Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Symposium of Transport Simulation (ISTS'18) and the International Workshop on Traffic Data Collection and its Standardization (IWTDCS'18)

Feature extraction of inter-region travel pattern using random matrix theory and mobile phone location data

Wataru Nakanishi[a,*], Hiromichi Yamaguchi[b], Daisuke Fukuda[a]

[a]*Department of Civil and Environmental Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro, Tokyo 152-8552, Japan*
[b]*Graduate School of Natural Science and Technology, Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan*

Abstract

Understanding inter-region travel patterns is an important issue for many reasons. This paper aims at utilising the data of the number of travellers, which nowadays can be obtained easily at any time by using mobile phone location data. In this study, an automatic feature extraction method is proposed with a random matrix theory-based principal component analysis (RMT-PCA), and its ability is confirmed by applying it to the data of the number of long-period inter-region travellers. The results show that some seasonal and weekly patterns, as well as economic and climatic situations, were revealed by the data, some of which might be missed by a conventional method. In addition, the selection of data, whether daytime or nighttime, brought about a different result in both the number of extracted features and their interpretation.

Keywords: Random matrix theory; Inter-region travel pattern; Feature extraction; Big data

1. Introduction

Understanding inter-region travel patterns is an important issue with regard to organising a campaign for tourism promotion and evaluating the impact of nationwide infrastructure construction such as high speed railway. Nowadays, data of the number of travellers can be obtained at any time thanks to mobile phone location data. The advantages of this data are 1) the large sample size and 2) spatio-temporal high resolution, contrary to the conventional traditional questionnaire survey. For example, "Mobile spatial statistics" (MSS) by NTT DOCOMO (Terada et al., 2013) provides the population of residential areas (location of permanence) and staying areas (location of temporality) pairs every hour. This population are estimated by following procedure using the records of periodical communications between mobile phones and base stations. At first, the number of mobile phones in each base station are obtained by

* Corresponding author. Tel.: +81-3-5734-2575 ; fax: +81-3-5734-2575.
E-mail address: nakanishi@plan.cv.titech.ac.jp

aggregating the record. Secondly, the population are calculated by multiplying by the reciprocal of the share of each attribute such as the place of residence. Since the communication records of more than 50 million mobile phones are being recorded through the service of NTT DOCOMO, we can obtain the population in any residential- and staying-area pairs at any time through this methodology [1]. Therefore, utilising this type of data will bring about new insights into travel patterns, e.g., grasping seasonal variations, discovering correlations and the time-series change in the pattern.

However, to specify a model or a system that generates this observed data, is a difficult task. This is because such a large dataset is generated by complicated interdependencies among many aspects, like a transport network, the economic situation, seasonal and weekly trends, differences in climate and the existence of promotions for travellers. In addition, although some periodic patterns in the data seem to exist, generally such interdependencies or even their components are not easy to determine. Given this scenario, the application of a data-driven feature extraction method, especially an unsupervised method, is a possible solution. Feature extraction applied to this study means discovering some ordinary and characteristic travel patterns associated with seasons and region pairs. Selecting the data-driven method leads to the utilisation of noisy and turbulent data. However, it implies an important requirement: that a method should be robust enough to handle noisy observations that the data may contain.

Regarding the application of a data-driven method, for example, Sun and Axhausen (2016) and Yao et al. (2015) have applied matrix and tensor decomposition methodologies to massive spatio-temporal mobility data to understand human mobility patterns. They explain that this data-driven understanding is effective for the modelling of urban structures or multi-context mobility prediction. Chen et al. (2018) and Tan et al. (2013) have suggested a methodology of complementing missing data from an understanding pattern obtained from a similar approach. Thus, over the past few years, a considerable number of studies have been conducted on data-driven methods for understanding intra-urban mobility patterns. Nevertheless, thus far, the studies have revealed only a superficial understanding of inter-region travel patterns. In inter-region travel, there is no distinct mobility pattern frequently held by the majority such as intra-urban commuting, shopping, and going home. Therefore, an approach is required to extract only statistically significant patterns out of the several weak features included in the dataset. This is one reason why a valid model has not been proposed for understanding inter-regional travel patterns.

