

Cyclic timetable improvement with train traffic data analysis

T. Graffagnino

*Swiss Federal Railways, Infrastructure, Timetable and Offer,
Switzerland*

Abstract

The Swiss Federal Railways (SBB) plans and operates one of the densest railway networks in the world. Analysis of train traffic data shows that overall punctuality has steadily increased since 2003. One challenge presented by a cyclic timetable is that the departure and arrival times for a given connection always remain the same during the day. In order to increase traffic density further while maintaining high punctuality, it is increasingly important to identify and manage delay fluctuations. The first part of this paper will be a short discussion about the problematic of a cyclic timetable in regard to the daily and seasonal variations. The second part of this paper will focus on setting a framework for the analysis of train traffic data by defining three dimensions: train, time and place. These three dimensions might at first seem obvious but are very powerful if used systematically. Based on these three dimensions, we will show some examples of different groupings of train traffic data used at SBB to identify potential punctuality problems. The third part presents a practical analysis of the seasonal delay fluctuations for the entire Swiss train network. In order to analyse the yearly fluctuation, we define a common punctuality measure, then compute ten years of train traffic data (about 14 million train arrivals). We will show that despite the fact that in these ten years the timetable changed different times, the seasonal fluctuation remain similar. Different questions such as the influence of winter, number of passengers or technical problems are discussed. Finally, the need for further data and other analysis to better understand the shown delay fluctuations will be addressed.

Keywords: cyclic timetable, punctuality, data analysis, train traffic, statistic.



1 Introduction

The quality of the SBB timetable is systematically analysed using train traffic data. Timetable analyses are carried out on a daily basis after the timetable change and then monthly during the year, thus allowing in-depth analysis of each yearly timetable and the definition of improvement measures which will then be integrated into the next timetable. Open Timetable is the main tool used by timetable schedulers when making these specific analyses of timetable stability [1, 2].

Over the years a three-dimensional structure for traffic data has proved increasingly helpful. This article will start by explaining these dimensions and proposing various possible ways of grouping data.

These ideas will then be applied to an actual traffic data analysis covering eleven years (from 6 January 2003 to 31 December 2013).

2 Considerations on cyclic timetable

The first Swiss cyclic timetable was introduced in 1982 after 10 years of preparation. The main reasons were the attractiveness for passenger, the gain of efficiency for the overall system and the difficulty to manage a non-cyclic tailor made timetable [3]. As a consequence, the departure and arrival times for a given train family (stopping at the same stations throughout the day) remain the same the whole day, without any regard for peak and non-peak hours. With adapted rolling stocks and extra peak hour train it is possible to counteract the increase of dwelling time due to the amount of passenger during peak hour. Another very important factor is to have a robust timetable that can swiftly recover from the peak hour delays. A key of success for the future lies in the understanding of the daily and seasonally punctuality variation. One future response will be to design the technical system to be less sensitive to the causes of the variations. Another idea to further increase the precision of the production will be to split the commercial (cyclic timetable for passenger) and the internal timetable (slightly non cyclic) in order to be able to take better into account the variations. In any case, it is very important to know thoroughly the causes of the variations, which is the goal of this study.

3 Dimensions and traffic data sets

The train traffic data used for this article were gathered over a period of eleven years. It was produced using two different rail traffic-monitoring systems. The data from January 2003 to March 2010 comes from the old monitoring system named SURF. Since March 2010 the data has been produced using a new system, called RCS [4]. The traffic monitoring systems gather data by recording rail track occupation, measured when trains cross signals. Each time a signal is crossed, the time is gathered and used to extrapolate the train's arrival and departure time for the corresponding station.



These data are then transmitted to the punctuality analysis system (called KVZ) every night. This second system then produces text files containing all the data for the day. Since the data format hasn't changed since January 2003, this gave us eleven years of data to analyze for this study.

3.1 Train traffic data format

The data for this study contains the following elements (see Table 1).

Table 1: Traffic data and its dimensions.

