An Approach to Extracting and Visualizing Daily Human Activity Patterns Using Principal Component Analysis

Weiying WANG*, Toshihiro Osaragi**

Abstract: Human activity patterns have raised broad attention in geography, urban planning, transportation, etc. In this paper, we present a method of extracting and visualizing daily activity patterns using principal component analysis. For each person, a day is divided into 48 slots (30 min for each), each of which is filled with the activity the person conducted during that period. Every list of slots is transformed into a one-dimensional binary matrix. We applied the principal component analysis to the matrices and extracted principal components (eigenvectors), which are regarded as activity patterns. Individual samples are projected on the vectors and further visualized on the map. Some interesting spreads of activity patterns over years can be observed, which may be correlated with the development of public transportation and the sprawling of urban areas.

Keywords: daily activity patterns, principal component analysis, time-series, visualization

1. Introduction

Our society is shaped by people's activity-travel patterns. They are complicated and affected by many factors, ranging from the urban environment and policies to individual attributes such as age, occupation, etc. Over the last seven decades, Japan has experienced fast economic growth and urbanization after the Pacific War, followed by the "lost decades". Significant changes have happened in the urban environment, economy, policies, as well as population and occupational structures. Such changes led to the transformation in daily lifestyle over years. Before understanding the relationship between the changes and the transformation of lifestyle, in this paper, we visualize what changes have happened in human daily activity patterns between 1978 and 2008.

Regarding human behaviors, discrete activities may easily appear random, but repeating patterns can be observed at the aggregative level. To capture characteristic activity patterns, we employ principal component analysis (PCA hereafter). This technic has been frequently applied to reduce the dimension of high-dimensional data and extract patterns. For example, Eagle and Pentland (2009) applied the method to a dataset containing 100 smartphone users to extract their behavioral patterns. In this research, we follow their steps for data preprocessing.

2. Dataset and six groups of people

2.1. Person trip survey data

Person Trip survey has been conducted every ten years by the Ministry of Land, Infrastructure, and Tourism of Japan in major urban areas. It is conducted to households and collects information about their travels on a given day. "Trips" defined in Person Trip survey data (hereafter PT data) are illustrated in Fig. 1. Place and time of departure and arrival, travel purpose, and means of trips, as well as personal attributes (age, gender, occupation, car ownership, etc.), are included in the dataset. In our research, the surveys were conducted in 1978, 1988, 1998, and 2008. Each survey covers an area of a circle with a radius of 70 km centered on the Tokyo Railway station (Osaragi and Kudo, 2020). Fig. 2 shows the 342 administrative regions in the survey. The day starts at 3:00 AM and ends at 3:00 AM the next day. We further filter out samples whose trip information, such as trip purpose,

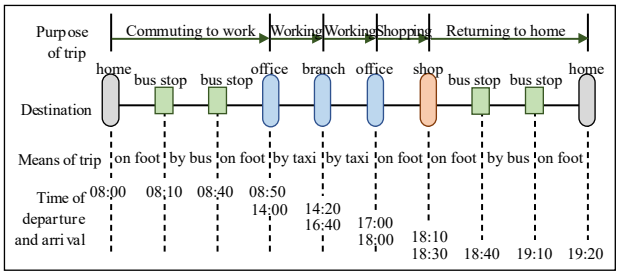


Fig. 1. Example of trips in Person Trip survey data.

* 学生会員 東京工業大学環境・社会理工学院（Tokyo Institute of Technology）
〒152-8550 東京都目黒区大岡山2-12-1 E-mail: wang.w.al@m.titech.ac.jp
** 正会員 東京工業大学環境・社会理工学院（Tokyo Institute of Technology）

travel time, and location, is missing.

2.2. Six groups of people

Based on age and occupation, we classify people into six groups to rule out the effect of population and occupational structures on activity patterns (Table 1). Fig. 3 shows their proportions over years.

