An algorithm to identify delay propagation routes based on visualization of association rules

Yasufumi Ochiai[1], Yuta Shibata[2], Norio Tomii[3]

[1]Odakyu Electric Railway Co., Ltd.
[2]Chiba Institute of Technology
[3]Nihon University

Summary

In urban railway lines where trains are running densely, even a small delay easily propagates to other trains. Thus, railway companies want to make their timetable more robust. They also want to evaluate if the timetable revision was successful or not from the viewpoint of robustness. In this paper, we propose an algorithm to identify primary delays and delay propagation which are very often occurring. Then we propose to visualize the results on a train graph. Thus, timetable planners can know how they should revise the timetable and judge if the revision was successful by comparing the results for the previous timetable and ones for the new timetable. The algorithm is developed based on the association rules. The algorithm takes historical train traffic data as an input and produces association rules which mean propagation of delays. We show some results of application to actual data to show how effective our algorithm was in making a timetable of an urban railway line more robust.

Keywords: timetable; historical train traffic records; association rules, delay propagation, robustness

1 Introduction

Although trains in Japan are known to be very punctual, one of the recent problems in urban railway lines is that small delays very often happen during morning rush hours. Because trains are operated very densely, even a small delay easily propagates to other trains and the delay tends to expand. Hence, railway companies are now keen to reduce occurrence and propagation of such small delays.

Delays are categorized into two types: one is a primary delay and the other is a secondary delay [1]. The latter is sometimes called a knock-on delay. A delay of a train is called a primary delay if the train itself is responsible for the delay. A secondary delay means that the cause of the delay is not the train itself but the delay was caused by a delay of another train. Timetable planners want to erase primary delays together with delay propagation. In particular, they want to prevent delay propagation to wider areas. In order to prevent delay propagation, in principle, it is necessary to add more margins so that delays are absorbed by the margins. For example, to give larger running time supplements, to increase intervals between trains to make the buffer time larger and to increase dwell times are effective to absorb delays. But of course, it is wasteful to increase margins everywhere. Thus, timetables have to be very carefully made so that adequate quantity of margins are included at proper places. Secondly, to show the routes of delay propagation is quite useful for timetable planners to figure out how they should improve the timetable. Thirdly, after they revised the timetable, obviously, they would like to evaluate if the revision was successful or not.

So, timetable planners have the requirements as follows:

1 To identify primary delays which cause secondary delays in wider area.

2 To identify the routes of delay propagation.

3 To evaluate if a timetable revision worked well in reducing delay propagation after they revise the timetable.

In addition, it is reasonable to consider frequency of delay propagation. If a propagation occurs say, once a month, the propagation might be negligible but a propagation occurs almost every day, the timetable has to be revised so that the propagation never or seldom occurs. Thus, it is necessary to consider the frequency of delay propagation.

In this paper, we propose an algorithm to extract primary delays together with delay propagation routes from historical train traffic data. We depict the results on a train graph. The algorithm is designed based on the association rules: one of the most popular techniques in the world of the data mining. The algorithm produces association rules which mean co-occurrence of delays.

In section 2, we explain why small delays happen during rush hours and what kinds of counter measures are taken by railway companies to reduce those delays. In section 3, we introduce our algorithm to identify delay propagation. In section 4, we show the results which we got by applying our algorithm to actual data of timetable revision.

2 Small delays and delay propagation

2.1 Small delays and delay reduction measures

In Japan, there is a big demand for railways in urban areas. As a matter of fact, in Tokyo area (the capitol of Japan), the total number of passengers of railways a day in average amounts to even 38 million. To transport such a massive amount of passengers, trains are operated very densely. In many railway lines in Tokyo, trains which consist of eight to ten (sometimes even 15) cars are running every two to three minutes on a double track line. This means that 20 to 30 trains are running per hour per direction on a double track.

Still, trains and stations are very congested. Sometimes, congestion rates of trains during peak hours exceed 150% or much more. The congestion of trains sometimes cause delays. That is, in order to run trains frequently, dwell times of trains are set to be as short as possible. In many stations, the dwell time is less than one minute and much less for smaller stations. Thus, if some trouble due to congestion happens and the dwell time becomes longer, the departure of the train is delayed and the delay propagates to succeeding trains. In addition, because trains have very small amount of running time supplements, the delayed train cannot resume the delay at next stations.

Railway companies are now very keen to reduce delays and they are making every effort to make their timetable more robust. Some of their countermeasures are improvement of infrastructure (tracks, signaling systems, rolling stock), deploy more staff on a platform, improvement of timetables etc. (for the details, please refer [2]).

2.2 Timetable revision to avoid delay propagation

To improve infrastructure needs a lot of time and budget. On the other hand, to revise a timetable does not cost such a long time and a big amount of money.

There exist several ideas of timetable revision to make it more robust.

1. To give more margins in running times.
2. To increase headways between trains.
3. To increase dwell times.
4. To change times, routes etc. so that conflicts of trains do not occur.
5. To change the line plan. For example, to change the destination of trains so that their delays are not brought to wider area.

These ideas, of course have to be implemented after careful observations and considerations. Furthermore, sometimes it is difficult or almost impossible to increase dwell times and headways if we want to keep current frequency of trains. We have to have knowledge how the delays are occurring and propagating and based on the observation, we should revise the timetable. In addition, we have to care about how frequent a delay propagates as we already mentioned.

It is also very important to be careful about the performance of the timetable. If we add large margins, the timetable might become very robust but its performance becomes very poor because journey times of passengers increase.

Thus, we have to be very careful in revising timetables having an observation where primary delay occurs, how the delay is propagating, how often the propagations are occurring and how far the delay is propagating and so on.

