Kim *et al. J Big Data (2018) 5:5*
https://doi.org/10.1186/s40537-018-0116-9

Journal of Big Data

RESEARCH

Open Access

A new approach to the space–time analysis of big data: application to subway traffic data in Seoul

Kwang-Yul Kim[1*], Chae-Young Lim[2] and Eunice J. Kim[3]

*Correspondence:
kwang56@snu.ac.kr
[1] School of Earth and Environmental Sciences, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
Full list of author information is available at the end of the article

Abstract

A prevalent type of big data is in the form of space–time measurements. Cyclostationary empirical orthogonal function (CSEOF) analysis is introduced as an efficient and valuable technique to interpret space–time structure of variability in a big dataset. CSEOF analysis is demonstrated to be a powerful tool in understanding the space–time structure of variability, when data exhibits periodic statistics in time. As an example, CSEOF analysis is applied to the hourly passenger traffic on Subway Line #2 of Seoul, South Korea during the period of 2010–2017. The first mode represents the weekly cycle of subway passengers and captures the majority (~ 97%) of the total variability. The corresponding loading vector exhibits a typical weekly pattern of subway passengers as a function of time and the locations of subway stations. The associated principal component time series shows that there are two occasions of significant reduction in the amplitude of the weekly activity in each year; these reductions are associated with two major holidays—lunar New Year and Fall Festival (called Chuseok in Korea). The second and third modes represent daily contrasts in a week and are associated with taking extra days off before or after holidays. The fourth mode exhibits an interesting upward trend, which represents a general decrease in the number of subway passengers during weekdays except for Wednesday and an increase over the weekends.

Keywords: CSEOF analysis, Space–time analysis, Big data analysis

Introduction

Living in the digital age, we have added and accumulated much refined location information on human activities. There are examples of road and air traffic analysis for logistics and human transportation, trending keywords by locale via social network services (SNS), real-time monitoring of payment services, to name a few. Understanding space–time structures of a complex system is an important aspect of big data analysis. Also, in understanding geosciences, space–time domain serves as the basis to analyze variability, and understanding the space–time structure of variability is an important scientific goal. Examples include concentration of PM10 (particulate matter of size less than 10 μm), precipitation, CO_2 concentration, snow coverage, wind speed and direction, and solar radiation, which all exhibit strong space–time variability.

Deep learning tools are developed and implemented [1] to identify patterns from a massive amount of data, and a space–time process is one example where its nonlinear

Springer Open

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

data generating mechanism can be learned through deep learning or multi-level decomposition. There are numerous examples in which space–time structure of variability provides valuable information in understanding and interpreting data and predicting the trend. For example, understanding the space–time patterns of a transit system use is crucial for transit operation such as adjusting service intervals to reduce the chance of a disruption by a large number of passengers on board and making an improved operation plan for comfortable ridership. Smart card data have been used to analyze the diurnal patterns of travel time, number of trips, and different means of transportation in Seoul, South Korea [2]. The transit ridership in New York City has also been analyzed by using cluster analysis [3] where the clusters exhibit distinct diurnal behaviors of transit ridership. Using smart card data of several transits, the daily spatio-temporal density of passengers, the counts of waiting and onboard passengers, and the trajectory of the trains have been analyzed for Singapore [4], Kochi City, Japan [5], and Brisbane, Australia [6] by different dimensions. A thorough review on smart card data use in public transit can be found in Pelletier et al. [7]. These earlier studies made use of simple summary statistics or provided a separate analysis of spatial and temporal variability. In order to understand space–time structure of variability, cluster analysis and principal component analysis (PCA) have been used in Morency et al. [8, 9] and Tsekeriset al. [10], respectively. Xing et al. [11] used the robust PCA technique to decompose highway traffic data into low-rank traffic matrix plus several residual traffic matrices.

While straightforward statistical analysis and conventional spatial analysis serve as useful tools, these analysis techniques are certainly limited in explaining the combined space and time structure. Often space–time variability has a specific direction of evolution. For example, traffic flow exhibits a specific direction in space and time. Namely, variability of traffic flow varies as a function of space and time. Therefore, it is necessary to examine spatial and temporal variability simultaneously in order to understand the direction of variability evolving in space and time. Conventional data analysis techniques in space or in time are not designed to resolve space–time variability together with its directivity, and separate analysis of spatial variability and temporal variability often leads to misleading interpretations of the nature of space–time variability. Analysis based on the assumption of cyclostationarity [12] is becoming a new research trend in climatology, electrical engineering, and signal processing [13].

In this study, the Cyclostationary Empirical Orthogonal Function (CSEOF) technique is introduced as an effective space–time analysis tool [14–16]. The CSEOF technique is applied to Seoul subway passenger data for Subway Line #2 for the purpose of analyzing the space–time variability of the subway passengers. The description of the data and the method of analysis are presented in "Data" and "Methods" section. The results of analysis are presented in "Results" section followed by discussion and concluding remarks in "Discussion" and "Conclusion".

Data

The data used represent hourly subway passengers archived by Seoul Metro Company (http://seoulmetro.co.kr); they are stored in excel (or CSV) files. The dataset contains the number of hourly boarding and alighting passengers for Subway Line #2 (also called the Green Line) in Seoul, Korea for the period of January 2010–March 2017. Subway Line

#2 has 50 stations managed by Seoul Metro Company. Therefore, the total size of the data analyzed is 50 stations × 24 h × 2647 days for both boarding and alighting passengers. We analyze the spatio-temporal structure of boarding and alighting passengers in order to understand the typical weekly variation of the number of subway passengers as a function of time and space (station).

Methods

In the present study, we are specifically interested in identifying space–time eigenfunctions of the subway passenger data, i.e., distinct modes of variability of subway passengers as a function of subway stations and times of the week. Since the statistical properties (mean and covariance) of the subway passenger data are nearly periodic with a distinct weekly cycle, we assume that the data fall under the category of cyclostationary random variables [12]. Thus, a space–time analysis technique called Cyclostationary Empirical Orthogonal Function (CSEOF) Analysis [14–16] is employed to identify eigenfunctions of subway passenger data.

Subway passenger data $T(r, t)$ is decomposed in terms of eigenfunctions as

$$T(r,t) = \sum_{n} B_n(r,t) T_n(t), \tag{1}$$

where the eigenfunctions, $B_n(r, t)$, are called cyclostationary loading vectors (CSLV) and the corresponding amplitude, $T_n(t)$, are called principal component (PC) time series. Because of the periodic nature of the statistics, CSLVs are also periodic, i.e.,

$$B_n(r,t) = B_n(r,t+d), \tag{2}$$

where d is a phenomenological parameter called the "nested period", which represents the periodicity of the underlying covariance statistics. That is,

$$C\left(r,t;r',t'\right) = E\left(T(r,t)T\left(r',t'\right)\right) = C\left(r,t+d;r',t'+d\right), \tag{3}$$

where $E(\cdot)$ stands for expectation. Equation (3) measures common variability shared by two points separated in space and time. We specifically assume that covariance function is periodic in time with the periodicity d: in the present study, $d = 168$ h (24 h × 7 days). This periodicity implies that behavioral patterns of subway passengers are strongly linked to the weekly cycle.