Description	Format	Dimension
Train number	Numeric integer between 1 and 99999	Train
Day of operation	YYYY-MM-DD	Time
Train arrival at station	YYYY-MM-DD-HH.MM.SS	Time
Train departure from station	YYYY-MM-DD-HH.MM.SS	Time
Delay upon arrival	Numeric integer	<i>Value</i>
Delay upon departure	Numeric integer	<i>Value</i>
Official abbreviation of the station's name	5-character alpha numeric XXXXX	Place
Track used	Text field	Place
Train type	Numeric integer	Train

When working on train traffic data, it is essential always to define which field of these three dimensions we are interested in. We must start by defining which trains we wish to analyse: train dimension. Then we need to define which stations we wish to analyse: place dimension. Finally, we must define which dates we are interested in: time dimension.

3.2 Possible calculated values

The basic raw data (Table 1) contains only the daily arrival and departure time with their respective delays. Based on these data, many values could be computed: 3-minutes punctuality upon arrival, planned arrival and departure time, planned and actual dwelling time, planned and actual journey time. These calculated values contain interesting information regarding the system, the variations and the possible causes, but won't be covered in this study. In this study, we will only use the 3-minute punctuality upon arrival (percent of arrival with delay ≤ 2 [min] 59 [sec]).

3.3 The three dimensions of the data

Traffic data represent for example the delay upon arrival of particular trains operating in particular stations on particular dates. Methods of naming trains or stations may vary from one Railway Company to another; however the intrinsic

information will remain the same. These three dimensions localize the delay values upon arrival in a 3 dimensional model, which can be as proposed by Ullius [5] represented as follows:

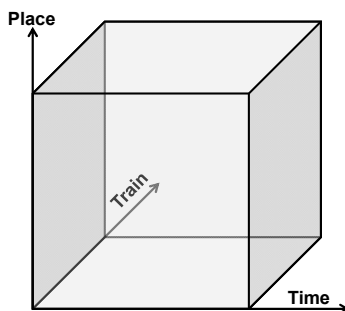


Figure 1: Cube of all available data.

Let us further examine each of these dimensions.

3.3.1 Place

The place dimension contains all of the information used to localize arrivals and departures. At SBB we have a short form name for every station and information about the track used: for example, “BN” for the station “Bern”.

3.3.2 Train

The train dimension contains all of the information used to specify which train is involved. At SBB, each train is defined by its number, which is unique for a given day. Furthermore, the type of train in numeric form shows what type of service the train provides: long-distance train or regional train for example.

3.3.3 Time

The time dimension contains the information of when the considered event took place.

3.4 Grouping the data in set

Each dimension may then be broken down in different ways. This allows categories or sets to be created in order to identify which one has noteworthy characteristics compared to others.

3.4.1 Data sets for the place dimension

The data set most commonly used for the place dimension consists of sorting stations by region. We can thus compare several stations at the same time. Stations can also be sorted according to significance or function: stop only, connection hub, etc.

Another possibility is a data set of all tracks serving one given station. For example, at Zurich’s main station (HB) tracks 21 to 24 are underground and used by regional trains (S-Bahn) only, while tracks 4 to 18 are at ground level and are mainly used by long-distance trains.

3.4.2 Data sets for the train dimension

At SBB, train numbers are allocated by family and the even/odd parity of the train number is used to determine the direction of operation. For example, trains 700 to 749 are long-distance trains that run between Geneva Airport and St. Gallen, passing through Lausanne, Berne and Zurich. Even numbers run (east to west) from St. Gallen to Geneva Airport and odd numbers run (west to east) from Geneva Airport to St. Gallen.

The data can be sorted into other sets using the information on train type. This allows different products, such as regional trains (S-Bahn), long-distance, etc., to be identified. For example it can be interesting to consider only the long distance trains which will lead to the following set on the train dimension:

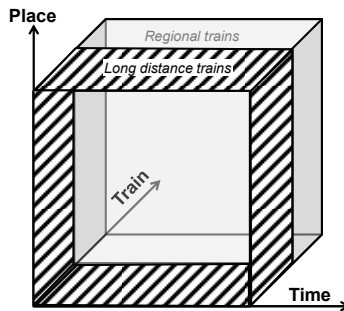


Figure 2: Set of long distance trains.

3.4.3 Data sets for the time dimension

The time dimension is the one most frequently used in data sets. We can start by grouping the data into years, months, weeks or days (e.g. 01.01.2011). Next, we can also group the data by day of the week, from Monday to Sunday.