3. Methodology

3.1. From activity sequences to activity vectors

Using PT data, individual activity at any time of the day can be inferred. Following most previous studies, we categorized activities into four types: staying at home (R), working/ educational (C), other activities (O, including shopping, entertainment, and personal activities), and having a trip (T). We divide one day into 30-min slots (48 slots for a day). Each of the 48 slots is filled with the activity a person was doing during that 30-min interval, and an activity sequence is constructed for every person (S in Fig. 4).

Following the work of Eagle and Pentland (2009), one activity sequence can be transformed into four binary vectors (x_i^R, x_i^C, x_i^O, and x_i^T), and each vector indicates one type of activity. If a person conducted activity R at the t-th time interval, the value at the t-th slot of x_i^R is one, and values at the t-th slots of other vectors are zero. The same applies to other activities. The four vectors are connected to form a new **activity vector**, x_i, with the length of 192 (Fig. 4, $x_i = x_i^R \oplus x_i^C \oplus x_i^O \oplus x_i^T$).

3.2. Applying principal component analysis

Many people share similar daily routines with others, and some activity patterns are frequently observed. This suggests that daily activities are not distributed randomly throughout the 192-dimensional space. Dimension reduction techniques can be applied to understand human activities. In this study, we apply PCA. In the original datasets, each person has a magnification value indicating the number of people this person represents. The weighted mean vector of one group of people is given by

$$\mu = \frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} m_i x_i \quad (1)$$

Table 1. Six groups of people.

Group ID	Attributes	Counts
Group 1	Workers	1,369,367
Group 2	Household wives/ husbands	462,866
Group 3	The unemployed	326,828
Group 4	Students aged 15 or above	124,293
Group 5	People under 15	368,201
Group 6	Others	79,774

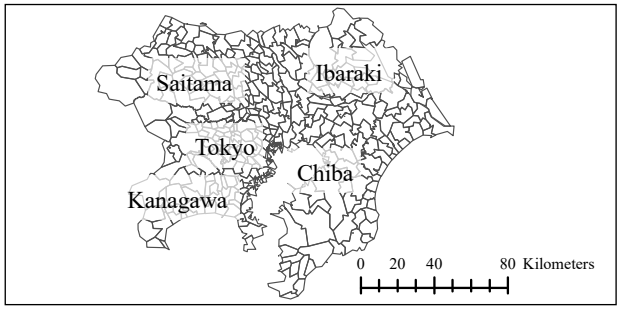


Fig. 2. 342 surveyed administrative regions.

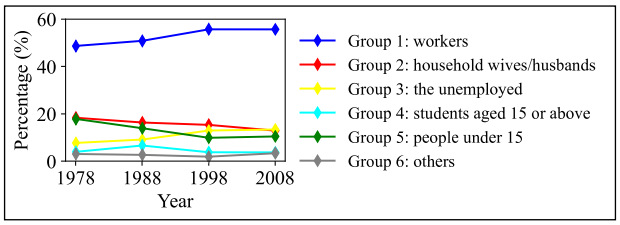


Fig. 3. Group proportions over years.

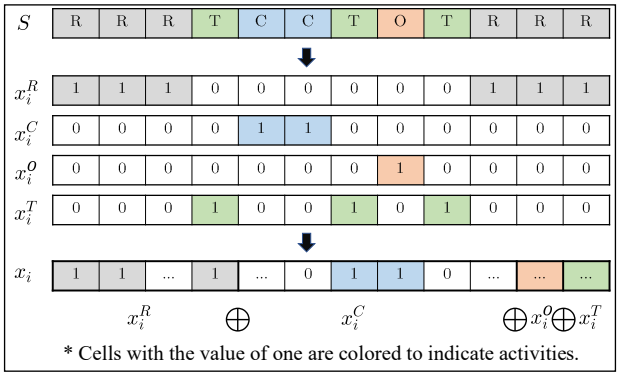


Fig. 4. How to configure activity vectors.