A specific requirement for $B_n(r, t)$ is that they are mutually orthogonal, i.e.,

$$(B_n(r,t) \cdot B_m(r,t)) = \frac{1}{Nd} \sum_{r=1}^{N} \sum_{t=1}^{d} B_n(r,t) B_m(r,t) = \lambda_n \delta_{nm}, \tag{4}$$

where $(A \cdot B)$ denotes dot (inner) product between A and B, N is the number of spatial points, λ_n represents the variance explained by mode n, and δ_{nm} is the Kronecker delta. A requirement for $T_n(t)$ is that they are mutually uncorrelated, that is,

$$(T_n(t) \cdot T_m(t)) = \frac{1}{M} \sum_{t=1}^{M} T_n(t) T_m(t) = \delta_{nm}, \tag{5}$$

where M is the number of temporal points. Thus, CSLVs are interpreted as mutually orthogonal space–time evolutions in data, of which the amplitude time series are mutually uncorrelated.

Equation (1) is our analytical model. A crucial motivation behind (1) is that we want to decompose variability of the subway passengers at 50 stations into weekly patterns that are mutually orthogonal and uncorrelated. Thus, each $B_n(r, t)$ describes a distinct weekly variation of subway passengers at 50 stations, and corresponding $T_n(t)$ describes its long-term amplitude change. In this way, different causes of passenger variability throughout the week can be investigated and understood. Therefore, a specific goal of the present study is to identify CSLVs and PC time series in (1) with the requirements of orthogonality of CSLVs and uncorrelatedness of PC time series as stipulated in (4) and (5).

Note that our analytical model (1) differs from its predecessor based on EOF analysis [17, 18]. In EOF analysis, subway passenger data is decomposed in the form

$$T(r,t) = \sum_n \phi_n(r) P_n(t), \tag{6}$$

where $\phi_n(r)$ are EOF loading vectors and $P_n(t)$ are corresponding PC time series. Unlike CSLVs, EOF loading vectors are functions only of space. In fact, EOF analysis is a special case of CSEOF analysis with the nested period of 1 ($d = 1$). Limitation, therefore, is inherent in delineating temporally evolving structures of variability in terms of EOF loading vector. Specifically, a temporally varying structure of subway passenger variability cannot be captured by a single EOF loading vector; several EOF loading vectors are required to describe it. Thus, the EOF analysis technique, based on the stationarity assumption [12], is not suited for describing subway passenger variability, which exhibits both strong diurnal and weekly cycles.

In order to obtain CSLVs and PC time series in (1), we used the algorithm called the CSEOF analysis technique [14–16]. In essence, CSLVs are obtained by solving

$$C\left(r,t;r',t'\right)B_n\left(r',t'\right) = \lambda_n B_n(r,t), \tag{7}$$

which is called the Karhunen–Loève equation [19]. A detailed solution procedure and more in-depth discussion on the CSEOF analysis technique can be found in Kim et al. [14, 16] and Kim and North [15] and will not be addressed in the present study. We have implemented the method in three different languages: FORTRAN, IDL (Interactive Data Language) and MATLAB (Matrix Laboratory).[1] The procedure for analysis is schematically depicted in Fig. 1.

Results

Seoul is a metropolis with a population of 10 million people and a density of 17,000 people km^{-2} according to the Seoul Metropolitan Government (http://stat.seoul.go.kr). Also, a large number of people commute to Seoul from the suburbs of Seoul. Subway Line #2, also called the Green Line, is one of the busiest lines among the 9 metropolitan subway lines and 4 special subway lines in Seoul (see Fig. 2). The total number of daily

---

[1] The memory required for running the algorithm is approximately 10 times the size of the covariance matrix. In the present study, the covariance matrix of size $(50 \times 24 \times 7) \times (50 \times 24 \times 7)$ required 2.8 GB memory.

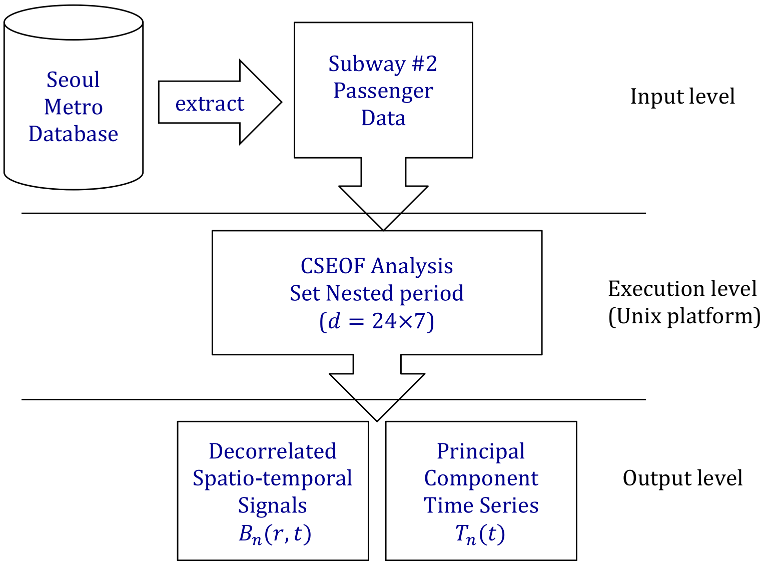

Fig. 1 A schematic diagram showing the analysis procedure for the Subway Line #2 passenger data

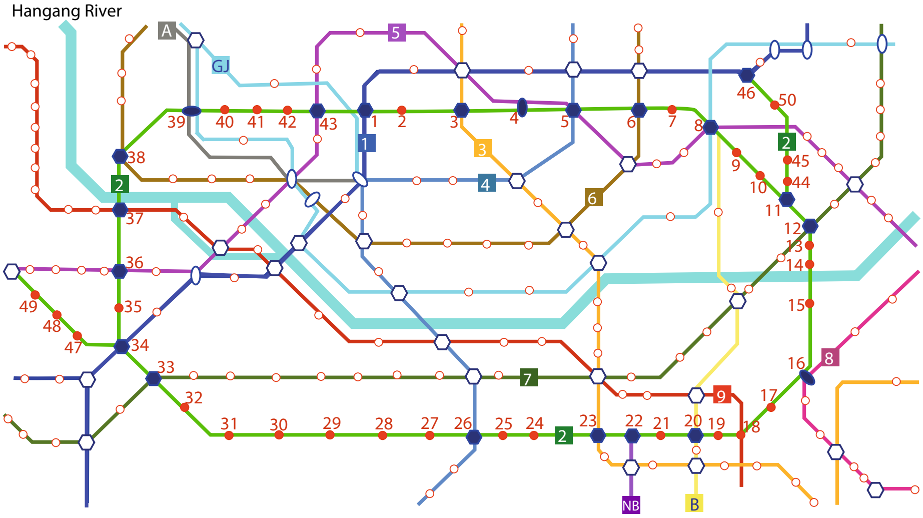

Fig. 2 Sketch of the nine subway lines in Seoul. The oval and hexagonal symbols denote subway transfer stations and circle symbols represent regular stations. The filled symbols apply to subway Line #2 (Green Line). The numbers in red denote station numbers for subway Line #2

passengers for Line #2 is over 2 million, which represents 31% of all subway passengers in Seoul. Figure 2 shows subway Line #2 with its 50 stations and all the connecting subways. There is a significant weekly cycle in the passenger data. During weekdays, a large number of people in Seoul use the subway to commute. During weekends, the number of passengers decreases significantly. Based on the common pattern, we set the nested period to be 1 week (168 h) in the present study.

