Einblick in Bahndaten aus der Schweiz: Analyse von Pünktlichkeit des öffentlichen Verkehrs

Ozan Tastekin
st178902@stud.uni-stuttgart.de
Universität Stuttgart
Baden-Württemberg, Deutschland

Zusammenfassung

Im öffentlichen Verkehr ist die Pünktlichkeit einer der größten Faktoren, was Kunden wichtig ist. Um zu wissen, warum ein Fahrzeug des öffentlichen Verkehrs sich verspätet sind Analysen von hohem Nutzen. Daher werden in diesem Paper drei Datensätze der Verkehrsbetriebe der Stadt Zürich (VBZ) untersucht und es werden zwei Hypothesen analysiert. In diesen geht es um die Frage, was dazu führt, dass ein Bus oder eine Bahn sich verspätet. In Hypothese 1 wird analysiert, ob die Anzahl an Ein- und Aussteigern einen Einfluss auf die Verspätungen hat, also ob mehr Passagiere auch mehr Verspätungen mit sich bringen. Dafür werden drei Buslinien und eine Bahnlinie der Hardbrücke Haltestelle in Zürich betrachtet. In Hypothese 2 wird auf das Wetter eingegangen und ob der Regen für Busse heißt, dass sie sich verspäten werden. Hierbei werden alle Busse des ganzen VBZ angeschaut und deren Verspätungen mit der Regendauer in Zürich analysiert.

Schlagwörter: Öffentlicher Verkehr; Verspätungen; Zürich; Fahrzeiten Soll und Ist; Passagierfrequenzen; Regendauer

1 Einleitung

Der öffentliche Verkehr ist für viele Menschen von wichtigem Nutzen, sie fahren sehr oft mit Bussen und Bahnen, sei es zur Arbeit, nach Hause, zum Einkaufen oder um Freunde und Familie zu besuchen. Dabei ist es ihnen wichtig, dass die Bahnen und Busse rechtzeitig ankommen und an der Zielhaltestelle sind [6], [10], [9]. Es kann viele Gründe geben, warum ein Zug oder Bus sich verspätet, so kann es an Personalmangel, einem Unfall oder weiteren Gründen liegen [8]. Durch Analysen kann herausgefunden werden, was zu einer Verspätung führt und wie genau und dementsprechend agiert werden sollte, um die Kundenzufriedenheit zu steigern und damit auch den Umsatz.

In diesem Paper werden zwei Faktoren für Verspätungen in Zürich, Schweiz angeschaut. Dabei werden öffentliche Datensätze des VBZ näher betrachtet. Es wird analysiert, ob eine erhöhte Anzahl an Passagieren dazu führt, dass ein Verkehrsmittel mehr Verspätungen hat. Außerdem noch, ob bei einer höheren Regendauer in einer Stunde Busse mehr verspätet sind.

In Abschnitt 2 werden die Daten beschrieben, die genutzt werden und um die Datenqualität für Analysen zu sichern werden Bereinigungsschritte gezeigt. In Abschnitt 3 werden dann die Analysen durchgeführt, wobei Abschnitt 3.1 Grundlagen beschreibt und 3.2 und 3.3 die Hypothesen einführt und in diesen Abschnitten die Ergebnisse vorgestellt. In 3.4 werden dann Limitationen beschreiben, die die Analysen betrifft. Schließlich gibt es in Abschnitt 3.5 eine Diskussion über die Ergebnisse, wobei hier darauf eingegangen wird, warum sich diese Ergebnisse mit denen von Nagy und Csiszár unterscheiden. Zuletzt wird in Abschnitt 4 ein Fazit gezogen.

2 Datenbeschreibung und Bereinigung

In diesem Abschnitt werden die Datensätze, welche für die Analysen benötigt werden, beschrieben und welche Erkenntnisse aus der Datenexploration gezogen wurden erklärt. Außerdem werden die Schritte, die für die Datenbereinigung durchgeführt wurden beschrieben.

Alle drei benötigten Datensätze für die Analysen sind auf der Webseite https://data.stadt-zuerich.ch zu finden und es handelt sich um Daten über die Stadt Zürich in der Schweiz.

2.1 Fahrzeiten Soll und Ist Vergleich

Beim ersten Datensatz geht es um Fahrzeiten Daten von 2016 bis 2022. In dem Datensatz werden Daten über Züge und Busse aufgezeichnet und dabei die tatsächlichen Ankunfszeiten und die Abfahrtszeiten der Verkehrsmittel zu den geplanten Ankunfts- und Abfahrtszeiten gegenüber gestellt.

In Zürich gibt es Bus-, Tram- und Seilbahnlinien. Die Nummer der Linie sagt aus, welche Art von Verkehrsmittel es ist. Dabei sind Tramliniennummern zwischen 2 und 17, Seilbahnlinien zwischen 18 und 28 und Buslinien ab 29. In diesem Datensatz gibt es keine Einträge zu Seilbahnen und die Anzahl an anderen Linien variiert pro Jahr.