2.3 Related works

Many papers have been published which deal with robustness of timetables.

Cacchiani [3] and Lusbya [4] are good survey papers for robustness of timetables. As for the algorithms concerning detection of delay propagation, Yamamura [5] introduced an idea to apply a longest path algorithm to trace the route of delay propagation.They express the train traffic records by a directed graph and introduced an idea to trace back the critical path on the graph from the delay in focus to its origin, which is regarded to be the cause of the delay, namely the primary delay. The algorithm works well if we can properly identify the critical path. But the problem is because the trains do not always run ideally (for example, trains' running times are not always exactly the same as the technically minimum running times), it is not an easy job to identify the critical path.

Conte [6] introduced an idea "a tri-graph" to analyze the delay propagation. Arrival and departure delays of trains are associated with random variables. But they assume that they know the delay distribution in advance. Flier et al. [7] introduced an algorithm to detect the two types of dependencies among delays, namely dependencies due to resource conflicts and due to maintained connections. But they only examine the delay dependency in a station and they are not dealing with the network-wide dependency of delays.

Büker and Seybold [8] introduced an algorithm to estimate delay propagation based on an analytical approach. Their approach is completely analytical and they do not resort to the Monte Carlo simulation. But on the other side, this means that they have to prepare the probabilistic distribution function of delays which properly reflect the real world.

Cule et al. [9] introduced an algorithm based on a technique which is often used in the data mining: that is "mining of frequent episodes." An episode is considered to be a set of events that reoccurs in the sequence within a window of specified length and one example they showed is "Trains A, B, and C, with C departing before A and B, are often delayed at a specific location, approximately at the same time." But they are more interested in the episode of each train (their algorithm identifies the train numbers) and do not deal with the group of trains, which is necessary in urban railways.

There also exist papers which propose algorithms to make a timetable more robust (such as [10]-[14]) and papers which introduce definition of robustness (such as [15], [16], [17]). There also exist papers which deal with historical train traffic data (such as [18], [19]). But as far as the authors know, there have been no papers which focus on improvement of robustness by applying datamining algorithm to historical train traffic data.

3 An algorithm to identify delay propagation routes

3.1 Association rules

Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n binary attributes called items. Let $T = \{t_1, t_2, ..., t_m\}$ be a set of transactions. Each transaction in T contains a subset of the items in I. An association rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$ [21]. The set of items X and Y are called the left hand side (LHS) and the right hand side (RHS) of the association rule respectively.

The **support** supp(X) of an item (or a set of items) X is defined as the proportion of transactions which contain X in the whole transactions. The support means how often the rule holds. The **confidence** of an association rule is defined as $\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$. The confidence means how reliable the rule is. Thus, by focusing on the values of the support and the confidence, we can sort out the meaningful rules.

3.2 The proposed algorithm

(1) Basic ideas

In this paper, we consider the following as an item:

- an arrival delay of a train at a station.

- a departure delay of a train at a station.

A set of items above mentioned for one train of one day is regarded as one transaction. Then, we try to obtain association rules from these data using a priori algorithm, which is a very popular algorithm to get association rules proposed by Agrawal and Srikant [21]. We expect to find an association rule such as "if a delay of a train at Station A is larger than α seconds, then the train's delay becomes larger than β seconds at Station B." Here, α and β are the threshold of delays. Namely if (actual arrival time at station X – planned arrival time at station X) >= α, we regard the arrival of the train is "delayed." This is the same for departure delays.

We also specify thresholds for the support and the confidence and we output association rules whose support and confidence are larger than the thresholds. This is because we want to get delay propagations which very often happen.

We use R [22], which is a free software developed for data mining and has various functions including a priori algorithm.

(2) Issues

One of the issues we have to be careful about is "correlation is not the same as causality." Even if we get a rule such as "if a delay of Train X at Station A ⇒ delay of Train Y at Station B," we cannot conclude that the delay of Train X at Station A is a cause of the delay of Train Y at Station B. It may be the case that these delays occurred independently.

In order to solve this problem, we assume the following:

(a) Because we are discussing delay propagation in a railway line where trains are running very densely, if a departure of a train from a station is delayed and the arrival of the succeeding train at the station is delayed, we assume the arrival delay was caused by propagation of the delay.
(b) Because dwell times are set to be quite short in the timetable, if the arrival of a train at a station is delayed and the departure of the train from the station is also delayed, we assume the departure delay was caused by propagation of the delay.
(c) Because running time supplement is very small, if a departure of a train is delayed and the arrival of the train at the next station is delayed, we assume the arrival delay was caused by propagation of the delay of the departure.

Thus, we select the association rules which satisfy either of the conditions (a), (b) and (c) and whose support and confidence are larger than the thresholds so that we may well expect the selected rules satisfy causality.

(3) Visualization of association rules on a train graph

We visualize association rules on a train graph so that we can intuitively grasp the route of delay propagation. We show examples in the next section.

3.3 How we apply the algorithm?

(1) To identify primary delays and delay propagation

Because a primary delay is caused mainly by an increase of a dwell time, its value is usually very small, such as 15 seconds or so. So, in order to find a primary delay, thresholds of delays (α and β) must be small (say, 15 seconds). But because we are particularly interested in delays which become larger, we propose the following approach.

1. Obtain association rules setting the delay thresholds as th_1.
2. Obtain association rules setting the delay thresholds as th_2, larger than th_1.
3. Obtain association rules setting the delay thresholds as th_3, larger than th_2.
4. Obtain association rules setting the delay thresholds as th_4, larger than th_3.
5. We draw the results of Step 1 on a train graph in blue.
6. We draw the results of Step 2 on the same train graph in green.
7. We draw the results of Step 3 on the same train graph in orange.
8. We draw the results of Step 4 on the same train graph in red.