Figure 3 shows the first CSEOF mode of hourly passengers for the 50 stations of Subway Line #2 in a week. Figure 3 from top to bottom shows the CSLVs of boarding

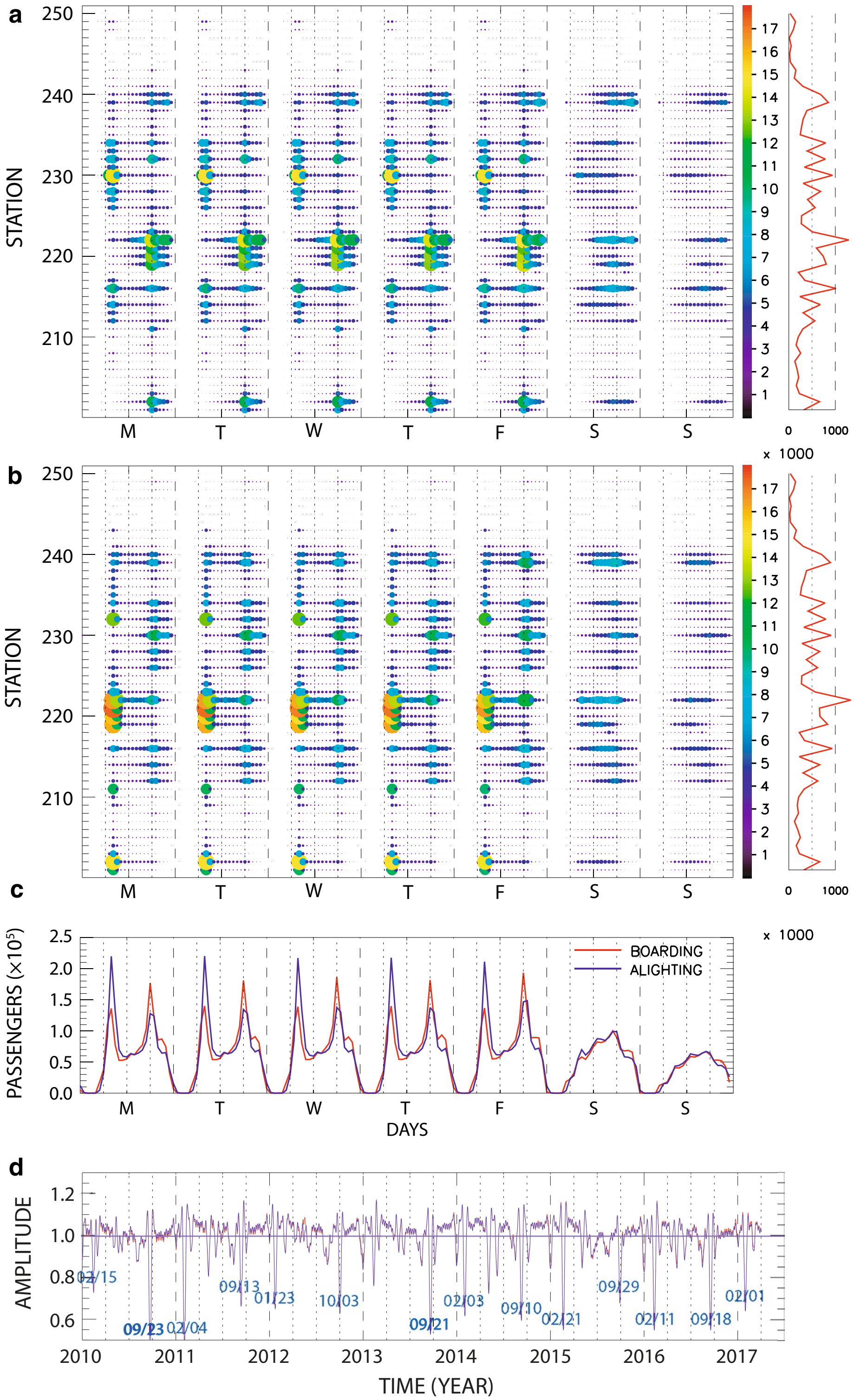

Fig. 3 The first CSEOF mode representing the weekly patterns of passengers for Subway Line #2: **a** hourly and total number of boarding passengers at each of the 50 subway stations numbered from 201 to 250; **b** hourly and total number of alighting passengers; **c** number of boarding (red) and alighting passengers in all stations; and **d** corresponding amplitude time series

passengers (Fig. 3a), alighting passengers (Fig. 3b), passengers added for all the stations (Fig. 3c), and the corresponding PC time series (Fig. 3d). This mode explains ~ 97% of the total variability.

The total boarding and alighting passengers respectively show two distinctive peaks that correspond to commuting times during weekdays (Fig. 3c); the morning peak is at around 7–8 a.m., and the evening peak is at around 6–7 p.m. The morning peak is larger for alighting passengers, whereas the evening peak is larger for boarding passengers. During 7–8 a.m., there are a large number of passengers boarding subways near Station 230 (Shilim). Around 6 p.m., there are also a large number of boarding passengers near Station 222 (Kangnam). In the morning (7–8 a.m.), a large number of alighting passengers are seen near Station 222. In the evening, however, the alighting pattern does not show any significantly crowded stations. It is not surprising that there is a large volume of passengers boarding near Station 230 in the morning because other subway lines, except for #2, are inconvenient to access and because there is a large community with a high density of population in this area. Stations 219–222 are located in the business center of the city with many companies. Note that Bundang Line (denoted as B in Fig. 2) and New Bundang Line (NB in Fig. 2) are connected with Subway Line #2 at Station 220 and Station 222, respectively. Alighting passengers from the Bundang Lines often use Stations 220 and 222 for exiting. Thus, there are large numbers of alighting passengers at these stations in the morning.

During weekends, the number of passengers is reduced and the morning and evening peaks are not seen. The total number of daily passengers is 3.0–3.2 million during weekdays and is reduced to ~ 2.4 million (78%) and ~ 1.7 million (54%) on Saturday and Sunday, respectively (Table 1).

Figure 4 shows the difference in the number between the alighting and boarding passengers. It is clear that the difference is more noticeable in the morning hours than in the evening. When averaged over 24 h, this imbalance in the number of boarding and alighting passengers is most pronounced at Stations 221–223; number of alighting passengers exceeds that of boarding passengers during weekdays (see also Table 1). The reason for this imbalance is not clear. One plausible explanation is that different modes of transportation are readily available in the evening, when commuting hours are more wide spread than in the morning. In the rest of the stations there is a tighter match in the number of boarding and alighting passengers.

Table 1 Daily total number of passengers (in units of million people) for Subway Line #2 in Seoul

	Boarding	Alighting	Total
Monday	1.498	1.514	3.012
Tuesday	1.534	1.550	3.084
Wednesday	1.537	1.555	3.092
Thursday	1.550	1.568	3.118
Friday	1.601	1.629	3.231
Saturday	1.210	1.232	2.442
Sunday	0.849	0.842	1.682
Total	9.771	9.890	19.661

