Method for Forecasting Fluctuation in Railway Passenger Demand for High-speed Rail Services

Ryosuke MATSUMOTO Daiki OKUDA Noriko FUKASAWA
Transport Planning and Marketing Laboratory, Signalling and Transport Information Technology Division

In order to plan high-speed rail transport services efficiently, it is necessary to be able to forecast fluctuations in passenger demand based on historical ridership data. Forecasting is difficult however, because of the number of components making up passenger demand. An effective way to forecast demand therefore should be to decompose these fluctuations into several independent demand components, which can then be forecast individually. This study applied an independent component analysis to decompose the fluctuation into several independent components. A method was then developed to forecast the fluctuation in passenger demand based on actual ridership data, calendar array, and number of people mobilized for large events.

Keywords: independent component analysis, time series analysis, demand fluctuations, demand forecasting, high-speed rail

1. Introduction

It can be said that high-speed rail in Japan (Shinkansen and limited express trains that connect cities at high speed) is a form of high-density transport. The Tokaido Shinkansen line connecting Tokyo, Nagoya, and Osaka’s three major metropolitan areas, which has the largest number of Japanese high-speed rail passenger, operates 365 train journeys per day, transporting 452 thousand passengers per day (data for 2016 fiscal year) [1]. Demand changes frequently due to calendar array and large events held near stations etc., so the ability to adapt operations, with extra trains etc., to these fluctuations is necessary.

Day-to-day rail services in Japan are carried out according to a transportation plan composed of train schedules, crew scheduling, and vehicle rostering. The basic transport plan, which is compiled during the annual train schedule revision, forms the basis of this transport plan. The basic plan specifies a transport plan for regular trains guaranteed to be operated every day with extra trains on standby with undetermined operating days. This means that the framework of the transport plan is already established when the basic transport plan is drafted. Margins are also provided in the basic transport plan, in order to allow increases in the number of trains to meet seasonal demand fluctuations or extra trains on a particular day. The merit of this approach is that it makes efficient use of resources, such as vehicles and crews, throughout the year. The drawback is that subsequent adjustments to the plan are difficult to make.

Therefore, in order to produce a high-quality high-speed rail transport plan, it is necessary to accurately predict fluctuations in passenger demand (changes in the number of passengers over a time series) based on historical ridership data. However, accurate predictions are difficult to make, because fluctuations depend on changes in passenger demand, and a variety of interlinked factors such as calendar array or large events taking place near stations, which generate surges in passenger volumes etc. In order to predict such complicated fluctuations precisely and effectively, they should be decomposed into several independent fluctuations that are then individually examined.

Independent Component Analysis (referred to as ICA hereafter) is a method for decomposing fluctuations made up of unknown multiple fluctuations, into several independent fluctuations. ICA has already been applied in various fields [2]. For example, in the voice analysis field, when a large number of people are talking at the same time, a specific person’s voice can be extracted. In the field of the brain wave analysis, by observing brain waves from the outside, several signals inside the brain can be separated. In the field of the image processing, noise in an image can be erased to improve its quality.

Several studies on the subject have already been conducted in Japan [3, 4, 5] in which ICA was applied to fluctuations in passenger demand. These studies aimed to clarify the characteristics of fluctuations in demand for cars at airports in relation to cancelled flights by applying ICA to fixed point observation data on automobiles flowing into one airport. However, to the knowledge of the authors of this paper, there are no published papers so far on ICA being applied to railway ridership data.

This study therefore applies ICA fluctuations in demand calculated from actual high-speed rail ridership data for a given set of origin and destination stations (referred to as OD hereafter). Then, by combining the calendar array and number of people mobilized for an event near a station with the ICA result, a method was built to predict future fluctuations in demand on a certain day.

2. Analysis method

2.1 Outline of ICA

ICA is a method for decomposing fluctuations composed of multiple unknown fluctuations into several independent fluctuations, to detect hidden factors and components from
multivariate data.

Let \( X(t) \) be the observation series which is the \( i \) th sample (the \( i \) th day) in the number of \( I \) observation series (\( I \) observation days) at the \( t \) th data (the \( t \) time period). When \( X(t) \) is obtained by the sum of products of several independent component series \( S_i(t) \) and mixing coefficient \( A_{ij} \) independent of \( t \), (1) holds for all \( t \).

\[
X_i(t) = \sum_j S_i^*(t)A_{ij}^* \quad (1)
\]

Here, by applying ICA to \( X(t) \), \( S_i^*(t) \) and \( A_{ij}^* \) can be estimated as \( S_i(t) \) and \( A_{ij} \) respectively. An image of ICA is shown in Fig. 1.

