert werden. Ob diese selektionierten Merkmale auch einen ursächlichen Einfluss auf die Zielgrösse ausüben, bleibt dabei dahingestellt.

6.2 Lehren aus den Fallbeispielen

Ziel der Fallbeispiele war die Demonstration der Anwendung des Data Mining Prozesses an typischen Fragestellungen aus der Verkehrsplanung mit Datensätzen, wie sie dem Verkehrsplaner üblicherweise zur Verfügung stehen. Das Entdecken *völlig neuer* Erkenntnisse aus den Daten zum Mobilitätsverhalten stand dabei – entgegen dem, was der Titel der Forschungsarbeit evtl. erwarten lässt – nicht im Vordergrund.

6.2.1 Datenaufbereitung

Die wichtigsten aus den Fallbeispielen gewonnenen Erkenntnisse hinsichtlich des Einsatzes der Data Mining-Methoden lassen sich wie folgt zusammenfassen:

- Wie bei allen Datenanalysen muss auch beim Data Mining eine gewisse Vertrautheit mit den zur Verfügung stehenden Daten vorhanden sein, d.h. es muss klar sein, welcher Art die Daten sind (Querschnitts- oder Zeitreihendaten, Repräsentativität der Stichprobe etc.) und was die Bedeutung der einzelnen Variablen ist.
- Die Daten müssen möglichst fehlerfrei und vollständig sein. Um dies sicherzustellen, sind Plausibilitätstests und unter Umständen umfangreiche Datenbereinigungen erforderlich.
- Die Daten müssen in die für die einzelnen Methoden geeignete Form (z.B. Matrix statt Liste) gebracht werden.
- Für gewisse Methoden, z.B. Clusteranalyse, sollten die Variablen der zu analysierenden Datensätze in vergleichbaren Skalen vorliegen. Dies ist bei relativen Merkmalen eher der Fall als bei absoluten. Um die Spannweite des Wertebereichs von absoluten Variablen (z.B. Dauer oder Distanz von Wegen) einzugrenzen, sind geeignete Transformationen wie Logarithmieren, Dividieren durch ein Vielfaches des Streubereiches usw.) ins Auge zu fassen.

Insgesamt haben die Fallbeispiele bestätigt, dass allein die Aufbereitung der Daten einen nicht zu unterschätzenden Anteil des Data Mining Prozesses ausmachen kann.

6.2.2 Anwendbarkeit der Methoden und Ergebnisse

In den Fallbeispielen wurden Antworten auf die folgenden Fragestellungen gesucht:

a) Welche Wegeketten-Typen treten innerhalb der Gesamtpopulation am häufigsten auf?
b) Wie gross ist der Anteil der am häufigsten auftretenden Wegeketten-Typen innerhalb der nach Altersklassen und Geschlecht differenzierten Personengruppen?
c) Lässt sich die Anzahl Wegeketten pro Person und Tag durch sozio-demographische Merkmale erklären?
d) Welche Mobilitätstypen bezüglich Verkehrsmittelwahl lassen sich unterscheiden?
e) Lässt sich die Zugehörigkeit einer Person zu einem bestimmten Mobilitätstyp mit deren sozio-demographischen Merkmalen und den Informationen zur Raumstruktur am Wohn- und am Zielort erklären?

Die Behandlung der Fragestellung a) bedurfte keiner speziellen Methode. Hier ging es primär um die Datenaufbereitung, bei welcher aus der in Listenform vorliegenden Wegedatei des Mikrozensus Verkehr für jede Person die durchgeführte Wegekette zu bilden war. Die Häufigkeiten der Wegeketten-Typen (charakterisiert durch die Abfolge der Wegezwecke) konnten dann durch einfaches Auszählen gewonnen werden. Interessant am Ergebnis ist, dass sich die am häufigsten vorkommenden Wegeketten-Typen aus lediglich 2 Wegen (Hin- und Rückweg) zusammensetzen und dass komplexe Wegeketten eher selten sind.

Die in Fragestellung b) gesuchte Zuordnung der Wegeketten-Typen zu den nach Geschlecht und Alter differenzierten Personengruppen bot methodisch keine Probleme. Hier ging es vor allem darum, die Möglichkeit zu demonstrieren, mit Mosaikplots den Inhalt von Kreuztabellen sehr anschaulich darzustellen.

Am Beispiel der Fragestellung c) wurde der Einsatz der Poisson-Regression demonstriert. Zu erklärende Variable (auch Zielvariable oder response variable genannt) war die Anzahl der Wegeketten, welche eine Person am Stichtag in Abhängigkeit ihrer sozio-demographischen Eigenschaften und des Wochentages durchgeführt hat. Die Variablenselektion für die schrittweise Regressionsrechnung erfolgte dabei nach einer ersten "Startvorgabe" von subjektiv wichtig erscheinenden Variablen nach rein statistischen Kriterien, ohne Vorgaben durch den Anwender. Dieses automatisierte Vorgehen ist einfach anwendbar und resultiert in einem Modell, welches die statistisch signifikanten Zusammenhänge aufzeigt. Bezüglich Kausalitäten resp. Plausibilitäten bleiben im vorliegenden Fall jedoch Fragen offen. Die gefundenen Zusammenhänge können aber wertvolle Hin-

weise im Hinblick auf die Erstellung eines praxistauglichen erklärenden Modelles liefern. In diesem Sinne ist die Poisson-Regression weiterhin eine wertvolle und in der Verkehrsplanung durchaus anwendbare Methode – wie jede Regression liefert sie aber keine Ergebnisse, welche unbesehen zur Vorhersage der Zielvariablen verwendet werden können.

Die Clusteranalyse, wie sie zur Beantwortung der Fragestellung d) eingesetzt wurde, ist eine schon bisher in der Verkehrsplanung verbreitet angewendete Methode, insbesondere auch zur Klassifikation von Verkehrsteilnehmern nach Verhaltensgruppen. Weniger bekannt ist die Methode der metrischen multidimensionalen Skalierung (MDS). Diese hat sich im Fallbeispiel als gut anwendbar und sehr geeignet für die graphische Darstellung der Clusterzugehörigkeiten erwiesen. Die MDS hat gegenüber der Verwendung von Dendrogammen den Vorteil der besseren Visualisierung und damit der einfacheren Unterscheidung der hauptsächlichen Cluster. Für die Erkennung der wichtigsten Unterscheidungsmerkale zwischen den Clustern ist die Darstellung der Mediane der Merkmale eine einfach anwendbare Methode mit gut interpretierbaren Ergebnissen.

Die Klassifikation einer Stichprobe resp. das Erkennen von Clustern allein bildet auch in der Verkehrsplanung selten ein Endresultat, sondern ist Ausgangspunkt für weitergehende Analysen. Von Interesse ist beispielsweise die Vorhersage der Zugehörigkeit einer Person zu einem bestimmten Mobilitätstyp (Cluster) mit erklärenden Variablen (auch Prädiktoren genannt), gemäss der Fragestellung e). Dieses Fallbeispiel diente dazu, die Anwendung der bisher in der Verkehrsplanung noch wenig bekannte Methode des Random Forest zu demonstrieren. Da diese Methode weitgehend automatisch funktioniert, handelt es sich um ein relativ einfach anwendbares aber mächtiges Instrument zur Gewinnung von Klassifikationsregeln. Ein interessantes Anwendungsgebiet für die Methode des Random Forest in der Verkehrsplanung ist beispielsweise die explorative Datenanalyse im Vorfeld der Etablierung von Verkehrsmodellen, wenn es darum geht, die massgeblichen im Modell zu berücksichtigenden Variablen zu identifizieren.

6.3 Empfehlungen

Die Literaturrecherche und die Fallbeispiele zeigen das breite Spektrum von Anwendungsmöglichkeiten und das Potential der vielfältigen Methoden von Data Mining. Im Rahmen der Verkehrsplanung ist die Erwägung eines Einsatzes von Data Mining namentlich für die folgenden Problemstellungen zu empfehlen:

- Klassifikation von Daten zum Mobilitätsverhalten, inkl. Ermittlung von Klassifikationsregeln.
- Visualisierung mehrdimensionaler Daten zur raschen Erkennung von Mustern resp. Clustern (z.B. Mosaikplot, MDS usw.)
- Rasche und automatische Erkennung der (aus statistischer Sicht) wichtigsten Prädiktoren für die Vorhersage einer Zielvariablen (z.B. mit Regression, künstlichen neuronalen Netzen oder Random Forest usw.)
- Analyse der Entscheidungsprozesse, z. B. bei der Aktivitätenplanung und der daraus folgenden Verkehrsteilnahme

Da weder die Klassifikationsregeln noch die Methoden zur automatischen Erkennung von statistisch signifikanten Prädiktoren die Kausalitäten zwischen diesen und der Zielvariablen berücksichtigen, können diese Data Mining Methoden die konventionellen Methoden zur Modellschätzung nicht ersetzen. Vielmehr wird – auch in der Literatur – empfohlen, konventionelle Modellansätze und Data Mining als sich ergänzende und sich gegenseitig unterstützende Methoden einzusetzen, wobei Data Mining insbesondere Hinweise auf Prädiktoren liefern soll, deren Berücksichtigung im Modell prüfenswert ist.

Wie für die traditionellen statistischen Methoden gilt auch für Data Mining, dass die erzielbaren Ergebnisse umso besser sein werden, je besser der Anwender die folgenden Voraussetzungen erfüllt:

- Vertrautheit mit den verkehrsplanerischen Fragestellungen (Business Understanding)
- Vertrautheit mit den vorhandenen Daten (Data Understanding)
- Vertrautheit mit den Data Mining-Methoden

Da für die meisten Verkehrsplaner Data Mining – auch wegen des damit verbundenen Aufwandes – kaum zur täglichen Praxis gehören wird, dürfte es wegen des mangelnden "Trainings" in den meisten Fällen an der notwendigen Vertrautheit mit den Werkzeugen von Data Mining fehlen. Auf der anderen Seite werden Data Mining Experten mit den verkehrsplanerischen Fragestellungen und den vorhandenen Daten eher weniger vertraut sein. Um den grössten Nutzen aus der Anwendung von Data Mining in der Verkehrsplanung zu erzielen, empfiehlt sich daher eine Zusammenarbeit zwischen Verkehrsplanern und Data Mining Experten.

6.4 Weiterer Forschungsbedarf

Die Entwicklung und Verbesserung der Data Mining Methoden ist ein laufender Prozess ausserhalb der Verkehrsforschung. Aufgrund der im Rahmen dieser Studie gemachten Erfahrungen sind insbesondere Methoden für den Umgang mit sehr umfangreichen Datenmengen resp. zur Lösung der bestehenden Speicherplatz-Probleme gesucht.

Im Übrigen besteht für die Anwendung von Data Mining in der Verkehrsplanung im Moment kaum Forschungsbedarf. Um Data Mining in Zukunft auch in der Verkehrsplanung nutzbringend einsetzen zu können, wären vielmehr möglichst viele weitere praktische Anwendungen erwünscht, aus denen sowohl Verkehrsplaner als auch Data Mining Experten lernen und die interdisziplinäre Zusammenarbeit sowie das gegenseitige Verständnis weiter vertiefen können.