Bei diesem Datensatz hat jedes der sieben Jahre 52 Dateien, also eine Datei pro Woche im Jahr. Pro Verkehrsmittel, welches an einer Station ankommt und abfährt wird die oben beschriebenen Zeiten aufgezeichnet. Zudem wird auch noch gespeichert, um welche Linie es sich handelt und an welcher Station und dessen Gleis/Punkt das Verkehrsmittel stoppt.

Zudem werden zwei weitere Dateien bereit gestellt, Haltestellen und Haltepunkte, welche als Verbindungstabellen dienen sollen. Haltepunkte sind dabei Gleise oder Punkte an denen das Verkehrsmittel anhält, damit Passagiere ein- und aussteigen können. Somit wird bei den Fahrzeiten ein Fremdschlüssel gespeichert, welcher die Haltepunkt ID beinhaltet,

diese Haltepunkt ID verweist dann auf die Haltepunkte Tabelle und in dieser ist ein Fremdschlüssel für Haltestellen. Das heißt jeder Eintrag in den Fahrzeiten ist einem Haltepunkt zugeordnet und dieser wiederum ist einer Haltestelle zugeordnet. Jede Haltestelle kann dabei mehrere Haltepunkte zu sich zugeordnet haben.

Pro Jahr hat der Datensatz der Fahrzeiten über 70 Millionen Datenpunkte gespeichert und es werden 34 Attribute, fortan Spalten genannt, gespeichert. Bei den Haltestellen sind es 5 Spalten und Haltepunkte haben 7 Spalten. Im Weiteren wird dieser Datensatz bereinigt.

2.1.1 Entfernen von Attributen. In den Fahrzeiten sind einige Spalten enthalten, die unnötigerweise doppelt gespeichert werden, nämlich durch die Verbindung zwischen Fahrzeiten mit Haltepunkten und Haltepunkten mit Haltestellen. So wird zum Beispiel in den Fahrzeiten die Spalte *halt_kurz* gespeichert, welche die Kurzform der Haltestelle enthält, und es gibt auch in den Haltestellen eine solche Spalte, die die gleiche Information enthält. Deshalb werden solche Spalten entfernt.

Außerdem gibt es in den Haltestellen eine Spalte *ist_aktiv*, die beschreibt, ob diese Haltestelle aktiv genutzt wird. Hierbei ist jedoch für jeden Datenpunkt dieser Wert auf *True* gesetzt, was bedeutet, dass in jedem Jahr jede aufgeführte Haltestelle aktiv ist. Da sich dieser Wert nie ändert und diese Information keinen Mehrwert bietet, wird auch diese Spalte entfernt.

2.1.2 Dateien zusammenführen. Um weitere Operationen effizienter zu gestalten, werden die 52 csv Dateien zu einer zusammengeführt, indem die Datenpunkte jeder Datei aneinander gehängt werden, sodass die Zeilenanzahl der neuen Datei der Summe der Zeilenanzahlen der einzelnen Dateien entspricht.

Die Haltestellen Dateien werden alle gelöscht, außer der des letzten Jahres, denn es kommen über die Jahre immer nur Haltestellen hinzu und es verändert sich für jede Haltestelle nur in wenigen Fällen die Werte für Spalten. So haben sich bei weniger als 1,5% der Datenpunkte über die Jahre hinweg die Spalten *halt_lang* und *halt_kurz* geändert. Also verändert sich nur der Name der Haltestelle oder seine Abkürzung. Dies wird aber eine Analyse nicht beeinflussen.

Bei den Haltepunkten werden alle Dateien wie bei den Fahrzeiten aneinander gehängt, da hier über die Jahre hinweg die Werte der Spalte *ist_aktiv* geändert werden, was für eine Analyse wichtig ist. Außerdem wird noch eine Spalte *jahr* hinzugefügt.

Nach dem Schritt gibt es also nur noch eine Fahrzeiten Datei pro Jahr und zwei zentrale Verbindungsdateien, Haltepunkte und Haltestellen.

2.1.3 Datenpunkte in ihr korrektes Jahr speichern. Die Fahrzeiten sollten idealerweise vom ersten Januar des Jahres starten und beim 31. Dezember des Jahres enden. Dies ist aber nicht der Fall für jedes Jahr, weshalb die Datenpunkte in ihr korrektes Jahr gespeichert werden und aus dem falschen Jahr entfernt werden. Zum Beispiel war vor diesem Schritt der 01. Januar 2022 im Jahr 2021 gespeichert.

2.1.4 Duplikate entfernen. In diesem Schritt werden duplizierte Datenpunkte entfernt. Bei den Haltestellen und Haltepunkte gibt es dies nicht, doch bei den Fahrzeiten schon. In Tabelle 1 ist zu sehen, wie viele Zeilen entfernt wurden in welchem Jahr, weil es diese dupliziert gab.