Another option is to group the data by time period: data from 06:00:00 to 06:59:59, from 07:00:00 to 07:59:59 and so on. It is also possible to define specific periods of the day and then use these to create data sets. At SBB, peak periods are defined as 06:30:00 to 08:59:59 in the morning and 16:00:00 to 18:59:59 in the evening. This obviously leaves two remaining periods of the day, from 09:00:00 to 15:59:59 and from 19:00:00 to 06:29:59. The first period, during the daytime, can be left as it is, but it makes sense to break the other period down into two parts: evening, from 19:00:00 to 00:59:59, and night, from 01:00:00 to 06:29:59.

3.5 Filtering the data with sets and combination of sets

Each set of data can be used to filter and get insights in the data. For example, we can analyse arrivals at a given station by splitting them into types of train (regional trains or long-distance) and periods of the day (not peak hours or peak hours) in order to create 4 distinct data subsets (as represented in Figure 3).

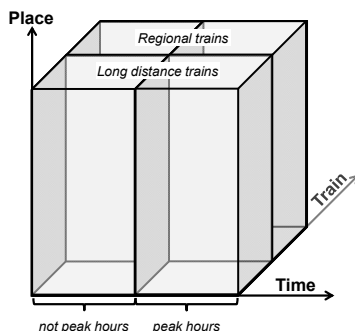


Figure 3: Example of 4 subsets.

4 Method of data processing used for the case study

4.1 Definition of punctuality for the purposes of this study

The official method of measuring punctuality has been adapted several times over the eleven years in question. For the purposes of this analysis, it was therefore necessary to define a fixed method of measurement for all the years so that the results are comparable. The punctuality figures calculated here may therefore vary slightly from SBB's official statistics. The punctuality analysed is the percentage of trains less than 2 minutes 59 seconds late on arrival. The arrivals are those into 13 major Swiss stations: Bellinzona, Biel/Bienne, Berne, Basel, Geneva, Lausanne, Lucerne, Olten, Sargans, St. Gallen, Visp, Winterthur and Zurich. The analysed trains are passenger trains only: long-distance trains and regional trains. They have been identified according to their type of train value (train types 39, 40, 41, 42, 43, 49, 50, 51, 53, 60 and 61). The study is made on the normal working days only (i.e. Monday to Friday), without taking account of bank holidays (in other words, Saturdays and Sundays have been systematically omitted).

4.2 Data processing method and tools used

The analysis of the eleven years of data had to be done in two stages. First, the data was extracted from flat files using Matlab. This pre-processing consisted of reading all the necessary text files, extracting the train events at the stations and computing the events in order to create the aggregated data.

The second stage consisted of producing the graphics using pivot tables in Excel.

5 Punctuality fluctuation analysis

5.1 Daily punctuality

The percentage of the train 3-minute arrival punctuality has continually increased over the eleven years under consideration:

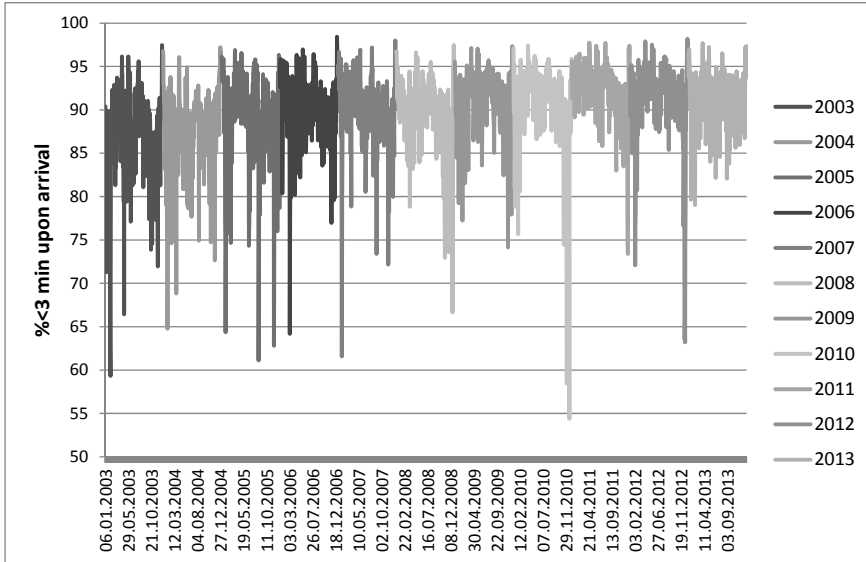


Figure 4: Daily punctuality in Switzerland.