Anhänge

I Verwendete Raumstrukturdaten

I.1 Raumstrukturdaten der Gemeindedaten

	Name	Typ	Spaltenfo	Dezim	Variablenlabel	W
1	ID	Numerisch	11	0		Kei
2	BFS2007	Numerisch	11	0	BFS Gemeindenummer (Gemeindestand 2007)	Kei
3	Gemeindename2007	String	24	0	Gemeindename (Gemeindestand 2007)	Kei
4	rg_VK_ARE	Numerisch	11	0	Verkehrliche Raumgliederung (V1-V5) des ARE	{1,
5	EW_2007	Numerisch	11	0	Einwoherzahl Stand 2007	Kei
6	Besch_2005	Numerisch	11	0	Anzahl der Beschaeftigten Stand 2005	Kei
7	Dist_Grossz	Numerisch	12	11	Distanz zum nächsten Grosszentrum	Kei
8	Fl_Gem	Numerisch	11	5	Totale Flaeche der Gemeinde (ha)	Kei
9	Fl_BZ_ueberbaut	Numerisch	11	3	ueberbaute_Bauzonenflaeche (ha)	Kei
10	Fl_OEV_Q_A	Numerisch	11	0	Flaeche_OEV_Qualitaet_A (ha)	Kei
11	Fl_OEV_Q_B	Numerisch	11	0	Flaeche_OEV_Qualitaet_B (ha)	Kei
12	Fl_OEV_Q_C	Numerisch	11	0	Flaeche_OEV_Qualitaet_C (ha)	Kei
13	Fl_OEV_Q_D	Numerisch	11	0	Flaeche_OEV_Qualitaet_D (ha)	Kei
14	Fl_OEV_Q_E	Numerisch	11	0	Flaeche_OEV_Qualitaet_E (ha)	Kei
15	Bes_EK	Numerisch	11	0	Anz_Beschaeftigte im Detailhandel	Kei
16	Bes_KSU	Numerisch	11	0	Anzahl Beschäftigte in K_S_U (BZ05)	Kei
17	Bes_FRZ	Numerisch	11	0	Anzahl Beschäftigte im Freizeitbereich ("Hauri2000)	Kei

I.1.1 Beschreibung der Variablen

BFS2007: BFS-Gemeindenummer (Gemeindestand 2007)

Gemeindename2007: Gemeindename (Gemeindestand 2007)

rg_VK_ARE: Verkehrliche Raumgliederung (V1-V5) des ARE aus dem Jahre 2002:
1: Grosszentren
2: Nebenzentren der Grosszentren oder Mittelzentren mit Anschluss ans nationale Bahnnetz
3: Nebenzentren der Grosszentren oder Mittelzentren ohne Anschluss ans nationale Bahnnetz
4: Gütelgemeinden der Agglomerationen
5: Ländliche Gemeinden (inkl. Kleinzentren)

EW_2007: Einwohnerzahl pro Gemeinde 2007, Gebietsstand: 31. Dezember 2007;
Quelle: ESPOP, Datei „Ständige Wohnbevölkerung nach Gden 1981-2007.xls“

Besch_2005: Anzahl der Beschäftigten 2005;
Quelle: Betriebszählung 2005,
Datei „Beschäftigte_2005.xls“

Dist_Grossz: Distanz zum nächstgelegenen Grosszentrum in der Schweiz in km;
Gemeindezentren gemäss Gebietsstand 2007

Fl_Gem: Totale Fläche der Gemeinde in Hektaren;
berechnet aus Gemeindeshape 2007;
Quelle: Geodaten BFS, Datei „gd-b-00.02-883-gg07g1“

Fl_BZ_ueberbaut:	Überbaute Fläche in der Bauzone 2007; Quelle: Bundesamt für Raumentwicklung ARE, Bauzonenstatistik Schweiz 2007; Datei „Bauzonenstatistik_Schweiz_2007_bereinigt.xls“
Fl_ÖV_Q_A:	Bauzonenfläche in ÖV-Güteklasse A (Sehr gute ÖV-Erschliessung); Quelle: Bundesamt für Raumentwicklung ARE, Bauzonenstatistik Schweiz 2007; Datei „Bauzonenstatistik_Schweiz_2007_bereinigt.xls“
Fl_ÖV_Q_B:	Bauzonenfläche in ÖV-Güteklasse B (Gute ÖV-Erschliessung); Quelle: Bundesamt für Raumentwicklung ARE, Bauzonenstatistik Schweiz 2007; Datei „Bauzonenstatistik_Schweiz_2007_bereinigt.xls“
Fl_ÖV_Q_C:	Bauzonenfläche in ÖV-Güteklasse C (Mittelmässige ÖV-Erschliessung); Quelle: Bundesamt für Raumentwicklung ARE, Bauzonenstatistik Schweiz 2007; Datei „Bauzonenstatistik_Schweiz_2007_bereinigt.xls“
Fl_ÖV_Q_D:	Bauzonenfläche in ÖV-Güteklasse D (Geringe ÖV-Erschliessung); Quelle: Bundesamt für Raumentwicklung ARE, Bauzonenstatistik Schweiz 2007; Datei „Bauzonenstatistik_Schweiz_2007_bereinigt.xls“
Fl_ÖV_Q_E:	Bauzonenfläche in ÖV-Güteklasse E (Marginale oder keine ÖV-Erschliessung); Quelle: Bundesamt für Raumentwicklung ARE, Bauzonenstatistik Schweiz 2007; Datei „Bauzonenstatistik_Schweiz_2007_bereinigt.xls“
Bes_EK:	Anzahl Beschäftigte im Detailhandel: Summe der Beschäftigten (Vollzeitäquivalente) im Detailhandel pro Gemeinde (Gebietsstand 2007); Quelle: Betriebszählung 2005, Hektarrasterdaten gemäss Datei „BZ05_AST_Besch_Abt_NOGA2002_V070630P“, Variable „B0552VZA“
Bes_KSU:	Anzahl Beschäftigte im „Freizeitbereich“: Summe der Beschäftigten (Vollzeitäquivalente) im Bereich Kultur, Sport und Unterhaltung pro Gemeinde (Gebietsstand 2007); Quelle: Betriebszählung 2005, Hektarrasterdaten gemäss Datei „BZ05_AST_Besch_Abt_NOGA2002_V070630P“, Variable „B0592VZA“

I.2 Raumstrukturdaten der Start- und Zielorte der Wege

Diese Datei beinhaltet verschiedene Variable zu den Start- und Zielorten der einzelnen Wege aus dem Mikrozensus Verkehr 2005.

Es sind nur Orte berücksichtigt, bei denen alle Variablen erfasst werden konnten. Damit fallen vor allem jene Start- und Zielorte weg, welche in der Datei „MZ05_Geodaten“ des Mikrozensus nicht vorhanden sind (Grund ungenaue oder fehlende Angaben zum Zielort bei der Erhebung der Wege oder Startort kommt nicht als Zielort vor).

	Name	Typ	Spaltenformat	Dezimalstell	Variablenlabel	Wer
1	Key_Ort	Numerisch	2	0	Code für Start- und Zielort	Kein
2	Key_Koordinaten	Numerisch	15	0	Key_Koordinaten = 1000000 * X_abs + Y_abs als Key	Kein
3	X_abs	Numerisch	9	0	X-Koordinaten absolut	Kein
4	Y_abs	Numerisch	9	0	Y-Koordinaten absolut	Kein
5	EW_300m	Numerisch	4	0	Anzahl Einwohner im Umkreis 300m (Stand 2000)	Kein
6	VZA_S2_300m	Numerisch	5	2	Anzahl Beschäftigte (Vollzeitäquivalente) Sektor 2 im Umkreis 300m (Stand 2005)	Kein
7	VZA_S3_300m	Numerisch	5	2	Anzahl Beschäftigte (Vollzeitäquivalente) Sektor 3 im Umkreis 300m (Stand 2005)	Kein
8	Abstand_Autobahn	Numerisch	5	2	Abstand zum nächsten Autobahnanschluss	Kein
9	Abstand_Gemeindezentrum	Numerisch	4	2	Abstand zum Gemeindezentrum	Kein
10	OeV_Gueteklasse	String	3	0	ÖV-Güteklasse	Kein
11	Besch_Detailhandel_800m	Numerisch	4	2	Anzahl Beschäftigte (VZA) im Detailhandel im Umkreis 800 m (Stand 2005)	Kein
12	Besch_Freizeit_800m	Numerisch	4	2	Anzahl Beschäftige (VZA) in Kultur, Sport und Unterhaltung im Umkreis 800 m (Stand 2005)	Kein

I.2.1 Beschreibung der Variablen

Key_Ort: Code für die Verknüpfung mit der Wegedatei (betrifft Start- und Zielorte)

Key_Koordinaten: Beschreibt den Ort mit den x- und y Koordinaten (jeweils Stellen nach dem Komma abgeschnitten mit der Funktion „Trunc“), dient nur zu Kontrollzwecken als zusätzliche Information zum Ort

X_abs: X-Koordinate des Ortes (Stellen nach dem Komma abgeschnitten), dient nur zu Kontrollzwecken als zusätzliche Information zum Ort

Y_abs: Y-Koordinate des Ortes (Stellen nach dem Komma abgeschnitten), dient nur zu Kontrollzwecken als zusätzliche Information zum Ort

EW_300m: Beschreibung der Einwohnerdichte in der nahen Umgebung des Ortes: Summe der Einwohner im Umkreis von 300 m des Ortes (alle Hektaren berücksichtigt, deren Mittelpunkt nicht weiter als 300m vom Ort entfernt sind); Quelle: Volkszählung 2000, Hektarrasterdaten gemäss Datei „Vz2000_dsview“, Variable „P00BTOT“

VZA_S2_300m: Beschreibung der Arbeitsplatzdichte (im Sektor 2) in der nahen Umgebung des Ortes: Summe der Beschäftigten (Vollzeitäquivalente) Sektor 2 im Umkreis von 300 m des Ortes (alle Hektaren berücksichtigt, deren Mittelpunkt nicht weiter als 300m vom Ort entfernt sind); Quelle: Betriebszählung 2005, Hektarrasterdaten gemäss Datei „BZ05_AST_Besch_S_NOGA2002_V070630P“, Variable „B05VZAS2“

