École Polytechnique Fédérale de Lausanne

Impact of the weather on Swiss railway punctuality

Master's Thesis

by Jordi Cluet i Martinell

Data Science Lab (dlab)
SBB CFF FFS

Thesis Supervisor
Prof. Robert West
dlab (EPFL)

Thesis Advisor
Manoel Horta Ribeiro
dlab (EPFL)

Company Supervisor
Pierre Dominguez
SBB FFS FFS

Lausanne, June 2023

This Master's Thesis was carried out by Jordi Cluet i Martinell, under the supervision of Professor Robert West, from the Data Science Lab (dlab) at École Polytechnique Fédérale de Lausanne (EPFL). It was advised by Manoel Horta Ribeiro, also from dlab, and Pierre Dominguez, from the Swiss Federal Railways (SBB-CFF-FFS).

The project took place within an exchange program from the Facultat d'Informàtica de Barcelona (FIB) of the Universitat Politècnica de Catalunya (UPC) to the Faculté Informatique et Communications (IC Faculty) of the École Polytechnique Fédérale de Lausanne (EPFL), as part of the Master in Data Science.

Acknowledgments

I would like to express my sincere gratitude to the Data Science Lab (dlab), particularly Prof. Robert West and Manoel Horta Ribeiro, for their invaluable support, guidance, and belief in me throughout both my last semester's project and this thesis. Their feedback and supervision have been crucial in shaping the outcome of my research.

I am also deeply grateful to the SBB team, especially Pierre Dominguez, for their guidance, insightful discussions, and for proposing this captivating project that allowed me to explore the fascinating realm of SBB. I take with me very good memories of this period and I hope to cross paths again in the future.

Of course, I would also like to thank all my friends for their unconditional support and encouragement throughout this year. My sincere gratitude to all of you for your patience, and for always being there for me, from Catalunya, Lausanne or any other place.

Una menzione speciale va ai miei amici italiani che mi hanno accompagnato in questo percorso. Sono profondamente grato a Giacomo, Lucia e Barbara per il loro sostegno incrollabile, il feedback costruttivo, e i momenti di gioia condivisi.

No puc acabar aquesta secció sense fer un agraïment a la meva família. És gràcies al seu suport que ha estat possible no només fer aquesta tesi de màster, sinó tota la meva etapa acadèmica. Els estic profundament agraït per la seva confiança, el seu suport incondicional, i per haver-me donat l'oportunitat de viure aquesta experiència a Suïssa.

Lausanne, June 22, 2023 Jordi Cluet i Martinell

Abstract

This thesis investigates the impact of meteorological phenomena on train delays in Switzerland. In particular, it analyses the effect of rain, snow, strong wind, and extreme temperatures on the SBB trains traveling between Bern and Zurich. The study utilizes open datasets spanning five years. The analyses reveal that all meteorological factors contribute to increased delays, with special attention on rain, as it is the most common phenomenon. A clustering method is proposed to identify snowy, sunny, and rainy days and characterize their expected delay.

The study also finds that weather-related disruptions persist throughout the year, with notable delays observed during the fall season, peak hours, and evenings. This negative impact is observed across all train stops, lines, and directions, with direct trains traveling westwards experiencing more significant delays. The analysis also demonstrates that delays occur between stops, while trains catch up at the stations.

Similar patterns are observed in other railway sections of the same line, which offers promising potential for applying the same methods to other regions and lines. Train cancellations are not concluded to be affected by weather conditions. The study also concludes that predicting train delays remains a complex challenge, no matter the predictors, target, or form of the model. However, meteorological features can help improve the performance of more complex prediction models.

This quantification of the impact of weather on delays provides relevant insights to SBB and other rail operators. This will help in the formulation of strategies to manage these delays effectively and ultimately improve the efficiency, reliability, and customer satisfaction of Swiss public transport.

Résumé

Cette thèse étudie l'impact des phénomènes météorologiques sur les retards des trains en Suisse. En particulier, elle analyse l'effet de la pluie, de la neige, des vents forts et des températures extrêmes sur les trains des CFF circulant entre Berne et Zurich. L'étude utilise des données ouvertes couvrant cinq ans. Les analyses révèlent que tous les facteurs météo contribuent à l'augmentation des retards, avec une attention particulière sur la pluie, car c'est le phénomène le plus courant. Une méthode de 'clustering' est proposée pour identifier les différents types de journées et caractériser leur retard attendu.

L'étude révèle également que les perturbations liées aux conditions météorologiques persistent tout au long de l'année, avec des retards notables observés à l'automne, aux heures de pointe et en soirée. Cet impact négatif est observé sur l'ensemble des arrêts, des lignes et des directions, les trains directs circulant vers l'ouest subissant des retards plus importants. L'analyse montre également que les retards se produisent entre les arrêts, tandis que les trains rattrapent leur retard aux gares.

Des tendances similaires sont observées dans d'autres tronçons ferroviaires de la même ligne, ce qui offre un potentiel prometteur pour appliquer les mêmes méthodes à d'autres régions et lignes. Les annulations de trains ne sont pas influencées par les conditions météorologiques. L'étude conclut également que la prévision des retards des trains reste un défi complexe, quels que soient les prédicteurs, le 'target' ou la forme du modèle. Toutefois, les caractéristiques météorologiques peuvent contribuer à améliorer les performances de modèles de prévision plus complexes.

Cette quantification de l'impact des conditions météorologiques sur les retards fournit des informations précieuses aux CFF et aux autres opérateurs ferroviaires. Cela permettra de formuler des stratégies pour gérer efficacement ces retards et, en fin de compte, d'améliorer l'efficacité, la fiabilité et la satisfaction de la clientèle des transports publics suisses.

Resum

Aquesta tesi investiga l'impacte dels fenòmens meteorològics en els retards dels trens a Suïssa. Concretament, analitza l'efecte de la pluja, la neu, el vent i les temperatures extremes en els trens d'SBB que viatgen entre Berna i Zúric. L'estudi utilitza dades obertes dels darrers cinc anys. Les anàlisis revelen que tots els factors meteorològics contribueixen a augmentar els retards, amb especial atenció a la pluja, que és el fenomen més freqüent. Es proposa un mètode de 'clustering' per identificar els dies de neu, assolellats i plujosos i caracteritzar el seu retard esperat.

L'estudi també revela que la meteorologia afecta els retards durant tot l'any, i especialment durant la temporada de tardor, les hores punta i els vespres. Aquest impacte s'observa a totes les estacions, línies i direccions de tren. Els trens directes i els que viatgen cap a l'oest del país són els que pateixen retards més importants. L'anàlisi també demostra que els retards es produeixen entre parades, mentre que els trens recuperen temps a les estacions.

L'anàlisi en altres trams ferroviaris de la mateixa línia mostra resultats similars, la qual cosa ofereix un potencial prometedor per aplicar els mateixos mètodes a altres regions i línies suïsses. No hi ha evidència que les cancel·lacions de trens es vegin afectades per les condicions meteorològiques. L'estudi també conclou que predir els retards dels trens és un repte complex, independentment dels predictors, el 'target' o el tipus del model. Tanmateix, les variables meteorològiques poden ajudar a millorar models de predicció més complexos.

La quantificació de l'impacte de la meteorologia en els retards ofereix conclusions rellevants per a SBB i qualsevol altre operador ferroviari. Això ajudarà en la formulació d'estratègies per gestionar aquests retards de manera eficaç i, en última instància, millorar l'eficiència, fiabilitat i satisfacció del client del transport públic suís.

Contents

Chapter 1

Introduction

Switzerland is renowned for its excellent performing rail network. SBB-CFF-FFS (Swiss Federal Railways, from now on, "SBB") is the primary railway company in the country. SBB prides itself on maintaining a high level of punctuality, as evidenced by their reported passenger service punctuality rate of 92.5% in 2022, which measures the percentage of trains arriving at their destination with less than three minutes' delay.

While these figures are impressive, achieving consistent punctuality across all train lines, regions, and types of trains remains a challenge. Various factors can negatively impact train punctuality, including accidents, construction works, and the main focus of this study: adverse weather conditions.

Maintaining a reliable and punctual rail network is of paramount importance, not only for the smooth functioning of daily commuting but also for the broader economic and social well-being. Delays in train services can disrupt productivity, impact various sectors of the economy, and affect customer satisfaction. While several factors contributing to train delays have been explored, the role of weather remains a key yet relatively unexplored aspect.

Switzerland experiences diverse and changeable weather, making it crucial for SBB to gain insights into the meteorological factors that affect train delays and the extent of their impact. Moreover, climate change is causing extreme weather events to become more frequent and severe, posing additional challenges to maintaining punctuality in the future.

Therefore, this project aimed to analyze the effect of adverse weather conditions on the punctuality of Swiss passenger train services, in order to provide valuable insights to SBB's punctuality team. The results show that rain and snow have a significant (negative) impact on train delays, while strong wind and extreme temperatures were not concluded to have a substantial effect. Rain is of particular interest, as it is much more frequent than the other phenomena (50% of days against less than 10% of days). This effect exists for all trains, all stops, all lines, and both train directions in the considered railway section, with differences in its magnitude.

Despite the apparent correlation between adverse weather conditions and train delays, weather parameters alone were not found to be reliable predictors of train delays. However, they could serve as valuable features when incorporated into more sophisticated models. Additionally, weather conditions did not demonstrate a significant impact on train cancellations.

These results allow SBB to confirm the intuition regarding the negative impact of meteorology on railway punctuality. This way, the team can focus on the most important factors in future work and continue the research to identify effective measures to mitigate such impact. Additionally, the findings of this study can contribute to the development of predictive models that leverage weather forecasts to anticipate and mitigate future delays, enabling more effective communication with customers and proactive measures to minimize this impact.

Furthermore, this thesis has broader implications beyond the scope of SBB and Switzerland. Its findings can be valuable to rail operators in other regions and countries willing to improve the punctuality and reliability of their services. Moreover, the methodology and tools used in this project can be applied to other railway networks and weather datasets, enabling further research in this field.

The thesis is organized as follows. Chapter 2 provides an overview of the data used in the project. Chapter 3 presents the analysis of this data and discusses the impact of weather on train punctuality. Chapter 4 describes the efforts to predict train delays given the weather forecast. Chapter 5 summarizes the findings and concludes the project. Finally, Chapter 6 discusses the limitations encountered during the research and outlines potential avenues for future work. The Appendices provide supplementary information that complements the main text.

1.1 Causality

Since this project aims to analyze the effect of weather on train punctuality, it is important to understand the different variables that could affect train delays and account for possible confounders. The causal graph in Figure 1.1 presents a simplified representation of the situation. The main question is whether any common factors cause adverse meteorological conditions and train delays.

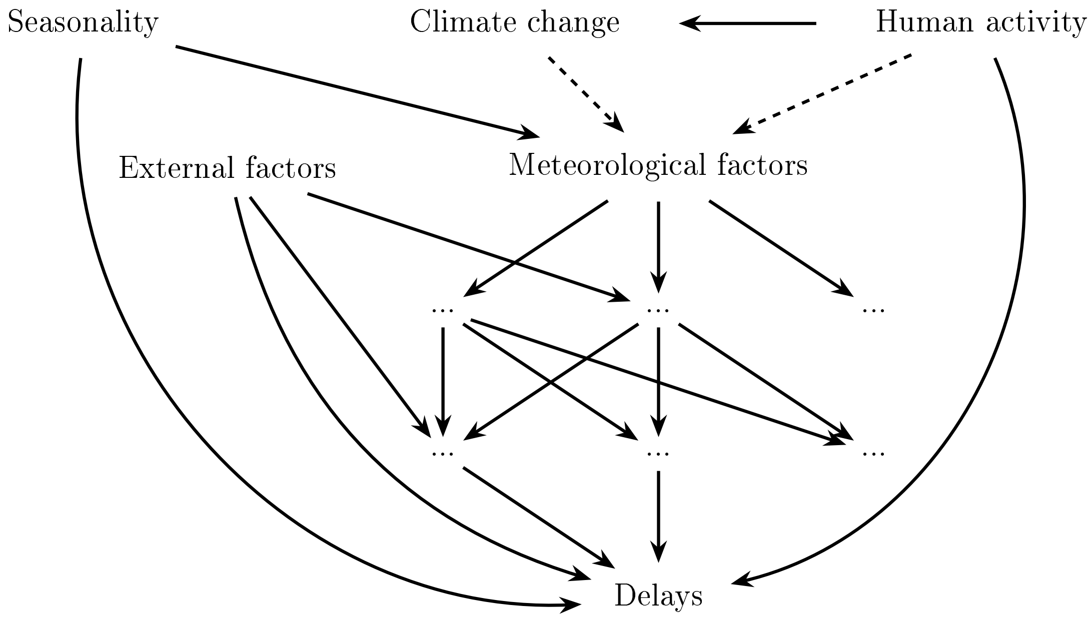


Figure 1.1: Causal graph (simplified version)

Referring to weather as the short-term atmospheric conditions in a specific region, we can generally consider it as an independent actor unaffected by other phenomena. However, some factors must be taken into account. First, seasonality clearly affects meteorological conditions (e.g., it only snows in winter) but can also affect delays through other phenomena (e.g., more people moving for vacation in summer). Therefore, it will be important to stratify the analysis by season. Second, climate change is included in the graph with a dashed line to indicate that it can influence long-term weather patterns (e.g., increasing temperatures and frequency of hurricanes) rather than immediate, unpredictable changes in weather. Human activity can also affect meteorology in the long term through climate change, or in the mid-term through activities like deforestation, pollution, or urbanization.

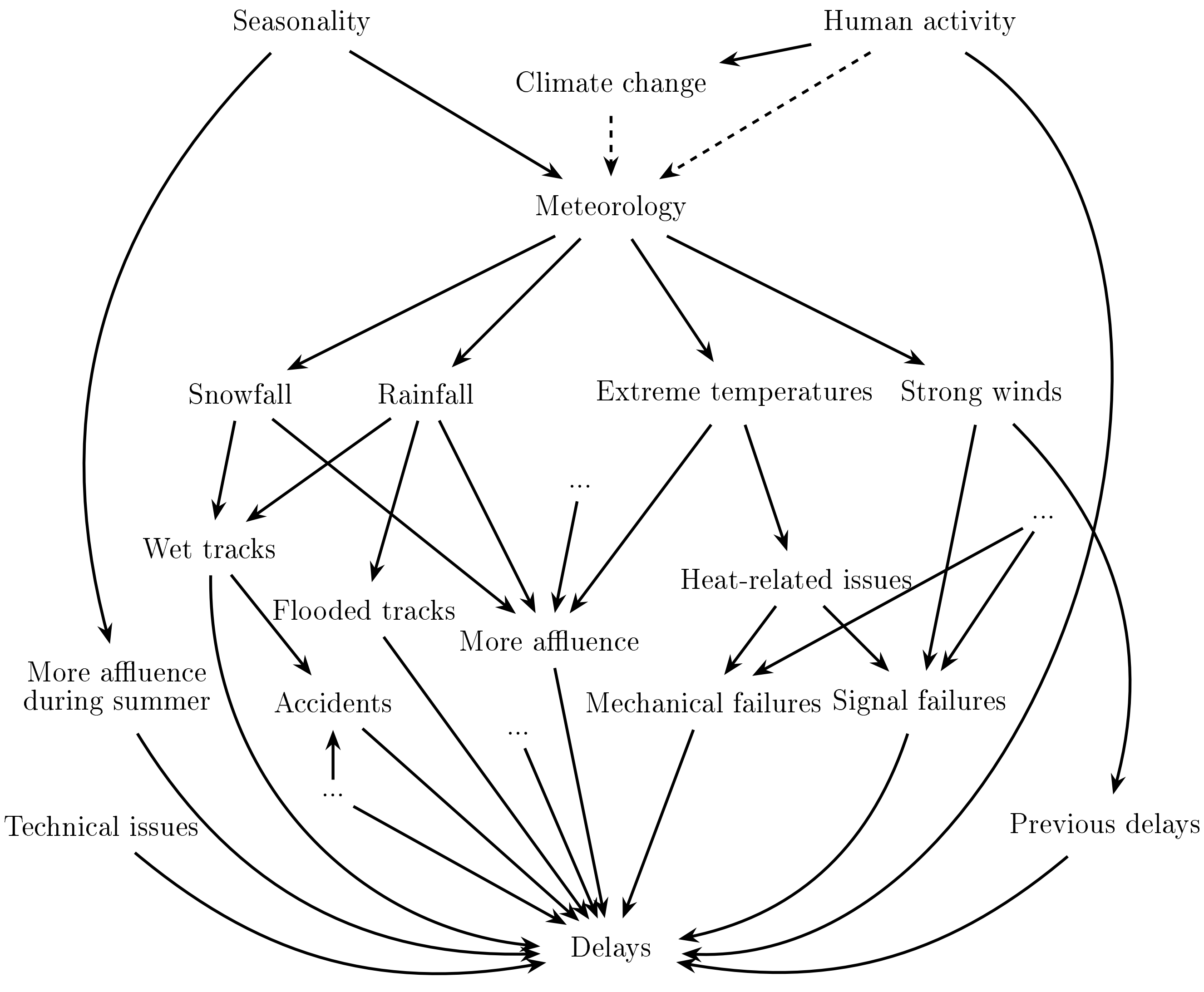


Figure 1.2: Causal graph (extended version)

However, it can only directly impact short-term weather in extreme events (e.g., a nuclear plant explosion), so I also use a dashed line. For these reasons, although climate change and human activity could potentially act as a confounder by affecting both weather and delays through other factors, I do not consider them in this project.

It is evident that there exists a complex chain of factors between meteorological conditions and train delays. These unobserved factors, represented by ellipsis in Figure 1.1, may interact with each other and be influenced by external factors unrelated to weather. All these factors can contribute to delays in various ways. For instance, meteorological conditions can lead to rainfall, which in turn can cause both flooding and more passengers boarding the train, subsequently resulting in more train delays. External factors such as previous delays on the same line, delays of trains coming from other countries, accidents, or technical issues (e.g., a broken door) may or may not be influenced by weather, and can also impact delays. The extended causal graph in Figure 1.2 provides this broader perspective, although it is not intended to be exhaustive.

In summary, a complex causal chain exists between weather and train delays, involving numerous intermediate factors. However, the key point is that weather and train delays are not simultaneously affected by other phenomena, except for seasonality. Therefore, we can focus on studying the effect of weather on train delays without the concern of other confounders that could impact both variables. From here onwards, I use the words 'impact', 'effect', 'influence', and 'relation' interchangeably to refer to the relationship between weather and train delays, subject to the assumptions in this section.

1.2 Previous work

Before this project, the punctuality team at SBB already had the intuition that bad weather conditions (such as rain, snow, or strong wind) negatively impact their train punctuality. However, only a few studies had been attempted until now, which failed to get any clear conclusions. To the best of the author's knowledge, no study had previously been done in Switzerland related to this subject outside of the team that commissioned it. However, there exist several previous projects in other countries aiming for the same goal.

[1] explored the factors that impacted the punctuality of trains in Sweden in 2015, including but not limited to weather variables (rain, snow, wind, and temperature). The authors found that punctuality drops with extremely hot or cold temperatures. They also observed that temperature variation during the journey negatively affects punctuality, as well as high wind speeds, rainfall, and snowfall. They used daily and hourly measurements. [2] studied the impact of weather on train delays in Norway from 2007 to 2016. They also used rain, snow, wind, and temperature as the weather variables on a daily and weekly basis. They found that extreme cold temperatures and snow depth are the most significant weather-related factors. [3] analyzed the impact of weather on train delays in the Greater Dublin Area (GDA) in eastern Ireland. They concluded that rain is the principal cause of observed delays, together with wind and temperature. They also observed a significant interaction between meteorological measurements. [4] show the negative impact of wind gusts, extreme temperatures, precipitation, snow, and leaves on Dutch passenger train punctuality and cancellation rate. The study distinguishes between the direct effects of weather conditions and the indirect effects through disruptions in the railway infrastructure. Snow, leaves, and high temperatures are found to almost only affect delays indirectly, while the other variables do also affect them directly. [5] employed a gradient-boosted regression tree predictive model for estimating train arrival delays at individual stations in a Chinese railway system, leveraging weather observations as a contributing factor. The model managed to predict the trend in delays but failed to predict its specific values accurately.

