

Big Train Data

Hannes Spitz

Abstract—Diese Arbeit analysiert die Zuverlässigkeit des deutschen Schienennverkehrs anhand eines Datensatzes von über 60 Millionen Einträgen im Zeitraum von Juli 2024 bis Januar 2026. Durch die Verknüpfung von Verspätungsdaten der Deutschen Bahn mit hochauflösenden Wetterdaten des Deutschen Wetterdienstes wird untersucht, inwiefern Geografie, Verkehrsaufkommen, Zugtyp und Temperatur Ausfälle und Verspätungen beeinflussen. Die Ergebnisse zeigen, dass insbesondere der Fernverkehr (ICE) verspätet ist. Während Schienenfahrzeuge eine signifikante Anfälligkeit für hohe Temperaturen aufweisen, hat der Busverkehr vor allem mit Frost zu kämpfen.

Index Terms—Big Data, Deutsche Bahn, Train Delays, Spark.

I. EINFÜHRUNG

VERSPÄTUNGEN der Bahn können unterschiedliche Gründe haben. Einige sind uns als Bahnfahrende ersichtlich, andere weniger. Wichtiger jedoch ist für uns ein Gefühl zu bekommen, wann welche Bahn verspätet ist, damit wir unsere Reisen besser planen können. Wer kennt das Gefühl nicht, zu der Haltestelle gerannt zu sein, gerade noch rechtzeitig angekommen zu sein, nur um dann zu merken dass die erhoffte Verbindung ausfällt oder 20 Minuten Verspätung haben soll. Natürlich bleibt es niemals bei den angegebenen 20 Minuten. Auch die Kommunikation, wenn ein Zug ausfällt kommt oft zu spät und erst wenn man bereits am Bahnsteig wartet.

Daher ist es unerlässlich, sich selbst ein Bild zu machen, welche Faktoren Verspätungen und Ausfälle beeinflussen. Eine ähnliche Auswertung wurde 2019 von David Kriesel beim CCC gehalten [1]. Diese beschränkt sich allerdings auf Daten bis 2019.

In dieser Arbeit betrachte ich die Einflüsse des Standortes eines Bahnhofs, der Wochentage, Tageszeiten und Zugtyp ebenfalls betrachtet, jedoch auf aktuellen Verspätungsdaten. Hinzukommt die Analyse des Verkehrsaufkommens und des Wetters als Einflussfaktoren.

II. METHODEN

A. Daten

Um die Analysen durchzuführen habe ich den Zeitraum Juli 2024 bis Januar 2026 gewählt, da meine Datenquelle für die Datenverpätungen "Deutsche Bahn Data" [2] diesen Zeitraum abdeckt. Die Daten wurden generiert durch Abfragen der Deutschen Bahn APIs StaDa [3] und DB Timetables [4]. Genutzte Spalten des Datensatzes sind station_name, delay_in_min, time, is_canceled, und train_type.

Um die Standorte der Haltestellen herauszufinden, habe ich einen weiteren Datensatz [5] einbezogen, der u.A. Informationen des Breiten- und Höhengrades, sowie die Haltestellennamen enthält.

Wetterdaten habe ich von dem Deutschen Wetterdienst dem HYRAS-DE-TAS [6] Datensatz entnommen. Dieser enthält Temperaturdaten seit 1951 im Tagesmittel in Deutschland. Die Werte liegen auf einem 1x1 km Raster vor, die aus verfügbaren Wetterstationsdaten interpoliert wurden.

B. Verarbeitung

Zur Bearbeitung der Daten wurde vor allem Spark [7] verwendet. Ein Problem dabei war dabei, dass timestamps in Nanosekunden angegeben waren aber Spark diese nicht verarbeiten kann. Daher habe ich diese mithilfe von Pandas in Mikrosekunden umgewandelt.

Im Anschluss habe ich den Haltestellendatensatz mit den Verspätungen vereinigt. Da die Haltestellennamen leichte Abweichungen enthalten, habe ich vorher diese normalisiert auf das erste Wort, Kleinbuchstaben und ohne Umlaute. In der Folge enthielten die meisten normalisierte Namen nur noch den Stadtnamen, was jedoch für die Zuordnung der Koordinaten kein großes Problem darstellt, da alle Koordinaten einer Stadt nahe beieinander liegen sollten. Wenn in dem Haltestellendatensatz mehrere Koordinaten existieren, habe ich jeweils den Median verwendet um Outlier weniger Gewicht zu geben.