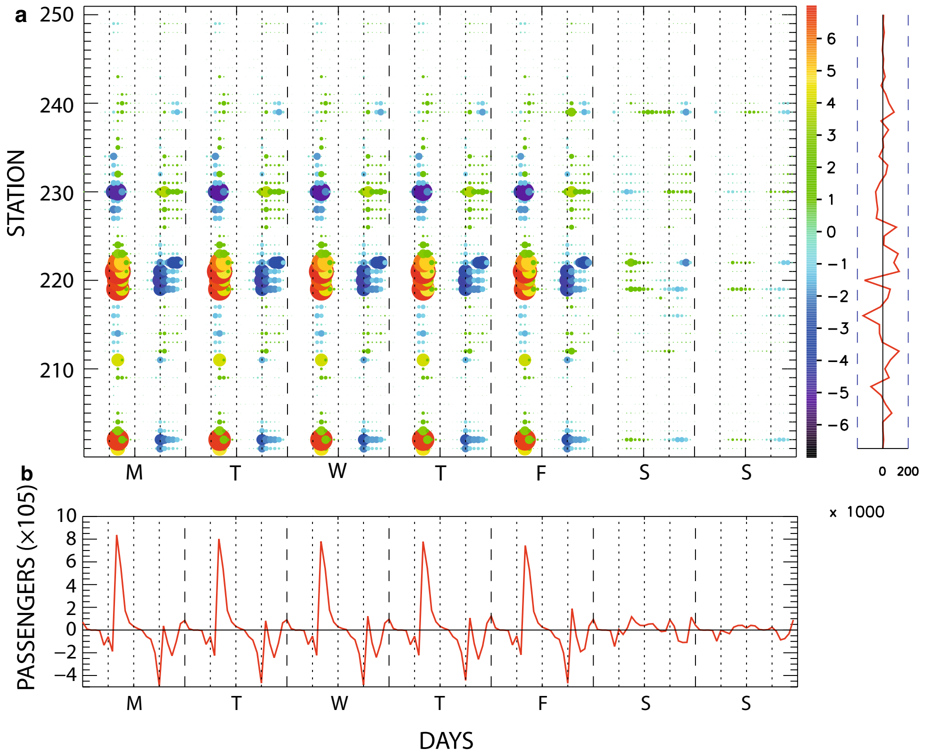

Fig. 4 The weekly patterns of the number of alighting minus boarding passengers for Subway Line #2: **a** hourly and total number of alighting minus boarding passengers at each of the 50 subway stations; **b** number of alighting minus boarding passengers in all stations

The corresponding PC time series shows that there is no noticeable trend in the number of passengers for Subway Line #2 (Fig. 3d). The amplitude of this weekly cycle fluctuates around the mean value of ~ 1.0 with a standard deviation of 0.094. The histogram is highly skewed with a long tail toward the low end (Fig. 5). There are a few occasions of significant reduction in the number of passengers in each year. The dates of reduction

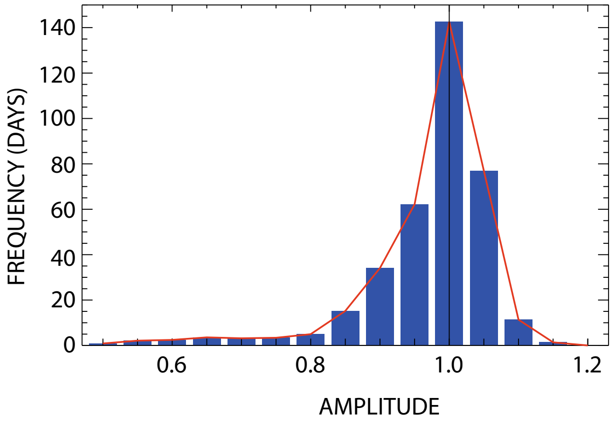

Fig. 5 Histogram of the first PC time series of the boarding passengers. The histogram is highly skewed with a long tail toward the low end

are shown in Table 2. Lunar New Year, in the early part of the calendar, and Chuseok or Mid-Autumn Festival, in September or October, are two major holidays in Korea. They each last for 3–4 days, and people in metropolitan areas, especially those working in Seoul visit their family living in the countryside. There is a general decline in the number of passengers around the New Year's Day.

Figure 6 shows the monthly averaged amplitudes of the PC time series of the weekly cycle over the record period. The monthly averaged amplitude, in general, varies by less than 10% except in September 2010. On average, the number of monthly passengers peaks in March and to a lesser extent in November. On average, the number of passengers decreases most significantly in September and to a less extent in August and February. The reduction of passengers in September is evidently associated with Chuseok, one

Table 2 Day (of the week) of the major peaks in the PC time series of the first CSEOF mode and the linked holidays

	Negative peaks		Remarks
	Peak day	**Holidays**	
1	10/09/23 (Thu)	09/21–23 (Tue–Thu)	Chuseok
2	11/02/04 (Fri)	02/02–04 (Wed–Fri)	Lunar New Year
3	13/09/21 (Sat)	09/18–20 (Wed–Fri)	Chuseok
4	16/09/18 (Sun)	09/14–18 (Wed–Sun)	Chuseok
5	16/02/11 (Thu)	02/06–10 (Sat–Wed)	Lunar New Year
6	15/02/21 (Sat)	02/18–20 (Wed–Fri)	Lunar New Year
7	14/09/10 (Wed)	09/07–10 (Sun–Wed)	Chuseok
8	14/02/03 (Mon)	01/30–02/01 (Thu–Sat)	Lunar New Year
9	12/10/03 (Wed)	09/29–10/01 (Sat–Mon)	Chuseok
10	17/02/01 (Tue)	01/27–01/30 (Fri–Mon)	Lunar New Year
11	12/01/23 (Mon)	01/22–24 (Sun–Tue)	Lunar New Year
12	11/09/13 (Tue)	09/11–13 (Sun–Tue)	Chuseok
13	15/09/29 (Tue)	09/26–29 (Sat–Tue)	Chuseok
14	10/02/15 (Mon)	02/13–15 (Sat–Mon)	Lunar New Year

These dates are also marked in Fig. 2d

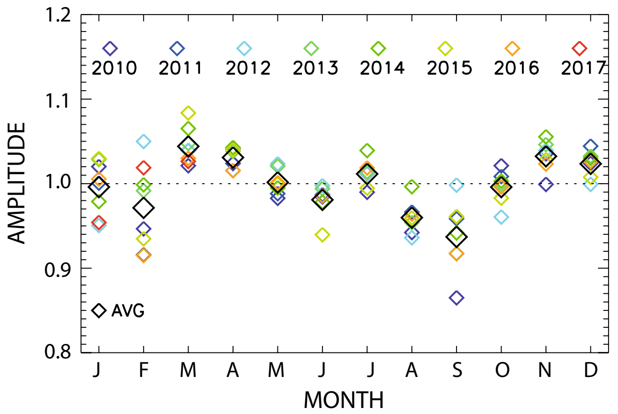

Fig. 6 The monthly averaged amplitude of the normalized PC time series of the weekly cycle during the data period. The black diamonds represent the averages of the monthly averaged amplitudes

of the two main holidays for family gathering. The reduction of passengers in February is not so dramatic as in September, since the lunar New Year was in January in 2012, 2014 and 2017, and the other years in February. This is reflected in the above normal or nearly normal amplitude of the weekly cycles in these 3 years, whereas the other years generally show a conspicuous reduction in the number of passengers in February. August is a typical vacation period in Seoul. There is an obvious reduction in the amplitude of the weekly cycle in August except in 2014. While variability of each monthly averaged amplitude of the weekly cycle is fairly sizable, there is a noticeable monthly trend in the number of subway passengers at heightened levels in March, April, November and December and at damped levels in February, August and September.

The second CSEOF mode, explaining 0.67% of the total variability, describes a reduction of passengers on Mondays and Tuesdays and an increase on Thursdays and Fridays when the amplitude is positive (Fig. 7). The PC time series exhibits several prominent positive and negative peaks (Table 3). The positive peaks are, in general, associated with a reduction of passengers due to holidays and an increase of passengers after holidays (Table 3). The negative peaks, in general, represent an increase of passengers before and a decrease after holidays (Table 3). This mode reflects a characteristic behavior of taking extra days off before or after holidays.

The third CSEOF mode explains 0.40% of the total variability and describes a reduction of passengers on Wednesday and an increase on Friday when its phase is positive (Fig. 8). The positive peaks are generally associated with an increase of passengers after holidays and a reduction of passengers prior to holidays (Table 4). The negative peaks are associated with a reduction of passengers after holidays and an increase of passengers before holidays (Table 4). This mode also reflects the behavior of taking extra days off before or after holidays. It should be noted that positive peaks are more prominent and frequent.

