Prognose von Verspätungen im öffentlichen Verkehr auf Basis empirischer Daten

Masterarbeit

Studiengang:	MAS Data Science
Autor:	Andreas Gutweniger
Betreuer:	Arno Schmidhauser
Datum:	12. März 2018

Management Summary

In der Schweiz ist die Pünktlichkeit im öffentlichen Verkehr traditionell hoch; auch hierzulande kommt es jedoch zu Verspätungen. Für Kunden und Mitarbeiter ist es hilfreich, darüber im Voraus informiert zu sein. IT-Systeme zur Verspätungsprognose sind daher bei allen grösseren Verkehrsunternehmen im Einsatz. Grundlage der Prognoserechnung bilden dabei theoretische Modelle über den zukünftigen Fahrtverlauf. Dass Verspätungsprognosen aufgrund von Betriebsdaten der Vergangenheit berechnet werden, ist dagegen nicht bekannt. Dabei wären die empirischen Grundlagen durchaus vorhanden: Heutige Leittechniksysteme generieren umfangreiche Protokolldaten. Teile davon sind in der Schweiz sogar als «open data» verfügbar.

In meiner Arbeit gehe ich der Frage nach, ob sich unter Verwendung von historischen Daten zum Zugsverkehr Verspätungsprognosen erzielen lassen, die jene der Bahn-Unternehmen mindestens in Teilbereichen übertreffen. Dabei ist ein starker Prognose-Bias einzuhalten: Das Überschätzen von Verspätungen ist unerwünscht und soll nicht häufiger auftreten, als dies bei den Systemen der Bahnen der Fall ist.

Die Arbeit führt zunächst ausführlich in Funktionsweise und Zusammenhänge des Bahnverkehrs ein und gibt einen Überblick über den Stand der relevanten Forschung.

Den Kern der Untersuchung bildet die Entwicklung eines Prognoseverfahrens, das mit Vergangenheitsdaten trainiert und dann auf beliebige Betriebslage-Konstellationen angewendet werden kann. Anforderungen an ein solches Verfahren werden diskutiert und verschiedene Machine Learning-Verfahren und Modelltypen evaluiert. Der von mir ausgewählte Ansatz basiert auf Entscheidungsbäumen, die mit Recursive Partitioning erzeugt werden.

Eine besondere Schwierigkeit resultiert aus dem Umstand, dass sich Verfügbarkeit und Stellenwert möglicher Prädiktor-Variablen im Zeitablauf ständig verändern. Um dem zu begegnen, wird eine grosse Zahl von Entscheidungsbäumen generiert, aus denen in der konkreten Vorhersagesituation die anwendbaren ausgewählt werden.

Das Verfahren wurde auf der bestehenden Systemarchitektur von puenktlichkeit.ch implementiert. Es kann auf historischen Daten simuliert oder in Echtzeit angewendet werden. Als Untersuchungsszenario für diese Arbeit dienten die Ankünfte aller taktmässig verkehrenden Normalspur-Züge am Bahnhof Bern. Simulationsrechnungen in diesem Szenario zeugen von der Validität des Verfahrens.

Seit Januar 2018 wird es auch im Echtzeit-Betrieb angewendet und liefert im Minutenrhythmus Verspätungsprognosen für die in Bern ankommenden Züge. Web-Applikationen erlauben eine Visualisierung der Prognosen, ihres zeitlichen Verlaufs und messen die Prognosequalität. Alle Ergebnisse sind unter www.puenktlichkeit.ch öffentlich einsehbar. Aktuelle Prognosen können auch auf dem Smartphone abgerufen werden.

Die Auswertung dieses Live-Experiments zeigt: In einem Zeithorizont bis etwa 5 Minuten vor Ankunft eines Zugs schneidet das Verfahren deutlich besser ab als die Prognosesysteme der Bahnen. Oft können Verspätungen bereits 25 Minuten im Voraus vorhergesagt werden. Lediglich bei den sehr kurzfristigen Prognosen sind die Systeme der Unternehmen überlegen – weil sie dabei Daten verwenden, die nicht öffentlich verfügbar sind.

Das Verfahren ist aber nicht nur für Reisende interessant, die sich um ihre pünktliche Ankunft sorgen. Die erzeugten Entscheidungsbäume können einfach visualisiert werden und geben Aufschluss über die Zusammenhänge im komplexen System der Bahnproduktion. Solche Erkenntnisse sind wertvoll in vielen Prozessen der Bahn-Branche: bei der Weiterentwicklung des Gleisnetzes, der gezielten Beseitigung von Engpässen und Verspätungsursachen, der Konstruktion noch robusterer Fahrpläne, der richtigen Reaktion beim Auftreten von Störungen und der Wahl einer energieeffizienten Fahrweise.

Inhaltsverzeichnis

1 Einleitung

«Werden wir pünktlich ankommen?». Dies gehört vermutlich zu den häufigsten Fragen, die einem Zugbegleiter gestellt werden. «Normalerweise sind wir pünktlich.» könnte die Antwort lauten. Der Zugbegleiter würde sich dabei auf sein Erfahrungswissen stützen: normalerweise sind Züge pünktlich (zumindest in der Schweiz).

Doch der Fahrgast hätte wohl kaum gefragt, wenn alles «normal» wäre. Vermutlich hat er festgestellt, dass sein Zug bereits verspätet unterwegs ist. Oder er hat von einer Störung erfahren, die sich auch auf seine Reise auswirken könnte. Lässt sich Erfahrungswissen auch in solchen Situationen nutzen, um eine Prognose abzugeben? Kann aus «ähnlichen» Situationen der Vergangenheit auf eine zukünftige Verspätung geschlossen werden?

Solche Fragen stehen im Zentrum der vorliegenden Arbeit. Und sie fordern das bestehende Paradigma heraus. Denn traditionelle Methoden der Verspätungsprognose im öffentlichen Verkehr sind stark theoriegeleitet und machen keinen direkten Gebrauch von empirischen Beobachtungen. Dies, obwohl die Verfügbarkeit von Daten über aktuelle und vergangene Betriebsabläufe stark zugenommen hat, sogar in Form von «open data». Schlägt Korrelation die Theorie? Mit meiner Arbeit möchte ich nicht zuletzt einen Beitrag zu dieser Diskussion liefern.[1]

1.1 Ausgangslage und Motivation

Pünktlichkeit gehört zu den wichtigsten Erfolgsfaktoren des öffentlichen Verkehrs. Sie ist in der Schweiz traditionell hoch; auch hierzulande kommt es jedoch zu Verspätungen. Darüber im Voraus informiert zu sein, ist für verschiedene Anspruchsgruppen von Wert:

a) Fahrdienstleiter und Disponenten können auf eine sich abzeichnende Verspätung reagieren, indem sie z.B.
   - Entscheide darüber treffen, ob Anschlussverkehre warten sollen oder nicht,
   - Fahrwege, Gleisbelegungen oder Kreuzungspunkte ändern, um Konflikte mit anderen Verkehren zu reduzieren,
   - bei sehr grossen Verspätungen Ersatzangebote aufbieten.

b) Kunden schätzen es ebenfalls, Verspätungen vorab zu kennen, weil es ihnen ggf. ermöglicht
   - andere Personen über ihre verspätete Ankunft zu informieren.
   - die Wartezeit sinnvoll zu nutzen,
   - auf andere Verbindungen auszuweichen.

Die Prognose von Verspätungen ist daher wichtiger Bestandteil von sowohl Dispositions- als auch Kundeninformationssystemen bei allen grösseren Verkehrsunternehmen. Ausgehend von der aktuellen Kenntnis der Betriebslage wird dabei in der Regel der früheste Zeitpunkt ermittelt, nach dem das betrachtete Verkehrsmittel theoretisch an den folgenden Stationen eintreffen kann. Häufig erfolgt dies, indem die verbleibende Fahrzeit unter Optimal-Bedingungen betrachtet wird. Komplexere Prognosemodelle berücksichtigen zusätzlich auch den Einfluss von anderen Verkehrsmitteln (z.B. Gleisbelegungen, Anschlussbeziehungen, Umlauf von Fahrzeugen und Personal). Fast immer basieren die Prognosen aber auf theoretisch hergeleiteten Modellen und Parametern und nicht auf empirischen Erhebungen über Betriebsabläufe der Vergangenheit. So kann es vorkommen, dass eine Prognose auf

[1] Chris Anderson, damaliger Chefreadakteur von «Wired», hat mit Blick auf den Siegeszug von Big Data und Machine Learning vor 10 Jahren das «Ende der Theorie» proklamiert - was noch immer kontrovers diskutiert wird. Vgl. Anderson (2008).

Basis von Annahmen getroffen wird, die – wenn auch vielleicht nicht grundsätzlich unrealistisch – unter vergleichbaren Bedingungen nie oder fast nie eingetreten sind.

Die empirischen Grundlagen wären dabei durchaus vorhanden: Sowohl im Strassenverkehr (Bus, Tram) als auch bei der Eisenbahn generieren die Leittechniksysteme umfangreiche Protokolldaten über den Betriebsablauf. Für die Schweiz ist ein beträchtlicher Teil davon seit Dezember 2016 über das Portal opentransportdata.swiss öffentlich zugreifbar.

1.2 Zielsetzung und Abgrenzung

Ziel der vorliegenden Masterarbeit ist, folgende Forschungsfrage zu beantworten:

> **Können unter Verwendung der empirischen Vergangenheitswerte von opentransportdata.ch mindestens in Teilbereichen bessere Verspätungsprognosen erzielt werden, als sie von den aktuell im öV Schweiz eingesetzten Kundeninformationssystemen geliefert werden?**

Die Prognosequalität kann dabei in mehreren Dimensionen betrachtet werden:

1. Wie genau kann das Ausmass (Anzahl Minuten) einer Verspätung vorhergesagt werden?
2. Wie häufig kann eine Verspätung im Sinne der Überschreitung eines gegebenen Grenzwerts (z.B. mehr als 5 Minuten) korrekt vorhergesagt werden?
3. Mit welchem zeitlichen Vorlauf können diese Vorhersagen erfolgen?

Dabei bestehen Wechselwirkungen zwischen Dimension 1 und 2 einerseits und Dimension 3 andererseits: Je kürzer der Vorhersagehorizont, desto genauere Prognosen werden möglich sein.

Als wichtige Rahmenbedingung gilt in jedem Fall, dass eine «Überschätzung» von Verspätungen vermieden werden muss: In der Praxis wäre es fatal, wenn Fahrgäste ihren Zug verpassen, weil sie sich auf eine Verspätungsprognose verlassen, die gar nicht eintritt. Aus diesem Grund sind existierende Prognosesysteme defensiv ausgelegt und liefern nur eine sehr geringe Rate von «false positives». Um Vergleichbarkeit zu gewährleisten, soll sich der alpha-Fehler der empirisch hergeleiteten Prognosen auf ähnlichem Niveau bewegen wie bei den «Theorie-geleiteten» Vergleichssystemen.

Die Prognose soll ausschliesslich auf Grundlage von Betriebsdaten durchgeführt werden, die über opentransportdata.ch bezogen werden. Die Verwendung weiterer Daten, die als Prädiktoren sinnvoll sein könnten (z.B. Wetter, Fahrgast-Zahlen, Staumeldungen), ist nicht vorgesehen.

Es ist nicht Ziel der Arbeit, die Gründe für das Auftreten von Verspätungen zu analysieren: erstellt werden sollen Prognosemodelle, nicht Erklärungsmodelle. Im Sinne eines Ausblicks soll jedoch die Nutzung der Modelle auch für die Ursachenanalyse kurz thematisiert werden.

1.3 Vorgehensweise und Aufbau dieses Berichts

Zur Beantwortung der Forschungsfrage wird ein Anwendungsszenario benötigt, in dem es möglich ist, Prognosen auf Basis empirischer Daten zu erstellen, deren Qualität zu messen und mit den Referenzsystemen zu vergleichen. Aus der Vielzahl der täglich in der Schweiz durchgeführten Verkehre ist dafür eine geeignete Auswahl zu treffen. Dabei sollen methodische Kriterien ebenso berücksichtigt werden wie die «Interessantheit» der Situation. Ebenso sind Zeithorizont der Prognose, Qualitätsmasse und das Niveau des akzeptierten alpha-Fehlers zu bestimmen. Kapitel 2 begründet zunächst, warum sich der Zugsverkehr für die Untersuchung besser eignet als andere Verkehre und liefert dann eine umfassende Einführung in die relevante Fachlichkeit. Es schliesst mit der Beschreibung eines Anwendungsszenarios.

Für fast aller Schritte dieser Arbeit – die Beschaffung und Aufbereitung der Daten; Implementierung, Training und Test des Prognoseverfahrens; die Durchführung von Prognosen; die Messung, Visualisierung und den Vergleich der Ergebnisse – wird Informationstechnik benötigt. Kapitel 3 gibt einen

Überblick über die verwendete IT-Architektur und beschreibt die einzelnen Komponenten und deren Zusammenspiel.

Gemäss Zielsetzung sollen ausschliesslich die «offenen» Daten von opentransportdata.ch verwendet werden. Kapitel 4 beschreibt die genutzten Daten, deren Qualität und Beschränkungen und erläutert, wie sie für die Zwecke der Untersuchung bezogen und aufbereitet werden.

Kern dieser Arbeit ist ein Prognoseverfahren, das mit empirischen Daten der Vergangenheit trainiert und dann auf neue Konstellationen (andere Vergangenheitsdaten oder Echtzeitdaten) angewendet werden kann. Um die Forschungsfrage positiv zu beantworten, muss dieses Verfahren gut genug sein, die Vergleichssysteme wenigstens in Teilbereichen zu übertreffen. Wie stark sie übertroffen werden, ist sekundär – zusätzliche Optimierungsschritte mögen interessant sein, bringen aber keinen weiteren Beitrag zur verfolgten Zielsetzung. Die Forschungsfrage negativ zu beantworten, ist dagegen nicht abschliessend möglich: auch wenn die auf empirischen Daten erzielten Prognosen hinter den Vergleichssystemen zurückbleiben sollten, könnte nicht ausgeschlossen werden, dass sich dies mit einem besseren Prognoseverfahren ändern liesse. In diesem Fall könnten höchstens überzeugende Argumente dafür geliefert werden, dass eine solche Verbesserung unwahrscheinlich ist – z. B. weil das verwendete Verfahren bereits sehr ausgefeilt ist und dem «state-of-the-art» entspricht oder weil die Diskrepanz zu den Vergleichssystemen sehr gross ist. Kapitel 5 geht ausgiebig auf die Gestaltung eines solchen Verfahrens, sowie auf Parametrisierung, Training und Test der verwendeten Modelle ein. Das vorgestellte Verfahren ist das Ergebnis zahlreicher Versuche und iterativer Verbesserungen. Die dabei getroffenen Design-Entscheide werden begründet und alternative Ansätze skizziert.

Das Prognoseverfahren lässt sich auf zweierlei Weise validieren: Einerseits kann es auf Vergangenheitsdaten angewendet werden. Dieses Vorgehen entspricht einer Simulation: Welche Prognose wäre jeweils erzielt worden, wenn man das Verfahren auf den Kenntnisstand eines vergangenen Zeitpunkts angewendet hätte? Und wie gut wäre die (in der Rückschau ja ebenfalls bekannte) Verspätung prognostiziert worden? Andererseits lässt sich das Prognoseverfahren auch auf Echtzeit-Daten anwenden. Dies ermöglicht einen direkten Vergleich mit den Referenzsystemen der Bahnunternehmen (deren vergangene Prognosen leider nicht bekannt sind). Beide Verfahren werden in Kapitel 6 beschrieben und die dabei verwendeten Qualitätsindikatoren vorgestellt.

Um den Verlauf der Untersuchung nachvollziehbar und überprüfbar zu machen und die erzielten Ergebnisse analysieren zu können, ist eine geeignete Visualisierung erforderlich. Im Sinne hoher Transparenz werden alle vom Verfahren gelieferten Prognosen öffentlich verfügbar gemacht. Um dabei den Ansprüchen unterschiedlicher Zielgruppen gerecht zu werden, kommen drei unterschiedliche Applikationen zum Einsatz, die in Kapitel 7 beschrieben werden.

Kapitel 8 dokumentiert die Ergebnisse der Untersuchung: Es liefert eine differenzierte Beantwortung der Forschungsfrage, bewertet das eingesetzte Verfahren und die verwendete Architektur und beschreibt weitere Erkenntnisse, die im Rahmen der Arbeit entstanden sind.

Der Bericht schliesst mit einer Zusammenfassung und einem Ausblick auf sinnvolle nächste Schritte.

2 Anwendungsdomäne und Untersuchungsszenario

Titel und Forschungsfrage dieser Arbeit beziehen sich auf den «öffentlichen Verkehr». Eine Einschränkung auf den öffentlichen *Personen*verkehr erscheint dabei legitim. Dieser umfasst sehr unterschiedliche Verkehrsmittel. In der Schweiz verbreitet sind Auto- und Trolleybusse, Trams, Schmalspur- und Normalspurbahnen, Zahnrad-, Stand- und Luftseilbahnen, Fluss- und Seeschifffahrt. Diese werden von über 250 Unternehmen betrieben.[2] Häufig wird dabei zwischen Ortsverkehr, Regionalverkehr und Fernverkehr unterschieden.[3] Abbildung 1 gibt einen Überblick über das Streckennetz.

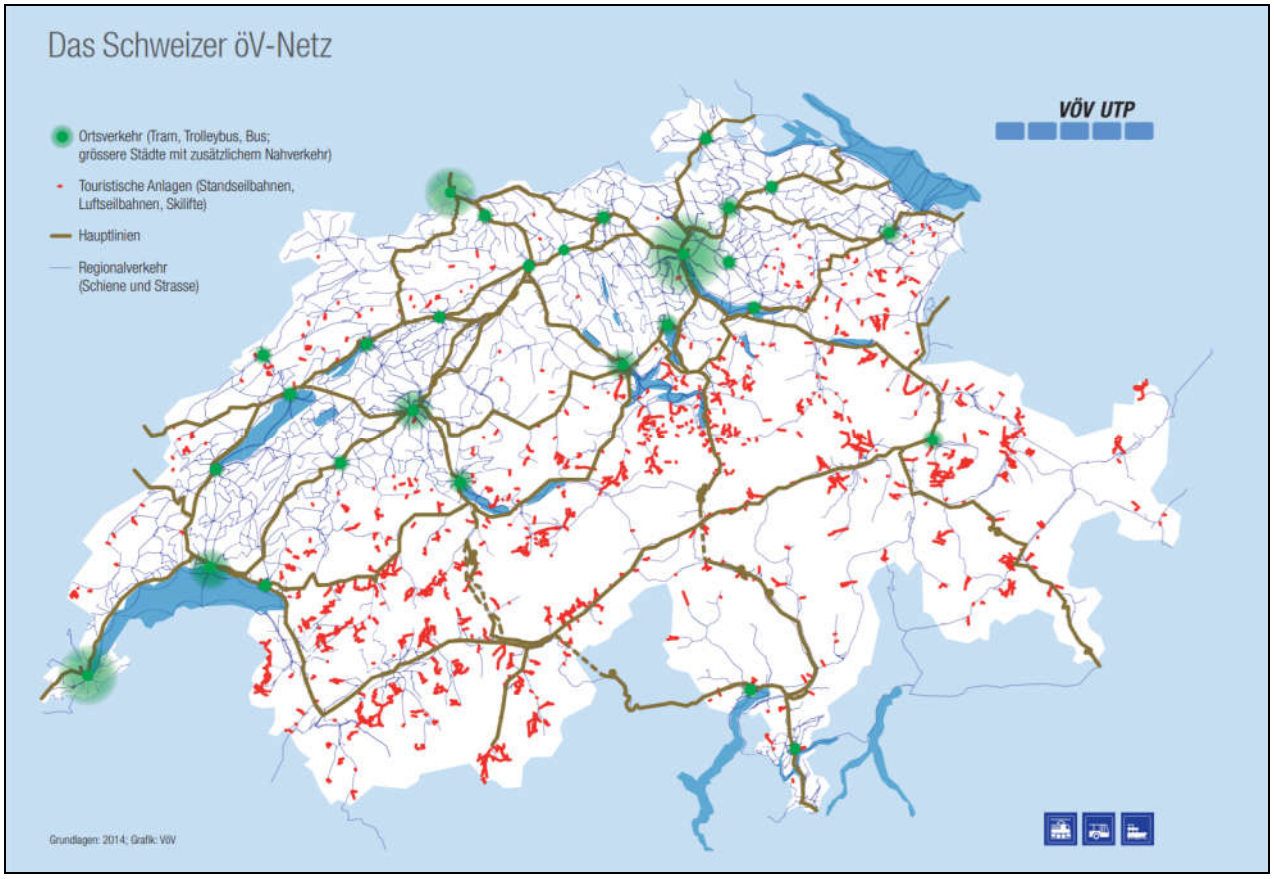


Abbildung 1: Das Schweizer öV-Netz[4]

Die Formulierung der Forschungsfrage[5] erlaubt es, die Untersuchung auf Teilbereiche einzugrenzen, was angesichts der vorherrschenden Vielfalt ratsam erscheint. Eine solche Eingrenzung könnte z.B. nach Art des Verkehrsmittels, nach Unternehmen, nach geographischen Kriterien oder auch unter Bezug auf ein konkretes Verkehrsangebot erfolgen. Ich habe mich entschieden, in einem ersten Schritt eine Eingrenzung auf den Zugsverkehr vorzunehmen. Dies aus folgenden Überlegungen:

[2] Das kleinste öV-Unternehmen der Schweiz ist übrigens die «Elektrischer Personenaufzug Matte-Plattform AG», die das Berner «Senkeltram» an der Münsterplattform betreibt. Das Streckennetz beträgt beachtliche 32m, davon 100% in der Vertikalen. Da es zwar Betriebszeiten, aber keinen eigentlichen Fahrplan gibt, sind Verspätungsprognosen dort kein Thema.

[3] Vgl. VöV (2016) und SBB (2017), S. 7.

[4] Quelle: VöV (2016).

[5] «Können [..] mindestens in Teilbereichen bessere Verspätungsprognosen erzielt werden, als sie von den aktuell im öV Schweiz eingesetzten Kundeninformationssystemen geliefert werden?

- Fahrtdauern von mehr als einer Stunde zwischen den Endhaltestellen erscheinen eher geeignet, Verspätungen auch mittelfristig vorhersagen zu können. Diese sind im Zugsverkehr häufig zu finden. Bei Ortsverkehren (Bus, Tram) und Seilbahnen sind Fahrten oft deutlich kürzer.
- Im Zugsverkehr treten Verspätungsübertragungen zwischen unterschiedlichen Fahrten häufiger auf als bei anderen Verkehrsmitteln: Einerseits, weil es mehr Situationen gibt, in denen Anschlüsse abgewartet werden (Umsteigebeziehungen); andererseits, weil es im spurgeführten Betrieb mehr Abhängigkeiten zwischen den Fahrwegen gibt (die Möglichkeiten für Kreuzungen, Überholungen, Zugsfolgen etc. werden durch die verfügbare Gleisanlage eingeschränkt). Auch dies könnte dazu beitragen, dass Verspätungen über einen längeren Zeitraum antizipiert werden können.
- Historische Daten und Echtzeitdaten zur Betriebslage sind für den Zugsverkehr seit Dezember 2016 nahezu flächendeckend vorhanden.[6] Für Regionallinien im Busverkehr (insbesondere Postauto) liegen Daten meist erst seit Mitte 2017 vor – und sind zudem auch heute noch lückenhaft.
- Zugsverkehr wird in der Regel auf separaten Verkehrswegen geführt, so dass es weniger systemexterne Störungseinflüsse gibt. Verspätungsursachen können daher womöglich leichter aus den Betriebslage-Daten erkannt werden. Insbesondere der Einfluss des Autoverkehrs entfällt.[7]
- Die Fahrpläne im Zugsverkehr weisen tendenziell eine höhere Frequenz und Regelmässigkeit auf als Regionalbusse und Schiffslinien, so dass zu einer gegebenen Prognosesituation deutlich mehr empirische Daten aus ähnlichen Situationen als «Erfahrungswissen» genutzt werden können.

2.1 Charakteristika des Zugsverkehrs[8]

Zu den offensichtlichen Eigenschaften des Zugsverkehrs gehört die Spurführung: Anders als Busse und Schiffe können die Fahrzeuge ihre Fahrtrichtung nicht autonom bestimmen. Sie folgen dem Gleisverlauf und werden auch an Abzweigen («Weichen» im Wortsinne) durch den Fahrweg «gelenkt». Diese Besonderheit hat diverse Konsequenzen:[9]

- Es können grosse Massen mit grossen Geschwindigkeiten und hoher Energieeffizienz[10] bewegt werden. Züge sind ein «Massenverkehrsmittel» - sowohl für Güter als auch für Personen. Dass die Schweiz ein «Bahn-Land»[11] ist, hat viel damit zu tun, dass Verkehrsströme dort oft linear entlang von Tälern, Pässen und Bergketten verlaufen – wohingegen die netzartigen Verkehrsströme in flacheren Regionen den Individualverkehr begünstigen.
- Aufgrund der geringeren Reibung sind die Bremswege sehr lang – oft deutlich länger als die Sichtdistanz. Dies erfordert umfangreiche Sicherungsmassenahmen. Offensichtlich ist, dass eine Begegnung von entgegengesetzt fahrenden Zügen auf dem gleichen Gleis verhindert werden muss («Gegenfahrschutz»). Auf eingleisigen Strecken ist eine Begegnung («Kreuzen») daher nur in Bahnhöfen oder an speziellen Ausweichstellen möglich.

[6] Einige Schmalspurbahnen liefern derzeit noch keine Daten, unter anderem der Regionalverkehr Bern-Solothurn (RBS).

[7] Bei Schmalspurbahnen gibt es zahlreiche Abschnitte, wo im sogenannten «Tram-Betrieb» öffentliche Strassen befahren werden – und somit eine starke Abhängigkeit zum Autoverkehr existiert. Bei der Definition des Prognoseszenarios in Abschnitt 2.5 wird darauf geachtet, solche Verkehre von der Betrachtung auszuschliessen.

[8] In diesem und den folgenden Abschnitten werden Sachverhalte des Schienenverkehrs dargestellt, die für diese Arbeit von Relevanz sind. Sofern es sich dabei um Aussagen handelt, die ich aus anderen Quellen übernommen habe, sind diese Quellen angeben. Überall dort, wo keine Quelle genannt ist, handelt es sich um Wissen aus meiner 12-jährigen Tätigkeit in der Branche.

[9] Vgl. Pachl (2016), S. 1.

[10] Dies aufgrund der wesentlich geringeren Reibung Stahlrad / Stahlschiene im Vergleich zu Gummireifen / Asphalt.

[11] Bei diversen Indikatoren nehmen die Schweizer Bahnen weltweit eine Spitzenposition ein, so z.B. bei der Dichte des Streckennetzes (Streckenkilometer pro Quadratkilometer), bei der Zugsdichte (Anzahl Züge pro Streckenkilometer und Tag), bei der Anzahl Fahrten / Personenkilometern pro Einwohner – und auch bei der Pünktlichkeit. Vgl. VöV (2016), S. 22 und SBB (2017).

- Zu den weit verbreiteten Grundprinzipien[12] eines sicheren Bahnbetriebs gehört das «Fahren im festen Raumabstand»:[13] Dabei darf ein Zug nur dann in einen Gleisabschnitt einfahren, wenn der Abschnitt als auch der dahinter liegende «Durchrutschweg»[14] frei sind (vgl. Abbildung 2).

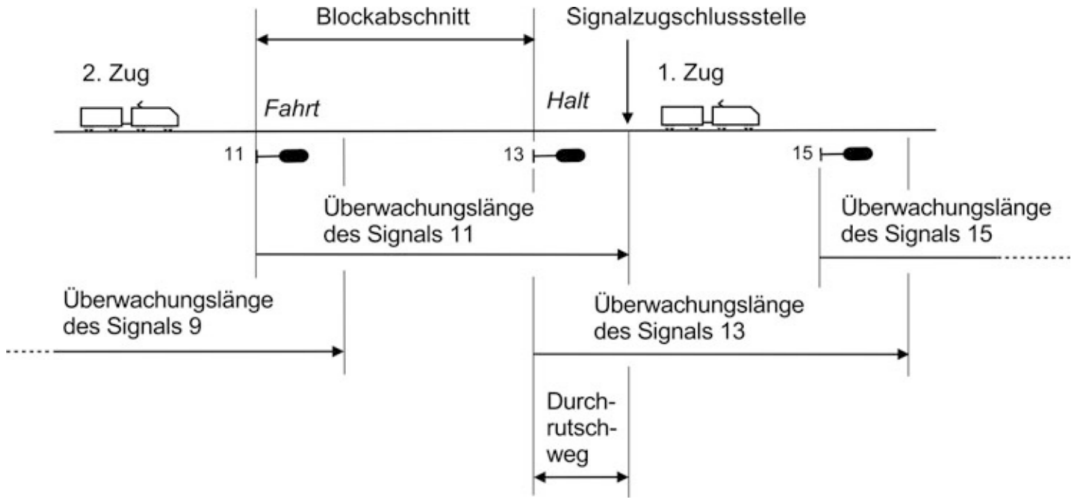


Abbildung 2: Sicherer Bahnbetrieb durch Fahren im festen Raumabstand.[15]

- Um den Raumabstand zu sichern, werden meist ortsfeste Aussensignale verwendet. Deren Anordnung und Ausgestaltung bestimmt den minimalen räumlichen und zeitlichen Abstand von aufeinanderfolgenden Zügen (Zugfolgedistanz bzw. Zugfolgezeit) – und damit die Kapazität der Strecke. Es existieren diverse, oft aufwendige signaltechnische Massnahmen, um die Zugfolgezeit gering zu halten.[16] Selbst bei Führerstandsignalisierung[17] sind im regulären Betrieb Abstände unter 2 Minuten bisher nicht möglich[18] – auf den meisten Abschnitten liegt sie heute deutlich höher.
- Aufgrund der langen Bremswege müssen Halt zeigende Signale vorab angekündigt werden: Die Sichtdistanz bis zum Signal würde schon bei mittleren Geschwindigkeiten nicht ausreichen, um einen Zug rechtzeitig zum Stehen zu bringen. Aus diesem Grund werden Vorsignale verwendet, die in ausreichender Distanz vor dem eigentlichen Signal («Hauptsignal») aufgestellt werden und dessen Anzeige vorankündigen (vgl. Abbildung 3).[19] Sieht der Lokführer am Vorsignal die Anzeige für «Halt erwarten», wird er so stark abbremsen, dass der Zug vor dem Hauptsignal anhalten kann. Wird der nachfolgende Abschnitt in der Zwischenzeit freigegeben, so kann er den Zug wieder beschleunigen, noch bevor er zum Stillstand gekommen ist.

[12] Mit Ausnahme von Situationen, in denen aufgrund geringer Geschwindigkeiten «auf Sicht» gefahren werden kann, ist dies das heute fast ausschliesslich verwendete Verfahren der Zugsicherung.

[13] Vgl. Pachl (2016), S. 35ff.

[14] Der «Durchrutschweg» ist der Bremsweg, welcher für eine automatische Notbremsung benötigt würde, wenn ein Halt zeigendes Signal am Ende des Abschnitts überfahren würde. Diese Länge ist u.a. von Geschwindigkeit und Bremsverhalten des Zugs abhängig, was zusätzliche Massnahmen erforderlich macht.

[15] Quelle: Pachl (2016), S: 40.

[16] Vgl. Pachl (2016), S. 38ff.

[17] Bei Führerstandsignalisierung werden keine Aussensignale verwendet. Die Fahrerlaubnis wird stattdessen elektronisch in die Lok übermittelt und dem Lokführer im Cockpit angezeigt.

[18] Vgl. SBB (2016). Noch geringere Distanzen können grundsätzlich mit Fahren im «Moving Block» erreicht werden. Dies ist seit vielen Jahren Gegenstand intensiver Forschungs- und Entwicklungsaktivitäten, die nicht zuletzt von den Schweizer Bahnen – aktuell im Programm «SmartRail» - vorangetrieben werden.

[19] Vgl. Pachl (2016), S. 45. Das Prinzip gilt in ähnlicher Weise auch bei heutiger Führerstandsignalisierung.

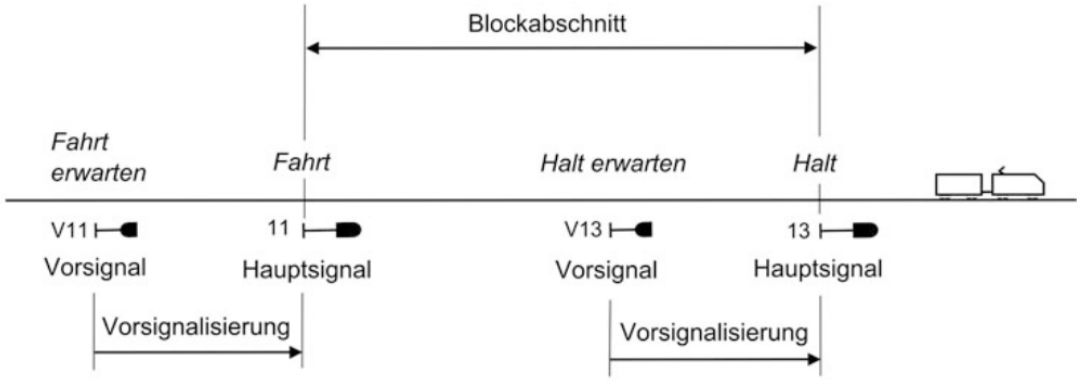


Abbildung 3: Prinzip der Vorsignalisierung.[20]

- Das Prinzip des Raumabstands gilt in verallgemeinerter Form auch für Bereiche, in denen Fahrwege verzweigen oder kreuzen: Es muss sichergestellt sein, dass kein Zug in einen Bereich fahren (oder «rutschen») kann, in denen sich möglicherweise noch ein anderes Fahrzeug (und sei es ein «entlaufener» Wagen) aufhält.[21] Dies schränkt die betriebliche Nutzung (und damit die Kapazität) der Gleisanlage ein: Züge müssen ggf. abbremsen und anhalten, bis ein anderer Zug den gemeinsam beanspruchten «Konfliktbereich» freigegeben hat (vgl. Abbildung 4).

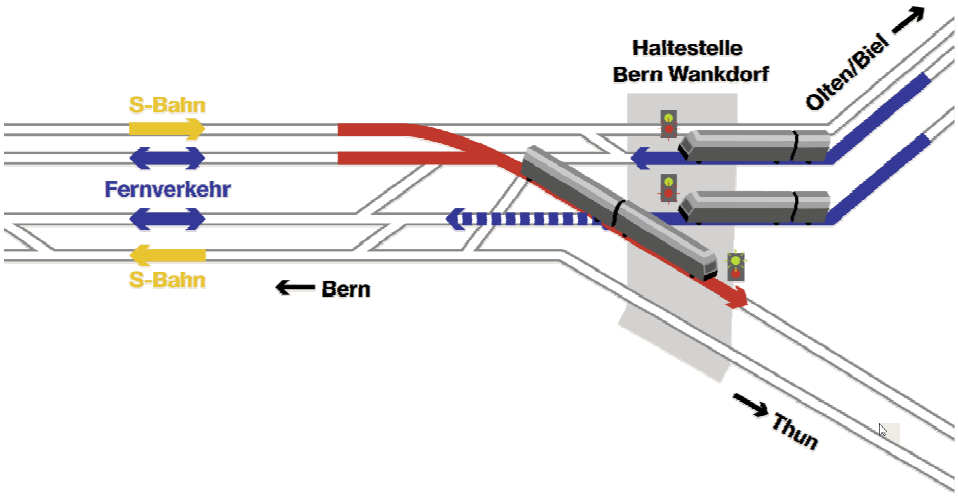


Abbildung 4: Abkreuzungskonflikt: Züge aus Olten müssen warten, bis jener aus Bern die Gleise freigibt.[22]

- Die für einen Streckenabschnitt benötigte Fahrzeit ergibt sich aus dem Geschwindigkeitsprofil des jeweiligen Zugs. Dies wiederum wird bestimmt durch die zulässigen Geschwindigkeiten (u.a. im Hinblick auf Fahrwegeigenschaften, Kurvenradien, Signalsichtdistanzen) einerseits und das Beschleunigungs- und Bremsvermögen des Zugs andererseits. Wichtige Parameter sind Antriebsleistung, Bremskraft, aber auch Länge und Gewicht der Fahrzeuge («Rollmaterial»).
- Der Erstellung von Fahrweg und Sicherungsanlagen ist aufwendig – sowohl in zeitlicher als auch in finanzieller Dimension: Änderungen erfordern grossen Vorlauf.[23] Und die erforderlichen Investitionen sind nur bei einer langfristigen Nutzung gerechtfertigt. Entsprechend vorausschauend müs-

[20] Quelle: Pachl (2016), S: 45.

[21] Vgl. Lüthi (2009), S. 13.

[22] Quelle: SBB (2014), S. 3.

[23] Vgl. Lüthi (2009), S. 10.

sen sie geplant werden. Zugleich ist der Entwurf einer «optimalen» Anlage aber nur unter Kenntnis der beabsichtigten Nutzung möglich. Einige Beispiele: Je nach Fahrzeit und Takt werden Kreuzungs- und Überholstellen an anderen Stellen benötigt. Die Haltepolitik und Länge der Züge beeinflusst die benötigten Bahnhofgleise. Überwerfungen sind nur dort erforderlich, wo Fahrwegkonflikte auch tatsächlich auftreten. Angebot, Rollmaterial und Infrastruktur beeinflussen sich also gegenseitig – in der Bahnbranche spricht man vom «Planungsdreieck».[24]

2.2 Planung des Zugsverkehrs: Fahrplan-Konstruktion

Planung ist die Festlegung eines Sollzustands. Durch die Konstruktion eines Fahrplans wird der angestrebte Betriebsablauf festgelegt.[25] Dem Fahrplan kommen drei Aufgaben zu:[26]

1. Die Koordination unterschiedlicher Nutzer einer Eisenbahn-Infrastruktur: Wie in den meisten europäischen Ländern, so sind auch in der Schweiz Bahninfrastruktur und Bahnverkehr organisatorisch getrennt und können von unterschiedlichen Unternehmen betrieben werden. So wird beispielsweise das Normalspurnetz im Raum Bern von Personenzügen der SBB, der BLS und der TPF befahren – sowie von diversen Güterbahnen. Die Aufteilung der vorhandenen Kapazität zwischen diesen Nutzern wird durch den Fahrplan geregelt. Dieser berücksichtigt zudem eine weitere «Nutzergruppe»: auch Unterhalts- und Umbauarbeiten nehmen Anlagenkapazität in Anspruch. Damit bei dieser aufwendigen Koordinationsaufgabe sowohl kurzfristige als auch langfristige Bedürfnisse Berücksichtigung finden, erfolgt die Planung in unterschiedlichen Zeithorizonten – von mehreren Jahren im Voraus bis kurz vor Abfahrt des Zuges. Am sichtbarsten für den Reisenden ist der Jahresfahrplan, der jeweils am zweiten Sonntag im Dezember wechselt: Seit dem 10. Dezember 2017 gilt die «Fahrplanperiode 2018».[27]
2. Die Bereitstellung der für den Betriebsablauf erforderlichen Informationen an alle Beteiligten: Zahlreiche Personen – im Zug, am Bahnhof, in Fernsteuerzentrum und Betriebszentrale, aber auch z.B. bei der Energieversorgung und der Personaleinteilung – sind an der Vorbereitung und Produktion einer Zugfahrt beteiligt. Ein wichtiges Ziel bei der Erstellung des Fahrplans besteht darin, dass der angestrebte Betriebsablauf auch umsetzbar ist: Loks, Wagen und Personal müssen verfügbar sein und die erforderlichen Eigenschaften aufweisen. Ebenso muss die Gleisanlage frei und für den geplanten Verkehr geeignet sein. Insbesondere wird darauf geachtet, dass Zugfolgezeiten eingehalten werden und es möglichst nicht zu Fahrwegkonflikten kommt:[28] Signalhalte sollen nach Möglichkeit vermieden werden, weil sie Reisezeit, Anlagenkapazität und Energie kosten. Der Fahrplan beschreibt also einen für alle Beteiligten «machbaren» Betriebsablauf. Und er sorgt dafür, dass sie einen einheitlichen Kenntnisstand über den Soll-Ablauf haben.
3. Die Information der Reisenden über das Verkehrsangebot: Während ein betrieblicher Plan sehr detailliert und mit zahlreichen Informationen versehen ist, besteht bei den Reisenden das Bedürfnis nach leichter Verständlichkeit und Einfachheit. Die meisten Bahnen unterscheiden daher zwischen «betrieblichem» und «kommerziellem»[29] Fahrplan. Diese Unterscheidung macht sich sowohl im Umfang als auch in der Genauigkeit der enthaltenen Angaben bemerkbar: Betriebliche Fahrpläne enthalten z.B. Zeitangaben zu Durchfahrten in allen Bahnhöfen, Abzweigen und Spurwechseln

[24] Vgl. Butler (2017), Folie 19. Zusammen mit Immobilien und Finanzierung ergibt sich sogar ein «Planungsfünfeck».

[25] Vgl. Scholz (2012), S. 127.

[26] Vgl. im Folgenden Pachl (2016), S. 185ff.

[27] Vgl. Caimi / Kroon / Liebchen (2017), S. 287.

[28] Aufgabe des Fahrplans ist es, einen möglichst reibungslosen Betriebsablauf zu ermöglichen. Die Verhinderung von Unfällen ist dagegen Aufgabe der Sicherungstechnik: Sollte es doch zu Fahrwegkonflikten kommen, so werden Kollisionen durch Signale und andere Schutzvorrichtungen verhindert. Es ist generell nicht möglich, dass durch einen Planungsfehler ein Unfall verursacht wird.

[29] Oder auch: «veröffentlichtem»

(«Betriebspunkte»). Den Reisenden interessieren dagegen nur die Ankunfts- und Abfahrtszeitpunkte an Einstiegs- und Ausstiegsorten. Betriebliche Fahrpläne sind in der Schweiz Zehntelminuten-genau und können Unregelmässigkeiten im Tagesverlauf (z.B. bedingt durch Unterschiede in Rollmaterial, Gleisbelegung, Fahrgastfrequenzen) und Jahresverlauf (z.B. bedingt durch Baustellen) abbilden. Im kommerziellen Fahrplan werden Zeiten in ganzen Minuten angegeben – und es wird eine hohe Stabilität angestrebt: «Die Liebe der […] Fahrgäste zu festen Taktzeiten geht so weit, dass man sie mitunter selbst dann noch verwendet, wenn sie erfahrungsgemäss nicht ganz eingehalten werden können […]. 'Leicht zu merken' ist manchmal wichtiger und sinnvoller als eine komplizierte empirische Wahrheit.»[30] Die Umrechnung der betrieblichen Zeiten in kommerzielle Zeiten erfolgt nach einem nicht ganz einfachen Regelwerk. Wichtige Anforderung ist dabei, dass ein Zug nach einem Halt niemals früher abfahren darf (→ betriebliche Zeit) als es den Fahrgästen kommuniziert wurde (→ kommerzielle Zeit): Verspätungen bei der Abfahrt sind *unerwünscht*, Verfrühungen sind *inakzeptabel*.[31]

Die Fahrt eines Zuges vom Abgangsbahnhof A zum Zielbahnhof Z besteht in der Regel aus mehreren Abschnitten und dazwischenliegenden Halten. Die Gesamt-Reisedauer vom A bis Z ergibt sich als Summe der Fahrzeiten auf allen Abschnitten und der Haltezeiten bei allen Halten. Diese können ihrerseits zerlegt werden in

- eine Minimal-Zeit, die auch im besten Fall nicht unterschritten werden kann und
- diverse Zuschläge und Reserven.

Abbildung 5 verdeutlicht den Sachverhalt: Die minimale Haltezeit ist hier diejenige Dauer, die für den Passagierwechsel in A benötigt wird. Bei der SBB gibt es hierzu vordefinierte Werte je Zugskategorie und Bahnhof. Hinzu kommen Aufschläge für das Abwarten von kreuzenden Zügen und Anschlüssen sowie den Abfahrprozess (u.a. Türschliessung; der Zeitbedarf ist abhängig vom eingesetzten Rollmaterial). Die Mindestfahrzeit auf dem folgenden Abschnitt wird durch das zulässige Geschwindigkeitsprofil und die Fahrdynamik des Zuges bestimmt. Sie repräsentiert die Fahrzeit unter Idealbedingungen – das Schnellste, was gemäss Vorschriften und Physik möglich ist. Aufgeschlagen werden

- ein prozentualer Wert für betriebliche Schwankungen,
- ein prozentualer Wert für «Langsamfahrstellen» (bedingt durch Baustellen),
- absolute Zuschläge, um der konkreten Situation Rechnung zu tragen, z.B. die Folge auf einen vorausfahrenden langsameren Zug.

Abbildung 5: Bestandteile von Haltezeiten und Fahrzeit bei einer Fahrt von A nach B[32]

[30] Scholz (2012), S. 143.

[31] Scholz (2012), S. 144.

[32] Quelle: Butler (2017), Folie 44. Eine etwas andere Systematik verwendet Lüthi (2009), S. 25.

Abbildung 6 zeigt am Beispiel einer S-Bahn-Fahrt zwischen dem Abzweig Riet (RT, bei Kloten) und Winterthur (W), wie diese Zusammenhänge in «NeTS», dem Fahrplansystem der SBB, abgebildet werden: Jede Zeile repräsentiert einen Betriebspunkt und (anders als oben) den *vorausgehenden* Fahrtabschnitt. Deutlich zu erkennen sind die Unterschiede zwischen betrieblichen (betrAn und betrAb) und kommerziellen Zeitpunkten (kommAn und kommAb). Betriebspunkte, in denen nicht gehalten wird, haben keinen Eintrag für betrAn. Die Fahrzeit (betrFz) setzt sich aus der minimalen Fahrzeit (tFz) und mehreren Zuschlägen (FzR, fr, zFzR) zusammen. Ebenso setzt sich die Haltezeit (betrHz) aus minimaler Haltezeit (minHz) sowie mehreren Zuschlägen zusammen (HzR, ZAZ). Wichtig zu verstehen ist, dass es sich bei Fahrplanangaben letztlich um stochastische Grössen handelt: die eingerechneten Reserven dienen ja gerade dazu, bei Bedarf Schwankungen auffangen zu können – die angegebenen Zeitpunkte sind also als «wahrscheinliche» Grössen zu verstehen.