Thus, we can get a train graph on which delay propagations are depicted by arrows and the colors of arrows are different reflecting the quantity of delays. From this train graph, we can intuitively grasp where the primary delays which are the cause of wide spreading secondary delays are and how the delays become larger.

As for the values of the thresholds of the support and the confidence, they should be decided based on an intention of users (ie. timetable planers of railway companies). For example, if they think it is problematic if a delay propagation is occurring three days a week (among weekdays) and more, the support should be 0.6.

(2) To evaluate if timetable revision was successful from the viewpoint of delay propagation

In order to evaluate if timetable revision was successful, we are more interested in how many primary delays exist and how far the delays are propagating. Hence we should just set the threshold of delays as an equal value (say, one minute) and obtain association rules. The same discussion holds for the value of the support and the confidence as we already discussed in the procedure to identify primary delays.

4 Application to actual cases

4.1 Timetable revision reflecting quadruplication

In this section, we show how we applied our algorithm to actual cases.

(1) Quadruplication in Odakyu Electric Railway Company

The target line is a railway line which connects suburban area and the center of Tokyo. Trains are running very densely but nevertheless trains were very congested. Hence, small delays very often happened during morning rush hours.

In order to decrease the congestion, quadruplication for a part of the line was planned and we were going to revise the timetable in March 2018 when the quadruplication work was expected to be finished (quadruplication work had been already completed for almost all the part and the construction for the final part – 1.6km – was finished in March 2018). We show a sketch of the whole track in Fig. 1 (only major stations are shown) and the detailed track layout between Station Y and Station U in *Fig. 2* and Fig. 3. *Fig. 2* is the track layout before the quadruplication work completed and Fig. 3 is one after the quadruplication work. As you see, The track between Station M and Station U is quadruple after March 2018.

In the quadruple part, it was required to increase the frequency of trains per hour per direction during the morning rush hour from 27 to 36 trains. In relation to this, it was also required to increase the frequency in the double track area connected to the quadrupled area (from Station Y to Station M) from 27 to 30. Also, some trains go (and come) directly to subway from Station U.

Although the capacity is expected to increase by quadruplication, we were very much anxious about the robustness because the frequency of trains increases (please note that frequency of trains to/from subway lines also increases) and a delay propagates from Odakyu line to the subway line and vice versa. So, we made the timetable very carefully [23] . We show the timetables in Fig. 4. Besides the difference of frequency of trains, we significantly changed the line plan. For example, trains from the subway lines went far in the old timetable whereas in the new timetable they do not go so farther. This is because we aimed to increase the frequency of trains which go directly to the subway lines in inner city area. We also expected delays of the subway lines do not widely propagate.

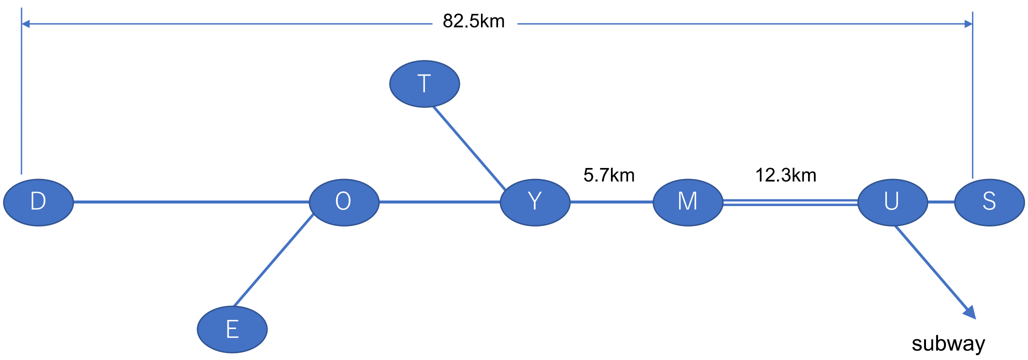


Fig. 1 Odakyu Electric Railway.

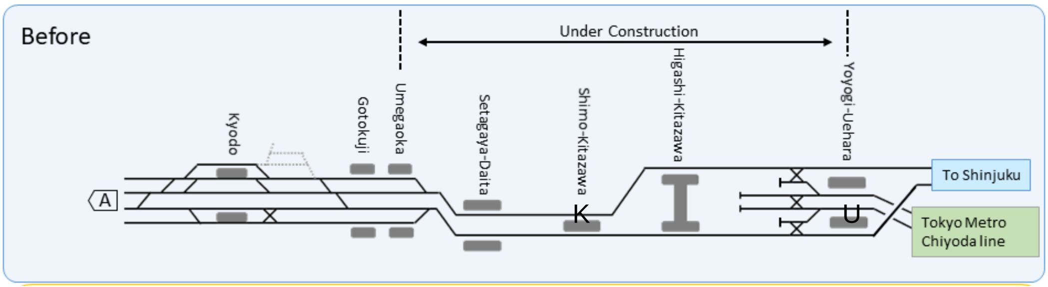


Fig. 2 Track layout before quadruplication (a part).