ICA is a method to estimate \( S_i(t) \) and \( A_{ij} \) simultaneously based only on \( X(t) \) when independence is assumed for \( X(t) \) of each \( i \).

By applying the principal component analysis (referred to as PCA hereafter) to the observation series, the number of independent component series can be determined based on the cumulative proportion of PCA results and the number of principal components corresponding to it.

2.3 Mixed coefficient prediction model

In constructing a mixed coefficient prediction model, the mixed coefficient \( A_{ij} \) is predicted on the basis of the calendar array and the number of people mobilized by an event on the \( i \) th day. The model is a multiple regression model. In the model, the calendar array are considered as a dummy variable by dummy coating. For modeling the behavior of people, and the crowds of people returning home at the same time after an event, a variable for the number of people mobilized by an event has to be prepared for each time band (18:00~18:59, 19:00~19:59 etc.) which is sum of the time of the end of an event and the time distance from the event venue to the station.

Let \( C(i) \) be the \( k \) th variable for the \( i \) th day, then the prediction value \( \hat{A}_{ij} \) of the mixing coefficient on the \( i \) th day for the \( j \) th independent component series can be expressed by (2).

\[
\hat{A}_{ij} = \alpha_i + \sum_j \beta_{ij} C(i) \quad (2)
\]

\( \alpha \): constant
\( \beta \): parameter

Table 1 shows the definitions of the calendar array in this study. For example, consecutive holidays when the number of visitors to sightseeing spots is expected to increase, and a weekday after consecutive holidays when demand is considered to be unusual compared to a normal weekday etc., are defined. The variables used in the mixed coefficient prediction model are indicated in Table 2 showing the resulting estimated parameters. Here, the dummy for the calendar array was based on a Wednesday.

3. Data

3.1 Data used in this study

As described in Chapter 1, in this study, actual high-speed rail ridership data was used (30 minute unit) for journeys to and from central stations in two central cities (referred to as station A and station B hereafter). In this paper, journeys from station A to station B are referred to as station A departures, and those from station B to station A as station B departures. The usual travel time by rail between station A and station B is 50 minutes.

Data was collected from April 1, 2006 to October 31, 2016, while data from April 1, 2016 to October 31, 2016 was used for prediction accuracy testing.

3.2 Monthly component decomposition

Data for each day and each time band was summed up into monthly data, to which component decomposition was applied.

The observation series for the monthly data were decomposed into trend components, monthly cycle components, and random components by applying the STL algorithm developed by Cleveland et al [6]. Trend components reveal tendencies over a long period of time; monthly cycle components reveal monthly tendencies over one year; while
the random series show the residuals obtained by subtracting the trend and the monthly cycle components from the observation series.

The data used in this study was long-term data covering 10 years. In order to extract the long-term trend components and the monthly cycle components reflecting seasonal fluctuations, component decomposition was applied to the monthly data.

The trend components and the monthly cycle components are shown in Fig. 2 and Fig. 3 respectively. The value of the trend component in April 2006 in Fig. 2 was set to 1 with the other values divided by the value of April 2006. In addition, the economic expansion period (January 2002 to February 2008, March 2009 to March 2012) and the economic recession period (February 2008 to March 2009, March 2012 to November 2012), set by the Cabinet Office in Japan [7], appear shaded red and blue respectively. The setting of these economic phases was based on post-verification, whereas for the phase following November 2012 there was no official statement.

The trend components did not appear to be influenced by the economic climate, and with a continuous gradual upward trend except for the period around the Lehman shock (September 2008). For this reason, future trend components were predicted using a simple time series linear regression. The twelve months following the Lehman shock were excluded, and the time series linear regression started in October 2009.

The monthly component was largest in August, which has 31 days including the long summer vacation (Obon festival in Japan), while it was smallest in February which has only 28 days and only one national holiday. From February to November, the components for station A departures and those from station B showed almost the same tendency, but the components of station A departures was larger than for station B departures in December, while the components station B departures were larger than station A departures in January. This is considered to reflect the surge in people returning home and travelling from large cities to local cities before the New Year’s holiday and with the opposite movement after the New Year’s holiday. In this study, the monthly cycle components are assumed to be constant even in the future, and past patterns are applied to the future.