The fourth CSEOF mode, explaining 0.28% of the total variability, exhibits an interesting trend. As seen in Fig. 9, the number of passengers keeps decreasing during the morning commuting hours on weekdays except for Wednesday. On the other hand, the number of passengers has increased after the morning commuting hours until the evening commuting hours (Fig. 9c). This seems to indicate that the number of people using the subway for commuting is decreasing and more people use it outside the rush hours. Note that the stations with heavy traffic in the weekly cycle (Fig. 3) are most significantly affected by this trend as should be expected. On Wednesdays, the number of boarding and alighting passengers has increased during evening hours (Fig. 9c). During weekends, the number of passengers has increased slightly except in the morning. It is of interest to note that there is a notably elevated number of boarding passengers during evening hours at Station 239 throughout the week. This station, called Hongik University Station, is a popular tourist attraction for young people with many street venders, shops, restaurants, bars and music cafés.

While the trend in the corresponding amplitude time series is independent of the weekly cycle, the detrended PC time series is rather highly correlated (corr $= -0.57$) with the PC time series of the weekly cycle (Fig. 10). This negative correlation implies that the decrease in ridership becomes more apparent when the pattern of weekly cycle

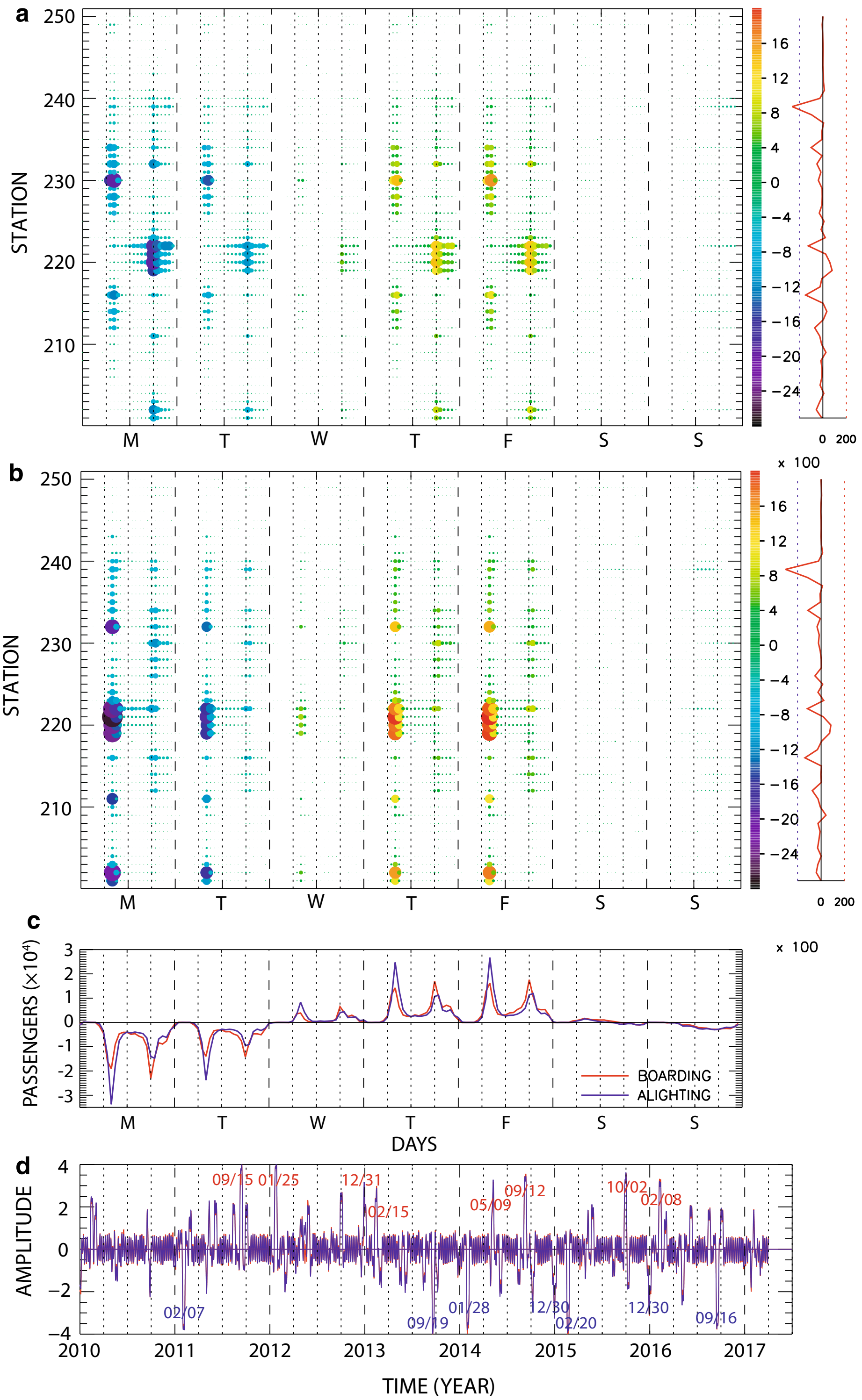

Fig. 7 The second mode of CSEOF representing the weekly patterns of passengers for Subway Line #2. **a** hourly and total number of boarding passengers at each of the 50 subway stations numbered from 201 to 250; **b** hourly and total number of alighting passengers; **c** number of boarding (red) and alighting passengers in all stations; and **d** corresponding amplitude time series. Four sets of stations show regular patterns during weekdays

Table 3 Day (of the week) of the major peaks in the PC time series of the second CSEOF mode and the linked holidays

	Positive peaks Reduction (Monday and Tuesday) Increase (Thursday and Friday)		Remarks
	Peak day	**Holidays**	
1	12/01/25 (Wed)	01/22–24 (Sun–Tue)	Lunar New Year
2	11/09/15 (Thu)	09/11–13 (Sun–Tue)	Chuseok
3	15/10/02 (Fri)	09/27–29 (Sun–Tue)	Chuseok
4	14/09/12 (Fri)	09/07–10 (Sun–Wed)	Chuseok
5	16/02/08 (Mon)	02/06–10 (Sat–Wed)	Lunar New Year
5	14/05/09 (Fri)	05/05–06 (Mon–Tue)	Children's Day/Budah Birthday
6	13/02/15 (Fri)	02/09–11 (Sun–Mon)	Lunar New Year
7	12/12/31 (Mon)	13/01/01 (Tue)	New Year Day

	Negative peaks Increase (Monday and Tuesday) Reduction (Thursday and Friday)		Remarks
	Peak day	**Holidays**	
1	13/09/19 (Thu)	09/18–20 (Wed–Fri)	Chuseok
2	15/02/20 (Fri)	02/18–20 (Wed–Fri)	Lunar New Year
3	14/01/28 (Tue)	01/30–02/01 (Thu–Sat)	Lunar New Year
4	11/02/07 (Mon)	02/02–04 (Wed–Fri)	Lunar New Year
5	16/09/16 (Fri)	09/14–18 (Wed–Sun)	Chuseok
6	14/12/30 (Tue)	15/01/01 (Thu)	New Year Day
7	15/12/30 (Mon)	16/01/01 (Fri)	New Year Day

These dates are marked in Fig. 7d

is subdued during holiday periods. In other words, there is a larger reduction in commuting during holidays.

These four CSEOF modes together explain 98.6% of the total variability of the Subway Line #2 passengers during the period of January 2010–March 2017. These four modes represent independent modes and allow reasonable explanation for the variability of boarding and alighting passengers for Subway Line #2.