To that end, in this study, we propose the application of random matrix theory-based principal component analysis (RMT-PCA) (e.g. Mehta, 2004) to day-to-day inter-region traveller population data. Then, the extracted features are interpreted as a spatio-temporal trend of inter-region OD patterns. RMT-PCA has two advantages. Firstly, like normal PCA by Pearson (1901) and Hotelling (1933), it is a totally unsupervised method. Secondly, an essential component and noise can be distinguished, theoretically. This is explained in detail in Section 3. Some applications of RMT to stock markets have demonstrated its ability to extract the essential features from noisy time-series data as in Laloux et al. (1999) and Plerou et al. (1999). However, to the best of the authors' knowledge, there is no application to transportation data. The contributions of this paper are the following. One is showing the possibility to explore inter-region travel patterns in high resolution with day-to-day samples over two and half years. The other is the introduction of random matrix theory into the transportation research field, which provides statistical measurements for the extraction of essential information from a given noisy dataset.

2. Data

In this study, the population of residential- and staying-area pairs at 3 A.M. and 1 P.M. every day are collected for 915 days (from 1st March 2014 to 31st August 2016) for each by MSS data. Although MSS provides the population data every hour on the hour, we use this dataset because the average number of travellers is the smallest at 3 A.M. and the largest at 1 P.M..

Regarding the definition of the zone, Japan is divided into 50 zones (these are almost consistent with prefectures). Next, to extract only travellers from the data, the population whose residential zones and staying zones are the same are replaced by 0. Then, these data are aggregated into 9 zones that are usually treated as a "region" in Japan (Fig. 1).

[1] If we regard the former area as "origin" and the latter area as "destination", this population pairs are similar to OD matrices in transportation field.

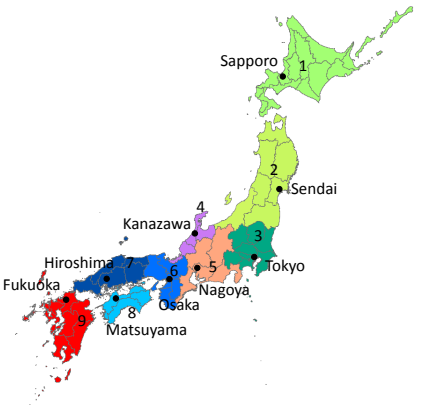

Fig. 1. 9 regions and representative cities. Okinawa prefecture is not drawn on this map because of the space limitation.

Table 1 shows the name and representative cities of each region. This division is defined mainly according to the National Spatial Strategies by the Ministry of Land, Infrastructure and Transport and Tourism (MLIT) of Japan [2]. The population of these 9×9=81 residential- and staying-zone pairs are regarded as the number of travellers in this study.

At last, the obtained matrices of 81 pairs × 915 days, denoted by $\boldsymbol{N}_{night}$ (3 A.M) and $\boldsymbol{N}_{day}$ (1 P.M.), are normalised to $\boldsymbol{B}_{night}$ and $\boldsymbol{B}_{day}$, respectively, with a mean for each row of 0 and a row variance of 1; let $\boldsymbol{n}_i$ be the ith row vector of $\boldsymbol{N}$ and $n_{i,j}$ the (i, j)-element of $\boldsymbol{N}$, then the same element of the normalised matrix $\boldsymbol{B}$, $b_{i,j}$, is calculated as $b_{i,j} = (n_{i,j} - \bar{n}_i)/\sigma_i$, where $\bar{n}_i$ represents the mean of $\boldsymbol{n}_i$ and σ_i the standard deviation of $\boldsymbol{n}_i$.

Table 1. Regions and representative cities.

Region	Name	Representative cities (three largest cities according to population)
1	Hokkaido	Sapporo, Asahikawa, Hakodate
2	Tohoku	Sendai, Niigata, Iwaki
3	National Capital Region	Tokyo, Yokohama, Kawasaki
4	Hokuriku	Kanazawa, Toyama, Fukui
5	Chubu	Nagoya, Hamamatsu, Shizuoka
6	Kinki	Osaka, Kobe, Kyoto
7	Chugoku	Hiroshima, Okayama, Kurashiki
8	Shikoku	Matsuyama, Takamatsu, Kochi
9	Kyushu	Fukuoka, Kita-Kyushu, Kumamoto

3. Methodology: random matrix theory

RMT (e.g. Mehta, 2004) is the theory that is based on the eigenvalue distribution of a matrix whose elements are random variables. Let $\boldsymbol{A}$ the $M \times T$ (assuming $M < T$ for simplicity) matrix and each element of $\boldsymbol{A}$ is generated independently according to the identical probability distribution whose odd-order moments are zero and second-order moment is finite. Many symmetric distributions, e.g., Gaussian distributions, hold this condition. Without a loss of generality, in the following, the variance of the distribution is considered to be one.