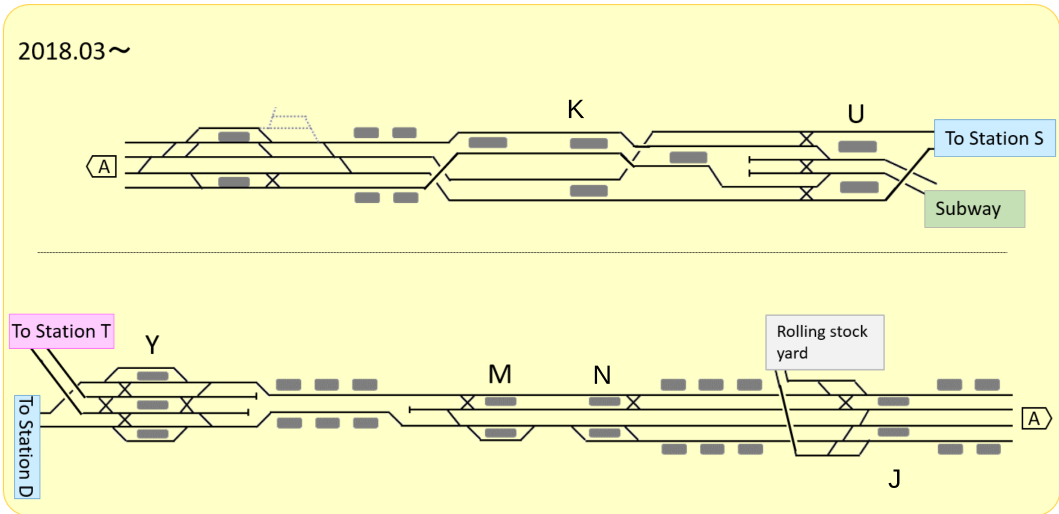


Fig. 3 Track layout after quadruplication.

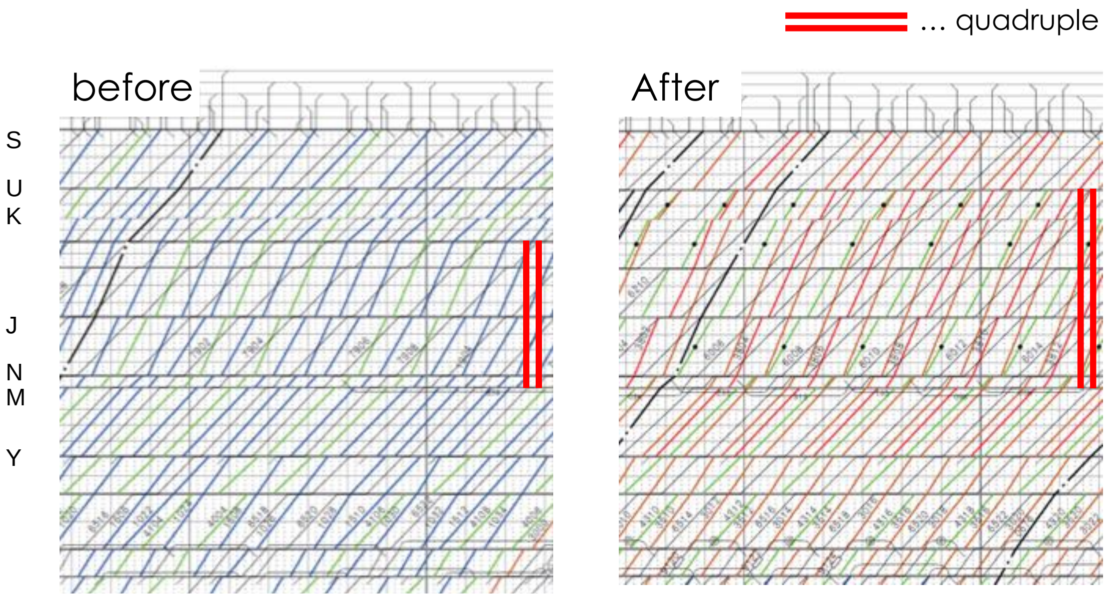


Fig. 4 Timetables: before (left) and after (right).

4.2 Results of applications

(1) To find primary delays and delay propagation

In order to clarify how the situation of delay emergence and propagation was improved (or worsened), we applied our algorithm to the historical train traffic data of weekdays from October to December of 2017 (namely, before the timetable revision) and those of weekdays from October to December of 2018 (namely, after the timetable revision) following the procedure we showed in section 3.3 (1). We show the results in *Fig.* 5 to Fig. 8. We set the both delay thresholds (α and β) as 15 seconds in Step1 (blue), 30 seconds in Step 2 (green), 60 seconds in Step 3 (orange) and 120 seconds in Step 4 (red), respectively. We set the threshold of the support as 0.6 and that of the confidence as 0.6.

(2) To confirm if the timetable revision was successful or not from the viewpoint of delay propagation

In order to know how the delays propagate, we applied our algorithm to the same historical train traffic data and visualized the results on a train graph as shown in Fig. 9 and Fig. 10. We set the delay thresholds as one minute. The thresholds of the support and the confidence are both 0.6. Please note in this case, the colors express the value of the support of the association rule (namely, how frequently the propagation occurs) to know how frequently the delay propagation is occurring.