Discussion

As demonstrated in this study, the CSEOF technique facilitates the analysis of complex data in terms of identifying main patterns and exploring the causes of space–time variability. There are some studies that analyze public transit data in Seoul [20, 21], but no study, as far as we know, has yet investigated spatial and temporal variability simultaneously. CSEOF analysis clearly reveals subway passenger variability as a function of location and time of the week. Not only so, the seasonal and long-term trends and erratic behaviors associated with national holidays are faithfully captured in the variation of the amplitude (PC) time series. Such information should be useful in predicting subway passengers and planning efficient train allocation schedules.

It should be pointed out that the regular PCA analysis extracts the time-averaged patterns of subway passengers on the right-hand side of Fig. 3a, b as the first mode (see Fig. 11). In order to explain temporal variability throughout the day, which differs

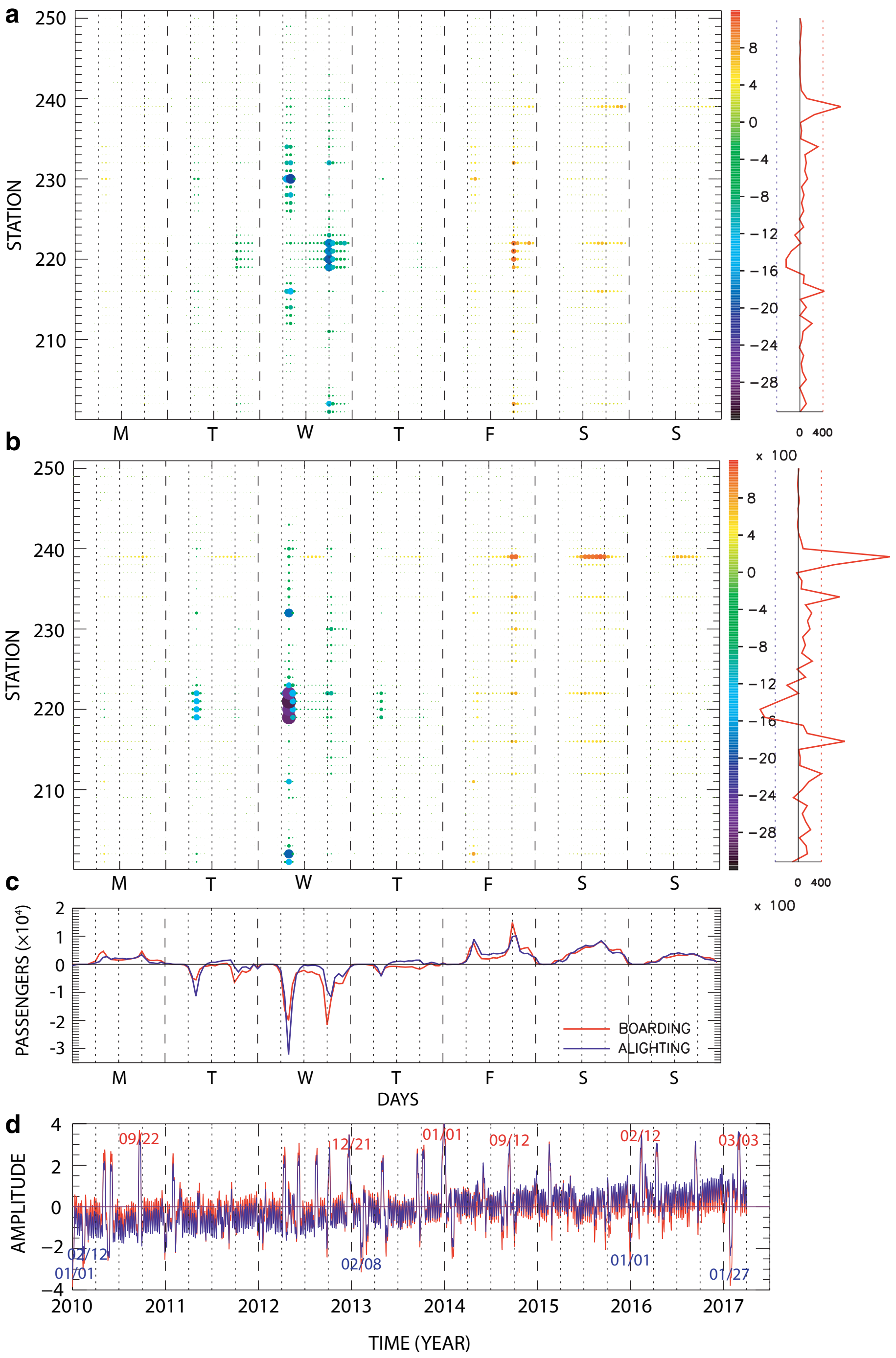

Fig. 8 The third mode of CSEOF representing the weekly patterns of passengers for Subway Line #2: **a** hourly and total number of boarding passengers at each of the 50 subway stations numbered from 201 to 250; **b** hourly and total number of alighting passengers; **c** number of boarding (red) and alighting passengers in all stations; and **d** corresponding amplitude time series. Heightened activities are noted on Wednesdays complementing the second CSEOF mode

from one station to another, several EOF modes are needed. For instance, the difference between the weekdays and weekends is partially resolved in the PC time series of the first EOF mode. Then, diurnal contrasts, distinct patterns at different stations, and

Table 4 Day (of the week) of the major peaks in the PC time series of the third CSEOF mode and the linked holidays

	Positive peaks Reduction (Wednesday) Increase (Friday)		Remarks
	Peak day	**Holidays**	
1	14/01/01 (Wed)	01/01 (Wed)	New Year Day
2	10/09/22 (Wed)	09/21–23 (Tue–Thu)	Chuseok
3	12/12/21 (Fri)	12/25 (Tue)	Christmas
4	16/02/12 (Fri)	02/06–10 (Sat–Wed)	Lunar New Year
5	17/03/03 (Fri)	17/03/01 (Wed)	March 1 Day
6	14/09/12 (Fri)	09/07–10 (Sun–Wed)	Chuseok
	Negative peaks Increase (Wednesday) Reduction (Friday)		**Remarks**
	Peak day	**Holidays**	
1	10/01/01 (Fri)	01/01 (Fri)	New Year Day
2	17/01/27 (Fri)	01/27–30 (Fri–Mon)	Lunar New Year
3	13/02/08 (Fri)	02/09–11 (Sat–Mon)	Lunar New Year
4	16/01/01 (Fri)	01/01 (Fri)	New Year Day
5	10/02/12 (Fri)	02/13–15 (Sat–Mon)	Lunar New Year

These dates are marked in Fig. 8d

detailed behavioral contrasts between weekend and weekdays are reflected in higher EOF modes (Fig. 11). As detailed, a PCA analysis requires several EOF modes to resolve the weekly cycle of subway passengers. Furthermore, the entire PC time series is fairly complicated with many peaks so that it is difficult to understand any peculiar features in the amplitude of the weekly cycle such as ridership reductions during major holidays.

We also used the Extended EOF and Periodically Extended EOF techniques [22] to analyze the same dataset (results not shown). A comparison between different Eigen techniques is beyond the scope of the present study. These techniques are developed for the purpose of examining spatio-temporal structures of variability as well. Unlike CSEOF analysis, these algorithms are not based on the cyclostationarity assumption, and the rendition of space–time covariance function is suboptimal in these techniques [22]. As a result, loading vectors and PC time series, particularly for higher modes, are often difficult to interpret or misleading.

It is clear that the weekly pattern of subway passengers differs from one station to another (Fig. 3). Thus, it is difficult to capture the complicated weekly pattern of subway passengers in terms of a few clusters. Undoubtedly, a large number of clusters are needed to capture the entire space–time variability as reflected in Fig. 3. Further, cluster analysis does not provide any information on the amplitude variation of clusters in time.