Figure 4 shows that general punctuality has steadily improved over the past eleven years. However, the worst daily figure was recorded in winter 2010 (17.12.2010) and the start of the year 2011 had a very high punctuality. One can see that despite the winter, punctuality levels are usually very high at the turn of the year. Despite the global trend of improved punctuality, it seems that the variations in punctuality repeat every year, being better in summer and worst in winter.

One way of analyzing the seasonality of the punctuality would be to use moving average methods [6]. In this study we will however extensively filter the data using data sets in different ways within the time dimension.

5.2 Monthly punctuality

Let us see how punctuality looks when the data is sorted monthly within the time dimension:

Figure 5 shows a repeated drop in punctuality (sometimes only slight) during June and another in autumn affecting the months of September, October and November. We can also note that the 4 first months of year 2011 were the best punctuality ever achieved.

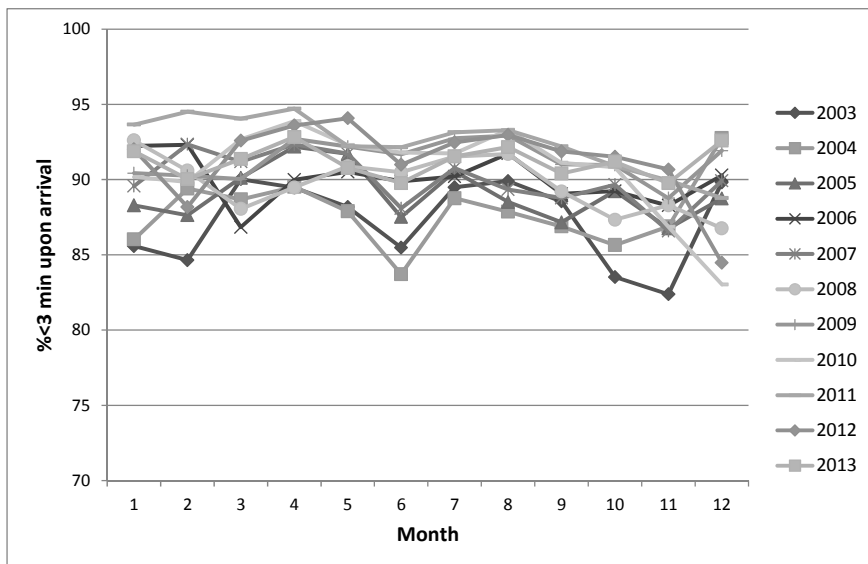


Figure 5: Monthly punctuality in Switzerland.

5.3 Punctuality according to time of day

For the purpose of this analysis, the following definitions have been defined for times of day:

Table 2: Times of day.

Description	Short name	From	To
Peak hour morning first part	PHM1	06:00:00	06:59:59
Peak hour morning second part	PHM2	07:00:00	08:59:59
Day first part	DAY1	09:00:00	12:29:59
Day second part	DAY2	12:30:00	15:59:59
Peak hour evening first part	PHE1	16:00:00	17:29:59
Peak hour evening second part	PHE2	17:30:00	18:59:59
Evening first part	EV1	19:00:00	21:59:59
Evening second part	EV2	22:00:00	00:59:59
Night	NIGHT	01:00:00	05:59:59

The different times of the day were split in order to better identify and understand when an effect occurs. Let us see how punctuality looks when the data is sorted by time of day within the time dimension.

By contrast, the daily variation over the years is surprisingly constant, as shown in Figure 6. The overall level seems depend on the year and thus to the annual timetable. A noticeable decrease in punctuality takes place during the morning and evening peak periods, as well as during the second half of the

evening. The highest punctuality levels are seen in the second part of the middle of the day.