where m_i is the magnification of the i-th person and n is the number of people in the group. μ is subtracted from all activity vectors, and we construct an $n \times 192$ matrix, X, that contains all these centered activity vectors of this group of people. The weighted covariance matrix is

$$K = \frac{1}{\sum_{i=1}^{n} m_i} X^T M X \quad (2)$$

where M is the diagonal matrix of magnifications. Eigen decomposition is applied to K to get 192 unit-eigenvectors ($v_1, v_2, \ldots, v_{192}$) and eigenvalues ($\lambda_1, \lambda_2, \ldots, \lambda_{192}$). Eigenvectors are orthogonal to each other. The first eigenvector, v_1, corresponding to the

largest eigenvalue, λ_1, is the direction where the projection variance is the largest. The second eigenvector, orthogonal to the first, corresponds to the second largest variance, and so forth. A linear combination of these vectors reconstructs people's activity vectors.

3.3. Visualization and spatial inference

As we have classified people into six groups based on age and occupation, population and occupational structures are treated as inherent characteristics of groups. The variation of activity patterns of a certain group of people is raised by external factors such as the development of transportation, urban environments, policies, and social norms. These factors show spatial differences. For example, public transportation is more developed in the city center than in suburban areas. Policies and living routines may vary as well. If such external factors have certain effects on people's daily behavior, we are likely to observe some structured spatial differences in activity vectors. Since activity vectors do not distribute randomly, a few eigenvectors capture most of the variance between them. For any group of people, a set of eigenvectors can be obtained using the method in Section 3.2. The activity vector of a person in this group can be represented in a low-dimensional space spanned by a limited number of eigenvectors. On the k-th axis (the k-th eigenvector), the coordinate of an activity vector in the new space is the projection of it on the vector:

$$y_i(k) = x_i v_k \tag{3}$$

where x_i is the centered activity vector (subtracted by μ). For any administrative region in Fig. 2, we look at the average projection value on the k-th eigenvector from people who live in the region:

$$z(r,k) = \frac{1}{\sum_{i=1}^{h} m_{i,r}} \sum_{i=1}^{h} m_{i,r} y_{i,r}(k) \tag{4}$$

where r is the region, $y_{i,r}(k)$ is the projection of the i-th person in this region, $m_{i,r}$ is the weight of the i-th person, and h is the number of people surveyed in the region. If the value shows structured patterns on the map that are similar to some external factors, such variation in activity patterns may be explained by these factors.

4. Results

For each group of people in Table 1, activity vectors form an activity matrix (X in Section 3.2, including all people in the group from 1978-2008). PCA is conducted and eigenvectors are obtained. Fig. 5 shows the proportions of the first five eigenvalues. On average, about 56.8% of the variances (i.e., proportions of eigenvalues) are captured by the first five eigenvectors, which is good considering the original space has 192 dimensions. This suggests that several repeating activity patterns dominate most people's daily activities. In the following subsections, we show the projections ($z(r,k)$ in Section 3.3) on the first several eigenvectors.