BP	Gleis	A..	Fahrweg	betrAn	betrAb	frühAn	frühAb	kommAn	kommAb	betrFz	tFz	FzR	fR	zFzR	Haltezwecke	betrHz	minHz	HzR	ZAZ	Haltepos
RT	83		RT83-509		07:32.2					0.7	0.7	0.1	0.0	-0.1						
KLB	1		KL3	07:32.8	07:33.3	07:32.4	07:32.7	07:32	07:32	1.7	1.4	0.1	0.0	0.2	1: 11	0.5	0.3	0.1	0.1	MitteAg
KL	3		KL22	07:35.0	07:36.1	07:34.2	07:36.0	07:35	07:36	1.8	1.6	0.1	0.0	0.1	1: 11	1.1	0.3	0.7	0.1	MitteAg
DORF	11		510-DORF11-11...		07:37.9					1.2	1.2	0.1	0.0	-0.1						
BSD	2		112-113-HUER991	07:39.0	07:39.5	07:38.8	07:39.1	*07:39*	07:39	2.3	2.3	0.2	0.0	-0.2	1: 11	0.5	0.3	0.1	0.1	MitteAg
HUER	991		114-EF1		07:41.8					2.2	1.2	0.1	0.0	0.9						
EF	1		EF81-118-120	07:44.0	07:45.1	07:42.7	07:45.0	07:44	07:45	2.9	2.6	0.2	0.0	0.1	1: 11	1.1	0.3	0.7	0.1	MitteAg
KE	2		KE2-121-TOEM24	07:48.0	07:48.5	07:47.7	07:48.0	*07:48*	*07:48*	1.8	1.6	0.1	0.0	0.1	1: 11	0.5	0.3	0.1	0.1	MitteAg
TOEM			TOEM95-924-W15		07:50.3					1.1	1.0	0.1	0.0	0.0						
WWE			W35-W45-W55-...		07:51.4					2.6	1.5	0.1	0.0	1.1						
W	6			07:54.0	07:55.1	07:52.1		07:54							1: 11	1.1	1.0	0.0	0.1	MitteAg

Abbildung 6: SBB-Planungswerkzeug «NeTS»: Fahrzeiten und Haltezeiten sowie ihre Bestandteile[33]

Die Kunst der Fahrplankonstruktion besteht gerade darin: Die Schwankungen zu antizipieren und im *richtigen* Ausmass zu berücksichtigen. Ziel ist es, ein ausgewogenes Verhältnis von Stabilität und Kapazität zu finden (vgl. Abbildung 7). Eine hohe Stabilität ist erzielbar, indem grosse Reservern eingeplant werden: Wenn Fahr- oder Haltezeiten einmal länger ausfallen als erwartet, können Reserven «konsumiert» werden und der Betrieb kehrt schnell zum planmässigen Zustand zurück. Reserven reduzieren aber häufig die Kapazität der Anlage: die «reservierte» Zeit kann nicht für andere Zugfahrten genutzt werden. Umgekehrt führen niedrige Reserven dazu, dass bereits kleine Störungen das System über längere Zeit aus dem Plan bringen. Die so entstehenden Verspätungen vernichten ebenfalls Kapziät. Im Extremfall schaukeln sie sich auf und führen zur Überlastung des Sytems.

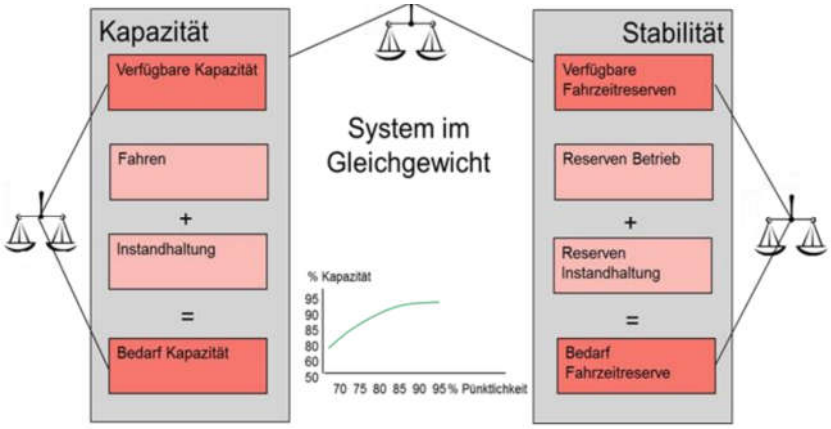


Abbildung 7: Ein guter Fahrplan bringt Kapazität und Stabilität ins Gleichgewicht[34]

[33] Quelle: SBB (2014), S. 3. Vorsicht beim Nachrechnen: das abgebildete Beispiel enthält Rundungsdifferenzen!

In der Schweiz gibt es seit 1982 einen flächendeckenden Taktfahrplan: auf den meisten Linien verkehren Züge in regelmässigen Abständen (meist stündlich, halbstündlich oder viertelstündlich). Schrittweise wurde dies zu einem integralen Taktfahrplan erweitert: in so genannten «Knoten-Bahnhöfen» treffen die Züge möglichst vieler Linien etwa gleichzeitig ein, warten einander ab und fahren auch etwa gleichzeitig wieder ab. Dieses «Rendez-Vous»-Prinzip ermöglicht den Reisenden gute Umsteigemöglichkeiten zwischen allen Linien – vorausgesetzt, die Anschlüsse werden auch gehalten.[35] Abbildung 8 zeigt ein Beispiel: Züge von 3 Linien (rot, rosa und blau) verkehren im Stundentakt zwischen A-city und B-city. A-city ist ein Knoten-Bahnhof («Hub»): Hier bestehen jeweils zur vollen Stunde Umsteigebeziehungen zwischen allen Linien und Fahrtrichtungen: je fünf Züge kommen kurz vor Minute :00 an, je fünf Züge fahren kurz nach Minute :00 weiter (die blaue Linie startet und endet in A-city, daher sind es nicht 6 Ankünfte und Abfahrten).

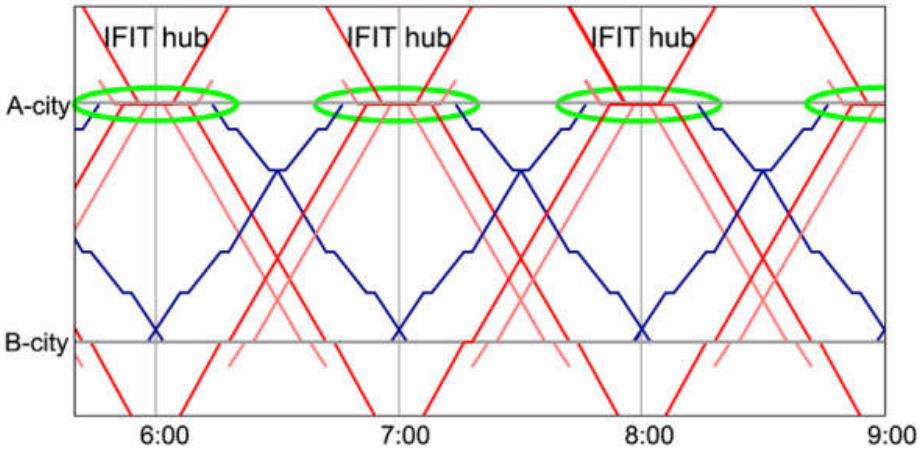


Abbildung 8: Integrierter Taktfahrplan mit Rendezvous in A-city[36]

Im Schweizer Fernverkehr gibt es heute 8 derartige Hubs: In Basel, Zürich, Olten, Bern und Chur treffen sich die Züge zur vollen und zur halben Stunde; in St. Gallen, Biel und Lausanne zur Minute :15 und :45 (vgl. Abbildung 9).

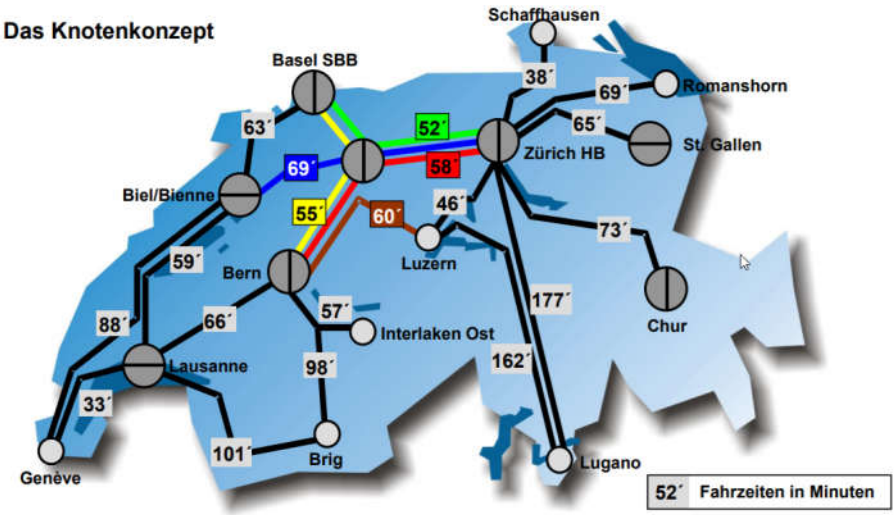


Abbildung 9: Heutige Fernverkehrs-Knoten in der Schweiz[37]

---

[34] Quelle: Butler (2017), Folie 39.

[35] Vgl. Pachl (2016), S. 213f. und Caimi / Kroon / Liebchen (2017), S. 290f.

[36] Quelle: Caimi / Kroon / Liebchen (2017), S. 291.

Das Rendezvous-Verfahren ist für die Reisenden vorteilhaft, betrieblich jedoch anspruchsvoll: Es führt dazu, dass jeweils viele Züge gleichzeitig oder dicht hintereinander in die Knoten ein- und aus ihnen ausfahren. Auch auf den Strecken (z.B. Olten-Bern) sind dichte Zugfolgen erforderlich. Das Risiko, dass sich Züge gegenseitig behindern und sich Verspätungen fortpflanzen, ist im integralen Taktfahrplan grundsätzlich höher.[38]

Um gleich gute Umsteigebeziehungen bei Hin- und Rückreise zu erzielen, versucht man, die Fahrplanzeiten von entgegenkommenden Zügen derselben Linie «symmetrisch» zu gestalten. In der Schweiz liegt die Symmetrieachse bei Minute :00: Wenn ein Zug *aus* X zur Minute :56 (= «4 vor») ankommt, dann fährt der Gegenzug *nach* X zur Minute :04 ab (= «4 nach»). Dieses Prinzip gilt nicht nur in den Knotenbahnhöfen, sondern strahlt auf die ganze Schweiz aus. Abweichungen bestehen dort, wo eine Umsetzung aus betrieblichen Gründen nicht möglich ist (z.B. bei Einspur-Betrieb). Abbildung 10 zeigt die Ankünfte (rot) und Abfahrten (blau) der Fernverkehrszüge im Bahnhof Bern.

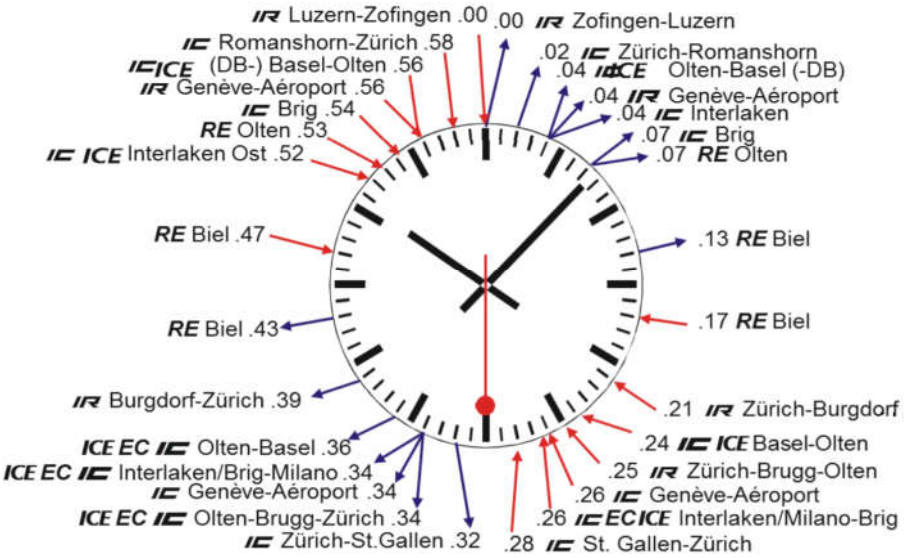


Abbildung 10: Abfahrt und Ankunft des Fernverkehrs in Bern (Fahrplanperiode 2016)[39]

2.3 Verspätungen und ihre Auswirkungen

Verspätungen sind Abweichungen vom Fahrplan. Sie ergeben sich als Differenz zwischen tatsächlichem Ereigniszeitpunkt und geplantem Ereigniszeitpunkt.[40] In dieser Arbeit wird die Kundensicht eingenommen, massgeblich sind hier also die dem Kunden kommunizierten «kommerziellen Zeitpunkte». Negative Abweichungen werden meist als Verfrühungen bezeichnet – sie sollen hier aber nicht weiter betrachtet werden.

Zugsverspätungen können unterschieden werden in:[41]

- Primärverspätungen (Urverspätungen): Verspätungen, die unmittelbar aufgrund einer Störung (z.B. Stellwerkstörung, Türstörung) oder eines äusseren Einflusses (z.B. Fahrgastaufkommen, Wetter) auftreten.[42]

---

[37] Quelle: Butler (2017), Folie 6.

[38] Vgl. Liebchen (2008).

[39] Quelle: Butler (2017), Folie 31.

[40] Vgl. Büker (2010), S. 25.

[41] Vgl. Flier / Gelashvili / Graffagnino / Nunkesser (2009), S. 355 und Büker (2010), S. 25.

[42] Einen Überblick über die Ursachen von Primärverspätungen liefert Ullius (2005), S. 34.

- Sekundärverspätungen (Folgeverspätungen): Verspätungen, welche von anderen Zügen übertragen werden, z.B. aufgrund von Zugfolgen, Fahrwegkonflikten, Anschlussbeziehungen.

Ob sich a) Verspätungen fortpflanzen und b) wie schnell sie wieder abgebaut werden, unterliegt zahlreichen Zusammenhängen, die sich trotz intensiver Forschungsbemühungen kaum vollständig beschreiben und modellieren lassen. In der Praxis können u.a. folgende Effekte beobachtet werden:

- Eine bestehende Verspätung kann während der Fahrt grundsätzlich über die bestehenden Fahrzeit- und Haltezeitreserven abgebaut werden. Abbildung 11 zeigt noch einmal das schon bekannte Beispiel aus dem Planungssystems «NeTS». Hervorgehoben sind die Fahrzeitreserven (rot, Summe = 2 Minuten) und die Haltezeitreserven (blau, Summe = 0.9 Minuten) zwischen Kloten (KL) und Winterthur (W). Ein in Kloten um 2.5 Minuten verspätet abgefahrener Zug könnte seine Verspätung demnach grundsätzlich bis zur Ankunft in Winterthur aufholen. Voraussetzung ist, dass sich die angenommenen Minimalwerte tatsächlich realisieren lassen – was einer Fahrt unter Idealbedingungen gleichkäme. Dazu gehört auch, dass der Lokführer – der grundsätzlich zu einer energieeffizienten Fahrweise angehalten ist – seinen Fahrstil entsprechend anpasst. Er wird dies zum Beispiel nicht tun, wenn er davon ausgeht, einen behindernden Zug vor sich zu haben.

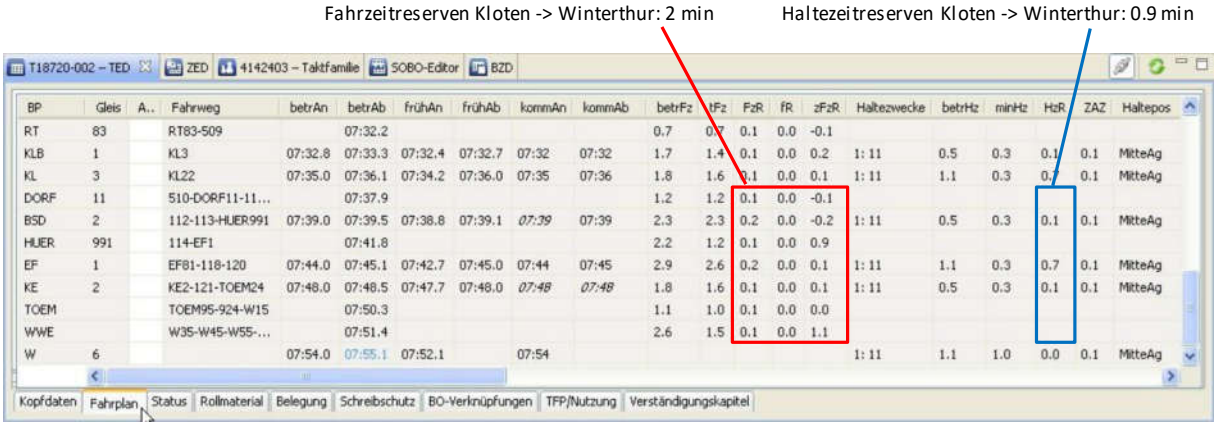


Abbildung 11: Nutzbare Reserven zwischen Kloten (KL) und Winterthur (W)[43]

- Wenn die Verspätung durch Einflüsse zustande gekommen ist, die am betrachteten Zug weiterhin bestehen, so ist nicht davon auszugehen, dass sie rasch aufgeholt wird. Sie könnte sich sogar noch vergrössern. Dies ist z.B. der Fall bei schlechter Witterung, hohem Fahrgastaufkommen, Einschränkungen von Leistung oder Kapazität des eingesetzten Rollmaterials.
- Wenn in einem Umsteigebahnhof ein Anschluss abgewartet wird, so übertragen sich Verspätungen auf den wartenden Zug. Dies jedoch nur in einem gewissen Bereich: kleine Verspätungen führen oft noch nicht dazu, dass gewartet werden muss; bei grossen Verspätungen wird der Anschluss nicht mehr abgewartet («Anschlussbruch»). Regeln darüber, wir lange welcher Zug auf welchen wartet, werden für jeden Bahnhof erstellt. Im Einzelfall kann davon abgewichen werden.
- Die eingesetzten Fahrzeuge und Mitarbeiter sind nach dem Ende einer Fahrt meist für eine Folgefahrt eingeplant. Wenn diese «Umläufe» knapp bemessen sind, übertragen sich Verspätungen der ersten Fahrt auf die Folgefahrt. Ein Bruch dieser Abhängigkeit ist nur möglich, wenn Ersatzpersonal bzw. Ersatzfahrzeuge einsetzbar sind oder indem eine «Kurzwende»[44] durchgeführt wird. Umlaufkonflikte können auch an einem Unterwegshalt auftreten, wenn dort das Personal gewechselt wird oder wenn Zugteile getrennt oder vereinigt werden.

[43] Abbildung basiert auf SBB (2014), S. 3.

[44] Bei einer Kurzwende fällt ein stark verspäteter Zug auf dem letzten Fahrtabschnitt aus. Die Fahrzeuge werden früher als geplant gewendet, um möglichst pünktlich die Rückfahrt antreten zu können. Vgl. Scholz (2012), S. 244.

- Bei Unterschreiten der Zugfolgezeit überträgt sich die Verspätung eines vorausfahrenden Zugs auf den Folgezug. Diese Situationen sind besonders häufig, wenn die Zugfolge dicht ist oder eine Strecke von Zügen mit unterschiedlicher Geschwindigkeit / Haltepolitik befahren wird (z.B. IC / S-Bahn oder Personenverkehr / Güterverkehr).
- Wenn die Ursache der Verspätung fortbesteht, aber nicht an einen Zug, sondern einen Ort gebunden ist (z. B. Ausfall oder Sperrung von Gleisen), so ist davon auszugehen, dass auch nachfolgende Züge davon betroffen werden, sobald sie dieselbe Stelle passieren. Hierbei handelt es sich nicht um Folge-, sondern um Primärverspätungen. Die Symptome sind aber ähnlich: nachfolgende Züge werden verspätet.
- Wie oben erläutert wurde, werden Fahrpläne so konstruiert, dass Signalhalte vermieden werden. Mit Auftreten von Verspätungen wird dieser geplante Zustand verlassen – das Risiko für Fahrwegkonflikte zwischen den Zügen nimmt zu. Ob es tatsächlich dazu kommt, und welche Auswirkungen sich ergeben, hängt von der Verspätungssituation aller beteiligten Züge ab. Ein Beispiel: Die Züge A und B passieren nacheinander eine Konfliktstelle – etwa aufgrund einer Abkreuzung, wie aus Abbildung 4 bekannt. Gemäss Fahrplan fährt A vor B und der zeitliche Abstand ist ausreichend gross. Wie Abbildung 12 zeigt, wirken sich Verspätungen von A und B *vor* der Konfliktstelle in einem komplexen Zusammenspiel auf die Verspätungen *hinter* der Konfliktstelle aus. Auf der X-Achse ist die Initialverspätung von A, auf der Y-Achse jene von B aufgetragen. Links ist die resultierende Verspätung von B als Rot-Schattierung dargestellt: Verkehrt A pünktlich, so reicht der zeitliche Puffer aus und B wird nicht zusätzlich verspätet, es bleibt bei der Initialverspätung. Ist A stärker verspätet als B, so wird der Lokführer von B eine Warnung am Vorsignal sehen und sein Tempo drosseln bis ihm wieder freie Fahrt angezeigt wird – er baut also zusätzliche Verspätung auf. Reicht dies nicht aus, so kommt es zum Signalhalt von B. Da danach aus dem Stillstand wieder auf Reisegeschwindigkeit beschleunigt werden muss, ist der Zeitverlust erheblich. Ist die Verspätung von A sehr viel grösser als jene von B, kommt es (von selbst oder durch Eingriff eines Fahrdienstleiters) zu einer Änderung der Reihenfolge: B verkehrt nun vor A und erhält freie Fahrt, verspätet sich also nicht zusätzlich. Rechts daneben ist in blau die resultierende Verspätung von A zu sehen: Solange die Reihenfolge beibehalten wird, ergeben sich keine zusätzlichen Verzögerungen. Bei einem Tausch der Reihenfolge kann es zu einem Signalhalt von A kommen. Bei noch grösserer Initialverspätung ist die Konfliktstelle bereits wieder frei, wenn A eintrifft.

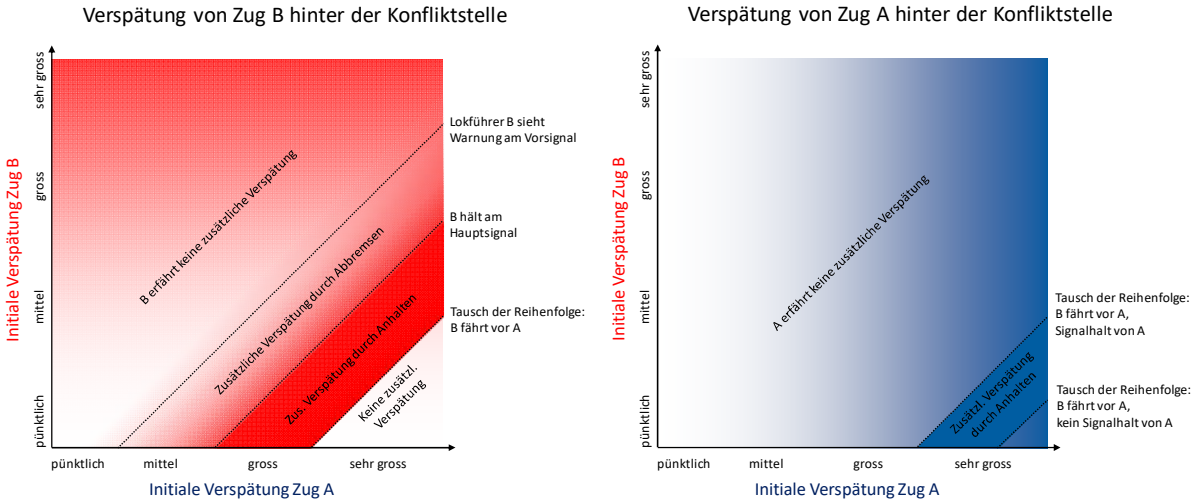


Abbildung 12: Auswirkungen der Initialverspätung zweier Züge bei einem Fahrwegkonflikt

- Die Kreuzung von Zügen auf eingleisigen Strecken (Einspurbetrieb) ist ein Spezialfall eines solchen Fahrwegkonflikts: Wenn sich entgegengesetzt fahrende Züge nur an Bahnhöfen oder Kreuzungsstellen begegnen dürfen, übertragen sich Verspätungen ebenfalls nach den vorstehend beschrie-

benen Mechanismen. Eine Reihenfolgeänderung («Kreuzungsverlegung») kann dabei je nach Länge des Streckenabschnitts zu grossen Verzögerungen des Gegenzugs führen.

- Sofern es die Gleisanlage zulässt, kann auf Verspätungssituationen auch durch Anpassung des Fahrwegs reagiert werden. Durch rechtzeitige Lenkung der Züge auf andere Gleise können zum Beispiel Überschneidungen im Fahrweg vermieden oder Überholungen realisiert werden. Auch dies hat jedoch Auswirkungen auf die Fahrzeit der Züge, da für die Fahrt über ablenkende Weichen die Geschwindigkeit reduziert werden muss.
- Der Schweizer Bahnverkehr wird in mehreren Betriebszentralen von Disponenten überwacht, die auf Planabweichungen entweder direkt oder durch Anweisung an die Fahrdienstleiter reagieren können. Das Brechen von Anschlüssen und Umläufen, die Änderung von Zugs-Reihenfolgen und Kreuzungsstellen sowie die Anpassung des Fahrwegs wurden bereits erwähnt. Weitere Massnahmen umfassen das Aufbieten von Ersatzzügen, die Umleitung auf andere Strecken oder den vollständigen oder abschnittsweisen Ausfall von Zügen zur Vermeidung von Überlastungen.[45] Solche Eingriffe sind grundsätzlich gut geeignet, um Verspätungen entweder abzubauen oder deren Fortpflanzung zu verhindern. Ihre Wirksamkeit hängt aber wesentlich davon ab, ob problematische Reaktionen rechtzeitig erkannt und in kurzer Zeit zielführende Entscheide gefällt werden können.[46] Dabei können Rückkopplungen auftreten: Die Prognose von Verspätungen beeinflusst die Entscheide der Disponenten. Im guten Fall wird sie zur selbst*zerstörenden*, im schlechten zur selbst*erfüllenden* Prophezeiung.

Die Zusammenhänge sind zweifellos komplex. Abschnittsweise können lineare Zusammenhänge vermutet werden (z.B. Weitergabe der Verspätung an einen unmittelbar folgenden Zug; Abwarten von Anschlüssen, Kreuzungen und Umläufen). Dazwischen treten aber auch Stellen auf, an denen der Funktionsverlauf nicht differenzierbar oder sogar unstetig ist (z.B. bei Anschlussbrüchen und Reihenfolgeänderungen). Viele Zusammenhänge sind von externen Faktoren abhängig, die sich kaum vorhersagen lassen (nicht zuletzt das Verhalten von Menschen und Eichhörnchen[47]). Da sowohl die Urverspätungen als auch verschiedene Wirkmechanismen stochastischer Natur sind, sind deterministische Modelle für die Beschreibung unzureichend.[48]

2.4 Prognose von Verspätungen

Zur Sicherung des Bahnverkehrs werden schon seit langer Zeit Gleisfreimeldeanlagen[49] eingesetzt, die das Befahren eines Gleisabschnitts detektieren und ans Stellwerk übermitteln. Aufgrund der im Stellwerk programmierten Abläufe können die Meldungen den einzelnen Zügen zugeordnet werden. Zum Zwecke der Fernbedienung sind heute schweizweit alle Stellwerke miteinander vernetzt. Über dieses Netzwerk wird jedes Mal, wenn ein Zug ein Hauptsignal passiert, ein «Zugnummern-Telegramm» mit Positionsangabe versendet. Auf diese Weise besteht jederzeit ein aktuelles und detailliertes Bild der Betriebslage. Durch Abgleich mit dem Fahrplan können Verspätungen ermittelt werden.

[45] Vgl. Ullius (2005), S. 36.

[46] Unter dem Stichwort «Rescheduling» wird seit vielen Jahren untersucht, inwieweit sich Kapazität und Stabilität dadurch erhöhen lassen, dass Fahrpläne automatisiert und in Echtzeit neu berechnet werden, um auf die aktuelle Betriebslage zu reagieren. Vgl. Lüthi et. al. (2007), Lüthi (2009) und Kecman / Corman / D'Ariano / Goverde (2013).

[47] Eine der grössten Störungen der letzten Jahre bei der Deutschen Bahn wurde durch ein Eichhörnchen verursacht, das beim Beklettern einer Oberleitung einen Kurzschluss ausgelöst hatte. Die Störung wirkte sich fast einen Tag lang landesweit aus. Vgl. Dittes (2010), S.124.

[48] Für einen Versuch, Verspätungen und deren Wechselwirkungen mit Hilfe von Verteilungsfunktionen und geeigneten Faltungen zu modellieren, siehe Büker (2010).

[49] Verbreitet sind zwei Formen der Realisierung: Entweder zählen «Achszähler» die die in einen Abschnitt einfahrenden und ausfahrenden Achsen - ist der Saldo 0, ist der Abschnitt frei. Bei Gleisstromkreisen wird eine elektrische Spannung zwischen linker und rechter Schiene angelegt. Befindet sich ein Fahrzeug auf dem Gleis, so fliesst ein Strom durch die Achsen und führt zu einem Spannungsabfall. Vgl. Pachl (2016), S. 66f.

Für die Prognose der Verspätungsentwicklung sind heute verschiedene Verfahren im Einsatz. Im einfachsten Fall werden bestehende Prognosen unverändert fortgeschrieben: Hat eine Fahrt derzeit x Minuten Verspätung, so wird angenommen, dass sie diese Verspätung bis zur Endhaltestelle weder abbauen noch steigern wird. Die Prognose für alle Folgehaltestellen lautet somit ebenfalls x Minuten. Dieses Verfahren ist häufig bei Bussen und Trams anzutreffen.[50] Im Bahnverkehr werden heute meist die aus dem Fahrplan bekannten Reservezeiten verwendet, um den möglichen Abbau von Verspätungen berücksichtigen zu können.[51] Teilweise werden dort auch kurzfristig bekannte Geschwindigkeitseinschränkungen eingerechnet.[52] Eine wesentliche Verbesserung hat die SBB mit ihrem 2009 eingeführten Dispositionssystem RCS eingeführt. Damit werden erstmals auch Abhängigkeiten zwischen mehreren Zugfahrten bei der Prognose berücksichtigt: In einem Event-Constraint-Modell werden Ankunfts- und Abfahrts-Ereignisse als Knoten eines Graphs modelliert, die dazwischenliegenden Mindestdauern bilden die Kanten (vgl. Abbildung 13) und werden aus dem Fahrplan abgeleitet. Der sich so ergebende gerichtete, azyklische Graph (DAG) kann effizient verarbeitet werden.[53] Es wird schweizweit für die Echtzeitprognose im Normalspurnetz[54] eingesetzt und seit einigen Jahren auch bei anderen europäischen Bahnen eingeführt.

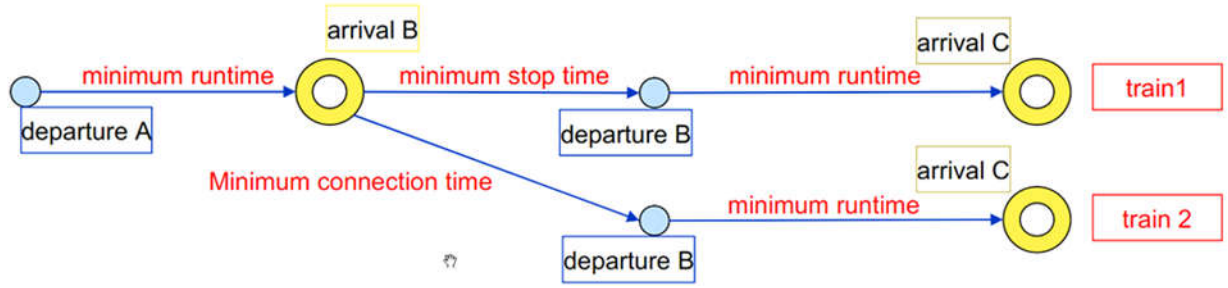


Abbildung 13: Event-Constraint-Graph zur Abbildung von Abhängigkeiten zwischen zwei Zügen in RCS[55]

Keines der beschriebenen Verfahren stützt sich auf empirische Beobachtungen zur tatsächlichen Entwicklung von Verspätungen in der Vergangenheit. Ein solcher Ansatz wurde nach meiner Kenntnis bisher in keinem produktiv und in Echtzeit betriebenen Prognosesystem umgesetzt.

Im Bereich von Forschung, Prototypen und Fahrplananalysen sind mir folgende Arbeiten bekannt:

- Carla Conte hat in ihrer Dissertation Verspätungsdaten der Deutschen Bahn verwendet, um Zusammenhänge bei der Fortpflanzung von Verspätungen zu identifizieren, die sich auf Engpässe in der Infrastruktur und daraus resultierende Fahrweg-Konflikte zurückführen lassen. Ereignisse und dazwischenliegende Aktivitäten werden auch hier als Graphen modelliert. Die daraus abgeleiteten Regressions-Modelle werden mit Methoden der stochastischen Analyse untersucht, um «kritische» Kanten zu identifizieren, welche Aufschluss auf Kapazitätsrestriktionen liefern.[56]
- Im Umfeld der SBB wurde 2009 ein Verfahren entwickelt, um auf effiziente Weise Abhängigkeiten zwischen Zügen in einer grossen Menge von historischen Daten zu finden. Untersucht wurden Anschluss- und Fahrwegkonflikte, die aufgrund ihrer charakteristischen «Pattern» in den Daten ge-

[50] Vgl. Scholz (2012), S. 280f.

[51] Dabei werden die Reserven ggf. nur teilweise in Anschlag gebracht, um der Tatsache Rechnung zu tragen, dass eine «Fahrt unter Idealbedingungen» unrealistisch ist.

[52] Vgl. Büker (2012), S. 34f.

[53] Vgl. Dolder / Krista / Völcker (2009), Büker (2012), S. 35, SBB (2017), SBB (2017b).

[54] Die Schweizer Schmalspurbahnen verwenden meist andere Prognosesysteme, die mit einfacheren Verfahren arbeiten. Der Verkehr der Rhätischen Bahn wird mit RCS überwacht, auch dort erfolgt die Prognoserechnung jedoch nach einem anderen Verfahren.

[55] Quelle: Dolder / Krista / Völcker (2009), Folie 9.

[56] Vgl. Conte (2007).

sucht wurden. Das entwickelte Verfahren trifft keine Annahmen zur Verteilung der verwendeten Variablen.[57]

- Eine ähnliche Fragestellung wurde bei der Belgischen Bahn untersucht: Durch die Suche nach Frequent Itemsets im Datenbestand von INFRABEL wurden Verspätungsübertragungen zwischen Zügen identifiziert.[58]
- Ein Team der Universität Halle hat ein stochastisches Modell für Zugsverspätungen entwickelt. Fahrt- und Haltezeiten wurden dabei als Wahrscheinlichkeitsverteilungen modelliert. Die Modelle wurden zumeist mit artifiziellen Werten getestet. Es kamen aber auch reale Verspätungsdaten von 2 Verkehrstagen der Deutschen Bahn zur Anwendung.[59]
- Pavle Kecman hat in seiner Dissertation historische Daten der niederländischen Bahnen verwendet, um ein Prognosemodell für Verspätungen zu erstellen. Mit einem Process Mining-Ansatz konnte er konfliktbereinigte Fahrzeiten und Haltezeiten in den Bahnhöfen abschätzen. Dabei wurden sowohl Regressionsverfahren als auch Entscheidungsbäume eingesetzt. Unter Verwendung dieser Werte wurde ein Echtzeitprognose-Verfahren implementiert, das auf einem gerichteten azyklischen Graphen basiert. In zwei Fallstudien wurde die Eignung des Verfahrens für ein Echtzeit-Rescheduling untersucht.[60]
- Ren Wang und Daniel B. Work haben in den USA ein Regressionsmodell für Zugverspätungen erstellt und an einem sehr umfangreichen Datensatz der AMTRAK getestet. Das Modell ist vergleichsweise einfach und unterstellt lineare Abhängigkeiten zwischen den Zügen.[61]
- Ein Forschungsteam an der Universität Genua hat Prognosemodelle für Zugfahrten in Italien entwickelt. Die verwendeten Regressionsmodelle wurden mit historischen Daten von etwa 1000 Zügen trainiert und ihre Prognosequalität mit jener der italienischen Bahnen (RFI) verglichen.[62]

Alle genannten Autoren waren darauf angewiesen, dass ihnen die benötigten Daten von den Bahnunternehmen zur Verfügung gestellt wurden – «Open Data» wurde nicht verwendet. Die Arbeit von Pavle Kecman ist die Einzige, in der Echtzeitprognosen erstellt wurden. Soweit ich es überblicke, ist in keinem Fall versucht worden, das Überschätzen von Verspätungen («alpha-Fehler») zu vermeiden. Bei den meisten Arbeiten (u.a. bei jenen mit Event-Graphen) sind umfangreiche Annahmen über den strukturellen Zusammenhang der Variablen in die Modelle eingeflossen.

2.5 Untersuchungsszenario: Ankunft Bahnhof Bern

Für meine eigene Untersuchung habe ich folgendes Szenario gewählt:

> *Vorhergesagt werden sollen die Ankunftsverspätungen aller im Takt verkehrenden Normalspur-Züge am Bahnhof Bern.*

Einen Überblick über den Verlauf der Zuglinien rund um Bern gibt die Netzgrafik in Abbildung 14. Dort sind alle grösseren Bahnhöfe als Rechtecke dargestellt, kleinere Halte als Punkte. Teilweise werden mehrere Haltestellen zu einem kleinen Kreis zusammengefasst; die Anzahl ist dann daneben angeschrieben. Intercity sind rot, Interregio, Regionalexpress und Schnell-S-Bahnen blau, Regio-Züge sowie S-Bahnen sind schwarz dargestellt. Grün wird für Züge verwendet, die nur an Werktagen verkeh-

[57] Vgl. Flier / Gelashvili / Graffagnino / Nunkesser (2009).

[58] Vgl. Cule et. al (2011). Ein indisches Plagiat dieses Beitrags wurde 2017 im «SSRG International Journal of Computer Science and Engineering» veröffentlicht und ist leider noch immer online zu finden unter http://www.internationaljournalssrg.org/IJCSE/2017/Special-Issues/ICEEMST/IJCSE-ICEEMST-P103.pdf. **Pfui!**

[59] Vgl. Berger / Gebhardt / Müller-Hannemann / Lemnian (2011).

[60] Vgl. Kecman (2014). Ein Vorläufer dieser Arbeit ist beschrieben in Hansen / Goverde / van der Meer (2010).

[61] Vgl. Wang / Work (2015).

[62] Vgl. Oneto / Fumeo / Clerico / Canepa / Papa / Dambra / Mazzino / Anguita (2017).

ren. Doppellinien beschreiben Halbstunden-, Vierfachlinien beschreiben Viertelstundentakte. Die bei den Bahnhöfen innen angeschriebene Zahl gibt die Ankunftsminute der jeweiligen Zugslinie an, die Äussere bezeichnet die Abfahrtsminute.

Lesebeispiele: Ein Regioexpress (blau) fährt stündlich um :49 in Kerzers ab und kommt um :07 in Bern an. Der Gegenzug fährt um :53 in Bern ab und kommt um :09 ein. Diese Züge halte nicht in Brünnen und Gümmenen (keine Zeitangaben dort). Zwischen Bern Europaplatz (Abfahrt :09 und :39) und Schwarzenburg (Ankunft :41 und :11) verkehrt halbstündlich eine S-Bahn mit 8 Unterwegshalten.

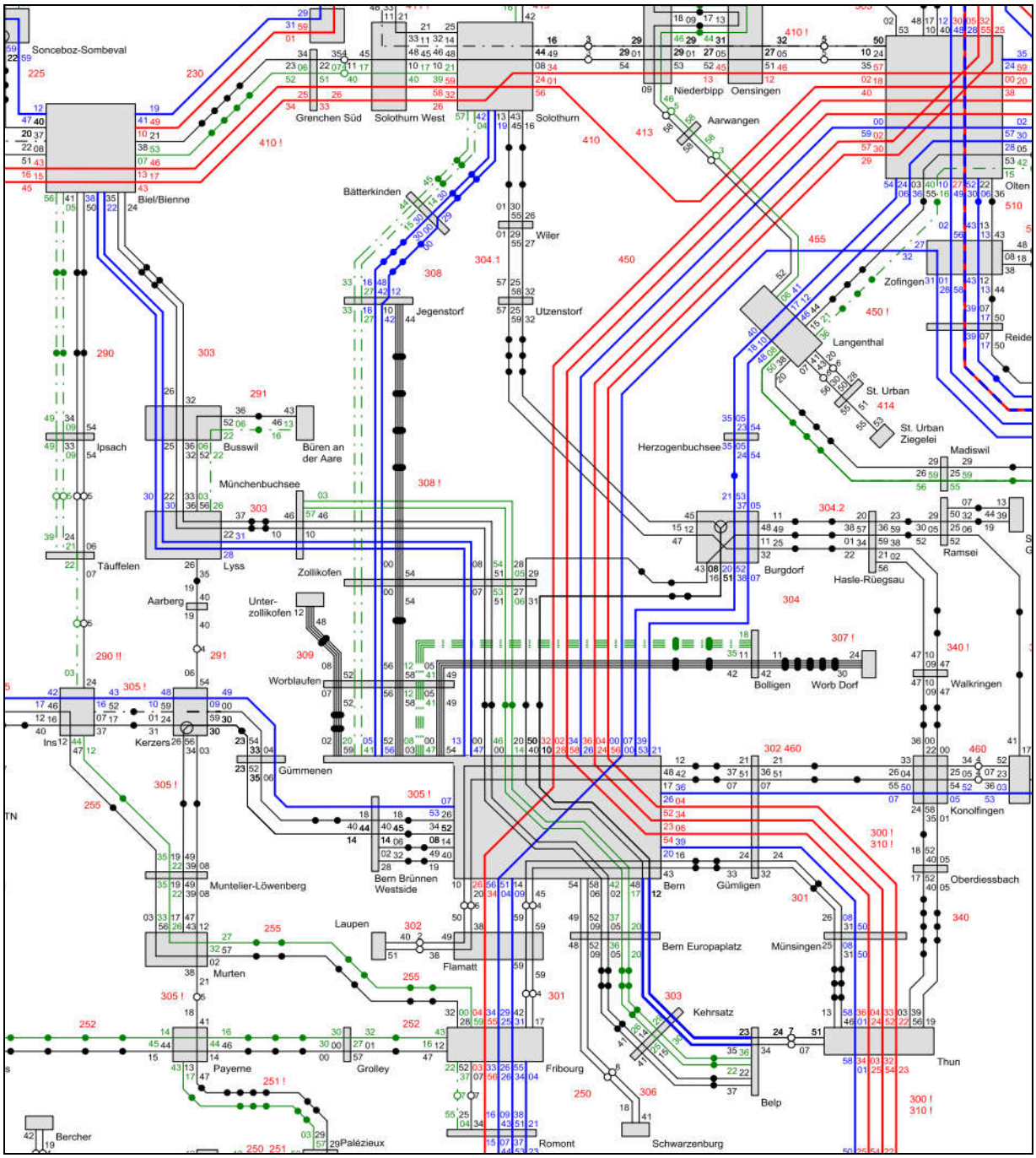


Abbildung 14: Netzgrafik mit Zugslinien im Umfeld von Bern (Fahrplanperiode 2018)[63]

[63] Quelle: Ausschnitt aus sma (2017).

Aus folgenden Gründen fiel meine Wahl auf Bern:

- Mit 45 Ankünften pro Stunde und Zügen aller Kategorien (S-Bahn, RE, IR, IC sowie Internationaler Fernverkehr) weist das Szenario eine angemessene Grösse und Heterogenität auf.
- Bern ist einer der sieben Fernverkehrs-Knoten der Schweiz mit Ankünften und Abfahrten im «Rendezvous-System». Dies führt zu einer starken Bündelung im Zulauf, was es ermöglicht, den Einfluss von Zugfolge- und Fahrweg-Abhängigkeiten zu berücksichtigen. Bekannte «Nadelöhre» existieren auf beiden Seiten des Bahnhofs (Weyermannshaus im Westen und Wankdorf im Osten). Auf den Strecken Olten-Bern und Thun-Bern gibt es zudem enge Zugfolgen.
- Es existieren Direktverbindungen von den Fernverkehrs-Knoten Olten, Biel und Lausanne, sowie zu zahlreichen Knoten des Regionalverkehrs. Kommt es dort zu Abfahrtsverspätungen aufgrund von Anschlussbeziehungen, so sollte sich dies bei der Ankunftsprognose für Bern bemerkbar machen.
- Da öffentliche Daten nur für die Schweiz verfügbar sind, können aus dem Ausland kommende Züge erst ab der Landesgrenze berücksichtigt werden. Bern ist mehr als 60 Fahrminuten von allen Grenzübergängen entfernt, so dass eine gute «Rundum-Sicht» gewährleistet ist.
- Im Zulauf nach Bern liegen mehrere Strecken mit Einspur-Abschnitten (Richtung Kerzers, Schwarzenburg, Belp und Konolfingen), so dass auch Sekundärverspätungen durch Zugskreuzungen beobachtet werden können.
- In Bern verkehren Züge von drei Normalspur-Bahnen (SBB, BLS und TPF). Dies trägt ebenfalls zur Heterogenität bei und macht diese Arbeit zudem für ein breiteres Publikum interessant.
- In Bern hat es beim Fahrplanwechsel 2017 / 2018 nur wenige Änderungen gegeben (siehe Abbildung 15). Prognosemodelle, die mit Daten aus 2017 trainiert wurden, können – mit Einschränkungen – eventuell auch für 2018 verwendet werden.