Die Wetterdaten habe ich mit für jeden Zeitpunkt und jede Koordinate des Verspätungsdatensatz abgefragt. Da dieser Lookup in konstanter Laufzeit möglich ist, und somit nicht von der Größe des Wetterdatensatzes abhängt, ist dies deutlich effizienter als ein Joinen. Um den Lookup effizient zu parallelisieren habe ich Sparks 'mapInPandas' Methode genutzt.

Der resultierende Datensatz enthält 60.701.100 Zeilen und die Spalten: Stationen Name (str), Normalisierter Name (str), Verspätung in Minuten (int), Ausfall (bool), Zugtyp (str), Zeit (timestamp), Höhengrad (float), Breitengrad (float), Temperatur (float).

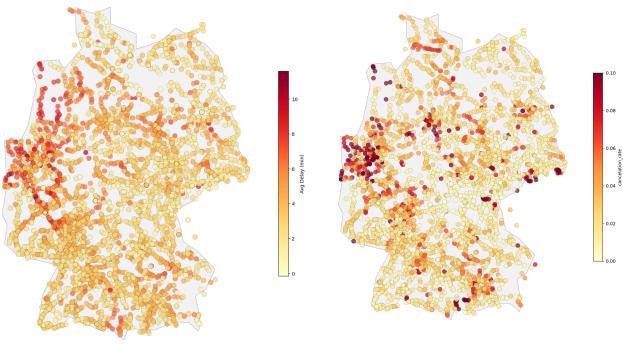
Für die Analysefragen habe ich die Daten jeweils in Spark aggregiert und im anschluss Grafiken mit Pandas, Geopandas und Matplotlib erstellt.

III. ERGEBNISSE

Die Analyse der geografischen Daten offenbart deutliche regionale Unterschiede in der Zuverlässigkeit des Schienennverkehrs.

A. Regionaler Vergleich: Verspätung vs. Ausfall

Die geografische Analyse (Abb. 1) zeigt eine deutliche Häufung von Verspätungen im Westen Deutschlands. Besonders in dicht vernetzten Regionen wie Nordrhein-Westfalen treten Kettenreaktionen auf (Abb. 1a). Ausfälle sind punktueller, konzentrieren sich jedoch ebenfalls auf das Ruhrgebiet (Abb. 1b).



(a) Durchschnittliche Verspätung

(b) Durchschnittliche Ausfälle

Fig. 1. Geografische Verteilung der Bahndaten. Die Häufung im Westen deutet auf hohe Netzauslastung hin.

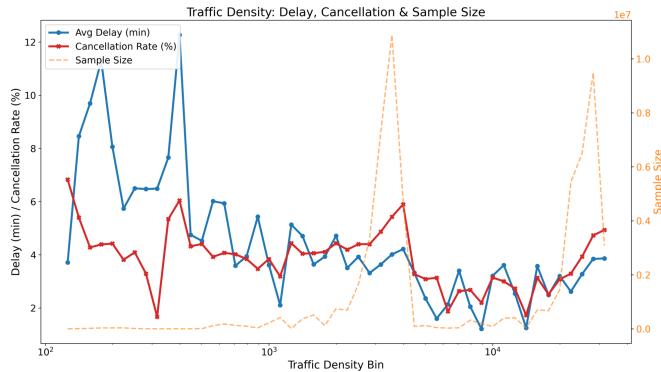


Fig. 2. Zusammenhang zwischen Verkehrsdichte eines Zeitpunkts (Zughalte in Deutschland pro Stunde) und Zuverlässigkeit des Netzes zu diesem Zeitpunkt. Aussagen können nur über Zeitpunkte getroffen werden, in denen ausreichend Daten zur Verfügung stehen (gelb gestrichelte Linie)