VZA_S3_300m	Beschreibung der Arbeitsplatzdichte (im Sektor 3) in der nahen Umgebung des Ortes: Summe der Beschäftigten (Vollzeitäquivalente) Sektor 3 im Umkreis von 300 m des Ortes (alle Hektaren berücksichtigt, deren Mittelpunkt nicht weiter als 300m vom Ort entfernt sind); Quelle: Betriebszählung 2005, Hektarrasterdaten gemäss Datei „BZ05_AST_Besch_S_NOGA2002_V070630P“, Variable „B05VZAS3“
Abstand_Autobahn:	Erreichbarkeit des Nationalstrassennetzes: Distanz zu Autobahnanschlüssen; Quelle: Mikrozensus Verkehrsverhalten 2005, Zuordnung von Raummerkmalen: Datei „MZ05_Zielort_Wegeinland_Geodaten.sav“, Variable „Dist_AB_AS“
Abstand_Gemeindezentrum:	Erreichbarkeit des Gemeindezentrums: Distanz zum Gemeindezentrum (Kirchturmkoordinaten); Quelle: Mikrozensus Verkehrsverhalten 2005, Zuordnung von Raummerkmalen: Datei „MZ05_Zielort_Wegeinland_Geodaten.sav“, Variable „Dist_Gmd_Zentrum“
ÖV_Gueteklasse:	Erreichbarkeit mit dem ÖV: ÖV-Güteklasse (A-D) oder keine Erschliessung (E) am Ort gemäss Fahrplan 2006/2007; Quelle: Mikrozensus Verkehrsverhalten 2005, Zuordnung von Raummerkmalen: Datei „MZ05_Zielort_Wegeinland_Geodaten.sav“, Variable „ÖV_Guete_06_07“
Besch_Detailhandel_800m:	Beschreibung der Erreichbarkeit von Einkaufsflächen in der Umgebung des Ortes: Summe der Beschäftigten (Vollzeitäquivalente) im Detailhandel im Umkreis von 800 m des Ortes (alle Hektaren berücksichtigt, deren Mittelpunkt nicht weiter als 800m vom Ort entfernt sind); Quelle Betriebszählung 2005, Hektarrasterdaten gemäss Datei „BZ05_AST_Besch_Abt_NOGA2002_V070630P“, Variable „B0552VZA“
Besch_Freizeit_800m:	Beschreibung der Erreichbarkeit von Freizeiteinrichtungen in der Umgebung des Ortes: Summe der Beschäftigten (Vollzeitäquivalente) im Bereich Kultur, Sport und Unterhaltung im Umkreis von 800 m des Ortes (alle Hektaren berücksichtigt, deren Mittelpunkt nicht weiter als 800 m vom Ort entfernt sind); Quelle Betriebszählung 2005, Hektarrasterdaten gemäss Datei „BZ05_AST_Besch_Abt_NOGA2002_V070630P“, Variable „B0592VZA“

I.2.2 Raumstrukturdaten des Wohnortes

Diese Datei beinhaltet Variablen zur Beschreibung der Wohnortsituation der Zielpersonen. Es sind wiederum nur Orte aufgeführt, bei denen alle Variablen erfasst werden konnten. Damit fallen vor allem jene weg, welche bei der Wohnortangabe nicht die Daten-Qualitätsstufe A erreicht haben.

Name	Typ	Spaltenfo...	Dezimals...	Variablenlabel	
Pers_ID	Numerisch	8	0	Personen_ID	K
HHNR	Numerisch	8	0	Haushaltsnummer	K
ZIELPNR	Numerisch	8	0	Zielpersonennummer	K
W_X	Numerisch	8	2	Wohnort: X-Koordinate	{-
W_Y	Numerisch	8	2	Wohnort: Y-Koordinate	{-
W_QAL	Numerisch	8	0	Wohnort: Erreichte Qualitätsstufe	{-
W_X_abs	Numerisch	8	0	X-Koordinaten absolut	K
W_Y_abs	Numerisch	8	0	Y-Koordinaten absolut	K
EW_300m	Numerisch	9	0	Anzahl Einwohner im Umkreis 300m (Stand 2000)	K
VZA_S2_300m	Numerisch	19	8	Anzahl Beschäftigte (Vollzeitäquivalente) Sektor 2 im Umkreis 300m (Stand 2005	K
VZA_S3_300m	Numerisch	19	8	Anzahl Beschäftigte (Vollzeitäquivalente) Sektor 3 im Umkreis 300m (Stand 2005)	K
Besch_Detailhandel_800m	Numerisch	9	0	Anzahl Beschäftigte (VZA) im Detailhandel im Umkreis 800 m (Stand 2005)	K
Besch_Freizeit_800m	Numerisch	9	0	Anzahl Beschäftige (VZA) in Kultur, Sport und Unterhaltung im Umkreis 800 m (S...	K
Dist_Hlstr_HS	Numerisch	20	6	Distanz zu Hauptstrassen	{-
Dist_Hlstr_erg	Numerisch	20	6	Distanz zu Strassen des Ergänzungsnetzes	{-
Siedlungsdichte	Numerisch	11	0	Siedlungsdichte in 1 Km Umkreis (Radius = 1 km)	{-
Dist_Siedlungszentrum	Numerisch	20	6	Lage im Siedlungsraum: Distanz zum nächsten Zentrum gemäss roCH_SPV	{-
Typ_naechst_SZ	Numerisch	11	0	Typ des nächstegelgenen Siedlungszentrums	{-
Dist_Gmd_Zentrum	Numerisch	20	6	Distanz zum Gemeindezentrum (Kirchturmkoordinaten)	{-
OeV_Guete_06_07	String	3	0	ÖV-Güteklasse A-D oder keine ÖV-Erschliessung (E), mit Fahrplan 2006/2007 b...	{-
Dist_Haltestellen	Numerisch	20	6	Distanz zur nächsten Haltestelle aus dem Tür2Tür-Fahrplan	{-
Dist_Superm	Numerisch	20	6	Distanz zum nächsten Lebensmittelgeschäft gemäss BZ Selektion "BZ_Superma...	{-
Dist_Banken	Numerisch	20	6	Distanz zur nächsten Bankfiliale	{-
Dist_Post	Numerisch	20	6	Distanz zur nächsten Poststelle	{-
Dist_Arztpraxen	Numerisch	20	6	Distanz zur nächsten Arztpraxis	{-
Dist_Spitaeler	Numerisch	20	6	Distanz zum nächsten Spital	{-
Dist_oblig_Schulen	Numerisch	20	6	Distanz zur nächsten Primar- oder Oberstufenschule sowie zu Kindergärten	{-
Dist_Sportanlagen	Numerisch	20	6	Distanz zur nächsten Sportanlage	{-
Dist_Fitness	Numerisch	20	6	Distanz zum nächsten Fitnesszentrum	{-
Dist_Zoos	Numerisch	20	6	Distanz zum nächsten Zoo, bot. Garten oder Naturpark	{-
Dist_Kinos	Numerisch	20	6	Distanz zum nächsten Kino	{-
Dist_Theater	Numerisch	20	6	Distanz zum nächsten Theater	{-
Dist_Museen	Numerisch	20	6	Distanz zum nächsten Museum	{-
Dist_Apotheken	Numerisch	20	6	Distanz zur nächsten Apotheke	{-
Dist_AB_AS	Numerisch	20	6	Distanz zum nächsten Autobahnanschluss	{-
Dist_Beck	Numerisch	20	6	Distanz zur nächsten Bäckerei	{-
Dist_EZ	Numerisch	20	6	Distanz zum nächsten Einkaufszentrum	{-
Dist_BZ_Det_80_150	Numerisch	20	6	Distanz zum nächsten Detailhandelsstandort mit 80-150 Beschäftigten	{-
Dist_BZ_Det_40_80	Numerisch	20	6	Distanz zum nächsten Detailhandelsstandort mit 40-80 Beschäftigten	{-
Dist_BZ_Det_20_40	Numerisch	20	6	Distanz zum nächsten Detailhandelsstandort mit 20-40 Beschäftigten	{-
Dist_BZ_Det_kl20	Numerisch	20	6	Distanz zum nächsten Detailhandelsstandort mit 1-20 Beschäftigten	{-
Dist_BZ_Restaurant	Numerisch	20	6	Distanz zum nächsten Restaurant	{-

I.2.3 Beschreibung der Variablen

Pers_ID: Code für die Verknüpfung mit der Personendatei

W_X_abs: X-Koordinate des Ortes (Stellen nach dem Komma abgeschnitten), dient nur zu Kontrollzwecken als zusätzliche Information zum Ort

W_Y_abs: Y-Koordinate des Ortes (Stellen nach dem Komma abgeschnitten), dient nur zu Kontrollzwecken als zusätzliche Information zum Ort

EW_300m:	Beschreibung die Einwohnerdichte in der nahen Umgebung des Ortes: Summe der Einwohner im Umkreis von 300 m des Ortes (alle Hektaren berücksichtigt, deren Mittelpunkt nicht weiter als 300m vom Ort entfernt sind); Quelle: Volkszählung 2000, Hektarrasterdaten gemäss Datei „Vz2000_dsview“, Variable „P00BTOT“
VZA_S2_300m:	Beschreibung der Arbeitsplatzdichte (im Sektor 2) in der nahen Umgebung des Ortes: Summe der Beschäftigten (Vollzeitäquivalente) Sektor 2 im Umkreis von 300 m des Ortes (alle Hektaren berücksichtigt, deren Mittelpunkt nicht weiter als 300m vom Ort entfernt sind); Quelle: Betriebszählung 2005, Hektarrasterdaten gemäss Datei „BZ05_AST_Besch_S_NOGA2002_V070630P“, Variable „B05VZAS2“
VZA_S3_300m:	Beschreibung der Arbeitsplatzdichte (im Sektor 3) in der nahen Umgebung des Ortes: Summe der Beschäftigten (Vollzeitäquivalente) Sektor 3 im Umkreis von 300 m des Ortes (alle Hektaren berücksichtigt, deren Mittelpunkt nicht weiter als 300m vom Ort entfernt sind); Quelle: Betriebszählung 2005, Hektarrasterdaten gemäss Datei „BZ05_AST_Besch_S_NOGA2002_V070630P“, Variable „B05VZAS3“
Besch_Detailhandel_800m:	Beschreibung der Erreichbarkeit von Einkaufsflächen in der Umgebung des Ortes: Summe der Beschäftigten (Vollzeitäquivalente) im Detailhandel im Umkreis von 800 m des Ortes (alle Hektaren berücksichtigt, deren Mittelpunkt nicht weiter als 800m vom Ort entfernt sind); Quelle Betriebszählung 2005, Hektarrasterdaten gemäss Datei „BZ05_AST_Besch_Abt_NOGA2002_V070630P“, Variable „B0552VZA“
Besch_Freizeit_800m:	Beschreibung der Erreichbarkeit von Freizeiteinrichtungen in der Umgebung des Ortes: Summe der Beschäftigten (Vollzeitäquivalente) im Bereich Kultur, Sport und Unterhaltung im Umkreis von 800 m des Ortes (alle Hektaren berücksichtigt, deren Mittelpunkt nicht weiter als 800 m vom Ort entfernt sind); Quelle: Betriebszählung 2005, Hektarrasterdaten gemäss Datei „BZ05_AST_Besch_Abt_NOGA2002_V070630P“, Variable „B0592VZA“

Die restlichen Variablen sind aus der Datenbank „Mikrozensus Verkehrsverhalten 2005, Zuordnung von Raummerkmalen“ entnommen.