### 3.3 Data to which ICA was applied

The period of data to which ICA was applied covered the fiscal years 2014 and 2015 (April 1, 2014 to March 31, 2016). There was almost no change in trend component variation during this time. However, since there were only three consecutive-day holidays consisting of Saturday, Sunday, and Monday during this period, the model corresponded more to the variable pattern of the calendar array by adding three consecutive holidays consisting of Saturday, Sunday, and Monday during this period, the model corresponded more to the variable pattern of the calendar array.
4. Analysis and discussion

4.1 Extraction of independent component series

The number of independent component series to be extracted in this study was determined on the basis of the cumulative contribution ratio of 80% in the PCA, and seven ($S_{A1}$ to $S_{A7}$) and six ($S_{B1}$ to $S_{B6}$) independent component series were extracted from the observation series for station A departures and station B departures, respectively. The extracted independent component series are shown sequentially in Section 4.3.

4.2 Parameter estimation for the mixed coefficient prediction model

Table 2 shows the parameter estimation results of the mixed coefficient prediction model for each independent component series. Concerts were selected as subjects for this study given the availability of information about start time and number of people mobilized for an event which was roughly equal to the capacity of the venue. In general, while the start times of concerts are announced, end times are not, so the end time was calculated by adding two hours to the start time.

This section describes a method for interpreting the parameters for the mixed coefficient prediction model. The constant term is a value which is multiplied with independent component series on all dates. On Wednesdays, which was selected as the standard weekday in this study, only the value of the constant term is multiplied with the independent component series. The value to be multiplied to each day of the three-consecutive day holiday (1st, 2nd and 3rd) can be calculated by adding the constant term, the parameter for each ‘day of the week’ and the parameter for ‘each day of the three consecutive day holiday.’ On the day of the event, in addition to the parameters of the calendar array, it is also multiplied by the event ‘end time’ parameter.

| Table 2 The parameter estimation result of the mixed coefficient prediction model |
|---------------------------------|---------------------------------|---------------------------------|
|                                 | Station A departures             | Station B departures             |
|                                 | $S_{A1}$ | $S_{A2}$ | $S_{A3}$ | $S_{A4}$ | $S_{A5}$ | $S_{A6}$ | $S_{B1}$ | $S_{B2}$ | $S_{B3}$ | $S_{B4}$ | $S_{B5}$ | $S_{B6}$ |
| Constant                        | -0.17   | 0.19   | 0.23   | 0.39   | 0.52   | 0.66   | 0.45   | 0.57   | 0.68   | 0.78   | 0.89   | 0.97   |
| Monday                          | -0.74   | -1.56  | -2.15  | -1.87  | -1.87  | -1.87  | -2.89  | -0.15  | -1.91  | -1.96  | -1.96  | -1.96  |
| Tuesday                         | -0.71   | -1.71  | -0.71  | -0.71  | -0.71  | -0.71  | -0.71  | -0.71  | -0.71  | -0.71  | -0.71  | -0.71  |
| Thursday                        | 0.67    | 0.96   | 0.67   | 0.96   | 0.67   | 0.96   | 2.67   | 2.67   | 2.67   | 2.67   | 2.67   | 2.67   |
| Friday                          | 6.11    | -3.94  | 1.31   | 9.76   | 1.35   | 1.71   | 8.91   | -3.40  | 3.33   | -2.03  | 15.67  | -1.13  |
| Saturday                        | 4.92    | -10.54 | -0.63  | 3.64   | 5.25   | 2.27   | 6.00   | 3.08   | 2.95   | 21.44  | 27.90  | -12.72 |
| Sunday                          | 10.73   | -34.37 | -1.67  | 20.81  | -11.72 | 9.09   | 2.67   | 2.67   | 2.67   | 2.67   | 2.67   | 2.67   |
| Three consecutive holidays      |        |        |        |        |        |        |        |        |        |        |        |        |
| First                           | 2.47    | 5.54   | 5.29   | -8.15  | 4.52   | -3.97  | -4.89  | 2.80   | -13.67 | 11.44  | -11.37 | 6.22   |
| Second                          | 9.28    | 9.28   | 9.28   | 9.28   | 9.28   | 9.28   | 3.21   | 3.21   | 3.21   | 3.21   | 3.21   | 3.21   |
| Third                           | 16.54   | 34.70  | 6.98   | 21.40  | -14.71 | 2.49   | 9.26   | 13.30  | -9.44  | 13.70  | 5.01   | 6.57   |
| Before consecutive holiday      |        |        |        |        |        |        |        |        |        |        |        |        |
| Saturday                        | 1.67    | -2.24  | -2.64  | -2.64  | -2.64  | -2.64  | -2.98  | -3.09  | -4.18  | -3.78  | -3.46  | -3.46  |
| Sunday                          | 4.18    | -1.73  | -1.95  | 3.86   | 2.04   | 3.77   | -5.56  | 5.93   | -2.37  | 2.76   |        |        |
| After consecutive holiday       |        |        |        |        |        |        |        |        |        |        |        |        |
| Saturday                        | -1.63   | 2.23   | -2.64  | 2.10   | -2.91  | -2.91  | -2.91  | -2.91  | -2.91  | -2.91  | -2.91  | -2.91  |
| Sunday                          | 2.12    | 4.47   | 3.90   | 4.41   | 3.63   | 5.72   | -8.50  | 12.13  | -4.23  | 5.00   |        |        |
| Single holiday                  | 13.01   | -17.73 | -13.99 | 7.29   | -4.21  | 4.63   | 10.58  | -9.96  | 9.53   | -11.34 | 15.97  | -7.99  |
| Single weekday                  | -3.39   | -4.94  | -0.39  | -0.39  | -0.39  | -0.39  | -3.32  | -5.99  | -5.99  | -5.99  | -5.99  | -5.99  |
| Two consecutive weekdays        |        |        |        |        |        |        |        |        |        |        |        |        |
| First                           | -3.60   | -4.31  | -3.60  | -3.60  | -3.60  | -3.60  | -3.60  | -3.60  | -3.60  | -3.60  | -3.60  | -3.60  |
| Second                          | -3.00   | -7.14  | -8.85  | 3.93   | 5.64   | 3.93   | 5.64   | 3.93   | 5.64   | 3.93   | 5.64   | 3.93   |
| Day after consecutive holiday   | -1.82   | -0.29  | -0.29  | -0.29  | -0.29  | -0.29  | -3.38  | 6.84   | -4.01  | -3.60  | 1.82   |        |
| Number of people mobilized for an event |        |        |        |        |        |        |        |        |        |        |        |        |
| Before 18:00                    | 7.91E-5 | -4.07E-5 | 4.35E-5 | 6.74E-5 | 1.17E-4 |
| 18:00~18:59                     | 1.45E-4 | 4.39E-5 | -4.96E-5 | 1.30E-4 |
| 19:00~19:59                     | 2.04E-5 | 5.22E-5 | 2.90E-5 | 2.01E-4 | 2.51E-4 |
| 20:00~20:59                     | 1.45E-5 | 4.71E-5 | 2.65E-4 | 5.35E-5 | 2.38E-3 |
| Number of people mobilized for an event |        |        |        |        |        |        |        |        |        |        |        |
| 21:00~21:59                     | 3.01E-5 | 4.56E-5 | 1.21E-4 |        |        |        |        |        |        |        |        |
| Adjusted R2                     | 72%     | 89%    | 29%    | 79%    | 64%    | 66%    | 73%    | 49%    | 53%    | 73%    | 77%    | 80%    |