B. Verkehrsaufkommen und zeitliche Einflüsse

Die Untersuchung des Verkehrsaufkommens (Abb. 2) zeigt eine deutliche Korrelation zwischen der Netzauslastung und der Zuverlässigkeit. In den zwei Datenclustern mit hoher Dichte – vermutlich repräsentativ für den Tag- und Nachtbetrieb – ist ein Anstieg der Verspätungen und Ausfallraten bei zunehmender Anzahl von Zügen pro Stunde erkennbar. Dies unterstreicht, dass das System bei hoher Auslastung empfindlicher auf Störungen reagiert. Zeitlich gesehen führt die geringere Last am Wochenende zu einer Stabilisierung (Abb. ??). Im Tagesverlauf (Abb. 3) akkumulieren sich Verspätungen bis in die Nachtstunden, was auf eine unzureichende Pufferkapazität des Netzes hindeutet.

C. Einfluss von Zugtyp und ICE-Analyse

Die statistische Auswertung nach Zuggattungen (Abb. 4) bestätigt den Verdacht, dass der Fernverkehr (ICE) überproportional für Unpünktlichkeit verantwortlich ist. Während die S-Bahn mit über $2,5 \cdot 10^7$ Datenpunkten die höchste Frequenz aufweist, verzeichnet der ICE trotz deutlich geringerer Stichprobengröße die höchsten Werte: Die durchschnittliche Verspätung liegt bei über 11 Minuten, die Aus-

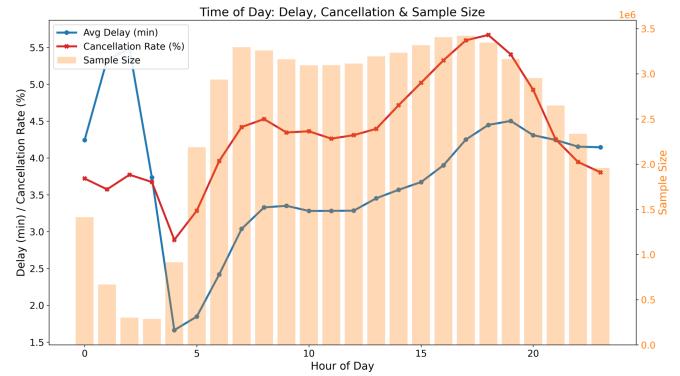


Fig. 3. Kumulation der Verspätungen über den Tagesverlauf.

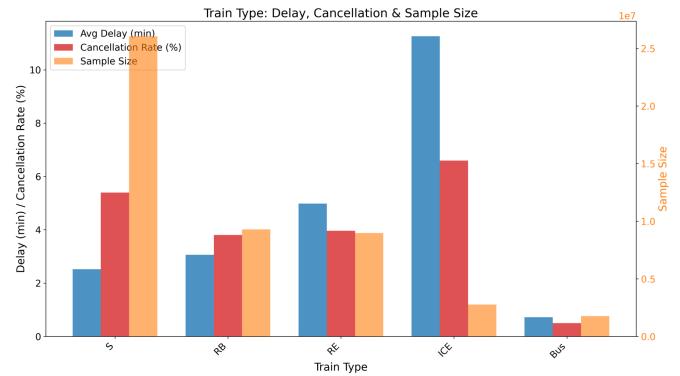
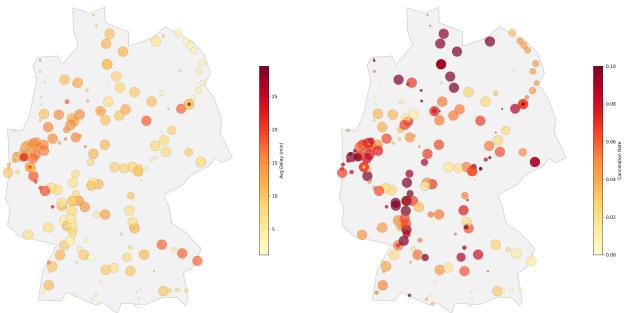


Fig. 4. Vergleich der Zuverlässigkeit nach Zugtyp.