Tabelle 1. Anzahl an duplizierten Zeilen in den Fahrzeiten

Jahr	Duplizierte Zeilen
2016	0
2017	4
2018	1
2019	1594
2020	1911
2021	8374
2022	246
Summe	12130

2.1.5 GPS-Daten hinzufügen. Nur die Haltepunkte Datei hat Datenpunkte mit Spalten, die leere Einträge haben. Es fehlen bei 7180 Zeilen (ungefähr 7%) alle GPS-Informationen: der Längengrad (longitude), Breitengrad (latitude) und die Ausrichtung (bearing). Die Ausrichtung kann dabei in einem Intervall von $[0 - 359]$ sein und sagt aus, in welche Richtung dieser Haltepunkt zeigt. Außerdem gibt es Datenpunkte, in denen die Ausrichtung allein fehlt, was 19401 Zeilen (ungefähr 18%) sind.

Im Open-Data-Plattform öffentlichen Verkehrs Schweiz ist eine Datei hinterlegt, welche die aktuellen Haltestellen auflistet und deren GPS-Daten beinhaltet. Dadurch werden zu den Haltestellen, welche keine GPS-Spalten vorher hatte solche hinzugefügt. Somit hat jede Haltestelle jetzt einen Standort auf der Karte.

Jeder Haltepunkt gehört zu einer Haltestelle und um obiges Problem, dass es leere Einträge gibt zu lösen, gibt es zwei Möglichkeiten, entweder die Spalten entfernen oder sie auffüllen. Zum Auffüllen konnten nicht GPS-Daten der Haltepunkte gefunden werden. Die nächstbeste Lösung ist daher, bei einem Haltepunkt, welches keine GPS-Koordinaten hat, die Koordinaten des zu dem Haltepunkt zugehörigen Haltestelle zu nehmen. Somit wird die Spalte behalten und die Werte sinnvoll aufgefüllt. Diese Werte sind aber nicht zu 100% genau, das wird bei zukünftigen Analysen beachtet falls nötig.

Bei der Ausrichtung kann so eine sinnvolle Auffüllung nicht durchgeführt werden, weil es keine Daten dazu gibt. Deshalb werden diese mit -1 gefüllt werden, was eigentlich

nicht in dem Wertebereich des Bearings liegt. Das signalisiert dann, dass hier der Wert aufgefüllt wurde und es keine Ausrichtung gefunden werden konnte.

2.2 Passagierfrequenzen der Hardbrücke-Haltestelle

Der Passagierfrequenz Datensatz handelt sich um eine Zählung an der Hardbrücke-Haltestelle in Zürich in den Jahren von 2020 bis 2023. Es wurden an zwei der vier Gleisen Sensoren eingebaut. Die Gleise sind gegenüberliegend, was heißt dass Passagiere nicht auf einer Plattform in Hin- und Rückrichtung einsteigen können, sondern es zwei Plattformen sind. Diese Gleise sind jeweils die Hin- und Rückrichtung der Buslinien 33, 72, 83 und der Tramlinie 8. Der Bereich, der von den Sensoren erfasst wird, ist in mehrere Abschnitte unterteilt. Dabei kann also unter anderem heraus gefunden werden, wie viele Fahrgäste vom Gleis in das Verkehrsmittel einsteigen, wie viele vom Verkehrsmittel aussteigen und auch wie viele Fahrgäste in den Bereich des Fahrradverleihs gehen. Die Sensoren zählen 5 Minuten lang und speichern diese 5 Minuten Intervalle dann ab.

Die Passagierfrequenzen speichern 4 Spalten ab und haben über 840.000 Datenpunkte pro Jahr. Im Folgenden wird die Datenbereinigung der Passagierfrequenzen beschrieben.

2.2.1 Fehlende Zeitstempel. Da dieser Datensatz von der Dimensionalität klein ist, ist auch das bereinigen einfacher als bei den Fahrzeiten. Der Datensatz besteht nur aus vier Spalten, einmal ein Zeitstempel, wie viele Personen eingestiegen sind, wie viele ausgestiegen sind und welcher Bereich damit gemeint ist.

In den Passagierfrequenzen sollte es alle 5 Minuten Datenpunkte geben laut Beschreibung, doch es fehlen einige. Dies betrifft 11641 Zeilen, was 11% vom Datensatz ist. Dabei ist kein auffälliges Muster zu sehen, warum diese Zeitstempel fehlen. Außerdem werden auf der Webseite des Datensatzes Ausfälle oder Störungen des Sensors aufgelistet und für den Zeitraum der hier gebraucht wird, 2022, wurde nichts derartiges notiert. Bei weiterem Betrachten der Daten fällt auf, dass es keine Zeilen gibt, wo gleichzeitig 0 Passagiere ein- und ausgestiegen sind. In Tabelle 2 wird genau das gezeigt, hierbei kann gesehen werden, wie viele Zeilen es pro Kombination von Ein- und Aussteigern gibt. Die Tabelle wird hier gekürzt, damit nur bis zu 5 Ein- und Aussteigern gesehen werden kann, um die Tabelle kleiner zu halten.

Dies lässt darauf schließen, dass wenn keiner ein- und aussteigt keine Zeile gespeichert wird im Datensatz. Somit werden diese Zeilen aufgefüllt, wo der Zeitstempel existieren sollte, aber das nicht tut und es werden bei den Spalten Ein- und Aussteigern eine 0 jeweils eingetragen.