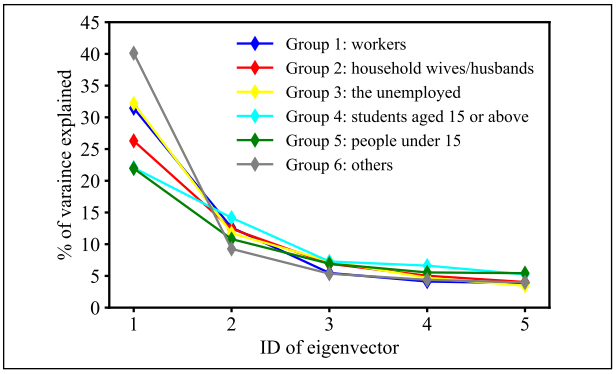
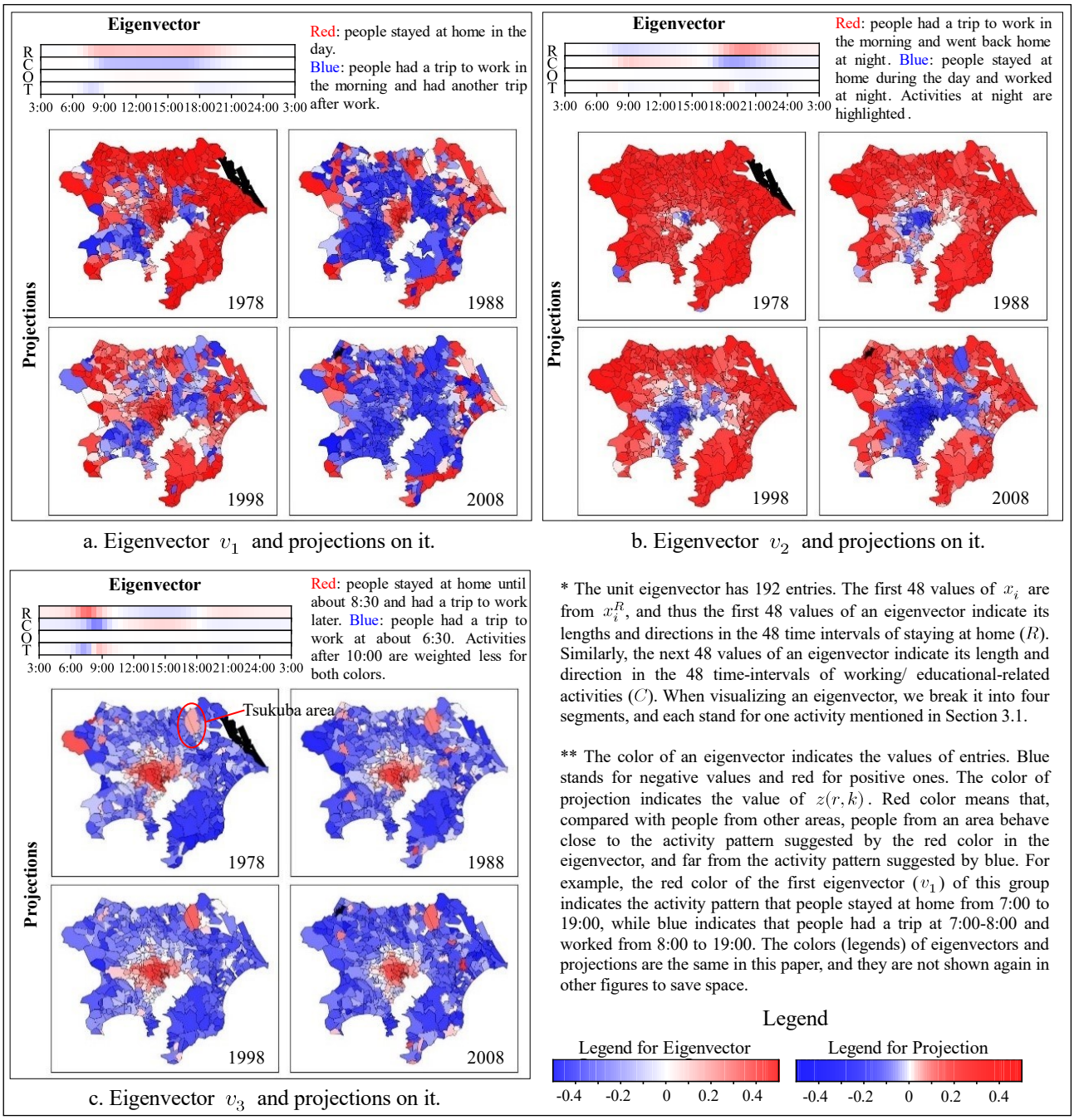


Fig. 5. Proportions of eigenvalues.