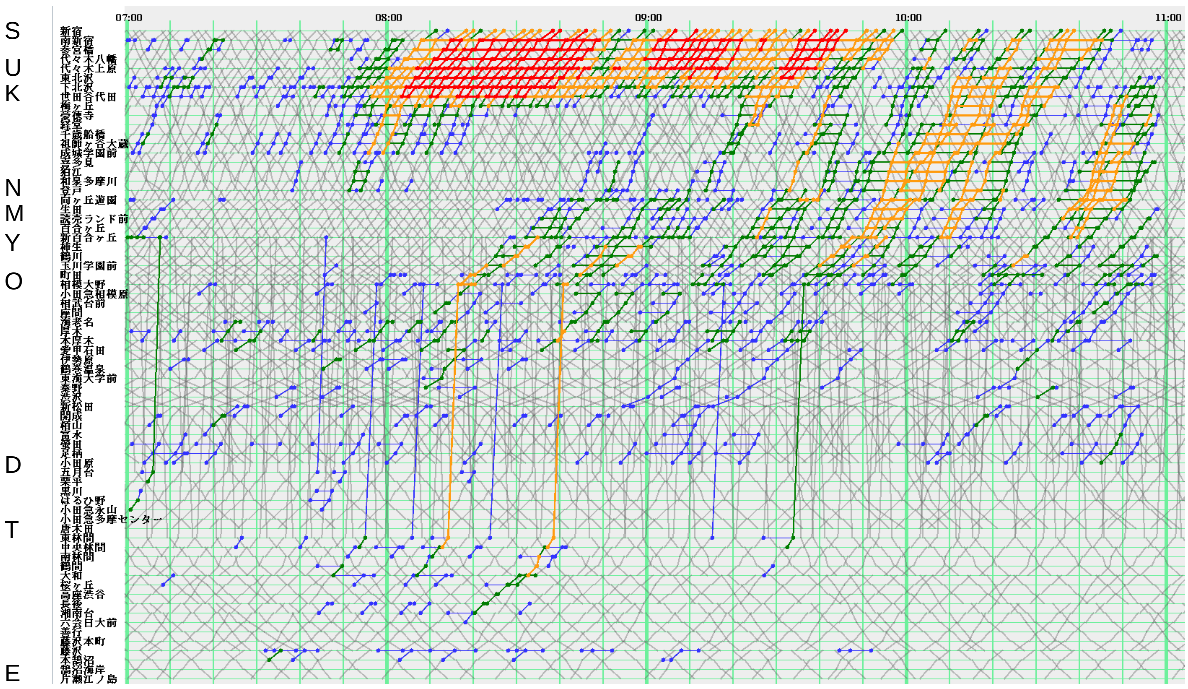


Fig. 5 Result for the historical data of the old timetable – inbound direction.

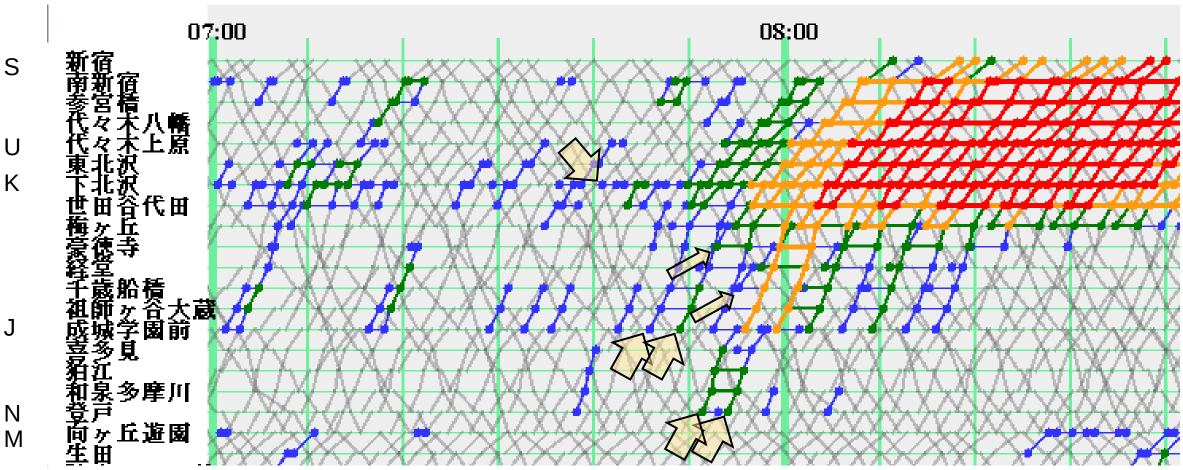


Fig. 6 Result for the historical data of the old timetable (enlarged).

4.3 Discussions

By comparing Fig. 5 and Fig. 7, we can know that the delay emergence and delay propagation pattern have changed significantly by the timetable revision. In the old timetable around 7:40 small primary delays occurr which cause larger secondary delays. In particular, the delays become quite large at Station K. This is because Station K is a very busy station where a lot of passengers get on and off and dwell times tend to increase and nevertheless the track was still double.

On the other hand, in the new timetable, although small primary delays are occurring, they do not become so serious. In particular, after 9 o'clock, no serious delays are occurring. We can observe delays become a bit large at Station U from 8 o'clock to 9 o'clock. This is because some trains go directly to subway lines at this station and because of the delay of the subway trains, trains of Odakyu line had to wait at this station and the delay propagate to trains to Station S.