Based on these previous studies, it is evident that rain, snow, wind, and temperature are crucial weather factors to consider when examining their impact on train punctuality. Although the extent of their influence varies, all studies consistently demonstrate a negative impact. Daily measurements show promising potential, while hourly values offer additional granularity. While predicting train delays solely based on weather conditions is an insurmountable task, it can aid in trend prediction or be incorporated into more comprehensive models. Lastly, the interaction between weather variables should also be acknowledged as a significant factor to contemplate when analyzing the effect of weather on train delays.

1.3 Scope of the project

The Swiss railway network is vast and complex. It contains many different regions, lines, types of transport, operators... Together with the SBB team, we decided to focus on passenger trains traveling from Bern to Zurich (or vice versa) to simplify the analysis and concentrate it on one of the most critical railway sections of the country, which is of great interest to the company.

Therefore, the study comprises all passenger trains that traverse both cities in any direction. They may start their journey before and end it after this section; the only condition is that they go through Bern and Zurich at some point. For example, an InterCity 1 (IC1) traveling from Bern to Zurich originally comes from Geneva, and the same train traveling from Zurich to Bern comes from St. Gallen (see Figure 1.3).

Figure 1.3: Map of the InterCity1 (IC1) trajectory

Regarding the time scope, I worked on the last five years (2018 to 2022, inclusive) since it aligns with the available open data for train delays and weather measurements (see section 2.1).

MeteoSwiss has an important number of measuring networks, stations, and parameters, so manyptions were considered. Following the previous works in other countries, I focused on rain, snow, wind, and temperature. However, I also included other parameters (e.g., humidity, pressure) in some parts of the analysis (e.g., section 3.6).

As for the measuring networks, after a thorough comparative study, I used data from the Swiss National Basic Climatological Network (Swiss NBCN). This network is a subset of the Automatic Measurement Network and connects the major ground-based stations within the MeteoSwiss monitoring system. The different MeteoSwiss networks and parameters can be compared in this interactive map.

Since I focused on the Bern–Zurich section, I used the data from its two main stations: Bern/Zollikofen (BER) and Zürich/Fluntern (SMA) (see Figure 1.4). They both measure all the chosen weather parameters during the whole study period.

Figure 1.4: Possible train routes from Bern to Zurich and location of the Bern/Zollikofen (BER) and Zürich/Fluntern (SMA) meteorological stations

Chapter 2

Data

2.1 Data sources

Two main sets of data were used in this project: train (or railway) punctuality data and weather (or meteorological) data.

Railway data

Most of SBB's data is open and can be found in their online portals. All railway data that was used was obtained from the SBB's Open Data Portal (data.sbb.ch) or the Swiss Mobility Open Data Platform (opentransport-data.swiss). Both of them are officially supported by SBB.

In particular, the primary dataset that was used to assess punctuality is the "Actual Data" dataset, which contains the target and actual departure and arrival times for all transports in Switzerland (including, but not limited, to all SBB's passenger trains). This data has been collected by SBB at least since 2018.

Meteorological data

SBB has its own weather stations spread all over the country. However, for this project, we decided to use the data provided by MeteoSwiss, the Federal Office of Meteorology and Climatology, as it is considered the official and most accurate source of meteorological data in the country.

MeteoSwiss provides part of its data in the Swiss Open Data portal. In particular, it provides metadata about their measuring networks, current (live) measurements, and daily records for several parameters. However, it does not provide more granular data for all the stations or the whole measuring period. Similarly, on its website, it only provides metadata about the measuring stations and weather data for the past few days.

Therefore, hourly data was obtained from their IDAWeb portal, which provides access to their complete records of data. This portal is private and only accessible under request and permission, but it is open for research purposes for students and teachers.

2.2 Data processing

Even if this section is relatively short, a significant part of the project's time consisted in finding, obtaining and preprocessing the data.

Railway data

As explained in section 2.1, train punctuality was computed from the Actual Data open dataset. This dataset contains the target (scheduled) and actual (real) arrival and departure times of each public transport vehicle in Switzerland, including buses, tramways, cargo trains, cable cars, and more.

Since the focus was on passenger trains from Bern to Zurich (or vice versa) (see section 1.3), important data filtering and preprocessing were needed. No API is provided for the whole Actual Data Archive (which ranges from 2018 to 2023), so the nearly 2000 zip files (one for each day) had to be downloaded, extracted, filtered, processed and joined one by one. All this was done using Python and bash scripts in remote servers.

This first processing resulted in a single unified dataset with over 250.000 trains from Bern to Zurich (or vice versa) from 2018 to 2023 (both included), containing around 0,5% of the whole data in the IstDaten Archive.

Then, a second processing was made to clean the dataset. I removed all additional trains (that is, extra trains that are added on special occasions) and all canceled trains (that is, trains that did not run or that stopped halfway). See section 3.8 for a dedicated analysis of the relationship between weather conditions and train cancellations.

Moreover, each departure and arrival time is accompanied by a status that indicates if that value was measured, was just a prediction, or is missing or unknown. I removed all trains that did not have complete 'real' times since they could provide wrong delay values.

The final result of the processing was a dataset with nearly 200.000 trains and around 775.000 stops, accounting for 80% of the uncleaned dataset.

Meteorological data

As explained in section 2.1, I used both daily measurements for the NBCN network available in the Swiss Open Data portal and hourly measurements from the IDAWeb portal. They are all provided in CSV format, so I downloaded and parsed them using Python scripts and notebooks.

2.3 Data exploration

This section shows a separate analysis of the railway punctuality and meteorological data. This is important to be able to understand the posterior combined analyses better. For conciseness, it is attached in Appendix A.

The following processing and analyses were mainly carried out using Python scripts, Pandas, and Jupyter Notebooks. For plotting, `matplotlib`, `seaborn`, `plotly`, and `folium` were used. For models, `statsmodels` and `scikit-learn` were used.

Chapter 3

Analysis

This chapter presents the analysis the effect of weather conditions on train delays, by merging both data sets. This makes it possible to obtain concrete numbers about this impact and show them in a graphical and intuitive way. This analytical part is already of great value to SBB, as it allows us to know where, when, and how to put efforts into reducing this impact or whether it exists at all.

I first carried out a separate analysis for each weather variable (rain, snow, temperature, and wind). The idea was to check for a significant difference in train delays between, for example, rainy and non-rainy days. Both daily and hourly weather values were used. The results are summarized in section 3.5.

Following some intuitions from the SBB team, I extended the analysis of rain to account for the effect of consecutive rainy or dry days (subsection 3.1.1) or the lag effect of recent rain (subsection 3.1.2).

Then, I ran different clustering models to find groups of days with similar weather conditions (see section 3.6). This allowed to check how different kinds of days (e.g., rainy, windy, sunny days) affect train delays instead of checking variable per variable, and account for the relationship and interaction between them.

Finally, I checked if the conclusions drawn from the Bern-Zurich section can also apply to other railway sections (see section 3.7), and whether the cancellation of trains is affected by weather conditions (see section 3.8).

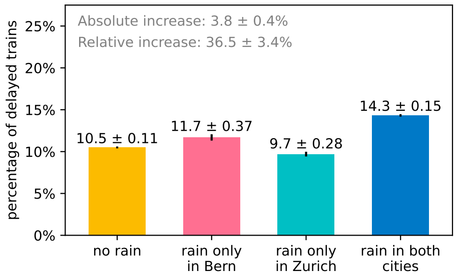
3.1 Rain

Does rain affect delays? Where and when?

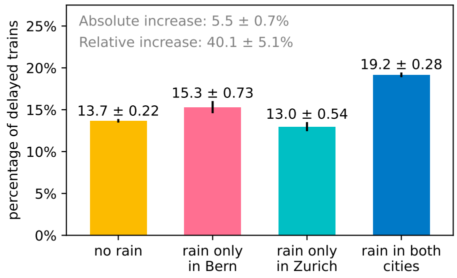
Figure A.15 showed how often each situation (no rain, rain only in Bern, rain only in Zurich, and rain in both cities) occurs. For each situation, one can compute the percentage of delayed trains (i.e., those arriving at their destination three or more minutes later than planned), as shown in Figure 3.1. Figure 3.1a shows the percentage of delayed trains for each situation, for all trains arriving at all stops. The percentage of delayed trains is higher when it rains in both cities than when it rains only in Bern, which in turn is higher than when it does not rain at all. Surprisingly, the lowest delay occurs when it only rains in Zurich. This might be because it does not rain much when it just rains in one of the two cities (see Figure A.18), and this situation is relatively uncommon (see Figure A.15). As written in Figure 3.1a, the absolute increase between the first and last bars is 3.8%, and the relative increase is 36.5%. This already indicates a (negative) effect of rain on train delays. The difference is more pronounced if we account only for arrivals at the destination (i.e., arrival to Bern or Zurich), and not at every intermediate stop (see Figure 3.1b).

If we separate between indirect and direct trains, we can observe that direct trains suffer from more delays in all situations (see Figure 3.1c and Figure 3.1d). This is expected, as indirect trains have more slack time in their schedule and thus are more likely to be able to recover from delays. Furthermore, if we separate by direction, we can observe that all delays are higher for trains traveling from Zurich to Bern than the other way around (see Figure 3.1e and Figure 3.1f). In the next subsection, I address the fact that trains already carry a certain delay when departing from Bern or Zurich, and, of course, this delay at origin depends on the direction. In any case, the relative effect of rain is similar for both directions, regardless of where it rains, which might be counterintuitive.

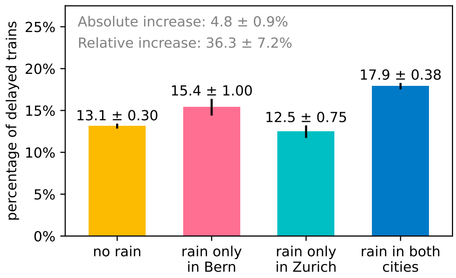
The same plots can be produced for the average arrival delay or the travel time difference (instead of the percentage of delayed trains). The conclusions remain the same, so for the sake of brevity, I do not include them in this report.



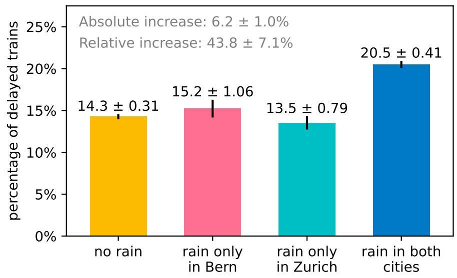
(a) All trains, arrival at all stops



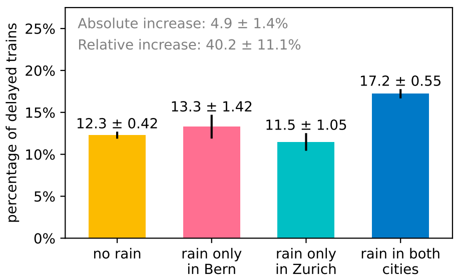
(b) All trains, arrival at destination



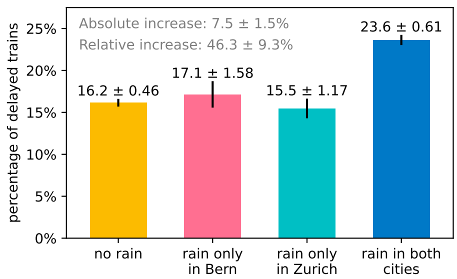
(c) Indirect trains, arrival at dest.



(d) Direct trains, arrival at dest.



(e) Direct trains from Bern to Zurich, arrival at destination



(f) Direct trains from Zurich to Bern, arrival at destination

Figure 3.1: Percentage of delayed trains for each situation (no rain, rain only in Bern, rain only in Zurich, and rain in both cities), for different subsets of data.

Effect of rain during the journey

Accounting only for the delay at the destination could be misleading since trains may already carry a certain delay when departing from Bern or Zurich. Returning to the example in Figure 1.3, an InterCity 1 (IC1) traveling from Bern to Zurich actually comes from Geneva. In contrast, the same train traveling from Zurich to Bern comes from St. Gallen (which is much closer). Therefore, depending on the direction and line, one could expect different delays at the origin.

First, Figure 3.2a shows the percentage of delayed trains in each location of the journey for all indirect trains. While trains manage to make up some of the delays at the stations (by reducing their scheduled dwell time), they incur delays between origin and destination. For example, an IC1 arriving late to Bern from Geneva may manage to leave Bern on time but not arrive at Zurich on time.

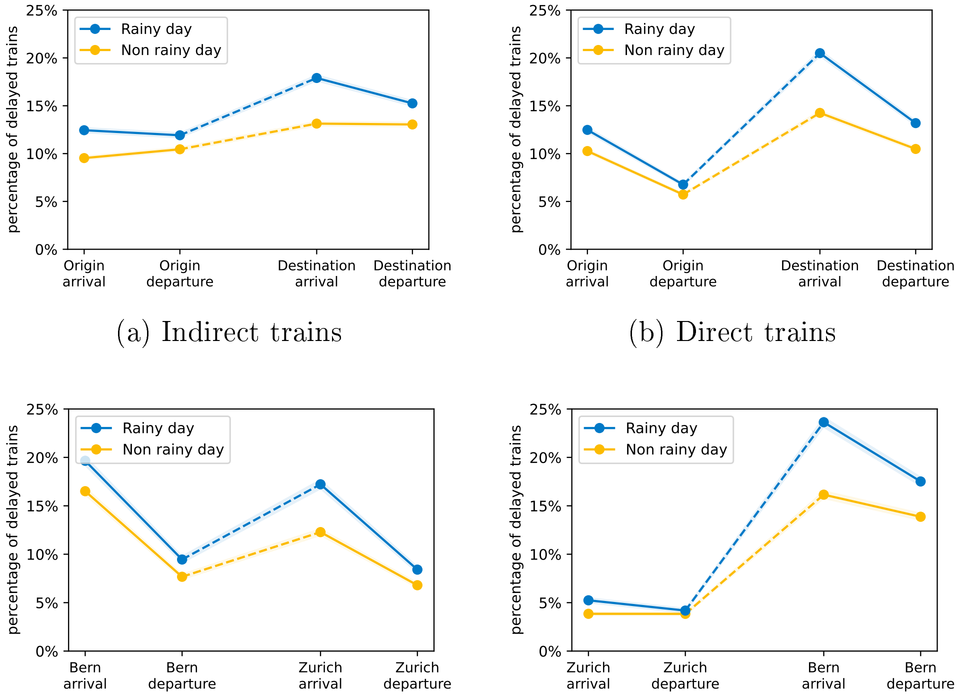
The negative impact of rain is clear here: the percentage of delays is higher in all journey locations. Moreover, the steeper slope in the middle tells that more trains get delayed during the travel time when it rains than when it does not rain.

Second, Figure 3.2b shows the percentage of delayed trains in each location of the journey for all direct trains. In this case, the delays at the origin are much lower than for indirect trains. This is expected, as direct trains usually stop for more time and thus are more likely to be able to recover from previous delays. However, the average delay at the destination is higher than for indirect trains. This is also expected, as direct trains do not stop in between and thus are more likely to be delayed during the journey. Here, the effect of rain during the journey is even more pronounced than for indirect trains (see the steeper slope in the middle).

Figure 3.2c and Figure 3.2d show the same plots separated by direction. Even if fewer trains are delayed when departing from Zurich than from Bern, more trains are delayed when arriving at Bern than at Zurich. In other words, the central slope is much steeper for the Zurich-Bern direction than for the other. Again, the effect of rain is clear in both directions.

Selecting other subsets of data only provides the same conclusion: trains catch up delays at stations but incur delays during the journey, and rain has a negative impact on train delays.

The same plots can be produced with the average or median delay (in minutes) at each location (instead of the percentage of delayed trains). The conclusions remain the same, so for the sake of brevity, I do not include them in this report.



(a) Indirect trains

(b) Direct trains

(c) Direct trains from Bern to Zurich

(d) Direct trains from Zurich to Bern

Figure 3.2: Percentage of delayed trains along the journey on rainy and non-raniy days, for different subsets of data.

Major delays

Another interesting question is how rain affects major delays, say, delays of more than 10 minutes. Figure 3.3 shows the percentage of trains with a delay greater than a certain threshold. As expected, the higher the delay threshold, the lower the percentage of trains with that delay. The proportion of delayed trains is always higher when it rains.

To see how this delay increase is related to the base (yellow) delay, Figure 3.4 shows the relative increase from non-rainy to rainy days (i.e., 'the difference between lines divided by the yellow line'). Rain always causes an increase of delays (between 30% and 60%), no matter the chosen threshold. The greatest delays (more than 20 minutes) notably increase when it rains.

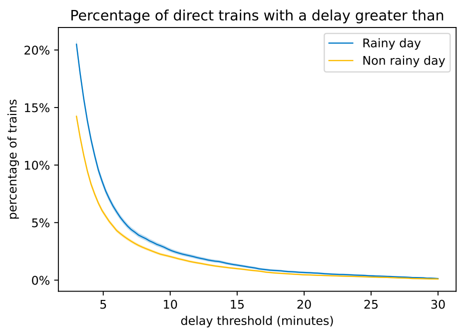


Figure 3.3: Percentage of direct trains with a delay greater than a certain threshold, when it rains and when it does not rain.

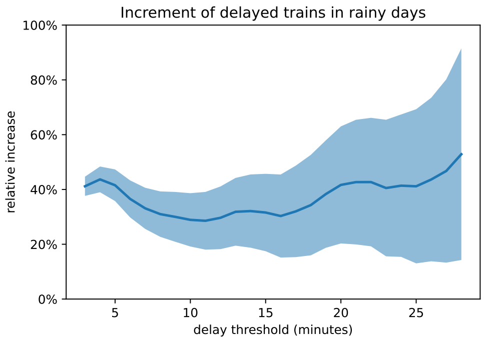


Figure 3.4: Relative increase of delayed trains on rainy days with respect to non-rainy days.

The same plots can be produced with other subsets of trains, but the conclusions remain the same. The conclusions remain the same, so for the sake of brevity, I do not include them in this report.

Amount of rain

Until here, I checked the impact of rain as a binary variable (i.e., it rained or it did not rain that day). However, another intuition is that the more it rains, the more delays should occur. This section uses the variable's continuous (and not the binary) version to check this.

Figure 3.5 shows the percentage of delayed trains per day according to rain. The fitted linear regression model indicates an increasing trend but is far from a perfect linear relationship. The correlation coefficient (0.23) suggests a weak positive linear association between the average daily rainfall and the percentage of delayed trains. However, the relatively low correlation indicates that the relationship is not very strong. The model's predictions exhibit significant error levels, as indicated by the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Additionally, the R-squared score of 0.05 indicates that only 5% of the variation in the percentage of delayed trains can be attributed to the average daily rainfall. Consequently, the data does not fit well with the model. However, it is important to note that this analysis solely considers rainfall and its right-tailed distribution, as depicted in Figure A.16.

Binning the rain in different categories and computing the average delay shows an increasing trend again (see Figure 3.6), but the confidence intervals are too large to draw clear conclusions. These results suggest that it is more critical whether it rains or not rather than how much it rains.

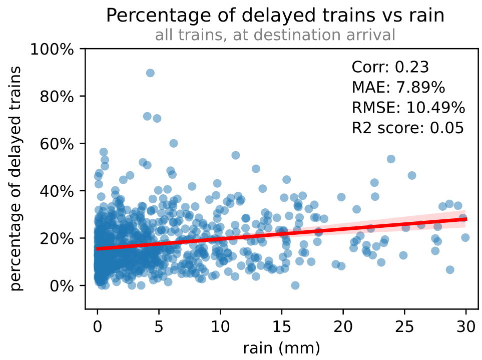


Figure 3.5: Percentage of delayed trains according to rain.

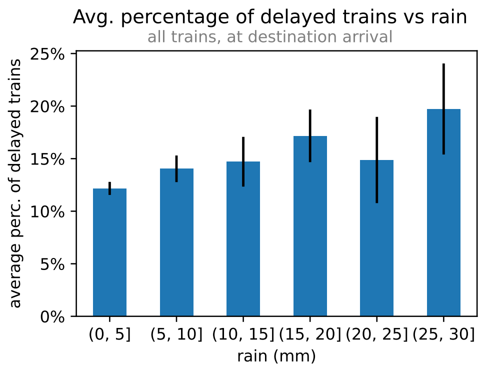


Figure 3.6: Average percentage of delayed trains according to rain.

Effect per stop and line

It might also be interesting to check whether the impact of rain exists equally in all train stops or lines. Figure 3.7 shows the percentage of delayed trains per stop according to rain. The (negative) effect of rain is similar in all stops, except for Turgi and Wynigen (where the CIs overlap), although they are also the less visited stops (see Figure A.12). Figure 3.8 shows the same plot per line. Again, there is no doubt about the negative effect of rain, which exists in all lines, and especially in the InterCity services, as commented above.