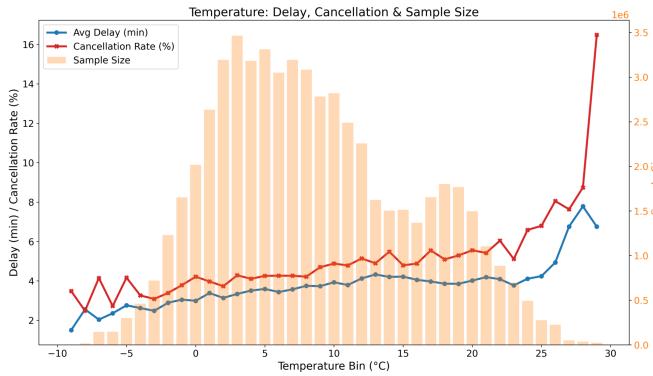
(a) ICE Verspätungen

(b) ICE Ausfälle

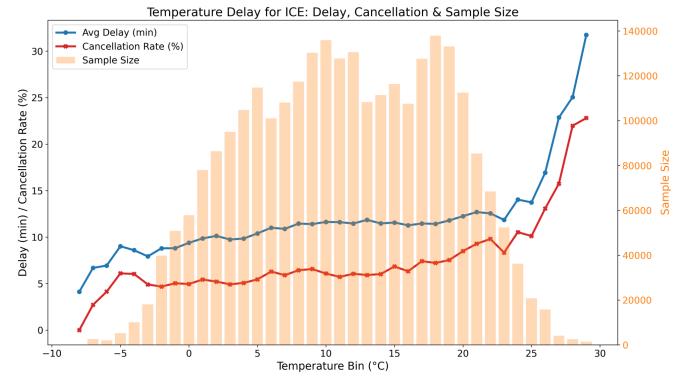
Fig. 5. Analyse des ICE-Verkehrs. Die Kreisgröße korreliert mit dem Zugaufkommen an den jeweiligen Knotenpunkten.

fallrate bei über 6,5 %. Im Vergleich dazu sind Regionalzüge (RB, RE) und insbesondere Busse deutlich zuverlässiger.

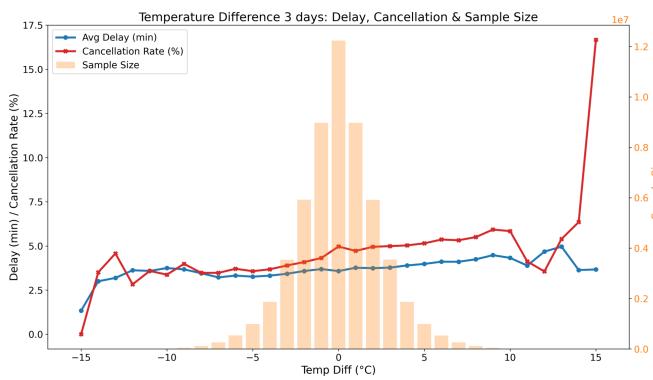
Eine isolierte Betrachtung des ICE-Netzes (Abb. 5) verdeutlicht die Belastung der Hauptrouten. Auffällig ist die Verbindung zwischen Karlsruhe und Frankfurt am Main: Diese weist eine überdurchschnittliche Ausfallrate auf (Abb. 5b), rangiert jedoch bei der reinen Verspätungszeit im Mittelfeld. Dies deutet darauf hin, dass Züge auf dieser hochbelasteten Strecke bei Problemen eher komplett gestrichen werden, um den restlichen Taktverkehr nicht zu gefährden.



(a) Einfluss der absoluten Temperatur



(a) Temperatur-Einfluss ICE



(b) Temperaturdifferenz (3 Tage)

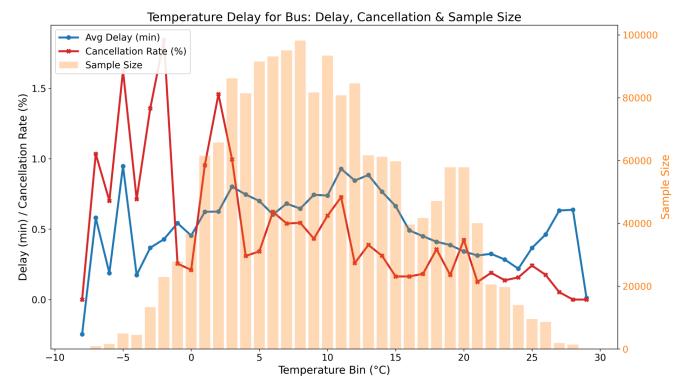
Fig. 6. Wetterbedingte Einflüsse: Während die absolute Temperatur mit der Unzuverlässigkeit korreliert, haben starke Temperaturschwankungen keinen signifikanten Zusatzeffekt.