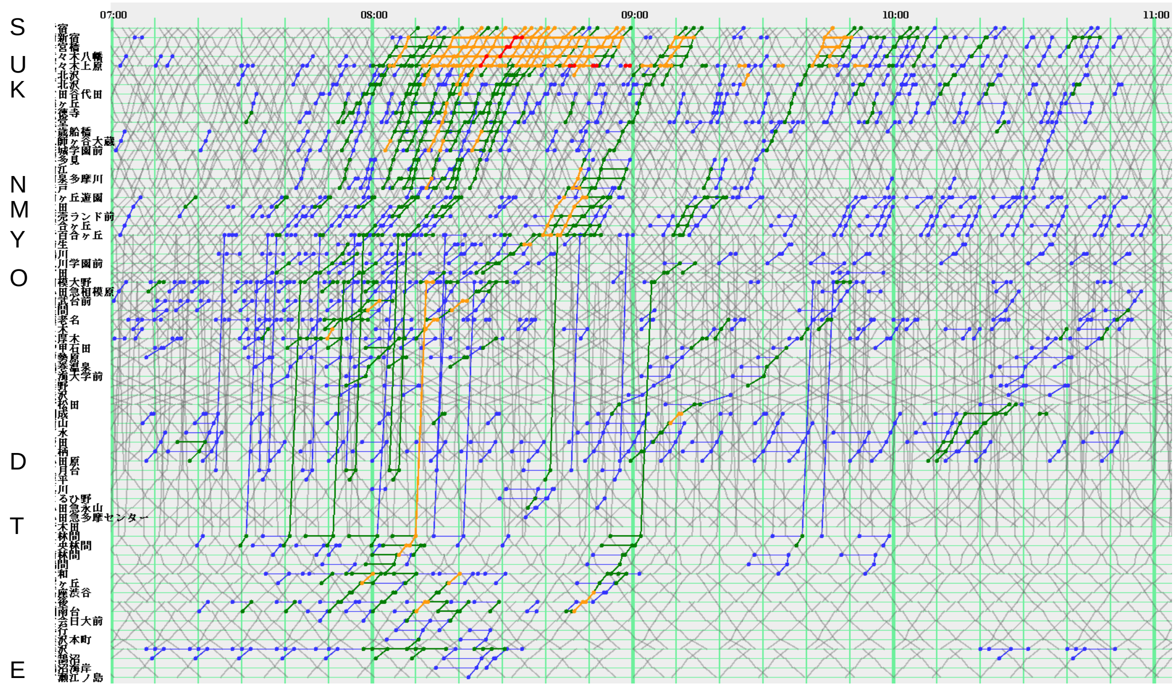


Fig. 7 Result for the historical data of the new timetable – inbound direction.

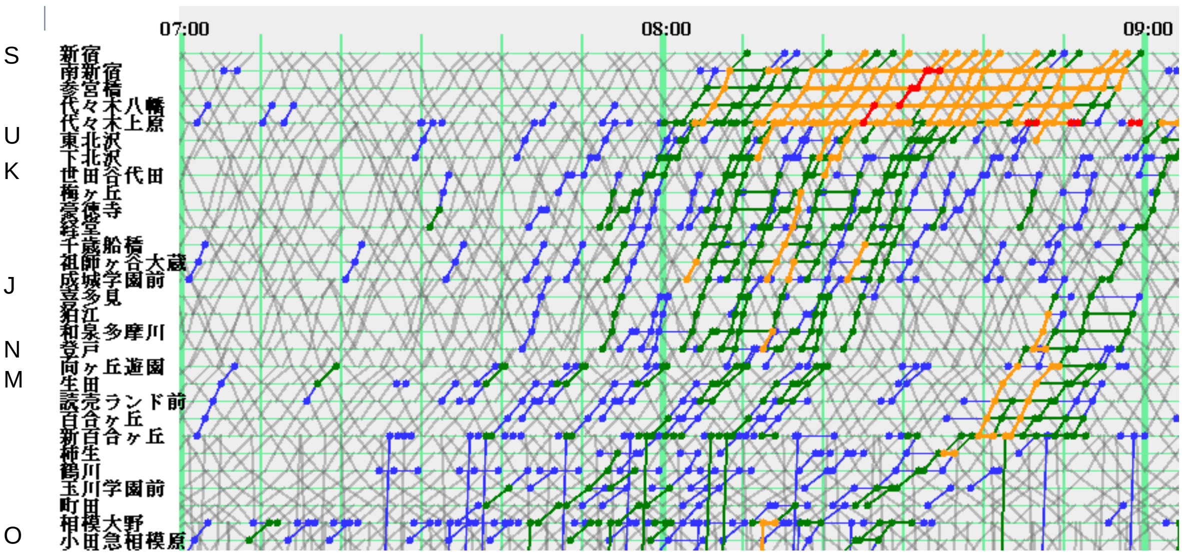


Fig. 8 Result for the historical data of the new timetable (enlarged).

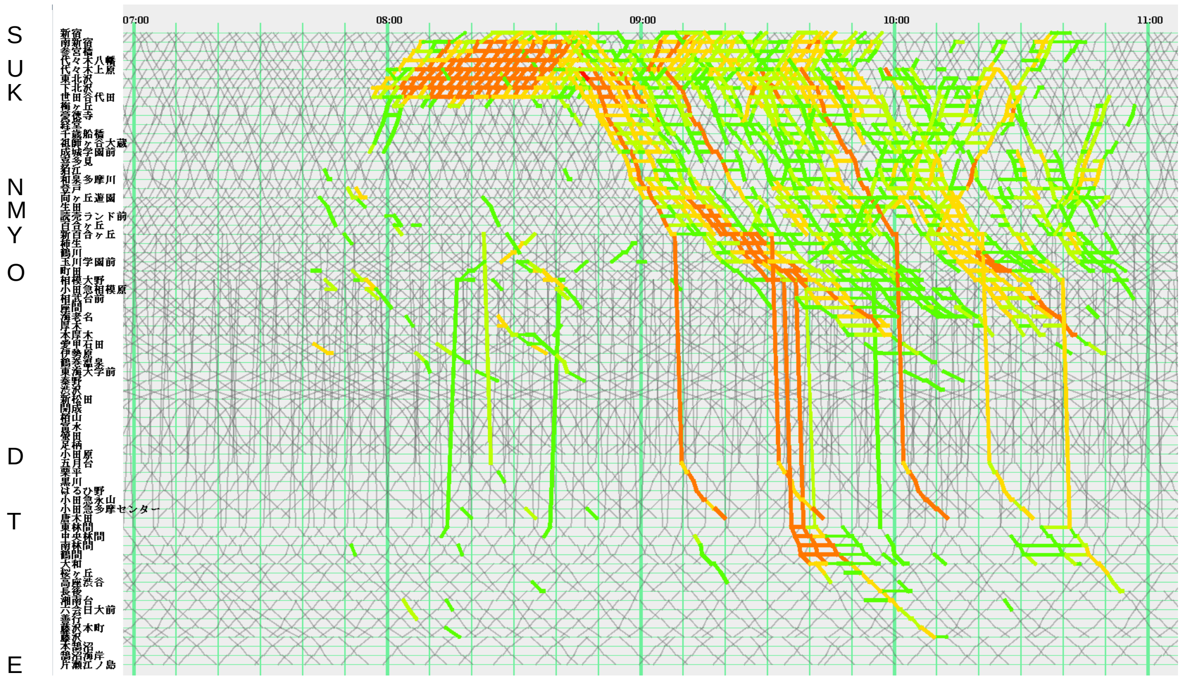


Fig. 9 Result for the historical data of the old timetable (both directions).