FP2017	FP2018
Zofingen (ab 32) - Bern (an 00)	Zofingen (ab 32) - Bern (an 00)
Bern Wankdorf (ab 54) - Bern (an 00)	Bern Wankdorf (ab 54) - Bern (an 00)
Kerzers (ab 49) - Bern (an 07)	Kerzers (ab 49) - Bern (an 07)
Bern Wankdorf (ab 05) - Bern (an 10)	Bern Wankdorf (ab 05) - Bern (an 10)
Bern Europaplatz (ab 03) - Bern (an 10)	Bern Europaplatz (ab 04) - Bern (an 10)
Bern Europaplatz (ab 07) - Bern (an 12)	Bern Europaplatz (ab 07) - Bern (an 12)
Bern Wankdorf (ab 08) - Bern (an 13)	Bern Wankdorf (ab 08) - Bern (an 13)
Bern Wankdorf (ab 09) - Bern (an 14)	Bern Wankdorf (ab 09) - Bern (an 14)
Bern Stöckacker (ab 09) - Bern (an 14)	Bern Stöckacker (ab 09) - Bern (an 14)
Bern Europaplatz (ab 09) - Bern (an 14)	Bern Europaplatz (ab 09) - Bern (an 14)
Lyss (ab 01) - Bern (an 17)	Lyss (ab 01) - Bern (an 17)
Bern Wankdorf (ab 12) - Bern (an 17)	Bern Wankdorf (ab 12) - Bern (an 17)
Belp (ab 05) - Bern (an 18)	Belp (ab 05) - Bern (an 18)
Münsingen (ab 08) - Bern (an 20)	Münsingen (ab 08) - Bern (an 20)
Burgdorf (ab 07) - Bern (an 21)	Burgdorf (ab 07) - Bern (an 21)
Thun (ab 04) - Bern (an 23)	Thun (ab 04) - Bern (an 24)
Olten (ab 57) - Bern (an 24)	Olten (ab 57) - Bern (an 24)
Bern Europaplatz (ab 19) - Bern (an 24)	Bern Europaplatz (ab 19) - Bern (an 24)
Olten (ab 59) - Bern (an 26)	Olten (ab 59) - Bern (an 26)
Konolfingen (ab 07) - Bern (an 26)	Konolfingen (ab 07) - Bern (an 26)
Fribourg/Freiburg (ab 04) - Bern (an 26)	Fribourg/Freiburg (ab 04) - Bern (an 26)
Bern Bümpliz Nord (ab 20) - Bern (an 26)	Bern Bümpliz Nord (ab 20) - Bern (an 26)
Zürich HB (ab 32) - Bern (an 28)	Zürich HB (ab 32) - Bern (an 28)
Bern Europaplatz (ab 22) - Bern (an 28)	Bern Europaplatz (ab 22) - Bern (an 28)

FP2017	FP2018
Bern Wankdorf (ab 24) - Bern (an 30)	Bern Wankdorf (ab 24) - Bern (an 30)
Bern Wankdorf (ab 35) - Bern (an 40)	Bern Wankdorf (ab 35) - Bern (an 40)
Bern Stöckacker (ab 35) - Bern (an 40)	Bern Stöckacker (ab 35) - Bern (an 40)
Bern Europaplatz (ab 33) - Bern (an 40)	Bern Europaplatz (ab 34) - Bern (an 40)
Bern Europaplatz (ab 37) - Bern (an 42)	Bern Europaplatz (ab 37) - Bern (an 42)
Bern Wankdorf (ab 38) - Bern (an 43)	Bern Wankdorf (ab 38) - Bern (an 43)
Bern Wankdorf (ab 39) - Bern (an 44)	Bern Wankdorf (ab 39) - Bern (an 44)
Bern Europaplatz (ab 39) - Bern (an 44)	Bern Europaplatz (ab 39) - Bern (an 44)
Lyss (ab 31) - Bern (an 47)	Lyss (ab 31) - Bern (an 47)
Bern Wankdorf (ab 43) - Bern (an 48)	Bern Wankdorf (ab 43) - Bern (an 48)
Belp (ab 35) - Bern (an 48)	Belp (ab 35) - Bern (an 48)
Fribourg/Freiburg (ab 29) - Bern (an 51)	Fribourg/Freiburg (ab 29) - Bern (an 51)
Thun (ab 33) - Bern (an 52)	Thun (ab 33) - Bern (an 52)
Bern Bümpliz Nord (ab 47) - Bern (an 52)	Bern Bümpliz Nord (ab 47) - Bern (an 52)
Burgdorf (ab 38) - Bern (an 53)	Burgdorf (ab 38) - Bern (an 53)
Thun (ab 36) - Bern (an 54)	Thun (ab 36) - Bern (an 54)
Bern Europaplatz (ab 49) - Bern (an 54)	Bern Europaplatz (ab 49) - Bern (an 54)
Olten (ab 29) - Bern (an 56)	Olten (ab 29) - Bern (an 56)
Fribourg/Freiburg (ab 34) - Bern (an 56)	Fribourg/Freiburg (ab 34) - Bern (an 56)
Zürich HB (ab 02) - Bern (an 58)	Zürich HB (ab 02) - Bern (an 58)
Bern Europaplatz (ab 52) - Bern (an 58)	Bern Europaplatz (ab 52) - Bern (an 58)

Abbildung 15: Ankünfte vertakteter Züge in Bern (Normalspur) in den Fahrplanperioden 2017 und 2018

- Da nur Daten zum Personenverkehr verfügbar sind, kann der Einfluss von Güterzügen nicht in der Prognose berücksichtigt werden. Es scheint daher ratsam, diese «Störgrösse» zu vermeiden. Im Bereich des Bahnhofs Bern gibt es vergleichsweise wenig Güterverkehr.
- Geographie, Bahnanlagen und Fahrplan sind mir im Raum Bern besser bekannt als an vielen anderen Orten, was die Arbeit mit diesem Szenario erleichtert.

Bedauerlich ist, dass die im Tiefbahnhof Bern verkehrenden Züge des Regionalverkehr Bern-Solothurn (RBS) nicht einbezogen werden können, weil hierzu bisher keine Daten veröffentlicht werden.

Ich habe mich auf die Prognose von Ankunfts-Verspätungen beschränkt, weil dies im Szenario Bern einfacher erschien: Aufgrund der Rendezvous-Situation haben viele Züge dort relativ lange Aufenthaltszeiten (z.B. IC Thun-Bern-Olten: Ankunft :52, Weiterfahrt :04). Entsprechend gross sind die Reserven, mit denen Ankunftsverspätungen kompensiert werden können. Gleichzeitig gibt es während eines Knoten-Halts viele externe Faktoren[64], die eine verzögerte Abfahrt herbeiführen können, ohne dass sich dafür Prädiktoren in den verfügbaren Daten finden. Aus diesen Gründen vermute ich, dass Abfahrtsverspätungen in Bern schwieriger zu prognostizieren sind als Ankunftsverspätungen. Ob diese Annahme richtig ist, habe ich bisher nicht untersucht. Es sollte aber möglich sein, denn das in dieser Arbeit entwickelte Prognoseverfahren ist grundsätzlich für Ankunft und Abfahrt einsetzbar.

Ich schliesse Verfrühungen von der aus, weil diese aus Kundensicht weniger interessant erscheinen. Verfrühungen werden von mir mit einer Verspätung von 0 Sekunden gleichgesetzt – sowohl bei den selbst durchgeführten Prognosen als auch beim Vergleich mit den Referenzsystemen der Bahnen.

Konform zur Zielsetzung werden zwei Arten von Vorhersagen betrachtet:

a) Der Umfang einer Verspätung in Minuten.

b) Das Überschreiten einer Mindestgrenze (1, 2, 3, 5, 10 oder 15 min).

In beiden Fällen wird untersucht, mit welchem zeitlichen Vorlauf korrekte Prognosen möglich sind.

In der Anfangsphase der Bearbeitung hatte ich die Absicht, neben Bern auch noch ein zweites Szenario zu untersuchen. Dafür wäre insbesondere die Rhätische Bahn (RhB) interessant gewesen (z.B. Betrachtung der Knoten Chur, Landquart, Klosters, Davos, Samedan und St. Moritz):

- Die Rhätische Bahn hat – u.a. aufgrund von Geographie und Tourismus – eine vergleichsweise geringe Pünktlichkeit (vgl. Abbildung 16). Es gibt dort also mehr vorherzusagen.

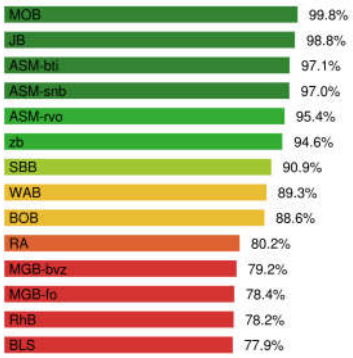


Abbildung 16: Pünktlichkeit von Schweizer Regionalzügen im Unternehmensvergleich[65]

- Die RhB setzt für die eigene Prognose ein weniger ausgefeiltes System ein als RCS. Der «Benchmark» wäre dort vermutlich einfacher zu schlagen.
- Es existieren sehr viele Einspurabschnitte und Kreuzungsstellen, was interessante Prognose-Konstellationen erwarten lässt.

Aufgrund der überraschend guten Ergebnisse aus ersten Untersuchungen mit Berner Daten habe ich im Verlauf der Bearbeitung entschieden, mich auf dieses eine Szenario zu konzentrieren.

[64] Insbesondere steht der Kunde an den Bahnhöfen im Mittelpunkt – und damit leider manchmal auch im Weg. Zu nennen sind weiterhin Personalwechsel, Rangiermanöver zum Aufstellen / Wegstellen / Verstärken / Schwächen von Zügen und das Hauptrisiko jedes Halts: die Türstörung.

[65] Zeitraum: 10. Februar 2017 bis 9. März 2018 (= 1 Jahr). Angegeben ist der Anteil von Ankünften mit weniger als 180 Sekunden Verspätung. Quelle: www.puenktlichkeit.ch.

3 IT-Architektur der Lösung

In allen Bearbeitungsschritten dieser Arbeit wird Informationstechnik benötigt. Dieses Kapitel gibt einen Überblick über die gewählte IT-Architektur und die verwendeten Technologien. Es handelt sich dabei um Erweiterungen der bestehenden Plattform von www.puenktlichkeit.ch[66]: Die dort verwendete Architektur hat sich grundsätzlich bewährt. Zudem lassen sich bei Datenbezug und Datenhaltung, aber auch bei der Visualisierung der Ergebnisse erhebliche Synergien nutzen.

3.1 Bausteinsicht: Komponenten der Lösung

Für die verfolgte Zielsetzung werden folgende Komponenten benötigt:

- Eine **Datenquelle** für historische Verspätungsdaten, mit denen die Prognosemodelle trainiert werden sollen,
- eine **Datenquelle** für aktuelle Daten zur Betriebslage, auf deren Grundlage Echtzeit-Prognosen durchgeführt werden können,
- ein **Prognoseverfahren** zur Vorhersage von Zugsverspätungen,
- diverse **Skripte** für das Trainieren, Validieren, Testen und Anwenden des Verfahrens,
- ein **automatisiertes Skript** zur Erstellung von Echtzeitprognosen,
- einen **Scheduler** zum zeitgesteuerten Anstoss von Datenbezug und Echtzeit-Prognose,
- ein **Datenbanksystem** zur Sammlung und Bereitstellung der erforderlichen Daten sowie zur Speicherung von Prognosen,
- **Applikationen** zur Visualisierung von Verspätungsprognosen und zur Auswertung der erzielten Prognosequalität,
- **Applikations-Server und Plattform** zur Ausführung von Skripten und Applikationen,
- ein **Web-Server** zur Bereitstellung und zum interaktiven Abruf der Auswertungen im Internet.

Die Datenquellen sind Gegenstand von Kapitel 4. Das Prognoseverfahren bildet den Kern dieser Arbeit. Seine Entwicklung und Anwendung wird in Kapitel 5 und 6 beschrieben. Die von mir entwickelten Applikationen zur Anzeige und Analyse der Ergebnisse beschreibe ich in Kapitel 7.

Für Entwicklung, Pflege und Betrieb der Lösung sind erforderlich:

- Eine **Entwicklungsumgebung** mit Editor, Compiler / Interpreter, Debugger etc.,
- geeignete **Libraries**, um gängige Funktionsbausteine nicht selbst entwickeln zu müssen (z.B. Datentransformation, DB-Zugriff, Diagrammerstellung, Machine Learning-Methoden),
- eine **Code- und Versionsverwaltung**,
- **Administrations- und Monitoring-Werkzeuge** für Server und Datenbanksystem,
- eine **Logging-Komponente** zwecks Kontrolle der Systemausführung.

3.2 Anbieter- und Technologieauswahl

Die gewählten Technologien und Anbieter entsprechen jenen, die ich schon bei den Semesterarbeiten und für mein Projekt www.puenktlichkeit.ch genutzt habe. Auf diese Weise konnten Einarbeitungs- und Installtionsaufwände klein gehalten und Risiken reduziert werden.

[66] Vgl. Gutweniger (2017).

Folgende Technologien werden eingesetzt:

- R als Programmiersprache,
- RStudio als Entwicklungsumgebung,
- Shiny als Web- und Applikationsserver,
- MySQL als Datenbankmanagement-System,
- MySQL-Workbench für die Administration der Datenbank,
- Windows 10 und Ubuntu 16 als Betriebssystem,
- Cron als Scheduler,
- Git als Code- und Versionsverwaltung.

Folgende R-Libraries werden verwendet:

- `httr` für Webservice-Abfragen,
- `jsonlite` und `xml2` für Parsing und Formatumwandlung,
- `RMySQL` für die Anbindung der MySQL-Datenbank,
- `data.table` für die Organisation der Daten (als schnellere Alternative zu Dataframes),
- `stringr` für String-Operationen,
- `lubridate` und `scales` für den Umgang mit Uhrzeiten und Prozentwerten,
- `caret` und `rpart` für Machine Learning-Algorithmen,
- `shiny` und `shinyjs` für die Web-Applikationen,
- `ggplot2`, `rpart.plot` und `leaflet` zur Visualisierung von Daten, Modellen und Landkarten,
- `DT` zur Umsetzung interaktiver Tabellen,
- `doParallel` zur parallelisierten Code-Ausführung,
- `futile.logger` für das Logging.

3.3 Verteilungssicht

Entwicklungsumgebung, Web- und Appplikationsserver sind auf einer Linux-Instanz bei Amazon Web Services (AWS) installiert. Die Datenbank wird ebenfalls als Cloud-Service von AWS bezogen.

Für aufwendige Berechnungen habe ich ausserdem die beiden Linux-Cluster der BFH nutzen dürfen. Hier konnten parallelisierte Berechnungen auf bis zu 88 Cores ausgeführt werden.

Um offline arbeiten zu können, habe ich zusätzlich eine lokale Installation von RStudio verwendet.

Zum Austausch des Quellcodes zwischen AWS, BFH-Cluster und lokalem Rechner dient ein zentrales Git-Repository bei Atlassian Bitbucket. Für den Datenaustausch kommt in erster Linie die Datenbank zum Einsatz: dort werden sowohl die historischen Daten gehalten als auch die Ergebnisse von Simulationen und Echtzeitprognosen gespeichert. Da es wiederholt Schwierigkeiten gab, grössere Datenmengen vom BFH-Cluster in die Datenbank zu schreiben, habe ich Simulationsergebnisse meist vor dem DB-Load per File-Transfer in die AWS-Umgebung übertragen.

Die erzeugten Prognosemodelle werden generell in Dateien verwaltet. Diese liegen aufgrund ihrer Grösse nicht im Repository, sondern werden per File-Transfer ausgetauscht.

Abbildung 17gibt einen Überblick über die verwendeten Komponenten und Umgebungen und zeigt die wichtigsten Datenflüsse.

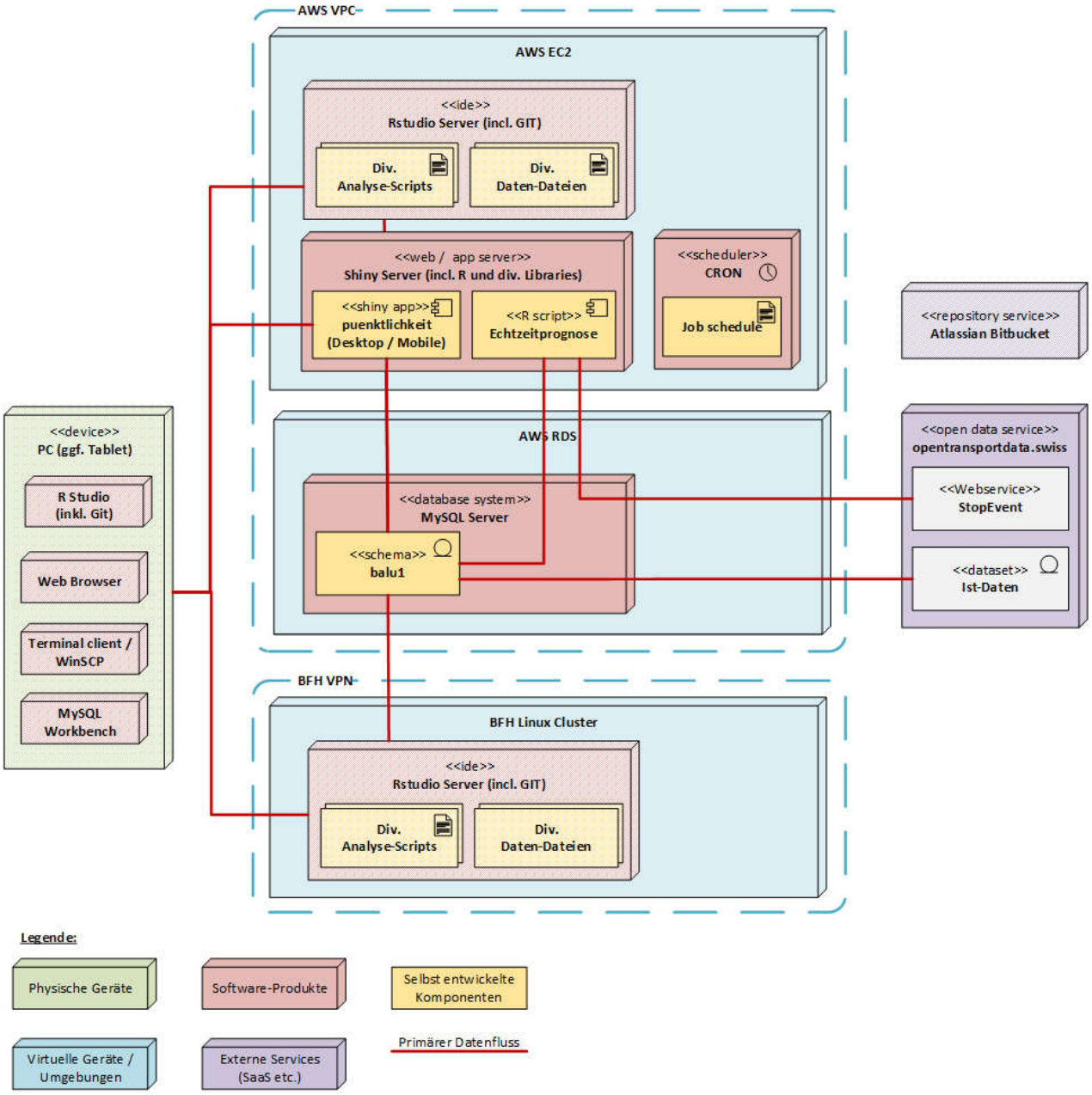


Abbildung 17: Architekturübersicht (Deployment-Diagramm)

4 Datenquellen und Datenbezug

Im Sinne der „Open Data Strategie“ des Bundesrats hat das Bundesamt für Verkehr die Schweizer öV-Unternehmen vor einiger Zeit beauftragt, gewisse Daten öffentlich bereit zu stellen. Dies soll es branchenfremden Akteuren ermöglichen, neue Systeme, Apps und Statistiken zum öffentlichen Verkehr zu entwickeln.[67] Erfahrungen im Ausland (z.B. in London) haben gezeigt, dass ein solcher Open Data-Ansatz hohe Innovationskraft freisetzen und zugleich die Verkehrsunternehmen von der Entwicklung eigener Informationssysteme entlasten kann.[68]

Seit Dezember 2016 steht die beauftragte Plattform www.opentransportdata.swiss bereit. Derzeit (März 2018) enthält Sie Daten von über 70 der ca. 250 Transportunternehmen der Schweiz, darunter alle Normalspurbahnen, die meisten Schmalspurbahnen, ein Teil des Postauto-Netzes und die Stadtverkehre von Basel, Bern, Luzern, Winterthur und Zürich. Zu den veröffentlichten Daten gehören:

- Die publizierten Fahrpläne der Unternehmen (nicht jedoch: die detaillierteren, innerbetrieblichen «Produktionspläne»),
- Das Verzeichnis aller Haltestellen und sonstigen «Betriebspunkte»,
- Anzahl der Generalabo- und Halbtax-Abonnenten pro Postleitzahl,
- Eine Gegenüberstellung aller geplanten und tatsächlich erfolgten Ankunfts- und Abfahrtszeiten der vergangenen 30 Tage (Historische Daten).
- Echtzeitdaten zu kürzlich erfolgten und in Kürze bevorstehenden Halten.

Für die Zwecke dieser Arbeit sind primär die beiden zuletzt genannten Punkte, historische Daten und Echtzeitdaten, von Interesse. Die folgenden Abschnitte gehen detailliert darauf ein. Abbildung 18 gibt eine Übersicht über die Verwendung dieser Daten und die damit erstellten Ergebnisse. Die einzelnen Schritte werden in den folgenden Abschnitten erläutert.

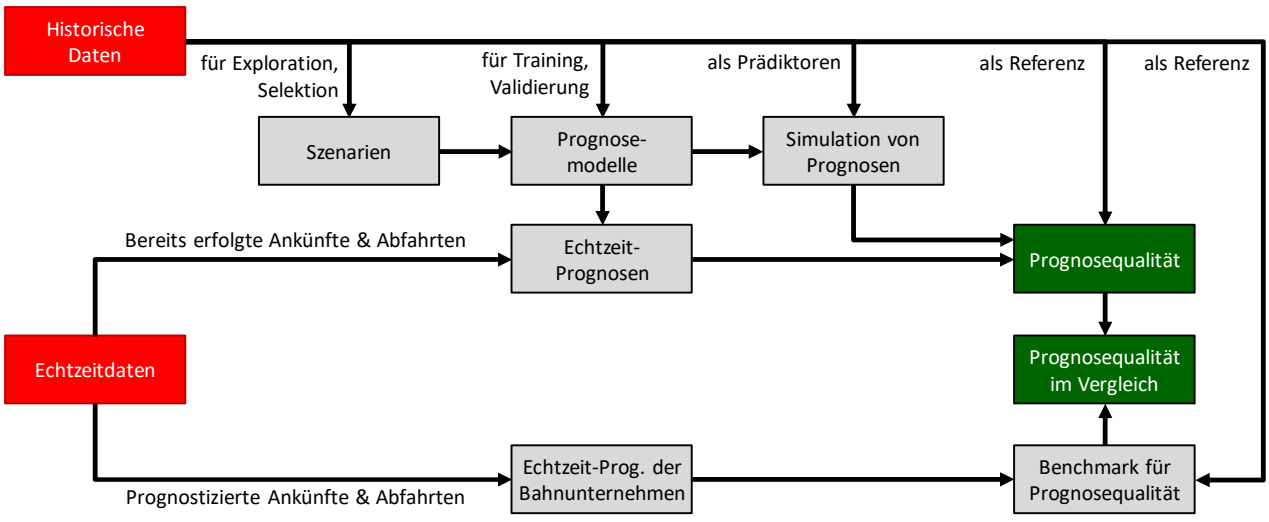


Abbildung 18: Verwendung von historischen Daten und Echtzeitdaten im Rahmen der Arbeit.

[67] Vgl. BAV (2016).

[68] Vgl. Gerny (2016).

4.1 Historische Daten

Historische Daten kamen bereits bei der Szenarienbildung zum Einsatz (vgl. Abschnitt 2.5), denn sie liefern Anhaltspunkte

- zum Verkehrsangebot: Zwischen welchen Haltestellen finden Fahrten statt? Mit welchen Verkehrsmitteln? Wie häufig und wie regelmässig?
- zur Realisierung dieses Angebots: Welchen Fahrtverlauf haben die einzelnen Linien? Wo und wie häufig halten sie? Wo gibt es Umsteigebeziehungen, wo kreuzen Verkehrsmittel?
- zur Pünktlichkeit des Angebots: Wo wurde in der Vergangenheit hohe Pünktlichkeit erzielt, wo gab es viele Verspätungen?
- zur Stabilität des Angebots: Wo kam es zu Veränderungen im Angebot (Fahrplanwechsel)?

Ich habe dafür Daten eines längeren Zeitraums (Dezember 2016 bis Januar 2018) analysiert.

Um für die Szenarien Prognosemodelle zu erstellen, werden ebenfalls historische Daten eines längeren Zeitraums benötigt:

- für das Trainieren der Modelle
- für das Validieren der Modelle

Dies ist Gegenstand von Kapitel 5.

Die so erstellten Modelle können auf eine separate – weder für Training noch für Validierung verwendete – Untermenge der historischen Daten angewendet werden, um die Durchführung von Prognosen im Nachhinein zu simulieren: Welche Prognosen hätte das Prognoseverfahren geliefert, wenn es zu einem vergangenen Zeitpunkt angewendet worden wäre? Dieses Vorgehen wird in der Literatur häufig als «Testen» bezeichnet.[69] Es wird in Abschnitt 6.1 beschrieben.

Im Gegensatz zu den nur minutengenauen Echtzeitdaten der Open Data Plattform sind die historischen Daten sekundengenau. Sie sind somit wesentlich besser geeignet, um die Qualität von Prognosen zu ermitteln. Sie werden daher verwendet zur Bewertung von

- Simulierten Prognosen des hier entwickelten Verfahrens,
- Echtzeit-Prognosen des hier entwickelten Verfahrens,
- Echtzeit-Prognosen der Bahn-Unternehmen.

Da die historischen Daten erst am jeweiligen Folgetag auf der Open Data Plattform verfügbar sind, ist die Bewertung in den letzten beiden Fällen erst mit entsprechender Verzögerung möglich: die Echtzeitprognosen werden zunächst aufgezeichnet. Ihre Qualität wird ermittelt, sobald die Referenzgrössen (tatsächliche Verspätungen) vorliegen.

Die Daten werden auf der Open Data Plattform als CSV-Tabelle «Ist-Daten» bereitgestellt. Für jeden Halt eines Verkehrsmittels ist darin eine Zeile mit den publizierten und tatsächlichen Zeitpunkten von Ankunft und Abfahrt an der Haltestelle aufgeführt.[70] Die Ankunfts- und Abfahrtsverspätung lässt sich daraus einfach durch Differenzbildung ermitteln. Zu beachten ist, dass es sich bei den tatsächlichen

[69] Die Terminologie hierzu ist uneinheitlich. Die verbreitete Unterscheidung zwischen «Training» und «Validierung» lässt häufig offen, inwieweit Erkenntnisse aus der Validierung für die Verbesserung des Verfahrens genutzt werden dürfen (z.B. durch Modifikation der verwendeten Modelle oder Hyperparameter). Verschiedene Autoren betonen, dass die abschliessende Bewertung auf Basis von Daten erfolgen sollte, die in keinerlei Weise bei Erstellung oder Tuning verwendet wurde (vgl. Russel / Norvig (2009), S. 709; Kuhn / Johnson (2013), S. 67). Empfehlenswert ist daher eine Dreiteilung in «Training» (anlernen / «fitten» des Modells), «Validierung» (im Rahmen von Verbesserung / «Tuning» des Modells) und «Test» (zur finalen Bewertung der Modell-Qualität), vgl. Ripley (1996), S. 354. Quellen zitiert nach Brownlee (2017a).

[70] Für eine detaillierte Beschreibung von Datenstruktur und -semantik siehe Open Data Plattform öV Schweiz (2016), Abschnitt «Ist-Daten».

Zeitpunkten meist um genäherte Werte handelt: Die Messpunkte der Leittechnik-Systeme befinden sich in der Regel vor dem effektiven Haltepunkt (z.B. bei der Bahnhofseinfahrt oder an einer Verkehrsampel). Durch Verwendung geeigneter Zu- und Abschläge wird daraus die tatsächliche Zeit geschätzt; entsprechend sind fast alle Einträge als «Prognose» oder «Geschätzt» gekennzeichnet.[71] Da es sich um die besten verfügbaren Daten handelt und die Unternehmen diese auch selbst für ihre Verspätungsprognosen, Kundeninformation und diverse Auswertungen verwenden, wird dieser Umstand toleriert. Es erscheint zudem plausibel, dass die Ungenauigkeiten relativ gering ausfallen.

Für den automatisierten Bezug und die Aufbereitung dieser «Ist-Daten» wird die bestehende Funktionalität von puenktlichkeit.ch verwendet.[72] Von den dabei angelegten Datenbank-Tabellen ist hier vor allem `fct_abschnitt` relevant. In Abbildung 19 ist beispielhaft ein Auszug dieser Tabelle mit Fahrten von Bern Wankdorf nach Bern dargestellt.

BETRIEBSTAG	BPUICA	AB_PLAN	AB_VERSP	BPUICB	AN_PLAN	AN_VERSP	DELTA	BTR_ID	VM_ID	FAHRT_BEZEICHNER
2018-02-15	8516161	07:05:00	465	8507000	07:10:00	416	22	24	28	85:33:16420:001
2018-02-15	8516161	07:08:00	253	8507000	07:13:00	189	-11	24	28	85:33:15124:001
2018-02-15	8516161	07:09:00	155	8507000	07:14:00	79	-18	24	28	85:33:16322:001
2018-02-15	8516161	07:12:00	226	8507000	07:17:00	176	4	24	28	85:33:15224:001
2018-02-15	8516161	07:24:00	150	8507000	07:30:00	93	-2	24	28	85:33:15324:001
2018-02-15	8516161	07:31:00	644	8507000	07:36:00	576	25	24	28	85:33:15024:001
2018-02-15	8516161	07:35:00	204	8507000	07:40:00	150	-2	24	28	85:33:15026:001
2018-02-15	8516161	07:35:00	181	8507000	07:40:00	298	241	24	28	85:33:15422:002

Abbildung 19: Auszug aus der Fakten-Tabelle «fct_abschnitt».

Für jeden Fahrtabschnitt zwischen zwei Halten ist ein Record enthalten mit folgenden Angaben:

- `BETRIEBSTAG`: Kalendertag, dem die Fahrt zugeordnet wurde. Bei Fahrten kurz nach Mitternacht kann dies der Vortag sein.
- `BPUICA`: Haltestelle am Beginn des Fahrtabschnitts. Angegeben ist der internationale Haltestellen-Code, 8516161 steht hier für «Bern Wankdorf».
- `AB_PLAN`: Uhrzeit, zu der die Abfahrt in `BPUICA` gemäss Fahrplan erfolgen sollte.
- `AB_VERSP`: Abfahrts-Verspätung in Sekunden. Im Fall der ersten Tabellenzeile ist der Zug also tatsächlich 465 Sekunden nach 7:05 Uhr, d.h. um 7:12:45 Uhr in Wankdorf abgefahren. Negative Werte stehen für eine verfrühte Abfahrt des Zuges.
- `BPUICB`: Haltestelle am Ende des Fahrtabschnitts. 8507000 steht hier für «Bern» (Hauptbahnhof).
- `AN_PLAN`: Uhrzeit, zu der die Ankunft in `BPUICB` gemäss Fahrplan erfolgen sollte.
- `AN_VERSP`: Ankunfts-Verspätung in Sekunden. Im Fall der ersten Tabellenzeile ist der Zug also tatsächlich 416 Sekunden nach 7:10 Uhr, d.h. um 7:16:56 Uhr in Bern angekommen. Negative Werte stehen für eine verfrühte Ankunft des Zuges.
- `DELTA`: Veränderung der Ankunfts-Verspätung gegenüber dem vorausgehenden Fahrtabschnitt desselben Zugs. Im Fall der ersten Tabellenzeile hat sich die Verspätung bei der Ankunft in Bern also um 22 Sekunden gegenüber jener bei der Ankunft in Bern Wankdorf erhöht.
- `BTR_ID`: Interne ID des Bahnunternehmens. 24 steht hier für die BLS AG.
- `VM_ID`: Interne ID für das Verkehrsmittel. 28 steht hier für S-Bahn.

[71] Für ausführlichere Erläuterungen zur Erhebungsmethodik und den in der Praxis auftretenden Konstellationen siehe Open Data Plattform öV Schweiz (2016), Unterabschnitt «Ist-Daten / Spezielle Effekte und ihre Ursachen».

[72] Die Website puenktlichkeit.ch ist im Rahmen einer Semesterarbeit im CAS Business Intelligence der BFH entstanden und wird seither von mir privat betrieben und weiterentwickelt. Architektur, Datenstrukturen und ETL-Prozesse sind beschrieben in Gutweniger (2017).

- `FAHRT_BEZEICHNER`: Die (innerhalb eines Betriebstages) eindeutige Bezeichnung der Fahrt – so, wie sie von der Open Data Plattform für die «Ist-Daten» geliefert wird. Darin ist unter anderem die Zugnummer (vorletzter Bestandteil) enthalten. Anhand dieses Bezeichners lässt sich der Verlauf einer Fahrt über alle ihre Abschnitte konstruieren.

Jede Tabellenzeile enthält somit Angaben zu zwei Ereignissen: zur Abfahrt an einer Haltestelle und zur Ankunft an der folgenden Haltestelle. Bei der Erstellung und Verwendung von Prognosemodellen werden daher häufig zwei Variablenwerte aus einem Tabelleneintrag abgeleitet.

Jeder Abschnitt wird durch zwei Halte begrenzt: Züge, die ohne Halt durch Bern Wankdorf fahren, sind folglich in Abbildung 19 nicht aufgeführt. Der Intercity aus Romanshorn wäre in der Tabelle stattdessen mit `BPUICA=8503000` (Zürich HB, letzter Halt vor Bern) und `BPUICB=8507000` (Bern) zu finden.

Pro Betriebstag enthält die Tabelle zwischen 520'000 und 830'000 Einträge, wovon etwa die Hälfte auf den Zugsverkehr entfallen. Während auf der Open Data-Plattform nur jeweils die Daten der letzten 30 Tage verfügbar sind, reicht der Datenbestand von puenktlichkeit.ch zurück bis zum 12.11.2016.

4.2 Echtzeitdaten

Die von der Open Data Plattform gelieferten Echtzeitdaten haben zwei unterschiedliche Bedeutungen:

- Sofern sie sich auf Ereignisse (Abfahrten oder Ankünfte) in der Zukunft beziehen, so stellen Sie die aktuelle Prognose der Bahnunternehmen für dieses Ereignis dar. Beispiel: Der Intercity aus Luzern soll gemäss Fahrplan um 19:00 Uhr in Bern eintreffen. Um 18:50 Uhr liefert die Open Data Plattform die Ankunftszeit 19:02 Uhr. Die SBB prognostiziert also eine Verspätung von 2 Minuten.
- Wenn sie sich auf Ereignisse in der Vergangenheit beziehen, so stellen sie den letzten bekannten Prognosestand zu diesem Ereignis dar. Da von den IT-Systemen bis unmittelbar vor dem Eintreten des Ereignisses solche Prognosen berechnet werden (in bestimmten Konstellationen werden sie sogar noch im Nachhinein korrigiert), kann dies als bestes verfügbares Wissen über den tatsächlichen Zeitpunkt aufgefasst werden. Beispiel: Der Intercity aus Luzern sollte gemäss Fahrplan um 18:32 Uhr in Zofingen abgefahren sein. Um 18:50 Uhr liefert die Open Data Plattform die Abfahrtszeit 18:37 Uhr. Offenbar hatte der Zug bei Abfahrt in Zofingen also 5 Minuten Verspätung.

In beiden Fällen werden die Daten im gleichen Format und über die gleichen Schnittstellen-Services geliefert. Im Rahmen dieser Arbeit werden die Fälle jedoch sehr unterschiedlich behandelt:

- Informationen über Ereignisse der Vergangenheit können und sollen für die Vorhersage von zukünftigen Ereignissen genutzt werden. Sie liefern Werte für die Prädiktor-Variablen – unter Anwendung der Prognosemodelle können damit Werte der Zielvariablen bestimmt werden. Wenn der Interregio in Zofingen 5 Minuten Verspätung hatte, so könnte z.B. daraus geschlossen werden, dass er mindestens 4 Minuten verspätet in Bern eintreffen wird.
- Informationen über zukünftige Ereignisse stellen dagegen selbst eine Prognose dar – und diese soll in keiner Weise die Prognose des hier entwickelten Verfahrens beeinflussen. Sie dürfen also nicht von den Modellen verwendet werden. Stattdessen liefern sie einen «Benchmark»: durch späteren Vergleich mit der tatsächlich eingetretenen Verspätung (z.B. 228 Sekunden), lässt sich feststellen, wie gut das hier entwickelte Modell im Vergleich zu den Systemen der Bahnunternehmen abgeschnitten hat. Im Beispiel hätte es die Verspätung um 12 Sekunden überschätzt, während die SBB-Prognose um 108 Sekunden zu niedrig gelegen hätte.

Echtzeitdaten werden über einen Request-Reply-Mechanismus (http-Protokoll) bereitgestellt. Anfrage und Antwort werden in einem XML-Format formuliert, das als Untermenge des VDV431-Standards[73] implementiert wurde. Dieses hat Eingang in den internationalen SIRI-Standard[74] gefunden.

[73] Vgl. VDV (2015).

Es stehen drei Abfrage-Services zur Verfügung, die identische Ergebnisse liefern sollten[75]:

- `StopEvent`[76] kann zu einem Zeitpunkt und einer Haltestelle die (maximal 40) nächsten Halte sowie den vorausgehenden und nachfolgenden Fahrtverlauf des Verkehrsmittels liefern.
  Beispiel: Wird nach Ankünften in Bern ab 18:42 gefragt, so wird aufgeführt: die S1 von Thun nach Fribourg, planmässig in Bern um 18:43 mit allen ihren Halten; die S1 von Fribourg nach Thun, planmässig in Bern um 18:44 mit allen ihren Halten; der RE von Biel nach Bern, planmässig in Bern um 18:47 mit allen seinen Halten etc. Dabei werden sowohl die planmässigen als auch die tatsächlichen / prognostizierten Ankunfts- und Abfahrts-Zeiten geliefert.
- `TripRequest`[77] kann zu Zeitpunkt, Start- und Zielhaltestelle die zugehörigen Verbindungen liefern. Im SIRI-Standard ist dieser Service als Routenplaner konzipiert, in der Umsetzung der Open Data Plattform werden leider nur umsteigefreie Verbindungen geliefert.
  Beispiel: Für eine Anfrage von Bern nach Luzern werden die durchgehenden Züge sowohl via Zofingen als auch via Langnau geliefert, nicht aber die Umsteigeverbindungen via Olten. Für eine Anfrage von Bern Wankdorf nach Luzern gibt es kein Ergebnis, da keine durchgehenden Züge verkehren. Für Start, Ziel und alle Unterwegshalte werden sowohl die planmässigen als auch die tatsächlichen / prognostizierten Zeitpunkte geliefert.
- `TripInfoRequest`[78] kann zu einer Fahrt alle Haltestellen mit Ankunfts- und Abfahrtszeiten (planmässig sowie tatsächlich / prognostiziert) liefern. Die Fahrt muss durch Angabe der sogenannten `journeyref` spezifiziert werden. Die `journeyref` ist leider nicht kompatibel zum Fahrt-Bezeichner aus den Ist-Daten und kann auch nicht aus anderen Eigenschaften (z.B. Zugnummer) abgeleitet werden. Zur Ermittlung muss also zunächst einer der anderen Services (`StopEvent` oder `TripRequest`) angefragt werden, was den Nutzen von `TripInfoRequest` erheblich reduziert.

In allen drei Echtzeit-Services wird die Weltzeit (UTC) verwendet, während die Historischen Daten in Mitteleuropäischer Zeit (mit Wechsel Winter-/Sommerzeit) angegeben werden.

Anders als bei den per Download erhältlichen historischen Daten, erfordert die Nutzung der Request-Reply-Schnittstelle eine Registrierung und das Lösen eines API-Keys. Die Zahl der Abfragen ist limitiert (gemäss Angabe auf der Web-Site auf 20'000 Requests pro Tag, tatsächlich sind es offenbar 25'000).

Ein weiterer Unterschied ist, dass tatsächliche / prognostizierte Zeiten bei den Echtzeit-Services auf ganze Minuten gerundet sind – in den «Ist-Daten»-Lieferungen des Folgetages sind sie dann sekundengenau. Warum das so ist, konnte auch auf Nachfrage nicht in Erfahrung gebracht werden. Für die Ermittlung von Verspätungs-Prognosen stellt dieser Umstand eine erhebliche Einschränkung dar – ich komme später noch darauf zu sprechen.

Alle drei Echtzeit-Services habe ich durch praktische Anwendung auf ihre Eignung untersucht. Dabei traten Inkonsistenzen zu Tage:

1. Es wurden wiederholt Fälle beobachtet, in denen die drei Services nicht dieselben Prognosewerte lieferten. Abbildung 20 zeigt ein Beispiel: Für die Ankunft von IC 889 in Zug wurden im Minutentakt Prognosen bei allen drei Services angefragt. Bis 21:09 Uhr waren diese jeweils identisch, ab 21:10 Uhr weichen die Ergebnisse bei `TripRequest` von den anderen beiden Services ab.

---

[74] Vgl. CEN (2015).

[75] So zumindest die Aussage in einer Mail der Fachstelle Open Data Plattform vom 16. November 2017.

[76] Vgl. Open Data Plattform öV Schweiz (2016), Abschnitt «Abfahrts-/Ankunftsanzeiger».

[77] Vgl. Open Data Plattform öV Schweiz (2016), Abschnitt «TripRequest».

[78] Vgl. Open Data Plattform öV Schweiz (2016), Abschnitt «Fahrtprognose».

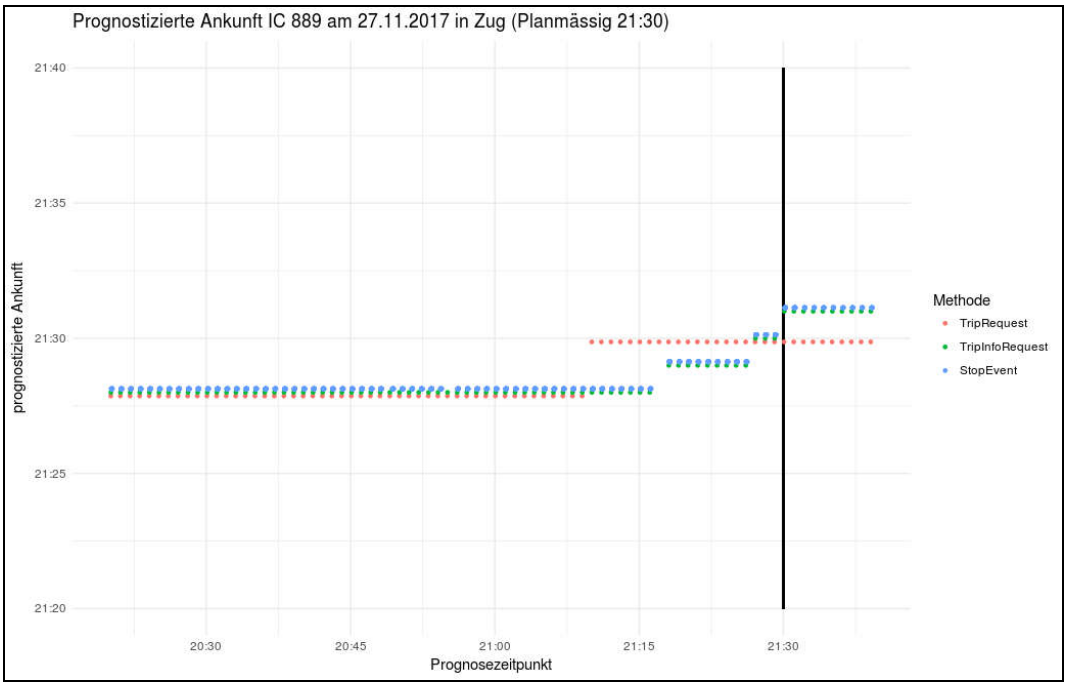


Abbildung 20: Inkonsistente Prognosen der drei Echtzeit-Services.

2. Es traten mehrfach Fälle auf, in denen die drei Services nicht die gleiche Fahrplan-Zeit lieferten. Abbildung 21 zeigt ein Beispiel: Im Minutentakt wurde die planmässige Ankunftszeit von ICN 691 in Bellinzona abgefragt. Während `TripInfoRequest` und `StopEvent` konstant dieselbe (und korrekte) Angabe liefern, änderte sich das Resultat von `TripRequest` im Zeitverlauf (was der Grundidee einer Fahrplan-Zeit widerspricht). Das Phänomen wurde nur an einigen Haltestellen und nur bei einigen Zügen beobachtet.

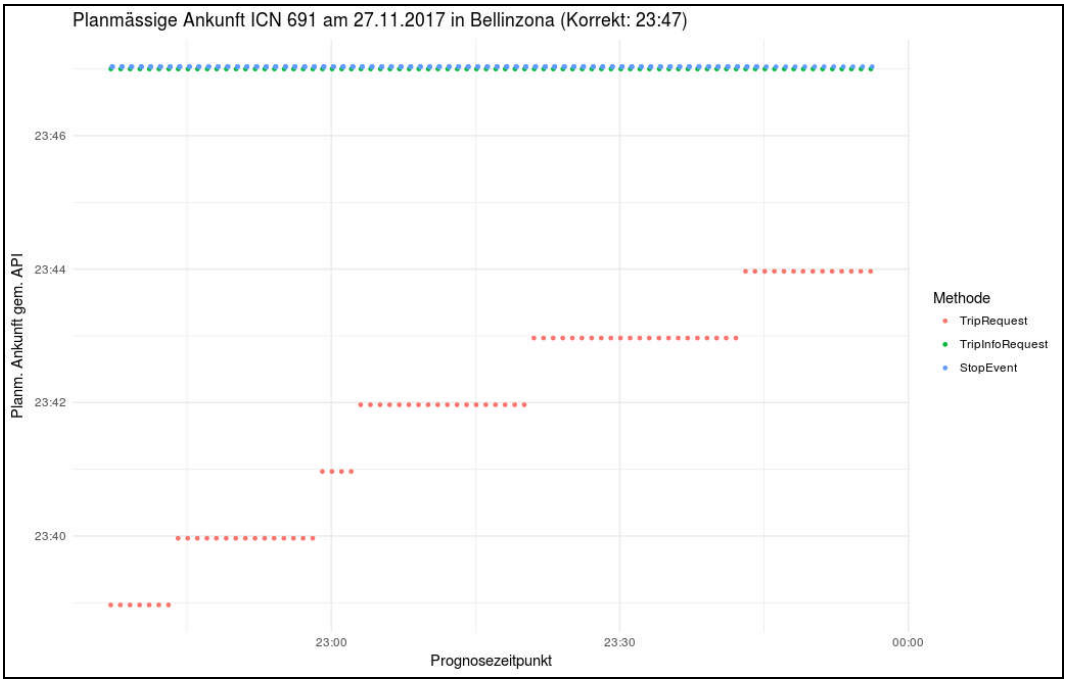


Abbildung 21: Inkonsistente Fahrplanangaben der drei Echtzeit-Services.

Beide Auffälligkeiten wurden an die Betreiber der Open Data Plattform gemeldet, konnten dort aber offenbar nicht nachvollzogen werden, so dass die Probleme potentiell weiterhin bestehen. Der betroffene Service `TripRequest` empfiehlt sich somit nicht für die Verwendung in dieser Arbeit.

`TripInfoRequest` benötigt wie oben erwähnt als Anfrageparameter die `journeyref`, was eine zweistufige Abfrage erforderlich machen würde. Dies ist nicht nur aufwendiger, sondern auch nachteilig im Hinblick auf die entstehenden Latenzen.