It should be pointed out that the CSEOF PC time series serves as instrumental information for predicting subway passengers. While loading vectors primarily represent independent patterns of variability repeating with the nested period, PC time series describe the amplitudes of corresponding loading vectors over the record period. Thus,

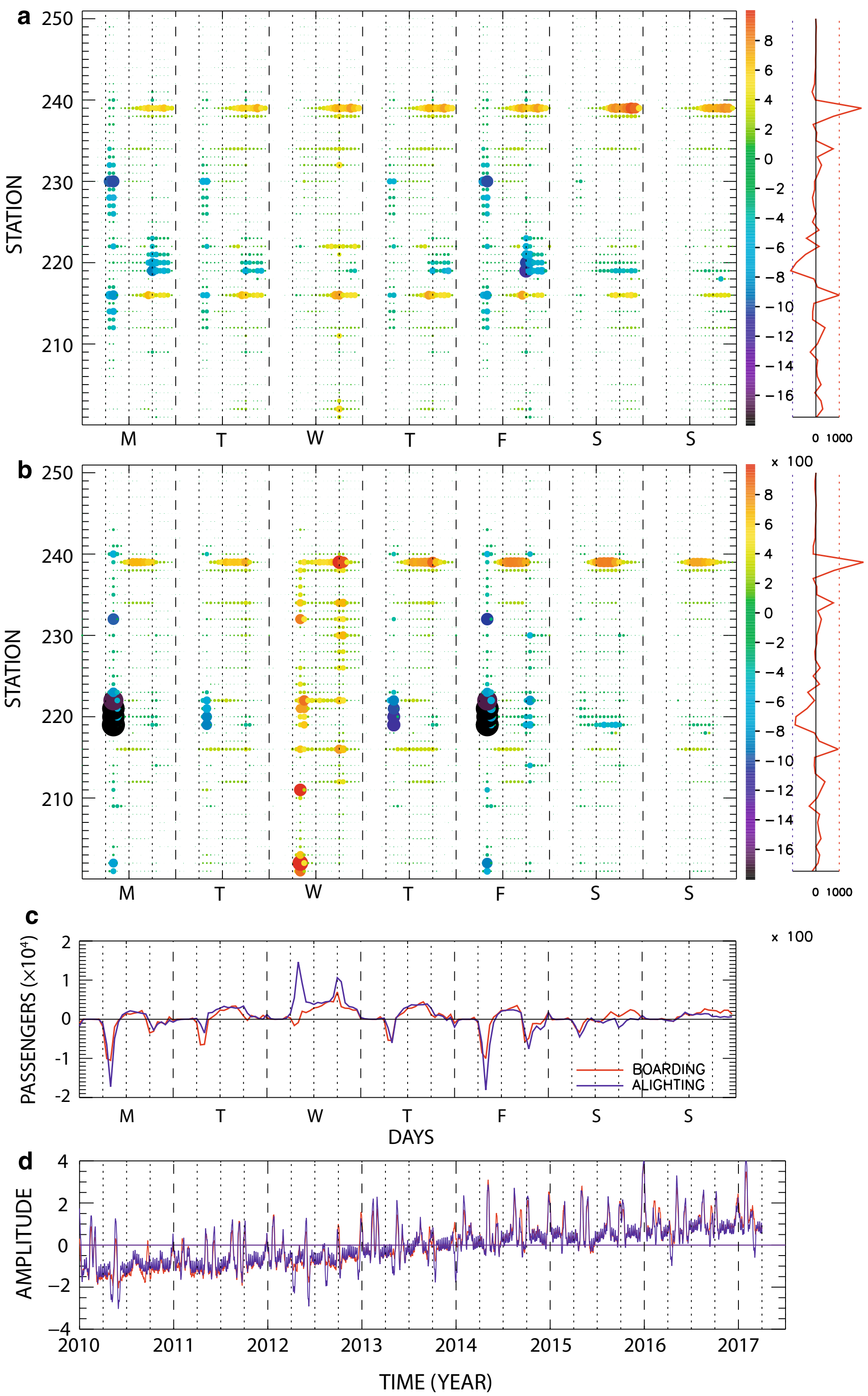

Fig. 9 The fourth mode of CSEOF representing the weekly patterns of passengers for Subway Line #2: **a** hourly and total number of boarding passengers at each of the 50 subway stations numbered from 201 to 250; **b** hourly and total number of alighting passengers; **c** number of boarding (red) and alighting passengers in all stations; and **d** corresponding amplitude time series. There are two major clusters and several distinct locations such as Station 239 and stations nearby Station 220 are surfaced in this mode throughout the week

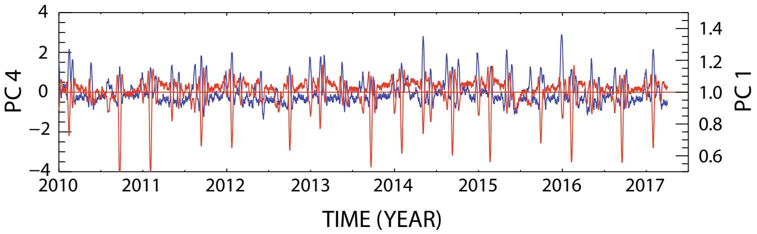

Fig. 10 Correlation between the PC time series of CSEOF mode 1 and the detrended time series of CSEOF mode 4. The two PC time series exhibits a significant negative correlation (— 0.57)

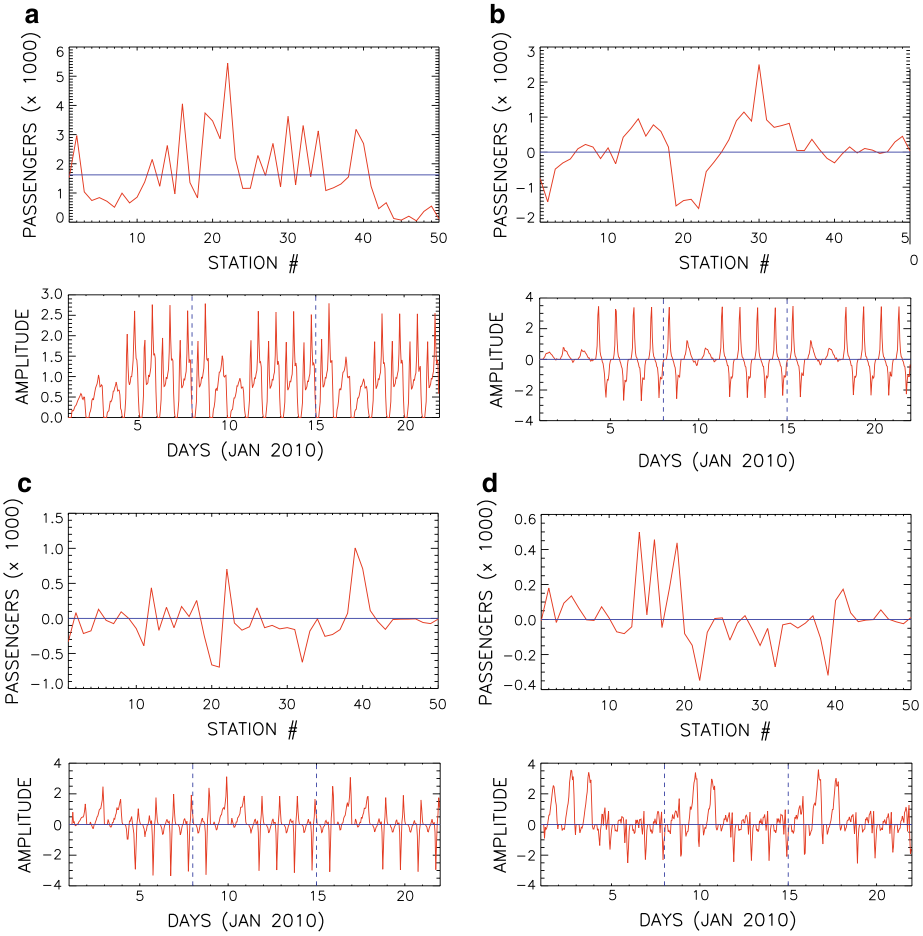

Fig. 11 The loading vector (upper panel) and PC time series (lower panel) of the first four (**a**–**d**) EOF modes

a forecast of the number of boarding and alighting subway passengers is possible by estimating the future amplitudes of each CSEOF mode.