II Anpassungen für das Wegekettenmodell

(siehe Kapitel 4.4)

Um auf die Werte in der Tabelle 2 in Kapitel 4.4 zu kommen, müssen die Werte in der Spalte "Estimate" der untenstehenden Tabelle mit der Exponential-Funktion (exp(x)) transformiert werden. Die Angaben zu den Standardabweichungen in den geschätzten Parametern (Spalte Std. Error) und die z- und Pr-Werte müssen mit grosser Vorsicht interpretiert werden. Dies deshalb, weil in diesem Fall das Modell aus einer Variablenselektion hervorgegangen ist, die ebenfalls zu statistischen Unsicherheiten führt. In den Standardausgaben (Tabelle 2) zu einer (Poisson-)Regression werden diese Unsicherheiten jedoch nie mitberücksichtigt, sondern nur die Unsicherheiten, die sich aus der Schätzung ergeben. Zweites muss bei einer Diskussion dieser Tabelle auch zwischen statistischer Signifikanz und fachlicher Relevanz unterschieden werden. Für die Prognosefähigkeit des Modells sind diese Aspekte jedoch nur von untergeordneter Bedeutung.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.45	0.03	-15.35	< 2e-16
Alter(0,16]	0.58	0.07	8.92	< 2e-16
Alter(16,20]	0.09	0.06	1.46	0.144347
Alter(20,35]	-0.08	0.03	-2.43	0.015123
Alter(35,64]	0			
Alter(64,100]	0.21	0.09	2.36	0.018070
Arbeitsstatus_anderes	0.41	0.09	4.45	0.000009
Arbeitsstatus_Arbeitslos	0.29	0.06	5.06	0.000000
Arbeitsstatus_Ausbildung(Schule,Studium,Lehre)	0.07	0.05	1.32	0.188636
Arbeitsstatus_Hausarbeit	0.36	0.05	7.41	0.000000
Arbeitsstatus_Keine Angabe/Weiss nicht	0.19	0.3	0.63	0.532208
Arbeitsstatus_RentnerIn	0.21	0.05	3.88	0.000105
Arbeitsstatus_Teilzeit | Anderes	0.13	0.23	0.57	0.570916
Arbeitsstatus_Teilzeit | Angestellt	0.22	0.03	7.57	0.000000
Arbeitsstatus_Teilzeit | Kader	0.24	0.06	4.4	0.000011
Arbeitsstatus_Teilzeit | Selbststaendig	0.29	0.06	5.26	0.000000
Arbeitsstatus_Vollzeit | Anderes	0.1	0.25	0.41	0.685838
Arbeitsstatus_Vollzeit | Angestellt	0			
Arbeitsstatus_Vollzeit | Kader	-0.04	0.03	-1.4	0.162805
Arbeitsstatus_Vollzeit | Selbststaendig	0.11	0.04	2.91	0.003572
MIVvorhImmer	0			
MIVvorhNach Absprache verfuegbar	-0.18	0.03	-5.98	0.000000
MIVvorhNicht verfuegbar	-0.24	0.03	-9	< 2e-16
MIVvorhNicht mobil/<18Jahre/kein Fuehrerausweis	-0.12	0.05	-2.59	0.009655
VELOvorhImmer	0			
VELOvorhNach Absprache Verfuegbar	-0.15	0.04	-3.55	0.000388
VELOvorhNicht verfuegbar	-0.14	0.02	-7.25	0.000000
VELOvorhWeiss nicht	-0.7	0.1	-7.38	0.000000
HHanzpers1	0.01	0.02	0.26	0.798114
HHanzpers2	0			
HHanzpers3	0.03	0.03	1.22	0.222808
HHanzpers4	0.14	0.02	5.76	0.000000

HHanzpers5	0.2	0.03	6.43	0.000000
HHanzpers6+	0.18	0.05	3.73	0.000194
ARBEITppautoJa, gratis	1			
ARBEITppautoJa, bezahlt	-0.18	0.04	-5.01	0.000001
ARBEITppautoNein	-0.07	0.03	-2.48	0.012975
ARBEITppautoKeine Angabe/Weiss nicht	-0.14	0.04	-3.7	0.000214
ArbeitstagFALSE	-0.01	0.03	-0.51	0.613142
Wochentag_Mo-Fr	0			
Wochentag_Sa	0.06	0.02	2.65	0.007944
Wochentag_So	-0.53	0.03	-16.96	< 2e-16
Alter(0,16]:arbeitstagFALSE	-0.78	0.05	-14.79	< 2e-16
Alter(16,20]:arbeitstagFALSE	-0.3	0.09	-3.35	0.000795
Alter(20,35]:arbeitstagFALSE	0.02	0.04	0.43	0.666402
Alter(35,64]:arbeitstagFALSE	0			
Alter(64,100]:arbeitstagFALSE	-0.31	0.09	-3.35	0.000812

Null deviance: 29802 on 27890 degrees of freedom

Residual deviance: 27972 on 27852 degrees of freedom

III Für die Klassifikation nach Mobilitätstypen verwendete Daten

(siehe Kapitel 4.5)

Variablenbezeichnung	Beschreibung	Herkunft
Zweck eines Weges	U Umsteigen, Verkehrsmittelwechsel, Auto abstellen A Arbeiten S Ausbildung, Schule E Einkaufen B Besorgungen, Inanspruchnahme von Dienstleistungen G Geschäftliche Tätigkeit R Rückkehr nach Hause bzw. auswärtige Unterkunft (wzweck2 = 'Nachhauseweg') a Anderes	aus Wegedatensatz
Personen im HH	Anzahl Personen im Haushalt: 1, 2, 3, 4, 5, 6 und mehr	aus Personendatensatz
Velo vorhanden	Ist ein Fahrrad vorhanden: Immer, nach Absprache, nicht verfügbar, nicht mobil	aus Personendatensatz
MIV vorhanden	Ist ein Auto vorhanden: Immer, nach Absprache, nicht verfügbar, nicht mobil	aus Personendatensatz
Wochentag	Stichtag, an welchem Wochentag wurden die Wege zurückgelegt: Mo-Fr, Sa-So	aus Personendatensatz
Parkplatz am Arbeitsort	Parkplatz am Arbeitsort vorhanden: Ja - gratis, ja - bezahlt, nein, weiss nicht	aus Personendatensatz
Arbeitsstatus	Berufstätigkeit: Ausbildung, Hausarbeit, weiss nicht, Rente, Teilzeit anderes, Teilzeit angestellt, Teilzeit Kader, selbstständig, Vollzeit anderes, Vollzeit angestellt, Vollzeit Kader, Vollzeit selbstständig, Anderes, arbeitslos	aus Personendatensatz
Alter falls Arbeitstag	Altersgruppe falls Stichtag ein Arbeitstag war: (0, 16], (16, 20], (20, 35], (35, 64], (64, 100]	aus Personendatensatz
Alter falls kein Arbeitstag	Altersgruppe falls Stichtag kein Arbeitstag war: (0, 16], (16, 20], (20, 35], (35, 64], (64, 100]	aus Personendatensatz
Andere VM: Etappe	Anteil der Etappen mit anderen Verkehrsmitteln	aus Etappendatensatz
Andere VM: Zeit	Anteil der Zeit (Dauer) mit anderen Verkehrsmitteln	aus Wegedatensatz
Andere VM: Distanz	Anteil der Distanz mit anderen Verkehrsmitteln	aus Wegedatensatz
ÖV: Etappe	Anteil der Etappen mit öffentlichem Verkehr	aus Etappendatensatz
ÖV: Zeit	Anteil der Zeit (Dauer) mit öffentlichem Verkehr	aus Wegedatensatz
ÖV: Distanz	Anteil der Distanz mit öffentlichem Verkehr	aus Wegedatensatz
MIV: Etappe	Anteil der Etappen mit motorisiertem Individualverkehr	aus Etappendatensatz
MIV: Zeit	Anteil der Zeit (Dauer) mit motorisiertem Individualverkehr	aus Wegedatensatz
MIV: Distanz	Anteil der Distanz mit motorisiertem Individualverkehr	aus Wegedatensatz

LV: Etappe	Anteil der Etappen mit Langsamverkehr	aus Etappendatensatz
LV: Zeit	Anteil der Zeit (Dauer) mit Langsamverkehr	aus Wegedatensatz
LV: Distanz	Anteil der Distanz mit Langsamverkehr	aus Wegedatensatz
Zivilstand	Zivilstand: ledig, verheiratet, getrennt, geschieden, verwitwet, keine Angabe/ weiss nicht/<16Jahre	aus Personendatensatz
ÖV-Abo vorhanden	Abo für den Öffentlichen Verkehr: Halbtaxabo, GA, beschränktes Abo (z.B. Streckenabo), kein Abo, Halbtax und beschränktes Abo, <6 Jahre	aus Personendatensatz
Fahrausweis Personenwagen	Fahrausweis Personenwagen ja/nein	aus Personendatensatz
Anzahl MIV	Anzahl Autos	aus Personendatensatz
Ausbildung	Zuletzt abgeschlossene Ausbildung: Keine Ausbildung abgeschlosen, Obligatorische Schule, Berufsschule, Maturität/PrimarlehrerInnenausbildung, höhere Ausbildung, keine Angabe/ weiss nicht/<16Jahre	aus Personendatensatz
Flexibilität bei Arbeitsstelle	Fest vorgegebene Zeiten für Arbeitsbeginn und Arbeitsende, feste Blockzeiten, fest vorgegebene Anzahl Stunden pro Woche oder Monat, total flexible Arbeitsstunden, keine Angabe/weiss nicht/arbeitet nicht	aus Personendatensatz
Position bei Arbeitsstelle und Tätigkeit	Position bei der Arbeit und Arbeitszeitmodell: andere Tätigkeit, andere Tätigkeit und Teilzeit, andere Tätigkeit und Vollzeit , Teilzeit angestellt, Vollzeit angestellt, arbeitslos, Ausbildung (Schule, Studium, Lehre), Hausarbeit, Kader und Teilzeit, Vollzeit-Kader, Teilzeit-Selbstständig, Vollzeit-Selbstständig, keine Angabe/weiss nicht, RentnerIn	aus Personendatensatz
Arbeitszeit pro Woche	Arbeitszeit pro Woche: [1,32), [32,41), [41,44), [44,100), keine Angabe/weiss nicht/arbeitet nicht	aus Personendatensatz
Arbeit von zu Hause möglich	Ein Teil der Arbeit kann zu Hause erledigt werden: Ja, manchmal, nein, keine Angabe/weiss nicht/arbeitet nicht	aus Personendatensatz
Vorhaben Arbeit, Ausbildung und/oder Einkaufen, Freizeit	War das Vorhaben dieses Stichtages mit Arbeit/Ausbildung, Einkaufen/Freizeit, oder Arbeit/Ausbildung UND Einkaufen/Freizeit verbunden?	aus Wegedatensatz
Kurze Distanz zurückgelegt	Distanz kleiner als 2 km, ja/nein	aus Wegedatensatz
Anz. Einwohner im Umkreis 300m	Anzahl Einwohner am Wohnort im Umkreis 300m	Wohnortdatensatz
Siedlungsdichte	Siedlungsdichte am Wohnort in 1 km Umkreis	Wohnortdatensatz
Distanz zu nächsten Lebensmittelgeschäft	Distanz zum nächsten Lebensmittelgeschäft am Wohnort	Wohnortdatensatz
Distanz zu Post	Distanz zur nächsten Post am Wohnort	Wohnortdatensatz
Distanz zu nächstem kleinen Geschäft	Distanz zum nächsten Detailhandelsstandort am Wohnort mit 1-20 Beschäftigten	Wohnortdatensatz
Anteil Wege mit Grosszentrum als Ziel	Anteil Wege mit Ziel Grosszentrum an allen Wegen	Raummerkmaledatensatz Wohnort
Anteil Wege mit ländlichen Raum als Ziel	Anteil Wege mit Ziel ländlicher Raum an allen Wegen	Raummerkmaledatensatz Wohnort