Effect throughout time

As explained in section 1.1, it is important to account for seasonality in the analysis. To clarify, one must check if the delay increase exists in all months or seasons to confirm the negative effect of rain on train delays (and not a mere correlation with other seasonality-related factors).

Figure 3.9 shows the effect of rain throughout the year (i.e., the percentage of delayed trains per month according to rain). Rain indeed affects delays throughout the whole year, especially in fall, which is also when there are more base delays. This shows that, although there is indeed certain seasonality in the delays, the impact of rain is always present.

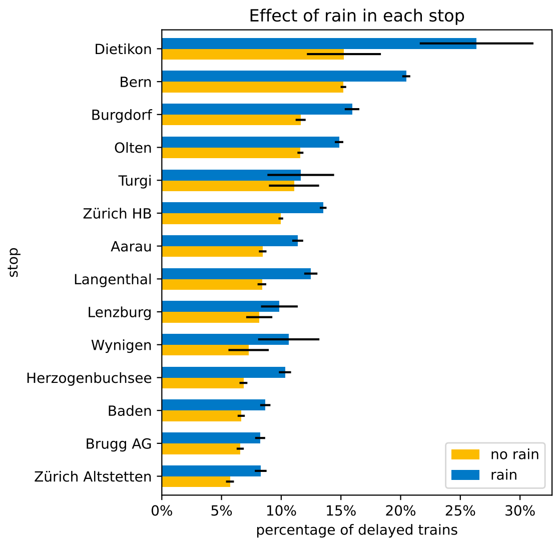


Figure 3.7: Effect of rain per stop

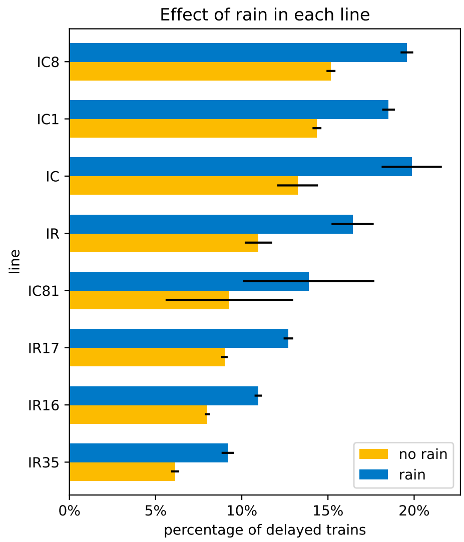


Figure 3.8: Effect of rain per line

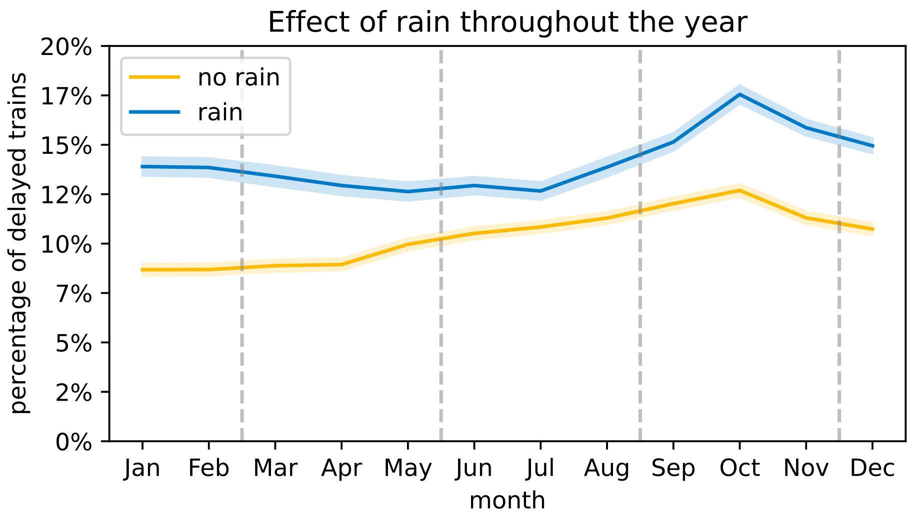


Figure 3.9: Effect of rain throughout the year

Figure 3.10 shows the effect of rain throughout the week. Weekends have significantly fewer delays than weekdays, both when it rains and when it does not. Surprisingly, the relative effect of rain seems to increase during the weekdays, starting at 20% on Mondays and ending at nearly 50% on Fridays, just to decrease again during the weekend.

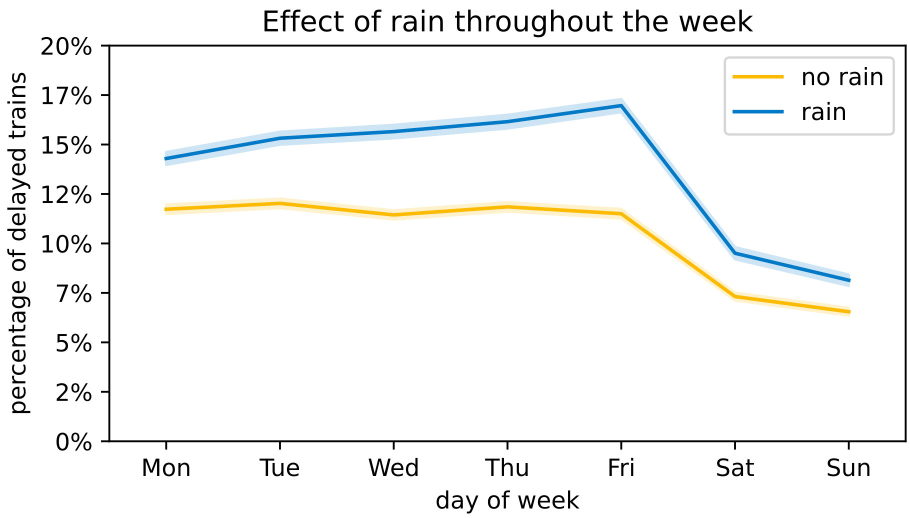


Figure 3.10: Effect of rain throughout the week

Finally, Figure 3.11 shows the percentage of delayed trains throughout the day, when it rains and when it does not. The delays vary considerably during the day: there are more delays at peak times (morning and late afternoon), and especially at night (from 21h). The absolute effect of rain is always similar (around 10% more delays).

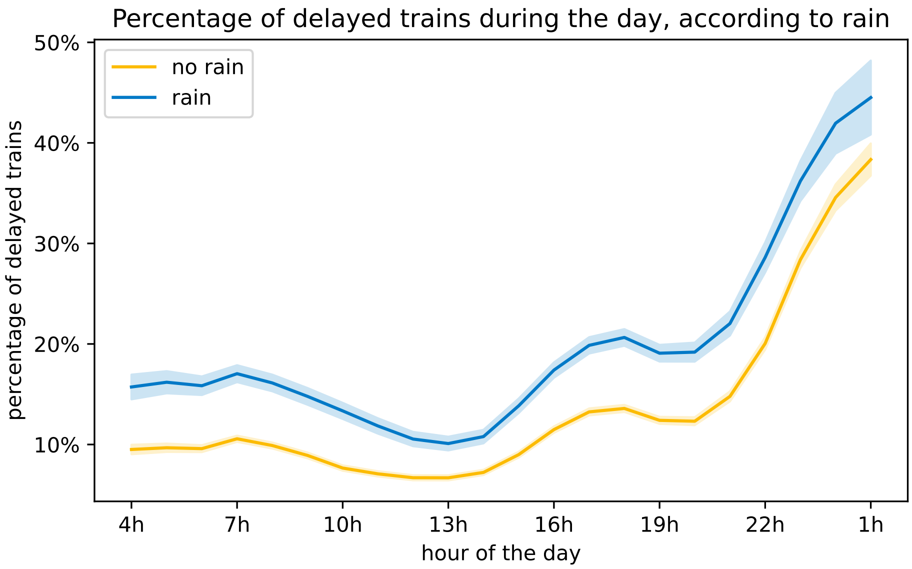


Figure 3.11: Percentage of delayed trains per hour according to rain.

3.1.1 Consecutive rain

The preceding analyses have shown that rain impacts train delays. A further intuition given by the SBB team is that consecutive rainy days (e.g., four days in a row) might have a different impact than isolated rain. Symmetrically, several consecutive “dry” (non-rainy) days may also have a cumulative effect (because the tracks get dirty or greasy, for example). Figure 3.12 shows the consecutive rainy days throughout the five years. Figure 3.13 shows the distribution of the length of the rainy and dry periods, respectively.

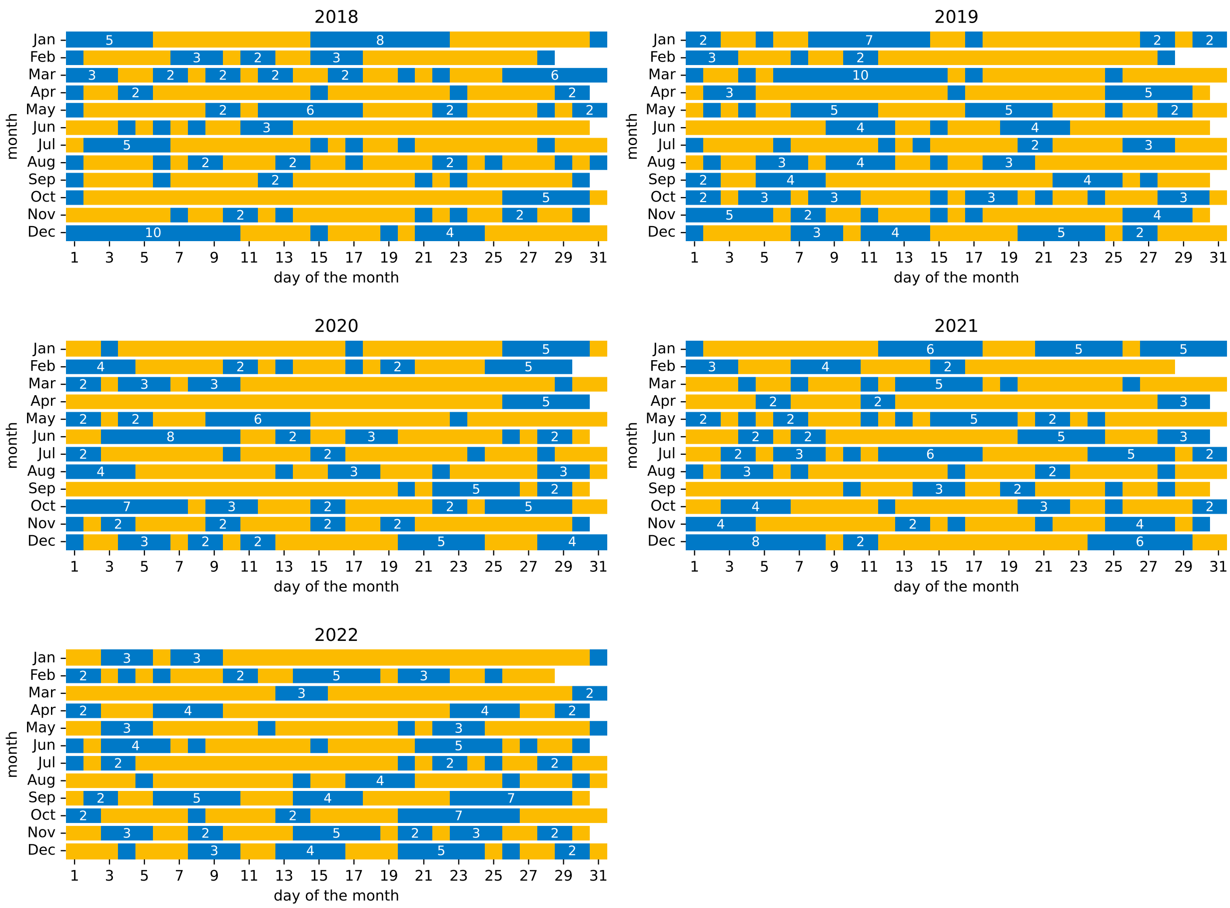
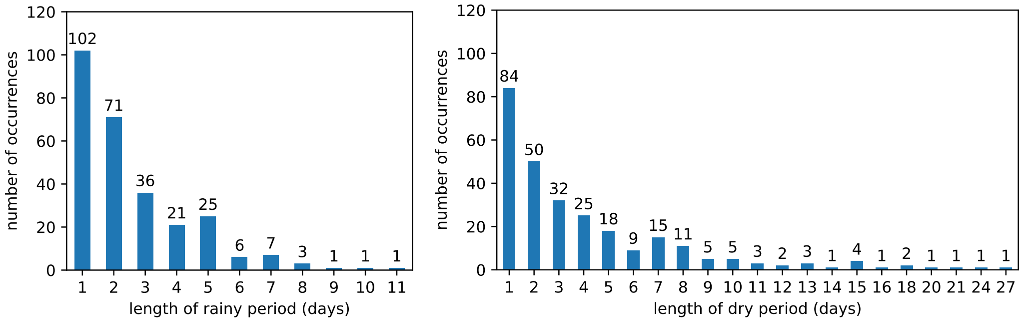


Figure 3.12: Series of consecutive rainy days over the five years.



(a) Rainy periods.

(b) Dry periods.

Figure 3.13: Distribution of the length of rainy and dry periods.

Figure 3.14 shows the average delay on the day previous to the rain, during the rainy period, and on the following day; for each period length. The only clear pattern is that the average delay during rainy days is higher than the day before and the day after (despite the overlapping confidence intervals). However, there is no apparent difference between the day before and after. Indeed, the independent two-sample t-test between the previous and the following day always resulted in a p-value over 0.05, while the test between the day before (or the day after) and during the rainy period always gave a p-value below 0.05. This suggests that the number of consecutive rainy days does not affect the posterior delays.

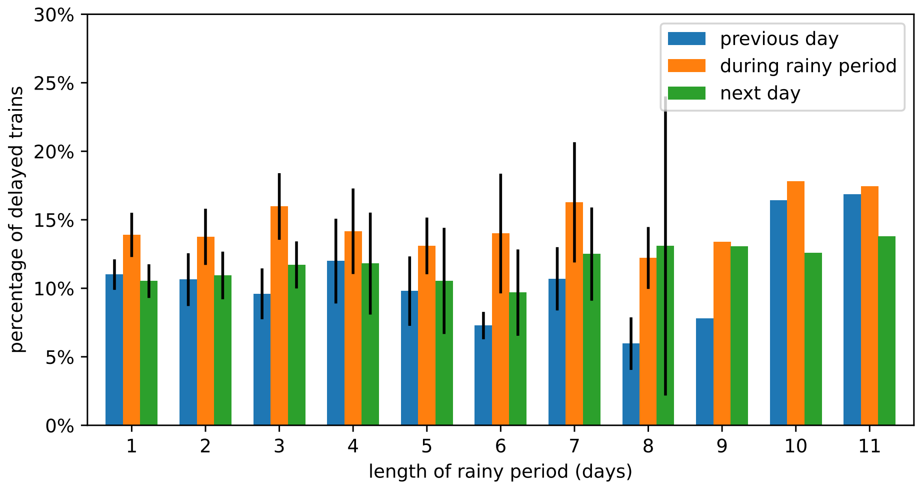


Figure 3.14: Average delay before, during, and after a rainy period.

Figure 3.15 shows the average delay on the first and last day of the period. Again, the plot and corresponding tests indicated no significant difference between both days. This suggests that the effect of rain is not cumulative.

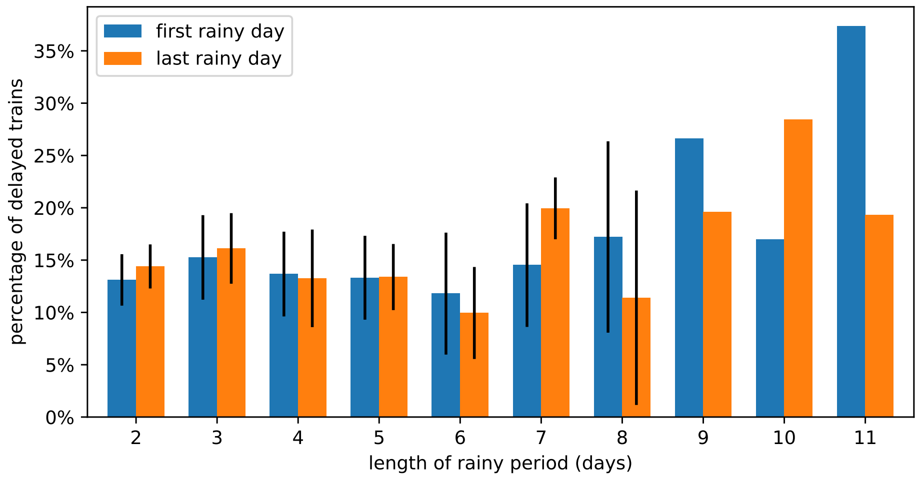


Figure 3.15: Average delay on the first and last day of the rainy period.

Symmetrically, Figure 3.16 replicates Figure 3.14 but for dry periods (limited to length 10). The average delay on the day before and after the dry period is similar, and the average delay during the dry period is lower than the day before and after, as expected.

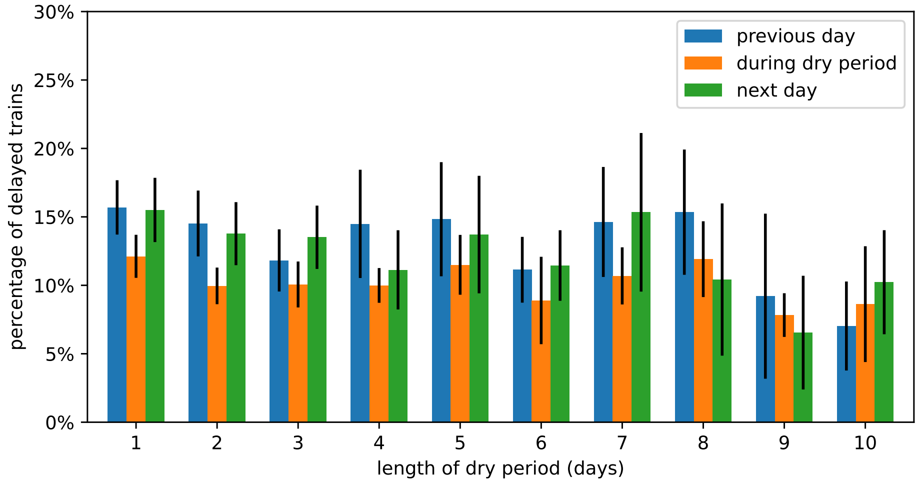


Figure 3.16: Average delay before, during, and after a dry period.

Finally, Figure 3.17 shows the average delay on the first and last day of the dry period. Again, the plot and corresponding t-tests indicated no significant difference between both days. This suggests that the number of consecutive dry days does not affect the delays either.

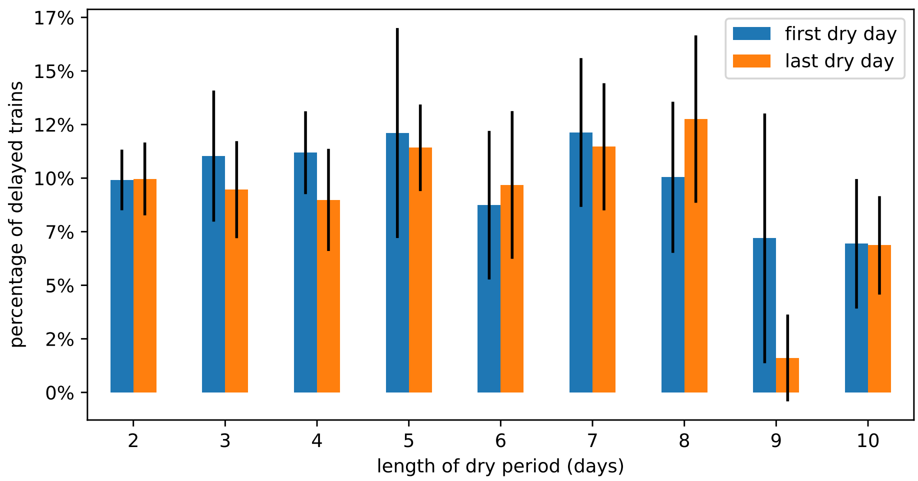


Figure 3.17: Average delay on the first and last day of the dry period.