2.3 Stündliche Wetterdaten

Der dritte Datensatz, befasst sich mit dem Wetter in Zürich von 2000 bis 2023. Hierbei wurden von 3 Wetterstationen die Messdaten gespeichert, in Abbildung 1 sind die Standorte

Tabelle 2. Anzahl an Zeilen der Kombinationen von Ein- und Aussteigern vor dem Auffüllen von Zeitstempel, hier nur Kombinationen von 0 bis 5 aufgezeigt

		Einsteiger					
		0	1	2	3	4	5
Aussteiger	0	**0**	3430	1063	462	308	246
	1	3993	2271	1053	562	395	270
	2	1469	1330	980	599	419	315
	3	796	852	700	559	447	331
	4	458	610	534	491	403	354
	5	343	480	433	414	334	318

der 3 Stationen zu sehen. Jede Station nimmt Temperatur, relative Luftfeuchtigkeit, Regendauer und weitere Attribute auf. Diese Messungen werden über eine Stunde durchgeführt und dann stündlich gespeichert.

Der stündliche Wetterdatensatz hat pro Jahr über 180.000 Datenpunkte und 7 Spalten.

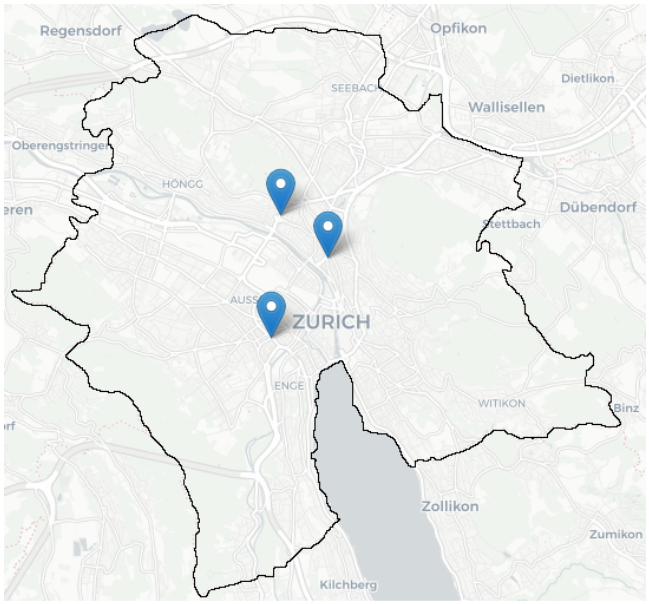


Abbildung 1. Standorte der Wetterstationen in Zürich

Beim Wetterdatensatz gibt es keine Probleme mit der Datenqualität und somit wurden nur kleinere Schritte unternommen, was zum Beispiel beinhaltet, den Zeitstempel des Datensatzes auf eine gleiche Form zu bringen wie die des Fahrzeiten Datensatzes um spätere Analysen leichter zu gestalten.

3 Analyse der Pünktlichkeit

Es werden folgende Hypothesen in diesem Abschnitt analysiert:

H1 : Im Jahr 2022 gibt es einen positiven linearen Zusammenhang zwischen ein- und aussteigenden Fahrgästen und aufgebauten Verspätungen an der Hardbrücke Haltestelle bei den Linien 33, 72, 83 und 8.

H2 : In der Mehrheit der Jahre von 2016 bis 2022 verzeichneten Busse in Stunden mit mehr als 30 Minuten Regen überdurchschnittlich hohe aufgebaute Verspätungen.

3.1 Grundbegriffe

In beiden Hypothesen handelt es sich um aufgebaute Verspätungen, dafür wird folgende Formel verwendet:

$$\text{Verspätung}_{\text{aufgebaut}} = \text{Verspätung}_{\text{abfahrt}} - \text{Verspätung}_{\text{ankunft}} \tag{1}$$

In Tabelle 3 sind Beispiele aufgezeigt und welche Werte daraus als aufgebaute Verspätung angenommen werden. Hierbei heißt ein positiver Wert, dass die Verspätung sich um so viel Minuten erhöht hat, also das Verkehrsmittel mehr verspätet ist. Bei einem negativen Wert ist das analog, also dass es weniger verspätet ist. Dabei beschreibt also die aufgebaute Verspätung die Verspätung, die der Zug an einer bestimmten Haltestelle hinzubekommen oder abgebaut hat.

Tabelle 3. Beispiel zur Berechnung der aufgebauten Verspätungen

		Ankunft		
		1 Min früh	Pünktlich	1 Min spät
Abfahrt	1 Min früh	± 0 Min	- 1 Min	- 2 Min
	Pünktlich	+ 1 Min	± 0 Min	- 1 Min
	1 Min spät	+ 2 Min	+ 1 Min	± 0 Min

3.2 H1 - Verspätungen wegen Fahrgästen

Auch Fahrgäste können Schuld an Verspätungen der Busse und Bahnen haben, wenn sehr viele Passagiere ein- oder aussteigen wollen, so muss das Verkehrsmittel auch dementsprechend lange warten an der Haltestelle. Dies wurde durch Nagy und Csiszár untersucht und herausgefunden [8]. Wenn also ein Verkehrsmittel länger an einer Haltestelle warten muss, wegen Passagieren die ein- oder aussteigen, als das was geplant ist, so ist es verspätet.