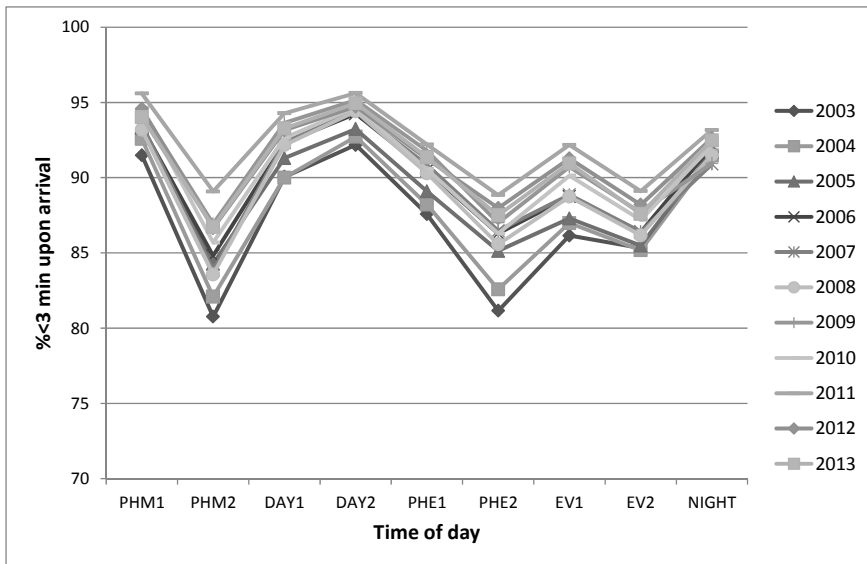


Figure 6: Punctuality by time of day for the different years.

5.4 Punctuality according to time of day filtered by month

Figure 6 shows surprising consistency, so let us see whether this continues to be the case throughout the year. In order to do so, we filter the data by month within the time dimension. This gives us a graph illustrating times of day for each month.

The month of January stands out because it shows a bigger variation in punctuality than for the year as a whole. The second half of the morning peak period has the worst punctuality levels of the day. Punctuality can also be seen to improve significantly during the evening and night.

In February, punctuality varies less than it does in January. The second half of the morning peak period has the worst punctuality levels of the day, not far off the figures recorded for the second half of the evening peak period.

In March, the morning deterioration in punctuality reduces further compared with the preceding months. By contrast, a new effect appears in the second half of the evening, when punctuality deteriorates again.

In April, punctuality continues to improve during the morning peak period. The overall variation in punctuality during the day reduces considerably. The second half of the evening continues to show an even greater deterioration in punctuality compared with the preceding months.

In May, the years 2003 and 2004 stand out as special cases. The introduction of the Rail 2000 timetable seems to have had a very positive effect on the

evening peak period. However, the general level of punctuality during May is slightly lower than in April.

During June, the general deterioration in punctuality noted during May continues. Furthermore, a new effect appears: the evening peak period has the day's worst punctuality figures, whereas it was the morning peak period which recorded the worst figures at the beginning of the year.

In July, punctuality improves overall and there is quite a marked change in the variation pattern. The peak periods are not particularly affected, while punctuality levels hit a low during the second half of the evening.

In August, deterioration in punctuality during the peak periods becomes noticeable again, even though the overall variation is similar to that of July.

September sees punctuality deteriorate again during the morning peak period, while the second part of evening has similar punctuality levels than the evening peak period. General punctuality deteriorates compared with July and August.