3.1.2 Recent rain

Another intuition is that rain might have a lagged effect, meaning that rain in the previous days can also affect current delays, either positively or negatively. In fact, the SBB team had the intuition that rain causes delays (as previously shown), but it can also help clean the tracks, which might ultimately help reduce delays.

Figure 3.18 shows the percentage of delayed trains according to the most recent rain in the previous week. It can be clearly observed that there is no carryover effect, at least on a daily basis, as all delays are similar except for when it rains the same day.

To check this effect in a shorter term (hourly granularity), Figure 3.19 shows the effect of recent rain over the last 24 hours. In this case, one can observe that train delays are especially affected by very recent rain. The apparent increase from 10 to 21 hours ago is not statistically significant (Independence TTest p-value: 0.20), but there is indeed a significant difference between rain during the trip and rain 10 hours before (p-value: 1.59e-18).

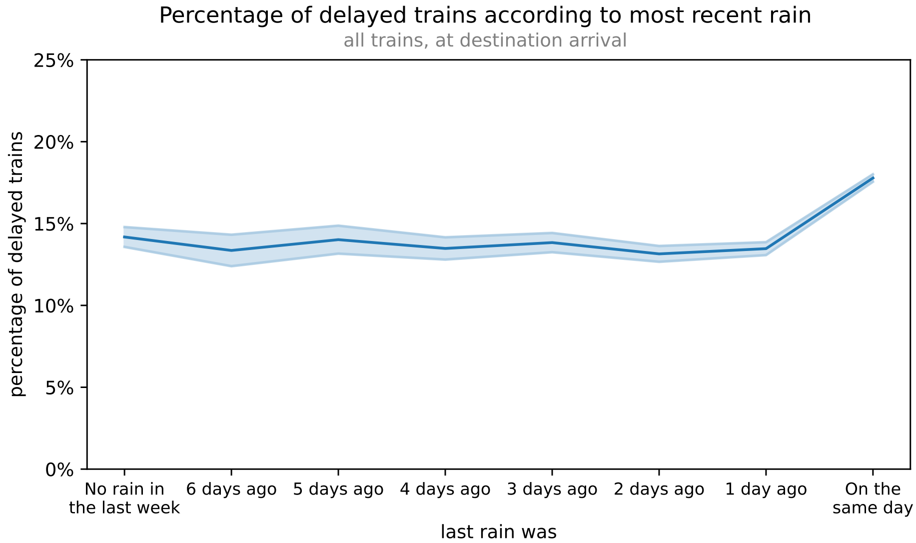


Figure 3.18: Percentage of delayed trains according to most recent rain in the previous week.

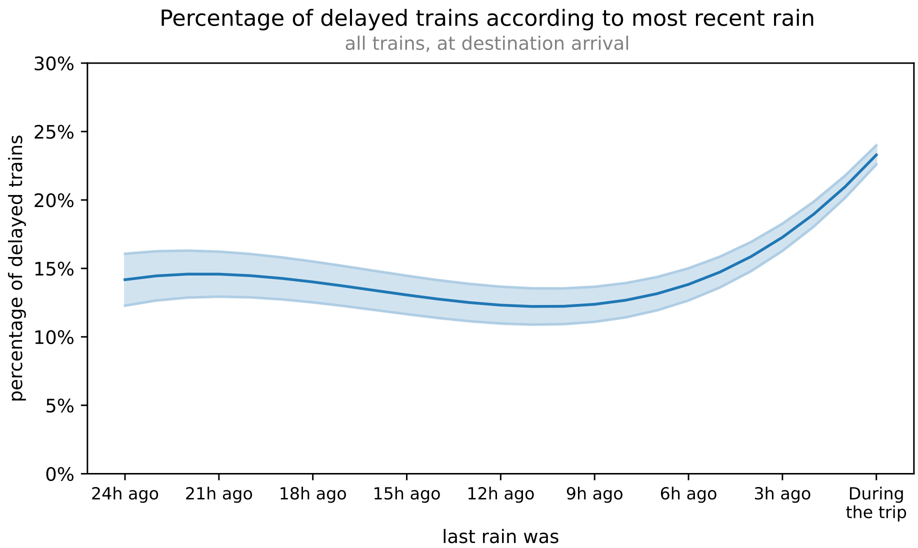


Figure 3.19: Percentage of delayed trains according to most recent rain in the last 24 hours.

To quantify this effect more precisely, Figure 3.20 shows the average delay for the last 6 hours, compared to the base delay (no rain in the last 24 hours). Only 6 hours are needed to 'go back to normality' (go back to the base delay).

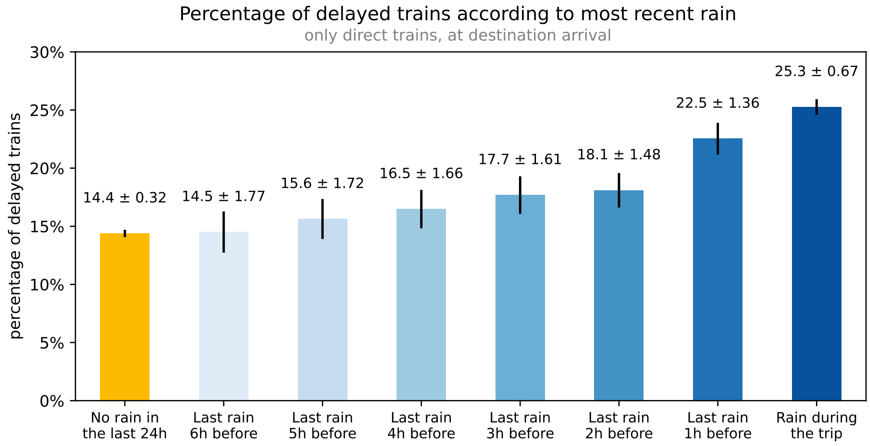


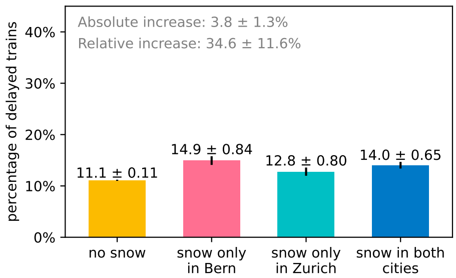
Figure 3.20: Percentage of delayed trains according to most recent rain in the last 6 hours.

3.2 Snow

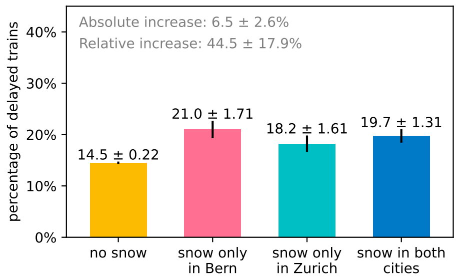
This section reproduces a similar analysis to that of section 3.1, but for snow. To account for the fact that snow occurs only in winter months (see section A.2), the following plots are limited to December, January, and February. For the sake of brevity and to avoid redundancy, only the most relevant plots are included here.

Figure A.20 showed how often each situation (no snow, snow only in Bern, snow only in Zurich, and snow in both cities) occurs. Analogously to Figure 3.1, Figure 3.21 shows the percentage of delayed trains for each of these situations. Again, delays are higher when it snows in at least one city than when it does not snow, and direct trains are more affected than non-direct ones. The relative effect of snow is almost the same in both directions. Surprisingly, trains are especially delayed when it snows in Bern (and not in both cities simultaneously), no matter the direction. The relative impact of snow is higher than for rain in all cases: up to 60% between no snow and snow in Bern.

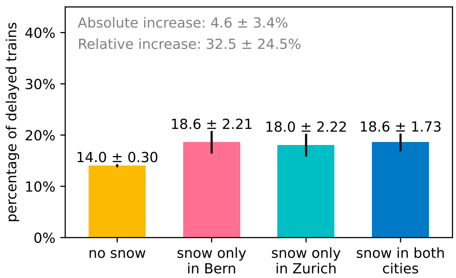
As for the effect throughout the journey, Figure 3.22 shows a very similar result to that in Figure 3.2. The conclusions are the same as for rain: delays are higher in all journey locations when it snows, and trains get delayed during the travel time while they catch up at the stations.



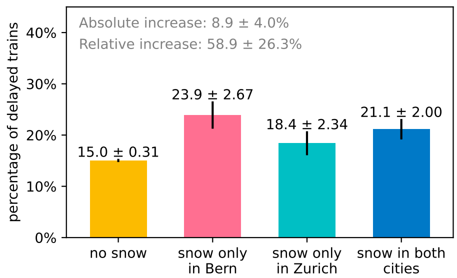
(a) All trains, arrival at all stops



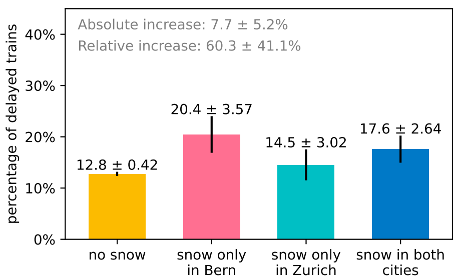
(b) All trains, arrival at destination



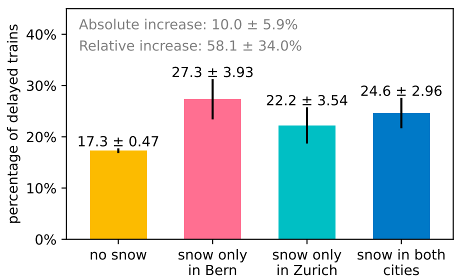
(c) Indirect trains, arrival at dest.



(d) Direct trains, arrival at dest.



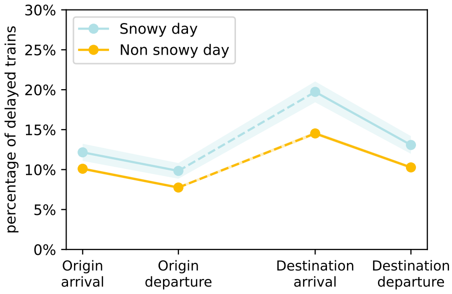
(e) Direct trains from Bern to Zurich, arrival at destination



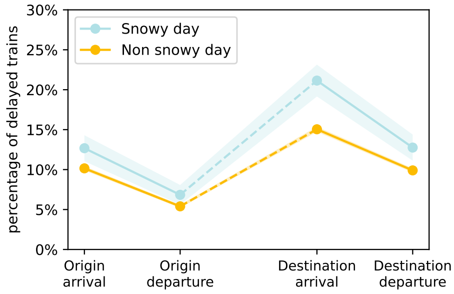
(f) Direct trains from Zurich to Bern, arrival at destination

Figure 3.21: Percentage of delayed trains for each situation (no snow, snow in Bern, snow in Zurich, and snow in both cities).

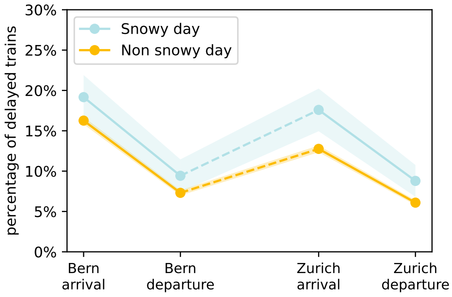
Regarding the major delays, Figure 3.23 shows the percentage of trains with a delay greater than a certain threshold. As for rain, the percentage of delayed trains is always higher when it snows. Figure 3.24 shows that snow always causes an increase in delays, up to more than double with respect to non-snow delays. Again, the greatest the delay, the more is affected by snow.



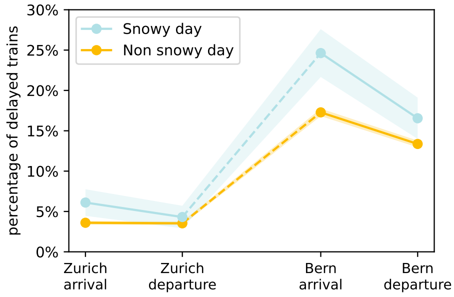
(a) Indirect trains



(b) Direct trains



(c) Direct trains from Bern to Zurich



(d) Direct trains from Zurich to Bern

Figure 3.22: Percentage of delayed trains along the journey on snowy and non-snowy days, for different subsets of data.

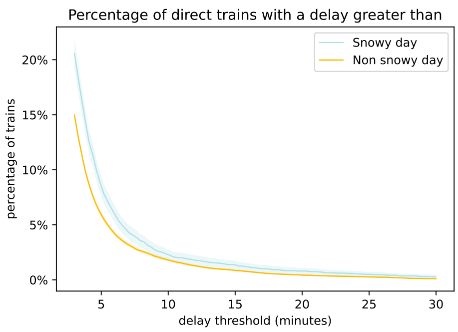


Figure 3.23: Percentage of direct trains with a delay greater than a threshold, with and without snow

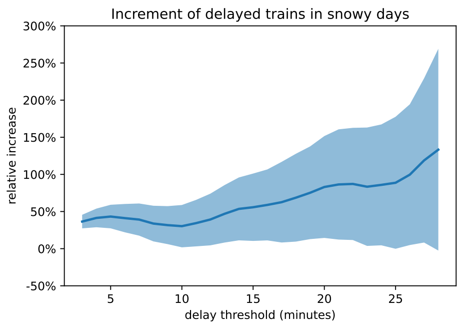


Figure 3.24: Relative increase of delayed trains on snowy days with respect to non-snowy days

The analysis of consecutive days does not apply to snow, as having two snowy days in a row is very unlikely. Likewise, no hourly analysis (including the lag effect) can be done, as snow measurements are only provided daily.

3.3 Temperature

Previous works claim that extremely hot or cold temperatures can cause delays (see section 1.2). In addition, they suggest that the temperature difference experimented on by a train can have an impact. First, Figure 3.25 shows the percentage of delayed trains with respect to the temperature. As expected, there are more delays when it is extremely cold (below -5ºC) or extremely hot (above 30ºC). However, as the confidence intervals are too large, it is hard to draw clear conclusions. Moreover, these extreme temperatures are rare (see Figure A.23).

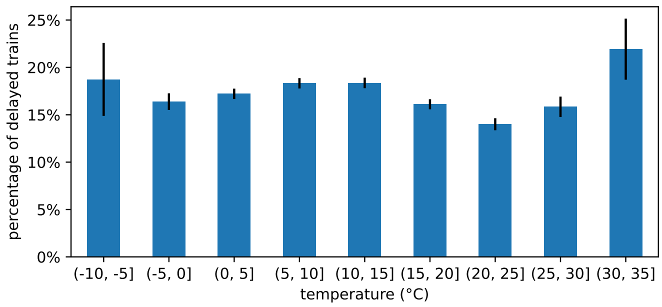
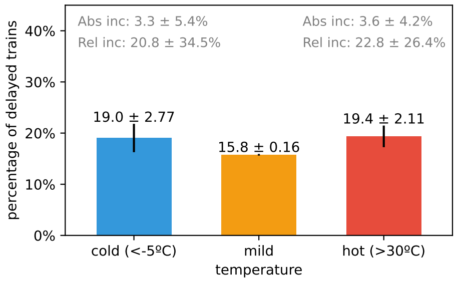


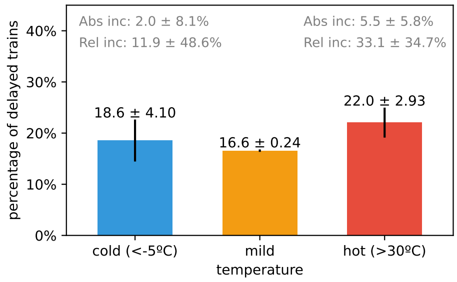
Figure 3.25: Percentage of delayed trains per temperature bin.

Figure 3.26 shows the average proportion of delayed trains on cold, mild, and hot days (using -5ºc and 30ºC as thresholds). In both directions, the average delay is indeed higher when it is cold or hot. This can be explained, for example, because rails can get deformed by very high temperatures or passengers might prefer to take public transport instead of cycling when it freezes. However, given the confidence intervals, extreme temperatures cannot be considered significantly different from mild temperatures.

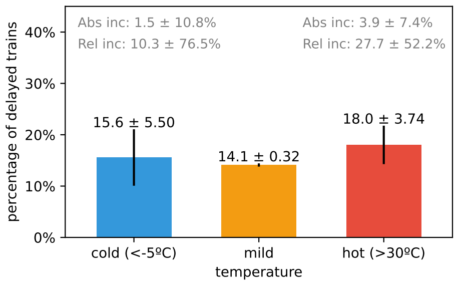
Regarding the temperature difference, Figure 3.27 shows that there is indeed a positive correlation between the temperature difference and the average delay. However, note that the effect is not symmetrical: the delays increase when there is a positive temperature difference (computed as the temperature in Bern minus the one in Zurich) but not when there is a negative one, no matter the direction of the train. This is probably because Zurich is usually warmer, so a higher temperature in Bern might be related to a cold front or another type of weather phenomenon that could cause delays.



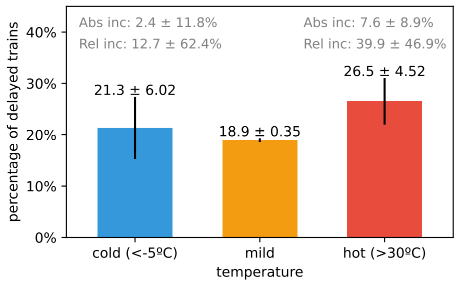
(a) All trains, arrival at destination



(b) Direct trains, arrival at dest.



(c) Direct trains from Bern to Zurich, arrival at destination



(d) Direct trains from Zurich to Bern, arrival at destination

Figure 3.26: Percentage of delayed trains according to temperature.

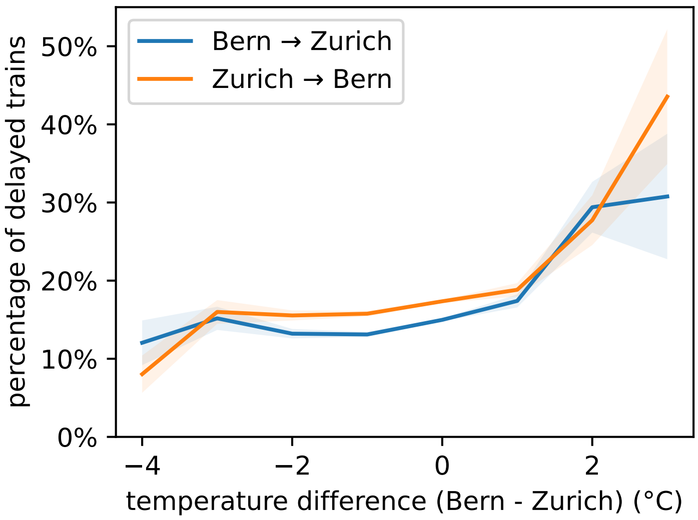
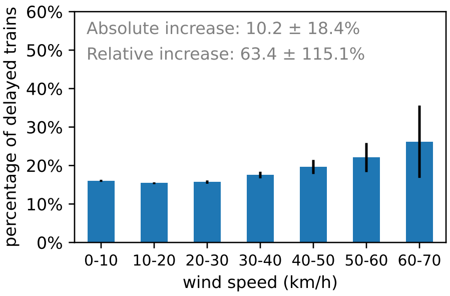


Figure 3.27: Delays with respect to temperature difference.

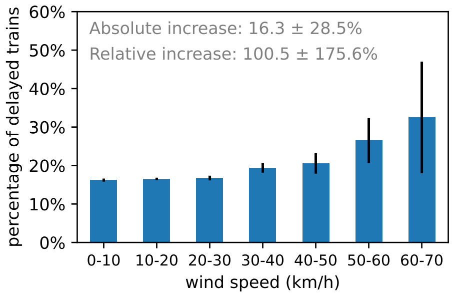
3.4 Wind

Wind can also be a factor in train delays. For example, strong winds can cause trees to fall on the tracks or make breaking harder. Figure 3.28 shows the average percentage of delayed trains according to wind speed. Its value is computed as the maximum 10-minute average wind registered during the train journey.