Somit wird hier analysiert, ob Fahrgastfrequenzen die Verspätungen an der Hardbrücke Haltestelle erhöhen, spezifisch, ob es einen positiven linearen Zusammenhang zwischen den beiden Metriken gibt.

Die beiden Datensätze schneiden sich in den Jahren 2020, 2021 und 2022. Da jedoch 2020 und 2021 Covid-19 prominent war und sich die Hypothesen deswegen verzerren könnten, wird als zu betrachtendes Jahr 2022 gewählt.

3.2.1 Vorbereitung der Daten.

Die Fahrzeiten werden auf das Jahr 2022 und auf die Datenpunkte, welche die richtigen Linien beinhaltet eingeschränkt. Die Passagierfrequenzen werden auch auf 2022 und auf den Bereich des Sensors eingeschränkt, der die Anzahl an Ein- und Aussteigern zählt. Die zwei Haltepunkte der Fahrzeiten werden mit den beiden Bereichen so zugeordnet, dass die Haltepunkt-ID's dem richtigen Bereich des Sensors gehört.

Die Fahrzeiten werden nun auf 5 Minuten aggregiert und mithilfe der Formel 1 die Verspätungen ausgerechnet. Die Ein- und Aussteiger der Passagierfrequenzen werden außerdem zu Passagieren aufsummiert, denn für die Hypothese ist nur wichtig, wie viele es insgesamt pro Zeitstempel sind.

Damit werden jeder Verspätung jetzt eine Passagieranzahl zugeordnet.

3.2.2 Spearman's Rangkorrelation.

Für die Hypothese wird eine Korrelationsanalyse benötigt. Sowohl die Verspätungen der Fahrzeiten, als auch die Ein- und Aussteiger der Passagierfrequenzen sind nicht normalverteilt, dafür wurde der Shapiro-Wilk Test verwendet [1]. In Tabelle 4 sind die Ergebnisse des Tests zu sehen, wobei der p-Wert für beide Attribute $\leq 0{,}05$ ist. Das bestätigt, dass beide Attribute nicht normalverteilt sind.

Attribut	p-Wert	Normalverteilt?
Verspätungen	$8{,}86 \cdot 10^{-168}$	✗
Passagiere	$3{,}52 \cdot 10^{-144}$	✗

Tabelle 4. Ergebnisse des Shapiro-Wilk Tests

Als gängige Korrelationsanalyse wird die Pearson Korrelation verwendet, doch im Paper von Myers et al. wird deutlich, warum diese Korrelationsanalyse bei nicht normalverteilten Daten nicht geeignet ist, sondern die Spearman's Rangkorrelation eine bessere Alternative ist [7]. Also wird für die Hypothese die Spearman's Rangkorrelation verwendet.

Bei der Rangkorrelation werden zuerst Ränge an die Datenpunkte für die beiden zu betrachtenden Variablen vergeben. Dann wird die Korrelation durch folgende Formel ausgerechnet:

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)}$$

wobei d_i die Differenz zwischen den Rängen der beiden Variablen und n die Anzahl an zu betrachtenden Datenpunkten ist [11]. Damit bekommt man dann einen Wert $-1 \leq r_s \leq 1$ heraus. Je näher das r_s an -1 und 1 jeweils ist, desto höher ist die Korrelation, wobei bei einem positiven Wert die Korrelation positiv ist, negativ bei negativem r_s. Es besteht eine positive Korrelation, wenn bei einem Anstieg des Wertes einer Variablen auch der Wert der anderen Variable ansteigt, analog bei einer negativen Korrelation.

3.2.3 Annahmekriterium. Laut Kuckartz et al. wird ein Wert von $0,5 \leq r_s < 0,7$ als eine hohe und $0,7 \leq r_s \leq 1$ als eine sehr hohe positive Korrelation bezeichnet [4]. Deshalb wird die Hypothese angenommen, wenn der Wert $\leq 0,5$ ist. Dies würde dann bedeuten, dass es einen hohen positiven linearen Zusammenhang zwischen der Anzahl an Fahrgästen und der Verspätung von Bahnen und Bussen der genannten Linien an der Hardbrücke Haltestelle gibt.

3.2.4 Ergebnisse der Rangkorrelation. Nach Durchführung des Tests hat sich ein Wert für r_s von rund 0,1529 ergeben, was nach Annahmekrierium heißt, dass die Hypothese abgelehnt wird. Im Buch von Kuckartz et al. wird dieser Wert als geringe positive Korrelation bezeichnet, was nicht hoch genug ist um diese Hypothese anzunehmen [4]. In Abbildung 2 wird in einem Scatterplot die Verspätung im Zusammenhang mit den Fahrgastfrequenzen gezeigt, wobei gesehen werden kann, dass es, wie oben beschrieben, nur eine sehr leichte positive Korrelation gibt.