Die Wahl fiel daher auf die Verwendung von `StopEvent`. Als zusätzlicher Vorteil ergibt sich dabei, dass mit einer einzigen Anfrage sehr viele Ereignisse abgefragt werden können, was das verfügbare «Request-Budget» schont: Jede Fahrt umfasst im Durchschnitt etwa 13 Halte (stark variierend je nach Zugslinie), was 24 Ereignissen entspricht (Abgangs- und Zielort haben jeweils nur ein Ereignis, alle anderen Halte haben Ankunft und Abfahrt). Mit einem Request lassen sich bis zu 40 Fahrten, also fast 1000 Ereignisse abfragen.

Für die Realisierung der http-Abfragen wird die R-Library `httr` verwendet, das Parsing der gelieferten Antwort erfolgt mit `xml2` – die zunächst verwendete Alternative `xml` erwies sich als sehr langsam.

Der Bezug der Echtzeitdaten ist als R-Funktion `ODPStopEvents` implementiert. Diese gibt eine Liste von Halten zurück. Aufrufparameter sind

- Eine Liste von `Abfragen`, bestehend jeweils aus einer Haltestelle und einem Minuten-Offset auf die aktuelle Uhrzeit. So können z.B. die in Zürich haltenden Züge, beginnend vor 60 Minuten, und die in Bern haltenden Züge, beginnend vor 30 Minuten, abgefragt werden. Dies ist sinnvoll, da für unterschiedliche Haltestellen mitunter unterschiedliche Zeithorizonte von Interesse sind.
- Die `ZeitfensterGroesse`. Dies gibt die Intervall-Länge (z.B. 60 Minuten) an, für die mindestens Daten geliefert werden sollen. In jedem Request werden Daten von 40 Fahrten bei der Plattform abgefragt. Reicht dies nicht aus, um das Intervall abzudecken, werden mehrere Requests gestellt.

`ODPStopEvents` ist nachfolgend in Pseudocode beschrieben. Der R-Code findet sich in der beigefügten Datei RT_PrognosenBern.R.

```
ODPStopEvents <- function(Abfragen, ZeitfensterGroesse) {
  Für alle Abfragen (bestehend je aus Haltestelle und benötigtem Vorlauf):
    Bis Zeitfenster abgedeckt ist:
      Formuliere Request zu Haltestelle, aktuellem Zeitpunkt, benötigtem Vorlauf
      Stelle Request an Open Data Plattform
      Für alle in der Response erhaltenen Fahrten:
        Ermittle Betriebstag
        Ermittle Zugsbezeichnung
        Für alle in der Fahrt enthaltenen Halte:
          Ermittle Fahrplan-Zeiten
          Ermittle Tatsächliche/prognostizierte Zeiten
          Füge einen Eintrag zur Resultatliste hinzu
   Gib Resultatliste zurück
}
```

Abbildung 22: Pseudocode: Abfrage von Echtzeitdaten

Diese Implementierung hat sich als ausreichend für die Zwecke dieser Arbeit erwiesen. Es bestehen jedoch auch einige Schwächen, die bei einer Weiterentwicklung angegangen werden sollten:

- Es kommt vor, dass die Open Data Plattform zu einem Request keine Echtzeitangaben, gar keine Daten oder eine Fehlermeldung liefert. Ursachen dafür können z.B. Überlastung, Netzwerkprobleme oder interne Fehler sein. In diesem Fall liegen die benötigten Daten nicht vor. Sinnvoll wäre es,

in solchen Fällen auf ein «Gedächtnis» zuzugreifen, mit dessen Hilfe die Ergebnisse früherer Requests wiederverwendet werden könnten.

- Das Handling solcher Störungen ist nur simpel implementiert: Das Skript bricht vollständig ab, statt den Request entweder zu wiederholen oder mit dem nächsten Request fortzufahren.
- Das Parsing der XML-Responses ist – obwohl bereits erhebliche Bemühungen erfolgt sind – noch immer recht langsam: Die Abfrage und Verarbeitung einer Haltestelle mit 40 Fahrten dauert in der verwendeten Umgebung ca. 2 Sekunden. Möglicherweise ist eine Beschleunigung unter Verwendung externer Tools (z.B. ein Pre-Processing mit xsltproc)[79] möglich.
- Die Abarbeitung der Abfragen für unterschiedliche Haltestellen erfolgt seriell. Eine Parallelisierung könnte die Gesamt-Durchlaufzeit verkürzen.

4.3 Weitere Daten

Zur Übersetzung der verwendeten fachlichen und technischen IDs in «sprechende Angaben» werden weiterhin einige Dimensionstabellen von puenktlichkeit.ch verwendet. Die darin enthaltenen Angaben stammen ebenfalls von der Open Data Plattform.[80] Es sind dies:

- `dim_bp`: Liste aller Haltestellen («Betriebspunkte»), inkl. Geo-Koordinaten
- `dim_btr`: Liste aller Verkehrsunternehmen («Betreiber»)
- `dim_vm`: Liste aller Verkehrsmittel

Zur Identifikation von Haltestellen gibt es neben dem internationalen Code auch eine «Betriebspunkt-Abkürzung», bestehend aus meist 2-4 Buchstaben. Diese ist innerhalb der Schweiz eindeutig, wird aber nur für Bahn-Haltestellen vergeben. Da die Abkürzung oft an die Namen der Haltestellen angelehnt ist (z.B. «ZUE» für «Zürich», «BN» für «Bern», «BNWD» für «Bern Wankdorf»), kann sie vom Menschen wesentlich einfacher verstanden werden. Dies erleichtert Prüfung und Interpretation der Daten erheblich, so dass im Rahmen der Arbeit die Haltestellen-Abkürzung als Identifier verwendet wird. Bei einer Übertragung der Methode auf andere Verkehrsmittel als die Bahn müsste hier ein Ersatz durch Haltestellen-Codes vorgenommen werden.

[79] Zu xsltproc siehe http://xmlsoft.org/XSLT/ . Dieser und weitere Vorschläge werden diskutiert auf https://stackoverflow.com/questions/44257890/parsing-large-and-complicated-xml-file-to-data-frame .

[80] Vgl. Gutweniger (2017), S. 14ff.

5 Prognoseverfahren

Gegenstand dieses Kapitel ist die Entwicklung und Implementierung eines Prognoseverfahrens, das in der Lage ist, auf Basis vorhandener Daten Verspätungsprognosen zu erstellen.

Zunächst werden die relevanten Rahmenbedingungen und Anforderungen identifiziert (5.1). Abschnitt 5.2 erläutert die grundsätzliche Eignung von «Recursive Partitioning» bezüglich dieser Anforderungen. Das Verfahren wird in einer explorativen Untersuchung (5.3) näher untersucht und alternativen Ansätzen gegenübergestellt (5.4). Von zentraler Bedeutung sind die Auswahl und Modellierung der Variablen (5.5) und die systematische Generierung von Modellen für unterschiedliche Variablenkonstellationen und Prognosehorizonte (5.6). Die beiden letzten Abschnitte widmen sich ausgewählten Aspekten bei der Implementierung des Verfahrens.

5.1 Rahmenbedingungen und Anforderungen

Gemäss dem in Abschnitt 2.5 formulierten Szenario sollen Prognosen für zwei Fragestellungen geliefert werden:

1. Wie gross wird das Ausmass einer Ankunftsverspätung für einen bestimmten Zug (z.B. IC 808) an einer bestimmten Haltestelle (Bahnhof Bern) sein? Dabei sind ganzzahlige Minutenangaben von ausreichender Genauigkeit: Dies entspricht dem, was Kunden und Bahnmitarbeiter von den bestehenden Systemen gewöhnt sind. Und auch die als Benchmark verwendeten Echtzeit-Prognosen der Open Data-Plattform beschränken sich auf ganzzahlige Minuten.
2. Wird das Ausmass einer Verspätung einen gegebenen Wert (z.B. 5 Minuten) überschreiten? Wird also z.B. IC 808 mit mehr als 5 Minuten Verspätung in Bern eintreffen?

Abbildung 23 zeigt exemplarisch die Verteilung von Ankunftsverspätungen im Bahnhof Bern. Zu erkennen ist, dass ein Grossteil der Ankünfte (ca. 54%) pünktlich oder sogar verfrüht erfolgt. Die Prognose von Verfrühungen ist meist nicht von Interesse und soll auch in dieser Arbeit ausgeklammert bleiben.[81] Die Verspätungen sind ausgesprochen schief verteilt: nur 1 Prozent aller Ankünfte erfolgt mehr als 10 Minuten, nur 1 Promille aller Ankünfte mehr als 25 Minuten verspätet. Bei Ausschluss von Extremfällen ist der Wertebereich der Zielvariablen also sehr klein.

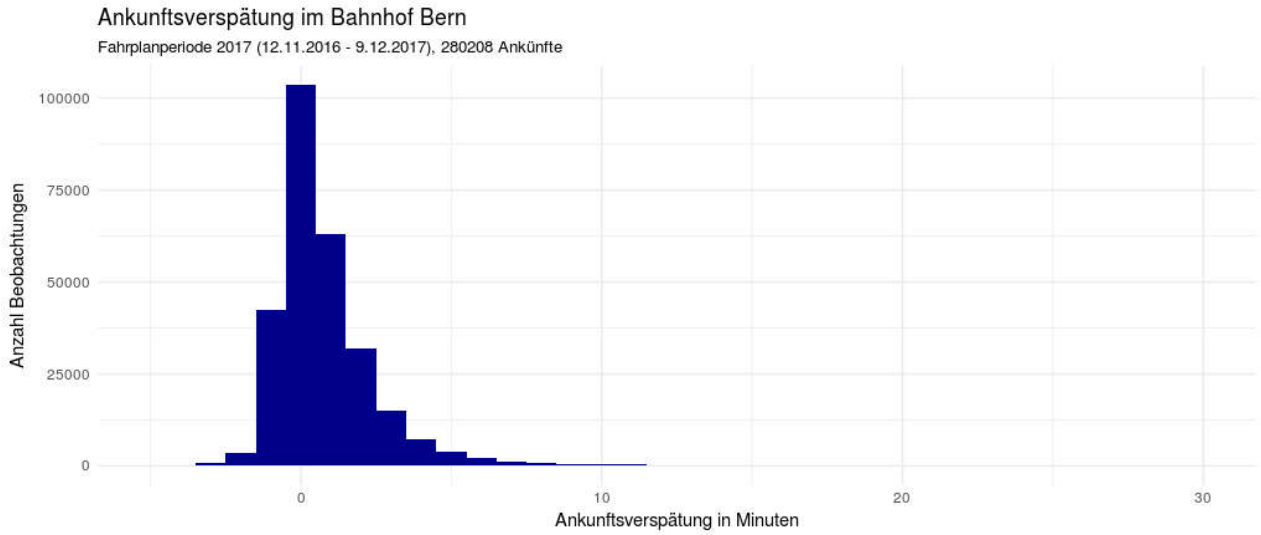


Abbildung 23: Verspätungsverteilung im Bahnhof Bern, Fahrplanperiode 2017.

[81] Vgl. Abschnitt 2.5. Grundsätzlich kann das hier entwickelte Verfahren auch für die Prognose von Verfrühungen eingesetzt werden.

Aus Kapitel 4 ist bekannt, welche Daten verfügbar sind und als Prädiktoren im Prognoseverfahren genutzt werden können: In Frage kommen in erster Linie zurückliegende tatsächliche Ankunfts- und Abfahrtszeitpunkte beliebiger Züge an beliebigen Haltestellen sowie daraus abgeleitete Werte (insbesondere die aufgetretenen Verspätungen). Dabei kann es sich sowohl um Ereignisse desselben Zugs an einer vorherigen Haltestelle als auch um Ereignisse eines anderen Zugs an derselben oder einer anderen Haltestelle handeln.[82] Ergänzend könnten z.B. Jahreszeit oder Wochentag (jeweils abgeleitet aus dem Betriebstag) Eingang in die Prognosemodelle finden.

An das Prognoseverfahren werden folgende Anforderungen gestellt:

- Mit dem Verfahren soll die **Prognosequalität** der Bahn-Systeme mindestens in Teilbereichen übertroffen werden.[83] Es ist wichtig, diese Teilbereiche (z.B. bestimmte Zugskategorien, Linien, Fahrtverläufe, Verspätungskonstellationen, Vorhersagehorizonte) identifizieren zu können und klare Aussagen über die erzielte Prognosequalität zu erlangen. Als Performance-Indikatoren dienen einerseits der Anteil korrekter Prognosen (in einem gegebenen Toleranzbereich) und andererseits der Anteil korrekt erkannter Verspätungen eines gegebenen Mindestumfangs.[84] Die Systeme der Bahnunternehmen definieren das Zielniveau, das für eine positive Beantwortung der Forschungsfrage erreicht werden muss. Es ist nicht notwendig, die Prognosequalität deutlich über dieses Niveau hinaus zu steigern – die Wahl des Verfahrens kann einem «good enough»-Prinzip folgen.
- Die Prognosen sollen sehr defensiv, d.h. stark «**biased**» sein: Eine Überschätzung der Prognosen ist zu vermeiden, da dies wesentlich gravierendere Auswirkungen haben könnte als eine Unterschätzung.[85] Der Anteil von «false positives» soll sich auf ähnlichem Niveau bewegen wie bei den Systemen der Bahnunternehmen.
- Die Erstellung valider Modelle muss mit dem vorhandenen **Datenumfang** möglich sein. Dieser ist begrenzt: Verfügbar sind historische Daten seit 12.11.2016, also für wenig mehr als 400 Tage. Bei Rückbehalt von Daten für Validierung und ggf. Test (vgl. Fussnote 69 auf S. 29), stehen entsprechend weniger Daten für das Training der Modelle zur Verfügung. Zudem ist zu beachten, dass es beim Fahrplanwechsel (zuletzt 9./10. Dezember 2017) zu relevanten Veränderungen gekommen sein kann, so dass ein im Vorjahr trainiertes Modell nicht uneingeschränkt auf die laufende Periode anwendbar ist. Ausgenutzt werden können möglicherweise die Eigenschaften des Taktfahrplans: Dort, wo Züge im Stundentakt verkehren, ergeben sich ca. 15 bis 20 sehr ähnliche Konstellationen pro Tag, so dass die dort gemachten Beobachtungen als Ausprägungen derselben Variablen interpretiert werden können.
- Die Erstellung valider Modelle sollte trotz sehr ungleich verteilter Datengrundlage möglich sein: Wie anhand von Abbildung 23 gezeigt wurde, treten kleine und grosse Verspätungen mit sehr unterschiedlicher Häufigkeit auf, wobei ausgerechnet die interessanten Fälle (grosse Verspätungen) selten sind. Diese im Zusammenhang mit Klassifikationsverfahren als «**Class Imbalance**» bekannte Situation kann die Performance von Machine Learning-Verfahren erheblich beeinträchtigen.[86]
- Ein **Kausalzusammenhang** zwischen Prädiktoren und Zielvariablen sollte mindestens plausibel sein. Vermieden werden sollen Modelle, die allein auf Korrelation in den Daten beruhen, ohne dass eine fachliche Begründung für den Zusammenhang möglich wäre. Dies wäre etwa der Fall, wenn ein Ereignis in Grindelwald als Prädiktor für eine Zugverspätung in Klosters benutzt würde.

[82] Ob es sich um denselben oder einen anderen Zug handelt, kann in den Daten anhand des «Fahrtbezeichners» (historische Daten) bzw. der «journeyref» (Echtzeitdaten) festgestellt werden.

[83] Vgl. Zielsetzung in Abschnitt 1.2.

[84] Vgl. Szenario aus Abschnitt 2.5. Die Qualitäts-Kriterien werden in Abschnitt 6.1 noch detaillierter behandelt.

[85] Vgl. Abschnitt 1.2.

[86] Das Problem ist an diversen Methoden untersucht worden, z.B. bei Entscheidungsbäumen (vgl. Cieslak / Chawla (2008)) und Neuronalen Netzen (vgl. Japkowicz (2000)). Einen guten Überblick über das Problem und mögliche Massnahmen liefern Guo / Yin / Dong / Yang / Zhou (2008).

Oder wenn ein Ereignis in Olten für eine Prognose in Bern verwendet würde, obwohl zwischen beiden Ereignissen nur 2 Minuten liegen.

- Die verwendeten Modelle sollen möglichst **frei von Annahmen über die Struktur des Zusammenhangs** von Prädiktoren und Zielvariablen sein. Wie in Kapitel 2 erörtert wurde, gibt es im Betriebsablauf der Bahnen zahlreiche Zusammenhänge, die nicht-linearer Natur sind. Häufig treten sogar Unstetigkeiten in Form von «Tipping Points» auf (z.B. Abwarten vs. Brechen eines Anschlusses; Änderung einer Zugskreuzung; Signalhalt vs. «Grüne Welle»). Da eine zufriedenstellende formale Beschreibung dieser Phänomene kaum möglich ist, sollte eine möglichst parameterfreie Modellierung angestrebt werden. Hierin unterscheidet sich meine Arbeit von den meisten bisherigen Publikationen zum Thema.
- Das Verfahren sollte eine grosse **Zahl von Variablen** verarbeiten können. Wie viele es sind, hängt ganz wesentlich von der gewählten Modellierung ab. Als Anhaltspunkt für die mögliche Komplexität der Aufgabenstellung kann aber folgende Überlegung dienen: innerhalb von 1 Stunde halten am Bahnhof Bern etwa 45 Züge, die auf ihrer Fahrt bis nach Bern zusammen etwa 500 Ereignisse (Ankunft oder Abfahrt) durchlaufen haben. Jeder dieser Züge hatte auf seiner Fahrt Interaktionen (z.B. Anschlüsse, Kreuzungen, Zugfolge) mit zahlreichen anderen Zügen, die ihrerseits bereits eine reiche «Ereignis-Geschichte» mit sich trugen.
- Das Verfahren sollte **robust gegenüber Missing Values** sein, da diese häufig auftreten. Beispiel: Sollte sich herausstellen, dass die Ankunft eines Zuges in Burgdorf ein guter Prädiktor für seine spätere Ankunft in Bern ist und zwischen diesen beiden Ereignissen üblicherweise etwa 18 Minuten liegen, so ist die entsprechende Variable ungeeignet für alle Prognosen, die einen längeren Vorlauf haben. Um die Ankunftszeit in Bern 20 Minuten im Voraus vorherzusagen, sollte das Verfahren auf eine oder mehrere geeignete «Ersatz-Variablen» ausweichen können.
- Das Verfahren sollte berücksichtigen, dass bestehende Prädiktoren obsolet werden können, wenn neue Informationen hinzukommen. Es sollte sich die **zum jeweiligen Zeitpunkt relevanten Prädiktoren** auswählen. Die Ankunftszeit eines Zuges in Burgdorf mag hohe Vorhersagekraft für seine Ankunft in Bern haben. Sobald jedoch auch seine Abfahrtszeit in Burgdorf bekannt ist, ist diese vermutlich besser geeignet. Sie wird den bisherigen Prädiktor obsolet machen.[87]
- Um für «Echtzeitprognosen» einsetzbar zu sein, sollte das Verfahren **effizient** auf eine gegebene Datenkonstellation angewendet werden können. Wenn man davon ausgeht, dass die Gesamt-Latenz unter einer Minute liegen soll (z.B. um eine Aktualisierung im Minutentakt zu ermöglichen) und für Bezug und Aufbereitung der notwendigen Daten 20-30 Sekunden benötigt werden, so sollte auch die Berechnung der Prognosen nicht mehr als 30 Sekunden dauern.
- Nicht notwendig, aber sehr wünschenswert ist, dass das Verfahren Modelle erzeugt, die **von Menschen interpretiert werden können**[88]. Dies erleichtert einerseits die Entwicklung und Evaluation, andererseits macht es die gelieferten Prognosen erklär- und nachvollziehbar. Im besten Fall kann es dazu führen, dass existierende Betriebsabläufe besser verstanden und Verspätungsursachen identifiziert werden können. Dies ist zwar nicht Ziel dieser Arbeit, wäre aber ein willkommenes Nebenprodukt.

[87] Unwahrscheinlich, aber grundsätzlich denkbar ist, dass in diesem Beispiel auch die Interaktion beider Variablen – Ankunft und Abfahrt in Burgdorf – relevant ist; etwa, weil die dadurch ausgedrückte Aufenthaltszeit in Burgdorf Aufschluss über das Passagieraufkommen und somit indirekt auch über den weiteren Fahrtverlauf liefert. In diesem Fall sollten beide Variablen als Prädiktoren verwendet werden.

[88] Die Eigenschaft der «Interpretierbarkeit» ist intuitiv verständlich, definitorisch aber äusserst schwer zu fassen, vgl. Doshi-Velez / Kim (2017). Häufig wird sie mit «White-Box»-Ansätzen gleichgesetzt, bei dem die vom Modell erzeugten Strukturen für den Menschen leicht zugänglich sind, etwa bei Entscheidungsbäumen oder Regressionsgleichungen vgl. Molnar (2018), Kapitel 4. Intepretierbarkeit kann aber auch im Nachhinein und unabhängig vom verwendeten Modelltyp hergestellte werden («Modell-agnostische Interpretierbarkeit», vgl. Molnar (2018), Kapitel 5).

Die Erfüllung dieser Anforderungen ist einerseits vom gewählten Machine Learning-Verfahren, andererseits von der Auswahl und Modellierung der Prädiktor-Variablen abhängig. Die beiden folgenden Abschnitte gehen darauf ein.

5.2 Auswahl von Machine Learning-Verfahren und Typ des Prognosemodells

In den letzten Jahrzehnten sind unzählige Machine Learning-Verfahren entwickelt worden – allein die populäre R-Library «Caret» unterstützt inzwischen fast 240 Verfahren und deren Varianten.[89] Eng verbunden mit dem Lern-Verfahren ist der Typ des dabei erzeugten Prognosemodells. Teilweise werden sogar gleiche Bezeichnungen für Verfahren und Modelltyp verwendet (z.B. «Linear Regression (Model)»). Häufig gibt es mehrere Algorithmen, die denselben Modelltyp erzeugen, sich aber in der Art der Berechnung unterscheiden (z.B. können Entscheidungsbäume mit ID3[90], C4.5[91], aber auch mit AdaBoost[92] und diversen anderen Verfahren erzeugt werden). Teilweise sind für die Bezeichnung der Verfahren auch die Namen der Produkte und Bibliotheken gebräuchlich, in die sie Eingang gefunden haben (z.B. rpart, CART, Weka).

Eine erste grobe Einteilung der Verfahren liefert die Unterscheidung in Klassifikation (metrische Zielvariable) und Regression (kategorielle Zielvariable). Die hier verfolgten Fragestellungen sind wie folgt einzuordnen:

1. Ob ein gegebener Verspätungs-Schwellwert übertroffen wird, kann mit einem Klassifikationsverfahren vorhergesagt werden. Dabei existieren nur 2 Kategorien: verspätet / nicht verspätet (binäre Klassifikation).
2. Die Frage nach dem Ausmass einer Verspätung impliziert grundsätzlich eine metrische Zielvariable. Wie oben erläutert wurde, ist jedoch eine Vereinfachung auf einen kleinen Bereich von diskreten Werten möglich: Ganzzahlige Angaben sind ausreichend, Verfrühungen brauchen nicht prognostiziert zu werden, Verspätungen über 15 Minuten sind vernachlässigbar aufgrund ihrer Seltenheit. Der Wertebereich der Zielvariablen ist somit überschaubar (keine Verspätung / 1 Minute Versp. / 2 Minuten Versp. / 3 Minuten Versp. / … / 14 Minuten Versp. / 15 Minuten oder mehr Verspätung). Es kann also auch hier ein Klassifikationsverfahren verwendet werden.

Im zweiten Fall besteht die Besonderheit, dass die Kategorien eine Ordnung aufweisen: Die Zielvariable ist ordinal skaliert. Diese Ordnung soll vom Klassifikationsverfahren berücksichtigt werden, es soll sich «monoton» verhalten.[93] Viele Verfahren wählen jedoch von mehreren Klassen jene, die die höchste Wahrscheinlichkeit hat. Dies kann zu Anomalien führen, wie das Beispiel in Abbildung 24 zeigt:

<u>Ausgangssituation:</u>
Wahrscheinlichste Prognose =
3 Minuten Verspätung

Klasse	WS
2 min	45%
3 min	**47%**
4 min	8%

<u>Modifikation:</u>
Indizien für eine noch höhere
Verspätung treten auf

Klasse	WS
2 min	**35%**
3 min	34%
4 min	31%

Abbildung 24: Beispiel für eine Verletzung der Monotonie bei Prognose einer ordinalen Zielvariable

[89] Vgl. Kuhn (2018), Kapitel 6.

[90] Vgl. Quinlan (1986).

[91] Vgl. Quinlan (1993).

[92] Vgl. Freund / Schapire (1996).

[93] Vgl. Potharst / Bioch / Petter (1997).

In der Ausgangssituation wird von (hier vereinfachend) drei möglichen Prognosen diejenige mit der höchsten Wahrscheinlichkeit (3 min) gewählt. In der modifizierten Situation gibt es stärkere Indizien für eine höhere Verspätung – die Wahrscheinlichkeiten verschieben sich in Richtung «4 min». Paradoxerweise führt dies dazu, dass nun «2 min» als wahrscheinlichste Klasse gewählt wird.

Dies kann durch Verwendung eines spezifischen Verfahrens für ordinale Klassifikation vermieden werden.[94] Ein sehr einfacher Ansatz besteht darin, die Einteilung in k Klassen als Kombination von k-1 binären Klassifikationsaufgaben aufzufassen: es wird sukzessive geprüft, ob eine Verspätung von *mindestens* 1 min, *mindestens* 2 min, *mindestens* 3 min etc. vorliegt. Die höchste positiv verlaufene Prüfung bestimmt die Höhe der prognostizierten Verspätung.[95]

Dieser Ansatz führt die Frage nach dem Ausmass der Verspätung somit auf jene nach dem Überschreiten bestimmter Verspätungsniveaus zurück: Für die k-1 benötigten binären Klassifikatoren können dieselben Modelle verwendet werden, die auch für die Mindestverspätung zum Einsatz kommen.

Für das Training der Modelle wird *Recursive Partitioning* in der Implementation der R-Library `rpart` verwendet.[96] Recursive Partitioning bietet folgende Vorteile:

- Das Verfahren liefert Entscheidungsbäume. Diese sind für Menschen leicht zu verstehen. Das ist hilfreich bei der Entwicklung und Evaluation des Verfahrens und führt zudem dazu, dass die erzeugten Prognosen nachvollziehbar werden.
- Das Verfahren ist non-parametrisch, d.h. es impliziert keinerlei Annahmen über die Struktur des Zusammenhangs von unabhängigen und abhängigen Variablen.
- Die Modelle können im Rahmen der Prognose sehr effizient angewendet werden.
- Der erwünschte «Bias» lässt sich beim Training über eine Loss-Matrix spezifizieren.

Ein verbreiteter Nachteil von Recursive Partitioning ist die Tendenz zum Overfitting. Essentiell ist daher das «Pruning», d.h. das «Zurückschneiden» des vorab erzeugten Entscheidungsbaums auf seine optimale Grösse. Bei `rpart` ist dieser Vorgang bereits in das Verfahren integriert: Nach Anlegen des Baums wird er vom Algorithmus soweit reduziert, dass der kreuzvalidierte Fehler minimal wird. Das Erzeugen kleiner Bäume ist im Rahmen dieser Arbeit noch aus weiteren Gründen wichtig:

- Die Verständlichkeit der Bäume nimmt mit zunehmender Komplexität rapide ab.
- Da für die Prognose des Verspätungsausmasses 15 Klassifikationen benötigt werden, soll jedes einzelne Modell klein sein, um eine ausreichende Geschwindigkeit sicher zu stellen.

5.3 Explorative Untersuchung

Um einen ersten Eindruck von der Eignung zu erhalten, wurde eine explorative Untersuchung durchgeführt. Dabei fand auch ein Vergleich mit anderen Verfahren statt (siehe dazu Abschnitt 5.3). Ziel war es, mit geringen Aufwand einen aussichtsreichen Kandidaten für das weitere Vorgehen zu identifizieren. Es wurde also nach einem Verfahren gesucht, dass mit hoher Wahrscheinlichkeit «gut genug» sein würde – nicht notwendig nach dem Besten. Im Vordergrund stand, eine erste Einschätzung von der Eignung der Verfahren bezüglich eines breiten Spektrums an Anforderungen zu erhalten, ohne dass damit bereits eine abschliessende Bewertung getroffen wäre.

Als Szenario wurde bewusst eine Situation gewählt, in der nicht-lineare Zusammenhänge vermutet werden können: In Olten kommen zu jeder halben Stunde (zwischen Minute :22 und :30) Züge an aus

[94] Vgl. Potharst / Bioch (1999), Archer (2010), Galimberti / Soffritti / Di Maso (2012).

[95] Ein sehr ähnliches Verfahren schlagen Frank und Hall vor, vgl. Frank / Hall (2001).

[96] Vgl. Therneau / Atkinson (2018). rpart implementiert im Wesentlichen das CART-Verfahren, vgl. dazu Breiman / Friedman / Olshen / Stone (1984).

Aarburg-Oftringen (S-Bahn), Langenthal (IR), Dulliken (RE), Olten Hammer (R), Liestal (IC), Luzern (IR), Zürich (IR), Basel (IC) und Bern (IC). Der IC von Bern fährt weiter nach Liestal. Bei der Ankunft in Olten ist er in 22% der Fälle mehr als 3 Minuten verspätet, bei der Abfahrt nach Liestal sind es 32% - der Zug baut also in Olten Verspätung auf. Fahrplan und Geographie sind in Abbildung 25 dargestellt.

Fahrt	An	Prognose	Von	Gleis/Kante/ Haltestelle
S 8 22836	**10:22**		**Sursee** Sursee 09:51 - St. Erhard-Knutwil 09:53 - Wauwil 09:56 - Nebikon 10:01 - Dagmersellen 10:03 - Reiden 10:07 • Brittnau-Wikon 10:09 - Zofingen 10:13 - Aarburg-Oftringen 10:16 - Olten 10:22	9
IR 17 2365	**10:24**		**Bern** Bern 09:39 - Burgdorf 09:53 - Herzogenbuchsee 10:05 - Langenthal 10:12 - Olten 10:24	3
RE 4764	**10:24**		**Wettingen** Wettingen 09:42 - Baden 09:48 - Turgi 09:53 - Brugg AG 09:59 - Wildegg 10:06 - Aarau 10:13 - Dulliken 10:18 - Olten 10:24	4
R 7815	**10:24**		**Biel/Bienne** Biel/Bienne 09:21 - Grenchen Süd 09:35 - Solothurn 09:49 - Niederbipp 10:01 - Oensingen 10:05 - Oberbuchsiten 10:08 • Hägendorf 10:14 - Wangen bei Olten 10:17 - Olten Hammer 10:19 - Olten 10:24	1
IC 61 965	**10:25**		**Basel SBB** Basel SBB 09:59 - Liestal 10:09 - Olten 10:25	11
IC 21 668	**10:27**		**Chiasso** Chiasso 07:12 - Balerna 07:17 - Mendrisio 07:24 - Lugano 07:42 - Bellinzona 08:13 - Flüelen 08:49 - Arth-Goldau 09:14 - Luzern 09:54 - Olten 10:27	10
IR 26 2319	**10:28**		**Basel SBB** Basel SBB 10:04 - Olten 10:28	12
IR 17 2364	**10:28**		**Zürich HB** Zürich HB 09:55 - Olten 10:28	9
IC 61 964	**10:30**		**Interlaken Ost** Interlaken Ost 09:00 - Interlaken West 09:05 - Spiez 09:22 - Thun 09:33 - Bern 10:04 - Olten 10:30	→ fährt weiter nach Liestal

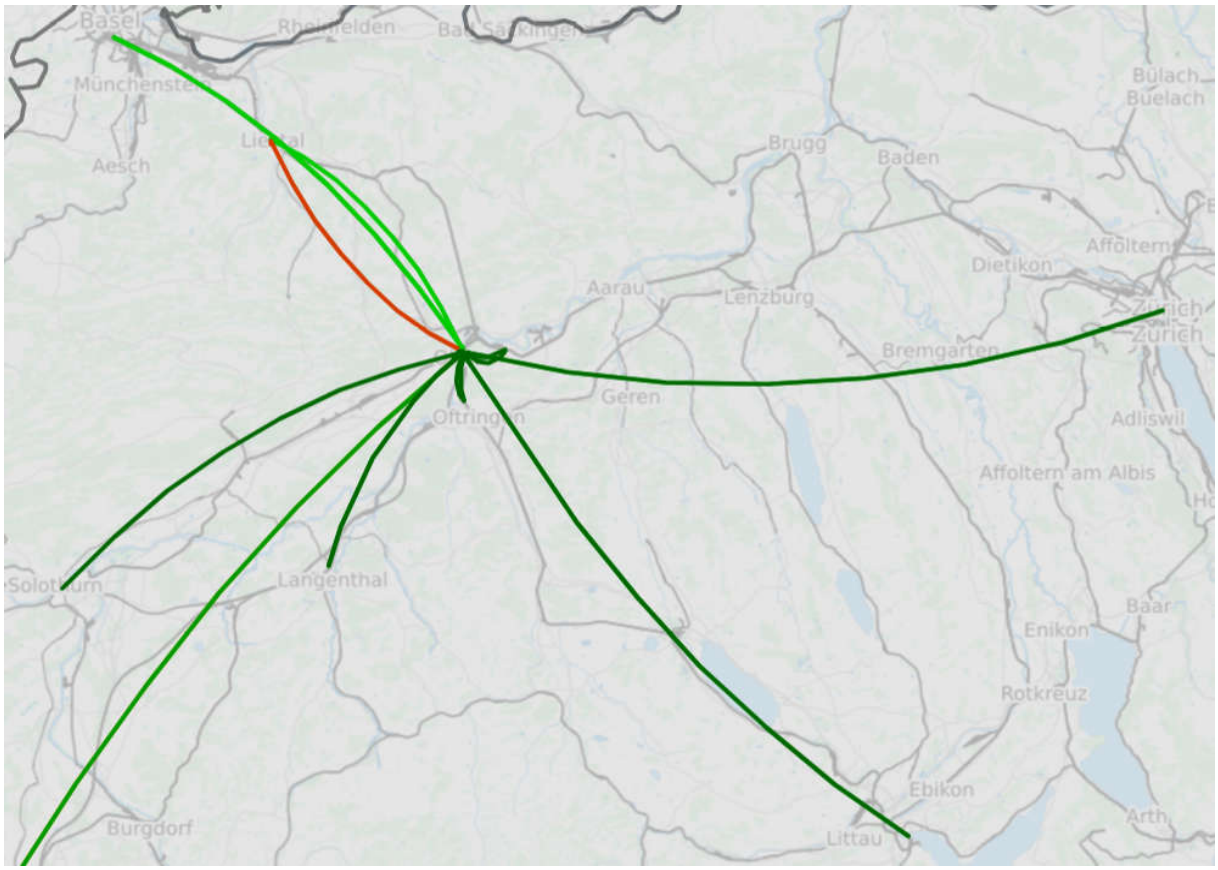


Abbildung 25: Szenario für die Exploration: Ankunft in Olten aus 9 Richtungen (grün) und Weiterfahrt nach Liestal (rot).

Es erscheint plausibel, dass Zusammenhänge mit einigen der anderen 8 Zügen bestehen, weil

- diverse Umsteigebeziehungen existieren und Anschlüsse teilweise abgewartet werden,
- bei der Ein- und Ausfahrt in Olten teilweise gleiche Anlagenteile befahren werden (Zugfolgen und sich kreuzende Fahrwege).

Im Szenario soll der Zusammenhang der insgesamt 9 Zugsankünfte und der Abfahrt des IC nach Liestal untersucht werden. Prädiktoren sind somit die 9 Ankunftsverspätungen in Olten, Zielvariable ist die Abfahrtsverspätung des IC nach Liestal, ebenfalls in Olten.

Das Verfahren aus Abschnitt 5.2 benötigt für jedes Verspätungsniveau (= ganzzahlige Minuten) Entscheidungsbäume. Diese können mit `rpart` erzeugt werden. Abbildung 26 zeigt beispielhaft einen solchen Baum und erläutert dessen Notation. Für Benennung der Variablen habe ich folgende Konvention getroffen (Trennung der 5 Komponenten jeweils durch Unterstrich «_»):

1. Kürzel der Haltestelle am Beginn des Abschnitts
2. Kürzel der Haltestelle am Ende des Abschnitts
3. `AB` für Abfahrtsverspätung am Beginn oder `AN` für Ankunftsverspätung am Ende des Abschnitts
4. Minute, zu der die Abfahrt bzw. Ankunft planmässig erfolgen sollte
5. Zeitabstand zwischen betrachtetem Ereignis und Prognose-Ereignis in Minuten. Dies wird benötigt, um Situationen zu berücksichtigen, bei denen mehr als 1 Stunde zeitlicher Abstand besteht (die Minutenangabe wäre dann nicht eindeutig).

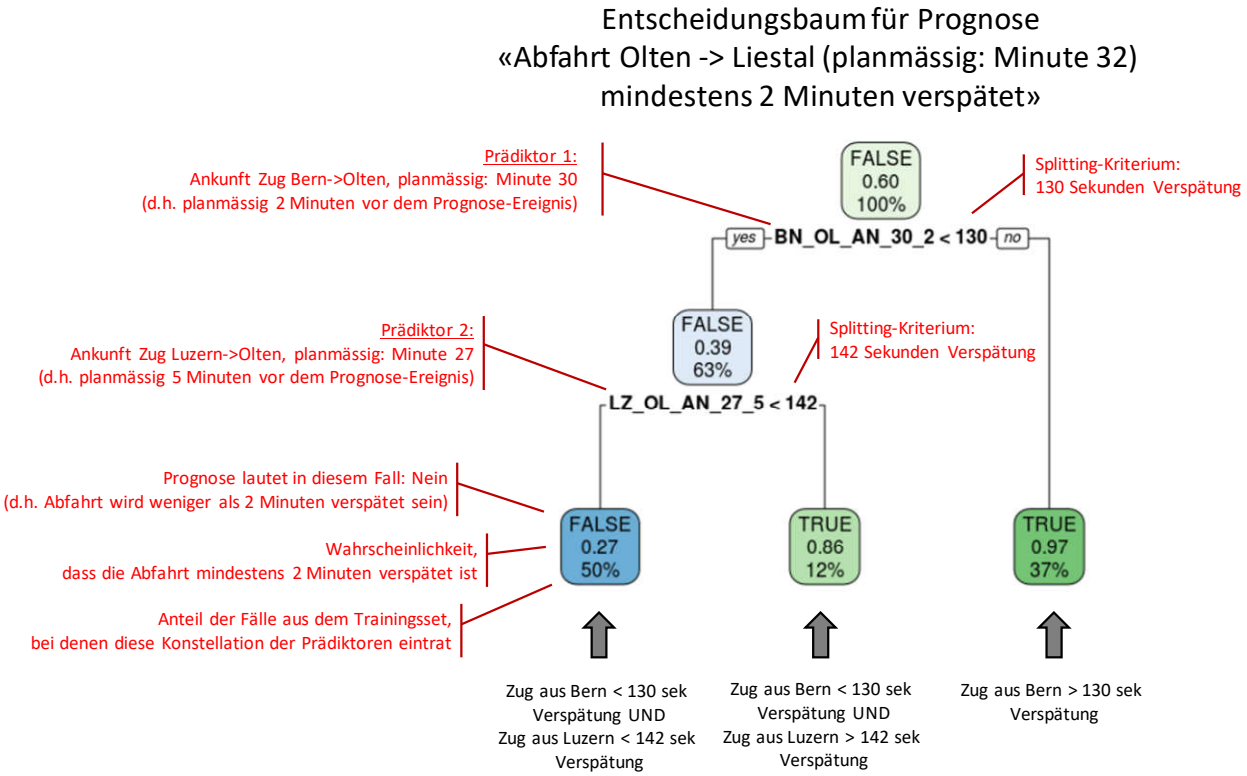


Abbildung 26: Entscheidungsbaum für die Prognose einer 2-minütigen Abfahrtsverspätung in Olten.

`rpart` hat in diesem Beispiel also 2 Prädiktoren als relevant identifiziert: Die Ankunft aus Bern (dies ist naheliegend; denn dies ist der Zug, der nach Liestal weiterfährt) und die Ankunft aus Luzern (ein anderer Zug; z.B., weil dessen Anschluss abgewartet wird oder weil sich die Fahrwege der Züge behindern). Der generierte Baum unterscheidet 3 Fälle: Eine verspätete Abfahrt (auf Niveau 2 Minuten

oder mehr) wird prognostiziert, wenn die Ankunft aus Luzern um mehr als 142 Sekunden verspätet ist oder die Ankunft aus Bern um mehr als 130 Sekunden. Trifft beides nicht zu, wird eine pünktliche Abfahrt prognostiziert (d.h. eine Verspätung von weniger als 2 Minuten).

Abbildung 27 zeigt die beiden Entscheidungsbäume, die von `rpart` für eine 3- bzw. 4-minütige Abfahrtsverspätung generiert wurden.

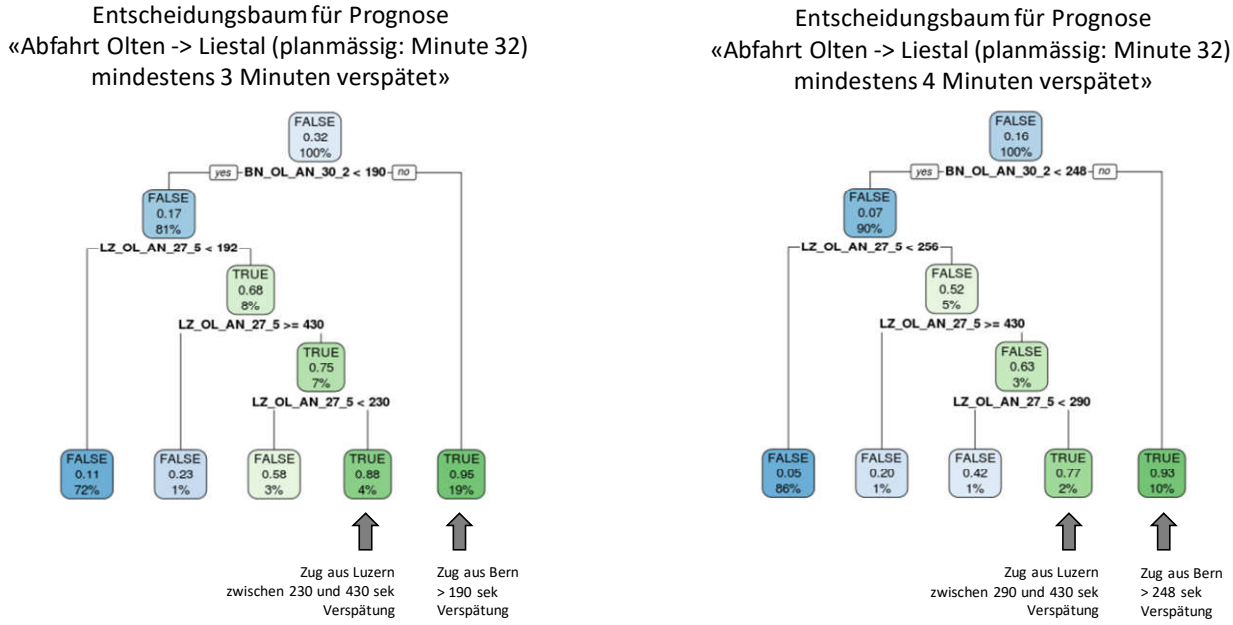


Abbildung 27: Entscheidungsbäume für die Prognose von 3- und 4-minütiger Abfahrtsverspätung in Olten.

Die relevanten Variablen sind hier dieselben wie im vorherigen Fall (Ankunft aus Bern und aus Luzern), der Zusammenhang aber etwas komplizierter: eine verspätete Ankunft aus Luzern führt hier nur in einem bestimmten Bereich (zwischen 230 bzw. 290 Sekunden und 430 Sekunden) zu einer verspäteten Abfahrt nach Liestal. Bei mehr als 430 Sekunden Verspätung wirkt sich der Luzerner Zug nicht aus. Die naheliegende Interpretation: bei kleinen Verspätungen muss der Anschluss für umsteigende Reisende nicht oder nur wenig abgewartet werden; bei mittleren Verspätungen wird er abgewartet und verspätet den Zug nach Liestal; bei grossen Verspätungen wird er nicht abgewartet. Das Verfahren hat also einen nicht-linearen (sogar: nicht-stetigen) Zusammenhang erkannt. Zugleich wird aber auch eine lineare Abhängigkeit entdeckt: Die Ankunftsverspätung des Zugs aus Bern pflanzt sich nahezu 1:1 fort: ab 130 Sekunden Ankunftsverspätung entstehen mindestens 2 Minuten (Abbildung 26), ab 190 Sekunden entstehen mindestens 3 Minuten, ab 248 Sekunden entstehen mindestens 4 Minuten Abfahrtsverspätung (Abbildung 27).

Die mittleren Zeilen in den Knoten des Baumes geben die Wahrscheinlichkeit für eine Verspätung an. Der Schwellwert, ab dem eine Verspätung prognostiziert wird, ist aber nicht notwendig 0.5. Dies ist am linken Baum in Abbildung 27 zu sehen: Hier trägt das mittlere der fünf Blätter eine Wahrscheinlichkeit von 0.58, dennoch plädiert der Baum für `FALSE` (d.h. keine Verspätung). Dies ist auf die verwendete Loss-Matrix zurückzuführen, mit der das Verhältnis von falsch-positiven zu falsch-negativen Fehlklassifikationen beeinflusst werden kann. In den Beispielen war die Vorgabe, dass falsch-positive Fälle doppelt so hoch «bestraft» werden sollen wie falsch-negative Fälle.

Die Auswahl von Prädiktoren und Splitting-Kriterien ist natürlich abhängig von den verwendeten Trainingsdaten, der Loss-Matrix und weiteren Einstellungen des Algorithmus. Dies wird in den folgenden Abschnitten noch erläutert werden. Die oben abgebildeten Entscheidungsbäume dienen lediglich der Veranschaulichung.

Aus den vorstehenden Ausführungen können Erkenntnisse über die Eignung des Verfahrens bezüglich einiger der in Abschnitt 5.1 genannten Anforderungen abgeleitet werden:

- Die **vorhandene Datenmenge** (in den Beispielen war es die Fahrplanperiode 2016) ist ausreichend, um valide Modelle zu erstellen (die Validitätsprüfung erfolgte durch Kreuzvalidierung im Rahmen des Prunings – wären die Bäume nicht valide, so wären sie auf ihre Wurzel zurückgeschnitten worden).
- Die «Bestrafung» von «False Positives» und die gezielte Herbeiführung eines **Bias** ist möglich.
- Das Verfahren macht **keine Annahmen über die Struktur des Zusammenhangs** von unabhängigen und abhängigen Variablen. Vielmehr ist es in der Lage, sowohl lineare als auch nicht-lineare und nicht-stetige Zusammenhänge zu identifizieren.
- Das Modell ist grundsätzlich in der Lage, aus mehreren Variablen (im Beispiel: 9) diejenigen zu identifizieren, die relevant sind. Ob das auch bei einer sehr grossen Variablenzahl möglich ist, kann an dieser Stelle noch nicht gesagt werden.
- Das Verfahren erzeugt Modelle, die für Menschen **verständlich und interpretierbar** sind.