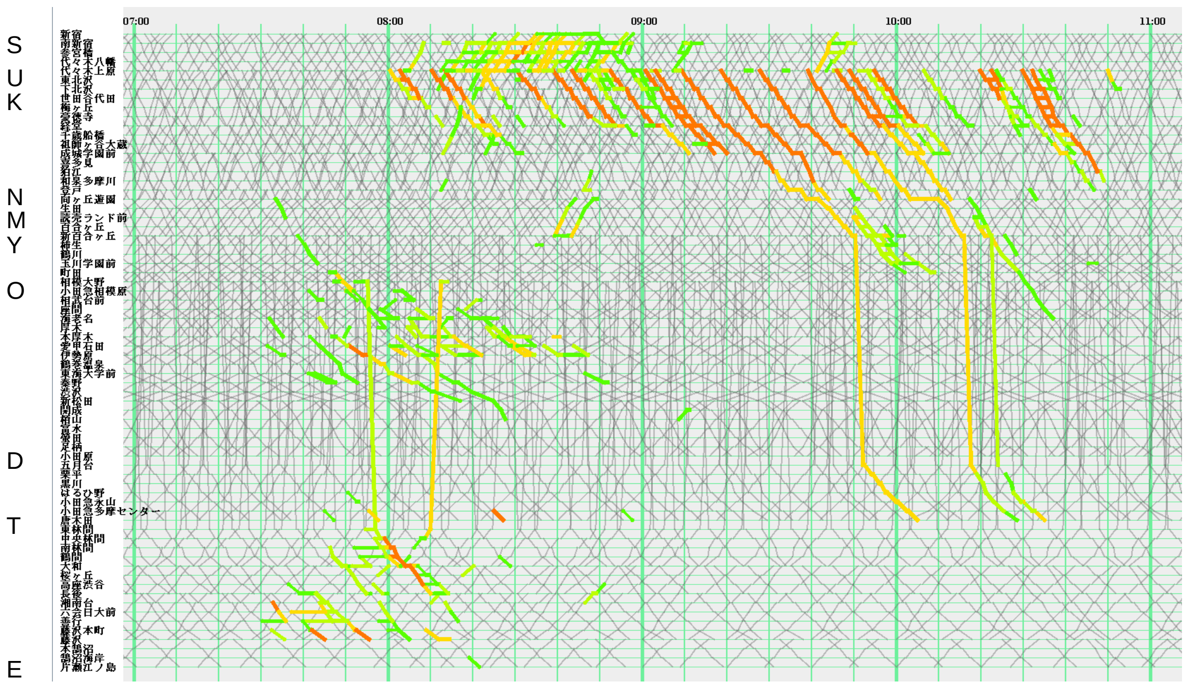


Fig. 10 Result for the historical data of the new timetable (both directions).

By comparing Fig. 9 and Fig. 10, we can know how the delay propagation patterns have changed by the revision of the timetable. In the old timetable (Fig. 9), we can observe that delays which occurred at Station K propagated to the terminal station whereas in the new timetable (Fig. 10), serious delays and propagations do not occur at Station K. This is because the new timetable was carefully made and the running speeds at critical points

are specified in the timetable[23]. In addition, delays of the outbound trains do not propagate so long. This is because the line plan was significantly changed so that delays are not brought to wider area.
So, as a consequence, from these results, we can conclude that the new timetable is far more robust than the old timetable.

5 Conclusions

We have proposed an algorithm to identify primary delays and delay propagation routes for railway lines where trains are running very densely. In our previous work [24], we showed an algorithm to find delay propagation routes for less busy lines but the proposed algorithm of this paper is applicable for busy railway lines, where the size of primary delays are usually very small. We also showed by applying our algorithm repeatedly using different value of the threshold and depict the results on a train graph, we can much easily find the primary delays which cause larger secondary delays.

One of our future work should be to quantitatively evaluate robustness of timetable from the viewpoint of delay propagation [25].

Acknowledgment

This research is partly supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C) 16K06428. The authors thank Ms. Ami Kono and Mr. Takuya Komuro for their contribution to this research.