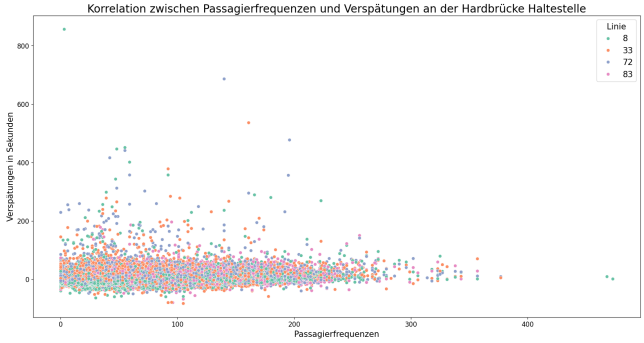


Abbildung 2. Scatterplot, welches die Korrelation zwischen Ein-und Aussteigern und Verspätungen zeigt, unterteilt in Linien

Außerdem kann in Abbildung 3 gesehen werden, dass aufsummiert alle der Linien im Jahr 2022 insgesamt negative aufgebaute Verspätungen hatten, wobei die Tram Linie 8 mit 8453 Minuten die höchsten abgebauten Verspätungen hat. Also bedeutet das, dass alle vier Linien auf das ganze Jahr gesehen mehr Verspätungen an der Hardbrücke Haltestelle abgebaut als aufgebaut haben.

3.3 H2 - Verspätungen wegen Wetterbedingungen

Nagy und Csiszár beschreiben zusätzlich in ihrem Paper, dass das Wetter auch Einfluss auf Verspätungen des öffentlichen Verkehrs haben kann [8]. Zudem führt Niederschlag zu mehr Verkehrsstaus, insbesondere bei Stoßzeiten, wie Koetse et al. berichten [3]. Durch Verkehrsstaus sind zwar Bahnen nicht viel beeinträchtigt, doch Busse fahren auf denselben Straßen, wie PKW's, was bei Staus auch Busse aufhält. Außerdem erhöht sich der Bremsweg eines Busses bei Regen deutlich. Ein Bus in Zürich hat ein Leergewicht von mindestens 7 Tonnen, was mit Fahrgästen dann noch viel mehr werden

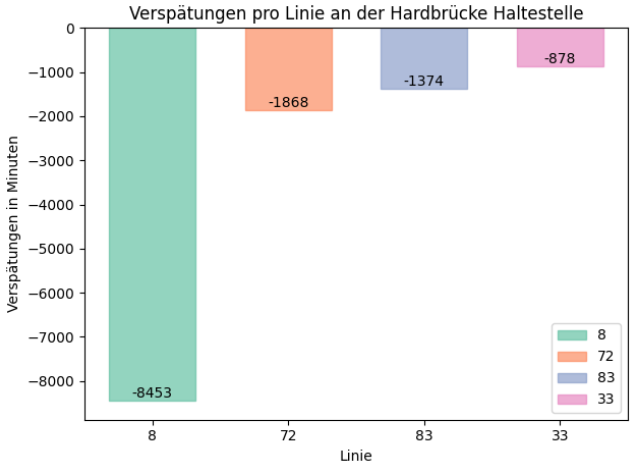


Abbildung 3. Aufsummierte Verspätungen pro Linie an der Hardbrücke Haltestelle

kann [12]. Durch das Gewicht alleine ist der Bremsweg eines Busses höher als die eines PKW's, doch bei Regen ist der Bremsweg noch höher [2]. Weshalb Busfahrer bei Regen mehr Acht geben müssen und daher auch langsamer fahren müssen als was sie sonst können.

Daher wird hier untersucht, ob der Regen für Busse heißt, dass sie Verspätungen aufbauen. Hierbei wird die Regendauer, wie es in dem Stündlichen Wetterdatensatz gespeichert wird, betrachtet und es wird angeschaut, ob wenn es bei mehr als der Hälfte der Stunde regnet, Busse überdurchschnittlich hohe aufgebaute Verspätungen verzeichnen. Auch hier werden die Verspätungen wie in Formel 1 beschrieben berechnet. In dieser Hypothese wird der Zeitraum von 2016 bis 2022 betrachtet, welche die Schnittmenge der aufgezeichneten Jahre der beiden Datensätze, Fahrzeiten und Wetter, ist.

3.3.1 Vorbereitung der Daten. Wie in der ersten Hypothese werden auch hier die Verspätungen ausgerechnet, doch diesmal werden die Datenpunkte der Fahrzeiten auf eine Stunde aggregiert, weil die Wetterdaten auch stündlich sind. Außerdem werden die Fahrzeiten auf Datenpunkte, welche Buslinien als Linie haben eingeschränkt.

Um die beiden Datensätze noch zu verbinden, muss die Frage gestellt werden, wie damit umgegangen wird, dass der Stündliche Wetterdatensatz in drei Wetterstationen unterteilt ist und jede Wetterstation seine eigene Messung hat. Dafür wird diese Hypothese in zwei Versionen aufgeteilt, $H2_a$ nimmt dabei die maximale Regendauer der Wetterstationen und teilt diese für jede Stunde den Verspätungen der einzelnen Haltepunkte zu. In $H2_b$ wird die nächste Wetterstation zu dem Haltepunkt ermittelt und dessen Messung für diesen Haltepunkt und seine Verspätung genommen.