4.1. Results for workers

In Fig. 6, the activity patterns indicated by red and blue are described briefly near the eigenvector diagrams. Detailed explanations are on the bottom right of Fig. 6. For Group 1 (workers), the first eigenvector (v_1 in Fig. 6-a) differentiates workers who spent more time at home from those who worked. A negative projection value (blue) suggests that a person may have spent more time working in the day and less time at home compared with others. It is worth noting that a negative value does not mean that a person spent more time working than staying at home. Also, a positive average value for an area does not suggest that more workers stayed at home than those who did not because activity vectors are centered. A positive or negative value just indicates the "relative location" compared with the "average". From the projections on the first vector, it can be observed that people spent less time at home in 1988 than in 1978. The time spent at home

increased from 1988 to 1998 while decreased again from 1998 to 2008. The second eigenvector (Fig. 6-b) differentiates those who worked in the day and relaxed at night from those who stayed at home in the day and worked at night. Activities at nighttime are weighted more (deeper color) than activities at the daytime. As time goes on, more and more people worked at night. This type of lifestyle started in the city center and expanded to suburban areas (blue). For the third vector (Fig. 6-c), activities in the morning are highlighted. A positive value (red) suggests that people stayed at home from about 6:30 to 8:30 and went to work later. A negative value (blue) suggests that people started the trip to work at this time. Significant differences between the city center and suburb areas can be spotted. Workers living in the city center started the trip later than workers from the suburban areas because the commuting time of the former is shorter. For the Tsukuba area in the northeast, workers started the trip late as well. Tsukuba is mainly a place for scientific research and workers living in the area may not often



a. Eigenvector v_1 and projections on it.

b. Eigenvector v_2 and projections on it.

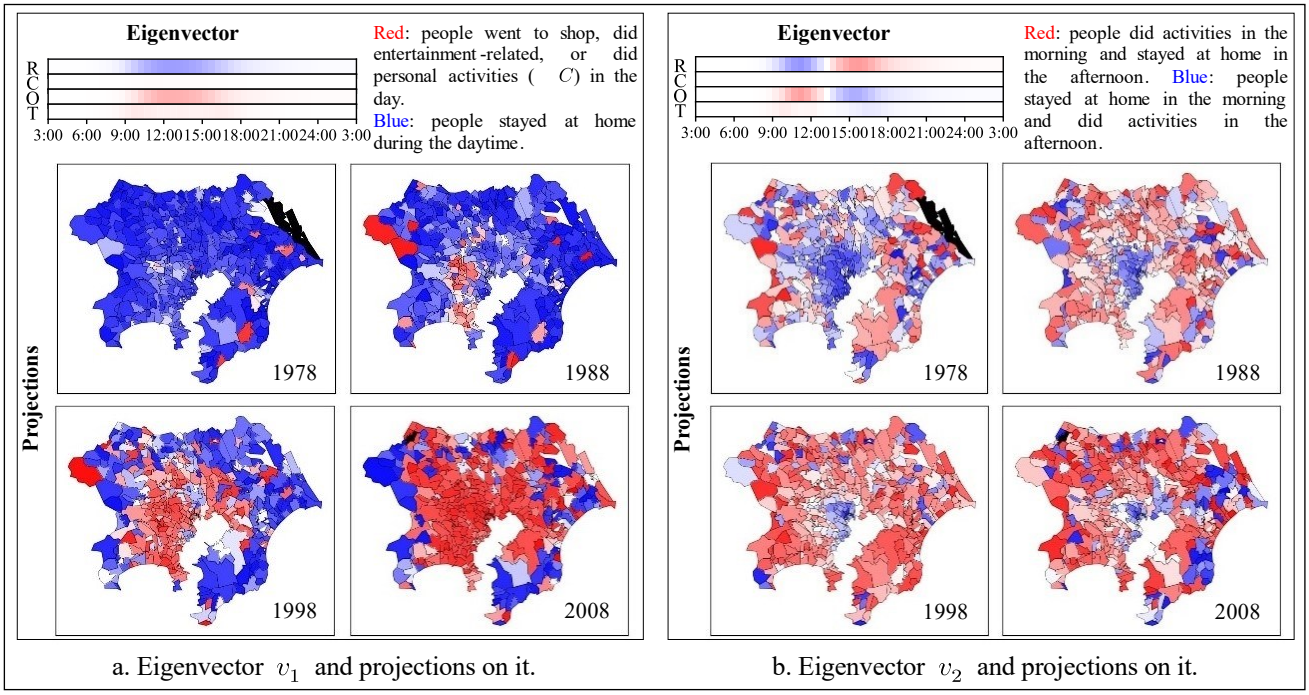
c. Eigenvector v_3 and projections on it.