[2] As the National Spatial Strategies does not include the Hokkaido area and Okinawa area, we regard the Hokkaido area as an individual region (Region 1) and the Okinawa area as being included in Kyushu region (Region 9), considering the number of travellers and area balance.

Then, let us consider the eigenvalue distribution of the correlation matrix $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^t$ where the superscript t represents the matrix transpose. Although $\boldsymbol{C}$ is randomly distributed, in the limit $M \to \infty, T \to \infty$ and $T/M \to Q < \infty$ (finite constant), following the Marcenko-Pastur distribution universally holds.

$$P(\lambda) = \frac{1}{2\pi T} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{1}$$

$$\lambda_\pm = (1 \pm \sqrt{Q})^2 \tag{2}$$

where λ represents an eigenvalue and $P(\lambda)$ represents their probability distribution. Feature extraction using RMT relies on the upper limit of the distribution in Eq. (2), that is, λ_+. If the sample data are assumed to be generated independently from an identical distribution, the eigenvalue distribution obtained by the data, for example $\boldsymbol{B}_{night}\boldsymbol{B}^t_{night}$, should be consistent with Eq. (1). Usually this hypothesis is rejected and some eigenvalues from the sample data are larger than the upper limit λ_+. These eigenvalues are extracted as the essential features of the data. At the same time, the eigenvalues that are smaller than λ_+ just show that the data that are not distinguishable from noise. Therefore, the number of extracted features is theoretically decided contrary to normal PCA. This entire process is referred to as RMT-PCA in this paper.

It is worth emphasising that one major advantage of RMT-PCA is the automatic determination process of the number of features. On the contrary, in normal PCA, the empirical value that has little meaning must be exogenously employed to decide the number of features: the threshold of the cumulative contribution ratio (e.g., if the cumulative contribution ratio is larger than 0.8, then the extraction is finished) and eigenvalues (e.g., eigenvalues larger than 1 are to be extracted). Once the eigenvalues are extracted, eigenvectors corresponding to each eigenvalue behave the same of those in normal PCA.

Nonetheless Eqs. (1) and (2) hold in the infinite limit, many empirical studies have shown that if the matrix size is sufficiently large (around 100 rows and columns), the eigenvalue distribution can be asymptotically regarded as a Marcenko-Pastur distribution. Thus, the matrix in this study, whose size is 81×915, can be handled with RMT.

4. Application

In this section, the data explained in Section 2 are applied to RMT-PCA from the previous section. First, let the matrix $\boldsymbol{B}$ explained in Section 2 be the sample data matrices: $M = 81$ and $T = 915$. By substituting them into Eqs. (1) and (2), $\lambda_+ = 19.02$ is obtained. Therefore, for each sample matrix $\boldsymbol{B}_{night}$ and $\boldsymbol{B}_{day}$, eigenvalues larger than the value of λ_+ correspond to the essential features. Once eigenvalues are extracted, then the travel pattern is understood in the following way. Firstly, the corresponding eigenvectors, the principal components, show the meaning of each feature. Secondly, the loading vector for each component, the correlation between eigenvectors and the original data for each day, show the importance of each feature on each day. If an element of a loading vector is around 1, the corresponding principal component is positively affected by the number of travellers, and if it is around -1, it is negatively affected. In other words, if the sign of a corresponding element (a day) composed of a loading vector and a principal component are the same, the traveller number increases on that day, and vice versa.

In the following subsections, the results are interpreted according to these two aspects for each sample matrix, during the nighttime and daytime.