D. Temperatur

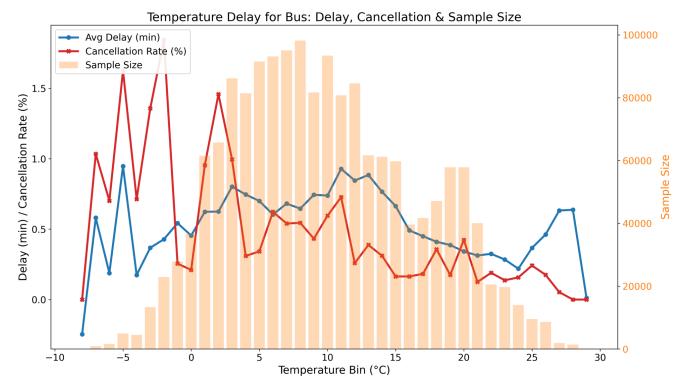
Die statistische Auswertung der Wetterdaten verdeutlicht, dass die absolute Umgebungstemperatur einen signifikanten Einfluss auf die Betriebsstabilität hat (Abb. 6), während Temperaturänderungen keine messbaren Auswirkungen im Bereich mit ausreichender Menge an Daten zeigen. Im Bereich zwischen -5°C und 25°C steigen Verspätungsminuten und Ausfallraten nahezu linear mit der Wärme an, wobei die Schienenverkehrsmittel, insbesondere der ICE, bei Extremtemperaturen über 25°C eine überproportionale Zunahme der Instabilität verzeichnen (Abb. 7a). Im Gegensatz dazu weist der Busverkehr bei Temperaturen unter dem Gefrierpunkt eine Diskrepanz zwischen einer hohen Ausfallquote und vergleichsweise geringen Verspätungen auf, zeigt jedoch mit zunehmender Erwärmung eine stetige Verbesserung der Zuverlässigkeit (Abb. 7b). Damit weisen die verschiedenen Verkehrsmittel im Datensatz unterschiedliche Sensitivitäten gegenüber thermischen Extremwerten auf, wobei die Schiene bei Hitze und die Straße bei Frost die jeweils höchsten Ausfallraten erreichen.

IV. KONKLUSION

Die datengestützte Analyse erlaubt eine klare Identifikation von Zeitfenstern und Bedingungen für eine möglichst zuverlässige Reiseplanung. Wer Pünktlichkeit anstrebt, sollte



(a) Temperatur-Einfluss ICE



(b) Temperatur-Einfluss Bus

Fig. 7. Vergleich der Temperatureinflüsse: Während Schienenfahrzeuge bei Hitze instabil werden, zeigen Busse eine kritische Anfälligkeit für Frost, stabilisieren sich jedoch bei Wärme.

Reisen auf das Wochenende legen, da die geringere Netzauslastung zu diesem Zeitpunkt weniger Verspätungen und Ausfälle provoziert. Innerhalb des Tagesverlaufs ist die Abfahrt in den frühen Morgenstunden (vor 06:00 Uhr) zu bevorzugen, da hier die über den Vortag kumulierten Verspätungsketten weitgehend abgebaut sind und das System noch nicht die im Tagesverlauf beobachtete Fehlerspirale erreicht hat.

In Bezug auf das Wetter und die Zuggattung zeigt die Analyse zwei kritische Szenarien: An Hitzetagen mit Temperaturen über 25°C steigt das Risiko massiver Verspätungen im Fernverkehr (ICE) überproportional an. In solchen Fällen oder bei engen Zeitplänen stellt der Regionalverkehr (RE, RB) trotz längerer Fahrtzeiten die stabilere Alternative dar, da dieser eine deutlich geringere Durchschnittsverspätung und höhere Hitzebeständigkeit aufweist als das ICE-Netz. Bei Frost ist vom Busverkehr abzuraten, da hier die Ausfallwahrscheinlichkeit ihr Maximum erreicht, während die Schiene in diesem Temperaturbereich stabiler operiert als bei Hitze. Zusammenfassend lässt sich festhalten: Die zuverlässigste Verbindung ist eine frühe Regionalfahrt außerhalb von Nordrhein-Westfalen an einem kühlen Wochenende.