Mittlere Fläche der ÖV-Güte A	Mittlere Fläche der ÖV-Güte A über alle Zielorte	Raummerkmaledatensatz Wohnort
Mittlere Fläche der ÖV-Güte B	Mittlere Fläche der ÖV-Güte B über alle Zielorte	Raummerkmaledatensatz Wohnort
Mittlere Fläche der ÖV-Güte C	Mittlere Fläche der ÖV-Güte C über alle Zielorte	Raummerkmaledatensatz Wohnort
Minimum der Einwohnerzahl	Minimum der Einwohnerzahl über alle Zielorte	Raummerkmaledatensatz Wohnort
Maximum der Einwohnerzahl	Maximum der Einwohnerzahl über alle Zielorte	Raummerkmaledatensatz Wohnort
Minimum der Anz. Beschäftigten	Minimum der Anzahl Beschäftigten über alle Zielorte	Raummerkmaledatensatz Wohnort
Maximum der Anz. Beschäftigten	Maximum der Anzahl Beschäftigten über alle Zielorte	Raummerkmaledatensatz Wohnort
Minimum der Distanz zum nächsten Grosszentrum	Minimum der Distanz zum nächsten Grosszentrum über alle Zielorte	Raummerkmaledatensatz Wohnort
Minimum der totalen Gemeindefläche	Minimum der totalen Fläche der Gemeinde über alle Zielorte	Raummerkmaledatensatz Wohnort
Maximum der totalen Gemeindefläche	Maximum der totalen Fläche der Gemeinde über alle Zielorte	Raummerkmaledatensatz Wohnort
Minimum der überbauten Bauzonenfläche	Minimum der überbauten Bauzonen-fläche über alle Zielorte	Raummerkmaledatensatz Wohnort
Maxmium der überbauten Bauzonenfläche	Maximum der überbauten Bauzonen-fläche über alle Zielorte	Raummerkmaledatensatz Wohnort
Minimum der Anz. Beschäftigten im Detailhandel	Minimum der Anzahl Beschäftigten im Detailhandel über alle Zielorte	Raummerkmaledatensatz Wohnort
Maximum der Anz. Beschäftigten im Detailhandel	Maximum der Anzahl Beschäftigten im Detailhandel über alle Zielorte	Raummerkmaledatensatz Wohnort
Minimum der Anz. Beschäftigten in Kultur, Sport und Unterhaltung	Minimum der Anzahl Beschäftigten im Kultur, Sport und Unterhaltung über alle Zielorte	Raummerkmaledatensatz Wohnort
Maximum der Anz. Beschäftigten in Kultur, Sport und Unterhaltung	Maximum der Anzahl Beschäftigten im Kultur, Sport und Unterhaltung über alle Zielorte	Raummerkmaledatensatz Wohnort

Abkürzungen

Begriff	Bedeutung
AB	AdaBoost
AIS-Algorithmus	Agrawal-Imielinski-Swami-Algorithmus
ARE	Bundesamt für Raumentwicklung
BFS	Bundesamt für Statistik
BIC	Bayes Information Criterion
C&RT	Classification and Regression Trees
CHAID	Chi-square Automatic Interaction Detectors
CRISP	Cross Industry Standard Prozess for Data Mining
CT	Classification Tree
DT	Decision Tree
EM	Expectation Maximization
FCM	Fuzzy C-Means
FP	Frequent Pattern
GA	Generalized Algorithm
GAM	Generalized Additive Model
GLM	Generalized Linear Model
GUI	Graphical User Interface
GVK-CH	Gesamtverkehrskonzeption Schweiz
HH	Haushalt
IIA	Independence of Irrelevant Alternatives
KEP	Kontinuierliche Erhebung Personenverkehr
KDD	Knowledge Discovery in Database
KI	Künstliche Intelligenz
LV	Langsamverkehr
MARS	Multivariate Adaptive Regression Splines
MDS	Multidimensionale Skalierung
MDSAM	Multi-Dimensional Sequence Alignment Method
MIV	Motorisierter Individualverkehr
MLP ANN	Mulitlayer Perceptron Artificial Neural Network
MNL	Multinomial Logit Model
NB	Naïve Bayes
NLM	Nested Logit Model
NM	Neural Network
OOB	Out-of- Bag
ÖV	Öffentlicher Verkehr
RBF	Radial Basis Function
SAM	Sequence Alignment Method
SBB	Schweizerische Bundesbahnen
SEMMA	Sample, Explore, Modify, Model, Assess
SOM	Self-Organizing Maps
SVM	Support Vector Machine

Literaturverzeichnis

Arentze T., F. Hofmanm, N. Kalfs and H. Timmermans (1999), *System for Logical Verification and Inference of Activity (SYLVIA) Diaries*. Transportation Research Record 1660, National Academy Press, Washington D.C., 1999, pp. 156-163

Biagioni J.P., P.M. Szezurek, P.C. Nelson und A. Mohammadian (2009), *Tour-based Mode Choice Modeling: Using An Ensemble of (Un-) Conditional Data-Mining Classifiers*. TRB 2009 Annual Meeting CD-ROM

Bishop Ch. M.(2007), *Pattern Recognition and Machine Learning*, Springer-Verlag, Berlin

Dennerlein R. und A. Gebert (1990), *Personenverkehr in der Schweiz, Verkehrselastizitäten*. Endbericht an den Stab für Gesamtverkehrsfragen des EVED, April 1990

Duda R.O., Hart P.E. und Stork D.G (2000), *Pattern Classification*, 2nd Ed, John Wiley & Sons

Gabriel R., Gluchowski P. und Pastwa A. (2009), *Datawarehouse und Data Mining*, W3l

Han J. and Kamber M. (2006), *Data Mining: Concepts and Techniques*, 2nd Ed, Morgan Kaufmann

Hastie T., Tibshirani R., und Friedman J. (2009), *The Elements of statistical Learning: Data Mining, Inference, and Prediction*, 3rd. Ed., Springer-Verlag, Berlin

Joh C.H., D. Ettema und H. Timmermans (2009), *Estimating Marginal Mental Efforts of Activity Schedule. Adjustment Operators by Using Sequence Alignment*. Transportation Research Record: Journal of the Transportation Research Board, No. 2134, Transportation Research Board of the National Academies, Washington D.C., 2009, pp. 171-177

Joh C.H., T. Arentze und H. Timmermans (2007), *Identifying Skeletal Information of Activity Patterns by Multidimensional Sequence Alignment*. Transportation Research Record: Journal of the Transportation Research Board, No. 2021, Transportation Research Board of the National Academies, Washington D.C., 2007, pp. 81-88

Joh C.H. T. Arentze und H. Timmermans (2006), *Measuring an Predicting Adaptation Behavior in Multi-Dimensional Activity-Travel Patterns*. Transportmetrika, Vol. 2, 2006, pp. 153-173

Joh C.H., T. Arentze und H. Timmermans (2001), *Pattern Recognition in Complex Activity Travel Patterns, Comparison of Euclidean Distance, Signal-Processing Theoretical and Multidimensional Sequence Alignment Methods*. Transportation Research Record 1752, National Academy Press, Washington D.C., 2001, pp. 16-22

Keuleers B., G. Wets, T. Arentze und H. Timmermans (2001), *Association Rules in Identification of Spatial-Temporal Patterns in Multiday Activity Diary Data*, Transportation Research Record 1752, National Academy Press, Washington D.C., 2001, pp. 32-37

Ma J. und K.G. Goulias (1996), *A Dynamic Analysis of Activity and Travel Patterns Using Data from the Pudget Sound Transportation Panel*. Transportation, Vol. 24, No. 1, 1996, pp. 1-23

Meyer D., F. Leisch and K. Hornik (2003), *The support vector machine under test*. Neurocomputing 55(1-2): 169-186, 2003; http://dx.doi.org/10.1016/S0925-2312(03)00431-4

Mohammadian A. and E.J. Miler (2002), *Nested Logit Models and Artificial Neural Networks for Predicting Household Automobile Choices*. Transportation Research Record 1807, National Academy Press, Washington D.C., 2002, pp. 92-100

Nisbet R., Elder J. and Miner G. (2009), *Handbook of Statistical Analysis and Data Mining Applications*, Academic Press

Runkler T.A. (2009), *Data Mining: Methoden und Algorithmen intelligenter Datenanalyse*, Vieweg & Teubner.

Schlich R., (2004), *Verhaltenshomogene Gruppen in Längsschnitterhebungen*, Dissertation, Eidgenössische Technische Hochschule, Zürich.

Strambi O. und K.-A. van de Bilt (1998), *Trip Generation Modeling Using CHAID, a Criteria-Based Segmentation Modeling Tool*. Transportation Research Record 1645, National Academy Press, Washington D.C., 1998, pp. 24-31

Wets G., K. Vanhoof, T. Arentze and H. Timmermans (2000), *Identifying Decision Structures Underlying Activity Patterns, An Exploration of Data Mining Algorithms*, Transportation Research Record 1718, National Academy Press, Washington D.C., 2000, pp.1-9

Wilson C. (1998), *Analysis of Travel Behavior Using Sequence Alignment Methods*, Transportation Research Record 1645, National Academy Press, Washington D.C., 1998, pp. 52-59

Witten I.H. and Frank E. (2005), *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Ed, Morgan Kaufmann

Yamamoto T., R. Kitamura and J. Fujii (2002), *Driver's Route Choice Behavior, Analysis by Data Mining Algorithms*. Transportation Research Record 1807, National Academy Press, Washington D.C., 2002, pp. 59-65

Xie C., J. Lu und E. Parkany (2003), *Work Travel Mode Choice Modeling with Data Mining, Decision Trees and Neural Networks*. Transportation Research Record 1854, National Academy Press, Washington D.C., 2003, pp. 50-61

Ye N. (2003), *Handbook of Data Mining*, Lawrence Erlbaum Associates, Inc, Mahwah, New Jersey

Projektabschluss

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement für
Umwelt, Verkehr, Energie und Kommunikation UVEK
Bundesamt für Strassen ASTRA

FORSCHUNG IM STRASSENWESEN DES UVEK — ARAMIS SBT

Formular Nr. 3: Projektabschluss

erstellt / geändert am:	7. Juli 2011
Grunddaten	
Projekt-Nr.:	**SVI 2004/014**
Projekttitel:	Neue Erkenntnisse zum Mobilitätsverhalten dank Data Mining?
Enddatum:	7. Juli 2011
Texte:	
Zusammenfassung der Projektresultate:	Die Studie beschreibt den Prozess von Data Mining und gibt einen Überblick über die wichtigsten Data Mining Methoden. Die Beschreibung von Anwendungsbeispielen aus der Literatur illustriert Einsatzgebiete und Methoden von Data Mining. An eigenen Fallbeispielen werden mit den Daten des Mikrozensus Verkehr 2005 und Raumstrukturdaten das Vorgehen und der Einsatz verschiedener Methoden beim Data Mining demonstriert. Die Studie kommt zum Schluss, dass Data Mining in der Verkehrsplanung nutzbringend anwendbar ist, dass aber nicht – wie vielleicht erhofft – automatisch auf alle Fragen gute Antworten gefunden oder ohne Dazutun des Anwenders aus vorhandenen Datensätzen neue Erkenntnisse gewonnen werden können. Data Mining Methoden können zur Zeit die konventionellen Modellansätze nicht ersetzten.
Zielerreichung:	Das Ziel der Forschungsarbeit, den praktisch tätigen Verkehrsingenieur mit dem Prozess und den Methoden von Data Mining vertraut zu machen und die Möglichkeiten und Grenzen von Data Mining als Hilfsmittel in der Verkehrsplanung auszuloten, konnte erreicht werden.
Folgerungen und Empfehlungen:	Es wird aber empfohlen, in der Verkehrsplanung vermehrt konventionelle Modellansätze und Data Mining als sich ergänzende und sich gegenseitig unterstützende Methoden einzusetzen. Dem praktisch tätigen Verkehrsingenieur, welcher Data Mining in der Regel nicht regelmässig anwendet, wird zur Sicherstellung der Effizienz und der Zuverlässigkeit der Ergebnisse die Zusammenarbeit mit einem Data Mining Experten empfohlen.
Publikationen:	Keine

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement für
Umwelt, Verkehr, Energie und Kommunikation UVEK
Bundesamt für Strassen ASTRA

Beurteilung der Begleitkommission:
Diese Beurteilung der Begleitkommission ersetzt die bisherige separate fachliche Auswertung.