4.1. 3 A.M. result

The eigenvalues of $\boldsymbol{B}_{night}\boldsymbol{B}^t_{night}$ are $\lambda_{n,1} = 626.82, \lambda_{n,2} = 109.16, \lambda_{n,3} = 59.28, \lambda_{n,4} = 25.86, \lambda_{n,5} = 14.84, \ldots$ in descending order. Therefore, as $\lambda_{n,4} = 25.86 > \lambda_+ = 19.02 > \lambda_{n,5} = 14.84$, four features are extracted. Let the corresponding eigenvectors $\boldsymbol{p}_{n,1}, \boldsymbol{p}_{n,2}, \boldsymbol{p}_{n,3}$ and $\boldsymbol{p}_{n,4}$. The values of their elements are shown in Fig. 2 with the coloured mesh [3]. In addition, the loading vectors for each eigenvector are shown in Fig. 2. If the conventional method of cumulative contribution ratio up to 0.8 is applied, only two features are regarded as meaningful, meaning that two

[3] Indeed eigenvectors are vectors of 1×81, they are drawn as an OD matrix in the figures for clarity.

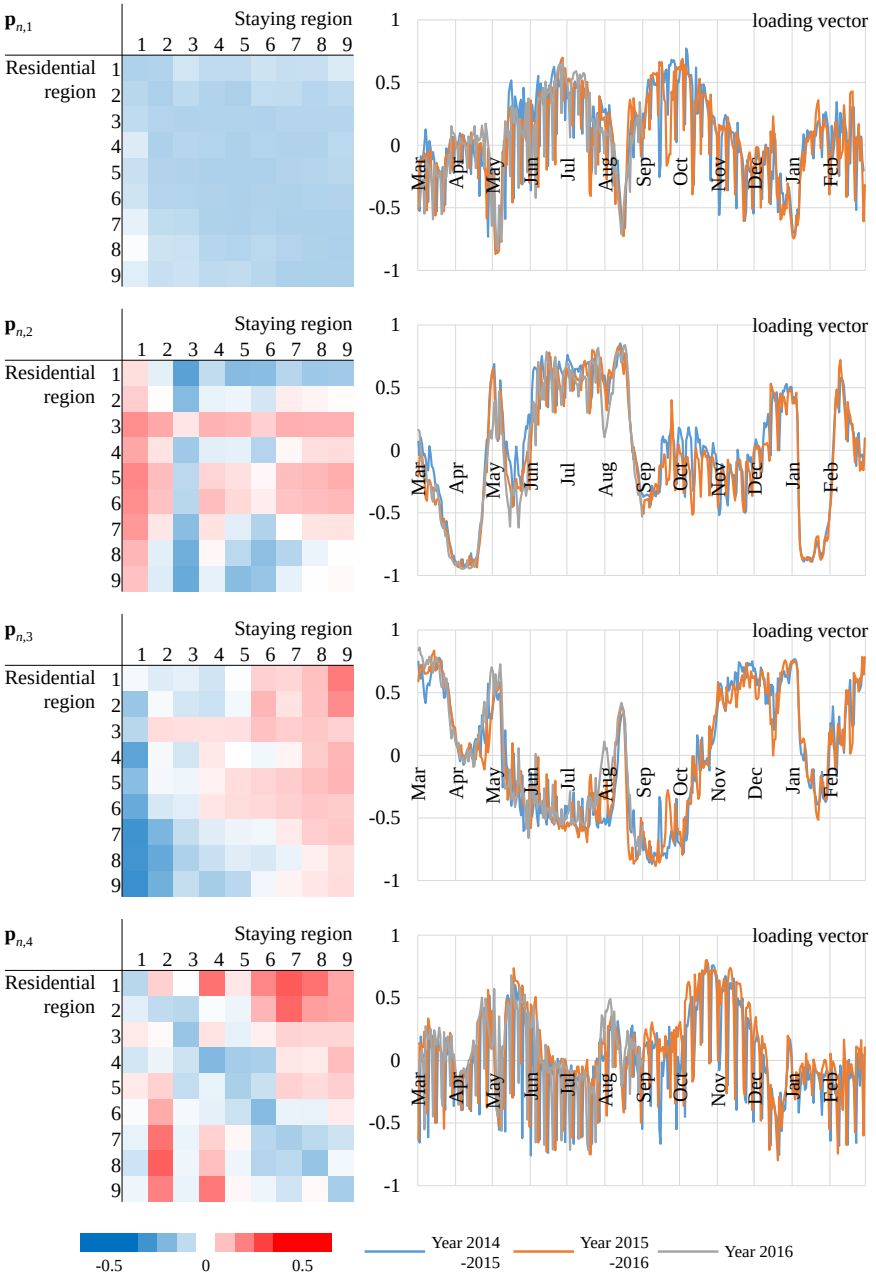

Fig. 2. Extracted features of 3 A.M. (left) eigenvector; (right) loading vector

essential features are missed. In a similar way, if the method of an eigenvalue larger than 1 is applied, 23 features are regarded as meaningful, meaning that 19 features which have statistically no meaning are falsely extracted.