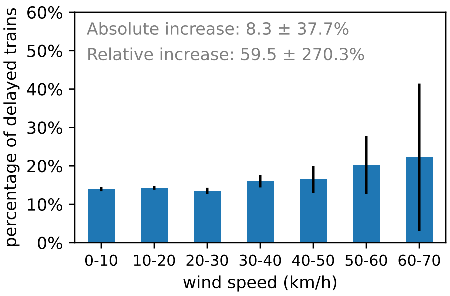
The average delay is higher when the wind speed is higher, suggesting that wind negatively impacts train delays. The difference is especially pronounced for very strong winds (over 50 km/h). However, the wide confidence intervals indicate notable variability, so as in the case of temperature, a significant effect cannot be concluded. The same pattern as rain and snow can be observed: direct trains are more affected than indirect ones, especially if they travel from Zurich to Bern.



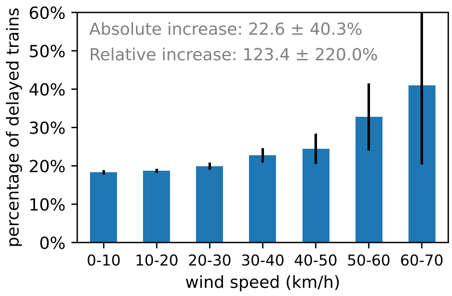
(a) All trains, arrival at destination



(b) Direct trains, arrival at dest.



(c) Direct trains from Bern to Zurich, arrival at destination



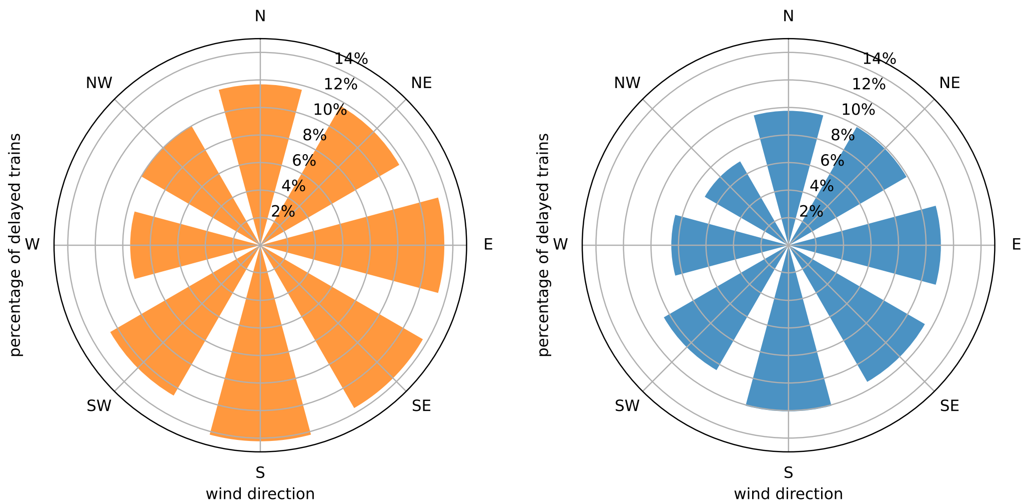
(d) Direct trains from Zurich to Bern, arrival at destination

Figure 3.28: Percentage of delayed trains according to wind speed, for different subsets of data.

Wind speed alone might not be accurate enough if its direction is not considered. For example, if the wind blows in the same direction as the train, it might speed it up, while if it blows in the opposite direction, it could slow it down. Crossed wind could also affect the train's stability.

Figure 3.29 shows the average percentage of delayed trains for each wind direction, separated by train direction. There is no observable difference between both train directions apart from the above-commented magnitude of delays.

Apparently, the wind does not help if it blows in the same direction as the train or hinders if it blows in the opposite direction. However, there are differences between wind directions. The wind blowing from the southeast (i.e., from the Alps) is correlated to more delays than the wind blowing from the northwest (i.e., from the Jura mountains), although the strongest winds come from the west and north (see Figure A.29). It is hard to tell if these are mere correlations or if there is a causality relationship, as wind direction can be directly related to other weather variables (e.g., rain or temperature) and, as said, suffers from high variability.



(a) From Bern to Zurich

(b) From Zurich to Bern

Figure 3.29: Percentage of delayed trains for each wind direction.

3.5 Summary of results

The following tables summarize the results presented in the previous subsections. They show the absolute and relative increase in delays for each weather variable. Table 3.1 shows the results for all direct trains, while Table 3.2 and Table 3.3 show the results for direct trains going from Bern to Zurich and from Zurich to Bern, respectively.

The values are taken from the preceding bar plots. For rain and snow, they indicate the increase between non-rainy (non-snowy) and rainy (snowy) days. For temperature, they indicate the increase between mild days and extremely cold (-5ºC) or hot (>30ºC) days. For wind, they indicate the increase between no wind and strong wind (>60km/h).

Wind stands as the (a priori) most significant factor, with an increase of more than 100% in delays. However, the confidence intervals are too large to draw clear conclusions. Even if the trends in the corresponding section look clear, the variability is too high to be confident about the results.

Snow is the second most significant factor, with a relative increase of nearly 60% in both directions. The corresponding confidence intervals are also relatively large, but its impact is unquestionable.

In the third place, rain provokes a relative increase of delays of around 45%. In this case, the confidence intervals are relatively small. This is probably the most interesting variable since it is the most common phenomenon (50% of days), while the other phenomena (snow, strong wind, extreme temperatures) occur in less than 10% of days.

Finally, temperature is the least significant factor, with a relative increase of less than 40% in both directions. Note that the increase during hot days is much higher than during cold days. However, as in the case of wind, the confidence intervals are too large to confidently draw conclusions.

This aligns partially with previous studies (see section 1.2), as rain and snow are indeed found to be significant factors, but wind and temperature are not. However, one must note that the previous studies were based in other countries (mostly Northern European countries), where weather conditions and railway systems probably have different characteristics.

Variable		Abs. increase (%)	Rel. increase (%)
Rain		6.2 ± 1.0	43.8 ± 7.1
Snow		8.9 ± 4.0	58.9 ± 26.3
Temperature	Cold	2.0 ± 8.1	11.9 ± 48.6
	Hot	5.5 ± 5.8	33.1 ± 34.7
Wind		16.3 ± 28.5	100.5 ± 175.6

Table 3.1: Absolute and relative delay increase for each weather variable, for all direct trains

Variable		Abs. increase (%)	Rel. increase (%)
Rain		4.9 ± 1.4	40.2 ± 11.1
Snow		7.7 ± 5.2	60.3 ± 41.1
Temperature	Cold	1.5 ± 10.8	10.3 ± 76.5
	Hot	3.9 ± 7.4	27.7 ± 52.2
Wind		8.3 ± 37.7	59.5 ± 270.3

Table 3.2: Absolute and relative delay increase for each weather variable, for direct trains going from Bern to Zurich

Variable		Abs. increase (%)	Rel. increase (%)
Rain		7.5 ± 1.5	46.3 ± 9.3
Snow		10.0 ± 5.9	58.1 ± 34.0
Temperature	Cold	2.4 ± 11.8	12.7 ± 62.4
	Hot	7.6 ± 8.9	39.9 ± 46.9
Wind		22.6 ± 40.3	123.4 ± 220.0

Table 3.3: Absolute and relative delay increase for each weather variable, for direct trains going from Zurich to Bern

3.6 Clustering

Weather variables are not independent of each other. As an intuitive example, rain is negatively correlated with sunshine but positively correlated with humidity, and they might interact with each other. Therefore, another valuable analysis is to check how different kinds of days (e.g., rainy, windy, or sunny days) affect train delays instead of checking variable per variable.

To do so, I run different clustering models to find groups of days with similar weather conditions. I used all the available daily weather parameters, including but not limited to the ones analyzed before. Figure 3.30 shows the relation between the measurements in Bern and Zurich. All of them are highly correlated (more than 0.8) except for rain and snow. I used the mean of both cities for the former, while I kept both the measurements from Zurich and Bern for the two latter.

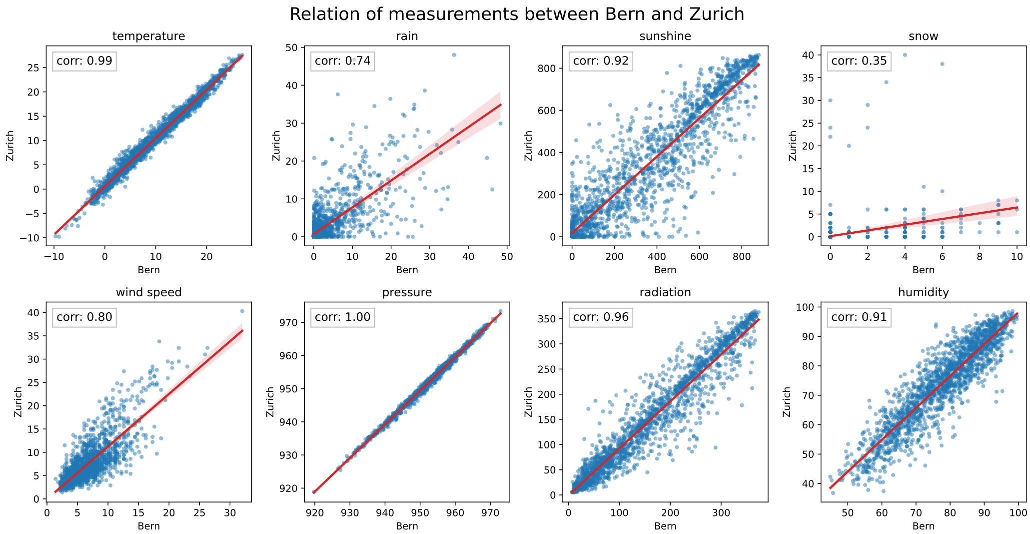


Figure 3.30: Correlation between the measurements in Bern and Zurich.

I tried two different clustering methods: KMeans and a Gaussian Mixture Model. Regardless of the clustering method, both the Silhouette score and the Elbow method indicated an optimal number of 3 clusters. Then, I ran different dimensionality reduction models to observe the clusters in 3D: PCA, MCA, and Factor Analysis of Mixed Data (FAMD). The latter (which is the most suitable for mixed data) is shown in Figure 3.31. In practice, all clustering methods resulted in the same 3 clusters, with just a few differences (<10). Figure 3.32 shows the size of each cluster.

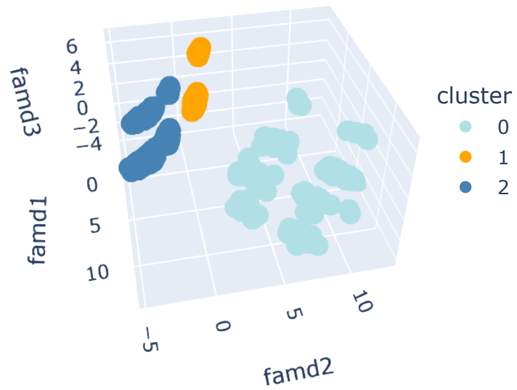


Figure 3.31: 3D representation of the clusters obtained with FAMD.

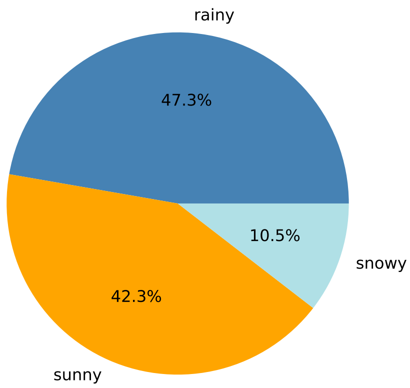


Figure 3.32: Size of each cluster.

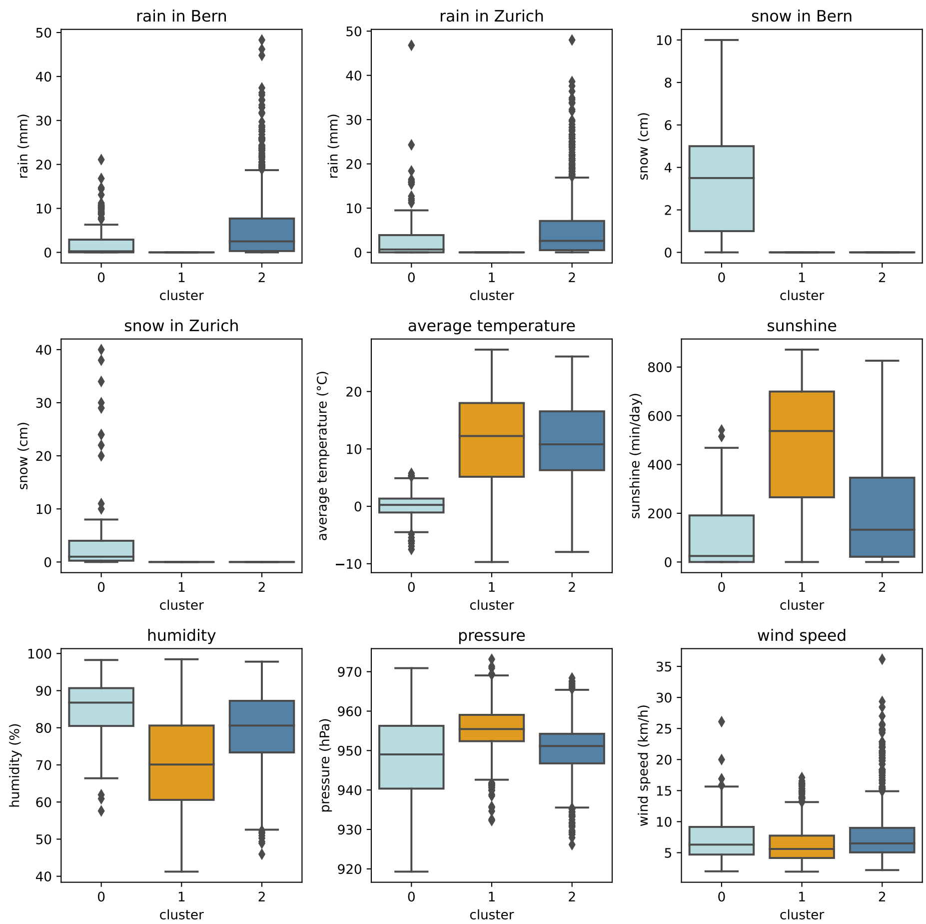


Figure 3.33: Overview of the weather in each cluster.

Figure 3.33 shows a boxplot for each cluster and each weather parameter to characterize each cluster. They can be described as follows:

- Cluster 0 includes cold, snowy days with no sunshine or radiation, very low temperatures, high humidity, and some rain.
- Cluster 1 includes sunny hot days with no rain or snow, high temperatures and pressure, low humidity, and low wind speed.
- Cluster 2 includes rainy (but non-snowy) days, with mid-high temperatures, medium humidity, and high wind speed.

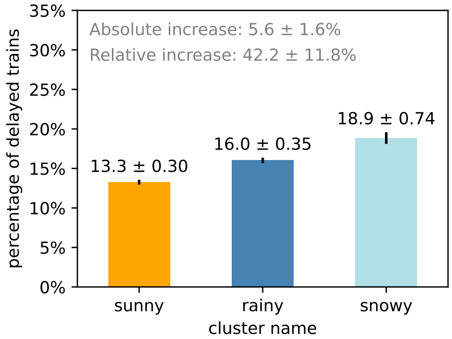
For simplicity, from now on I call them *snowy*, *sunny*, and *rainy*, respectively.

Figure 3.34 shows the average percentage of delayed trains for each cluster. As expected from the previous results, snowy days have more delays than rainy days, which in turn have more delays than sunny days. The same pattern as before can be observed: direct trains are more affected than indirect ones, especially if they travel from Zurich to Bern.

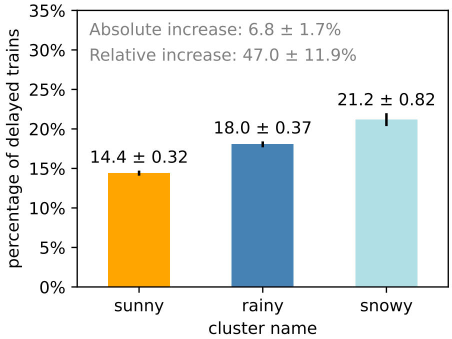
Figure 3.35 shows the average percentage of delayed trains in each location of the journey for each cluster. The conclusions are the same as for the previous sections: delays are higher in all journey locations when it snows or rains, and trains get delayed during the travel time while they catch up at the stations.

As for major delays, Figure 3.36 shows the percentage of trains with a delay greater than a certain threshold for each cluster. Again, the percentage of delayed trains is always higher when it snows or rains, although there is no significant difference between clusters from 10 minutes onwards.

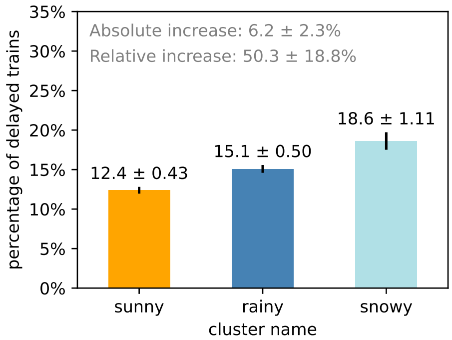
Finally, Figure 3.37 shows that the impact of rain and snow remains throughout the whole year. However, note that snowy days have an especially high impact in months where snow is not common (October and March). The same effect can be observed in all stops and all train lines, but the corresponding plots are not shown for brevity.



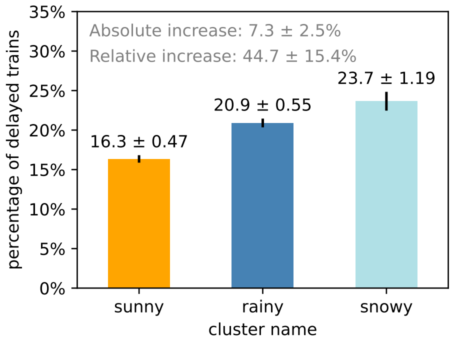
(a) Indirect trains, arrival at dest.



(b) Direct trains, arrival at dest.



(c) Direct trains from Bern to Zurich, arrival at destination



(d) Direct trains from Zurich to Bern, arrival at destination

Figure 3.34: Percentage of delayed trains for each cluster, for different subsets of data.

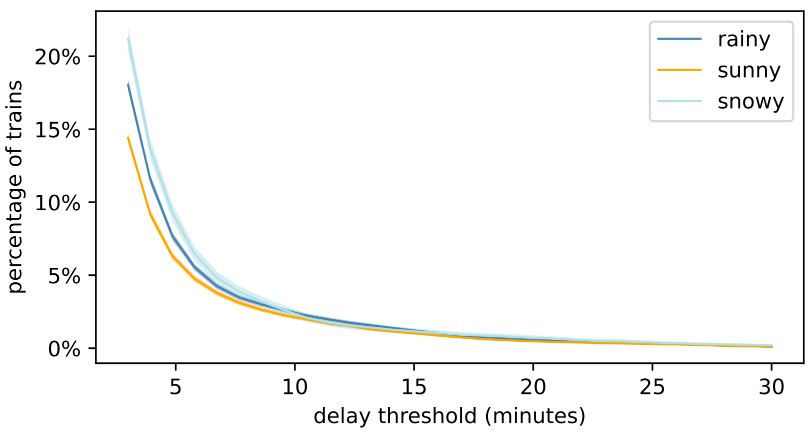
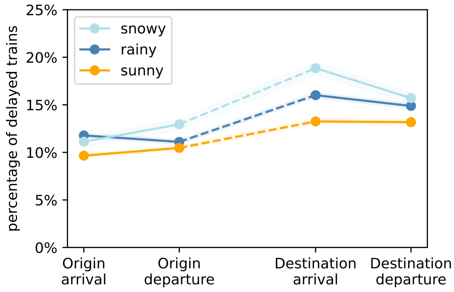
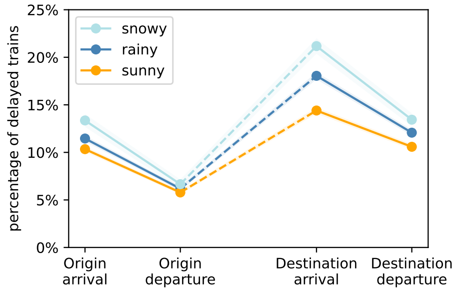


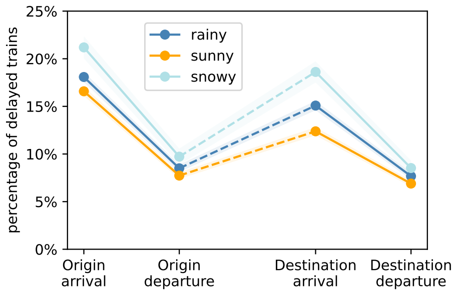
Figure 3.36: Percentage of trains with a delay greater than a certain threshold for each cluster.