Beurteilung:	Die Forschungsarbeit vermittelt dem mit Data Mining noch nicht vertrauten Leser einen guten und verständlichen Überblick über den Prozess und die Methoden von Data Mining. Die Fallbeispiele demonstrieren anschaulich das Vorgehen und die Anwendung verschiedener Methoden. Gleichzeitig zeigen sie neben den Möglichkeiten auch die Grenzen von Data Mining auf. Die Empfehlung, in der Verkehrsplanung Data Mining und konventionelle Modellansätze als sich ergänzende und gegenseitig unterstützende Methoden einzusetzen, wird unterstützt. Mit der Forschungsarbeit liegt für den praktisch tätigen Verkehrsingenieur eine wertvolle Einführung in Data Mining vor, welche ihm zeigt, wie er dieses Datenanalyse-Instrument nutzbringend – in der Regel in Zusammenarbeit mit einem Data Mining Experten – einsetzen kann. Aus der Sicht der Begleitkommission wurde damit das Ziel der Forschungsarbeit erreicht.
Umsetzung:	Um Data Mining zukünftig auch in der Verkehrsplanung nutzbringend einsetzen zu können, bedarf es keiner weiteren Forschung. Vielmehr sind möglichst viele praktische Anwendungen erwünscht, mit denen Verkehrsplaner und Data Mining Experten in interdisziplinärer Zusammenarbeit Erfahrungen sammeln und weitergeben können.
weitergehender Forschungsbedarf:	Im Moment keiner
Einfluss auf Normenwerk:	Keiner

Präsident Begleitkommission:

Name:	Axhausen	Vorname:	Kay
Amt, Firma, Institut:	IVT, ETH Zürich		
Strasse, Nr.:	Wolfgang-Pauli-Strasse 15		
PLZ:	8093	Email:	axhausen@ivt.baug.ethz.ch
Ort:	Zürich	Telefon:	044 6333943
Kanton, Land:	Zürich	Fax:	044 6331057

Unterschrift Präsident Begleitkommission:

Gez.: K. W. Axhausen

Verzeichnis der Berichte der Forschung im Strassenwesen

Bericht-Nr.	Projekt Nr.	Titel	Datum
1323	VSS 2008/205	Ereignisdetektion im Strassentunnel *Détection d'incidents dans les tunnels routiers* *Incident Detection in Road Tunnels*	2011
1327	VSS 2006/601	Vorhersage von Frost und Nebel für Strassen *Prévision de gel et de brouillard pour les routes* *Prediction of frost and fog for roads*	2010
1328	VSS 2005/302	Grundlagen zur Quantifizierung der Auswirkungen von Sicherheitsdefiziten *Principes pour la quantification des effets des déficits de la sécurité* *Basis for the quantification of the effects of safety deficits*	2011
1329	SVI 2004/073	Alternativen zu Fussgängerstreifen in Tempo-30-Zonen *Alternatives aux passages pour piétons dans les zones 30* *Alternatives to zebra crossings in 30km/h zones*	2010
1330	FGU 2008/006	Energiegewinnung aus städtischen Tunneln; Systemeevaluation *Energy extraction from urban tunnels, evaluation of systems* *Extraction d'énergie géothermique de tunnels urbains;*	2010
1331	VSS 2005/501	Rückrechnung im Strassenbau *Analyse inverse pour la construction routière* *Inverse analysis in Road Geotechnics*	2011
1311	VSS 2000/543	Viabilite des projets et des Installations annexes *Kontrolle der Befahrbarkeit von Strassen und Nebenanlagen* *Viability of road projects and secondary facilities*	2010
1332	VSS 2006/905	Standardisierte Verkehrsdaten für das verkehrsträgerübergreifende Verkehrsmanagement *Standadisation des données de trafic pour gestion intermodale du trafic* *Standardised traffic data for intermodal trafic management*	2011
1333	SVI 2007/001	Standards für die Mobilitätsversorgung im peripheren Raum *Standards for mobility supply in peripheral regions* *Standards pour l'offre de mobilité dans l'espace périphérique*	2011
1334	ASTRA 2009/009	Was treibt uns an ? Antriebe und Treibstoffe für die Mobilität von Morgen *Transports de l'avenir ?* *Moteurs et carburants pour la mobilité de demain* *What drives us on ?* *Drives and fuels for the mobility of tomorrow*	2011
1335	VSS 2007/502	Stripping bei lärmmindernden Deckschichten unter Überrollbeanspruchung im labormasstab *Désenrobage des enrobés peu bruyants des couches de roulement sous sollicitation de roulement en laboratoire* *Stripping of Low Noise Surface Courses during Laboratory Scaled Wheel Tracking*	2011
1336	ASTRA 2007/006	SPIN-ALP: Scanning the Potential of Intermodal Transport on Alpine Corridors *SPIN-ALP: Abschätzung des Potentials des Intermodalen Verkehrs auf Alpenkorridoren* *SPIN-ALP: Estimation du potentiel du transport intermodal sur les axes transalpins*	2010
1339	SVI 2005/001	Widerstandsfunktionen für Innerorts- Strassenabschnitte ausserhalb des Einflussbereiches von Knoten *Fonctions de résistance pour des tronçons routiers urbains en dehors de la zone d'influence de carrefours* *Capacity restraint functions for urban road sections not affected by intersection delays*	2010

1325	SVI 2000/557	Indices caractéristiques d'une cité-Vélo. Méthode d'évaluation des politiques cyclables en 8 indices pour les petites et moyennes communes. *Die charakteristischen Indikatoren einer Velostadt. Evaluationsmethode der Velopolitiken anhand von 8 Indikatorgruppen für kleine und mittlere Gemeinden* *Characteristic indices of a Bike City. Method of evaluation of cycling policies in 8 indices for small and medium-sized communes*	2010
1337	ASTRA 2006/015	Development of urban network travel time estimation methodology *Temps de parcours en réseau urbain*	2011
1338	VSS 2006/902	Wirkungsmodelle für fahrzeugseitige Einrichtungen zur Steigerung der Verkehrssicherheit *Modèles d'impact d'équipements de véhicules pour améliorer la sécurité routière* *Modelling of the impact of in-vehicle equipment for the enhancement of traffic safety*	2009
1341	FGU 2007/005	Design aids for the planning of TBM drives in squeezing ground *Entscheidungsgrundlagen und Hilfsmittel für die Planung von TBM-Vortrieben in druckhaftem Gebirge* *Critères de décision et outils pour la planification de*	2011
1343	VSS 2009/903	Basistechnologien für die intermodale Nutzungserfassung im Personenverkehr *Basic technologies for detecting intermodal traveling passengers* *Les technologies de base pour l'enregistrement automatique*	2011
1340	SVI 2004/051	Aggressionen im Verkehr *L'aggressivité au volant* *Aggressive Driving*	2011
1344	VSS 2009/709	Initialprojekt für das Forschungspaket "Nutzensteigerung für die Anwender des SIS" *Projet initial pour le paquet de recherche "Augmentation de l'utilité pour les usagers du système d'information de la route"* *Initial project for the research package "Increasing benefits for the users of the road and transport information system"*	2011
1345	SVI 2004/039	Einsatzbereiche verschiedener Verkehrsmittel in Agglomerationen *Application areas of various means of transportation in agglomerations* *Domaine d'application de different moyen de transport dans*	2011
1342	FGU 2005/003	Untersuchungen zur Frostkörperbildung und Frosthebung beim Gefrierverfahren *Investigations of the ice-wall grow and frost heave in artificial ground freezing* *Recherches sur la formation corps gelés et du soulèvement au gel pendant la procédure de congélation*	2010
647	AGB 2004/010	Quality Control and Monitoring of electrically isolated post-tensioning tendons in bridges *Qualitätsprüfung und Überwachung elektrisch isolierter Spannglieder in Brücken* *Contrôle de la qualité et surveillance des câbles de précontrainte isolés électriquement dans les ponts*	2011
1348	VSS 2008/801	Sicherheit bei Parallelführung und Zusammentreffen von Strassen mit der Schiene *Sécurité en cas de tracés rail-route parallèles ou rapprochés* *Safety measures to manage risk of roads meeting or running close to railways*	2011
1349	VSS 2003/205	In-Situ-Abflussversuche zur Untersuchung der Entwässerung von Autobahnen *On-site runoff experiments on roads* *Essai d'écoulements pour l'évacuation des eaux des*	2011

1350	VSS 2007/904	IT-Security im Bereich Verkehrstelematik *IT-Security pour la télématique des transports* *IT-Security for Transport and Telematics*	2011
1352	VSS 2008/302	Fussgängerstreifen (Grundlagen) *Passage pour piétons (les bases)* *Pedestrian crossing (basics)*	2011

Publikationsliste SVI

Forschungsberichte auf Antrag der Vereinigung Schweizerischer Verkehrsingenieure (SVI)
Rapports de recherche sur proposition de l'Association suisse des ingénieurs en transports
(erschienen im Rahmen der Forschungsreihe des UVEK / parus dans le cadre des recherches du DETEC)

1980 **Velo- und Mofaverkehr in den Städten**
(R. Müller)
1980 **Anleitung zur Projektierung einer Lichtsignalanlage**
(Seiler Niederhauser Zuberbühler)
1981 **Güternahverkehr, Gesetzmässigkeiten**
(E. Stadtmann)
1981 **Optimale Haltestellenabstände beim öffentlichen Verkehr**
(Prof. H. Brändli)
1982 **Entwicklung des schweizerischen Strassenverkehrs ***
(SNZ Ingenieurbüro AG)
1983 **Lichtsignalanlagen mit oder ohne Uebergangssignal Rot-Gelb**
(Weber Angehrn Meyer)
1983 **Güternahverkehr, Verteilungsmodelle**
(Emch + Berger AG
1983 **Modèle Transyt 8: Traffic Network Study Tool; Programme Pretrans**
(...)
1983 **Parkraumbewirtschaftung als Mittel der Verkehrslenkung ***
(Glaser + Saxer)
1984 **Le rôle des taxis dans les transports urbains (franz. Ausgabe)**
(Transitec)
1984 **Park and Ride in Schweizer Städten ***
(Balzari & Schudel AG)
1986 **Verträglichkeit von Fahrrad, Mofa und Fussgänger auf gemeinsamen Verkehrsflächen ***
(Weber Angehrn Meyer)
1986 **Transyt 8 / Pretrans; Modell Programmsystem für die Optimierung von Signalplänen von städtischen Strassennetzen**
(...)
1987 **Verminderung der Umweltbelastungen durch verkehrsorganisatorische und –technische Massnahmen ***
(Metron AG)
1987 **Provisorischer Behelf für die Umweltverträglichkeits-Prüfung von Verkehrsanlagen ***
(Büro BC, Jenni + Gottardi AG, Scherrer)
1988 **Bestimmungsgrössen der Verkehrsmittelwahl im Güterverkehr ***
(Rapp AG)
1988 **EDV-Anwendungen im Verkehrswesen**
(IVT, ETH Zürich)
1988 **Forschungsvorschläge Umweltverträglichkeitsprüfung von Verkehrsanlagen**
(Büro BC, Jenni & Gottardi AG, Scherrer)
1989 **Vereinfachte Methode zur raschen Schätzung von Verkehrsbeziehungen ***
(P. Widmer)
1990 **Planungsverfahren bei Ortsumfahrungen**
(Toscano-Bernardi-Frey AG)
1990 **Anteil der Fahrzeugkategorien in Abhängigkeit vom Strassentyp**
(Abay & Meyer)
1991 **Busbuchten, ja oder nein?***
(Zwicker und Schmid)
1991 **EDV-Anwendung im Verkehrswesen, Katalog 1990**
(IVT, ETH Zürich)
1991 **Mofa zwischen Velo und Auto**
(Weber Angehrn Meyer)
1991 **Erhebung zum Güterverkehr**
(Abay & Meier, Albrecht & Partner AG, Holinger AG, RAPP AG, Sigmaplan AG)
1991 **Mögliche Methoden zur Erstellung einer Gesamtbewertung bei Prüfverfahren***
(Basler & Partner AG)
1992 **Parkierungsbeschränkungen mit Blauer Zone und Anwohnerparkkarte**
(Jud AG)
1992 **Einsatzkonzepte und Integrationsprobleme der Elektromobile***
(U. Schwegler)