All elements of the first principal component, $\boldsymbol{p}_{n,1}$, are negative. Thus, this feature is interpreted as the average variance, that is, the magnitude of inter-region travel. Compared with the loading vector, the number of travellers increases in early May, the middle of August (both are long holidays in Japan) and the New Year's Holiday and decreases in weekdays in June and October.

The second principal component $\boldsymbol{p}_{n,2}$ shows the entry to and exit from regions 3, 5 and 6 (National Capital, Chubu and Kansai regions), in which the biggest cities in Japan are located. Compared with the loading vector, many travellers go out of these regions for summer vacations (and especially go to region 1 (Hokkaido)), and in January and April many travellers come to these regions. It might be understood that sightseeing activity in Japan is mainly done by people in these regions during the summer season, and people in the other regions seldom go out of their regions for sightseeing. This also implies that inter-region travel in January and April are mainly business trips. These interpretations could be consistent with the current economic situation in Japan, in which only big cities are considered vital.

The third principal component $\boldsymbol{p}_{n,3}$ clearly shows the travel direction. Many people go north in September and south in November to March according to the loading vector. This might be understood as being due to the climatic effect because people, especially tourists, like to go north during the summer season to pass the summer, and vice versa in the winter season. In addition, the loading vector shows the reverse pattern in the middle of August, as traditionally many Japanese go back to their hometown regardless of the climatic condition during that season.

The fourth principal component $\boldsymbol{p}_{n,4}$ shows the difference of intra- and inter-region travel behaviour. As the low peaks in the loading vector are mainly on Sundays, it shows that intra-region travel increases on Sundays and inter-region travel increases in October. It might be understood that intra-region travel increases according to the increase in tourists on weekends. Actually, the national inter-regional survey (MLIT, 2010) also presented that most of the leisure travel on weekends is directed to the neighbouring prefectures (intra-region), which is shorter than most business trips. On the contrary, as business trips might be distinguishable in October, inter-region trips appear to increase.

By using this methodology, four types of seasonal or weekly patterns are extracted. In addition, they are understood as an average variance and economic and climatic effect on travel patterns. At the same time, these tendencies are also confirmed by the actual data.

4.2. 1 P.M. result

The eigenvalues of $\boldsymbol{B}_{day}\boldsymbol{B}^{t}_{day}$ are $\lambda_{d,1} = 583.94, \lambda_{d,2} = 104.14, \lambda_{d,3} = 60.12, \lambda_{d,4} = 46.77, \lambda_{d,5} = 23.06, \lambda_{d,6} = 14.64, \ldots$ in descending order. Therefore, as $\lambda_{d,5} = 23.06 > \lambda_{+} = 19.02 > \lambda_{d,6} = 14.64$, five features are extracted. Let the corresponding eigenvectors $\boldsymbol{p}_{d,1}, \boldsymbol{p}_{d,2}, \boldsymbol{p}_{d,3}, \boldsymbol{p}_{d,4}$ and $\boldsymbol{p}_{d,5}$. The values of their elements are shown in Fig. 3 with the coloured mesh. In addition, the loading vectors for each eigenvector are shown in Fig. 3. If the conventional method of cumulative contribution ratio up to 0.8 is applied, only three features are regarded as meaningful, meaning that two essential features are missed. In a similar way, if the method of eigenvalue larger than 1 is applied, 26 features are regarded as meaningful, meaning that 21 features which have statistically no meaning are falsely extracted.

First three components, $\boldsymbol{p}_{d,1}$, $\boldsymbol{p}_{d,2}$ and $\boldsymbol{p}_{d,3}$, are understood in almost the same way as the 3 A.M. result. One difference is that in the first component, $\boldsymbol{p}_{d,1}$, has positive elements in the intra-region trip of regions 3 and 6. This is quite consistent with the actual situation that there are many inter-zone and intra-region commuters in these regions. Therefore, compared to 3 A.M. result, the variance between weekdays and weekends is large.