APPENDIX

In diesem Abschnitt werden die restlichen typenspezifischen Temperaturkurven aufgeführt (S-Bahn, Regional-Express und Regionalbahn).

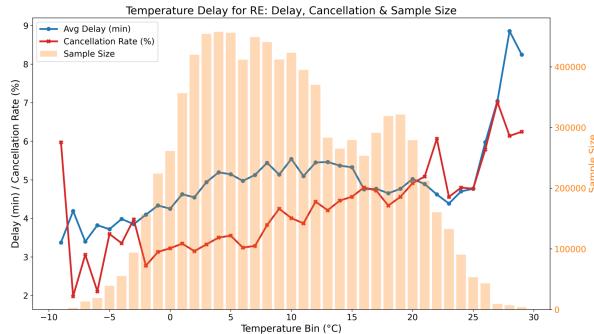


Fig. 8. Temperatureinfluss auf den Regional-Express (RE).

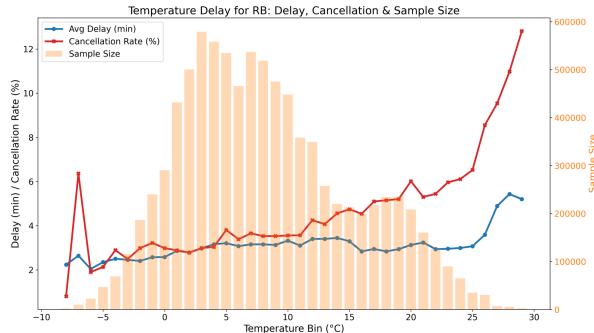


Fig. 9. Temperatureinfluss auf die Regionalbahn (RB).

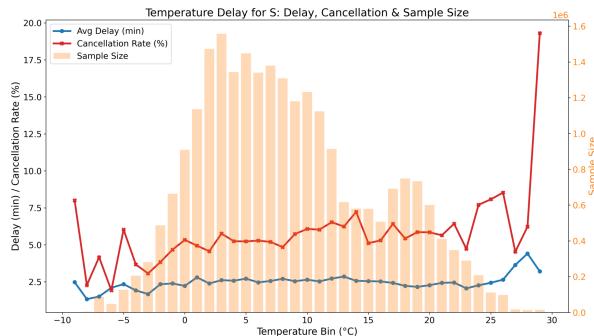


Fig. 10. Temperatureinfluss auf die S-Bahn (S).

REFERENCES

- [1] D. Kriesel, “BahnMining - Pünktlichkeit ist eine Zier,” 36th Chaos Communication Congress (36C3), 2019. [Online]. Available: https://media.ccc.de/v/36c3-10652-bahnmining_-_punktlichkeit_ist_eine_zier
- [2] P. Brömmel and Contributors, “Deutsche Bahn Data,” 2025. [Online]. Available: <https://github.com/piebro/deutsche-bahn-data>
- [3] Deutsche Bahn AG, “StaDa - Station Data.” [Online]. Available: <https://developers.deutschebahn.com/db-api-marketplace/apis/product/stada>
- [4] Deutsche Bahn AG, “Timetables.” [Online]. Available: <https://developers.deutschebahn.com/db-api-marketplace/apis/product/timetables>
- [5] DELFI e.V. und WVI, “Zentrales Haltestellenverzeichnis Version 2.0.” [Online]. Available: <https://zhv.wvigmbh.de>
- [6] Raster der Mitteltemperatur in °C für Deutschland - HYRAS-DETAS v6-1, Version v6-1. Deutscher Wetterdienst. [Online]. Available: https://opendata.dwd.de/climate_environment/CDC/grids_germany/daily/hyras_de/air_temperature_mean/
- [7] M. Zaharia *et al.*, “Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing,” in *Proc. 9th USENIX Conf. Networked Systems Design and Implementation (NSDI)*, 2012, pp. 2-2.