Die stündlichen Verspätungen bekommen einen Messwert, je nach Version zugeordnet, dieser Messwert sagt aus, wie viele Minuten es in dieser Stunde geregnet hat. Dann werden

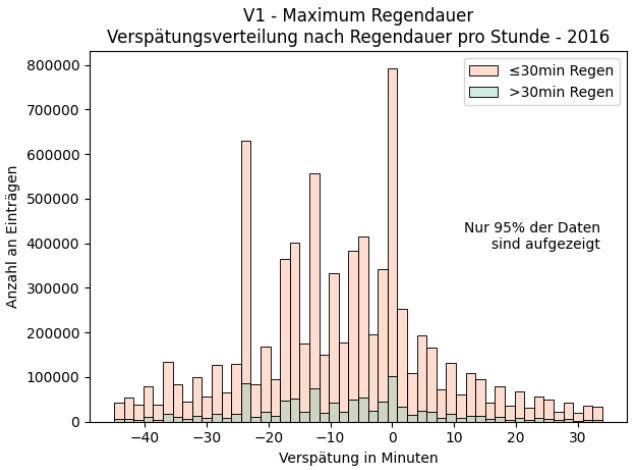
Tabelle 5. Ergebnisse des Mann-Whitney U Tests

	$H2_a$ - Maximale Regendauer		$H2_b$ - Nächste Wetterstation	
Jahr	p-Wert	< 0,05	p-Wert	< 0,05
2016	0,3679	✗	$0{,}8 \cdot 10^{-4}$	✓
2017	0,9999	✗	1,0	✗
2018	0,9933	✗	1,0	✗
2019	1,0	✗	1,0	✗
2020	$0{,}2 \cdot 10^{-59}$	✓	0,9601	✗
2021	1,0	✗	1,0	✗
2022	$0{,}25 \cdot 10^{-4}$	✓	0,0942	✗

zuletzt die Datenpunkte in zwei Kategorien unterteilt, mehr als 30 Minuten Regen und weniger oder gleich 30 Minuten Regen. Diese werden im weiteren zueinander verglichen.

3.3.2 Mann-Whitney U Test. Der Mann-Whitney U Test, welcher auch als Wilcoxon bekannt ist, ist das Gegenstück zum parametrischen T-Test, welcher wie Spearman's Rangkorrelation keine Normalverteilung der Daten in den untersuchten Variablen voraussetzt [5]. Da die Fahrzeiten Daten, wie in Abbildung 4 am Beispiel $H2_a$ und 2016 gesehen nicht Normalverteilt sind, wird also dieser Test gewählt für diese Hypothese.

Abbildung 4. Verteilung der Verspätungsdaten für H2



3.3.3 Annahmekriterium. Beim Mann-Whitney U Test handelt es sich um einen Signifikanztest, welcher einen p-Wert berechnet. Dabei ist der p-Wert größer, wenn dieser Zusammenhang eher ein Zufall ist. Als Annahmekriterium wird also etwas kleines gewählt und in der Stochastik wird hier oft ein Signifikanzniveau $\alpha = 0,05$ gewählt, was auch hier als Annahmekriterium gilt. Hierbei wird aber für jedes Jahr einzeln der Test durchgeführt und wenn $p \leq 0,05 = \alpha$ für mindestens 4 der 7 Jahre gilt, so wird die Hypothese angenommen.

3.3.4 Ergebnisse der Analyse. In Tabelle 5 sind die Ergebnisse des Mann-Whitney U Tests pro Jahr aufgezeigt. In $H2_a$ ist es in den Jahren 2020 und 2022 der Fall, dass bei über 30 Minuten Regen in der Stunde es heißt, dass die Busse mehr Verspätungen haben und in $H2_b$ ist es im Jahr 2016 der Fall. In beiden Versionen aber wird die Hypothese abgelehnt, weil das Annahmekriterium nicht bei mindestens 4 der 7 Jahre zutrifft.

3.4 Limitationen der Analysen

In der Berechnung der Verspätungen gibt es Fälle in denen ein Verkehrsmittel Verspätungen abbaut indem es zu früh abfährt (siehe erste Zeile in Tabelle 3). Dies ist laut Formel gut, doch in Realität ist das nicht das, was stattfinden sollte. Ein Fahrgast, welcher pünktlich an der Haltestelle ist und seinen Zug verpasst hat, weil dieser zu früh abgefahren ist, ist auch nicht zufrieden, wie wenn das Fahrzeug zu spät ist. Das wurde hier in der Hypothese nicht beachtet und sollte in Betracht gezogen werden bei weiteren Analysen.

Außerdem wird in der Analyse nicht zwischen den Kategorien der abgebauten Verspätungen differenziert:

- Das Verkehrsmittel ist 1 Minute zu spät angekommen und pünktlich abgefahren
- Es ist 2 Minuten zu spät angekommen und 1 Minute zu spät abgefahren

Analog auch mit aufgebaute Verspätungen.