The fourth and fifth components, $\boldsymbol{p}_{d,4}$ and $\boldsymbol{p}_{d,5}$, show the difference of intra- and inter-region travel behaviour as in the fourth component in the 3 A.M. result. Although it is not easy to provide clear interpretations for them, one possible explanation is as follows. $\boldsymbol{p}_{d,4}$ shows the dominant factor for each OD pair, either business or tourism, as the loading vector shows the difference between weekdays and weekends. Intra-region trips, except in regions 3 and 6, increase on weekends, meaning that many day trip activities exists. On the contrary, most inter-region travellers are on business on weekdays. $\boldsymbol{p}_{d,5}$ shows the difference in long holiday travel direction between the New Year's Holiday and early May. As Japanese tend to go back to their hometown during the New Year's Holiday and not necessarily in early May, travel destinations differ from each other. The former is decided by the nationwide population distribution, while the latter can be changed by infrastructure construction and tourism promotion.

In addition, one important implication from this result is that if we use the daytime data, commuting travel behaviour becomes more dominant than the nighttime data and even the number of features to be extracted is changed. Therefore, generally we should carefully select the data to be utilised to meet the purpose of an analysis.

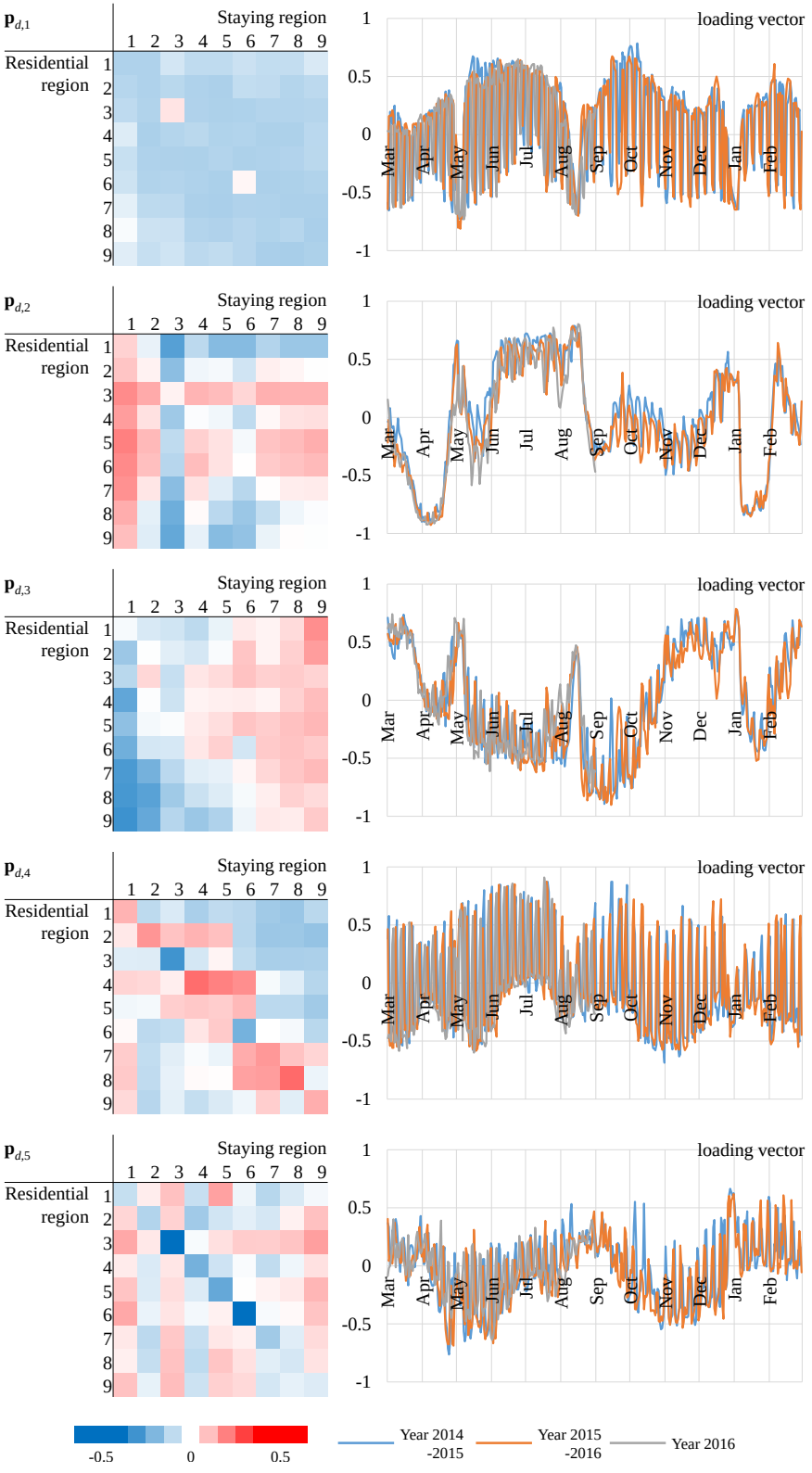

Fig. 3. Extracted features of 1 P.M. (left) eigenvector; (right) loading vector

5. Concluding remarks

In this study, an automatic feature extraction method was proposed with RMT-PCA, and its ability was demonstrated through the application to the data of long-period inter-region traveller numbers. As a result, the major tendencies that we could easily understand were extracted, and these results ensured the proposed methodology of RMT-PCA. The proposed method determines the essential information that the data contains statistically, meaning the extent of data utilisation. In concrete, some seasonal and weekly patterns as well as economic and climatic situation were revealed by the data. It is also shown that the selection of the data, whether daytime or nighttime, brings about a different result in both the number of extracted features and their interpretations.

The interpretation and validation method of extracted features should be much improved in the future. Some difficulties are 1) as RMT-PCA is just a statistical method, it cannot necessarily distinguish the self- and cross-correlation within the data and bring about practical meanings, and 2) some important information like trip purpose and trip chain (trajectory) are not included in the MSS data. At the same time, the result could be utilised in the following