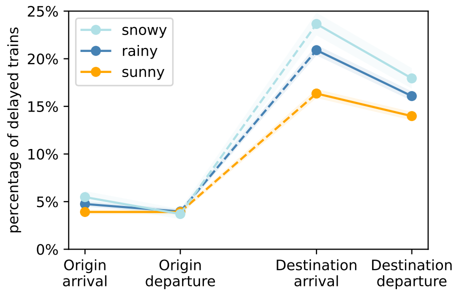
(a) Indirect trains



(b) Direct trains



(c) Direct trains from Bern to Zurich



(d) Direct trains from Zurich to Bern

Figure 3.35: Percentage of delayed trains along the journey for each cluster, for different subsets of data.

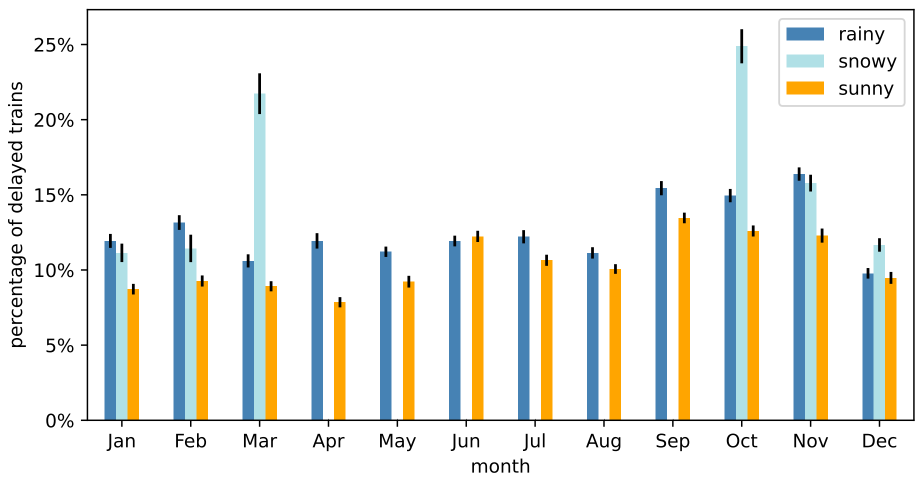
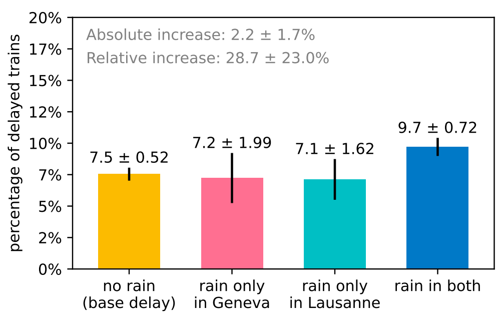


Figure 3.37: Percentage of delayed trains per month, for each cluster.

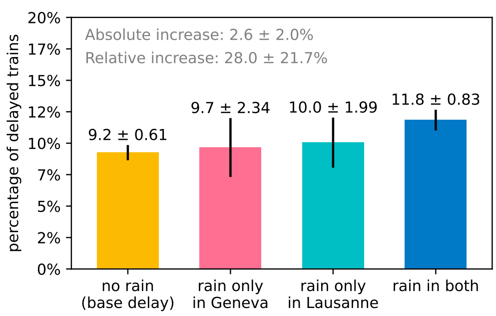
3.7 Effect on other railway sections

All the previous analysis was done on a single section of the railway (Bern-Zurich). One may wonder if these results and conclusions also apply to other sections or regions in Switzerland. With the same railway data used for the previous analysis, and retrieving new weather data, I checked if it could also be applied to another section of the same line: from Geneva to Lausanne.

Figure 3.38 show the percentage of delayed trains in each situation (no rain, rain only in Geneva, rain only in Lausanne, and rain in both cities), separated by direction. Note that this only includes IC1 trains, since they are the only ones that go through both cities and that were already included in the railway dataset. The conclusions are the same as in section 3.1: the percentage of delayed trains is higher when it rains in both cities than when it rains only in one, which in turn is higher than when it does not rain at all. Again, the trains going west (from Lausanne to Geneva) are more affected by rain than the other way around, just as in the Bern-Zurich section.



(a) From Geneva to Lausanne



(b) From Lausanne to Geneva

Figure 3.38: Percentage of delayed IC1 trains according to the location of rain, separated by direction.

The same results are obtained if this exercise is repeated for the Lausanne-Bern or the St.Gallen-Zurich section. It is interesting to see that, in all sections of this line, trains going west (and thus, generally, downwards) get more delayed on average when it rains. Although this is probably just a correlation and not a causation, a hypothesis could be that rain wettens the tracks, thus making it more difficult for the trains to break, thus making them go slower and ultimately get more delayed. However, this is just a hypothesis; more analysis outside this line or region would be needed to confirm it.

3.8 Cancellation of trains

Around 9% of trains are canceled either before departure or during their journey, according to the used data. These trains were not included in the preceding analysis, as their delay cannot be computed. However, it is also interesting to know if train cancellation is affected by weather conditions.

Figure 3.39 shows the percentage of canceled trains in each cluster presented in section 3.6. Surprisingly, fewer trains were canceled when it snowed, while there was the same number of cancellations on rainy and sunny days. This could be explained by the fact that many cancellations are actually scheduled, e.g., for maintenance reasons, and thus not affected by weather. Another possible explanation could be that SBB may prepare the network better before snowy days, knowing in advance that more issues may occur. Also, keep in mind that snowy days represent a relatively small proportion of days (see Figure 3.32).

The cancellation of trains is not affected by rain, contrary to the effect on delays presented in the previous sections. This could be explained by the fact that SBB non-scheduled cancellations usually occur because of malfunctioning, human accidents, or other major issues that are not directly related to weather, at least in the short term. However, it is important to note that the data used for this analysis does not include the reason for cancellation, so this is just a hypothesis.

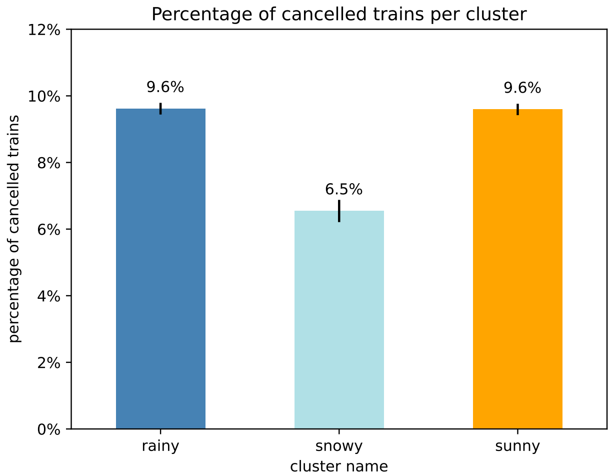


Figure 3.39: Percentage of canceled trains per day according to weather.

Chapter 4

Prediction

This chapter describes the efforts to predict train delays using meteorological data. First, it is worth mentioning that predicting the delay of a train is a very complex task. A myriad of factors can have an influence, and meteorology is only one of them. Many efforts have been put in both within and outside of SBB to predict delays, but it remains challenging even if many other factors are considered. Therefore, these models do not aim to predict the delay of a train with high accuracy, but to check if meteorological data can be used to improve delay prediction models.

Train-level prediction

Minutes of delay

First, some ambitious models were built to predict the minutes of delay for each train. I built three versions, changing the predictor variables: first, one only with weather variables; second, one only with previous delays (departure delay of the train, arrival delay of the previous train, and average delay in the same line during the past 6 hours); and third, one combining both sets of variables. I tried both a Random Forest Regressor and a Gradient Boosting Regressor. Table 4.1 shows the metrics obtained with each version and model.

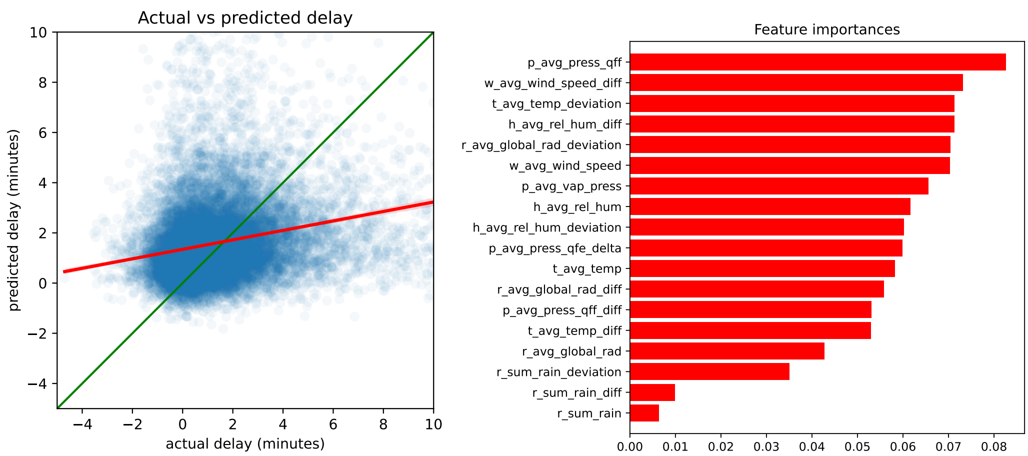
The best-performing model was Gradient Boosting when including both weather and delay variables, but it was still poor. In the best case, it managed to explain 40% of the variability of the data.

Model	Metric	W	D	D&W
Random Forest	MAE	1.81	1.55	1.48
	RMSE	3.00	2.46	2.37
	R2	0.03	0.31	0.36
Gradient Boosting	MAE	1.67	1.44	1.41
	RMSE	2.89	2.35	2.32
	R2	0.05	0.37	0.40

Table 4.1: Metrics for the prediction of minutes of delay

Note how the models that only include weather variables are very poor, and R2 does not exceed 0.05. However, weather features can indeed improve the model's performance if the previous delays are already used. Also, one must take into consideration that the variables concerning previous delays might already include information about meteorological conditions, so the weather variables might not be adding their full information in the last version of the models.

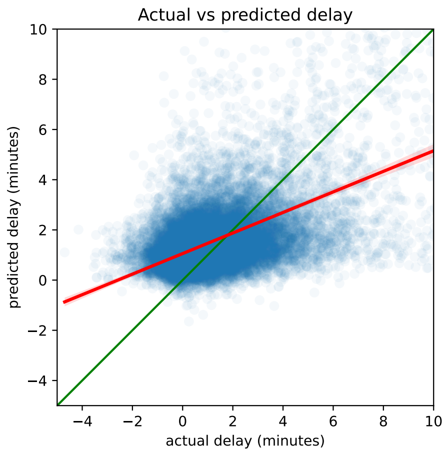
Figure 4.1 shows the results for the best model with only weather. Figure 4.2 shows the results for the best model with weather and delays.



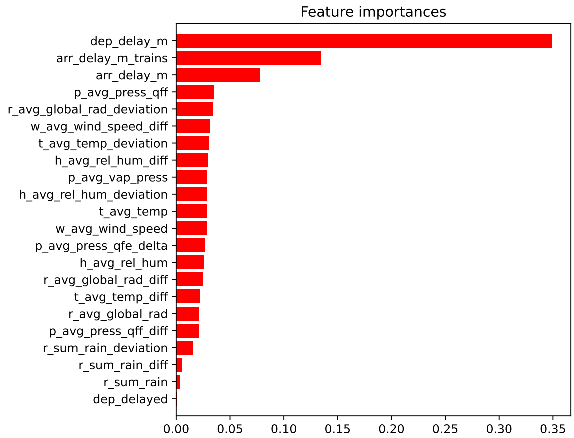
(a) Actual vs predicted delay (b) Feature importance

Figure 4.1: Results with only weather variables

Figure 4.1b shows that the most important variables are pressure, wind, and temperature if only weather variables are considered. However, note that the model is very poor so these results should be taken with caution.



(a) Actual vs predicted delay



(b) Feature importance

Figure 4.2: Results with both weather and previous delays

Figure 4.2b shows that, when previous delays are included, the most important variables are the previous delays, as expected. The weather variables have a similar (low) importance, but pressure, radiation, wind, and temperature are the best ones.

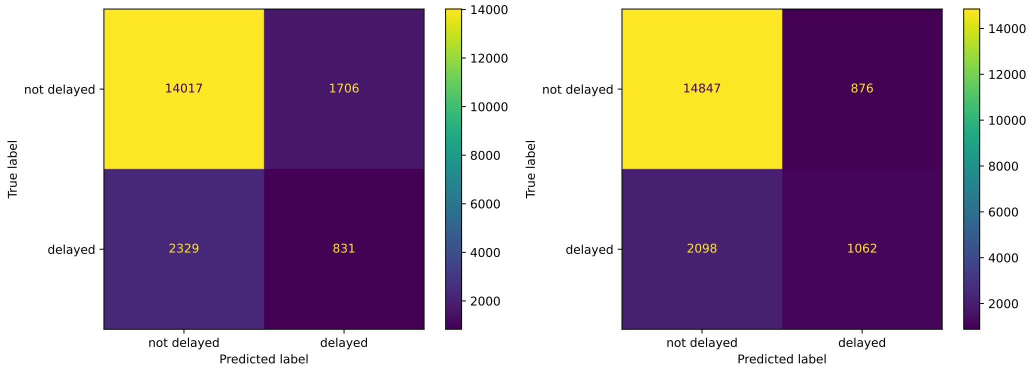
Delayed or not

In addition, I changed the target of the models to predict whether the train would be delayed or not. The metrics are shown in Table 4.2. The models suffered from a low recall, meaning that they were able to predict the non-delays but not the delays, even if the model was weighted to account for the imbalance in the target.

Model	**Metric**	**W**	**D**	**D&W**
Random Forest	Accuracy	0.78	0.83	0.84
	Precision	0.33	0.52	0.55
	Recall	0.26	0.30	0.34
	F1	0.29	0.38	0.42
Gradient Boosting	Accuracy	0.83	0.85	0.86
	Precision	0.44	0.74	0.72
	Recall	0.04	0.25	0.27
	F1	0.07	0.37	0.39

Table 4.2: Metrics for the prediction of whether a train will be delayed or not

Again, including both sets of features provided the best results. Both types of models performed similarly, but Random Forests gave a higher overall F1 score. The confusion matrices are shown in Figure 4.3.



(a) Only weather varia

(b) Weather and previous delays data

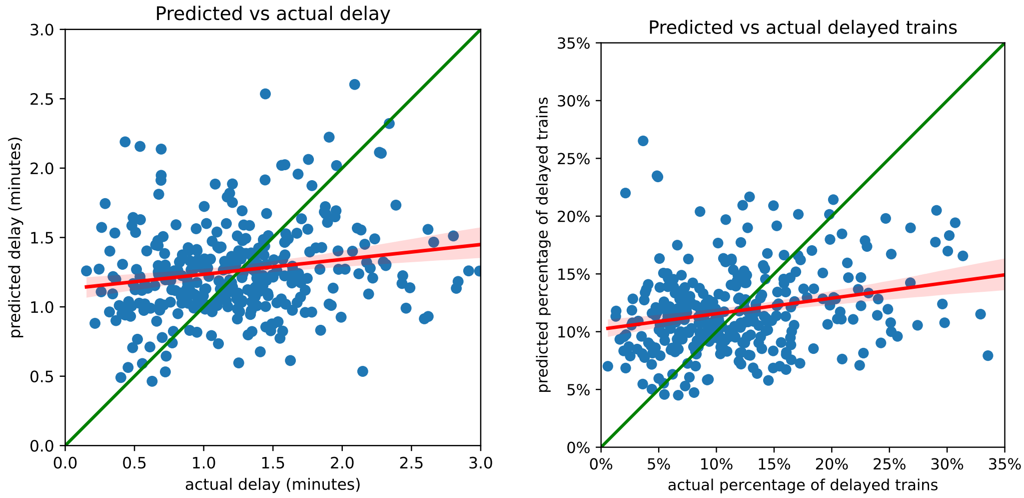
Figure 4.3: Confusion matrix for delayed trains

A Neural Network and a Long short-term memory (LSTM) were also tried, but they did not improve the results of the other models. Therefore, they were discarded due to their complexity.

Day-level prediction

Since predicting the delay for each train is a very challenging task, I simplified the target by aggregating the delays and predicting the line's average delay (or the percentage of delayed trains) during the day. In this case, I only used the daily weather prediction. After removing weather and delay outliers, several types of models were tried (Linear Regression, Random Forests Regressor, and Gradient Boosting Regressor), but the results were the same: the models were incapable of adequately predicting the delays, as one would expect. Even if all possible combinations of daily weather parameters (and their derived variables) were tried, the R2 score did not exceed 0.10. The relation between actual and predicted values is shown in Figure 4.4, both for the percentage of delayed trains and the average delay. A slightly increasing trend is indeed found if a regression line is fitted, but the model is not able to capture the variability of the data.

Doing the same on the hour level was discarded, since there are not enough trains per hour in this section so as to aggregate them in significant groups.



(a) Daily average delay (minutes)

(b) Daily percentage of delayed trains

Figure 4.4: Prediction of delays with daily weather data

The conclusion is that simplifying the target does not help in predicting the delays. However, note that this model only includes weather data.

Conclusions

These results confirm that predicting delays is a challenging task, no matter the form of the target. The obtained models are very poor even if data from previous delays (including the departure delay of the train) are included. In fact, as shown in Figure A.8, a low correlation exists between a train's origin departure and destination arrival delay. In other words, not even by knowing the train's delay at a station's departure, it is easy to predict its delay upon arrival.

Reducing the data to only direct trains, a single direction, a single line, only during weekdays, or during a single month or season did not improve the results significantly. Even if the preceding analyses were clear about meteorological factors' impact on delays, there is too much variability to accurately use them as predictors. However, meteorology features could be used to improve the prediction if other factors are considered in more complex models. This is out of the scope of this thesis and should be explored in future work.

Chapter 5

Conclusion

This thesis aimed to investigate the impact of meteorological phenomena on train delays in Switzerland. The analyses provide compelling evidence of a correlation between adverse weather conditions and train delays. Assuming the causal graphs and remarks presented in the introduction, it can be concluded that there is a causal relationship between weather and delays.

Snow emerges as the most significant contributor to delays. However, rain can be of greater interest, as it is a much more common phenomenon and also has a considerable effect. Extreme temperatures and strong wind show a positive correlation with delays and, on average, a negative effect on punctuality, but their high variability hinders us from establishing a statistically significant effect. These findings emphasize the importance of considering adverse weather conditions when assessing and managing train operations.

The strength of the results is reinforced by the extensive datasets utilized, spanning five years and encompassing a substantial number of train journeys and weather measurements. As an added value, it should be noted that all the work has been done exclusively with open data. The detailed granularity of the weather data, both daily and hourly, aligns with previous studies and effectively captures the impact of meteorological phenomena on train delays.

The negative impact of adverse weather conditions is observed across all train stops, lines, and directions, with notable delays for trains traveling from Zurich to Bern and, generally, westward along the St. Gallen-Lausanne line. Direct trains are more susceptible to delays due to their limited opportunities for recovery.

Weather-related disruptions persist throughout the year, with significant effects observed during the fall season, peak hours, and evenings. Delays of all magnitudes, ranging from 3 to 30 minutes, are increased. Moreover, the study reveals that the influence of recent rain on train delays can extend up to 6 hours. In other words, it takes approximately 6 hours for the Swiss network to recover from rain-related disruptions and go back to normality.

Predicting train delays remains a complex challenge, and solely relying on weather data for predictions yields limited accuracy, as expected. The best model, which included both weather and previous delay data, was able to explain only 40% of the variance in the delays. However, incorporating meteorology-related factors on more sophisticated models may enhance predictive capabilities, an avenue worth exploring in future research.

A notable contribution of this thesis is the development of clustering techniques that allow the classification of days into distinct categories based on weather predictions. This clustering approach serves as a valuable tool for the punctuality team at SBB, enabling proactive preparations for days expected to experience adverse weather conditions.

Furthermore, the outcomes obtained from examining other railway segments within the same line offer promising indications of the generalizability of the findings across different regions. This work sets the methods and tools to extend this analysis in future research.

Finally, the study found no evidence that weather conditions directly affect train cancellations. This aligns with SBB's expectations, as cancellations often result from scheduled adjustments or technical issues unrelated to weather conditions. Additionally, the analysis did not uncover a significant impact of consecutive rainy days on delays compared to isolated days.