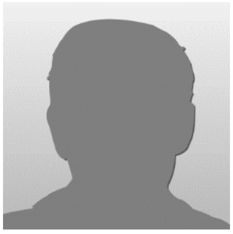
Literature

1. Hansen, I. & Pachl, J. eds., *Railway Timetabling and Operations - Analysis, Modeling, Optimization, Simulation, Performance Evaluation*, 2nd ed., Eurail Press, 2014.
2. Yamamura, A., Koresawa, M., Adachi, S, Tomii, N., How we have succeeded in regaining Punctuality in Tokyo Metropolitan Railway Network, *WCRR-World Congress on Railway Research*, Sydney, Australia, 2013.
3. Cacchiani, V. & Toth, P., Nominal and robust train timetabling problems, *European Journal of Operational Research* 219, 727–737, 2012.
4. Lusbya, R. M. and Larsen, J., Bull, S., A survey on robustness in railway planning, *European Journal of Operational Research* 266, 1–15, 2018.
5. Yamamura A., Koresawa, M., Adachi, S., Tomii, N., Identification of Causes of Delays in Urban Railways, *COMPRAIL2012 - 13th International Conference on Design and Operation in Railway Engineering*, New Forest, UK, 2012.
6. Conte, C. and Schöbel, A., Identifying dependencies among delays, *RailHanover – 2nd International Seminar on Railway Operations Research*, Hanover, Germany, 2007

7. Flier, H., Gelashvili, R. Graffagnino, T., Nunkesser, M., Mining Railway Delay Dependencies in Large-Scale Real-World Delay Data, *Robust and Online Large-Scale Optimization*, Lecture Notes in Computer Science Volume 5868, 2009.
8. Büker, T., Seybold, B, Stochastic modelling of delay propagation in large networks, *Journal of Rail Transport Planning & Management*, Volume 2, Issues 1–2 , 2012.
9. Cule, B., Goethals, B, Tassenoy, S., Verboven, S., Mining Train Delays, *Advances in Intelligent Data Analysis X*, Lecture Notes in Computer Science Volume 7014, 2011.
10. Andersson, E. V., Peterson, A., Krasemann, J. T., Quantifying railway timetable robustness in critical points, *J. Rail Transport Planning & Management*, 3(3), 2013.
11. Andersson, E. V., Peterson, A., Krasemann, J. T., Reduced railway traffic delays using a MILP approach to increase Robustness in Critical Points, *Journal of Rail Transport Planning & Management*, Volume 5, Issue 3, November 2015.
12. Fischetti, M., Salvagnin, D., Zanette, A., Fast Approaches to Improve the Robustness of a Railway Timetable, *Transportation Science* 43(3):321-335, 2009.
13. Jovanovic´, P., Kecman, P., Bojovic´, N., Mandic´, D., Optimal allocation of buffer times to increase train schedule robustness, *European Journal of Operational Research*, 256, 44–54, 2017.
14. Kroon, L. G., Helmrich, M. M., Vromans, M., Dekker, R., Stochastic improvement of cyclic railway timetables, *Transportation Research* Part B, vol. 42, issue 6, 2008.
15. Carey, M., Ex ante heuristic measures of schedule reliability, *Transportation Research* Part B 33 pp. 473-494, 1999.
16. Takeuchi, Y., Tomii, N. and Hirai, C., Evaluation Method of Robustness for Train Schedules, *Quarterly Report of RTRI*, vol. 48, no. 4, 2007.
17. Salido, M. A. et al., Robustness in Railway Transportation Scheduling, *Proc. 7th World Congress on Intelligent Control and Automation*, Chongqing, China, 2008.
18. Kecman, P, Corman, F., Peterson, A, Joborn, M., Stochastic prediction of train delays in real-time using Bayesian networks, *CASPT2015 - Conference on Advanced Public Transport*, Rotterdam, the Netherlands, 2015.
19. Labermeier, H., On the Dynamic of Primary and Secondary Delay, *RailCopenhagen - 5th International Seminar on Railway Operations Modelling and Analysis*, Copenhagen, Denmark, 2013.
20. Berry, M. J. A and Linhoff, G., *Data Mining Techniques – For Marketing, Sales and Customer Support*, John Wiley & Sons, Inc., 1997.
21. Agrawal, R. and Srikant, R., Fast Algorithms for Mining Association Rules, *Proc. 20th VLDB Conference*, Santiago, Chile, 1994.
22. R (2015) http://www.r-project.org/
23. Ochiai, Y. & Tomii, N., A Novel Timetabling Procedure which considers Running Speeds of Trains, *Proc. RailNorrkoping – 8th International Conference on Railway Operations Research*, Norrkoping, Sweden, 2019.

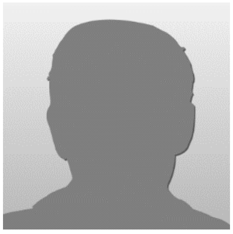
24. Kono, A. and Tomii, N., Identifying the Cause and the Propagation Route of Delays of Trains using Association Rules, *IRSA2017-International Railway Symposium Aachen*, Aachen, Germany, 2017.
25. Kono, A. and Tomii, N., Ex-Post heuristic measures of timetable robustness, EASTS Conference 2017 – *The 12th International Conference of Eastern Asia Society for Transportation Studies*, Ho Chi Minh City, Vietnam, 2017.

Authors



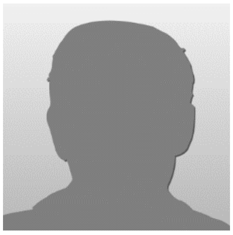
Ochiai, Yasufumi

After experience as a driver and a timetable planner of Odakyu Electric Railway Co., Ltd, Mr. Yasufumi Ochiai is currently working in MaaS department. Mr. Ochiai was a research student of Chiba Institute of Technology in 2018 and engaged in research to realize more robust timetables.



Shibata, Yuta

Mr. Yuta Shibata was a student of Chiba Institute of Technology majoring in Computer Science. In Chiba Institute of Technology, he worked to develop an algorithm to improve robustness of timetables applying datamining technique.



Tomii, Norio

Professor Tomii worked for Japan National Railways (JNR) and Railway Technical Research Institute (RTRI). After he worked as a professor in Chiba Institute of Technology, he is now a professor of Nihon University and is Deputy Director of Center for Railway Research. His current interests are applying optimization technique to railway scheduling, applying datamining technique to realize robust timetable etc. Professor Tomii is currently President of International Association of Railway Operations Research.