Die erwünschte **Plausibilität des Kausalzusammenhangs** wird nicht durch den Algorithmus (rpart) erreicht, sondern durch die Auswahl der betrachteten Variablen: Im Szenario wurden nur Ereignisse betrachtet, die sich am gleichen Ort (Olten) und kurze Zeit vor dem Prognose-Ereignis (planmässig in den Minuten 22 bis 30) ereignen, so dass ein Zusammenhang plausibel erscheint. Wie später noch erläutert werden wird, können durch geeignete Vorauswahl der Variablen (siehe Abschnitt 5.4) sowie durch situationsspezifische Auswahl der anzuwendenden Entscheidungsbäume (siehe Kapitel 6) noch weitere Anforderungen erfüllt werden:

- Umgang mit sehr vielen Variablen
- Robustheit gegenüber Missing Values
- Eine vom Prognosezeitpunkt abhängige Auswahl der anzuwendenden Prädiktoren

Unklar verbleibt vorerst die Anforderungserfüllung hinsichtlich

- der Prognose-Performance, absolut und im Vergleich zu den Referenzsystemen,
- dem Umgang mit der in den Daten vorhandenen «Class Imbalance»,
- der Prognose-Geschwindigkeit (Effizienz).

5.4 Alternative Ansätze

Mehrere andere Machine Learning-Verfahren und Prognose-Modelle wurden im gleichen Szenario untersucht:

- Lineare Regressionsmodelle (`lm` aus Library `stats`): diese Modelle unterstellen einen durchgängig linearen Zusammenhang zwischen den Variablen. Diese Annahme ist jedoch für verschiedene Wirkungsmechanismen im Bahnbetrieb unzutreffend, wie in Kapitel 2 erläutert wurde. Es hat daher nicht überrascht, dass die erzielten Ergebnisse wenig aussagekräftig waren: neben der Ankunft aus Bern wurden meist noch mehrere weitere Variablen als signifikant identifiziert. Deren Koeffizienten im Modell waren aber stets klein, was fachlich kaum begründet werden kann: Warum sollte sich eine Ankunftsverspätung aus Zürich zu 5% auf einen anderen Zug übertragen? Entsprechend war das Bestimmtheitsmass (R^2) der Modelle durchwegs niedrig (ca. 0.5).
- Multivariate Adaptive Regression Splines (Library `Earth`): "MARS" ist ein non-parametrisches Regresssionsverfahren, das Nicht-Linearitäten und Interaktionen zwischen Variablen berücksichtigt.[97]

[97] Vgl. Friedman (1991).

Relevante Variablen werden (ähnlich wir bei `rpart`) automatisch identifiziert und ausgewählt. Das Ergebnis ist ein abschnittsweise lineares Modell: So genannte «hinge»-Funktionen sorgen dafür, dass die Steigung der Geraden sich an «Cut-Points» ändern kann.[98] Abbildung 28 visualisiert ein solches Modell: gezeigt wird der Zusammenhang der Ankunftsverspätung aus Luzern und der Abfahrtsverspätung nach Liestal. Die verschiedenen Farben repräsentieren unterschiedlich ausgeprägte Ankunftsverspätungen aus Bern. Auch hier werden Anschlussaufnahme (ab 98 Sekunden) und Anschlussbruch (zwischen 395 und 548 Sekunden) erkannt. Statt einer Unstetigkeit wie bei rpart wird hier aber ein linearer Übergang unterstellt. Dies kann so interpretiert werden, dass die Wahrscheinlichkeit für einen Anschlussbruch in diesem Intervall stetig zunimmt. Generell erschien das Verfahren sehr mächtig, aber auch anspruchsvoller in der Anwendung. Die erzeugten Modelle sind schwieriger zu interpretieren als Entscheidungsbäume. Sehr nachteilig ist, dass sich eine «biased prediction» nicht ohne Weiteres forcieren lässt.

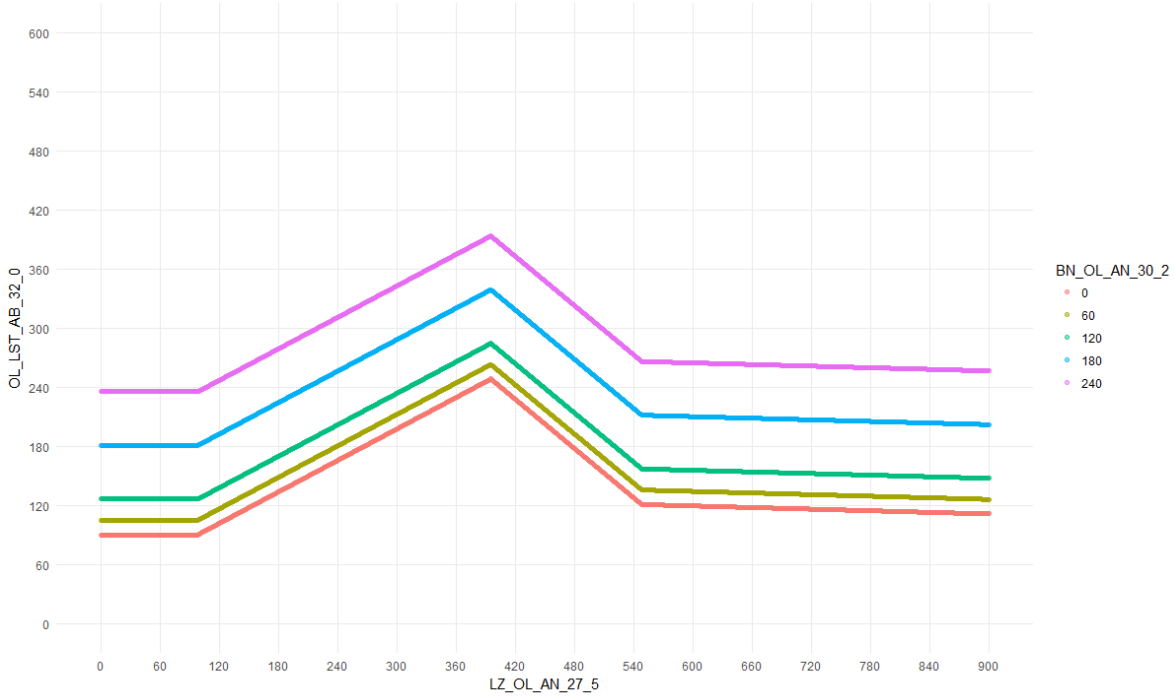


Abbildung 28: Spline-Modell für das Szenario.

- k-nearest Neighbour (`knn`): k-nearest Neigbour ist ein parameterfreies Verfahren, das sowohl für Klassifikation als auch für Regression verwendet werden kann. Beides wurde im Szenario ausprobiert. Die dabei erzeugten Prognose-Ergebnisse waren meist besser als bei Spline. Das Verfahren ist instanzbasiert, d.h. es findet keine Generealisierung zu Modellen statt. Stattdessen werden die Prädiktoren zur Laufzeit mit den gespeicherten Instanzen verglichen («lazy learning»). Die Prognose erfolgt dann auf Basis «ähnlicher» Fälle. Dieses Vorgehen impliziert mehrere Nachteile: 1. Statt einem kompakten Modell müssen alle Instanzen gespeichert werden; 2. Die Durchführung von Prognosen ist rechenintensiv und langsam; 3. es gibt keine interpretierbaren Modelle.
- Cubist (`cubist`): Cubist kombiniert diverse Ideen und Ansätze aus anderen Verfahren. Es ist ebenfalls ein nicht-parametrisches Verfahren mit automatischer Variablenselektion. Die erstellten Modelle haben eine Baum-Struktur, wobei jedes Blatt ein lineares Modell enthält. Das Verfahren verwendet einen Ensemble-Ansatz (ähnlich Boosting), bei dem mehrere Modelle («Committees») erstellt und miteinander kombiniert werden. Das Verfahren kann auch instanz-basierte Mechanis-

[98] Vgl. Kuhn / Jonson (2013), S. 146.

men einbeziehen.[99] Es erweckt somit den Eindruck einer «eierlegenden Wollmilchsau» – was es dem Anwender erschwert, eine für die jeweilige Situation angemessene Nutzung und Parametrisierung zu finden. Die erzeugten Modelle sind sehr umfangreich (siehe Abbildung 29) und kaum mehr interpretierbar. Erst recht, da in der R-Implementation keine Möglichkeit für eine angemessene Visualisierung existiert.[100]

```
Rule 9/5: [764 cases, mean 177.0, range 48 to 832, est err 31.2]

  if
      Bern > 77
      Bern <= 848
      Luzern <= 56
  then
      outcome = 9.6 + 0.93 Bern + 0.11 Luzern

Rule 9/6: [1787 cases, mean 180.9, range 46 to 576, est err 29.6]

  if
      Bern > 77
      Bern <= 380
      Luzern > 56
      Luzern <= 308
  then
      outcome = 23.3 + 0.8 Bern + 0.18 Luzern

Rule 9/7: [223 cases, mean 229.3, range 58 to 420, est err 32.2]

  if
      Bern <= 380
      Luzern > 208
      Luzern <= 308
  then
      outcome = -116.8 + 1.26 Luzern + 0.32 Bern
```

Abbildung 29: Ausschnitt aus einem von Cubist generierten Regelwerk.

Zahlreiche weitere Verfahren sowie Varianten der beschriebenen Ansätze erscheinen grundsätzlich interessant für die Problemstellung, konnten aber aus Zeitgründen nicht evaluiert werden, u.a. Regression Trees, Segmented Regression / Breakpoint Analysis und Random Forest.

Von den im Szenario untersuchten Ansätzen erschien Recursive Partitioning als besonders aussichtsreich, da damit alle untersuchten Anforderungen zufriedenstellend abgedeckt werden können, namentlich

- Keine Annahmen über strukturelle Zusammenhänge
- Gut nachvollziehbar und anschaulich
- Forcierung eines Bias ist möglich
- Schnell in der Anwendung

Die Auswahl von `rpart` für das weitere Vorgehen folgt der Prämisse, dass eine ausreichend gute Abdeckung aller Anforderungen im Hinblick auf die Forschungsfrage wesentlich wichtiger ist, als die Optimierung einzelner Kriterien auf Kosten der Nichterfüllung von anderen Punkten.

5.5 Modellierung und Auswahl der Variablen

Für Training, Validierung und Test der Prognosemodelle wird eine möglichst grosse Fallzahl benötigt. Eine Fahrplanperiode umfasst aber nur 52 oder 53 Wochen, d.h. jedes Ereignis kann maximal 371 Mal beobachtet werden. Die meisten davon weisen eine hohe Pünktlichkeit auf, so dass für grössere Ver-

[99] Vgl. Quinlan (1992), Quinlan (1993a) und Kuhn / Jonson (2013), S. 208ff.

[100] Vgl. Diskussion auf Stackoverflow «Plotting rules as a tree for Cubist package in R", https://stackoverflow.com/questions/41296038/plotting-rules-as-a-tree-for-cubist-package-in-r

spätungen kaum Beobachtungen vorhanden sind. Die Fallzahl lässt sich jedoch deutlich erhöhen, wenn man den in der Schweiz verbreiteten Taktfahrplan ausnutzt: die meisten Züge, namentlich jene in den Szenarien Olten und Bern verkehren im 1-Stunden-Takt[101]. Zwischen Mitternacht und 5:00 Uhr morgens verkehren nur weniger Züge. In den Tagesrandstunden verkehren weniger Züge, teilweise gibt es dort auch Abweichungen vom normalen Takt. Für die meisten Linien verbleiben 15-19 gleichartige Durchführungen pro Tag, so dass eine Fahrplanperiode zu den meisten Variablen zwischen 5000 und 7000 Beobachtungen enthält.

Diese Ausnutzung des Stundentakts schlägt sich auch in der Variablen-Benennung nieder, die bereits in Abschnitt 5.3 vorgestellt wurde: Eine Variable wird durch Beginn des Fahrtabschnitts, Ende des Fahrtabschnitts, Ereignis (Ankunft oder Abfahrt) und Minutenangabe gemäss Fahrplan charakterisiert. Stunde und Datum gehen nicht in den Variablennamen ein, denn diese unterscheiden lediglich Beobachtungen derselben Variable. Haltestellen werden durch ihr Kürzel bezeichnet[102], die Bestandteile des Variablennamens durch Unterstrich getrennt. Dient eine Variable als Prädiktor wird an den Namen der (fahrplanmässige) Vorlauf auf das zu prognostizierende Ereignis angehängt, bei Zielvariablen lautet die Endung «00». Abbildung 30 führt einige Beispiele auf.

Ereignis	**als Zielvariable**	**als Prädiktor für ein Ereignis um 14:01**	**als Prädiktor für ein Ereignis um 15:01**
Ankunft EC aus Bern in Olten um 13:30	BN_OL_AN_30_00	BN_OL_AN_30_31	BN_OL_AN_30_91
Ankunft RE aus Langenthal in Olten um 13:54	LTH_OL_AN_54_00	BN_OL_AN_30_7	BN_OL_AN_30_67
Ankunft IR aus Bern in Olten um 14:00	BN_OL_AN_00_00	BN_OL_AN_30_1	BN_OL_AN_30_61
Ankunft IC aus Bern in Olten um 14:02	BN_OL_AN_02_00	Nicht möglich	BN_OL_AN_02_59
Ankunft IR aus Langenthal in Olten um 14:24	LTH_OL_AN_24_00	Nicht möglich	LTH_OL_AN_24_37
Ankunft IC aus Bern in Olten um 14:30	BN_OL_AN_30_00	Nicht möglich	BN_OL_AN_30_1

Abbildung 30: Beispiele für die Benennung von Variablen im Rahmen dieser Arbeit.

Beim Auftreten von sehr grossen Verspätungen werden auf den folgenden Abschnitten häufig Ersatzzüge eingesetzt, die mit anderer Zugnummer, aber gleichem Soll-Fahrplan verkehren. In diesen Fällen tritt dasselbe «Fahrplan-Ereignis» doppelt auf. Beispiel: Der in der ersten Zeile von Abbildung 30 aufgeführte EC ist bereits in Spiez 25 Minuten verspätet. Ab Bern wird ein Ersatzzug eingesetzt. Bezogen auf ein Ereignis um 14:01 liefern beide Züge den Prädiktor `BN_OL_AN_30_31` – allerdings mit stark unterschiedlichen Werten (hohe Verspätung beim «Original», hoffentlich niedrige Verspätung beim Ersatzzug). Da eine sinnvolle Zuordnung in diesen Fällen nicht möglich ist, wird keiner der beiden Züge verwendet (die Beobachtung wird als «fehlend» taxiert, «missing value»).

[101] Aus Kundensicht gibt es oft sogar einen Halbstundentakt (teilweise sogar Viertelstundentakt). Bei genauer Betrachtung stellt sich aber häufig heraus, dass die damit verbundenen Betriebsabläufe signifikante Unterschiede aufweisen, weil sich a) Linienverläufe unterscheiden (mehrere unterschiedliche Linien mit Stundentakt überlagern sich auf einem Streckenabschnitt, z.B. ergibt sich der Halbstundentakt Bern-Zürich aus den Stundentakten Genf-Bern-Zürich-St. Gallen und Brig-Bern-Zürich-Romanshorn), b) der Rhythmus nicht exakt 30-minütig ist («hinkende Takte», z.B. Bern->Basel zur Minute 04 und 36) oder c) Anschlussbeziehungen und Fahrwegkonflikte nicht gleich sind (z.B. Anschluss von Zofingen in Bern nur zur vollen, nicht zur halben Stunde).

[102] Dies funktioniert nur im Schweizerischen Zugsverkehr. Bei anderen Verkehrsmitteln oder im Ausland müsste stattdessen der UIC-Code verwendet werden, der jedoch weniger intuitiv verständlich ist, vgl. Abschnitt 4.3.

Ergänzend zu Zugs-Verspätungen könnten als Prädiktoren auch weitere «Features» dienen, die sich aus den verfügbaren Daten ableiten lassen, z.B. Jahreszeit, Wochentag, Tageszeit, Verspätungsaufbau auf einem Abschnitt / in einer Haltestelle. Diese Möglichkeiten wurden rudimentär im Rahmen der beschriebenen Exploration untersucht. Sie erwiesen sich meist nicht als relevant (d.h. sie wurden bei der Variablen-Selektion sehr selten berücksichtigt bzw. zeigten niedrige Signifikanz in den Regressionsmodellen), so dass diese Ideen nicht weiterverfolgt wurden. Ebenfalls nicht weiterverfolgt wurde die Idee, dass die Verspätung eines Zuges als Prädiktor für die folgende Taktlage (also den 60 Minuten später verkehrenden Zug der gleichen Linie) dienen könnte.[103] Damit ist nicht ausgeschlossen, dass diese Ansätze bei elaborierterer Anwendung oder in anderen Szenarien von Nutzen sein können.

5.6 Prognosemodelle für unterschiedliche Konstellationen und Vorlaufzeiten

Die in Abschnitt 5.3 vorgestellten Entscheidungsbäume verwendeten für die Vorhersage der Abfahrtsverspätung nach Liestal die beiden Prädiktoren `BN_OL_AN_30_2` und `LZ_OL_AN_27_5`. Wie an den Variablennamen zu erkennen ist, treten die damit verbundenen Ereignisse planmässig 2 bzw. 5 Minuten vor dem Prognose-Ereignis auf, d.h. es sind nur sehr kurzfristige Prognosen möglich: Zur Minute 28 liegt der Wert von `BN_OL_AN_30_2` meist noch nicht vor, die vorhandenen Modelle sind also nicht anwendbar. Um zu diesem Zeitpunkt dennoch eine Prognose abgeben zu können, müssen Modelle erstellt werden, die ohne `BN_OL_AN_30_2` auskommen, aber dennoch eine valide Vorhersage für die Zielvariable liefern. Möglichkeiten dafür sind

1. Entfernen der Variable: es wird ein Entscheidungsbaum erstellt, der nur `LZ_OL_AN_27_5` enthält. Dies ist nur möglich, falls andere Prädiktoren im ursprünglichen Modell vorhanden sind.
2. Ersetzen der Variablen: an Stelle von `BN_OL_AN_30_2` werden eine oder mehrere Ereignisse ins Modell aufgenommen, die zeitlich früher liegen und für die plausibel ist, dass sie ihrerseits als Prädiktoren für `BN_OL_AN_30_2` dienen können. Naheliegend ist z.B. die Verwendung von `BN_OL_AB_04_28`: es darf vermutet werden, dass die Abfahrtsverspätung des Zuges in Bern hohen Einfluss auf seine Ankunftsverspätung in Olten hat (und somit dann auch auf die Abfahrtsverspätung nach Liestal).

In beiden Fällen ist zu prüfen, ob die neuen Modelle valide sind. Bei Verwendung von `rpart` erfolgt dies implizit im Rahmen des Pruning: Sollte sich bei der Kreuzvalidierung herausstellen, dass keine der erzeugten Entscheidungsregeln valide ist, so wird er auf seine Wurzel zurückgeschnitten.

Nacheinander wird für jeden Prädiktor versucht, ob sich durch Entfernen und / oder Ersetzen neue valide Modelle erzeugen lassen – unabhängig davon, wie deren planmässige Abfolge ist (denn die tatsächliche Abfolge kann vom Plan abweichen). Im Beispiel würden also auch Modelle betrachtet, in denen `BN_OL_AN_30_2` beibehalten und `LZ_OL_AN_27_5` entfernt oder ersetzt wird.

Für den Ersatz erwies sich im Rahmen von diversen Versuchen die Betrachtung folgender Kandidaten als sinnvoll:

a) Falls es sich beim zu ersetzenden Prädiktor um ein *Ankunfts*-Ereignis handelt:
   - Die vorhergehende Abfahrt desselben Zugs (wie bereits oben erwähnt).
   - Die Abfahrt von anderen Zügen, die zu einer *ähnlichen* Zeit an der Haltestelle ankommen, im Beispiel also die Abfahrten von allen Zügen, die *ungefähr* zur Minute 30 in Olten ankommen. Die fachliche Begründung ist, dass sich Züge, die zu einer ähnlichen Zeit am gleichen Ort ankommen sollen, gegenseitig beeinflussen können, weil sie möglicherweise nacheinander dieselben Gleise befahren. Als Operationalisierung von *ungefähr* wird hier das Intervall von 10

[103] Ähnlich wie bei der «Yesterday's Weather»-Methode geht dies von der Annahme aus, dass Ursachen einer Verspätung (z.B. Baustellen, Störungen, Passagieraufkommen, Witterung) mit einer gewissen Wahrscheinlichkeit auch eine Stunde später noch vorhanden sind.

Minuten vor bis 2 Minuten nach dem zu ersetzenden Prädiktor gewählt, bezogen auf die planmässige Ankunft.[104]

b) Falls es sich beim zu ersetzenden Prädiktor um ein *Abfahrts*-Ereignis handelt:

- Alle Ankünfte, die 10 Minuten vor bis 2 Minuten nach dem zu ersetzenden Prädiktor geplant sind. Auch hier kann die Beeinflussung in Fahrwegkonflikten begründet sein. Zusätzlich kommen Anschlussbeziehungen als Ursache in Betracht.

Das Verfahren wird abermals auf die neu erstellten Modelle angewendet, so dass sich sukzessive weitere Modelle mit tendenziell zunehmendem zeitlichen Vorlauf auf das Prognose-Ereignis ergeben.[105] Das Verfahren bricht ab, wenn keine validen neuen Modelle mehr erstellt werden können oder eine gegebene Maximalzahl von Modellen erreicht ist. Abbildung 31 illustriert das Vorgehen am bekannten Beispiel: Entfernen-Vorgänge sind rot, Ersatz-Vorgänge sind grün dargestellt. Die beim Ersetzen neu aufgenommenen Variablen sind grün unterstrichen. Jede Zeile der Abbildung steht für einen Entscheidungsbaum.

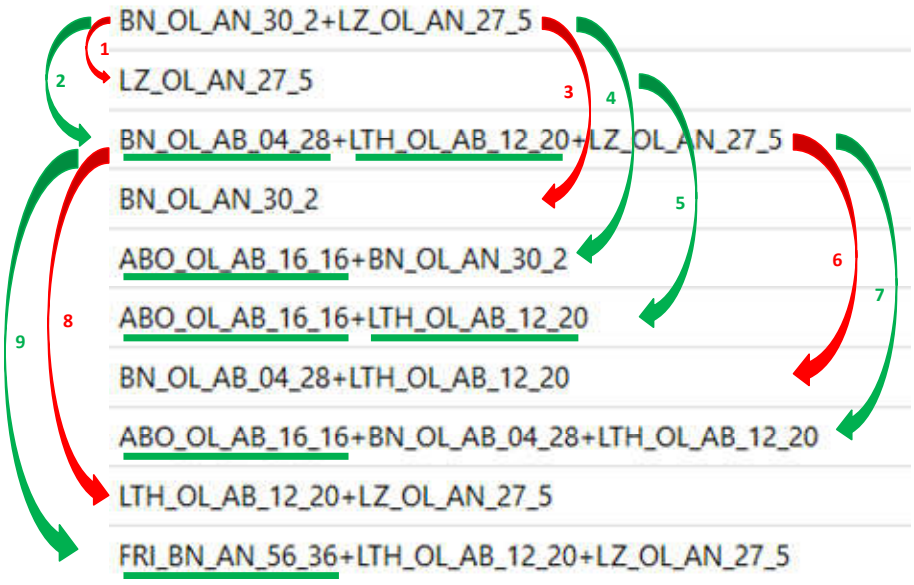


Abbildung 31: Sukzessive Generierung von Prognosemodellen für die Zielvariable OL_LST_AB_32_00.

Schritt 1: `BN_OL_AN_30_2` aus dem ersten Modell wird ersatzlos entfernt.

Schritt 2: `BN_OL_AN_30_2` aus dem ersten Modell wird durch zwei andere Variablen ersetzt.

Schritt 3: `LZ_OL_AN_27_5` aus dem ersten Modell wird ersatzlos entfernt.

Schritt 4: `LZ_OL_AN_27_5` aus dem ersten Modell wird durch eine andere Variable ersetzt.

Schritt 5: `LZ_OL_AN_27_5` aus dem zweiten Modell wird durch zwei andere Variablen ersetzt (ein ersatzloses Entfernen ist hier nicht möglich, da kein weiterer Prädiktor existiert).

[104] Dass auch Prädiktoren in Betracht gezogen werden, die zeitlich nachfolgend sind, mag zunächst paradox erscheinen. Grund dafür ist, dass sich die verwendeten Zeiten auf den Fahrplan beziehen: Tatsächlich kann sich ein Ereignis B, das *planmässig* 2 Minuten nach A auftreten soll, sehr wohl vor A ereignen und somit auch einen Einfluss auf den Zeitpunkt von A haben.

[105] Das Verfahren erinnert an «Event-Activity-Graphen». Allerdings bestehen wichtige Unterschieden gegenüber den in Kapitel 2 erwähnten Verfahren: a) der Aufbau des Graphen erfolgt «rückwärts», ausgehend vom zu prognostizierenden Ereignis, b) die «Kanten» werden nicht mit Prozesszeiten belegt, sondern mit Entscheidungsbäumen, b) «Kanten» werden nur angelegt, wenn sich der Zusammenhang als valide herausstellt, c) eine «Kante» kann mehrere Vorgänger-Ereignisse haben (ist also nicht wirklich ein Kante im Sinne der Graphentheorie), d) da auch Ereignisse als «Vorgänger» in Betracht kommen, die gemäss Fahrplan zeitlich leicht zurückliegen, können Zyklen auftreten.

Schritt 6: `LZ_OL_AN_27_5` aus dem dritten Modell wird ersatzlos entfernt.

Schritt 7: `LZ_OL_AN_27_5` aus dem dritten Modell wird durch eine andere Variable ersetzt.

Schritt 8: `BN_OL_AB_04_28` aus dem dritten Modell wird ersatzlos entfernt.

Schritt 9: `BN_OL_AB_04_28` aus dem dritten Modell wird durch eine andere Variable ersetzt.
Dass dieses Modell zur Anwendung kommt, ist möglich, aber unwahrscheinlich, weil die referenzierten Ereignisse gemäss Fahrplan grossen zeitlichen Abstand haben: In den meisten Fällen wird es so sein, dass entweder LZ_OL_AN_27_5 nicht vorliegt oder aber FRI_BN_AN_56_36 nicht benötigt wird, weil bereits «aussagekräftigere» Nachfolge-Ereignisse bekannt sind.[106]

Das Verfahren ist nachfolgend in Pseudocode beschrieben:

```
modelle_erstellen <- function() {
  Lege Trainingsset an mit Werten der Zielvariablen
  Lege Modelliste an bestehend aus "Dummy-Eintrag": Zielvariable ist Prädiktor
  Bis Ende der Modelliste (oder anderes Abbruchkriterium) erreicht:
    Wähle das nächste Modell der Modelliste
    Lege Trainingsset an mit Werten von Zielvariable und allen Prädiktoren
    Für alle Prädiktoren des Modells:
      Falls es nicht der einzige Prädiktor ist:
        Trainiere neues Modell ohne den Prädiktor
          (inkl. Kreuzvalidierung und Pruning)
        Falls valides Modell gefunden, das noch nicht in der Modelliste enthalten:
          Füge es am Ende der Modelliste hinzu
      Ermittle alle Ersatzkandidaten für den Prädiktor
        Erweitere das Trainingsset um die Werte dieser Kandidaten
        Trainiere neues Modell ohne den Prädiktor und mit den Kandidaten
          (inkl. Kreuzvalidierung und Pruning)
        Falls valides Modell gefunden, das noch nicht in der Modelliste enthalten:
          Füge es am Ende der Modelliste hinzu
  Gib die Modelliste zurück
}
```

Abbildung 32: Pseudocode: Generierung von Prognosemodellen

5.7 Parametrisierung der Modellerstellung

Das vorstehende beschriebene Verfahren wurde so implementiert, dass bei der Modellerstellung folgende Parameter[107] angegeben werden können:

- Die **Zielvariable** (`zf`) der zu erstellenden Entscheidungsbäume.
- Die Menge der **Betriebstage** (`betriebstage`), die für das Training verwendet werden sollen. Dies ermöglicht einerseits, Modelle für spezifische Zeitintervalle zu erstellen (z.B. nur Fahrplanperiode 2017). Andererseits kann so auch eine Hold-Out-Validierung durchgeführt werden. Ein 80/20-Splitting der Fahrplanperiode 2017 kann in R z.B. erzielt werden mit

[106] Kapitel 6 beschreibt das Verfahren zur Auswahl der anzuwendenden Modelle in einer konkreten Prognose-Situation. Tatsächlich stellen grosse Zeitunterschiede zwischen den Prädiktoren einen möglichen Ansatzpunkt zur Reduzierung der Modellzahl und somit zur Optimierung des Verfahrens dar.

[107] Es handelt sich hier um die Parameter des Trainings-Skripts. Diese sind nicht zu verwechseln mit den Modell-Parametern. Häufig spricht man bei Parametern, die nicht aus den Daten «gelernt», sondern zur Beeinflussung des Trainingsprozesses festgelegt werden, auch von «Hyper-Parametern» oder «Tuning Parametern», vgl. Brownlee (2017). Die Aufzählung hier ist noch etwas weiter gefasst und enthält auch Vorgaben zu Zielvariablen und Trainings-Sample.

```
alle_tage <- seq.Date(as.Date('2016-12-10'), as.Date('2017-12-09'), by=1)
in_train <- sample(alle_tage, size = 0.8*length(alle_tage))
```

- Die Liste der **Verspätungsniveaus** (`vniveaus`), zu denen Entscheidungsbäume erstellt werden sollen. Im Rahmen der Arbeit wurden meist 60-Sekunden-Schritte gewählt im Intervall 60 Sekunden bis 900 Sekunden.
- Der **Loss-Koeffizient** (`lossfaktor`): Dieser gibt an, wie stark False Positive-Klassifikationen im Verhältnis zu False Negative-Klassifikationen bestraft werden sollen. Die verwendete rpart-Implementierung erwartet hierzu eine Loss-Matrix, die aus dem Skalar-Parameter `lossfaktor` erzeugt werden kann mit `matrix(c(0,lossfaktor,1,0), byrow=TRUE, nrow=2)`. Verschiedene Werte für den Loss-Koeffizienten wurden sowohl in Simulationen als auch in Echtzeit-Anwendung erprobt. Mit einem Wert von 14 konnte die Häufigkeit von False-Positives im Szenario Bern auf das Niveau der Referenzsysteme gebracht werden. Dies wird in Abschnitt 6.1 noch näher erläutert.
- Die **Höchstgrenze** (`maxiter`), bis zu der weitere Modelle durch Weglassen oder Ersetzen von Prädiktoren erstellt werden. Dies soll sicherstellen, dass die Modellerstellung in vernünftiger Zeit terminiert, auch wenn noch sehr viel mehr Modelle generiert werden könnten. Häufig wurde eine Grenze von 1000 Modellen verwendet. Das Erreichen dieser Grenze ist in der Regel ein Indikator dafür, dass es zu einer «Explosion» der Modelliste kommt, weil andere Parameter der Modellerstellung ungünstig gewählt sind.
- Ob beim Pruning statt des Standardverfahrens die **1SE-Regel** (`prune1SE`) angewendet werden soll: Das Pruning verwendet einen Komplexität-Parameter `cp`, bei dessen Unterschreiten Äste des Entscheidungsbaums «abgeschnitten» werden. Im Standard-Verfahren wird diejenige Ausprägung von `cp` verwendet, die in der Kreuzvalidierung zu einem minimalen Fehler führt. Tatsächlich ist es jedoch häufig so, dass die Fehlerkurve im Bereich ihres Minimums sehr flach verläuft, so dass die Auswahl von `cp` als beinahe zufällig angesehen werden kann – was potentiell zum Einschluss «zufälliger» Entscheidungsregeln führt. Ein alternatives Kriterium ist die so genannte «1SE-Rule», die `cp` um einen Standardfehler höher ansetzt – und somit in den weniger «zufälligen» Verlauf der Fehlerkurve. Tendenziell sind die so gestutzten Entscheidungsbäume etwas kleiner und weisen höhere Validität auf (um den Preis einer leicht reduzierten Vorhersagegenauigkeit). Das Risiko für «Overfitting» wird kleiner.[108] Abbildung 34 zeigt den R-Code zur Ermittlung des cp-Werts nach Standard-Verfahren und nach 1SE-Rule. Die in den folgenden Kapiteln verwendeten Modelle wurden mit der 1SE-Rule erstellt.

```
if(prune1SE) {
  cptable <- as.data.table(m$cptable)
  cut_xerror <- cptable[which.min(m$cptable[,"xerror"]), xerror+xstd]
  cpprune <- cptable[xerror<cut_xerror][1]$CP
} else {
  cpprune <- m$cptable[which.min(m$cptable[,"xerror"]),"CP"]
}
if(!is.na(cpprune)) m <- prune(m, cp= cpprune)
```

Abbildung 33: R-Code zur Ermittlung des Komplexitätswerts, Standardverfahren und 1SE-Rule.

- Die **maximale Tiefe** (`maxtiefe`) der Entscheidungsbäume. Hierbei handelt es sich um einen «Pre-Pruning-Parameter» des rpart-Verfahrens: während beim eigentlichen Pruning (teils auch als «Post-Pruning» bezeichnet) bereits erzeugte Bäume «zurückgeschnitten» werden, steuert dieser Parameter, bis zu welcher Grösse Bäume initial erstellt werden. Der Parameter beeinflusst die Komplexität

[108] Vgl. Therneau / Atkinson (2018), S. 14.

der erstellten Modelle und damit indirekt auch die Zahl der verwendeten Prädiktoren – und letztlich auch die Zahl der erstellten Modelle.

- Die **maximale Zahl** (`maxpreds`) der verwendeten Prädiktor-Variablen. Dieser Parameter dient ebenfalls dem Pre-Pruning und hat eine ähnliche Wirkung wie die maximale Baum-Tiefe.
- Die **minimale Knotengrösse**, die im Trainingsset auf ein Blatt des Entscheidungsbaums entfallen muss. Diese kann entweder als absolute Zahl (`minbuck`) oder als Anteil (`minbuckratio`) angegeben werden. Letzterer bezieht sich auf die Anzahl der positiven Ausprägungen der Zielvariable im Trainingsset.[109] Indem sie Grösse und Komplexität der erstellten Bäume beeinflussen, handelt es sich ebenfalls um Pre-Pruning-Parameter. Sie richten das Augenmerk auf die Vermeidung eines Overfitting, weil damit Regeln vermieden werden, die sich auf eine geringe Fallzahl stützen.

Auch für die Pre-Pruning-Parameter wurden diverse Versuche in Simulationsrechnungen und Echtzeit-Anwendung durchgeführt. Ziel war es, die Menge und Komplexität der Modelle so weit zu begrenzen, dass eine stabile Echtzeit-Prognose für das Szenario Bern im Minutentakt möglich ist. Gewählt wurden schliesslich folgende Werte: maximale Baumtiefe = 10, maximale Zahl von Prädiktoren = 10, minimale Knotengrösse für Split = 5% aller positiven Ausprägungen.

Innerhalb des Codes wurden weiterhin folgende Festlegungen implementiert:

- Es werden nur Modelle generiert, wenn im Trainingsset in mindestens 1 Promille der Fälle Verspätungen des betrachteten Niveaus vorhanden sind. Dies soll verhindern, dass Prognosemodelle angelegt werden für Verspätungssituationen, die in der Vergangenheit kaum aufgetreten sind.
- Der maximal betrachtete zeitliche Vorlauf auf ein Prognose-Ereignis sind 90 Minuten (massgeblich ist hier wieder die zeitliche Distanz der Ereignisse gemäss Fahrplan).

5.8 Training der Modelle

Zur Erstellung der einzelnen Modelle wird die Funktion `rpart()` aufgerufen. Diese erstellt einen Entscheidungsbaum und führt auch sogleich ein Pruning (auf Basis von Kreuzvalidierung) durch. Zu beachten ist dabei, dass das erstellte Modellobjekt nach wie vor den gesamten Entscheidungsbaum (vor Pruning) enthält. Darin sind möglicherweise noch Prädiktor-Variablen enthalten, die nach Pruning gar nicht mehr benötigt werden. Dies führt zu Schwierigkeiten, wenn das reduzierte Modell später angewendet werden soll: beim Aufruf von `predict()` werden möglicherweise fehlende Variablen reklamiert, die im reduzierten Modell gar nicht benötigt werden. Um dies zu verhindern, wird beim Training mehrstufig vorgegangen:[110]

1. Erstellen eines Modells für alle Prädiktor-Kandidaten
2. Ermitteln der benötigten Variablen nach Pruning
3. Erstellen eines Modells für die benötigten Variablen

Das sukzessive Erstellen der Modelle wurde so implementiert, dass es sowohl sequentiell (auf der AWS-Instanz) als auch parallelisiert (auf dem R-Cluster der BFH) durchgeführt werden kann. Für die Parallelisierung wurde die R-Library `doParallel` verwendet. Damit kann die Berechnung für unterschiedliche Zielvariablen, Verspätungs-Niveaus und Loss- Koeffizienten auf mehrere Knoten verteilt werden. Folgende begleitende Massnahmen waren notwendig:

[109] Die positiven Ausprägungen der Zielvariable repräsentieren die verspäteten Fälle, es handelt sich in der Regel um die kleinere der beiden Klassen. Beispiel: Bei 5000 Beobachtungen, von denen 500 eine Verspätung repräsentieren, führt eine «minbuckratio» von 0.05 zu einer Mindestzahl von 25 Fällen pro Blatt des Entscheidungsbaums.

[110] Vgl. http://r.789695.n4.nabble.com/rpart-package-why-does-predict-rpart-require-values-for-quot-unused-quot-predictors-td4638777.html.

- Der Stand der Berechnungen der einzelnen Knoten wird in eine gemeinsame Datei geloggt, so dass der Fortschritt leicht nachvollzogen werden kann. Dafür wird die Library `futile.logger` verwendet.
- Die benötigten Daten (Fahrten und Takte) werden statt aus der Datenbank aus einer vorbereiteten Datei gelesen. Dieses Verfahren wurde gewählt, da sich der gleichzeitige Zugriff zahlreicher Cores auf dieselbe (entfernte) Datenbank als Engpass herausgestellt hat, der grosse Teile des Parallelisierungs-Vorteils zunichtemacht.
- Um ein gutes Verhältnis von Rechenzeit zu Parallelisierungs-Overhead zu erreichen, können in einem einzelnen Funktions-Aufruf Modelle für mehrere Verspätungs-Niveaus errechnet werden.

Die vom `rpart()` erzeugten Objekte benötigen viel Speicher, weil darin ausser den erzeugten Modellen auch die Aufrufparameter sowohl die verwendeten «Environments» als auch die Trainingsdaten abgelegt werden. Um eine effiziente Verwaltung der grossen Zahl benötigter Modelle zu ermöglichen, werden einige dieser «Slots» explizit auf `null` gesetzt.

Für die Holdout-Validierung kann ein zufälliges Sample von Betriebstagen erstellt werden (Abbildung 34). Es werden dann nur diejenigen Betriebstage zum Training der Modelle verwendet, die im Sample enthalten sind – die übrigen dienen später dem Test des Modells.

```
alle_betriebstage <-
  seq.Date(from=as.Date('2016-12-10'), to=as.Date('2017-12-09'), by=1)
in_train <- sample(alle_betriebstage, size = 0.8*length(alle_betriebstage))
```

Abbildung 34: Bilden eines Trainings-Samples aus den Betriebstagen der Fahrplanperiode 2017.

Der vollständige Source-Code für das Training der Modelle ist in den beigefügten Dateien `Fahrten_vorbereiten.R`, `Modelle_erstellen.R` und `Aufruf_Modellerstellung.R` enthalten.

6 Anwendung des Prognoseverfahrens

Die Erstellung eines Prognoseverfahrens ist die eine Seite der Medaille – seine Anwendung die andere. Erst beide Teile zusammen ermöglichen es, Verspätungen vorherzusagen. Und erst bei Vorliegen beider Teile kann beurteilt werden wie gut es gelingt. In diesem Kapitel geht es um die Anwendung der Modelle. Zunächst werden Performance-Kriterien hergeleitet, mit denen das verwendete Verfahren bewertet und mit den Referenzsystemen der Bahnen verglichen werden kann. Abschnitt 6.2 beschreibt, wie in einer konkreten Prognosesituation aus den zahlreich generierten Entscheidungsbäumen die anzuwendenden ausgewählt und die Ergebnisse miteinander verknüpft werden. Schon mehrfach ist darauf hingewiesen worden, dass die verwendeten Verspätungsdaten teilweise sekundengenau, teilweise nur minutengenau vorliegen. Daraus resultiert die Frage, wie mit Rundungsdifferenzen umzugehen ist. Dies wird in Abschnitt 6.3 behandelt. Die letzten beiden Abschnitte beschreiben, wie das Verfahren für Simulationen (6.4) und bei der Echtzeitprognose (6.5) angewendet wird und gehen dabei auch auf Implementierungs-Aspekte ein.

6.1 Performance-Kriterien

Es wurde bereits mehrfach erwähnt, dass bestimmte Parameter «durch Ausprobieren» festgelegt wurden oder dass sich bestimmte Vorgehensweisen «bewährt» haben. Ich habe noch nicht erklärt, welche Beurteilungskriterien dabei zu Grunde gelegt wurden. Die Frage, wie die Qualität eines Prognoseverfahrens zu beurteilen ist, soll hier erörtert werden. Sie ist nicht nur für das Parameter-Tuning und die Auswahl von Design-Varianten relevant, sondern wird auch dazu dienen, das hier entwickelte Verfahren mit den Prognosesystemen der Bahnunternehmen vergleichen zu können.

Die Zielsetzung dieser Arbeit nennt drei unterschiedliche Qualitätsaspekte:

1. Wie häufig kann eine Verspätung im Sinne der Überschreitung eines gegebenen Grenzwerts (z.B. mehr als 5 Minuten) korrekt vorhergesagt werden?
2. Wie genau kann das Ausmass (Anzahl Minuten) einer Verspätung vorhergesagt werden?
3. Mit welchem zeitlichen Vorlauf können diese Vorhersagen erfolgen?

Im Sinne von Punkt 1 kann eine Prognose als korrekt angesehen werden, wenn eine vorhergesagte Überschreitung des Grenzwerts tatsächlich eingetreten ist oder wenn eine vorhergesagte Unterschreitung des Grenzwerts tatsächlich eingetreten ist. Dabei ist der Grenzwert eine ganzzahlige Minutenangabe, die (ab dem Folgetag) mit der sekundengenauen tatsächlichen Verspätung verglichen werden kann. Wie bereits diskutiert wurde, sind dabei zwei Fehlertypen («false positive» und «false negative») zu unterscheiden (vgl. Abbildung 35): Die Rate der «false positives» (Ausfallrate, das Verhältnis von falsch erkannten Verspätungen zu allen Verspätungen) soll sich auf gleichem Niveau wie bei den Referenzsystemen bewegen. Zugleich soll der Anteil korrekter Prognosen (Accuracy) maximiert werden.

Gegeben: Grenzwert (z.B. 5 Minuten)	**Tatsächlich:** Verspätung > Grenzwert	**Tatsächlich:** Verspätung <= Grenzwert
Prognose: Grenzwert überschritten	Korrekte Prognose	False Positive (auf Niveau der Referenzsysteme halten!)
Prognose: Grenzwert nicht überschritten	False Negative	Korrekte Prognose

Abbildung 35: «Confusion Matrix» für die Vorhersage von Mindest-Verspätungen

Im Sinne von Punkt 2 der Zielsetzung kann eine Prognose als korrekt angesehen werden, wenn die Differenz zwischen Prognosewert und tatsächlichem Wert eine gegebene Toleranz nicht überschrei-

tet.[111] Da Verspätungen hier minutengenau vorhergesagt werden, können selbst bei einer «perfekten» Prognose Abweichungen von bis zu +/- 30 Sekunden auftreten. Es erscheint plausibel, weitere +/- 30 Sekunden als «tolerierbaren Fehler» zuzulassen, so dass gemäss Abbildung 36 bewertet werden kann:

Prognostiziert < Tatsächlich – 60 sek	Tatsächlich – 60 sek <= Prognostiziert <= Tatsächlich + 60 sek	Prognostiziert > Tatsächlich + 60 sek
Verspätung unterschätzt	Korrekte Prognose	Verspätung überschätzt (auf Niveau der Referenzsysteme halten!)

Abbildung 36: Bewertung einer Vorhersage zum Ausmass der Verspätung

Punkt 3 der Zielsetzung ruft in Erinnerung, dass die Qualität von Vorhersagen in der Regel stark davon abhängt, wieviel Zeit zwischen Prognose und prognostiziertem Ereignis (hier: tatsächliche Ankunft bzw. Abfahrt eines Zuges) liegt. Bei einem Vergleich von zwei Prognoseverfahren wäre etwa zu fragen, in welchem Vorlauf-Intervall die Genauigkeit (Accuracy) des einen Verfahrens jene des anderen übertrifft – wobei wiederum die Ausfallrate / der Anteil Überschätzungen auf fixem Niveau zu halten ist.

In jedem Fall ist eine Kalibrierung an die Qualität der bei den Bahnen eingesetzten Systeme vorzunehmen. Dabei ist festzustellen, dass die beiden Kriterien «Ausfallrate» und «Anteil Überschätzungen» zwar ähnlich, aber nicht identisch sind (vgl. Abbildung 39).

	Ausfallrate («False Positive rate»)	Anteil Überschätzungen
Verursachendes Ereignis (Zähler)	Prognose einer Verspätung, die nicht auftritt	Überschätzung einer Verspätung
Bezugsgrösse (Nenner)	Alle Fälle ohne Verspätung	Alle Fälle (mit und ohne Verspätung)
Toleranzwert	Nein	Ja, hier auf +/-60 sek festgelegt
Parameter	Grenzwert, ab dem ein Ereignis als verspätet gilt	Keiner

Abbildung 37: Unterschiede zwischen Ausfallrate und Anteil Überschätzungen

Da hier ein einziges, einheitlich parametrisiertes Verfahren verwendet werden soll, muss eine Auswahl zwischen den beiden Kalibrierungskriterien getroffen werden. Die Wahl fällt auf den Anteil Überschätzungen, weil dieser unabhängig von einem Parameter (Grenzwert) festgelegt werden kann.[112]

Abbildung 38 zeigt den Anteil Überschätzungen, wie er bei den Referenzsystemen im Zeitraum 17. Januar bis 27. Februar (=6 Wochen) im Szenario Bern beobachtet wurde. Grundlage sind über 27'000 Ankünfte (ca. 650 pro Tag) in Bern, für die die Prognose im Minutentakt festgehalten wurde. Verwendet wird das Intervall von 60 Minuten bis 1 Minute vor tatsächlicher Ankunft. Von den mehr als 1.6 Mio untersuchten Prognosen entfiel ein Anteil von 0.05% auf Überschätzungen. Dieser Überschätzungsanteil dient als Anhaltspunkt bei der Festlegung der Loss-Matrix (vgl. Abschnitt 5.7).[113]

[111] Alternativ könnte zur Bewertung auch ein Distanzmass, z.B. mittlerer quadratischer Fehler oder mittlerer absoluter Fehler verwendet werden. Da die Prognosegranularität (ganze Minuten) relativ grob ist und zugleich ein Grossteil der Verspätungen im niedrigen einstelligen Minutenbereich liegt, erscheint das hier nicht sinnvoll.