Blue bold: 1% significant, Bold: 5% significant
4.3 Interpretation of independent component series

The independent component series of departures from station A and those from station B are interpreted on the basis of results from the analysis described in Section 4.1 and Section 4.2.

Independent component series $S_a$ and $S_b$ with peaks in demand in the morning and independent component series $S_{A3}$ and $S_{B3}$ peaks in demand in the evening are shown in Fig. 4 and Fig. 5, respectively.

$S_{A3}$ and $S_{B3}$ are fluctuations peaks in demand at 8:30 and 9:00 respectively. In addition, $S_{A3}$ also has a negative peak in demand at 15:30.

Both $S_{A3}$ and $S_{B3}$ were observed on weekdays (Monday to Friday) and Saturday, which suggests that they both reflect demand from commuting and business passengers and leisure passengers in the morning. However, since $S_{A3}$ is greater on weekdays than on Saturdays, it can be assumed that it reflects commuting and business passenger demand more than leisure demand. $S_{B3}$ however is higher on Saturdays than weekdays, indicating that it shows leisure traveler demand more than commuter and business demand. In addition, $S_{A3}$ has a negative peak in demand on Sundays and on the third day of three consecutive-day holidays, so it can be deduced that this reflects demand from leisure travelers returning home.

$S_{B3}$ and $S_{B5}$ show peaks in demand at 15:30 and 20:00 respectively. In addition, $S_{B3}$ shows a negative peak in demand at 10:30.

$S_{B3}$ and $S_{B5}$ were observed on weekdays and on Sundays, which suggests that they both reflect evening homeward bound commuter, business and leisure passengers. $S_{B5}$ is particularly high on Friday, which suggests a large component of homeward bound weekly commuters who work away from home during the week. Also, $S_{B5}$ showed a negative peak on Saturdays, which could be a reflection of leisure passenger demand.