1992 **UVP bei Strassenverkehrsanlagen, Anleitung zur Erstellung von UVP-Berichten***
(Büro BC, Jenni & Gottardi AG, Scherrer)
erschienen auch als Mitteilungen zur UVP Nr. 7/Mai 1992 des BUWAL

1992 **Von Experten zu Beteiligten - Partizipation von Interessierten und Betroffenen beim Entscheiden über Verkehrsvorhaben***
(J. Dietiker)

1992 **Fehlerrechnung und Sensitivitätsanalyse für Fragen der Luftreinhaltung: Verkehr - Emissionen – Immissionen ***
(INFRAS)

1993 **Indikatoren im Fussgängerverkehr ***
*(RAPP AG)*1993

1993 **Velofahren in Fussgängerzonen***
(P. Ott)

1993 **Vernetztes bzw. ganzheitliches Denken bei Verkehrsvorhaben**
(Jauslin + Stebler, Rudolf Keller AG)

1993 **Untersuchung des Zusammenhanges von Verkehrs- und Wanderungsmobilität**
(synergo, Jenni + Gottardi AG)

1993 **Einsatzmöglichkeiten und Grenzen von flexiblen Nutzungen im Strassenraum**
(Sigmaplan AG)

1993 **EIE et infrastructures routières, Guide pour l'établissement de rapports d'impact ***
(Büro BC, Jenni + Gottardi AG, Scherrer)
erschienen als Mitteilungen zur UVP Nr. 7(93) / Juli 1993 des BUWAL/parus comme informations concernant l'étude de l'impact sur l'environnement EIE No. 7(93) / juillet 1993 de l'OFEFP

1993 **Handlungsanleitung für die Zweckmässigkeitsprüfung von Verkehrsinfrastrukturprojekten, Vorstudie**
(Jenni + Gottardi AG)

1994 **Leistungsfähigkeit beim Fahrstreifenabbau auf Hochleistungsstrassen**
(Rutishauser, Mögerle, Keller)

1994 **Perspektiven des Freizeitverkehrs, Teil 1: Determinanten und Entwicklungen***
(R + R Burger AG, Büro Z)

1995 **Verkehrsentwicklungen in Europa, Vergleich mit den schweizerischen Verkehrsperspektiven**
(Prognos AG / Rudolf Keller AG)
erschienen als GVF-Auftrag Nr. 267 des GS EVED Dienst für Gesamtverkehrsfragen / paru au SG DFTCE Service d'étude des transports No. 267

1996 **Einfluss von Strassenkapazitätsänderungen auf das Verkehrsgeschehen**
(SNZ Ingenieurbüro AG)

1997 **Zweckmässigkeitsbeurteilung von Strassenverkehrsanlagen ***
(Jenni + Gottardi AG)

1997 **Verkehrsgrundlagen für Umwelt- und Verkehrsuntersuchungen**
(Ernst Basler + Partner AG)

1998 **Entwicklungsindices des Schweizerischen Strassenverkehrs ***
(Abay + Meier)

1998 **Kennzahlen des Strassengüterverkehrs in Anlehnung an die Gütertransportstatistik 1993**
(Albrecht & Partner AG / Symplan Map AG)

1998 **Was Menschen bewegt. Motive und Fahrzwecke der Verkehrsteilnahme**
(J. Dietiker)

1998 **Das spezifische Verkehrspotential bei beschränktem Parkplatzangebot ***
(SNZ Ingenieurbüro AG)

1998 **La banque de données routières STRADA-DB somme base de modèles de trafic**
(Robert-Grandpierre et Rapp SA / INSER SA / Rosenthaler & Partner AG)

1998 **Perspektiven des Freizeitverkehrs. Teil 2: Strategien zur Problemlösung**
(R + R Burger und Partner, Büro Z)

1998 **Kombinierte Unter- und Überführung für FussgängerInnen und VelofahrerInnen**
(Büro BC / Pestalozzi & Stäheli)

1998 **Kostenwirksamkeit von Umweltschutzmassnahmen**
(INFRAS)

1998 **Abgrenzung zwischen Personen- und Güterverkehr**
(Prognos AG)

1999 **Gesetzmässigkeiten im Strassengüterverkehr und seine modellmässige Behandlung**
(Abay & Meier / Ernst Basler + Partner AG)

1999 **Aktualisierung der Modal Split-Ansätze**
(P. Widmer)

1999 **Management du trafic dans les grands ensembles**
(Transportplan SA)

1999 **Technology Assessment im Verkehrswesen : Vorstudie**
(RAPP AG Ing. + Planer Zürich)

1999 **Verkehrstelematik im Management des Verkehrs in Tourismusgebieten**
(ASIT / IC Infraconsult AG)

1999 **„Kernfahrbahnen“ Optimierte Führung des Veloverkehrs an engen Strassenquerschnitten ***
(Metron Verkehrsplanung und Ingenieurbüro AG)

2000 **Sensitivitäten von Angebots- und Preisänderungen im Personenverkehr**
(Prognos AG)

2000 **Dephi-Umfrage Zukunft des Verkehrs in der Schweiz**
(P. Widmer / IPSO Sozial-, Marketing- und Personalforschung)

2000 **Der Wert der Zeit im Güterverkehr**
(Jenni + Gottardi AG)

2000 **Floating Car Data in der Verkehrsplanung**
(Rudolf Keller & Partner Verkehrsingenieure AG + Rosenthaler + Partner AG)

2000 **Verlässlichkeit als Entscheidvariable: Experimente mit verschiedenen Befragungssätzen**
(IVT - ETHZ)

2001 **Aktivitätenorientierte Personenverkehrsmodelle, Vorstudie**
(P. Widmer und K.W. Axhausen)

2001 **Zeitkostenansätze im Personenverkehr**
(G. Abay und K.W. Axhausen)

2001 **Véhicules électriques et nouvelles formes de mobilité**
(Transitec Ingénieurs-Conseils SA)

2001 **Besetzungsgrad von Personenwagen:** Analyse von Bestimmungsgrössen und Beurteilung von Massnahmen zu dessen Erhöhung
(RAPP AG Ingenieure + Planer)

2001 **Grobkonzept zum Aufbau einer multimodalen Verkehrsdatenbank**
(INFRAS)

2001 **Ermittlung der Gesamtleistungsfähigkeit (MIV + OEV) bei lichtsignalgeregelten Knoten**
(büro S-ce Simon-consulting-engineering)

2001 **Besteuerung von Autos mit einem Bonus/Malus-System im Kanton Tessin**
(U. Schwegler Büro für Verkehrsplanung)

2001 **GIS als Hilfsmittel in der Verkehrsplanung**
(büro widmer)

2001 **Umgestaltung von Strassen im Zuge von Erneuerungen**
(Infraconsult AG + Zeltner + Maurer AG)

2001 **Piloterhebung zum Dienstleistungsverkehr und zum Gütertransport mit Personenwagen**
(Prognos AG, Emch+Berger AG, IVU Traffic Technologies AG)

2002 **Parkplatzbewirtschaftung bei publikumsintensiven Einrichtungen - Auswirkungsanalyse**
(Metron AG, Neosys AG, Hochschule Rapperswil)

2002 **Probleme bei der Einführung und Durchsetzung der im Transportwesen geltenden Umweltschutzbestimmungen; unter besonderer Berücksichtigung des Vollzugs beim Strassenverkehrslärm**
(B+S Ingenieur AG)

2002 **Nachhaltigkeit und Koexistenz in der Strassenraumplanung**
(Berz Hafner + Partner AG)

2002 **Warum steht P. Müller lieber im Stau als im Tram?**
(Planungsbüro Jürg Dietiker / MOVE RAUM P. Regli / Landert Farago Davatz & Partner / Dr. A. Zeyer)

2002 **Nachhaltigkeit im Verkehr**
(Jenni + Gottardi AG)

2002 **Massnahmen zur Erhöhung der Akzeptanz längerer Fuss- und Velostrecken**
(Arbeitsgemeinschaft Büro für Mobilität / V. Häberli / A. Blumenstein / M. Wälti)

2002 **Carreiseverkehr: Grundlagen und Perspektiven**
(B+S Ingenieur AG / Gare Routière de Genève))

2002 **Potentielle Gefahrenstellen**
(Basler & Hofmann / Psychologisches Institut der Universität Zürich)

2003 **Evaluation kurzfristiger Benzinpreiserhöhungen**
(Infras / M. Peter / N. Schmidt / M. Maibach)

2002 **Verlässlichkeit als Entscheidungsvariable, Vorstudie**
(ETH Zürich, Institut für Verkehrsplanung und Transportsysteme IVT)

2002 **Mischverkehr MIV / ÖV auf stark befahrenen Strassen**
(Verkehrsingenieurbüro TEAMverkehr)

2003 **Vorstudie zu den Wechselwirkungen Individualverkehr – öffentlicher Verkehr infolge von Verkehrstelematik-Systemen**
(Abay & Meier, Zürich)

2003 **Strassen mit Gemischtverkehr: Anforderungen aus der Sicht der Zweiradfahrer**
(WAM Partner, Planer und Ingenieure, Solothurn)

2003 **Erfolgskontrolle von Umweltschutzmassnahmen bei Verkehrsvorhaben**
(Metron Landschaft AG, Brugg / Quadra GmbH, Zürich / Metron Verkehrsplanung AG, Brugg)