[112] Hier könnte eingewendet werden, dass auch der Toleranzwert ein Parameter ist. Dieser wird aber im Rahmen der Arbeit als stabil angesehen, wohingegen für den Grenzwert unterschiedliche Ausprägungen betrachtet werden sollen.

[113] Ähnliche Untersuchungen vom November 2017 zeigten (bei deutlich kleinerer Stichprobe) einen etwas höheren Überschätzungsanteil von 0.07%, weshalb Echtzeitprognosen bis Ende Februar mit einem Loss-Koeffizienten von 10 berechnet wurden.

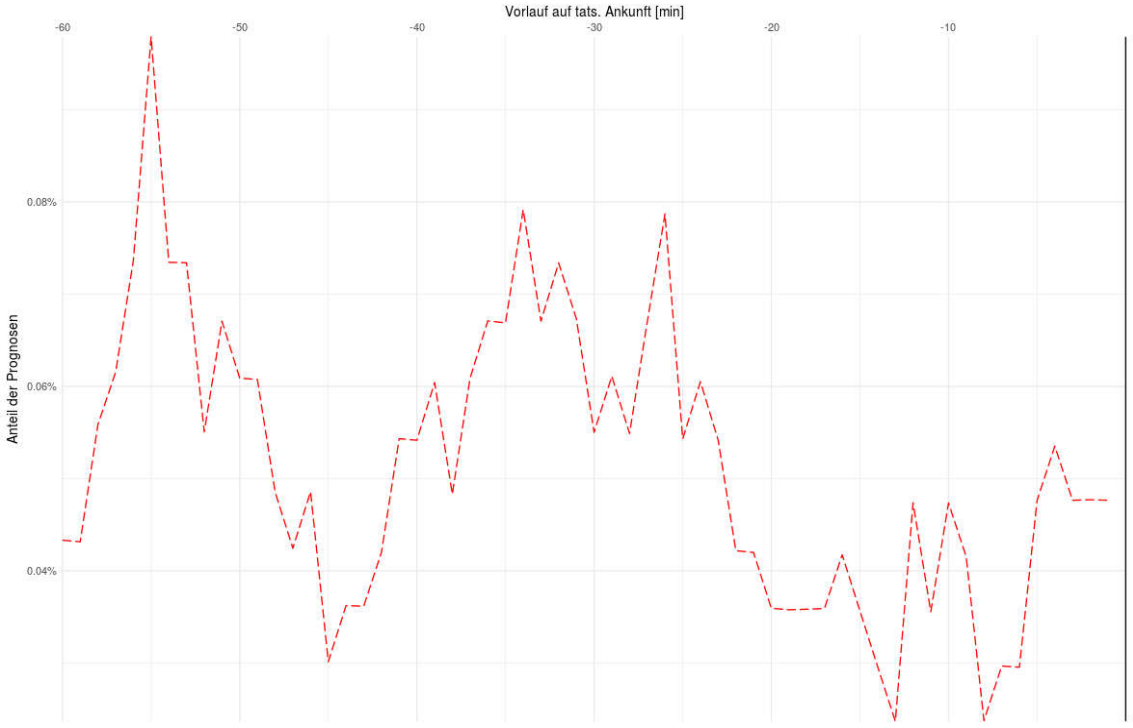


Abbildung 38: Anteil Überschätzungen bei den Referenzsystemen 60 min bis 1 min vor Ankunft in Bern

Wie in Abbildung 39 zu erkennen ist, sind diese Überschätzungen recht ungleich auf die Zugslinien verteilt. Am häufigsten treten Sie auf bei Ankünften aus Richtung Freiburg (FRI), Konolfingen (KF), Münsingen (MS), Thun (TH) und Zürich (ZUE). Daraus muss geschlossen werden, dass der Wert von 0.05% möglicherweise nicht verallgemeinerbar ist. Für das Szenario Bern bildet er den beobachteten Mittelwert über alle Zugsankünfte. Für andere Szenarien sollte eine separate Kalibrierung vorgenommen werden.

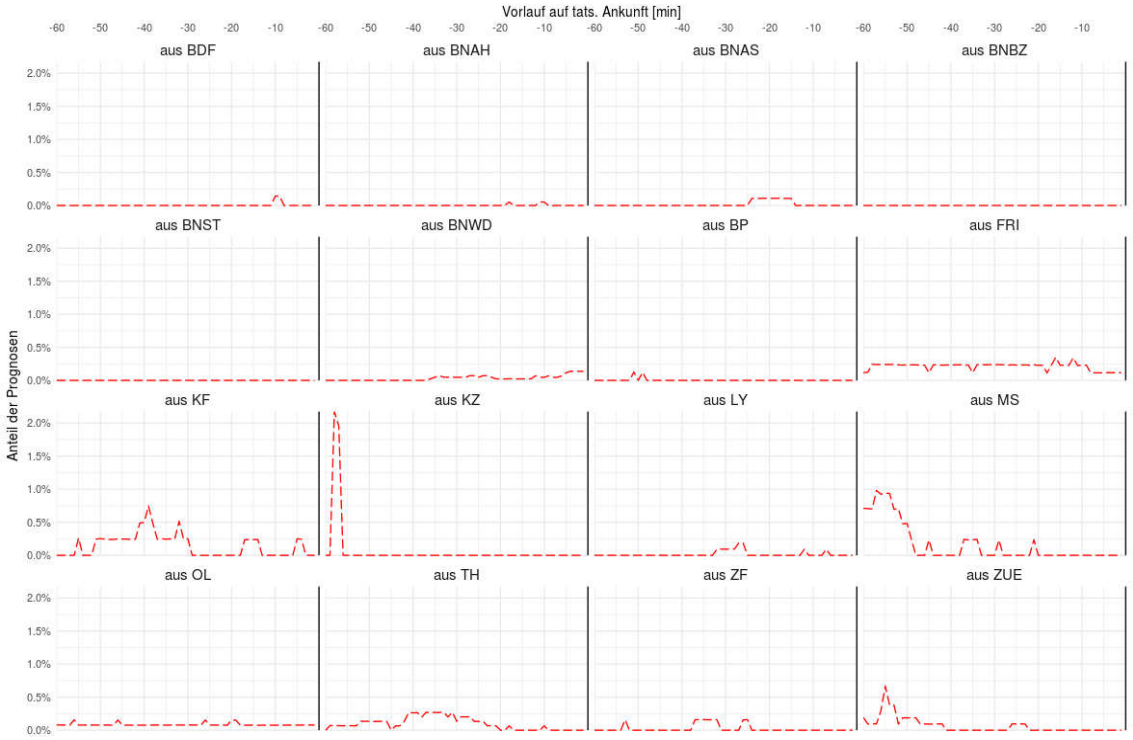


Abbildung 39: Anteil überschätzter Verspätungen bei den Referenzsystemen, nach Herkunft des Zuges

Für ein Verfahren, dass im Prognoseszenario Bern eingesetzt wird, können somit folgende Beurteilungskriterien festgelegt werden: Vorausgesetzt, dass der mittlere Anteil von Überschätzungen im Intervall von 60 Minuten bis 1 Minute vor tatsächlicher Ankunft des Zuges einen Wert von ungefähr 0.05% aufweist, wird die Performance gemessen:

1. Bei der Prognose von Mindestverspätungen zu einem gegebenen Grenzwert:
   als Anteil korrekter Prognosen an der Anzahl verspäteter Züge in einem Intervall von 60 bis 1 Minuten vor tatsächlicher Ankunft des Zuges.
2. Bei der Prognose des Verspätungs-Ausmasses:
   als Anteil korrekter Prognosen an der Gesamtzahl der Prognosen in einem Intervall von 60 bis 1 Minuten vor tatsächlicher Ankunft des Zuges. Dabei gilt eine Toleranz von +/- 60 Sekunden.
3. Bei der Beurteilung des zeitlichen Vorlaufs:
   als Zeitintervall, in dem ein Verfahren ein anderes übertrifft, gemessen an einem der beiden vorstehenden Kriterien.

Als Anhaltspunkt seien hier noch die Performance-Werte angegeben, die bei den Prognosesystemen der Bahnen im genannten Zeitraum 17. Januar bis 27. Februar beobachtet werden konnten:

Abbildung 40 zeigt die *Trefferquote* der Referenzsysteme bei der Vorhersage von Mindestverspätungen. Wie man leicht erkennt, nimmt sie mit steigendem Verspätungsniveau stetig zu. Erneut sieht man auch, dass grosse Verspätungen wesentlich seltener auftreten als kleine.

	Anzahl Verspätungen	Trefferquote Referenzsysteme
>1 min Verspätung	9078	16.2%
>2 min Verspätung	4713	18.9 %
>3 min Verspätung	2535	23.0%
>4 min Verspätung	1501	27.8%
>5 min Verspätung	985	32.9%
>6 min Verspätung	722	35.2%
>7 min Verspätung	536	39.3%
>8 min Verspätung	416	43.8%
>9 min Verspätung	349	46.3%
>10 min Verspätung	299	49.2%

Abbildung 40: Verspätungen im betrachteten Zeitraum und Trefferquote der Referenzsysteme

Das *Ausmass* der Verspätungen wurde von den Bahnen in 71.0% der Fälle korrekt vorhergesagt.[114] Dieser Wert mag sehr hoch erscheinen, beachtet werden muss aber, dass 68.4% alle Ankünfte pünktlich erfolgten – was den «Normalfall» einer konservativen Prognose darstellt. Bei den verspäteten Ankünften (ab 60 Sekunden) wurde das Ausmass dagegen nur in 5.0% der Fälle richtig erkannt.

Die Prognosequalität im Zeitverlauf ist in Abbildung 41 dargestellt – links für alle Ankünfte des betrachteten Zeitraums, rechts für jene mit mindestens 60 Sekunden Verspätungen. Sehr deutlich sieht man, dass die Accuracy erst 5-6 Minuten vor dem tatsächlichen Ankunftsereignis ansteigt. Vor diesem Zeitpunkt wird fast nie eine Verspätung gemeldet – was dank der hohen Pünktlichkeit der Bahnen ja in 68.4% der Fälle eine zutreffende Prognose ist.

[114] Toleranz +/- 60 sek; Mittel über alle Prognosen im Vorlauf von 1 bis 60 Minuten vor tatsächlicher Ankunft.

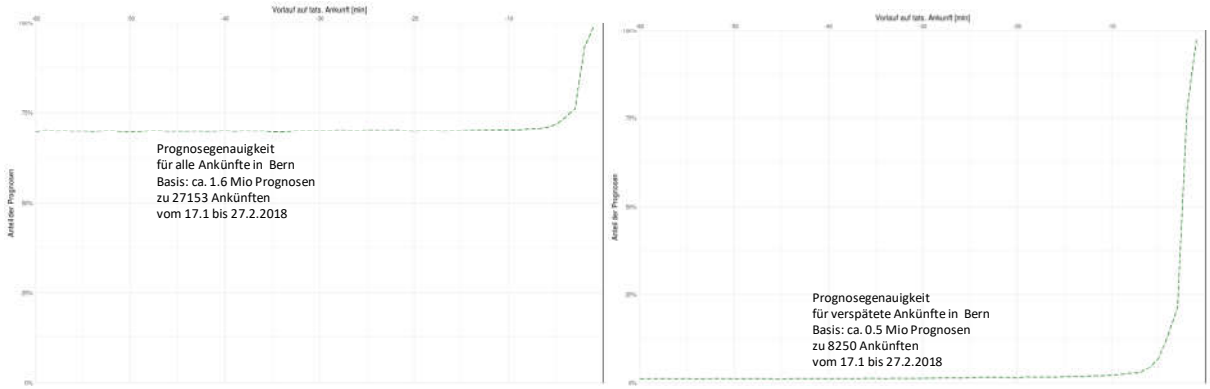


Abbildung 41: Accuracy der Referenzsysteme im Zeitverlauf, alle Ankünfte (links) und verspätete Ankünfte (rechts)

6.2 Auswahl der anzuwendenden Modelle

Um in einer konkreten Situation eine Prognose zu ermitteln, müssen aus der Vielzahl der generierten Entscheidungsbäume (Modelle) diejenigen ausgewählt werden, die zur Anwendung kommen sollen. Dies ist abhängig von den vorliegenden Prädiktoren: Liegen für ein Modell nicht alle benötigten Prädiktoren vor, weil die entsprechenden Ereignisse noch nicht eingetreten sind, so kann dieses Modell nicht für die Prognose verwendet werden. Abbildung 42 verdeutlicht dies an einem Beispiel:

Prädiktoren:
- Abfahrt Zug B in Spiez (:25)
- Ankunft Zug A in Thun (:32)

Prädiktoren:
- Abfahrt Zug B in Spiez (:25)
- Abfahrt Zug A in Spiez (:22)

Prädiktoren:
- Abfahrt Zug B in Spiez (:25)
- Ankunft Zug C in Thun (:24)

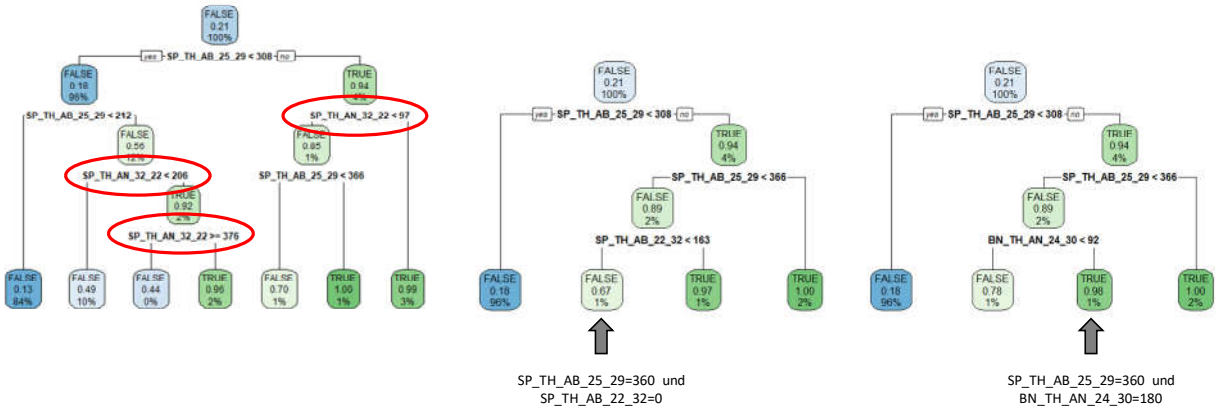


Abbildung 42: Drei Bäume für die Prognose einer 3-minütigen Ankunftsverspätung von Zug B in Bern.

Gesucht ist ein Prognose über eine mindestens 3-minütige Ankunftsverspätung des IC Spiez-Thun-Bern mit planmässiger Ankunft in Bern zur Minute :54. Dieser wird nachfolgend als Zug B bezeichnet. Seine Pünktlichkeit wird beeinflusst von Zug A, der auf der gleichen Strecke planmässig 2-3 Minuten vorausfährt. Zug B wird weiterhin beeinflusst vom entgegengesetzt fahrenden Zug C, der in Thun wenige Minuten vorher das gleiche Haltegleis (Gleis 2) benutzt. Die Abbildung zeigt 3 Entscheidungsbäume mit jeweils 2 Prädiktoren. Um 11:31 Uhr liegen folgende Informationen vor:

- Abfahrt Zug A in Spiez war pünktlich (`SP_TH_AB_22_32=0`), seine Ankunft in Thun ist noch nicht bekannt (`SP_TH_AB_32_22=NA`).
- Abfahrt Zug B in Spiez war 6 Minuten verspätet (`SP_TH_AB_25_29=360`).
- Ankunft Zug C in Thun war 3 Minuten verspätet (`BN_TH_AN_24_30=180`).

Der linke Baum ist in dieser Situation nicht anwendbar, da für die Ankunft von Zug A in Thun noch kein Wert vorliegt. Die Abfahrt von Zug B in Spiez ist dagegen bekannt, so dass der mittlere Baum

angewendet werden kann. Er weist eine 67%-ige Wahrscheinlichkeit für eine Verspätung aus. Da sehr konservativ prognostiziert werden soll (hoher Loss-Koeffizient), lautet die Vorhersage «keine Verspätung». Der rechte Baum stützt sich auf die Ankunft von Zug C in Thun, die ebenfalls bekannt ist. Das Resultat des rechten Baums lautet «Verspätung».

Die Resultate aller anwendbaren Bäume werden ODER-verknüpft: Wenn mindestens eine Konstellation entdeckt wurde, die mit ausreichender Konfidenz eine Verspätung erwarten lässt, so wird eine Verspätung prognostiziert. Im Beispiel wird die Prognose also lauten: Zug B kommt mit mindesten 3 Minuten Verspätung in Bern an. Zu prüfen sind nun noch die Modelle für höhere Verspätungsniveaus (mindestens 4 Minuten, mindestens 5 Minuten etc.).

Es kommt häufig vor, dass für eine gegebene Situation sehr viele Modelle anwendbar sind. Um die Anwendung zu beschleunigen, werden folgende Heuristiken verwendet:

- Die Modelle werden aufsteigend sortiert nach Verspätungsniveau angewendet (d.h. zunächst wird nach mindestens 1-minütigen Verspätungen gesucht, dann nach mindestens 2-minütigen Verspätungen etc.).
- Sobald eine positive Prognose für ein Niveau vorliegt, wird auf dem nächst höheren Niveau fortgefahren (aufgrund der ODER-Verknüpfung ist es in diesen Fällen nicht notwendig, weitere Kriterien zu prüfen).
- Innerhalb eines Verspätungsniveaus werden die Modelle in Reihenfolge der planmässigen Vorlaufzeit ihrer Prädiktoren angewendet: Der Einfluss von Ereignissen auf die Prognosevariable nimmt gemeinhin mit zunehmendem zeitlichen Abstand ab. Die jeweils jüngsten vorliegenden Prädiktor-Variablen haben daher tendenziell den grössten Einfluss – führen sie zu einer positiven Prognose, so können die Modelle mit «älteren» Prädiktoren übersprungen werden.
- Wenn für mehrere Verspätungsniveaus nacheinander kein einziges Modell eine positive Prognose zeitigt, so wird das Verfahren vorzeitig abgebrochen: Es ist sehr unwahrscheinlich, dass ein Zug 10 Minuten Verspätung haben wird, wenn keinerlei Hinweis auf eine 2-minütige Verspätung gefunden werden konnte. Zu betonen ist, dass es sich hier um eine Heuristik handelt: es kann aufgrund ungünstiger Konstellationen in den Trainingsdaten tatsächlich vorkommen, dass auf einem Niveau kein einziger Entscheidungsbaum «anschlägt», auf dem nächsthöheren dann aber sehr wohl. Es hat sich bewährt, dann abzubrechen, wenn in zwei aufeinanderfolgenden Niveaus keine positive Prognose aufgetreten ist.

Als prognostiziertes Ausmass der Verspätung wird das höchste Niveau verwendet, bei dem eine positive Prognose aufgetreten ist. Das Verfahren ist nachfolgend in Pseudocode beschrieben. Es wird sowohl bei der Anwendung auf Vergangenheitsdaten (Abschnitt 6.3) als auch bei Echtzeit-Prognose (Abschnitt 6.5) verwendet.

```
berechne_prognosen <- function(modelle, praediktoren) {
  Für jede Zielvariable:
    Für jedes Modell, sortiert nach 1) Verspätungsniveau, 2) Vorlaufzeit:
      Falls alle Prädiktoren des Modells vorliegen:
        Wende das Modell an
        Falls Prognose positiv (= Verspätung wird vorhergesagt):
          Überspringe alle weiteren Modelle desselben Verspätungsniveaus
      Falls zwei Niveaus ohne positive Prognose durchlaufen wurden:
        Überspringe die Prüfung weiterer Modelle für diese Zielvariable
  Gib für jede Zielvariable die prognostizierte Verspätung zurück als das höchste
  Verspätungsniveau, bei dem eine positive Prognose aufgetreten ist
}
```

Abbildung 43: Pseudocode: Ermittlung von Verspätungsprognosen durch Anwendung der Modelle

6.3 Umgang mit gerundeten Zeitangaben

Die von der Open Data Plattform bezogenen Vergangenheitsdaten sind sekundengenau. Die damit erzeugten Entscheidungsbäume enthalten Splitting-Kriterien, die ebenfalls sekundengenau sind (vgl. z.B. Abbildung 42).

In Echtzeit stellt die Open Data-Plattform jedoch nur minutengenaue Angaben zur Verfügung. Dabei wird stets abgerundet, d.h. die Sekundenangaben werden abgeschnitten. Es entsteht somit ein Rundungsfehler zwischen 0 und 60 Sekunden. Werden die abgerundeten Werte als Prädiktoren verwendet, fallen die prognostizierten Verspätungen systematisch geringer aus. Im Vergleich zu den ungerundeten Prognosen

- nimmt der Anteil korrekter Prognosen ab,
- nimmt der Anteil überschätzter Prognosen ab.

Um dem entgegen zu wirken, kann auf den gerundeten Wert ein fixer Sekundenbetrag aufaddiert werden. Es liesse sich etwa argumentieren, dass jeweils 30 Sekunden aufgeschlagen werden sollen, weil dies dem Mittel der möglichen Rundungsfehler entspricht. Ein solcher Wert wird von mir im Folgenden als «Rundungsgrenze» bezeichnet. Ein Wert von 30 würde der Anforderung eines stark konservativen Prognose-Bias zuwiderlaufen: wenn der Fehler einer Überschätzung gravierender wiegt als der einer Unterschätzung, dann ist ein «gemittelter» Aufschlag zu hoch. Zudem ist die Wahrscheinlichkeitsverteilung der Prädiktor-Variablen unbekannt und es sollte nicht unterstellt werden, dass diese symmetrisch zur Rundungsgrenze verläuft. Ein angemessener Wert kann also nur durch Ausprobieren ermittelt werden. Dabei ist die Wechselwirkung mit dem Loss-Koeffizienten zu beachten: Die Rundungsgrenze beeinflusst sowohl den Anteil korrekter Prognosen (= zu maximierende Grösse) als auch den Anteil überschätzter Prognosen (= einzuhaltende Rahmenbedingung; Zielwert = 0.05%). In einem «Grid-Search» wurden daher zunächst Modelle für verschiedene Loss-Koeffizienten generiert und dann mit verschiedenen `rundungsgrenze`-Werten simuliert.[115] Als gute Parameter-Kombination erwiesen sich ein Loss-Koeffizient von 14 und eine Rundungsgrenze von 8 Sekunden.

6.4 Anwendung auf Vergangenheitsdaten

Durch Anwendung des Prognoseverfahrens auf Vergangenheitsdaten lässt sich im Nachhinein simulieren, welche Verspätungs-Vorhersagen in einer bestimmten Situation getroffen worden wären. Dies ist nützlich, um

a) ein Tuning der (Hyper-)Parameter durchzuführen: Es kann geprüft werden, welche Ergebnisse sich bei unterschiedlicher Parametrisierung des Lern-Verfahrens ergeben (z.B. Loss-Koeffizient, Tiefe der Bäume, Höchstzahl der Prädiktoren). Durch wiederholtes Durchlaufen der Schritte Modellerstellung und Modellsimulation lassen sich die Auswirkungen der Parameter untersuchen und geeignete Konstellationen finden. Auch die Auswirkung verschiedener Rundungsverfahren bei der Modellanwendung lässt sich untersuchen.

b) das Verfahren in seiner Gesamtheit zu testen: Es kann geprüft werden, ob das Zusammenspiel der einzelnen Modelle im beschrieben Algorithmus zu validen Resultaten führt. Mit einer Anwendung auf Vergangenheitsdaten lässt sich ermitteln, welche Prognosen sich an jenen Betriebstagen ergeben hätten, die nicht Bestandteil des Trainingssets waren (Holdout-Verfahren).[116]

c) zu untersuchen, welche Faktoren die Prognosequalität beeinflussen: Die Zielsetzung dieser Arbeit nennt explizit die Möglichkeit, dass das Verfahren möglicherweise nur in «Teilbereichen» den Re-

[115] Hierbei wurden getrennte Samples für Training und Validierung gewählt. Das für den späteren Test reservierte Sample wurde dabei nicht angerührt, vgl. nächster Abschnitt.

[116] Das hierbei verwendete Testset sollte ausschliesslich der finalen Beurteilung des Modells dienen und weder beim Tuning noch beim Training zum Einsatz kommen, vgl. Russell / Norvig (2009), S. 709.

ferenzsystemen überlegen ist. Die Suche nach solchen Teilbereichen ist möglich, indem z.B. Simulationsergebnisse zu unterschiedlichen Zugslinien, mit unterschiedlichen Vorlaufzeiten oder unterschiedlichen Verspätungsniveaus verglichen werden.

d) das Verfahren auf neue Situationen anzuwenden: Es lässt sich z.B. untersuchen, wie sich die mit Daten der Fahrplanperiode 2017 trainierten Modelle nach dem Fahrplanwechsel verhalten.

Bei Durchführung der Simulation sind die Vergangenheitsdaten danach zu unterscheiden, ob sie zum simulierten Zeitpunkt bereits bekannt waren. Nur dann können sie als Prädiktoren in den Modellen verwendet werden.

Abbildung 44 zeigt das aus dem vorherigen Abschnitt bekannte Beispiel Spiez-Thun-Bern. Dargestellt sind die Fahrten der drei Züge A, B und C mit ihren planmässigen und tatsächlichen Ankunfts- und Abfahrtszeiten. Zur Vereinfachung sind alle Angaben im Beispiel minutengenau. Prognostiziert werden soll die Ankunft von B in Bern (Zielvariable, rot dargestellt). Die linke Spalte stellt die Situation um 11:31 Uhr dar, die auch im Beispiel oben verwendet wurde. Zu diesem Zeitpunkt sind die Abfahrten von A und B in Spiez sowie die Abfahrt von C in Bern und seine Ankunft in Thun bekannt (fett hervorgehoben). Nur diese Ereignisse dürfen verwendet werden, um eine Prognose zum Zeitpunkt 11:31 zu simulieren. Entsprechend sind die Modelle auszuwählen, deren Anwendung dann z.B. zur Vorhersage einer 3-minütigen Verspätung führen könnte. Eine Minute später, um 11:32 sind zwei weitere Ereignisse bekannt: die Ankunft von A in Thun und die Abfahrt von C in Thun. Die Hinzunahme dieser Information ermöglicht vermutlich die Anwendung zusätzlicher Modelle, die tendenziell zu einer besseren Prognose führen könnten (z.B. 4-minütige Verspätung von Zug B in Bern). Beide Prognosen können unmittelbar auf ihre Qualität untersucht werden, da der tatsächliche Wert der Zielvariablen (11:59 Uhr) ja bereits bekannt ist.

Simulation von Zeitpunkt 11:31

Zug	Ereignis	Planmässig	Tatsächlich
A	Abfahrt in Spiez	11:22	**11:22**
A	Ankunft in Thun	11:32	11:32
A	Abfahrt in Thun	11:33	11:34
A	Ankunft in Bern	11:52	11:52
B	Abfahrt in Spiez	11:25	**11:31**
B	Ankunft in Thun	11:34	11:40
B	Abfahrt in Thun	11:36	11:41
B	Ankunft in Bern	11:54	11:59
C	Abfahrt in Bern	11:06	**11:07**
C	Ankunft in Thun	11:24	**11:27**
C	Abfahrt in Thun	11:25	11:32
C	Ankunft in Spiez	11:34	11:40

Simulation von Zeitpunkt 11:32

Zug	Ereignis	Planmässig	Tatsächlich
A	Abfahrt in Spiez	11:22	**11:22**
A	Ankunft in Thun	11:32	**11:32**
A	Abfahrt in Thun	11:33	11:34
A	Ankunft in Bern	11:52	11:52
B	Abfahrt in Spiez	11:25	**11:31**
B	Ankunft in Thun	11:34	11:40
B	Abfahrt in Thun	11:36	11:41
B	Ankunft in Bern	11:54	11:59
C	Abfahrt in Bern	11:06	**11:07**
C	Ankunft in Thun	11:24	**11:27**
C	Abfahrt in Thun	11:25	**11:32**
C	Ankunft in Spiez	11:34	11:40

Abbildung 44: Verwendbarkeit von historischen Daten für zwei simulierte Zeitpunkte

Das Vorgehen kann iterativ für alle Minuten eines Betriebstags wiederholt werden. Es werden dabei nur zukünftige Ereignisse prognostiziert. Die Grösse des «Prognosefenster», d.h. der zeitliche Vorlauf der Prognose, ist wählbar. Das Verfahren ist nachfolgend in Pseudocode beschrieben:

```
prognose_simulation <- function(Betriebstag) {
  Ermittle alle predevents des Tages, welche einem Prädiktoren entsprechen
  Wende das gewünschte Rundungsverfahren auf alle predevents an
  Ermittle alle progevents des Tages, welche einer Zielvariable entsprechen
  Für alle Minuten des Tages:
    Wähle die predevents aus, deren Wert zur betrachteten Tagesminute bekannt ist
    Wähle die progevents aus, die zur betrachteten Minute im Prognosefenster liegen
    Für alle ausgewählten progevents:
      Berechne eine Prognose auf Basis der ausgewählten predevents
  Gib alle ermittelten Prognosen zurück
}
```

Abbildung 45: Pseudocode: Simulation der Prognosen eines Betriebstages mit Vergangenheitsdaten

Auch dieser Algorithmus wurde so implementiert, dass eine parallelisierte Ausführung auf dem R-Cluster möglich ist. Hierzu waren folgende Vorkehrungen notwendig:

- Der Bezug der erforderlichen Daten aus der (entfernten) Datenbank dauert recht lange. Für weiderholte Berechnungen werden die einmal aufbereiteten `predevents` und `progevents` daher in Dateien abgelegt. Um eine Überlastung der Datenbank durch parallele Anfragen zu vermeiden, habe ich diese Dateien jeweils vor dem Start der parallelisierten Simulation angelegt.
- Die erzeugten Prognosen können für spätere Analysen in die Datenbank geschrieben werden. Während dies innerhalb der AWS-Plattform reibungslos funktionierte, kam es beim schreibenden Zugriff vom BFH-Cluster immer wieder zu Abbrüchen. Um dies zu vermeiden, wurden die Prognosen vom BFH-Cluster zunächst in Dateien geschrieben, später auf die AWS-Plattform transferiert und von dort in die Datenbank geladen.

Der Source-Code für die Anwendung der Modelle auf Vergangenheitsdaten ist in den beigefügten Dateien `Modelle_anwenden.R` und `Aufruf_Simulationsrechnungen.R` enthalten.

6.5 Echtzeit-Anwendung

Die Durchführung von Echtzeitprognosen gliedert sich in drei Schritte:

1. Durch wiederholten Aufruf des `StopEvents`-API werden die aktuellen Prognosen der Referenzsysteme und möglichst viele der benötigten Prädiktoren von der Open Data Plattform bezogen. Dies geschieht mit Hilfe der in Kapitel 4.2 beschriebenen Funktion `ODPStopEvents()`.
2. Auf die so erhaltenen Echtzeitinformationen wird das bekannte Prognoseverfahren angewendet (Funktion `berechne_prognosen()`, vgl. Abbildung 43). Die ganzzahligen Minutenangaben werden dabei um 8 Sekunden (`rundungsgrenze`) erhöht, um dem Rundungsfehler entgegenzuwirken.
3. Die ermittelten Prognosen werden zusammen mit den Prognosen der Referenzsysteme in die Datenbank geschrieben. Da die Echtzeitprognose auf der AWS-Plattform läuft, ist dies problemlos möglich.

Für Echtzeitprognosen in der laufenden Fahrplanperiode sind die mit den Daten aus FP2017 trainierten Modelle unzureichend: bei einem Zug aus Thun hat sich die planmässige Ankunft in Bern geändert (und somit die Zielvariable), bei anderen Zügen ist es zu Veränderungen einiger Prädiktoren gekommen. Für die Echtzeitprognosen wurden daher neue Prognosemodelle trainiert mit allen Daten der FP2017 (ohne Bildung eines Holdout-Samples) und den bereits verfügbaren Daten der FP2018. Dies wurde erstmalig am 10. Januar vorgenommen (d.h. 1 Monat nach Fahrplanwechsel) und am 4. März wiederholt.

Bei der Implementierung wurde darauf geachtet, dass alle 3 Schritte zusammen eine Laufzeit von weniger als 60 Sekunden haben. Dies erlaubt es, über einen Scheduler Echtzeitprognosen im Minuten-

rhythmus erstellen zu lassen. Um diese Laufzeitanforderung zu erfüllen, musste einerseits die Zahl der Prognosemodelle begrenzt werden. Dies wurde durch geeignete Wahl der Pre-Pruning-Parameter erreicht, wie bereits in Abschnitt 5.7 ausgeführt wurde. Andererseits musste die Zahl der API-Abfragen begrenzt werden. Abbildung 46 zeigt für das Szenario Bern alle Fahrtabschnitte, deren Daten von mindestens einem Prognosemodell benötigt werden. Grün dargestellt sind jene Abschnitte, die durch Abfrage der «StopEvents» für Bern, Thun, Konolfingen, Burgdorf, Zofingen, Olten, Lyss, Freiburg, Lausanne, Kerzers, Langenthal, Mittelhäusern, Spiez und Wohlhusen abgedeckt werden können. Auf die Daten der roten Abschnitte muss aus Laufzeitgründen verzichtet werden.

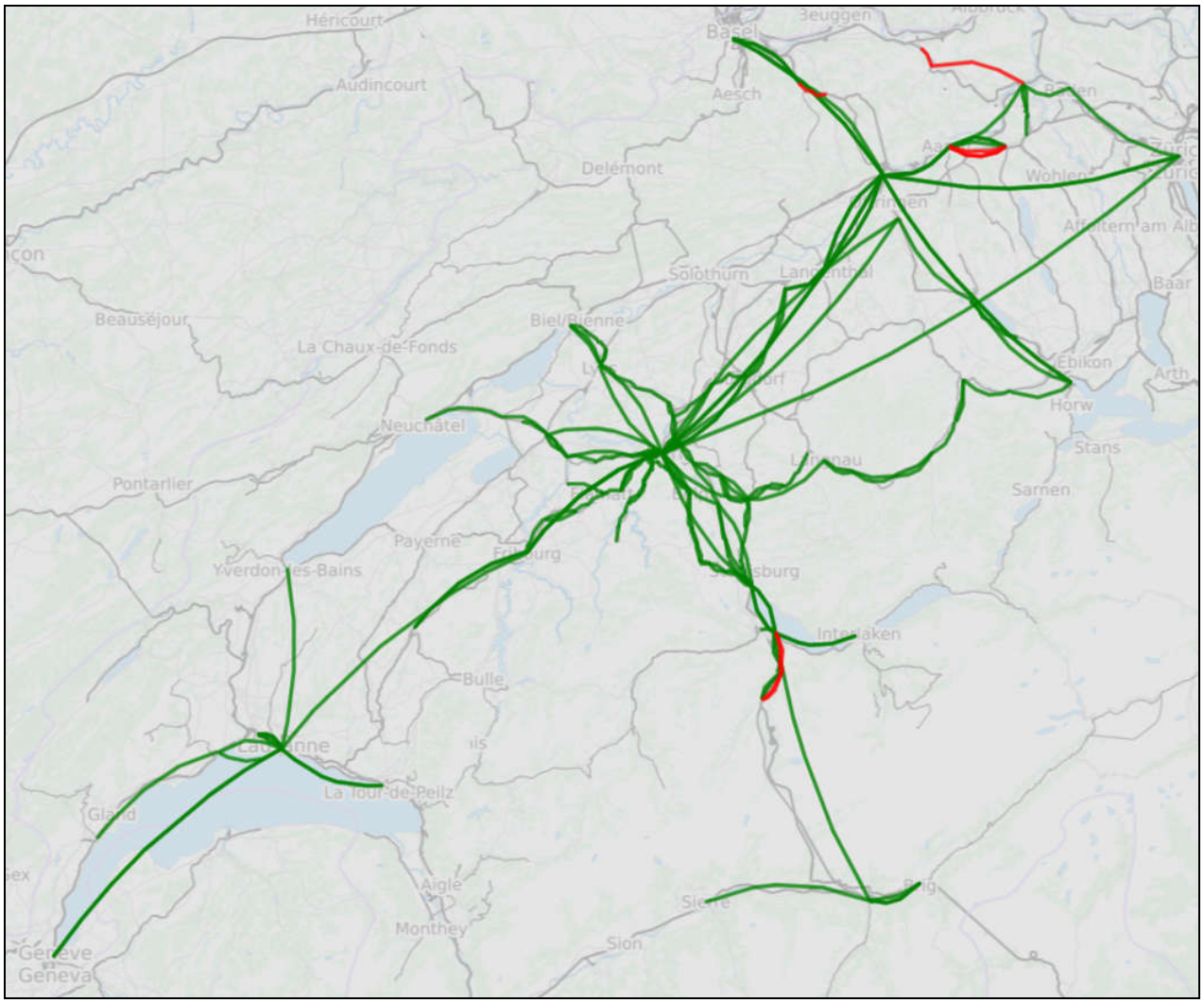


Abbildung 46: Abdeckung der benötigten Echtzeit-Informationen durch Abfrage von 15 Haltestellen (Ausschnitt)

Der Source-Code für die Echtzeitprognose ist in der Datei `RT_PrognosenBern.R` enthalten.

7 Visualisierung der Ergebnisse

Eine Prognose ist nichts wert, wenn sie nicht bekannt wird. Und eine bekannte Prognose ist nichts wert, wenn sie nicht geglaubt wird. Um Vertrauen in die mit meinem Verfahren erstellten Prognosen zu schaffen, habe ich folgende Massnahmen vorgesehen:

1. Jede Prognose wird sofort öffentlich kommuniziert.
2. Alle Prognosen werden jenen der (vertrauenswürdigen) Referenzsysteme gegenübergestellt.
3. Alle Prognosen werden begründet; es wird ersichtlich, wie sie zustande gekommen sind.
4. Alle Prognosen werden aufbewahrt und im Nachhinein geprüft.
5. Das Ergebnis der Prüfung wird öffentlich bekannt gemacht.

Die Umsetzung erfolgt mit Hilfe von drei Applikationen, die in diesem Kapitel beschrieben werden:

- *Analyse* bietet umfangreiche Auswertungsmöglichkeiten («Experten-Modus»).
- *Publikum* ist in www.puenktlichkeit.ch eingebunden und bewusst einfach gehalten.
- *Mobil* ermöglicht es, Verspätungsprognosen auf dem Smartphone abzurufen.

Die Implementierung erfolgte jeweils – wie schon bei meinen früheren Studienprojekten –mit R Shiny. Alle drei Applikationen sind für jedermann über das Web aufrufbar.

Die folgenden Abschnitte beschränken sich auf die Beschreibung der Informationsvisualisierung. Da es sich um «Mittel zum Zweck» handelt, gehe ich nicht auf die Implementierung ein. Für meine Erfahrungen mit R Shiny verweise ich auf die Semesterarbeiten aus den CAS Business Intelligence und Data Visualization.[117]

7.1 Applikation «Analyse»

Ziel der ersten Applikation ist es, aktuelle und historische Prognosen möglichst umfangreich darstellen und auswerten zu können. Programmieraufwand für individuelle Analysen soll reduziert und die Darstellung von wiederkehrenden Auswertungen – wie sie z.B. in Kapitel 8 dieser Arbeit verwendet werden – standardisiert werden. Sie wird primär von mir selbst genutzt, ist aber öffentlich aufrufbar unter www.puenktlichkeit.ch/prognose – allerdings weder dokumentiert noch selbsterklärend. Es lassen sich alle in der Datenbank vorhandenen Prognosewerte aufrufen – zu belieben Zügen, Daten, Zeitpunkten; aus Echtzeit-Prognose und Simulation:

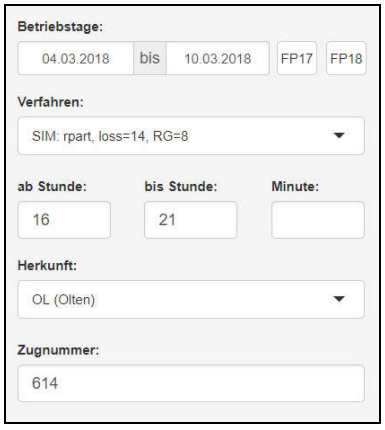
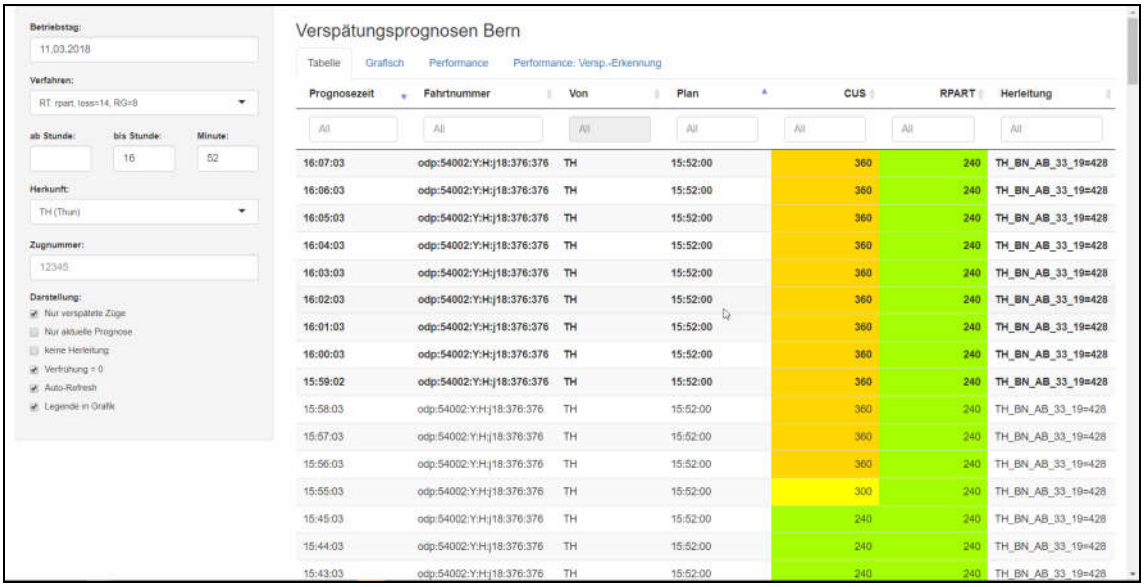


Abbildung 47: Analyse-Applikation: Auswahl der zu visualisierenden Daten

[117] Vgl. Gutweniger (2017) und Gutweniger (2017a).

Die Applikation umfasst vier Sichten:

- In der tabellarischen Darstellung werden Prognosen ähnlich angezeigt, wie sie auch in der Datenbank stehen. Es gibt umfangreiche Filter- und Sortier-Funktionen. Die Verspätungswerte der beiden Prognosequellen («CUS» für die Systeme der Bahnen, «RPART» für mein Verfahren) sind farblich codiert, um Unterschiede und Entwicklungen schnell erfassen zu können. In der Spalte Herleitung ist zu sehen, durch welche Prädiktor-Werte eine Prognose ausgelöst wurde: Im Beispiel von Abbildung 48 erfolgte die Abfahrt eines Zugs aus Thun um 439 Sekunden verspätet.



Prognosezeit	Fahrtnummer	Von	Plan	CUS	RPART	Herleitung
16:07:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
16:06:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
16:05:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
16:04:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
16:03:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
16:02:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
16:01:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
16:00:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
15:59:02	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
15:58:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
15:57:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
15:56:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	360	240	TH_BN_AB_33_19=428
15:55:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	300	240	TH_BN_AB_33_19=428
15:45:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	240	240	TH_BN_AB_33_19=428
15:44:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	240	240	TH_BN_AB_33_19=428
15:43:03	odp:54002:Y:H:j18:376:376	TH	15:52:00	240	240	TH_BN_AB_33_19=428

Abbildung 48: Analyse-Applikation, Tabellen-Ansicht

- In der grafischen Darstellung wird die Entwicklung der Prognosen im Zeitverlauf visualisiert. Auf der x-Achse ist die Prognosezeit, auf der y-Achse die prognostizierte Zeit aufgetragen. Prognosen der Bahn-Systeme sind mit gestrichelten Linien, Prognosen meines Verfahrens mit durchgezogenen Linien visualisiert. Farben für unterschiedliche Herkunftsorte helfen dabei, die Züge zu unterscheiden.

Abbildung 49: Analyse-Applikation, Grafische Ansicht

- Im Reiter «Performance» kann die Qualität von Prognosen zum Ausmass einer Verspätung evaluiert werden. Sie wird im Zeitverlauf (Vorlauf auf die tatsächliche Ankunft) dargestellt. Wiederum sind diverse Selektionskriterien vorhanden. Auch Schwellwerte und andere Auswertungsparameter können eingestellt werden.