The independent component series $S_{A2}$, $S_{B4}$ and $S_{B6}$ show peaks in demand in the morning and independent component series $S_{A1}$, $S_{A4}$, $S_{A6}$ and $S_{A7}$ with peaks in demand in the evening are shown in Fig. 6 and Fig. 7, respectively.

$S_{A1}$, $S_{A4}$, $S_{A6}$ and $S_{A7}$ are fluctuations with peaks in demand at 20:00, 18:00, 22:00 and 21:00 respectively.

Since $S_{B2}$ was largely observed on Mondays and Saturdays, it is considered that it mainly shows demand from weekly commuters and leisure travelers in the morning. $S_{B4}$ and $S_{B6}$ appeared on weekdays in general, but especially on Saturdays and on the first day of three consecutive-day holidays, which suggests a heavy component of leisure passenger demand.

The independent component series $S_{A3}$ and $S_{B1}$, showing
smaller peaks in demand than other independent component series, are shown in Fig. 8 and Fig. 9, respectively.

Both $S_{A3}$ and $S_{B1}$ were fluctuations with positive values in the time bands from the morning to evening and whilst they displayed negative values for the other time bands. The constant term values were larger in these fluctuations than in other independent component series, and their values were less influenced by calendar array that affect schedules, which suggest that they represent the fundamental fluctuations observed every day.

5. Accuracy verification

In this research, the accuracy of the constructed method was verified by checking reproducibility and prediction accuracy. Reproducibility and prediction accuracy were confirmed by examining the correlation coefficients of the series obtained through observation and those estimated with the model (reproduction series and prediction series), as shown in Fig. 10 and Fig. 11. The demand fluctuation estimated by the constructed method was the sum of all the trend series, monthly cycle series, and irregular series.

Statistically high reproducibility was confirmed for both station A and station B departures: on 604 days the correlation coefficients were 0.9 or more, and on 610 days (total target reproduction days) the correlation coefficients were 0.8 or more.

Statistically high prediction accuracy was confirmed for both station A and station B departures: on 170 days the correlation coefficients were 0.9 or more, and on 172 days (total target prediction days) the correlation coefficients were 0.8 or more.

This therefore demonstrates the statistically high accuracy of the constructed method in terms of both reproducibility and prediction accuracy.

6. Conclusions

In this study, a model for forecasting fluctuations in passenger demand was constructed by applying ICA to an OD between the central stations of a large city and a local city, and combining calendar array and events attracting large crowds, with the results obtained after applying ICA. This study demonstrated that the model based on the calendar array and event information, had high statistical accuracy.

In order to make the model presented in this study more versatile, the explanatory variables of the mixed coefficient estimation model focused on the calendar array and large scale events attracting large crowds. In order to make the model more practical, future work should try to identify increases in demand specific to each OD (for example, forecasts for the blooming of cherry blossoms).

References

[1] Central Japan railway Company, Annual report 2017, 2017.
[2] Hyvärinen, A. et al. (Japanese translation: Nemoto, I.), [Detailed explanation] Independent component analy-
sis. Tokyo Denki University Press, 2005.

[3] Inoue, S. et al., “Analysis on Non-ordinal Traffic to local Airport by using Independent Component Analysis,” Infrastructure planning review, Vol. 26, No. 3, pp. 583-593, 2009 (in Japanese).

[4] Inoue, S. et al., “Analysis on Unusual Traffic demand based on Continuous Traffic Observation at Hiroshima airport” Proceedings of infrastructure planning, Vol. 37 No. 54, 2008 (in Japanese).

[5] Tsukai, S. et al., “Analysis of time distribution pattern of Hiroshima airport inflow vehicle by data mining,” Proceedings of transport engineering, Vol. 28, pp. 269-272, 2008 (in Japanese).

[6] Cleveland, R. et al., “STL: A Seasonal-Trend Decomposition Procedure Based on Loess,” Journal of Official Statistics, Vol.6, No.1, pp.3-73, 1990.

[7] Cabinet Office (Japan), Turning point of business cycle, http://www.esri.cao.go.jp/jp/stat/di/150724hiduke.html

Authors

Ryosuke MATSUMOTO
Researcher, Transport Planning and Marketing Laboratory, Signalling and Transport Information Technology Division
Research Areas: Transport Planning, Transport Economics

Daiki OKUDA
Assistant Senior Researcher, Transport Planning and Marketing Laboratory, Signalling and Transport Information Technology Division
Transport Planning, Transport Economics

Noriko FUKASAWA
Senior Chief Researcher, Laboratory Head, Transport Planning and Marketing Laboratory, Signalling and Transport Information Technology Division
Transport Planning, Travel Behavior Analysis