The findings of the study align with previous research. However, it provides the first empirical confirmation of these effects in Switzerland, shedding light on the unique characteristics of the Swiss railway network.

In conclusion, this thesis contributes to understanding how meteorological phenomena influence train delays in Switzerland. The demonstrated correlations, impact quantification, and clustering techniques offer valuable insights to inform decision-making and improve punctuality in the face of adverse weather conditions. By building upon these findings, SBB and other transport operators can develop more robust strategies for reducing weather-related disruptions and improving their networks' efficiency and reliability.

Chapter 6

Limitations and future work

This thesis has provided valuable into the relationship between meteorological conditions and disruptions in the railway system. However, it is important to acknowledge the limitations of this study and identify potential areas for future research.

To begin with, future work could extend the analysis to include other Swiss lines or regions. By broadening the scope, a more comprehensive understanding of how weather influences train delays in Switzerland could be achieved. Examining different lines or regions would allow for a comparison of the effects of weather on diverse railway networks, accounting for variations in geographical and climatic factors. The present thesis is a valuable foundation for such extensions, as the code and methodologies employed can be readily replicated or adapted to incorporate new data.

Moreover, studying the interaction of trains, including accumulated delays and other complex factors, could provide valuable insights into the dynamics of disruptions in the railway system. Although this thesis focused on the direct impact of weather on train delays, exploring the broader intricacies of delays and their propagation throughout the network could enhance our understanding of the overall system's resilience and vulnerability.

One limitation of this study was the absence of data regarding the number of passengers on each train, as it is not open. Including this information in future research would enable a more comprehensive investigation into the connection between weather-induced delays and their implications for commuters. Exploring the interaction between passenger volume and weather-related disruptions can provide insights for developing strategies to minimize the impact on individuals and improve overall system performance.

Similarly, future studies could aim to include data on incidents or construction work that may affect train operations. Although the acquisition and analysis of this data present challenges, exploring its influence on train delays could contribute to a more comprehensive understanding of the factors contributing to disruptions in the railway system. While this study assumed a uniform impact of incidents or construction work on all trains, future research could delve into the specific ways in which these events interact with weather conditions and affect train operations.

In addition, investigating the impact of weather on other modes of transport, such as buses or cars, could provide valuable comparative insights. Understanding how weather conditions influence various transportation systems would help identify common challenges and develop strategies for improving overall transportation resilience in the face of adverse weather events.

Finally, it is important to note that this thesis has focused on confirming and quantifying the negative effect of weather conditions on train delays. Further investigation is needed to understand the factors that connect weather conditions to train delays and gain insights into the underlying processes. For instance, one hypothesis is that rain prompts passengers to seek sheltered areas of the stations, making boarding and alighting from trains more challenging and resulting in longer stops. Also, rain can wet and slippery the tracks, potentially affecting braking. However, this study does not delve into these aspects due to a lack of technical knowledge. Gaining a deeper understanding of the mechanisms and causal pathways by which weather events lead to disruptions would inform the development of targeted strategies and interventions to mitigate the impact of adverse weather on train operations.

By addressing these limitations and pursuing the suggested avenues for future research, scholars, policymakers, and especially SBB and other public transport operators in Switzerland can leverage the findings of this study. This will enable them to develop measures for managing weather-related disruptions, enhancing the railway network reliability, and improving the commuting experience for passengers.

Bibliography

[1] Carl-William Palmqvist, Nils Olsson, and Lena Winslott-Hiselius. ?Some influencing factors for passenger train punctuality in Sweden? In: *International Journal of Prognostics and Health Management* 8 (Oct. 2017), p. 13. DOI: 10.36001/ijphm.2017.v8i3.2649.

[2] Ghazal Zakeri and Nils Olsson. ?Investigating the effect of weather on punctuality of Norwegian railways: a case study of the Nordland Line? In: *Journal of Modern Transportation* 26 (June 2018). DOI: 10.1007/s40534-018-0169-7.

[3] William Brazil et al. ?Weather and rail delays: Analysis of metropolitan rail in Dublin? In: *Journal of Transport Geography* 59 (2017), pp. 69–76. ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2017.01.008. URL: https://www.sciencedirect.com/science/article/pii/S0966692316304409.

[4] Yuanni Xia et al. ?Railway infrastructure disturbances and train operator performance: The role of weather? In: *Transportation Research Part D: Transport and Environment* 18 (2013), pp. 97–102. ISSN: 1361-9209. DOI: https://doi.org/10.1016/j.trd.2012.09.008. URL: https://www.sciencedirect.com/science/article/pii/S1361920912001101.

[5] Pu Wang and Qing-peng Zhang. ?Train delay analysis and prediction based on big data fusion? In: *Transportation Safety and Environment* 1.1 (Feb. 2019), pp. 79–88. ISSN: 2631-4428. DOI: 10.1093/tse/tdy001. eprint: https://academic.oup.com/tse/article-pdf/1/1/79/29227827/tdy001.pdf. URL: https://doi.org/10.1093/tse/tdy001.

Appendix A

Data exploration

A.1 Railway data

General description

As explained in section 2.2, the railway dataset contained nearly 200.000 trains from Bern to Zurich (or vice versa). The same number of trains traveled in each direction. Around half of them were direct (no intermediate stops), while the other half stopped a variable amount of times between both cities. The trains stopped in around 775.000 stops (or train stations), giving an average of 3.9 stops per train (or 2.0 stops per direct train and 5.8 stops per non-direct train). Figure A.1 shows the distribution of the number of stops per train.

The trains are part of several lines, grouped into InterCity (IC) and InterRegio (IR) services. Note that each line stops at different stations with different frequencies, and trains of the same line may also stop in different places. All InterRegio services stop between Bern and Zurich, while most of the InterCity trains are direct. Figure A.2 shows the percentage of trains and stops for each line. Regional services stop more than the InterCity ones, even if they represent a smaller part of the trains.

As explained in section 1.3, not all trains start or end in Zurich or Bern. Instead, they may start in a previous stop and/or end in a posterior one (see example in Figure 1.3). Figure A.3 illustrates each possible combination and shows the corresponding percentage of trains.

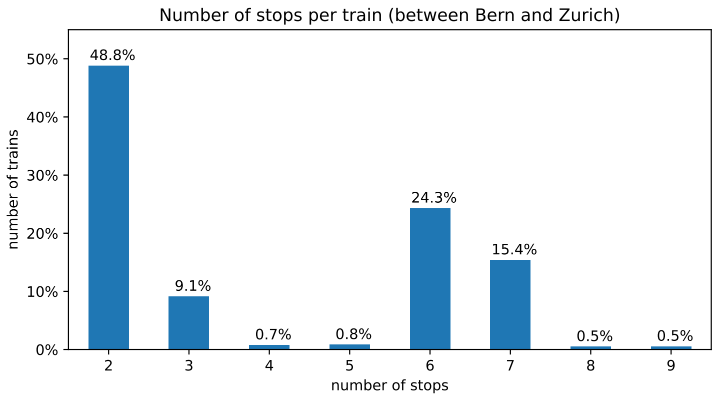


Figure A.1: Number of stops per train (between Bern and Zurich)

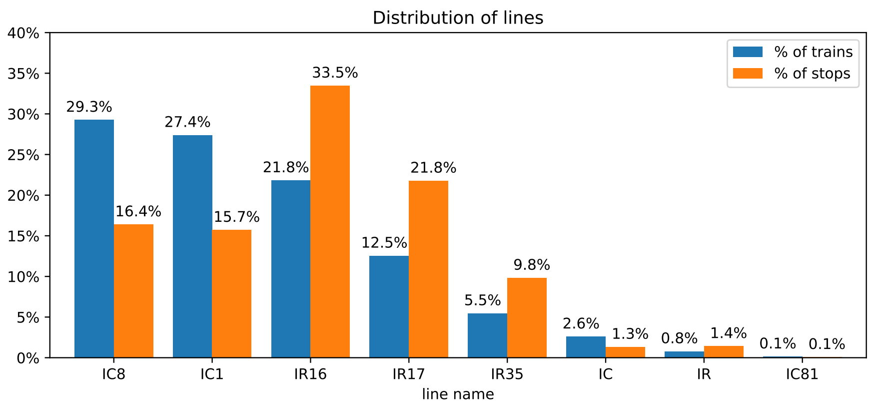


Figure A.2: Distribution of lines

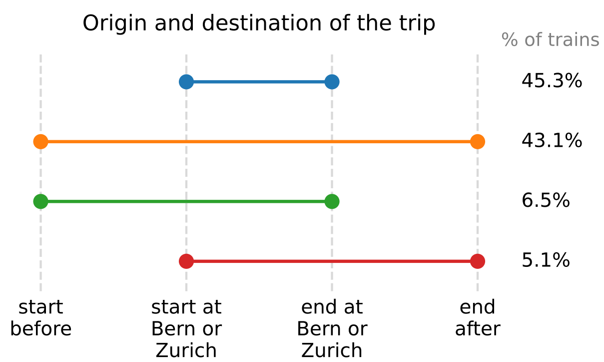


Figure A.3: Origin and destination of the trip

Regarding the time span, the trains were differently distributed over time. Although the considered period was from 2018 to 2022 included (5 years or 1.825 days), some dates were missing (see Figure A.4), leaving around 1700 days with data. This is due to two reasons: first, possible changes in the collection or annotation of the data over time, and second, changes in the train schedule (including the number of trains) since the SBB timetables change every year. However, this is not critical for this project since there is enough data distributed over different years, stations and months so as to be able to capture the difference of the effect over these periods.

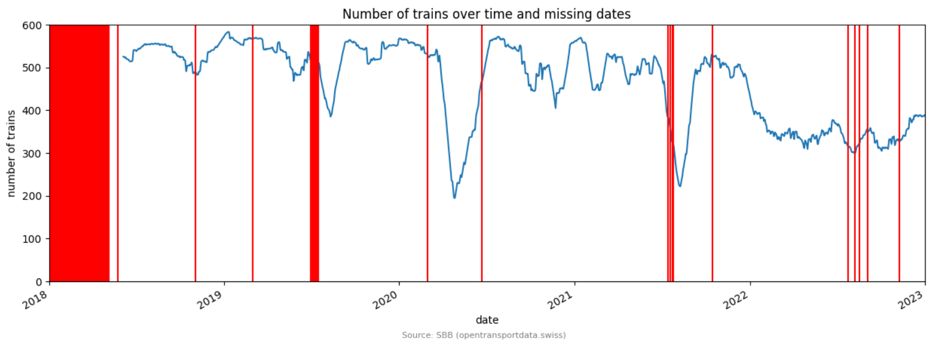


Figure A.4: Number of trains per line and missing dates

Also, note that the number of trains oscillates during the day and that there are no night trains. Most travel from 6h to 19h (see Figure A.5).

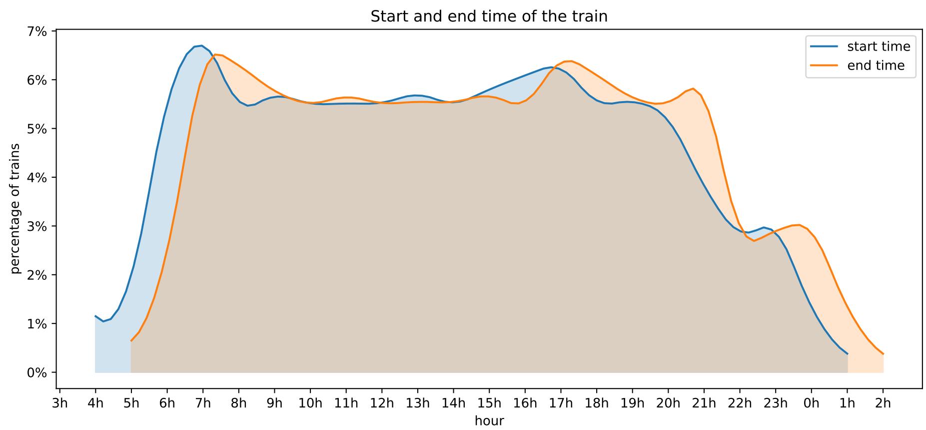


Figure A.5: Distribution of the start and end hours of the trains

Departure and arrival delays

Regarding the delays, first note that one can compute the ***departure delay*** (that is, the difference between the *target departure time* and the *actual departure time* from a station), and the ***arrival delay*** (that is, the difference between the *target arrival time* and the *actual arrival time* to a station). Both delays can be computed at each stop, but we usually focus on the *origin departure delay* and the *destination arrival delay*.

Figure A.6 shows a Kernel Density Estimate plot of the origin departure and destination arrival delays. One can see that the departure delay is rarely negative (trains do not depart in advance) and is usually very low (under the 3-minute threshold). As for the arrival delay, it is much more widespread, meaning that trains sometimes arrive in advance, but also arrive quite later than 3 minutes. Indeed, the median arrival delay is much lower and the standard deviation much higher than those of the departure delay, even if the mean is nearly identical.

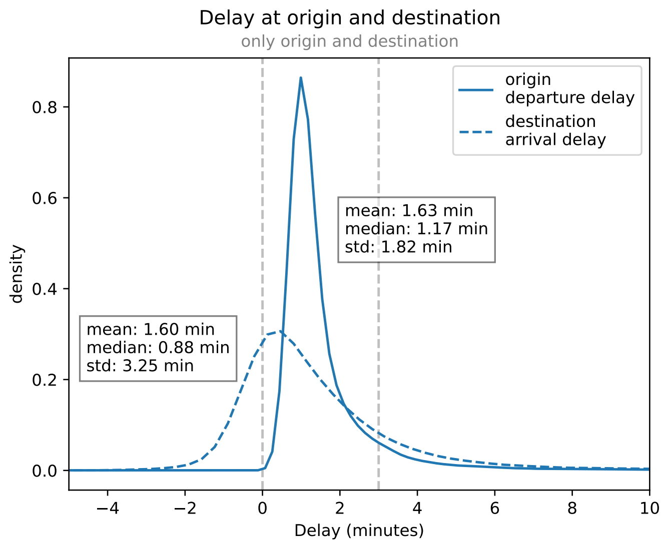


Figure A.6: Origin departure and destination arrival delays

Figure A.7 shows the same but separated by direction. The departure delay is lower for trains traveling from Zurich to Bern, but the arrival delay is slightly higher, as confirmed by the corresponding t-tests.

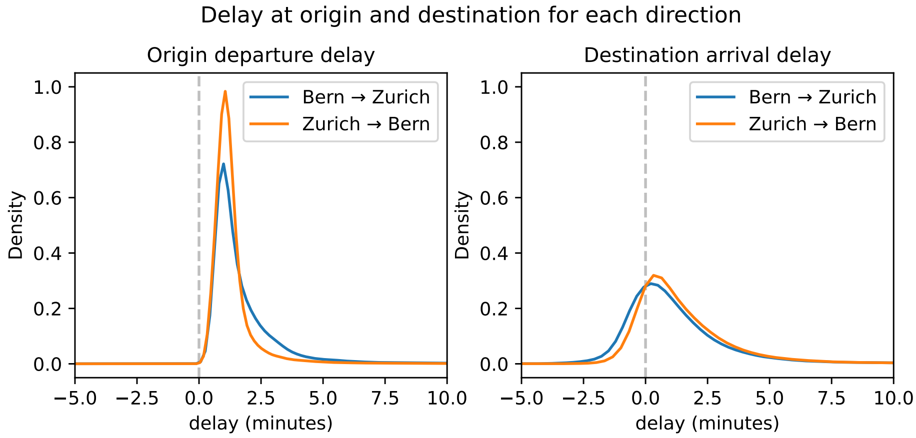


Figure A.7: Origin departure and destination arrival delays, by direction

Figure A.8 shows the correlation between arrival and departure delays, at the origin, during the trip, and destination. Arrival and departure delays at origin and destination are highly correlated and approach a diagonal shape. However, the regression lines are lower than the corresponding diagonals, meaning that the departure delay tends to be lower than the arrival one. In other words, trains catch up by stopping less than scheduled. The correlation during the journey, instead, is much lower (below 0.5), meaning that these delays are very unpredictable, even given the delay at departure.

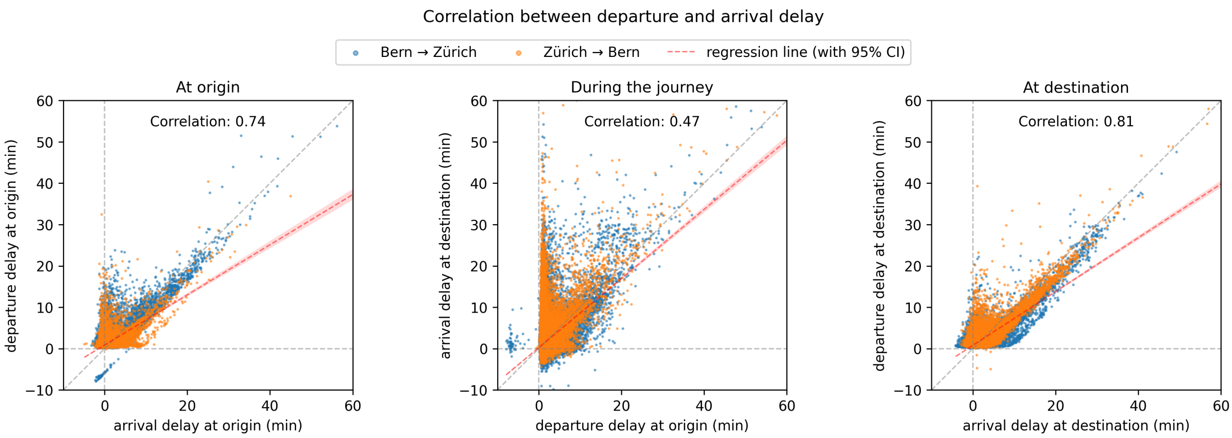


Figure A.8: Correlation between arrival and departure delays

Table A.1 shows the difference between directions, also illustrated in Figure A.8. Although the trend is similar, trains traveling from Zurich to Bern have a lower correlation in all cases, meaning they get delayed more easily.

Direction	At origin	During the journey	At destination
Bern-Zurich	0.78	0.51	0.83
Zurich-Bern	0.61	0.43	0.79

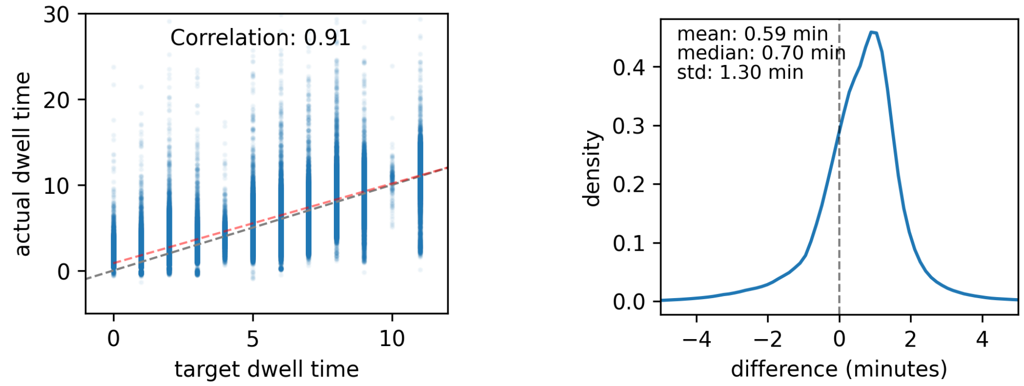
Table A.1: Correlation between arrival and departure delays

Dwell time and travel time

Apart from arrival and departure delays, it may also be interesting to check the elapsed time at a station (*dwell time*) and the elapsed time between two stations (*travel time*). This way, one can check if a train incurs a delay while stopped at a station or during the journey, regardless of its previous delay.

Figure A.9a shows the relation between the *target dwell time* and the *actual dwell time* in a station. The regression line is shown in red, and the perfect linear relation is shown in gray. They are highly correlated, but the actual dwell time is slightly higher than scheduled.

Figure A.9b shows a Kernel Density Estimate plot of the *dwell time difference*. That is the difference between the actual dwell time and the target dwell time of the train in a given station. Again, one can observe that trains tend to stop more than scheduled (both median and mean are positive), although they sometimes stop less than expected, probably to catch up.