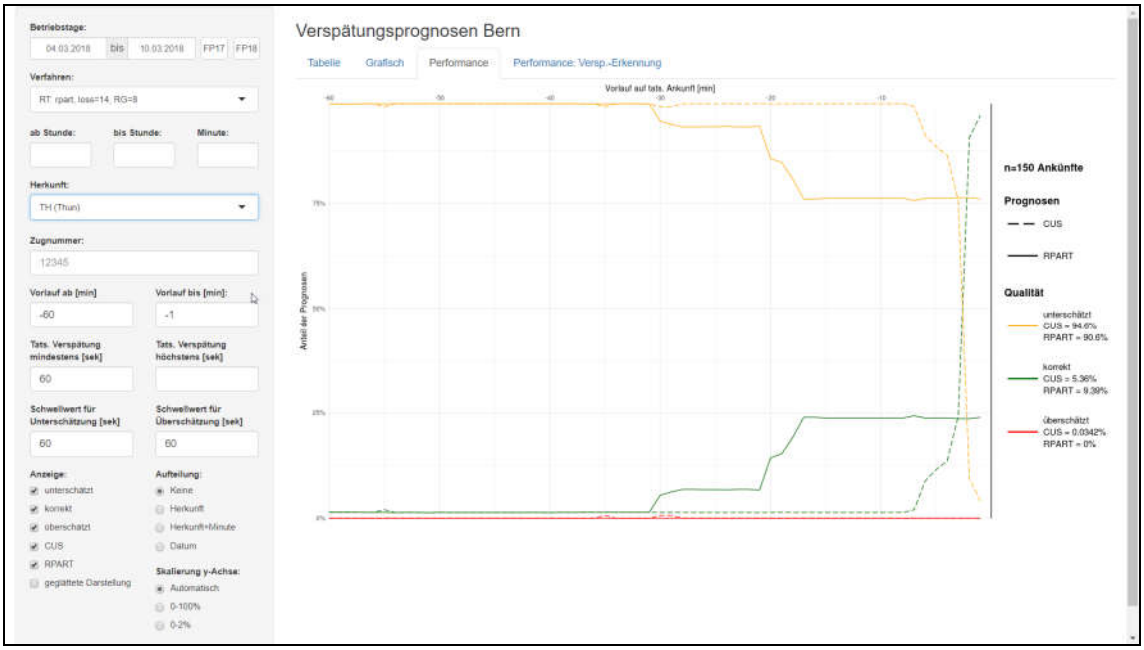


Abbildung 50: Analyse-Applikation, Performance-Ansicht (Verspätungsausmass)

- Performance-Auswertungen zur Prognose von Mindestverspätungen sind in der vierten Ansicht zu finden. Um die beiden Performance-Auswertungen leicht auseinanderhalten zu können, sind die verwendeten Farben bewusst unterschiedlich gewählt: dunkelgrün für Verspätungsausmass, hellgrün für Mindestverspätung.

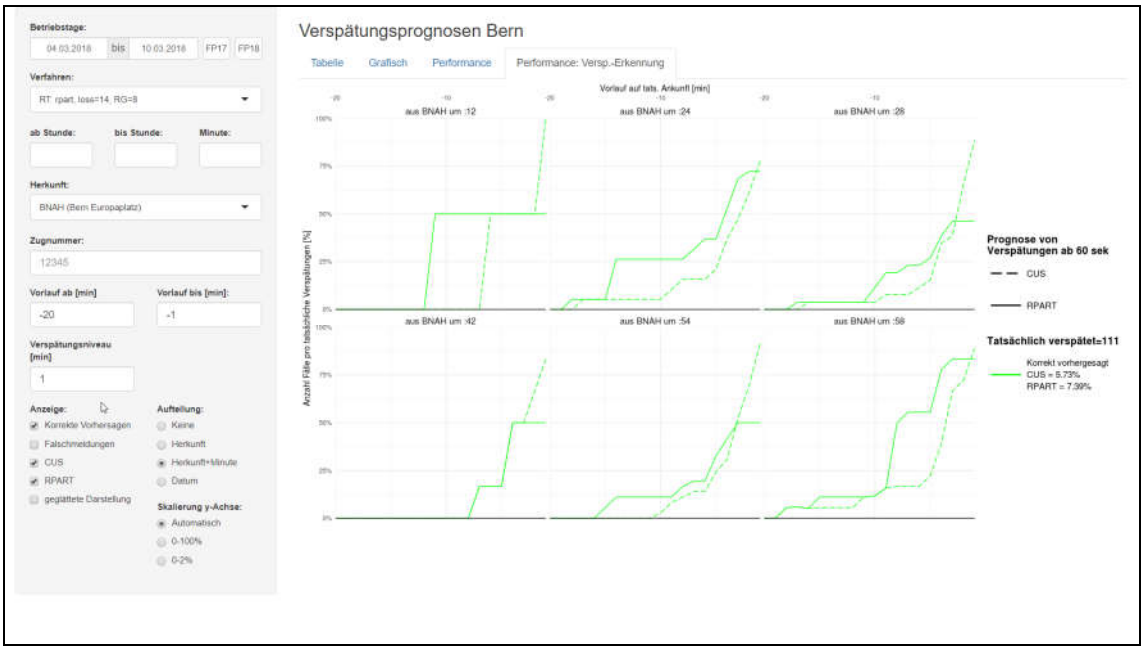


Abbildung 51: Analyse-Applikation, Performance-Ansicht (Mindestverspätung)

7.2 Applikation «Publikum»

Um meinem Projekt auch einem breiten Publikum zugänglich zu machen, habe ich vereinfachte Darstellungen in die bestehende Site www.puenktlichkeit.ch integriert. Die Informationen werden dort so aufbereitet, dass sie für interessierte Laien verständlich sind. Die Auswahlmöglichkeiten sind stark eingeschränkt, um leichte Bedienbarkeit zu gewährleisten. Erläuternde Texte vermitteln grundlegende Informationen zur Funktionsweise.

Die Applikation umfasst drei Sichten:

- Der erste Reiter zeigt die aktuellen Prognosen von Bahnunternehmen und RPART-Verfahren an. Verspätungen werden auch hier farblich hervorgehoben. Der Zugfahrt und Prognosebegründung werden textuell beschrieben.

Abbildung 52: www.puenktlichkeit.ch, Aktuelle Prognosen

- Die Prognose-Entwicklung der letzten 100 Minuten kann im zweiten Reiter grafisch nachvollzogen werden. Mit der Maus lassen sich Detail-Informationen abrufen.

Abbildung 53: www.puenktlichkeit.ch, Prognosen im Zeitverlauf

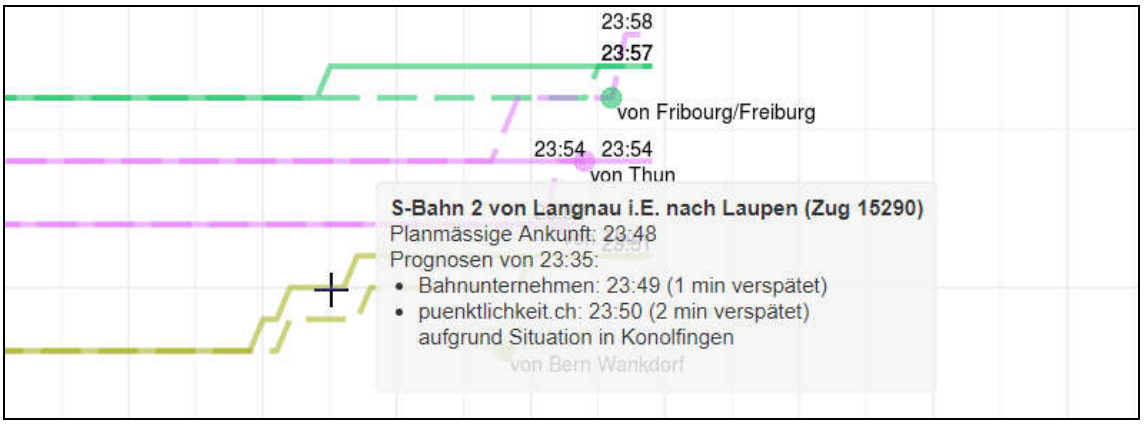


Abbildung 54: www.puenktlichkeit.ch, Detail-Informationen zum Prognoseverlauf

- Die Darstellung der Prognosequalität ist Gegenstand des dritten Reiters. Es wird die Vorhersage des Verspätungsausmasses analysiert. Auch hier gibt es rudimentäre Auswahlmöglichkeiten sowie eine Mouseover-Funktion für Detailinformationen.

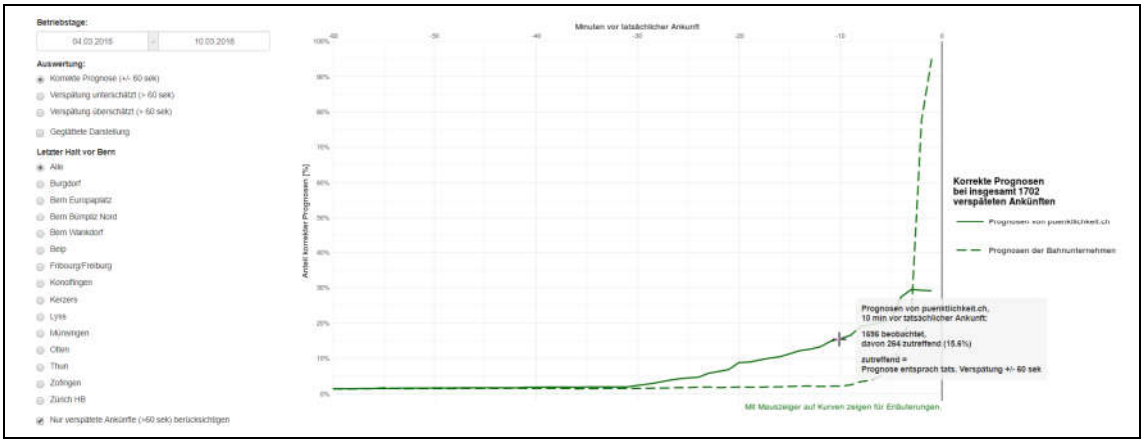


Abbildung 55: www.puenktlichkeit.ch, Prognosequalität

7.3 Applikation «Mobil»

Zugsverspätungen finden nicht daheim oder im Büro statt, sondern unterwegs. Es erscheint daher reizvoll, meine Verspätungsprognosen auch dort abrufen zu können. Ich habe daher die Realtime-Prognose auch in die Smartphone-Version von www.puenktlichkeit.ch integriert. Damit wird es möglich, das Eintreten von Prognosen «live» und vor Ort zu überprüfen.

Abbildung 56: Echtzeitprognosen auf dem Smartphone

8 Erkenntnisse

In diesem vorletzten Kapitel werden die bei der Bearbeitung entstandenen Erkenntnisse zusammengetragen. Ich beginne mit den Ergebnissen aus Simulation (Abschnitt 8.1) und Echtzeitprognose (8.2), um darauf aufbauend die Forschungsfragen zu beantworten (8.3). Das hier verwendete, auf Recursive Partitioning und Entscheidungsbäumen beruhende Verfahren, stellt eine von vielen Möglichkeiten dar, Verspätungsprognosen auf Basis von Open Data zu erzeugen. In der Rückschau kann einiges über Stärken und Schwächen des gewählten Ansatzes gesagt werden (8.4) – ebenso wie über seine Umsetzung und die dabei verwendeten Technologien (8.5). Lassen sich aus der Prognose von Verspätungen auch Erkenntnisse über Fahrplan und Betriebsabläufe der Bahnen gewinnen? Die Frage liegt jenseits des Scopes dieser Untersuchung, jedoch sind im Rahmen der Bearbeitung Dinge aufgefallen, die erwähnenswert erscheinen. Eine Auswahl solcher «Zufallsfunde» wird im letzten Abschnitt präsentiert.

8.1 Ergebnisse aus der Simulation

Um die Forschungsfragen zu beantworten, wurde das auf einem Sample der Fahrplanperiode 2016 trainierte Modell auf die zurückgehaltenen Daten angewendet (Holdout-Validierung).

Die Generierung der Entscheidungsbäume erfolgte mit folgenden Parametern:

- 45 Zielvariablen gemäss Szenario Bern und Fahrplanperiode 2017
- 15 Verspätungsniveaus («> 1 min» bis «>15 min»)
- Trainingsdaten: 291 zufällig gewählte Betriebstage der Fahrplanperiode 2017 (= 80%)
- Loss-Koeffizient = 14
- Pruning gemäss 1SE-Rule
- maximale Baumtiefe = 10
- maximale Zahl von Prädiktoren = 10
- minimale Knotengrösse für Split = 5% aller positiven Ausprägungen.

Die parallele Berechnung auf 80 Nodes des BFH-Clusters dauerte etwa 1 Stunde. Das resultierende Modell-Set weist folgende Charakteristika auf:

- 5124 Modelle (zwischen 7 und 1200 pro Zielvariable)
- Für 10 Zielvariablen wurden zu allen Verspätungsniveaus (d.h. bis «>15 min») Modelle generiert
- Verwendung von 892 verschiedenen Prädiktor-Variablen (zwischen 2 und 206 pro Zielvariable)
- Maximaler Vorlauf auf ein Ereignis: 89 Minuten
- Höchste Anzahl Prädiktoren nach Pruning: 5 (in 0.6% aller Modelle);
  1 Prädiktor in 61%, 2 Prädiktoren in 16%, 3 Prädiktoren in 16%, 4 Prädiktoren in 6% aller Modelle

Bei der Anwendung dieses Modell-Sets auf die 73 zurückgehaltenen Betriebstage wurden die Zeitangaben der Prädiktoren zunächst abgerundet und dann um 8 Sekunden erhöht (Parameter `rundungsgrenze=8`). Dies führte zu folgenden Ergebnissen:

Der Anteil Überschätzungen lag mit 0.05% exakt bei der Vorgabe. 69.9 % aller Ankünfte wurden korrekt vorhergesagt; von den mindestens 60 Sekunden verspäteten Ankünften waren es 6.8%. Das liegt nahe bei den Werten, die bei den Referenzsystemen beobachtet wurde (71.0% bzw. 5.0%). Allerdings ist die Vergleichbarkeit eingeschränkt, weil die beobachteten Tage und auch der zu Grunde liegende Fahrplan (Fahrplanperiode 17 vs. 18) unterschiedlich waren.

Interessant ist ein Vergleich der Vorhersagegenauigkeit im Zeitverlauf (vgl. Abbildung 57): Bei den Referenzsystemen bleibt sie bis ca. 6 Minuten vor tatsächlicher Ankunft auf sehr niedrigem Niveau und steigt dann steil an. 1 Minute vor tatsächlicher Ankunft werden 97% aller verspäteten Ankünfte korrekt vorhergesagt (+/- 60 Sekunden). Verspätungen werden also recht spät, kurz vor Ankunft aber sehr genau erkannt. Beim Recursive Partitioning-Verfahren (RPART) beginnt die Vorhersagegenauigkeit schon 30 Minuten vor Ankunft zu steigen und verharrt ab 3 Minuten Vorlauf auf dem erreichten Niveau von 30%. In diesen letzten Minuten vor Ankunft kann die Prognose nicht mehr verbessert werden, da keine neuen Informationen mehr auf der Open Data Plattform verfügbar werden: alle Züge haben den letzten Halt bereits passiert.

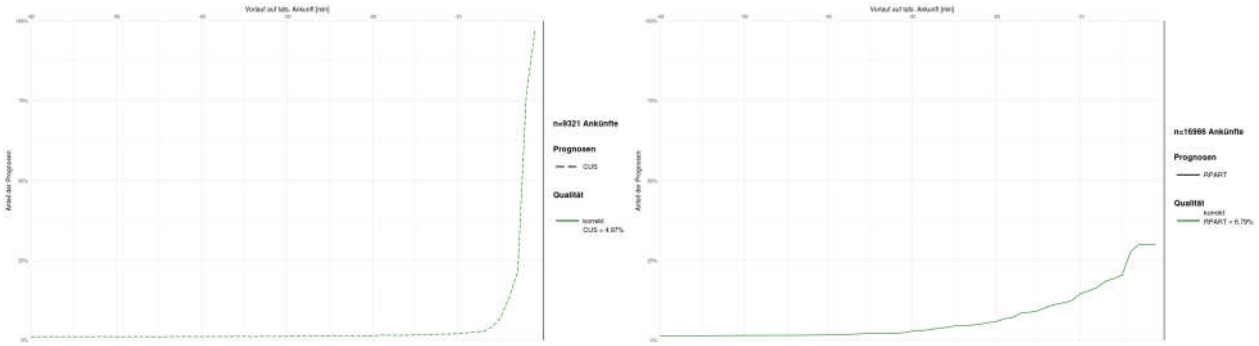


Abbildung 57: Accuracy von Referenzsystemen (FP18, links) und Simulation (FP17, rechts) bei verspäteten Ankünften[118]

Die oben dargestellten Kurven ergeben sich aus der Betrachtung aller in Bern ankommenden Züge. Schlüsselt man diese nach «Taktfamilie», d.h. nach Herkunft und Fahrplanminute auf, ergibt sich ein heterogenes Bild: Die erzielte Prognosequalität und ihre Entwicklung im Zeitverlauf sind sehr unterschiedlich.

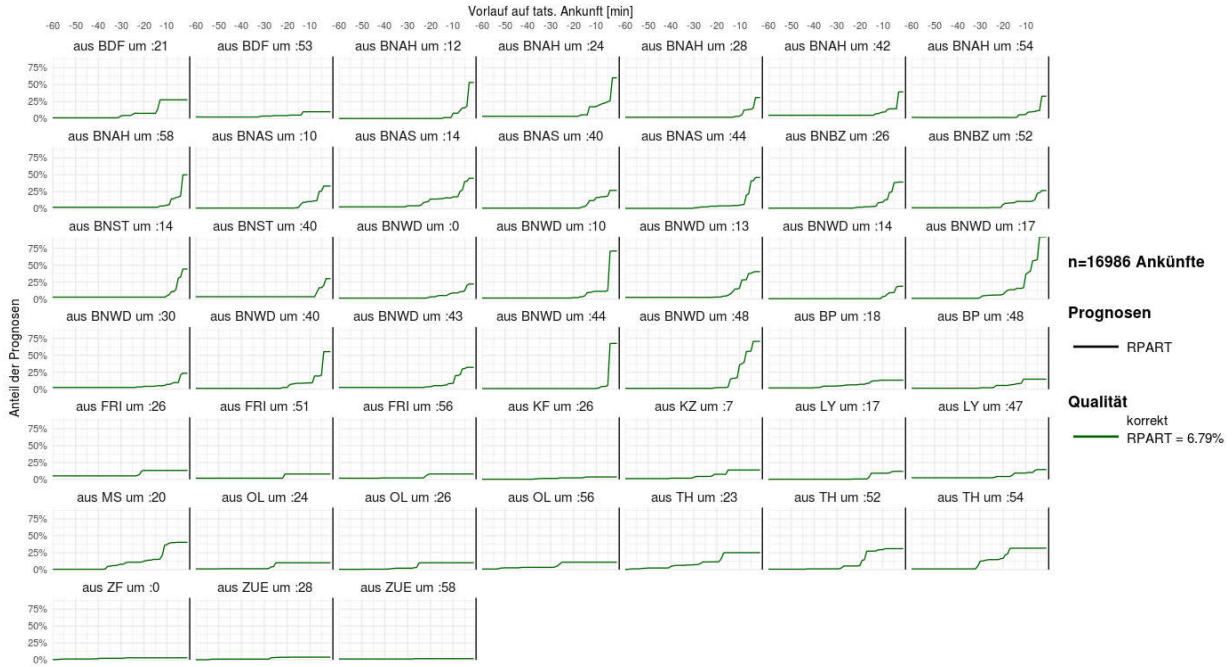


Abbildung 58: Ergebnisse der Simulation, unterschieden nach Herkunft und Fahrplanminute des Zuges[119]

[118] Die beiden Vorläufe sind hier bewusst in separaten Diagrammen dargestellt, weil sie aufgrund unterschiedlicher Beobachtungszeiträume nur eingeschränkt vergleichbar sind. Berücksichtigt sind nur Ankünfte ab 60 sek Verspätung.

[119] BDF = Burgdorf, BNAH und BNAS = Bern Europaplatz, BNBZ = Bern Bümpliz Nord, BNST = Bern Stöckacker, BNWD = Bern Wankdorf, BP = Belp, FRI = Freiburg, KF = Konolfingen, KZ = Kerzers, LY = Lyss, MS = Münsingen, OL = Olten, TH = Thun, ZF = Zofingen, ZUE = Zürich.

Bei den meisten S-Bahnen (aus BNAH, BNAS, BNBZ, BNST, BNWD) verbessert sich die Prognosequalität 5 bis 10 Minuten vor Ankunft noch deutlich – bei der S2 aus Langnau (:17) auf über 90%. Dies ist möglich, weil dank der häufigen Halte auch in diesem Zeitraum relevante Informationen zur Betriebslage verfügbar werden.[120] Die Prognosen mit mehr als 15 Minuten Vorlauf sind bei den S-Bahnen von geringer Qualität. Plausible Erklärungen hierfür sind:

a) die Züge sind dann noch gar nicht am Ausgangsort abgefahren (z.B. S31 aus Münchenbuchsee, Bern an :14 und :44)

b) aufgrund der häufigen Halte sind mehr Reserven im Fahrplan enthalten, d.h. Verspätungen werden schneller abgebaut

c) viele Verspätungsursachen (z.B. durch hohes Fahrgastaufkommen, Türstörungen) treten erst kurzfristig auf.

Umgekehrt ist die Prognosequalität bei RE-, IR- und IC-Zügen (aus BDF, FRI, KF, KZ, TH, LY, OL, ZF, ZUE) deutlich geringer: die besonders «wertvollen» Informationen in den letzten 10 Minuten vor Ankunft sind hier nicht verfügbar. Eine generell schlechte Prognosequalität weisen die Züge aus Zofingen (ZF) und Zürich auf. Für diese Züge konnten bereits in der Trainings-Phase nur wenige valide Modelle gebildet werden, weil kaum aussagekräftige Prädiktoren vorhanden sind. Zum Vergleich: Beim IC aus Zürich findet das letzte auf der Open Data Plattform publizierte Ereignis (= Abfahrt in Zürich) 58 Minuten vor Ankunft in Bern statt. Die Bahn-internen Systemen erhalten Zugpositionsmeldungen grundsätzlich beim Passieren jedes Hauptsignals – zwischen Zürich und Bern sind das etwa 140.

Bei anderen RE/IR/IC-Linien – z.B. aus Thun (TH) und Freiburg (FR) – sind dagegen gute Prognosen auch schon mit grossem Vorlauf (30 Minuten und mehr) möglich.

Bei vielen Diagrammverläufen sind ausgeprägte Stufen zu erkennen. Diese treten zu Zeitpunkten auf, wo relevante Informationen bekannt werden. Oft handelt es sich um Halte des betrachteten Zugs.

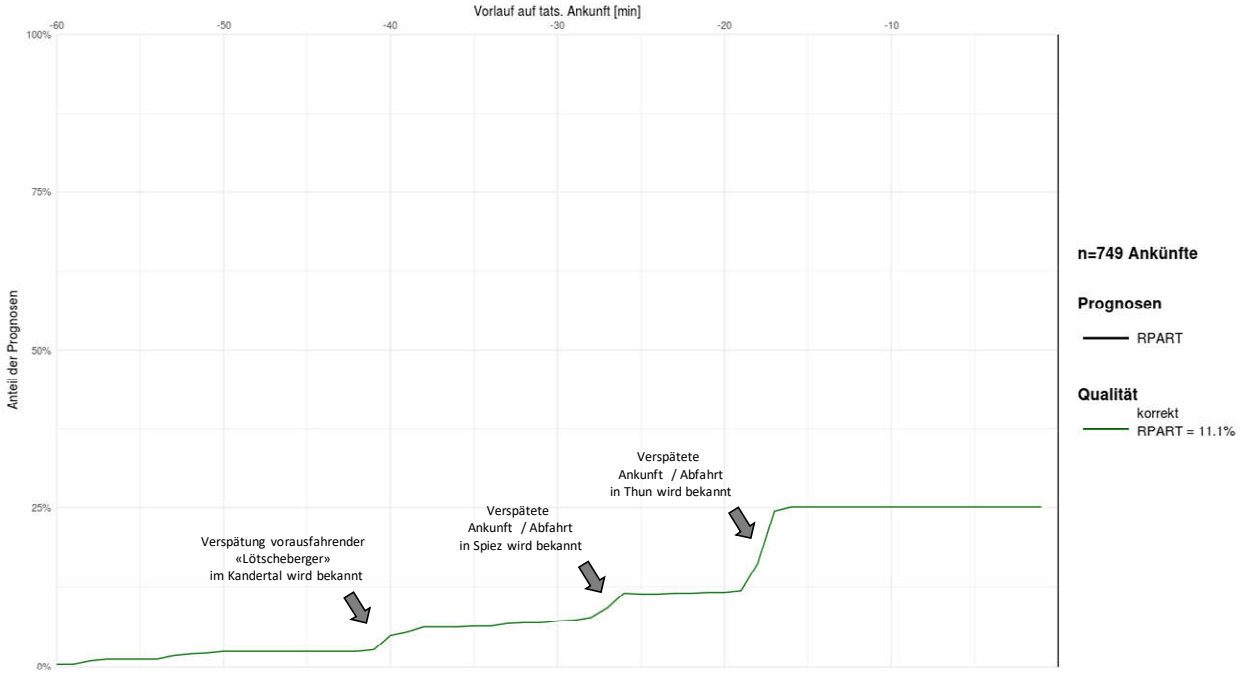


Abbildung 59: Simulation IC aus Thun (:23) in FP2017: neue Informationen verbessern die Prognose[121]

[120] Eine Ausnahme bildet die S4 aus Belp (BP, :18 und :48), die 13 Minuten vor Ankunft in Bern ihren letzten Halt hat.

[121] Berücksichtigt sind nur Ankünfte ab 60 Sekunden Verspätung.

Können mit den auf der Fahrplanperiode 2017 trainierten Modellen auch Vorhersagen in 2018 getroffen werden?

Abbildung 60 zeigt die simulierte Anwendung auf die ersten 64 Tage der Fahrplanperiode 2018. Die Accuracy fällt nicht viel geringer aus als in Abbildung 58 (6.2% vs. 6.8%, bezogen auf Ankünfte mit mindestens 60 Sekunden Verspätung). Der IC aus Thun mit Ankunft um :23 ist wegen der Fahrplanänderung nicht mehr aufgeführt; sein «Nachfolger» (Ankunft um :24) kann nicht prognostiziert werden, weil hierfür keine Entscheidungsbäume vorliegen. Für die meisten anderen Takte ist die Prognosequalität ähnlich wie in der Vorperiode. Kurios ist die Situation beim IC aus Freiburg um :26: Dieser wird (mit dem gleichen Modell) nun besser prognostiziert als im Vorjahr. Die Ursache hierfür habe ich bisher nicht ausfindig machen können. Möglicherweise lässt sich ein in FP2017 gelerntes «Verspätungspattern» in FP2018 besonders häufig anwenden.

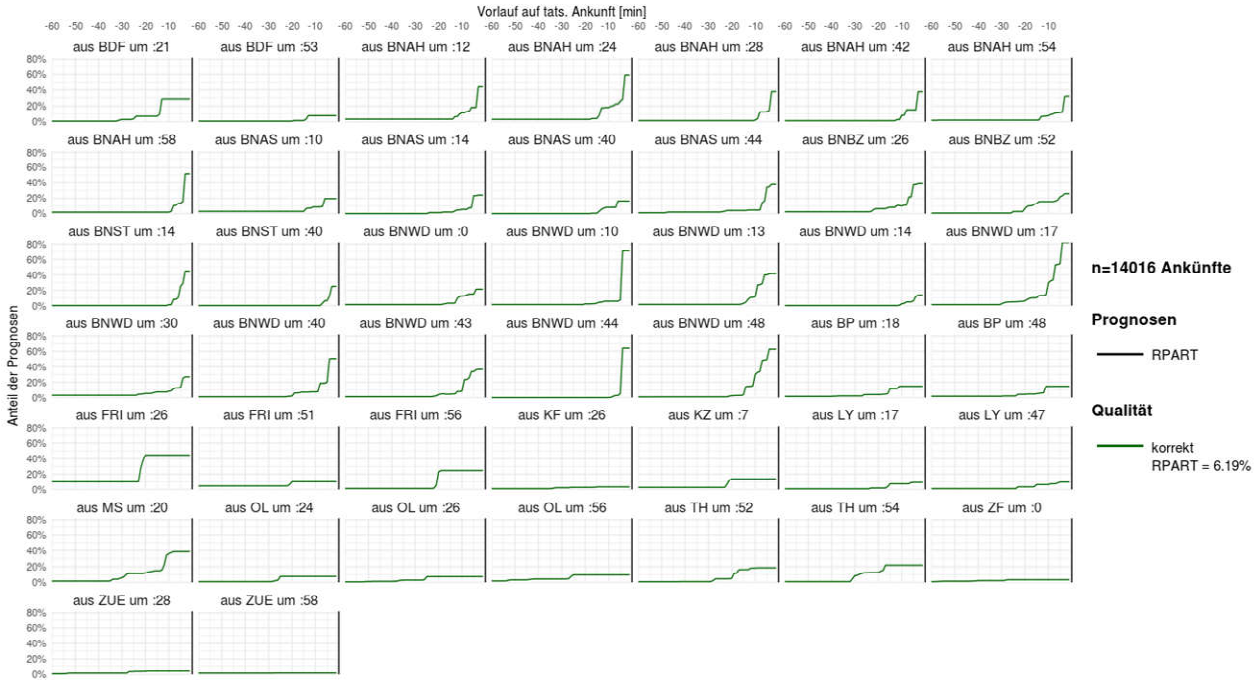


Abbildung 60: Simulierte Anwendung der Modelle aus FP2017 auf den Zeitraum 10.12.2017 bis 11.02.2018[122]

Die bisherigen Auswertungen betrachteten den Anteil korrekter Prognosen (+/- 60 Sekunden) bei Vorhersagen zum Ausmass von Verspätungen. Und wie häufig kann die Überschreitung eines gegebenen Grenzwerts (z.B. mehr als 5 Minuten) korrekt vorhergesagt werden?

Abbildung 61 stellt die Beobachtungen bei den Referenzsystemen (17. Januar bis 27. Februar 2018) der Simulationsrechnung (Hold-Out-Sample der Fahrplanperiode 2018) gegenüber. Erneut muss beachtet werden, dass die Vergleichbarkeit aufgrund unterschiedlicher Beobachtungszeiträume eingeschränkt ist.

[122] Berücksichtigt sind nur Ankünfte ab 60 Sekunden Verspätung.

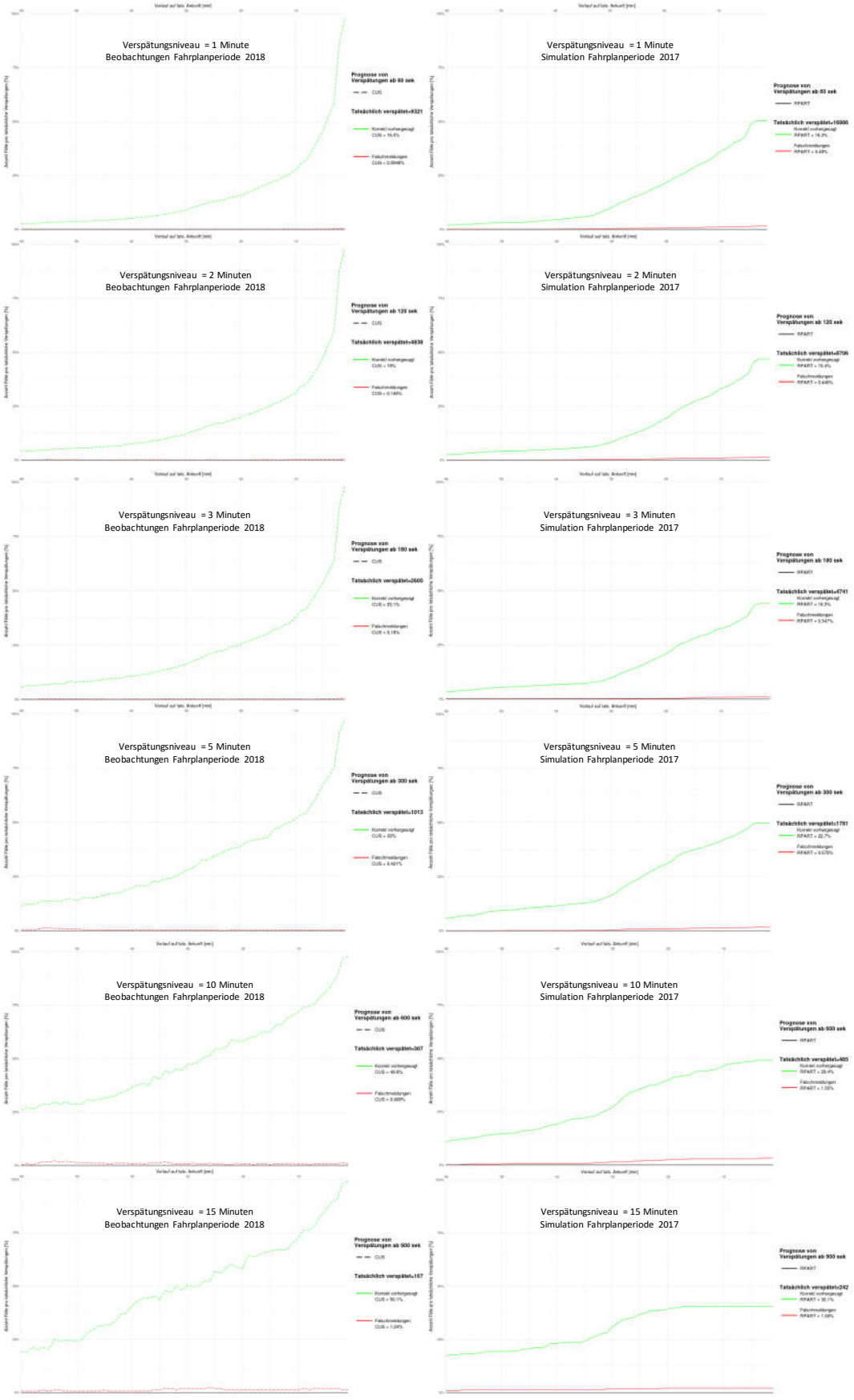


Abbildung 61: Prognose von Mindestverspätungen: Referenzsysteme (FP18, links) und Simulation (FP17, rechts)

Offensichtlich schneiden die Referenzsysteme der Bahnen diesmal besser ab als das hier entwickelte Verfahren – und zwar umso mehr, je höher das Verspätungsniveau ist. Sie scheinen grosse Verspätungen zuverlässiger vorhersagen zu können als kleine. Da grosse Verspätungen in der Schweiz selten sind, schlug dieser Vorteil bei der Bewertung des Verspätungsausmasses nicht zu Buche.

8.2 Ergebnisse aus der Echtzeit-Prognose

Für die Echtzeit-Prognosen wurden über einen Zeitraum von 10 Wochen mehrfach Modelle generiert, wobei Parameter und Trainingsdaten an den jeweils vorhandenen Erkenntnis- und Datenstand angepasst wurden. Im Folgenden werden die Ergebnisse der letzten Iteration beschrieben. Grundlage ist ein mit folgenden Parametern erstelltes Modell:

- 45 Zielvariablen gemäss Szenario Bern und Fahrplanperiode 2018
- 15 Verspätungsniveaus («> 1 min» bis «>15 min»)
- Trainingsdaten: alle 447 Betriebstage vom 11. Dezember 2016 bis zum 2. März 2018
- Loss-Koeffizient = 14
- Pruning gemäss 1SE-Rule
- maximale Baumtiefe = 10
- maximale Zahl von Prädiktoren[123] = 3
- minimale Knotengrösse für Split = 5% aller positiven Ausprägungen.

Das resultierende Modell-Set weist folgende Charakteristika auf:

- 5039 Modelle (zwischen 12 und 376 pro Zielvariable)
- Für 20 Zielvariablen wurden zu allen Verspätungsniveaus (d.h. bis «>15 min») Modelle generiert
- Verwendung von 783 verschiedenen Prädiktor-Variablen (zwischen 2 und 136 pro Zielvariable)
- Maximaler Vorlauf auf ein Ereignis: 80 Minuten
- Höchste Anzahl Prädiktoren war durch Vorgabe beschränkt auf 3; es wurden verwendet: 1 Prädiktor in 76%, 2 Prädiktoren in 17%, 3 Prädiktoren in 7% aller Modelle

Bei der Anwendung dieses Modell-Sets auf die minutengenauen Echtzeitdaten der Open Data Plattfrom wurden erneut 8 Sekunden addiert (Parameter `rundungsgrenze=8`).

Im Zeitraum 4. bis 10. März 2018 (= 1 Woche) wurden 4646 Ankünfte beobachtet. Mein Verfahren überschätzte Verspätungen in 0.05% der Fälle, was erneut dem Zielwert entspricht. Bei den Bahn-Systemen waren es ungewöhnlich hohe 0.1%. Mein Verfahren lieferte 66.7% korrekte Prognosen (+/- 60 sek), die Bahnen 65.9%. Bei Eingrenzung auf Ankünfte mit mindestens 60 Sekunden Verspätung ergibt sich eine Accuracy von 7.4% (mein Verfahren) gegenüber 5.4% (Bahnen).

Die Unterschiede sind also gering, sehr deutlich ausgeprägt sind aber wieder die Unterschiede im zeitlichen Verlauf (vgl. Abbildung 62). Es zeigt sich dasselbe Bild wie bei der Simulation: Die Referenzsysteme (gestrichelte Linien) liefern erst wenige Minuten vor Ankunft gute Prognosen, können diese aber bis zuletzt steigern. Die Prognosequalität steigt bei RPART (durchgezogene Linie) schon deutlich früher, verharrt in den letzten 5 Minuten aber auf dem erreichten Niveau.

[123] Der sonst verwendete Wert von maximal 10 Variablen führte in diesem Setting dazu, dass für eine der Zielvariablen weit über 1000 Modelle generiert wurden. Dies erwies sich einerseits als ineffizient in der Anwendung, andererseits entstand auch kein nenenswerter Nutzen, da die Modelle in hohem Masse redundant waren (geringfügige Variationen und Permutation zu einer zudem eher unwahrscheinlichen Verspätungskonstellation). Der Wert wurde daher auf 3 reduziert.

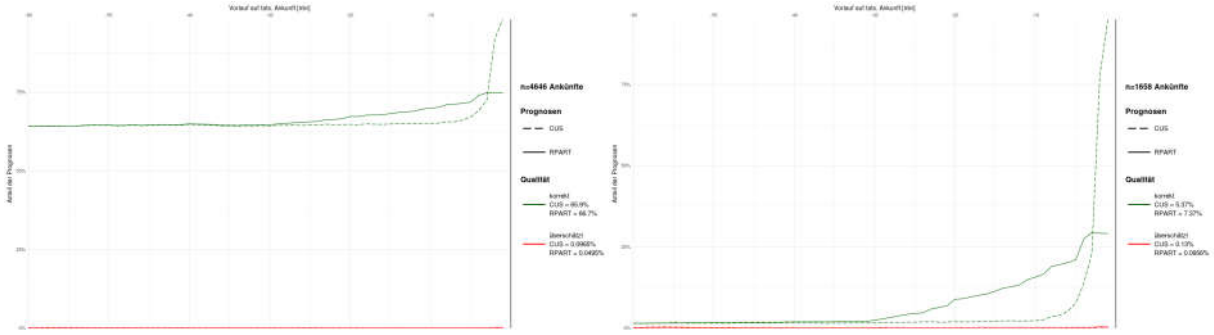


Abbildung 62: Qualität der Echtzeitprognosen: Alle Ankünfte (links) und Ankünfte ab 60 sek Verspätung (rechts)

Dieses Muster ist an allen Tagen zu beobachten, lediglich die Werte variieren leicht (Abbildung 63). Die Ergebnisse stehen auch im Einklang mit den Erfahrungen aus vorherigen Iterationen (die Echtzeitprognose werden von mir seit 17. Januar 2018 mit unterschiedlichen Modellen und Parametern durchgeführt).

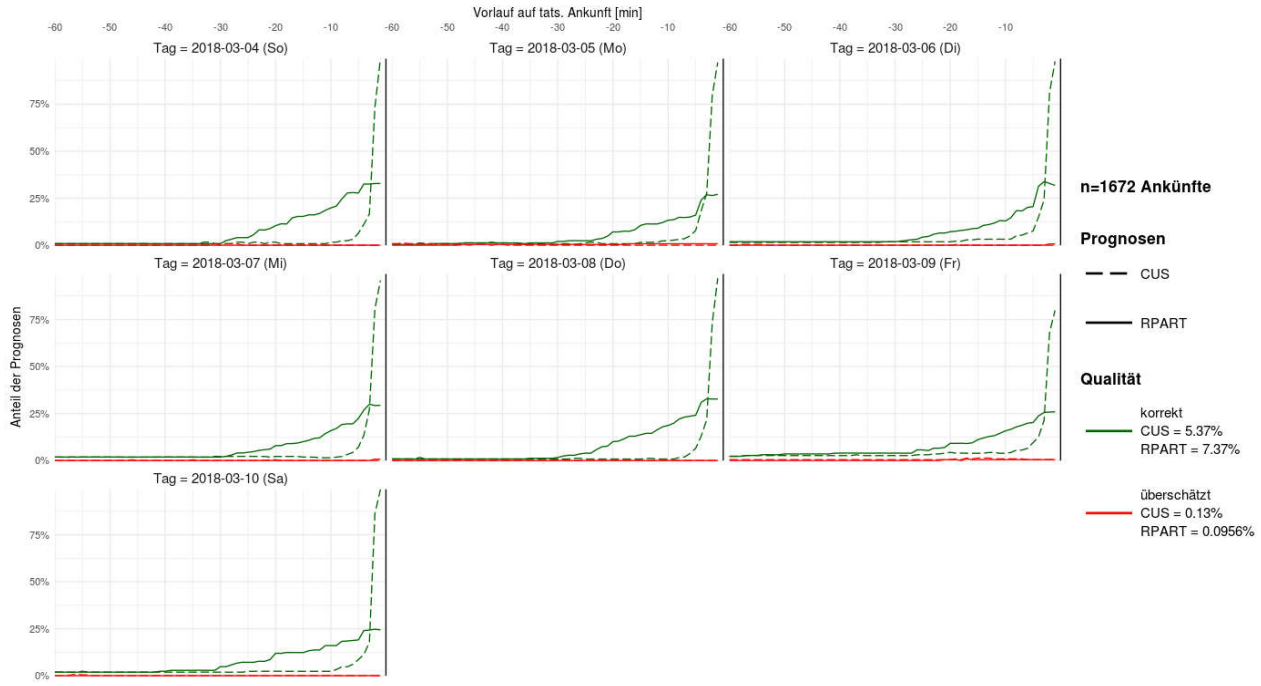


Abbildung 63: Qualität der Echtzeitprognosen an verschiedenen Tagen (Ankünfte ab 60 sek Verspätung)

Auch bei der Vorhersage von Mindestverspätungen bestätigen sich die Ergebnisse aus der Simulation: Die Bahn-Systeme schneiden meist besser ab – und zwar umso mehr, je höher das Verspätungsniveau ist. Lediglich für «>1 min» liefert RPART im Vorlauf 30 bis 5 Minuten vor Ankunft etwas bessere Vorhersagen. Abbildung 64 zeigt die Ergebnisse für Niveaus von 1, 2, 3, 5, 10 und 15 Minuten. Zu erkennen ist auch, dass bei den hohen Niveaus die Fallzahlen sehr klein sind.

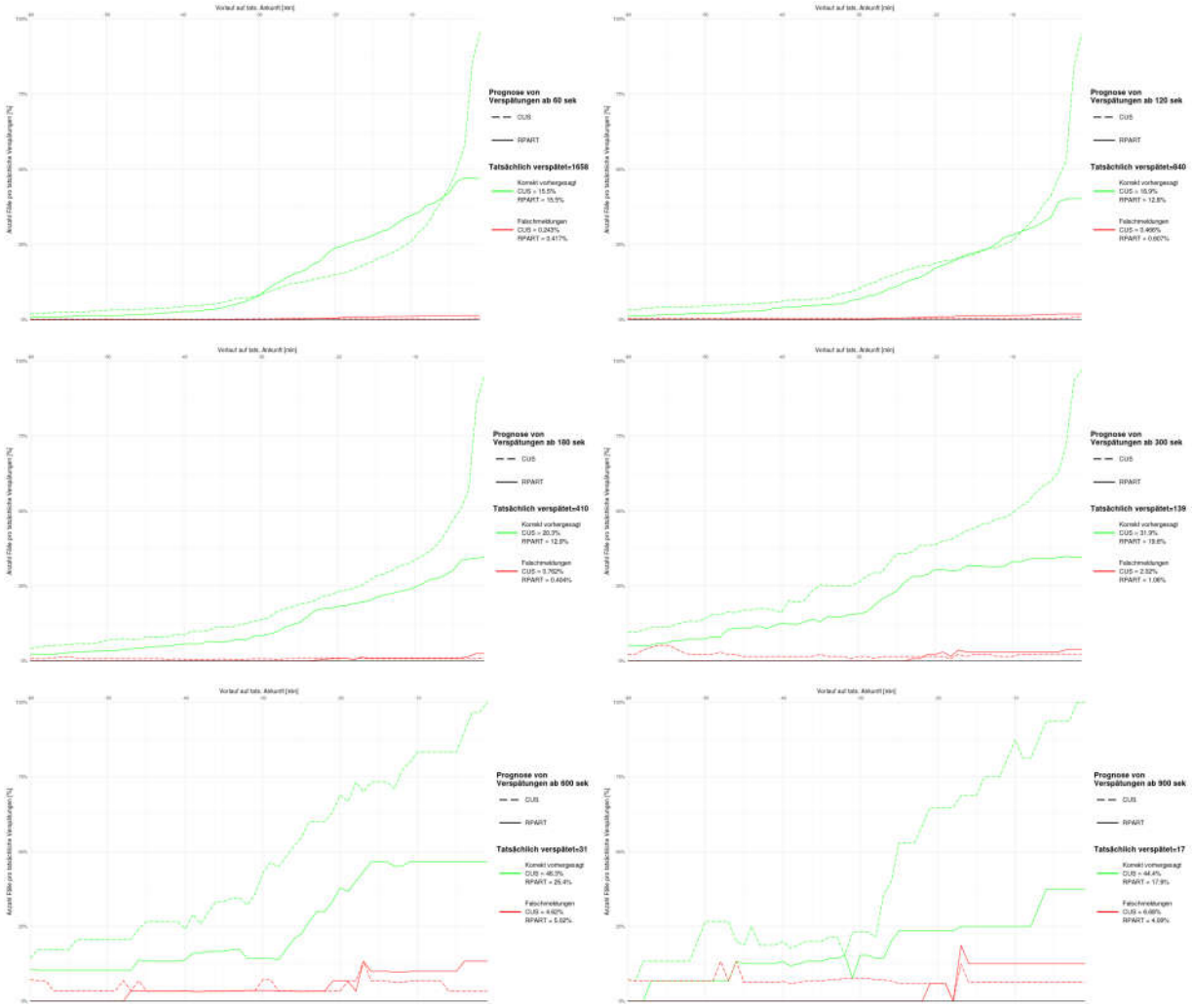


Abbildung 64: Treffer- und Fehlerquote bei der Vorhersage von Mindestverspätungen (1, 2, 3, 5, 10 und 15 min)

8.3 Beantwortung der Forschungsfragen

Die präsentierten Ergebnisse erlauben die Beantwortung der Forschungsfrage

> **Können unter Verwendung der empirischen Vergangenheitswerte von opentransportdata.ch mindestens in Teilbereichen bessere Verspätungsprognosen erzielt werden, als sie von den aktuell im öV Schweiz eingesetzten Kundeninformationssystemen geliefert werden?**

Ja, wie gezeigt wurde, ist eine bessere Verspätungsprognose möglich

- im Szenario Bern
- mit dem hier beschriebenen Verfahren
- bei der Vorhersage des Verspätungsumfangs
- in einem Zeitraum von ca. 30 bis 5 Minuten vor Ankunft des Zuges.