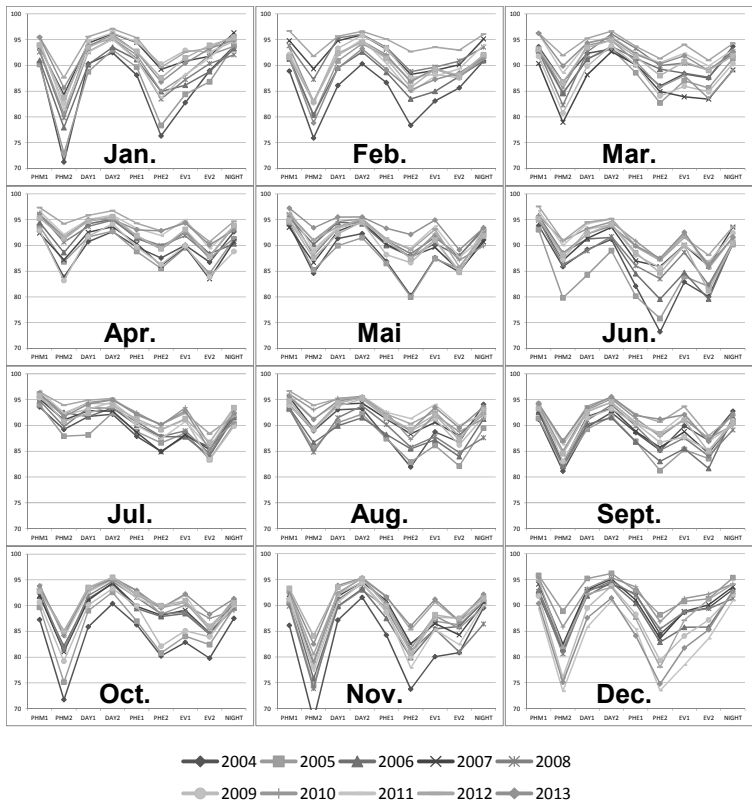


Figure 7: 3-minute punctuality upon arrival by time of day for each month of the year.



In October, there is a marked deterioration during the morning peak period, which again becomes the time of day with the worst punctuality figures. Here again, 2011 can be identified as one of the best years.

November sees a general worsening in punctuality. The peak periods again become the times of day with the lowest punctuality levels.

In December, general punctuality improves again. The peak periods remain the times of day with the worst punctuality figures. As in January, there is a steady improvement in punctuality during the evening.

5.4.1 Conclusions regarding the monthly analysis of times of day

For each month there are strong similarities between the different years; the punctuality level seems to follow the same trend every year. Sorting the data by time of day, filtered by month, has furthermore highlighted different phenomena that are repeated every year. In winter, the second half of the morning peak period has the worst punctuality levels. It may be seen that the first half of the morning peak period remains relatively good, which suggests that the deterioration during the winter is simply due to the greater number of customers who decide not to use the railway instead of their bicycles, motorcycles and cars because of the difficult winter conditions.

Similarly, punctuality deteriorates markedly during the evening in summer. Again, this is the time of year when customers are most likely to return home late in the evening and use the late running trains.

In May and June, a new decline in punctuality is evident, whereas March and April seemed to have brought an improvement in punctuality. Once again, customer behavior would seem to explain this: in Switzerland, May and June are the months when a lot of school trips take place. These two months see thousands of groups of pupils crisscrossing Switzerland to visit new areas. The deterioration is slightly noticeable during the day, but is particularly marked during the evening peak period. Since all the schoolchildren want to go back home at about 17:30, that is when the railway system is busiest, whereas in the morning they leave slightly later than commuters on their way to work.

6 Conclusions

After having introduced the concepts of dimensions and data grouping at the start of this article, we have been able to use them in a practical case where data covering eleven years were analysed. These concepts of dimensions and data grouping, which may at first sight appear trivial, are in practice extremely useful because they help prevent researchers becoming bogged down in a large quantity of data and possible analysis. The example of analysis presented in this article could be refined further through the use of the other dimensions of place and train. It would, for example, be possible to rework all the graphs so as to differentiate long-distance from regional services or to analyse each station separately. Further analyses with other data sources will allow us to refine the hypotheses advanced here on the possible causes. In the future, we will use data of passenger, track possessions, weather and technical defects to find out the

most probable causes or combination of in order to better understand the sources of punctuality variation.

References

- [1] Graffagnino, T., “Ensuring timetable stability with train traffic data”, In: Brebbia, C.A., Tomii, N., Mera, J.M., Ning, B., Tzieropoulos, P. (eds.), *Computers in Railways XIII*, pp. 427-438, WIT Press, Southampton, 2012.
- [2] Open Timetable, www.via-con.de/ott
- [3] Stähli, S., “Grundfragen der Fahrplangestaltung”, In: *Monatsschrift der Internationale Eisenbahn-Kongress-Vereinigung*, pp. 445-456, Juli-August 1969.
- [4] Rail Control Center, <http://www.sbb.ch/en/group/sbb-as-business-partner/offers-for-rus/onestopshop/systems/rail-control-system-rcs.html>
- [5] Ullius, M., “Verwendung von Eisenbahnbetriebsdaten für die Schwachstellen- und Risikoanalyse zur Verbesserung der Angebots- und Betriebsqualität”, Ph.D. thesis, ETH Zürich, 2005.
- [6] Wieser, Ph., *Logistique et analyse prévisionnelle*, Script EPFL-IML, Lausanne, 2000.