3.5 Diskussion

In Nagy und Csiszár's Paper schreiben sie, dass Passagiere ein Grund für Verspätungen sind. Doch in diesem Paper werden Gründe durch das Personal zugewiesen, was heißt, dass es mehrere Gründe geben kann, welche das Personal nicht beachtet haben. Also es kann dazu führen, dass eine Mischung von Gründen als ein Grund klassifiziert wird, weil dieser prominent war.

Ein weiterer Grund, warum die Ergebnisse der beiden Authoren von den Ergebnissen dieses Papers abschweifen

ist, dass hier zwei verschiedene Städte aus verschiedenen Ländern betrachtet werden. Die Analyse von Nagy und Csiszár bezieht sich auf die Stadt Győr in Ungarn, wobei dieses Paper sich Zürich, Schweiz anschaut. Die Infrastruktur und der öffentliche Verkehr von Land zu Land unterscheidet sich drastisch. Das kann auch zu anderen Ergebnissen wie in diesem Fall führen.

Zudem werden in diesem Paper die beiden Gründe von Verspätungen nur einzeln angesehen, eine Analyse, welche mehrere Gründe zusammen betrachtet ist auch hilfreich für die Verbesserung der Service Qualität des öffentlichen Verkehrs.

4 Fazit

In diesem Paper wurden drei Datensätze des VBZ bereinigt, exploriert und analysiert und es wurde herausgefunden, dass es keinen positiven linearen Zusammenhang zwischen Ein- und Aussteigern und Verspätungen gibt im Jahr 2022 für die Haltestelle Hardbrücke, bei den Buslinien 33, 72, 83 und der Tramlinie 8.

Außerdem gibt es nur in der Minderheit der Jahre einen signifikanten Unterschied, bei den Verspätungen einer Stunde, unter den beiden Kategorien, dass es mehr als 30 Minuten in dieser Stunde geregnet hat und weniger oder gleich 30 Minuten geregnet hat.

Die Datensätze der Stadt Zürich bieten eine Vielzahl an Analysen, die durchgeführt werden können. Hier wurden nur drei der vielen Datensätze verwendet und auch mit diesen drei Datensätzen können noch viel mehr Schlüsse gezogen werden. So kann zum Beispiel die Kombination der beiden Hypothesen analysiert werden oder es kann auf eine der Hypothesen mehr eingegangen werden.

Im Ganzen ist die Analyse von Verspätungsgründen ein wichtiger Teil des öffentlichen Verkehrs und könnte dazu führen, dass Bahnunternehmen pünktlicher sind, denn das ist ein Teil der Kundenzufriedenheit im öffentlichen Verkehr.

Literatur

[1] Elizabeth González-Estrada and Waldenia Cosmes. 2019. Shapiro–Wilk test for skew normal distributions based on data transformations. *Journal of Statistical Computation and Simulation* 89, 17 (2019), 3258–3272.

[2] Poul Greibe. 2007. Braking distance, friction and behaviour. *Trafitec, Scion-DTU* (2007).

[3] Mark J Koetse and Piet Rietveld. 2009. The impact of climate change and weather on transport: An overview of empirical findings. *Transportation Research Part D: Transport and Environment* 14, 3 (2009), 205–221.

[4] Udo Kuckartz, Stefan Rädiker, Thomas Ebert, and Julia Schehl. 2013. *Statistik: eine verständliche Einführung*. Springer-Verlag.

[5] Patrick E McKnight and Julius Najab. 2010. Mann-Whitney U Test. *The Corsini encyclopedia of psychology* (2010), 1–1.

[6] Arnoud Mouwen. 2015. Drivers of customer satisfaction with public transport services. *Transportation Research Part A: Policy and Practice* 78 (2015), 1–20.

[7] Leann Myers and Maria J Sirois. 2004. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences* 12 (2004).

[8] Enikő Nagy and Csaba Csiszár. 2015. Analysis of delay causes in railway passenger transportation. *Periodica Polytechnica: Transportation Engineering* 43, 2 (2015), 73–80.

[9] Piet Rietveld, Frank Reinier Bruinsma, and Daniel J Van Vuuren. 2001. Coping with unreliability in public transport chains: A case study for Netherlands. *Transportation Research Part A: Policy and Practice* 35, 6 (2001), 539–559.

[10] Dea Van Lierop, Madhav G Badami, and Ahmed M El-Geneidy. 2018. What influences satisfaction and loyalty in public transport? A review of the literature. *Transport Reviews* 38, 1 (2018), 52–72.

[11] Chengwei Xiao, Jiaqi Ye, Rui Máximo Esteves, and Chunming Rong. 2016. Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience* 28, 14 (2016), 3866–3878.

[12] VBZ ZüriLine. 2024. *Fahrzeuge*. https://www.stadt-zuerich.ch/vbz/de/index/die_vbz/fahrzeuge.html Zuletzt zugegriffen: April 2024.