Die Qualität der Prognosen war je nach Linie unterschiedlich, übertraf im genannten Vorlauf-Zeitraum aber stets jene der Referenzsysteme. Besonders gute Prognosen wurden geliefert, wenn

a) durch Unterwegs-Halten ausreichend häufig Informationen über den Zug bekannt wurden,

b) viele Erfahrungswerte aus der Vergangenheit zum Training verwendet werden konnten,

c) es sich um kleine bis mittlere Verspätungen handelte.

8.4 Bewertung des gewählten Verfahrens

Das hier verwendete Prognoseverfahren basiert auf Entscheidungsbäumen, die mit Recursive Partitioning erzeugt werden. Somit wird ein seit über 30 Jahren bekannter Algorithmus verwendet, der für ein breites Spektrum von Anwendungen geeignet ist. Der Nachteil, dass die erzielbare Performance möglicherweise unter jener von anspruchsvollerer Verfahren (z.B. Ensemble Learning) zurückbleiben könnte, wurde angesichts anderer Vorteile (Einfachheit, vermutete Effizienz, Parameterfreiheit, Interpretierbarkeit) bewusst in Kauf genommen. Dies auch deshalb, weil angesichts der verfolgten Zielsetzung eine breite Abdeckung aller Anforderungen wesentlich wichtiger erschien, als die Optimierung der Prognosequalität. Auch die mit der Methodenwahl verbundene Beschränkung auf kategorielle (ordinale) Zielgrössen erschien vertretbar. Zu den häufig genannten Nachteilen von Recursive Partitioning gehört eine starke Tendenz zum Overfitting. Dem konnte erfolgreiche begegnet werden, in dem die Entscheidungsbäume mit Pre-Pruning und Post-Pruning / Kreuzvalidierung klein gehalten wurden.

Die Wahl von Recursive Partitioning hat sich grundsätzlich bewährt:

- Die erzielte **Prognosequalität** übertrifft jene der Bahn-Systeme in vielen Situationen. Dort, wo sie es nicht tut, ist dies häufig auf das Fehlen aktueller Betriebslage-Daten zurückzuführen (bei Zügen mit nur wenigen Halten sowie im Zeitintervall unmittelbar vor Ankunft eines Zuges) oder auf ein unzureichendes Trainingsset (bei Zügen mit Fahrplanänderungen oder wenigen Taktlagen).
- Der erforderliche «**Bias**» der Prognosen konnte durch den Loss-Koeffizienten erzielt werden und liess sich leicht kalibrieren. Schwierigkeiten bereitete lediglich die Bestimmung des Zielwerts angesichts der grossen beobachteten Varianz bei den Referenzsystemen. Hier wurden erhebliche Schwankungen zwischen Zugslinien und Betriebstagen beobachtet.
- Der **Datenumfang** erwies sich als ausreichend für das Erzeugen valider Modelle.
- Die Anwendung des Verfahrens ist **effizient**: Prognosen können ausreichend schnell berechnet werden, um eine Echtzeit-Anwendung zu ermöglichen. Die Erstellung der zahlreichen Modelle ist deutlich aufwendiger, bei parallelisierter Berechnung aber in vertretbarer Zeit möglich.
- Von grossem Wert für Entwicklung und Qualitätssicherung war die **Interpretierbarkeit** der erstellten Entscheidungsbäume: die gelieferten Ergebnisse liessen sich leicht nachvollziehen, Software-Tests waren mit vertretbarem Aufwand durchführbar, Unstimmigkeiten in Daten und Algorithmus konnte einfach auf den Grund gegangen werden. Es gehörte zwar nicht zum Scope dieser Arbeit, Erklärungsmodelle zu erzeugen. Die entstandenen Entscheidungsbäume würden sich aber vermutlich gut dafür eignen, Verspätungsursachen zu analysieren.
- Bewährt hat sich auch der weitgehende **Verzicht auf Annahmen zu strukturellen Zusammenhängen** zwischen den Variablen. Tatsächlich konnten mehrfach Fälle von «unerwarteten» Prädiktoren beobachtet werden, deren Bedeutung erst bei genauerer Analyse der Situation offenkundig wird. Ein Beispiel wurde in Abschnitt 6.2 behandelt: Der Einfluss des ICs von Bern nach Spiez, der bei Verspätung in Thun ein benötigtes Gleis belegt und den Zug in Richtung Bern verspätet.
- Der wohl gravierendste Nachteil liegt in der Schwierigkeit, mit der «**Class Imbalance**» in den Daten umzugehen: Eine Folge der Schweizer öV-Pünktlichkeit ist, dass nur wenige Trainingsdaten für die besonders interessanten Fälle zur Verfügung stehen: Während sich für kleine Verspätungen zahlreiche Modelle erzeugen liessen, konnten Verspätungsniveaus oberhalb von 10 Minuten vielfach nicht abgedeckt werden. Für den Umgang mit solchen Situationen werden in der Literatur verschiedene Massnahmen vorgeschlagen[124], die sich im Zeitrahmen dieser Arbeit jedoch leider nicht mehr erproben liessen.

[124] Häufig genannt werden Down-Sampling, Up-Sampling, SMOTE und ROSE. Vgl. Kuhn (2017), Chawla / Bowyer / Hall / Kegelmeyer (2002), Menardi / Torelli (2014).

Während für den Kern des Machine Learnings auf eine Standardmethode gesetzt wurde, widmete sich das Hauptaugenmerk dieser Arbeit auf die Adaptierung des Verfahrens an den dynamischen Kontext: Von einer grossen Zahl potentiell relevanter Variablen sind zu einem gegebenen Zeitpunkt stets nur einige verfügbar. Deren relative Bedeutung kann sich zudem von Minute zu Minute ändern: wenn neue Informationen zur Betriebslage hinzukommen, werden andere obsolet. Zudem sollten nur solche Variablen in Beziehung gesetzt werden, deren Kausalzusammenhang plausibel erscheint.

Zu diesem Zweck wurden Algorithmen für die Variablenauswahl, die Erstellung von Modellen, die Auswahl passender Modelle und den Zusammenzug der Prognoseergebnisse entworfen. Die praktische Anwendung dieser Algorithmen in Simulationsrechnungen und in Echtzeit-Situationen hat deren Tauglichkeit unter Beweis gestellt: Die Auswahl aus zahlreichen Prädiktoren, der Umgang mit Missing Values und die Berechnung von validen Prognosen kann damit effizient bewerkstelligt werden. Das Verfahren ist seit Mitte Januar nahezu kontinuierlich in Betrieb und hat sich als robust erwiesen.

8.5 Bewertung von Systemarchitektur und Technologien

Auch die gewählte Systemarchitektur hat sich bewährt. Die eingesetzten Technologien und ihr Zusammenspiel waren gut geeignet, um die vielfältigen Aufgaben abzudecken, die zur Beantwortung der Forschungsfrage notwendig waren: Beschaffung, Speicherung und Aufbereitung von Daten, Entwicklung von Algorithmen; Trainieren und Anwenden von Prognosemodellen; Prüfung und Validierung von Ergebnissen; Visualisierung von Ergebnissen gegenüber unterschiedlichen Zielgruppen.

In der gegenwärtigen Konfiguration ist das System leistungsfähig genug, um für die stündlich 45 Ankunftsereignisse in Bern und mit einem Vorlauf von 90 Minuten Prognosen im Minutentakt zu erstellen (es werden also 4000 Prognosen pro Stunde berechnet). Die Latenz (Durchlaufzeit von der ersten Betriebsdaten-Anfrage bis zum Vorliegen aller Prognosen) beträgt 40-50 Sekunden, wovon etwas mehr als die Hälfte auf Bezug und Verarbeitung der Echtzeitdaten entfällt.

In diesem Bereich besteht aber auch Verbesserungspotential: Die derzeitige Implementierung ist «gedächtnislos»: sie bezieht jede Minute aufs Neue alle benötigten Daten zur Betriebslage. Eine stärkere Entkopplung von Datenquelle (Open Data Plattform) und Datenverarbeitung (Prognosemodelle) wäre wünschenswert, insbesondere im Hinblick auf die Robustheit: kommt es zu Fehlern beim Bezug der Daten, ist keine Prognose möglich; kommt es zu fehlenden Werten in der gelieferten Antwort, verschlechtert sich der Prognose-Input. In beiden Fällen wären die gesuchten Werte evtl. aus dem vorherigen Durchlauf bekannt. Die Situation liesse sich durch Separierung von Datenbezug und Prognoseberechnung verbessern. Als zwischengeschalteter Cache könnte zum Beispiel eine Redis-Datenbank dienen. Auf diese Weise würde ausser der Robustheit auch die Effizienz erhöht: Da Echtzeitdaten von der Open Data Plattform in einem komplexen und etwas überladenen XML-Format bereitgestellt werden, ist das Parsen der Daten aufwendig. Die Zahl der abgefragten Haltestellen ist aktuell begrenzt werden, damit genügend Zeit für die anschliessende Prognoserechnung bleibt. Mit einem Cache liessen sich die Prozesse parallelisieren, eine Synchronisierung wäre nicht nötig. Auch könnten Haltestellen je nach Wichtigkeit in unterschiedlichen Intervallen abgefragt werden.

Verbesserungswürdig ist auch die Analyse-Applikation: diese hat sich als hilfreich erwiesen, stösst aber angesichts der wachsenden Menge von Daten aus Simulation und Echtzeit-Prognose an ihre Grenzen. Das Datenmodell müsste angepasst werden, um schnellere Abfragen zu ermöglichen. Auch sind verschiedene weitere Features wünschenswert (z.B. Visualisierung der Betriebslage; Darstellung der generierten Entscheidungsbäume; Auswertungen über die Häufigkeit, in der Prädiktoren angewendet werden). Für die einfache Übertragung des Verfahrens auf andere Anwendungsszenarien wären Hilfsmittel gut, mit denen die Auswahl der abzufragenden Haltestellen erleichtert wird.

Sehr wertvoll war die Möglichkeit, im Rahmen der Arbeit den R-Cluster der BFH für aufwändige Berechnungen verwenden zu dürfen. **Einen grossen Dank möchte ich an dieser Stelle Daniel Baumann und dem Team der BFH-Systemadministratoren aussprechen, die – auch zu ungewöhnlichen Tageszeiten – mehrfach tollen und schnellen Support geleistet haben!**

8.6 Erkenntnisse zu Fahrplan und Betriebsablauf

Die für Prognosezwecke erstellten Modelle lassen sich grundsätzlich auch für die Analyse von Fahrplan und Betriebsabläufen nutzen. Exemplarisch sollen einige Beispiele gezeigt werden, die «als Nebenprodukt» der Arbeit angefallen sind und interessante Einblicke über fachliche Zusammenhänge vermitteln. Die meisten der Beispiele sind «Zufallsfunde». Eine fachliche Validierung mit Experten der Bahnunternehmen hat noch nicht stattgefunden.

Abbildung 65 zeigt einen sehr einfachen Entscheidungsbaum. Zielvariable ist die um mindestens eine Minute verspätete Ankunft der S-Bahn 2 aus Langnau (planmässig zur Minute 17).

Entscheidungsbaum für Prognose der S2 von Langnau:
«Ankunft Wankdorf -> Bern (planmässig: Minute 17)
mindestens 1 Minute verspätet»

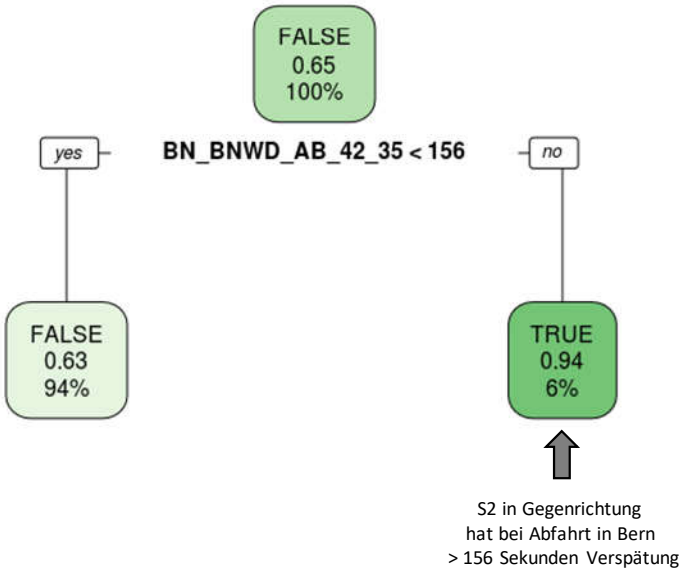


Abbildung 65: Entscheidungsbaum für die S2: Verspätung lässt sich aufgrund Pünktlichkeit des Gegenzugs vorhersagen

Der Baum verwendet nur einen einzigen Prädiktor. Interessant ist, dass es sich dabei nicht um ein Ereignis aus dem Fahrtverlauf der S2 von Langnau handelt, sondern um einen Zug in Gegenrichtung. Interessant ist weiterhin, dass der Prognosehorizont relativ gross ist: Bereits 35 Minuten im Voraus kann eine Verspätung prognostiziert werden, wenn die Abfahrt des Gegenzugs in Bern um mehr als 156 Sekunden verspätet ist. Der Zusammenhang ist zunächst nicht offensichtlich. Erst ein Studium von Fahrplan und Gleisanlage zeigt: der Zug von Langnau muss in Worb SBB die Kreuzung mit jenem aus Bern abwarten[125], wobei kaum Reserven bestehen. Offensichtlich kann eine verspätete Abfahrt in Bern bis Worb nicht aufgeholt werden, so dass sich die Verspätung auf den Gegenzug überträgt.

Ähnliche Zusammenhänge zeigen sich bei der S-Bahn-Linie 6: Zwischen Bern und Schwarzenburg existieren zahlreiche einspurige Abschnitte, so dass auch dort häufig die Kreuzung eines Gegenzugs abgewartet werden muss. Eine dieser Kreuzungsstellen ist in Köniz. Diesmal ist die Situation ein wenig komplizierter, wie Abbildung 66 zeigt:

[125] Zwischen Tägertschi und Worb SBB ist die Strecke eingleisig.

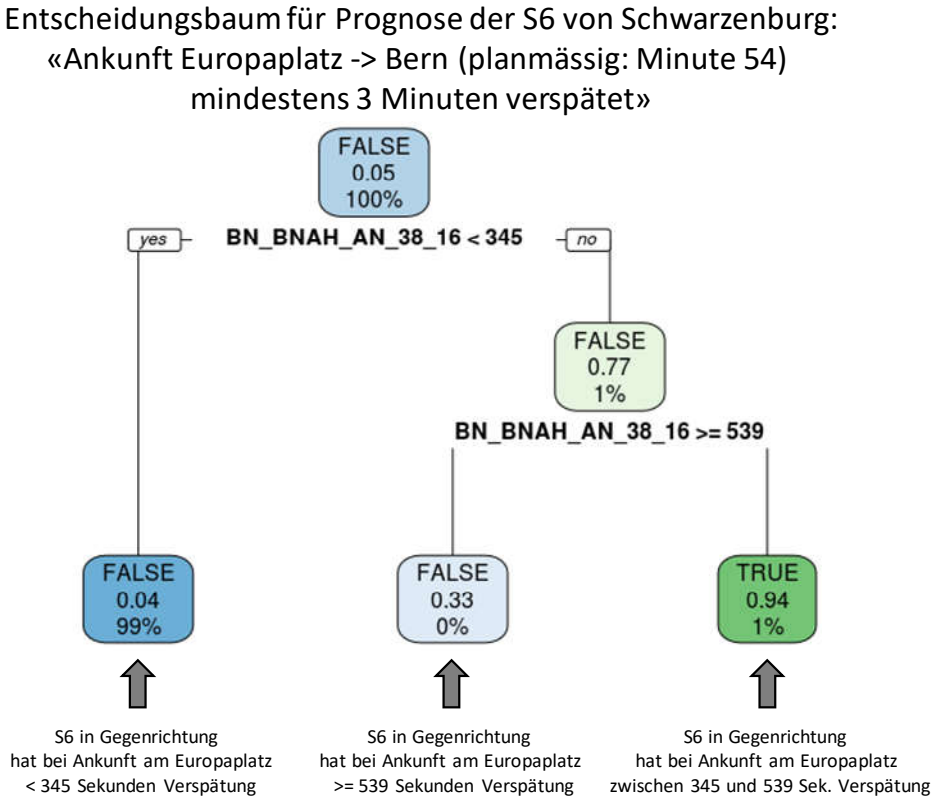


Abbildung 66: Entscheidungsbaum für die S6: Bei starker Verspätung wird die Kreuzungsstelle verlegt

Zielvariable ist hier eine mindestens 3-minütige Ankunftsverspätung der S6 (planmässig zur Minute 54). Prädiktor ist die Ankunft des Gegenzugs an der Haltestelle Europaplatz[126]. Verspätungen unterhalb von 345 Sekunden haben keinen Einfluss auf die Zielvariable. Höhere Verspätungen wirken sich dagegen deutlich aus – jedoch nur bis zu einem Umfang von 539 Sekunden. Jenseits dieses Werts wird erneut `FALSE` (also: keine Verspätung) prognostiziert. Die vermutete Ursache für diese «Unstetigkeit» besteht darin, dass bei grossen Verspätungen die Kreuzungsstelle der beiden Züge verschoben wird: Der aus Schwarzenburg kommende Zug befährt den eingleisigen Abschnitt dann ausnahmsweise vor dem stark verspäteten Gegenzug aus Bern.

Wie können solche Zusammenhänge systematisch gefunden werden? Einen ersten guten Überblick erhält man, in dem man die Prädiktoren einer Zielvariablen auf der Landkarte verortet. So zeigt Abbildung 67 alle Fahrtabschnitte, die Einfluss auf das Ereignis «Ankunft IC von Olten zur Minute 26» haben. Die Darstellung liefert Indizien dafür, dass in Aarau und Brugg Verspätungen von anderen Zügen übertragen werden. Die entsprechenden Entscheidungsbäume könnten dann genutzt werden, um dies genauer zu analysieren.

[126] Die Abkürzung BNAH steht für «Bern Ausserholligen». Die Haltestelle wurde vor einiger Zeit in «Bern Europaplatz» umbenannt; das Kürzel aber beibehalten.

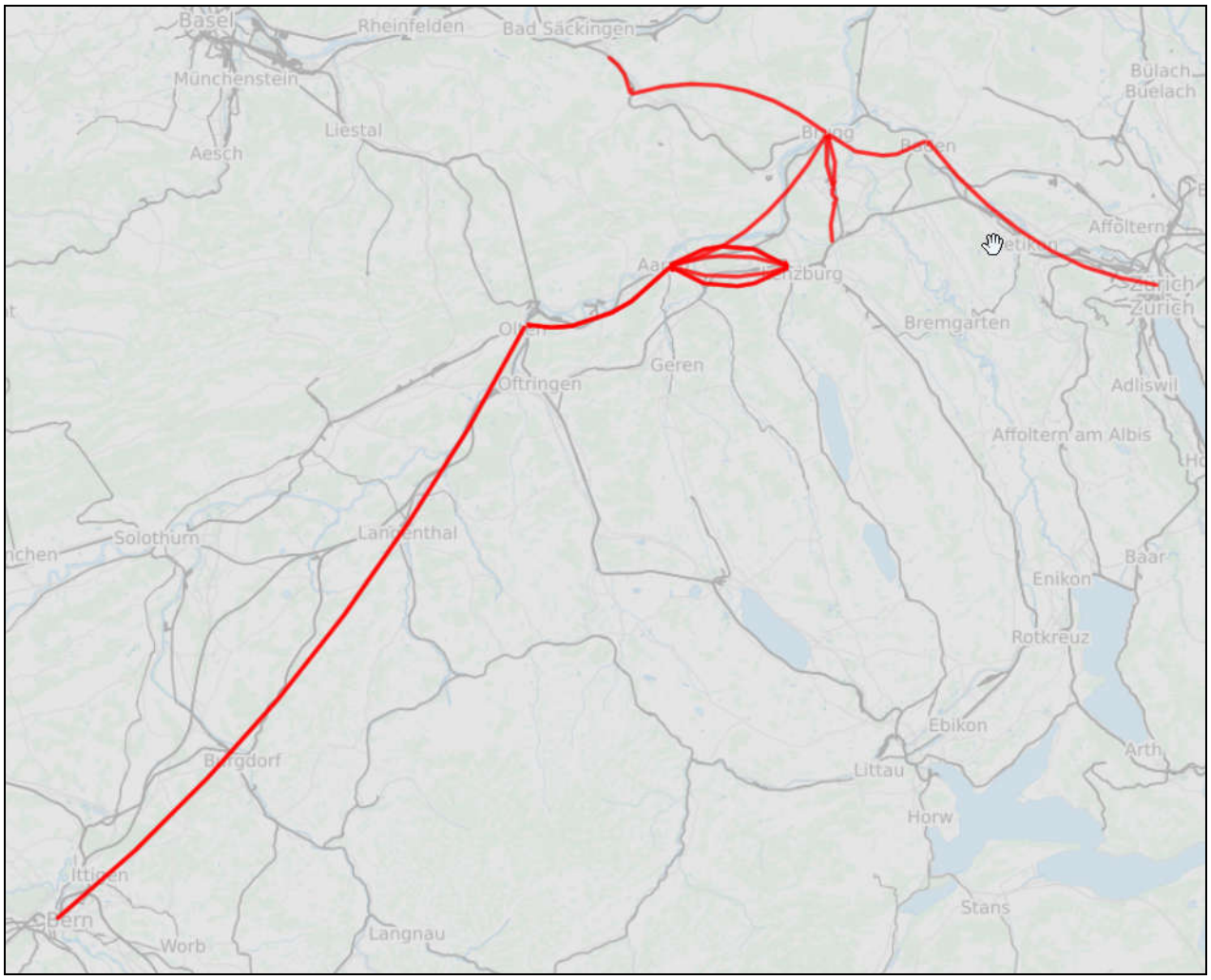


Abbildung 67: Verortung von Prädiktoren auf der Landkarte: Fahrtabschnitte mit Einfluss auf den IC aus Olten um :26

Neben den Modellen könnte auch deren Anwendung im Rahmen der Prognose (simuliert oder Echtzeit) wertvolle Hinweise geben: Welche Modelle haben oft dazu geführt, dass eine Verspätung prognostiziert wurde? Welche Prädiktoren kamen dabei am häufigsten zur Anwendung? Welche Züge beeinflussen sich gegenseitig besonders stark? Die bei der Prognose erstellen Logs bieten reichlich Material für die Analyse derartiger Fragen.

9 Zusammenfassung und Ausblick

Ziel dieser Arbeit war die Beantwortung der folgenden Forschungsfrage:

> **Können unter Verwendung der empirischen Vergangenheitswerte von opentransportdata.ch mindestens in Teilbereichen bessere Verspätungsprognosen erzielt werden, als sie von den aktuell im öV Schweiz eingesetzten Kundeninformationssystemen geliefert werden?**

Zu diesem Zweck wurde ein Vorhersageverfahren entwickelt, das auf Recursive Partitioning basiert. Da sich Verfügbarkeit und Stellenwert der Prädiktoren im Zeitablauf stark verändern, wird eine grosse Zahl von Entscheidungsbäumen generiert, aus denen in der konkreten Vorhersagesituation die anwendbaren ausgewählt werden. Die dafür notwendigen Algorithmen wurden ausführlich beschrieben.

Das Verfahren wurde auf der bestehenden Systemarchitektur von www.puenktlichkeit.ch implementiert. Es kann auf historische Daten und in Echtzeit angewendet werden. Als Untersuchungsszenario dienten die Ankünfte aller taktmässig verkehrenden Normalspur-Züge am Bahnhof Bern.

Die erstellten Skripte und Applikationen bietet die Möglichkeit

- Prognosemodelle zu erstellen,
- Die Anwendung der Modelle auf historischen Daten zu simulieren,
- Echtzeitprognosen zu erstellen,
- Echtzeitprognosen von den Referenzsystemen der Bahnunternehmen zu beziehen,
- Die Qualität von Prognosen (Echtzeit und Simulation) zu bewerten,
- Die Qualität der Prognosen mit jenen der Bahnunternehmen zu vergleichen.

Abbildung 68 zeigt noch einmal die Zusammenhänge.

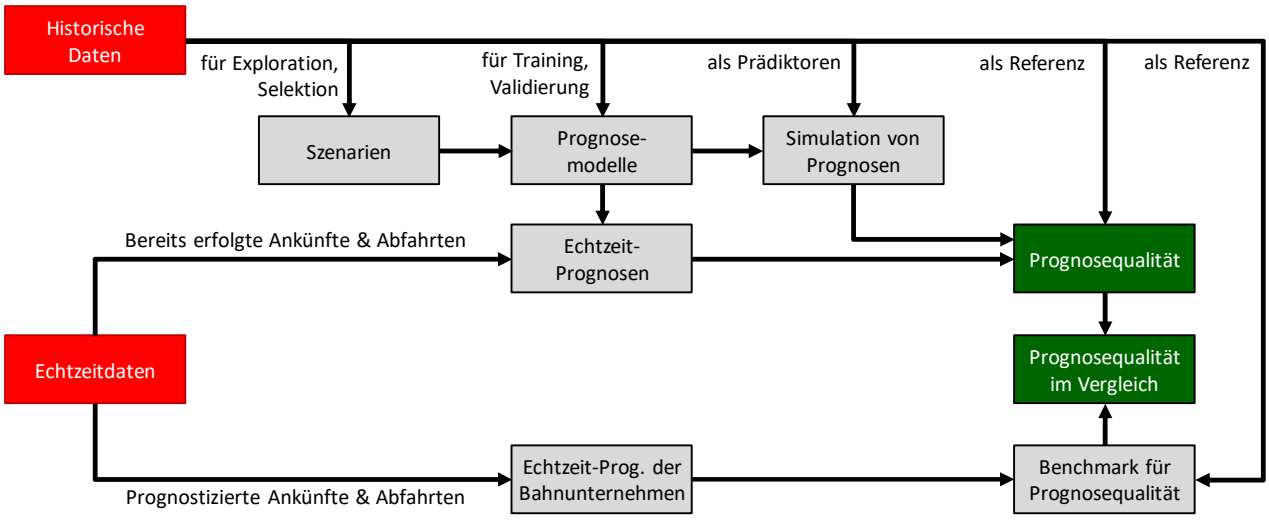


Abbildung 68: Verwendung von historischen Daten und Echtzeitdaten im Rahmen der Arbeit.

Die Echtzeitprognose ist seit Mitte Januar in Betrieb und liefert im Minutenrhythmus Verspätungsprognosen für die in Bern ankommenden Züge. Die Vorhersagen sind öffentlich auf der Website www.puenktlichkeit.ch einsehbar und können mit jenen der Bahnunternehmen verglichen werden.

Die Forschungsfrage kann klar mit «Ja» beantwortet werden: Das Verfahren liefert unter Verwendung öffentlich verfügbarer Daten Verspätungsprognosen, deren Qualität jene der Bahnunternehmen bis ca. 5 Minuten vor Ankunft des Zuges deutlich übertreffen, sofern

a) aufgrund von Unterwegs-Halten ausreichend häufig Informationen über den betrachteten Zug bekannt sind (spätestens 30 Minuten vor dem betrachteten Ereignis),

b) ausreichend viele Erfahrungswerte aus der Vergangenheit vorliegen (mindestens 2 Monate bei stündlich verkehrenden Zügen),

c) es sich um kleine bis mittlere Verspätungen handelt (bis etwa 10 Minuten).

Die dabei verwendeten Daten sind stark limitiert im Vergleich zu jenen, die den Bahnunternehmen intern zur Verfügung stehen: Die zeitliche Auflösung von Echtzeit-Informationen ist nur minutengenau. Während der Fahrt eines Zuges sind keine Positionsmeldungen verfügbar. Es fehlen Informationen zum eingesetzten Rollmaterial, Dispositions-Entscheidungen und Fahrwegen. Güter- und Dienstzügen sind gar nicht enthalten.

Wie ist es möglich, dass dennoch bessere Prognosen erzielt werden können? Und dies, obwohl die Schweizer Bahnen über Jahrzehnte sehr stark in ihre IT-Systeme investiert haben?

Ich habe keine abschliessende Antwort auf diese Frage, aber ich vermute, dass es mit einem zu schwach ausgeprägten «Feedback-Loop» zusammenhängt: Die Komplexität der verwendeten Fahrplan- und Prognosemodelle hat über die Jahre stetig zugenommen. Die heute im Einsatz befindlichen Systeme und Verfahren sind äussert elaboriert. Jede Verfeinerung eines Modells geht aber auch mit zusätzlichen Prämissen einher: Implizit oder explizit fliessen Annahmen und Parameter über die tatsächlichen Wirkungszusammenhänge in die Modelle ein. Und diese Annahmen werden möglicherweise zu wenig überprüft. Die Bahn-Branche hat eine starke Ingenieurs-Tradition. Dies hat die Entwicklung exzellenter Technologien, Verfahren und Modelle ermöglicht. Etwas weniger ausgeprägt ist möglicherweise die «empirische Tradition»: Es erscheint wünschenswert, die Ergebnisse der vorhandenen Modelle systematischer mit dem zu vergleichen, was tatsächlich im Bahnbetrieb passiert. Die dafür erforderlichen Daten sind in grossem Umfang vorhanden. Ich bin der festen Überzeugung, dass Verkehrsunternehmen noch sehr viel mehr Nutzen daraus ziehen könnten, als sie es bisher tun.

Das hier entwickelte Verfahren zielt auf die Untersuchung der Forschungsfrage. Es ist bei weitem nicht geeignet, um die Prognosemodelle der Bahnen zu ersetzen. Doch die drei oben genannten Limitationen können angegangen werden: Informationen zu fahrenden Zügen sind bei den Unternehmen vorhanden und lassen sich nutzen. Benötigt werden weiterhin Mechanismen, die eine zufriedenstellende Prognose selbst dann ermöglichen, wenn die Datenlage für den empirischen Ansatz unzureichend ist. Eine Kombination von klassischem und empirischem Ansatz erscheint dafür sinnvoll. Bereits mit der simplen Regel «Nimm von beiden Verfahren die höhere Prognose» liesse sich die Prognosequalität deutlich erhöhen – und die False-Positive-Rate läge noch immer unter 1 Promille. Auch die Kombination mit Regressionsmodellen könnte lohnend sein: damit liessen sich empirisch gestützte Werte für Fahrt- und Haltedauern ermitteln. Diese könnten immer dann für die Prognose herangezogen werden, wenn keine «besseren» Regeln vorhanden sind.

Prognosemodelle, die auf Basis empirischer Daten trainiert wurden, können aber noch mehr leisten: Wenn – wie in dieser Arbeit – ein Modelltyp gewählt wird, der leicht visualisiert und interpretiert werden kann, so lassen sich damit auch neue Zusammenhänge entdecken. Solche Einsichten sind wertvoll in vielen Prozessen der Bahn-Branche: bei der Weiterentwicklung des Gleisnetzes, der gezielten Beseitigung von Engpässen und Verspätungsursachen, der Konstruktion noch robusterer Fahrpläne, der richtigen Reaktion beim Auftreten von Störungen, der Wahl einer energieeffizienten Fahrweise.

Begründen Machine Learning-Algorithmen das Ende der Theorie? Hoffentlich nicht. Wenn sie genutzt werden, um bestehende Annahmen zu hinterfragen und unbekannte Zusammenhänge aufzudecken, dann schliessen sie vielmehr den «Feedback Loop» und befruchten die Entwicklung besserer Theorien. Und das wäre noch viel wertvoller als der Gewinn von einigen Prozent Prognosequalität.

10 Abbildungsverzeichnis

11 Literaturverzeichnis

Anderson (2008): The End of Theory: The data deluge makes the scientific method obsolete, Wired, 23. Juni 2008. Online: https://www.wired.com/2008/06/pb-theory/.

Archer Kellie J. (2010): rpartOrdinal: An R Package for Deriving a Classification Tree for Predicting an Ordinal Response. Journal of Statistical Software, 34(7), 1–17. Online: http://www.jstatsoft.org/v34/i07/.

BAV (2016): Bund schafft Voraussetzungen für Apps mit Echtzeitinformationen zum öffentlichen Verkehr, Medienmitteilung des Bundesamts für Verkehr vom 1. Dezember 2016. Online: https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-64744.html.

Berger, Annabell; Gebhardt, Andreas; Müller-Hannemann, Matthias; Lemnian, Martin (2011): Stochastic Delay Prediction in Large Train Networks. OpenAccess Series in Informatics, 2011 (20), S. 100-111.

Breiman L.; Friedman J. H.; Olshen R. A.; Stone, C. J. (1984): Classification and Regression Trees. Wadsworth.

Brownlee Jason (2017): Machine Learning Mastery: What is the Difference Between a Parameter and a Hyperparameter?. Online: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/

Brownlee, Janson (2017a): What is the Difference Between Test and Validation Datasets?. Online: https://machinelearningmastery.com/difference-test-validation-datasets/.

Büker, Thorsten (2010): Ausgewählte Aspekte der Verspätungsfortpflanzung in Netzen. Dissertation, RWTH Aachen. Online: http://publications.rwth-aachen.de/record/63831/files/3469.pdf.

Butler, Heidrun (2017): Wie entsteht der Fahrplan. Vortrag bei der Gesellschaft der Ingenieure des öffentlichen Verkehrs (GDI). Online: http://www.gdi-adi.ch/uploads/tx_webuseevents/170301_Referat_GdI_-_wie_entsteht_der_Fahrplan.pdf.

Caimi, Gabrio; Kroon, Leo; Liebchen, Christian (2017): Models for railway timetable optimization: Applicability and applications in practice. Journal of Rail Transport Planning & Management. 6 (2017) 285-312.

CEN (2015): CEN TS 15531, Service Interface for Realtime Information (SIRI), Beuth Verlag, 2015.

Chawla, Nitesh V.; Bowyer, Kevin W.; Hall, Lawrence O.; Kegelmeyer, Philip W. (2002): SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16 (2002), S. 321-357.

Cieslak, David A.; Chawla, Nitesh V. (2008): Learning Decision Trees for Unbalanced Data. In: Daelemans W., Goethals B., Morik K. (Hrsg.), Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008. Lecture Notes in Computer Science, vol 5211. Springer, 2008. Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.151.3829&rep=rep1&type=pdf.

Conte, Carla (2007): Identifying Dependencies Among Delays. Dissertation, Universität Göttingen, 2007. Online: http://webdoc.sub.gwdg.de/diss/2008/conte/.

Cule, Boris, Goethals Bart, Tassenoy Sven, Verboven Sabine (2011): Mining Train Delays. In: Gama J., Bradley E., Hollmén J. (Hrsg.): Advances in Intelligent Data Analysis X. Springer, 2011.

Dittes, Frank-Michael (2012): Komplexität. Warum die Bahn nie pünktlich ist. Springer Vieweg, 2012.

Dolder, Urs; Krista, Matthias; Völcker, Marcus (2009): Real Time Train Run Forecast. Präsentation beim 3rd International Seminar on Railway Operations Modelling and Analysis. RailZurich 2009. Online: http://webarchiv.ethz.ch/ivt/news/railzurich2009/presentations/d2_dolder.pdf.

Doshi-Velez, Finale; Kim, Been (2017): Towards A Rigorous Science of Interpretable Machine Learning. Online: http://arxiv.org/abs/1702.08608.

Flier Holger; Gelashvili Rati; Graffagnino Thomas; Nunkesser Marc (2009): Mining Railway Delay Dependencies in Large-Scale Real-World Delay Data. In: Ahuja R.K., Möhring R.H., Zaroliagis C.D. (Hrsg.), Robust and Online Large-Scale Optimization. Lecture Notes in Computer Science 5868, S. 354-368.

Frank Eibe; Hall Mark (2001) A Simple Approach to Ordinal Classification. In: De Raedt L., Flach P. (eds) Machine Learning: ECML 2001. ECML 2001. Lecture Notes in Computer Science, vol 2167. Springer. Online: https://www.cs.waikato.ac.nz/~eibe/pubs/ordinal_tech_report.pdf .

Freund Yoav, Schapire Robert (1996): Experiments with a New Boosting Algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, S. 148–156.

Friedman, Jerome H. (1991): Multivariate Adaptive Regression Splines. In: The Annals of Statistics, 19, S. 1-67.

Galimberti, Giuliano; Soffritti, Gabriele; Di Maso, Matteo (2012): Classification Trees for Ordinal Responses in R: The rpartScore Package. Journal of Statistical Software. 47. 1-25. Online: https://www.jstatsoft.org/article/view/v047i10/v47i10.pdf.

Gerny, Daniel (2016): Das Smartphone wird zum wichtigsten Verkehrsmittel, in: Neue Zürcher Zeitung vom 2. Dezember 2016, online: https://www.nzz.ch/schweiz/aktuelle-themen/open-data-das-smartphone-wird-zum-wichtigsten-verkehrsmittel-ld.131923.

Guo, Xinjian; Yin, Yilong; Dong, Cailing; Yang, Gongping; Zhou, Guangtong (2008): On the Class Imbalance Problem, Fourth International Conference on Natural Computation, 2008. Online: http://sci2s.ugr.es/keel/pdf/specific/congreso/guo_on_2008.pdf.

Gutweniger, Andreas (2017): puenktlichkeit.ch. Publikation von Pünktlichkeits-Auswertungen zum Öffentlichen Verkehr auf Basis von Open Data, Semesterarbeit im CAS Business Intelligence der BFH, HS 2016/17, unveröffentlicht. (eine elektronische Version ist dieser Arbeit beigefügt)

Gutweniger, Andreas (2017a): puenktlichkeit.ch. Analysen zum öV Schweiz, Autorenprojekt im CAS Data Visualization der BFH, FS 2017, unveröffentlicht. (eine elektronische Version ist dieser Arbeit beigefügt)

Hansen, Ingo A.; Goverde, Rob; van der Meer, Dirk J. (2010): Online train delay recognition and running time prediction. Proceedings of the IEEE Conference on Intelligent Transportation Systems, S. 1783-1788.

Japkowicz, Nathalie (2000): Class Imbalance Problem: Significance & Strategies, International Conference on Artificial Intelligence (ICAI). (2000) 111–117. Online: https://pdfs.semanticscholar.org/907b/02c6322d0e7dff6b0201b03e3d2c6bc1d38f.pdf.

Kecman, Pavle (2014): Models for Predictive Railway Traffic Management. Dissertation, Delft University of Technology, 2014. Online: https://repository.tudelft.nl/islandora/object/uuid:539e96b3-18d7-4f6f-b662-ce22ae269f2a/datastream/OBJ/download.

Kecman Pavle; Corman Francesco; D'Ariano, Andrea; Goverde, Rob M.P. (2013): Rescheduling models for railway traffic management in large-scale networks, Public Transport, 5 (2013), S. 95-123.

Kuhn, Max (2017): The Caret Package. Online: http://topepo.github.io/caret/index.html.

Kuhn, Max; Johnson, Kjell (2013): Applied Predictive Modelling, Springer, 2013.

Liebchen, Christian (2008): Optimization of passenger timetables: are fully-integrated, regular-interval timetables really always the best? European Rail Technology Revue 48 (4), S. 13-19.

Luthi, Marco (2009): Improving the Efficiency of Heavily Used Railway Networks through Integrated Real-Time Rescheduling. Dissertation, ETH Zurich, 2009. Online: https://pdfs.semanticscholar.org/19a6/ae0ce24397e9506df86878bd6f0bf538b743.pdf.

Lüthi, Marco; Nash, Andrew; Weidmann, Ulrich; Laube, Felix; Wüst, Raimond (2007): Increasing Railway Capacity and Reliability through Integrated Real-Time Rescheduling. In: Proceedings of the 11th World Conference on Transport Research. Online: https://www.research-collection.ethz.ch/handle/20.500.11850/64806.

Menardi Giovanna, Torelli Nicola (2014): Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery. 28 (1), S. 92-122.

Molnar, Christoph (2018): Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Online: https://christophm.github.io/interpretable-ml-book/.

Oneto, Luca; Fumeo, Emanuele; Clerico, Giorgio; Canepa, Renzo; Papa, Federico; Dambra, Carlo; Mazzino, Nadia; Anguita, Davide (2017): Delay Prediction System for Large-Scale Railway Networks Based on Big Data Analytics. In: Angelov P., Manolopoulos Y., Iliadis L., Roy A., Vellasco M. (Hrsg.): Advances in Big Data. Springer, 2017.

Delay Prediction System for Large-Scale Railway Networks Based on Big Data Analytics." INNS Conference on Big Data (2016).

Open Data Plattform öV Schweiz (2016): Cookbook. Online: https://opentransportdata.swiss/de/cookbook/.

Pachl, Jörn (2016): Systemtechnik des Schienenverkehrs: Bahnbetrieb planen, steuern und sichern. Springer Vieweg, 8. Auflage, 2016.

Potharst, Rob, Bioch, Jan C., Petter, Thijs C. (1997): Monotone decision trees. Technical Report EUR-FEW-CS-97-06, Erasmus University Rotterdam. Online: https://repub.eur.nl/pub/522/fewinf19971006113242.pdf.

Potharst Rob, Bioch Jan C. (1999): A Decision Tree Algorithm for Ordinal Classification. In: Hand D.J., Kok J.N., Berthold M.R. (Hrsg.) Advances in Intelligent Data Analysis. IDA 1999. Lecture Notes in Computer Science, vol 1642. Springer. Online: https://link.springer.com/chapter/10.1007/3-540-48412-4_16.

Quinlan, John Ross (1986): Induction of Decision Trees. Machine Learning 1, 1 (März 1986), S. 81–106.

Quinlan, John Ross (1992): Learning with Continuous Classes. Proceedings of the 5th Australian Joint Conference On Artificial Intelligence, S. 343–348.

Quinlan, John Ross (1993): C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

Quinlan, John Ross (1993a): Combining Instance-Based and Model-Based Learning. Proceedings of the Tenth International Conference on Machine Learning, S. 236–243.

Russel, Stuart; Norvig, Peter (2009): Artifical Intelligence: A Modern Approach, Pearson, 3. Auflage, 2009.

SBB (2010): NeTS. Handbuch Knotenplanung. Online: https://www.sbb.ch/content/dam/sbb/de/pdf/sbb-konzern/sbb-als-geschaeftspartner/NeTS/NeTS_NeTS-Plan_Benutzerhandbuch%20NeTS-Plan_Kurzhandbuch%20Knotenplaner_D.pdf.

SBB (2014): Entflechtung Wylerfeld. Mehr Zug für Bern und die Schweiz. Online: https://company.sbb.ch/content/dam/sbb/de/pdf/sbb-konzern/ueber-die-sbb/projekte/ausbau-schienennetz/Broschuere_Wylerfeld_2014.pdf.

SBB (2016): Zugsicherung ETCS. Online: https://company.sbb.ch/de/ueber-die-sbb/verantwortung/die-sbb-und-ihre-kunden/die-sbb-und-die-sicherheit/zugsicherung-etcs.html.

SBB (2017): Rail Control System (RCS) – Das Bahnverkehrskonzept der Zukunft. White Paper. Online: https://www.sbbrcs.ch/fileadmin/user_upload/bilder/broschuere/RCS_Whitepaper_de.pdf.

SBB (2017b): RCS-Dispo – Das Dispositionssystem für alle Fälle. Online: https://www.sbbrcs.ch/fileadmin/user_upload/bilder/broschuere/Broschuere_DISPO_DE_Web.pdf.

Scholz, Gero (2012): IT-Systeme für Verkehrsunternehmen. dpunkt.verlag 20112.

sma (2017): Netzgrafik Fahrplan Schweiz 2018. Online: http://www.sma-partner.com/images/Downloads/NGCH-2018.pdf.

Therneau, Terry M.; Atkinson, Elizabeth J. (2018): An Introduction to Recursive Partitioning Using the RPART Routines. Online: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf .

Ullius, Markus (2005): Verwendung von Eisenbahnbetriebsdaten für die Schwachstellen- und Risiko-analyse zur Verbesserung der Angebots- und Betriebsqualität. Dissertation, ETH Zürich, 2005.

VDV (2015): Echtzeit Kommunikations- und Auskunftsplattform EKAP, Teil 2: EKAP-Schnittstellenbeschreibung V1.1, Verband Deutscher Verkehrsunternehmen, 2015. Online: https://www.vdv.de/431-2sds-v1.1.pdfx.

VöV (2016): Fakten zum öffentlichen Verkehr der Schweiz 2016/2017. Verband öffentlicher Verkehr. Online: https://www.voev.ch/de/Service/content_index.php?section=downloads&download=2207.

Wang Ren; Work, Daniel B. (2015): Data Driven Approaches for Passenger Train Delay Estimation. Proceeedings of the IEEE 18th International Conference on Intelligent Transportation Systems, Las Palmas, 2015, S. 535-540.

12 Anhang: Beilagen zur Master Thesis

Gemeinsam mit dieser Arbeit werden die folgenden Beilagen abgegeben.

Source Code

Datei	*Verwendung*
Fahrten_vorbereiten.R	Skript zum Auslesen und Aufbereiten der benötigten Trainingsdaten (Fahrten und Takte) aus der Datenbank.
Modelle_erstellen.R	Funktion (parametrisierbar), mit der Entscheidungsbäume für unterschiedliche Variablen-Konstellationen und Vorlaufzeiten erstellt werden.
Aufruf_Modellerstellung.R	Skript zum parallelisierten Aufruf der Modellerstellung im R-Cluster.
Modelle_anwenden.R	Funktion (parametrisierbar), mit der die Modellanwendung für einen vollständigen Betriebstag simuliert werden kann.
Aufruf_Anwendung.R	Skript zum parallelisierten Aufruf der Modellanwendung / Simulation im R-Cluster.
RT_PrognosenBern.R	Ermittelt Echtzeit-Prognosen im Szenario Bern; wird 1x pro Minute ausgeführt
app.R	Shiny-Applikation «Analyse»
prognose_util.R	Enthält API Credentials und häufig gebrauchte Hilfsfunktionen.

Unveröffentlichte Quellen

Datei	*Inhalt*
CAS_BI.pdf	Gutweniger, Andreas (2017): puenktlichkeit.ch. Publikation von Pünktlichkeits-Auswertungen zum Öffentlichen Verkehr auf Basis von Open Data, Semesterarbeit im CAS Business Intelligence der BFH, HS 2016/17.
CAS_DATAVIZ.pdf	Gutweniger, Andreas (2017a): puenktlichkeit.ch. Analysen zum öV Schweiz, Autorenprojekt im CAS Data Visualization der BFH, FS 2017.

13 Selbständigkeitserklärung

Ich bestätige, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der im Literaturverzeichnis angegebenen Quellen und Hilfsmittel angefertigt habe. Sämtliche Textstellen, die nicht von mir stammen, sind als Zitate gekennzeichnet und mit dem genauen Hinweis auf ihre Herkunft versehen.

Andreas Gutweniger, Bremgarten bei Bern, 12. März 2018