2004 **Perspektiven für kurze Autos**
(Ingenieur- und Planungsbüro Bühlmann, Zollikon)
2004 **Lange Planungsprozesse im Verkehr**
(BINARIO TRE, Windisch)
2004 **Auswirkungen von Personal Travel Assistance (PTA) auf das Verkehrsverhalten**
(Ernst Basler und Partner AG, Zürich)
2004 **Methoden zum Erstellen und Aktualisieren von Wunschlinienmatrizen im motorisierten Individualverkehr**
(ETH Zürich, Institut für Verkehrsplanung und Transportsysteme IVT)
2004 **Zeitkostenansätze im Personenverkehr**
(ETH Zürich, Institut für Verkehrsplanung und Transportsysteme IVT / Rapp Trans AG, Zürich)
2004 **Determinaten des Freizeitverkehrs: Modellierung und empirische Befunde**
(ETH Zürich, Institut für Verkehrsplanung und Transportsysteme IVT)
2004 **Verfahren von Technology Assessment im Verkehrswesen**
(Rapp Trans AG, Zürich / IKAÖ, Bern / Interface, Luzern)
2004 **Mobilitätsdatenmanagement für lokale Bedürfnisse**
(SNZ, Zürich / TEAMverkehr, Cham / Büro für Verkehrsplanung, Fischingen)
2004 **Auswirkungen neuer Arbeitsformen auf den Verkehr - Vorstudie**
(INFRAS, Bern)
2004 **Standards für intermodale Schnittstellen im Verkehr**
(synergo, Zürich / ILS NRW, Dortmund)
2005 **Verkehrsumlegungs-Modelle für stark belastete Strassennetze**
(büro widmer, Frauenfeld)
2005 **Wirksamkeit und Nutzen der Verkehrsinformation**
(B+S Ingenieure AG, Bern / Ernst Basler + Partner AG, Zürich / Landert Farago Partner, Zürich)
2005 **Spezialisierung und Vernetzung: Verkehrsangebot und Nachfrageentwicklung zwischen den Metropolitanräumen des Städtesystems Schweiz**
(synergo, Zürich)
2005 **Wirkungsketten Verkehr - Wirtschaft**
(ECOPLAN, Altdorf und Bern / büro widmer, Frauenfeld)
2005 **Cleaner Drive**
Hindernisse für die Markteinführung von neuen Fahrzeug-Generationen
(E'mobile, der Schweizerische Verband für elektrische und effiziente Strassenfahrzeuge, Urs Schwegler)
2005 **Spezifische Anforderungen an Autobahnen in städtischen Agglomerationen**
(Ingenieur- und Planungsbüro Dr. Walter Berg, Zürich)
2005 **Instrumente für die Planung und Evaluation von Verkehrssystem-Management-Massnahmen**
(Jenni + Gottardi AG, Zürich / Universität Karlsruhe)
2005 **Trafic de support logistique de grandes manifestations (Betriebsverkehr von Grossanlässen)**
(Ecole Polytechnique Fédérale de Lausanne, EPFL)
2005 **Verkehrsdosierungsanlagen, Strategien und Dimensionierungsgrundsätze**
(Ingenieurbüro Walter Berg, Zürich)
2005 **Angebote und Erfolgskriterien im nächtlichen Freizeitverkehr**
(Planungsbüro Jud, Zürich)
2005 **Vor- und Nachlauf im kombinierten Ladungsverkehr**
(Rapp Trans AG, Zürich)
2005 **Finanzielle Anreize für effiziente Fahrzeuge - Eine Wirkungsanalyse der Projekte VEL2 (Tessin) und NewRide in Basel und Zürich**
(Rapp Trans AG, Zürich / Interface, Luzern)
2006 **Reduktionsmöglichkeiten externer Kosten des MIV am Beispiel des Förderprogramms VEL2 im Kanton Tessin**
(Università della Svizzera Italiana, Lugano / Eidgenössische Technische Hochschule, Zürich)
2006 **Nachhaltigkeit im Verkehr**
Indikatoren im Bereich Gesellschaft
(Ernst Basler + Partner AG, Zollikon / Landert Farago Partner, Zürich)
2006 **Früherkennung von Entwicklungstrends zum Verkehrsangebot**
(Interface - Institut für Politikstudien, Luzern)
2006 **Publikumsintensive Einrichtungen PE: Planungsgrundlagen und Gesetzmässigkeiten**
(Metron Verkehrsplanung AG, Brugg / Transitec Ingenieurs-Conseils SA, Lausanne / Fussverkehr Schweiz, Zürich)
2006 **Erhebung des Fuss- und Veloverkehrs**
(IRAP, Hochschule für Technik, Rapperswil / Fussverkehr Schweiz, Zürich / Pestalozzi & Stäheli, Basel / Daniel Sauter, Urban Mobility Research, Zürich)
2006 **Verkehrstechnische Beurteilung multimodaler Betriebskonzepte auf Strassen innerorts**
(S-ce Simon consulting experts, Zürich)
2006 **Beurteilung von Busbevorzugungsmassnahmen**
(Metron Verkehrsplanung AG, Brugg)

2006 **Error Propagation in Macro Transport Models**
(Systems Consult, Monaco / B+S Ingenieur AG, Bern)
2007 **Fussgängerstreifenlose Ortszentren**
(Ingenieurbüro Ghielmetti, Winterthur / IAP, Zürich)
2007 **Kernfahrbahnen auf Ausserortsstrecken**
(Frossard GmbH, Zürich)
2007 **Road Pricing Modelle auf Autobahnen und in Stadtregionen**
(INFRAS, Zürich / Rapp Trans AG, Basel)
2007 **Entkopplung zwischen Verkehrs- und Wirtschaftswachstum**
(INFRAS, Zürich / Università della Svizzera Italiana, Lugano)
2007 **Genderfragen in der Verkehrsplanung Vorstudie**
(SNZ Ingenieure und Planer AG, Zürich)
2007 **Konfliktanalyse beim Mischverkehr**
(Sigmaplan AG, Bern)
2007 **Verfahren zur Berücksichtigung der Zuverlässigkeit in Evaluationen**
(Ernst Basler + Partner AG, Zürich / Eidgenössische Technische Hochschule, Zürich)
2007 **Überlegungen zu einem Marketingansatz im Fuss- und Veloverkehr**
(Büro für Mobilität AG, Bern/Burgdorf / büro für utopien, Burgdorf/Berlin / LP Ingenieure AG, Bern / Masciardi communication & design AG, Bern)
2008 **Einbezug von Reisekosten bei der Modellierung des Mobilitätsverhaltens**
(Institut für Verkehrsplanung und Transportsysteme (IVT) ETH, Zürich / TRANSP-OR EPF Lausanne, Lausanne / IRE USI, Lugano)
2008 **Ausgestaltung von multimodalen Umsteigepunkten**
(Metron AG, Brugg / Universität Zürich Sozialforschungsstelle, Zürich)
2008 **Überbreite Fahrstreifen und zweistreifige Schmalfahrbahnen**
(IRAP HSR Hochschule für Technik, Rapperswil)
2008 **Fahrten- und Fahrleistungsmodelle: Erste Erfahrungen**
(Hesse+Schwarze+Partner, Zürich / büro widmer, Frauenfeld)
2008 **Quantitative Auswirkungen von Mobility Pricing Szenarien auf das Mobilitätsverhalten und auf die Raumplanung**
(Verkehrsconsulting Fröhlich, Zürich / TransOptima GmbH, Olten / Ernst Basler + Partner AG, Zürich)
2008 **Organisatorische und rechtliche Aspekte des Mobility Pricing**
(Ernst Basler + Partner AG)
2008 **Forschungspaket "Güterverkehr", Initialprojekt "Bestandesaufnahme und Konkretisierung des Forschungspakets"**
(Eidgenössische Technische Hochschule, Zürich - ETH / Università della Svizzera Italiana / Universität St. Gallen)
2008 **Freizeitverkehr innerhalb von Agglomerationen**
(Hochschule Luzern - Wirtschaft, Luzern / ISOE, Frankfurt am Main / Interface Politikstudien, Luzern)
2008 **Gesetzmässigkeiten des Anlieferverkehrs**
(Sigmaplan AG / Rudolf Keller & Partner Verkehrsingenieure AG)
2009 **Modal Split Funktionen im Güterverkehr**
(Rapp Trans AG, Zürich / IVT ETH, Zürich)
2009 **Mobilitätsmuster zukünftiger Rentnerinnen und Rentner: eine Herausforderung für das Verkehrssystem 2030?**
(büro widmer Frauenfeld / Institut für Psychologie, Universität Bern)
2008 **Mobilitätsmanagement in Berieben - Motive und Wirksamkeit**
(synergo, Zürich / Tensor Consulting AG, Bern)
2009 **Monitoring und Controlling des Gesamtverkehrs in Agglomerationen**
(Ecoplan, Altdorf und Bern / Ernst Basler + Partner, Zürich)
2009 **Wie Strassenraumbilder den Verkehr beeinflussen**
(Zürcher Hochschule für angewandte Wissenschaften zhaw, Winterthur / Jenni + Gottardi AG, Thalwil)
2009 **Nettoverkehr von verkehrsintensiven Einrichtungen (VE)**
(Berz Hafner + Partner AG, Bern / Hornung Wirtschafts- und Sozialstudien, Bern / Künzler Bossert + Partner GmbH, Bern / Roduner BSB + Partner AG, Schliern)
2009 **Verkehrspolitische Entscheidfindung in der Verkehrsplanung**
(synergo, Mobilität - Politik - Raum, Zürich / Institut für Politikwissenschaft/Uni Bern, Bern / Büro Vatter, Bern / Büro für Mobilität AG, Bern)
2009 **Einsatz von Simulationswerkzeugen in der Güterverkehrs- und Transportplanung**
(Rapp Trans AG, Zürich / ZHAW, Wädenswil, IAS Institut für Angewandte Simulation)
2009 **Multimodale Verkehrsqualitätsstufen für den Strassenverkehr - Vorstudie**
(Institut für Verkehrsplanung und Transportsysteme (IVT), ETH Zürich)
2010 **Optimierung der Stassenverkehrsunfallstatistik durch Berücksichtigung von Daten aus dem Gesundheitswesen**
(Rapp Trans AG, Zürich)

2010 **Systematische Wirkungsanalysen von kleinen und mittleren Verkehrsvorhaben**
(B,S,S. Volkswirtschaftliche Beratung AG, Basel / Basler & Hofmann AG, Zürich)

2011 **Zeitwerte im Personenverkehr: Wahrnehmungs- und Distanzabhängigkeit**
(Institut für Verkehrsplanung und Transportsysteme (IVT), ETH Zürich)

2011 **Hindernisfreier Verkehrsraum - Anforderungen aus Sicht von Menschen mit Behinderung**
(Pestalozzi & Stäheli, Basel / Schweiz. Fachstelle für behindertengerechtes Bauen, Zürich)

2011 **Der Verkehr aus Sicht der Kinder: Schulwege von Primarschulkindern in der Schweiz**
(Interfakultäre Koordinationsstelle für Allgemeine Ökologie (IKAÖ), Bern / Interface Politikstudien Forschung und Beratung, Luzern / verkehrsteiner, Bern)

2011 **Alternativen zu Fussgängerstreifen in Tempo-30-Zonen**
(Ingenieurbüro Ghielmetti, Chur / Pestalozzi & Stäheli, Basel / verkehrsteiner, Bern)

2011 **Standards für die Mobilitätsversorgung im peripheren Raum**
(Ecoplan, Bern / Metron, Brugg)

2011 **Widerstandsfunktionen für Innerorts-Strassenabschnitte ausserhalb des Einflussbereiches von Knoten**
(büro widmer ag, Frauenfeld / Rudolf Keller & Partner AG, Muttenz)

** vergriffen: Diese Exemplare können auf Wunsch nachkopiert werden*
**épuisé: Selon désir, ces rapports peuvent être copiés*

Die Berichte können bezogen werden bei / Les rapports peuvent être commandés au:
VSS, Sihlquai 255, 8005 Zürich,
Tel. 044 / 269 40 20, Fax. 044 / 252 31 30, info@vss.ch