ways as a part of the next step. Firstly, the extracted features could be taken as explanatory variables of the time-series regression model of the original dataset. Secondly, the residual of such model could be input to the proposed RMT-PCA. This process would help us extract weaker but meaningful features. In addition, the travel patterns of extracted features could be applied to anomaly detection and change point detection problems (Takeuchi and Yamanishi, 2006). Although the result in this study seems steady and there is no significant change point in the time-series, how and when tourism promotion and infrastructure construction affect travel patterns could be understood through this type of further exploration over longer periods of data.

Acknowledgements

This work is partially supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (A) # 17H01297 and (B) # 26289171, and by the Committee on Advanced Road Technology (CART), Ministry of Land, Infrastructure, Transport, and Tourism, Japan.

References

Chen, Z., He, Z., Wang, J., 2018. Spatial-temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition, *Transportation Research Part C: Emerging Technologies*, 86, 59–77.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, 24.6, 417–441.

Laloux, L., Cizeau, P., Bouchaud, J.P., Potters, M., 1999. Noise dressing of financial correlation matrices, *Physical Review Letters*, 83, 1467–1470.

Mehta, M. L., 2004. *Random Matrices (3rd edition)*, Academic Press.

Ministry of Land Infrastructure and Transport, 2010. 2010 Inter-Regional Travel Survey in Japan, `http://www.mlit.go.jp/common/001005633.pdf`, last access Feb. 20th 2018.

Pearson, K., 1901. On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine*, 2.11, 559–572.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Stanley, H.E., 1999. Universal and nonuniversal properties of cross correlations in financial time series, *Physical Review Letters*, 83, 1471–1474.

Sun, L., Axhausen, K.W., 2016. Understanding urban mobility patterns with a probabilistic tensor factorization framework, *Transportation Research Part B: Methodological*, 91, 511–524.

Takeuchi, J., Yamanishi, K., 2006. A unifying framework for detecting outliers and change points from time series, *IEEE Transactions on Knowledge and Data Engineering*, 18.4, 482–492.

Tan, H., Feng, G., Feng, G., Wang, W., Zhang, J., Li, F., 2013. A tensor-based method for missing traffic data completion, *Transportation Research Part C: Emerging Technologies*, 28, 15–23.

Terada, M., Nagata, T., Kobayashi, M., 2013. Population Estimation Technology for Mobile Spatial Statistics. NTT DOCOMO Technical Journal, 14.3, 10–15.

Yao, D., Yu, C., Jin, H., Ding, Q., 2015. Human mobility synthesis using matrix and tensor factorizations, *Information Fusion*, 23, 25–32.