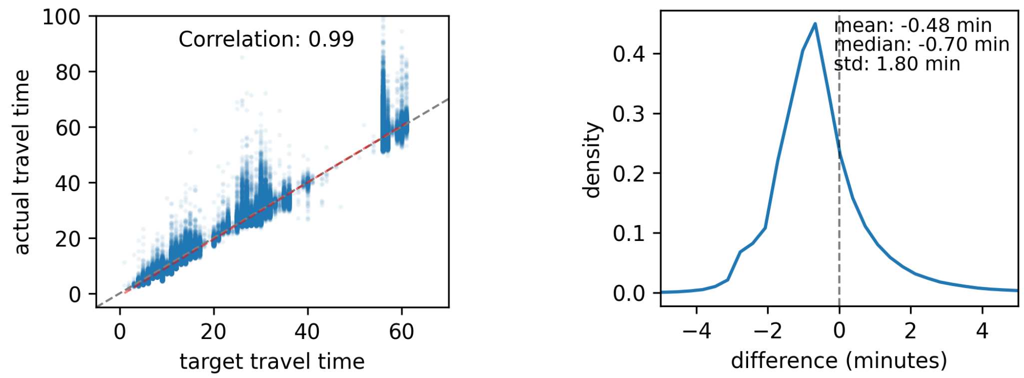


(a) Target vs actual dwell time

(b) Dwell time difference KDE

Figure A.9: Dwell time analysis

Similarly, one can produce the same plots for the ***travel time***. Figure A.10a shows the relation between the *target travel time* and the *actual* *travel time*. Figure A.10b shows a Kernel Density Estimate plot of the *travel time difference* (the difference between the actual travel time and the target travel time). Both show that the actual travel time tends to be lower than the target one, but they have an almost perfect linear relationship.

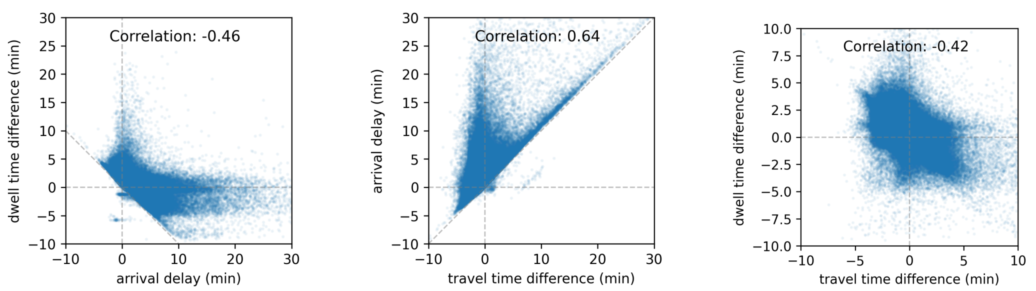


(a) Target vs actual travel time

(b) Travel time difference KDE

Figure A.10: Travel time analysis

Finally, Figure A.11 shows the relationships between the arrival delay, dwell time difference, and travel time difference. Figure A.11a shows that the arrival delay and the dwell time difference are negatively correlated, meaning that trains stop for a shorter time when they are delayed to catch up and be on time again. Figure A.11b, instead, shows that the arrival delay and the travel time difference are positively correlated (delay during the journey provokes an arrival delay at the next stop). Finally, Figure A.11c shows that travel and dwell time differences are negatively correlated, meaning again that trains stop for less time when they get delayed during the journey.



(a) Arrival delay vs dwell time difference

(b) Arrival delay vs travel time difference

(c) Dwell time difference vs travel time difference

Figure A.11: Relation between delays and dwell/travel time differences

Stops

Figure A.12 shows the number of occurrences per stop. All trains stop in Bern and Zurich, half in Olten, and around a quarter in Aarau and other intermediate stops. Very few stop in Lenzburg, Wynigen, Turgi and Dietikon. Figure A.14 shows the average arrival and departure delay for one of them, by direction (for stops in both directions only). Especially noticeable is how Bern has a much higher delay than Zurich, no matter the direction. Finally, Figure A.13 shows the average arrival delay according to the number of stops between Bern and Zurich. There is not a clear increasing or decreasing pattern, but the most delayed trains are those that stop between 3 and 5 times. See Figure A.1 for the distribution of the number of stops per train.

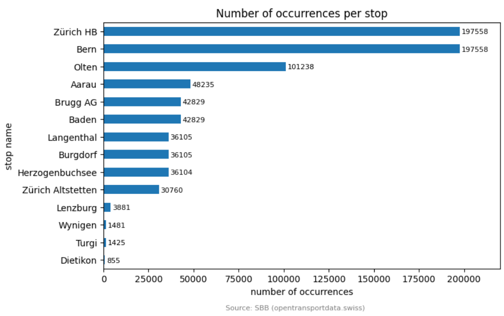


Figure A.12: Number of occurrences per stop

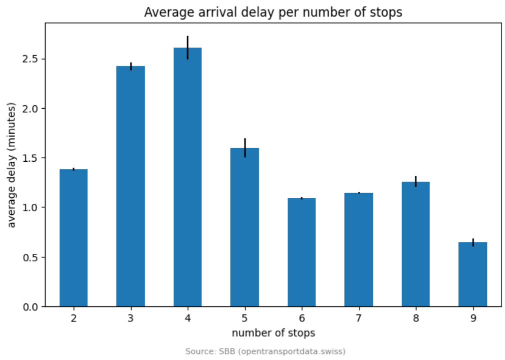


Figure A.13: Average arrival delay per number of stops

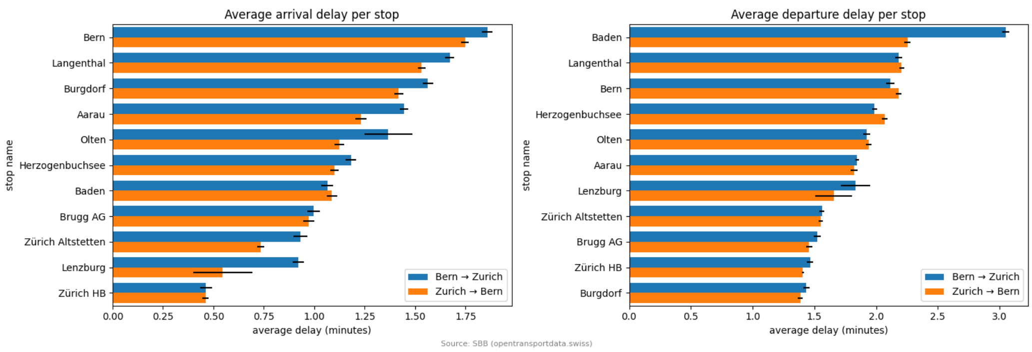


Figure A.14: Average arrival and departure delay per stop, by direction

A.2 Meteorological data

As mentioned, the analysis is focused on four weather parameters: rain, snow, wind, and temperature. This section shows a quick overview of each one to understand, for example, how often it rains or snows in Bern and Zurich or how often there are extreme temperatures or strong wind gusts.

Rain

Rain is a very common weather phenomenon in Switzerland. During the considered period, it rained once every two days in at least one of the cities, and almost 40% of days in both (see Figure A.15). In 88% of days, both cities coincided (either it rained in both or did not rain at all). In fact, the rainfall measurements in both cities had a 73.65% correlation (see Figure A.17). As one would expect, rainfall had a right-tailed distribution (see Figure A.16).

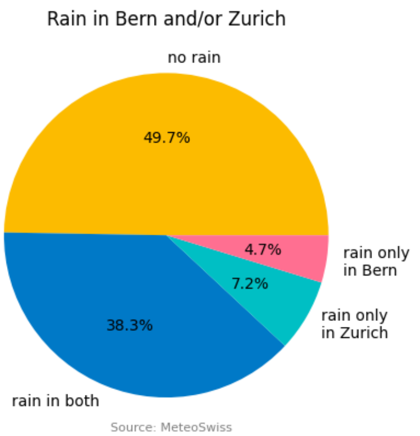


Figure A.15: Distribution of rain locations

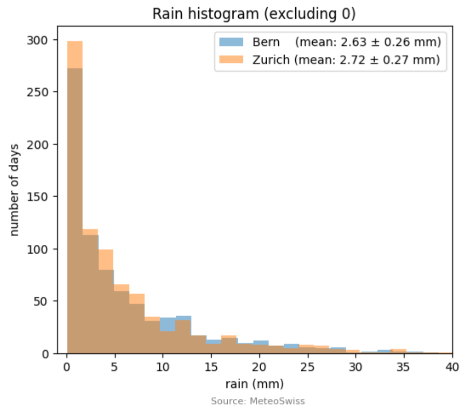


Figure A.16: Rainfall histogram, per city

Figure A.18 shows that the amount of rain is much higher when it rains in both cities than when it rains only in one. This can be explained by the fact that rain in both cities is probably related to worse generalized weather (therefore with more rain) in the whole country.

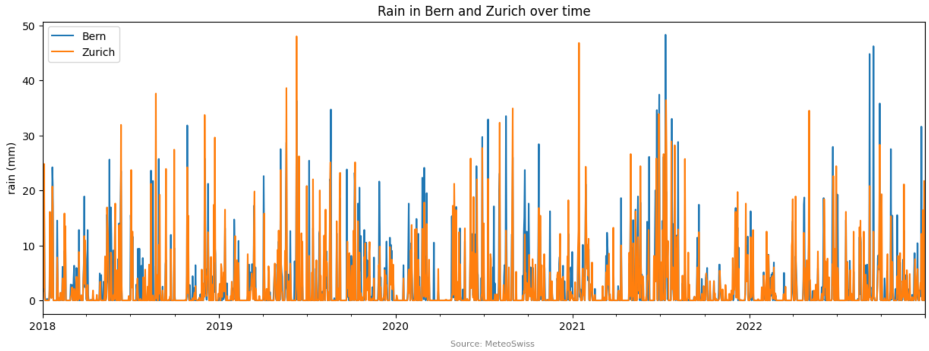


Figure A.17: Rainfall in Bern and Zurich

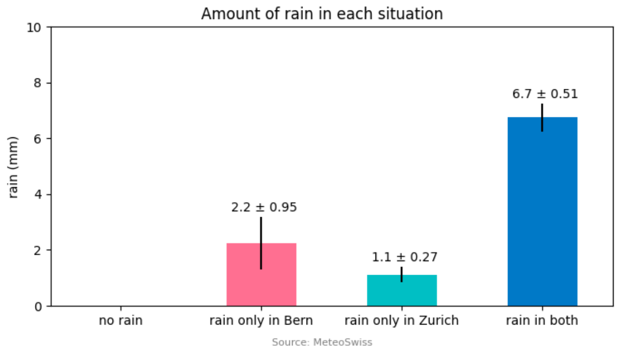


Figure A.18: Average amount of rain in each situation

Snow

As one can guess, snow is rare and only occurs in winter. In Bern and Zurich, snow is relatively common from December to February and exceptional in November, March, and April, as shown in Figure A.19. Over the last five years, it snowed 12% of the days from December to February. Figure A.20 shows how often it snowed in one or both cities.

Figure A.21 shows the right-tailed distribution of snowfall in both cities. Even if the mean in Bern is higher, Zurich suffered from some significant outliers (more than 20 cm of snow in one day). As in the case of rain, Figure A.22 shows that the amount of snow is much higher when it snows in both cities than when it snows only in one.

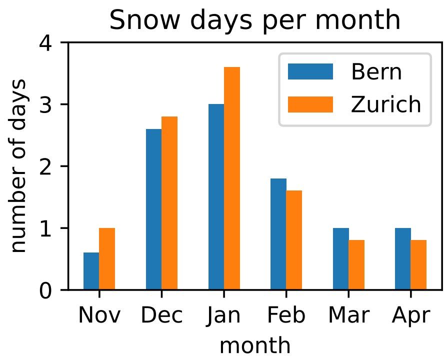


Figure A.19: Snow days per month

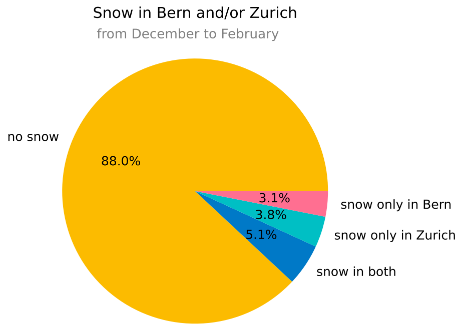


Figure A.20: Distribution of snow locations

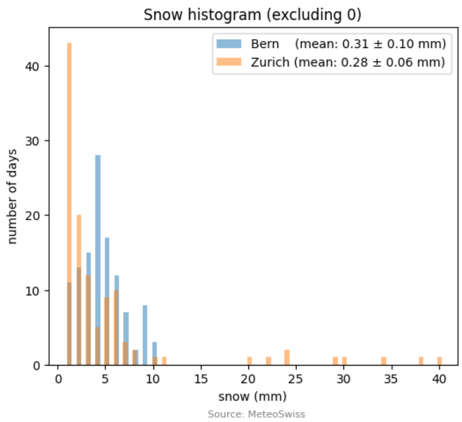


Figure A.21: Snowfall histogram, per city

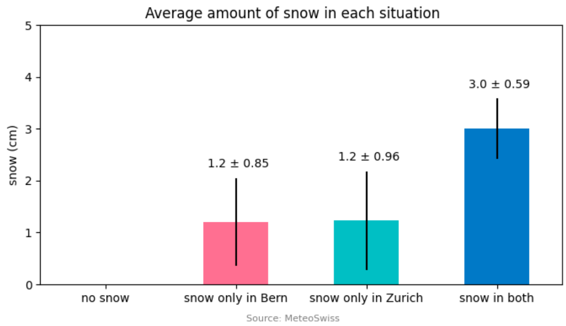


Figure A.22: Average snowfall in each situation

Temperature

As for temperature, Figure A.25 shows the expected seasonal behavior. The correlation between Bern and Zurich measurements was 0.99. Figure A.23 shows the histogram of temperatures in both cities. Daily average temperatures range from -10ºC to nearly 30ºC, being slightly higher in Zurich than in Bern. Temperatures below 0ºC and above 20ºC are relatively rare. The temperature difference on the same day (Bern - Zurich) is usually negative but can range from -5ºC to 4ºC (see Figure A.24).

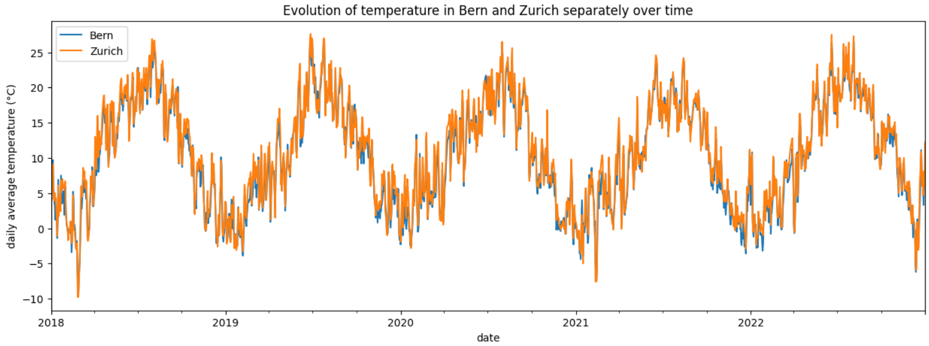


Figure A.25: Temperature evolution in Bern and Zurich

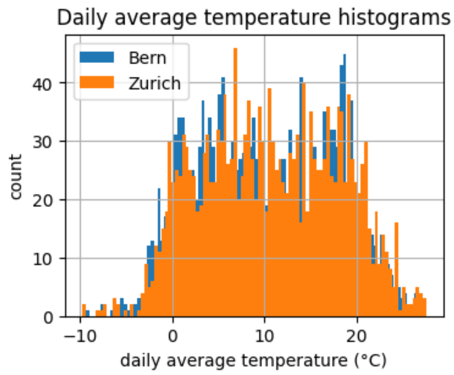


Figure A.23: Temperature histogram in Bern and Zurich

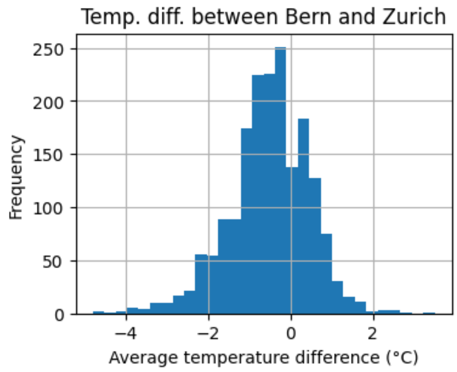
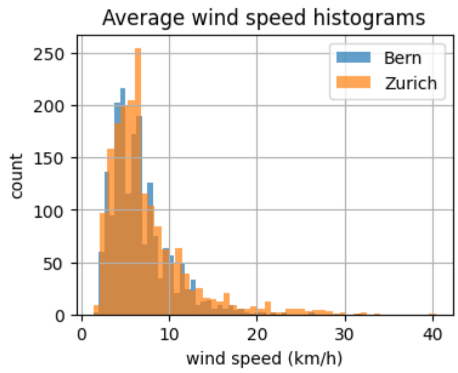


Figure A.24: Temperature difference histogram

Wind

Finally, Figure A.28 shows the wind speed in Bern and Zurich. The correlation between both cities was 0.81. Figure A.26 shows the wind speed histogram in both cities. The daily average is usually below 10 km/h, but it can reach up to 40 km/h. As for the wind gusts, Figure A.27 shows the histogram of the maximum wind speed registered daily. Note that wind speed can reach up to 200km/h, although wind gusts above 50km/h are rare. Figure A.29 shows the average wind speed for each wind direction. The strongest winds blow from the west and northwest, while the weakest ones blow from the south and southeast.



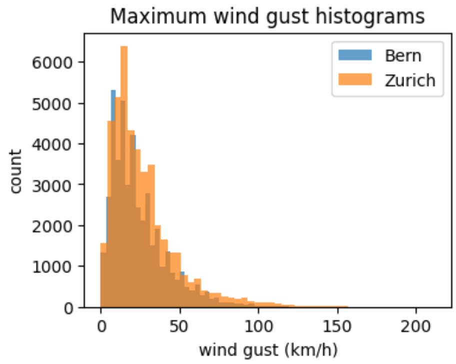


Figure A.26: Wind speed histogram in Bern and Zurich

Figure A.27: Maximum wind gust histogram in Bern and Zurich

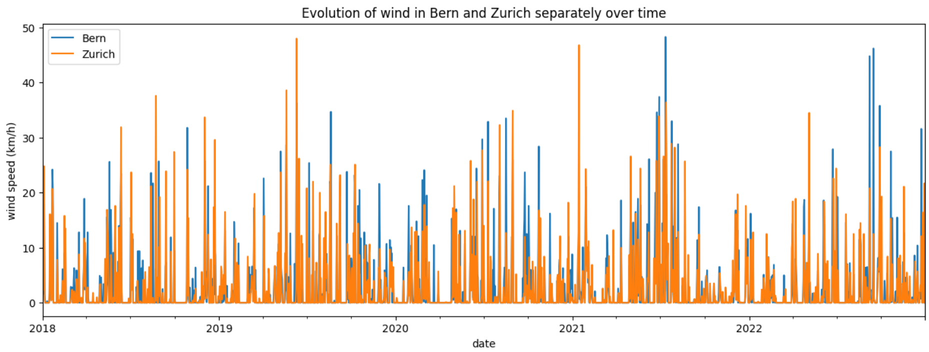


Figure A.28: Wind speed in Bern and Zurich

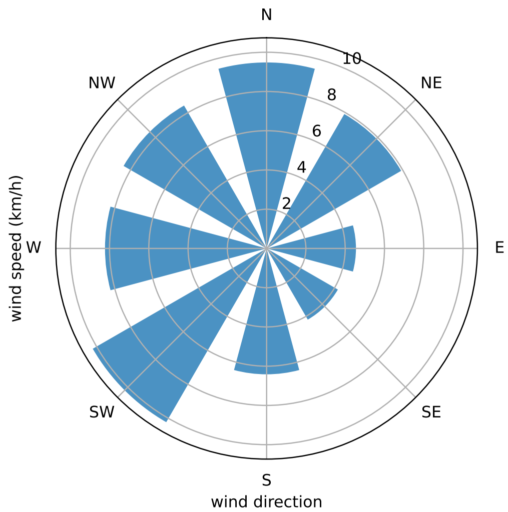


Figure A.29: Average wind speed per wind direction.