While CSEOF analysis is capable of identifying flow patterns in space and time, the flow pattern analysis was not carried out because of the absence of the origin-and-destination information per passenger. The data used in the present study are not at a passenger level but rather at a station level, which would not distinguish the in-bound (clockwise) and out-bound (counterclockwise) flows.

Given the number of in-bound and out-bound passengers as a function of time, a flow map can easily be produced via a CSEOF analysis. This may allow us to provide a more interesting analysis of subway ridership at the passenger level rather than the station level.

Finally, one issue in dealing with big data is a reduction of dimensions [23, 24]. The CSEOF analysis has reduced the facets of spatio-temporal variability; majority (~ 99%) of subway passenger variability is nicely summarized in terms of the first four CSEOF modes. While the CSEOF loading vectors depict space–time patterns of variability, PC time series show how the amplitudes of corresponding loading vectors vary on a long-term basis. Therefore, it is much more advantageous and efficient to use CSEOF PC time series for developing any classification or prediction algorithms instead of using raw data.

Conclusion

Extracting useful information from complex or unstructured data is a key issue in big data analysis and an important goal of deep learning. The CSEOF technique is an excellent tool for analyzing variability in space–time or multi-dimensional data. Although its main applications are in climatological and geophysical sciences, it can be applied to a wide spectrum of data including trading, traffic, and communication to name a few. As long as the process exhibits periodic nature in the data generating mechanism, distinct modes of space–time variability can be extracted easily. The code runs on a Unix-based (personal) computer.

As an example, we extracted space–time structures (modes) of variability from a subway passenger dataset by applying CSEOF analysis and interpreted each mode of variability based on the structure of the loading vector and the corresponding amplitude time series. We demonstrated that CSEOF analysis is useful in understanding the variability in space–time process data with periodicity in time, which is an important technical contribution of the present study.

Authors' contributions

KYK conceived the research in the present study and conducted data research in this study. CYL participated in writing the manuscript and provided scientific feedback on the interpretation of the outputs. EJK participated in writing/editing the manuscript and provided scientific feedback on the interpretation of the outputs. All authors read and approved the final manuscript.

Author details

[1] School of Earth and Environmental Sciences, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea. [2] Department of Statistics, Seoul National University, Seoul 08826, Republic of Korea. [3] iCIMS, 101 Crawfords Corner, Holmdel, NJ 07733, USA.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. 2016R1A2B4008237).

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The data used for this work are freely available at Seoul Metro Company (http://seoulmetro.co.kr). The data analysis programs used in this study can be acquired by contacting the corresponding author (kwang56@snu.ac.kr).

Consent for publication

We content to the publication of this paper once it is accepted. This study is original and has not been published elsewhere. We will submit the copyright form once the paper is accepted.

Ethics approval and consent to participate

We strictly adhere to the ethical standard of the journal and consent to the participation in the journal's policy of ethical standard.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 17 November 2017 Accepted: 22 January 2018
Published online: 07 February 2018

References

1. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. J Big Data. 2015;2:1. https://doi.org/10.1186/s40537-014-0007-7.
2. Park JY, Kim D-J, Lim Y. Use of smart card data to define public transit use in Seoul, South Korea. Trans Res Rec. 2008;2063:3–9.
3. Chen C, Chen J, Barry J. Diurnal pattern of transit ridership: a case study of the New York City subway system. J Transp Geogr. 2009;17(3):176–86.
4. Sun L, Lee D-H, Erath A, Huang X. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. In: Proceedings of the ACM SIGKDD international workshop on urban computing. New York City: ACM; 2012. p. 142–8.
5. Nishiuchi H, King J, Todoroki T. Spatial–temporal daily frequent trip pattern of public transport passengers using smart card data. Int J Intell Transp Syst Res. 2013;11(1):1–10.
6. Tao S, Rhode D, Corcoran D. Examining the spatial–temporal dynamics of bus passenger travel behavior using smart card data and the flow-comap. J Transp Geogr. 2014;41:21–36.
7. Pelletier MP, Trépanier M, Morency C. Smart card data use in public transit: a literature review. Transp Res Part C. 2011;19:557–68.
8. Morency C, Trépanier M, Agard B. Analysing the variability of transit users behavior with smart card data. In: Proceedings of 2006 IEEE intelligent transportation systems conference. New York: IEEE; 2006. p. 44–9.
9. Morency C, Trépanier M, Agard B. Measuring transit use variability with smart-card data. Transp Policy. 2007;14:193–203.
10. Tsekeris T, Stathopoulos A. Measuring variability in urban traffic flow by use of principal component analysis. J Transp Stat. 2006;9(1):49.
11. Xing X, Zhou X, Hong H, Huang W, Bian K, Xie K. Traffic flow decomposition and prediction based on robust principal component analysis. In: IEEE 18th international conference on intelligent transportation systems. New York: IEEE; 2015. p. 2219–24.
12. Newton HJ. TIMESLAB: a time series analysis laboratory. Pacific Grove: Wadsworth and Brooks/Cole; 1988.
13. Napolitano A. Cyclostationarity: new trends and applications. Signal Process. 2016;120:385–408.
14. Kim K-Y, North GR, Huang J. EOFs of one-dimensional cyclostationary time series: computations, examples and stochastic modeling. J Atmos Sci. 1996;53:1007–17.
15. Kim KY, North GR. EOFs of harmonizable cyclostationary processes. J Atmos Sci. 1997;54:2416–27.
16. Kim KY, Hamlington BD, Na H. Theoretical foundation of cyclostationary EOF analysis for geophysical and climatic variables: concepts and examples. Earth Sci Rev. 2015;150:201–18.
17. Lorenz EN. Empirical orthogonal functions and statistical weather prediction, Statistical Forecasting Project Rep. 1. MIT Department of Meteorology. 1956.
18. Hannachi A, Joliffe I, Stephenson D. Empirical orthogonal functions and related techniques in atmospheric science: a review. Int J Climatol. 2007;27:1119–52.
19. Loève M. Probability theory II. 4th ed. Berlin: Springer; 1978.
20. Lee K, Jung W-S, Park JS, Choi MY. Statistical analysis of the Metropolitan Seoul Subway System: network structure and passenger flows. Physica A. 2008;387:24. https://doi.org/10.1016/j.physa.2008.06.035.
21. Choi J, Lee YJ, Kim T, Sohn K. An analysis of metro ridership at the station-to-station level in Seoul. Transportation. 2012;39:705–22.
22. Kim KY, Wu Q. A comparison study of EOF techniques: analysis of nonstationary data with periodic statistics. J Clim. 1999;12:185–99.
23. Kumar S, Toshniwal K. A novel framework to analyze road accident time series data. J Big Data. 2016;3:8. https://doi.org/10.1186/s40537-016-0044-5.
24. Kaur A, Datta A. A novel algorithm for fast and scalable subspace clustering of high-dimensional data. J Big Data. 2015;2(17):1–24.