Fig. 6. Eigenvectors and the average value of projections of each area over years for Group 1 (Workers).

commute to the city center to work. Other similar areas in the northeast can be identified. Red areas are slightly shrinking, probably because of the increase in commuting time (more people in suburban areas commute to the city center).

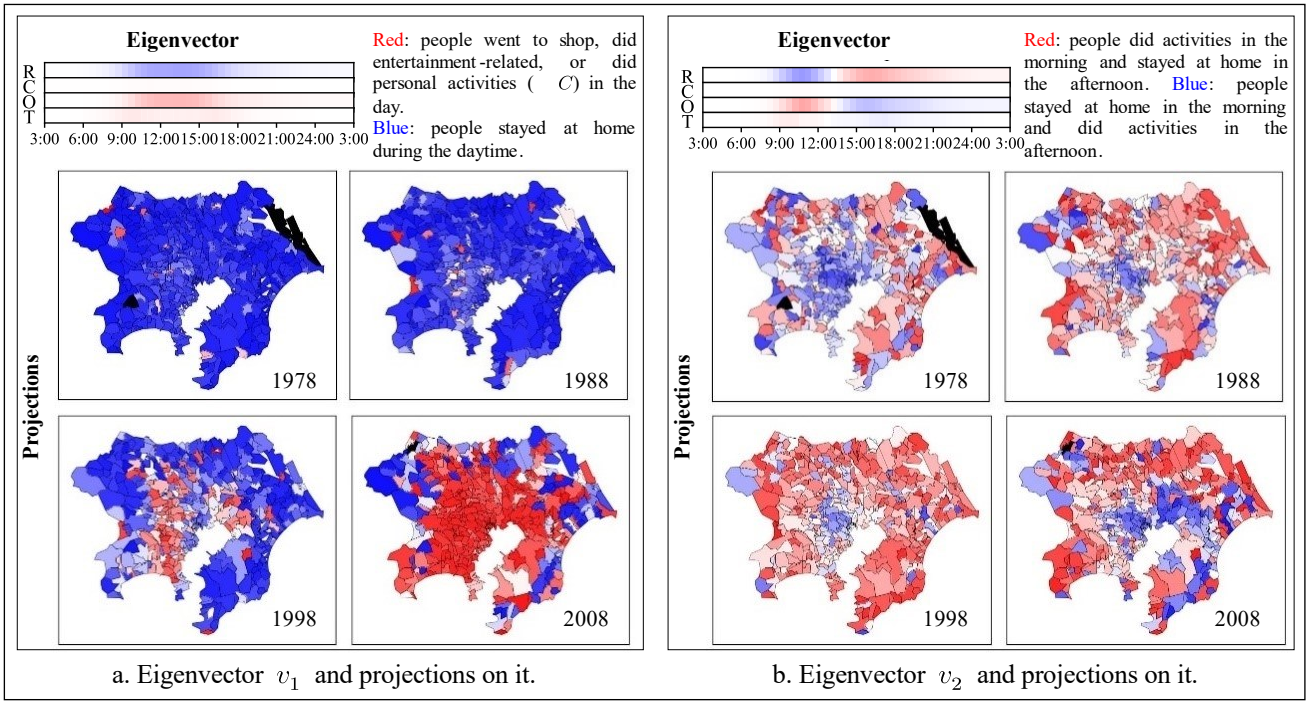
4.2. Results for household wives/ husbands and the unemployed

For Group 2 (household wives/ husbands) and Group 3 (the unemployed), not only the eigenvectors that differentiate people, but also how lifestyles have changed over the years, are similar. This may suggest that the relationships between external factors and lifestyles are similar for the two groups. However, we do not combine the two groups into one because their activity vectors show differences (similar eigenvectors do not imply similar activity vectors). A simple combination leads to



a. Eigenvector v_1 and projections on it. b. Eigenvector v_2 and projections on it.

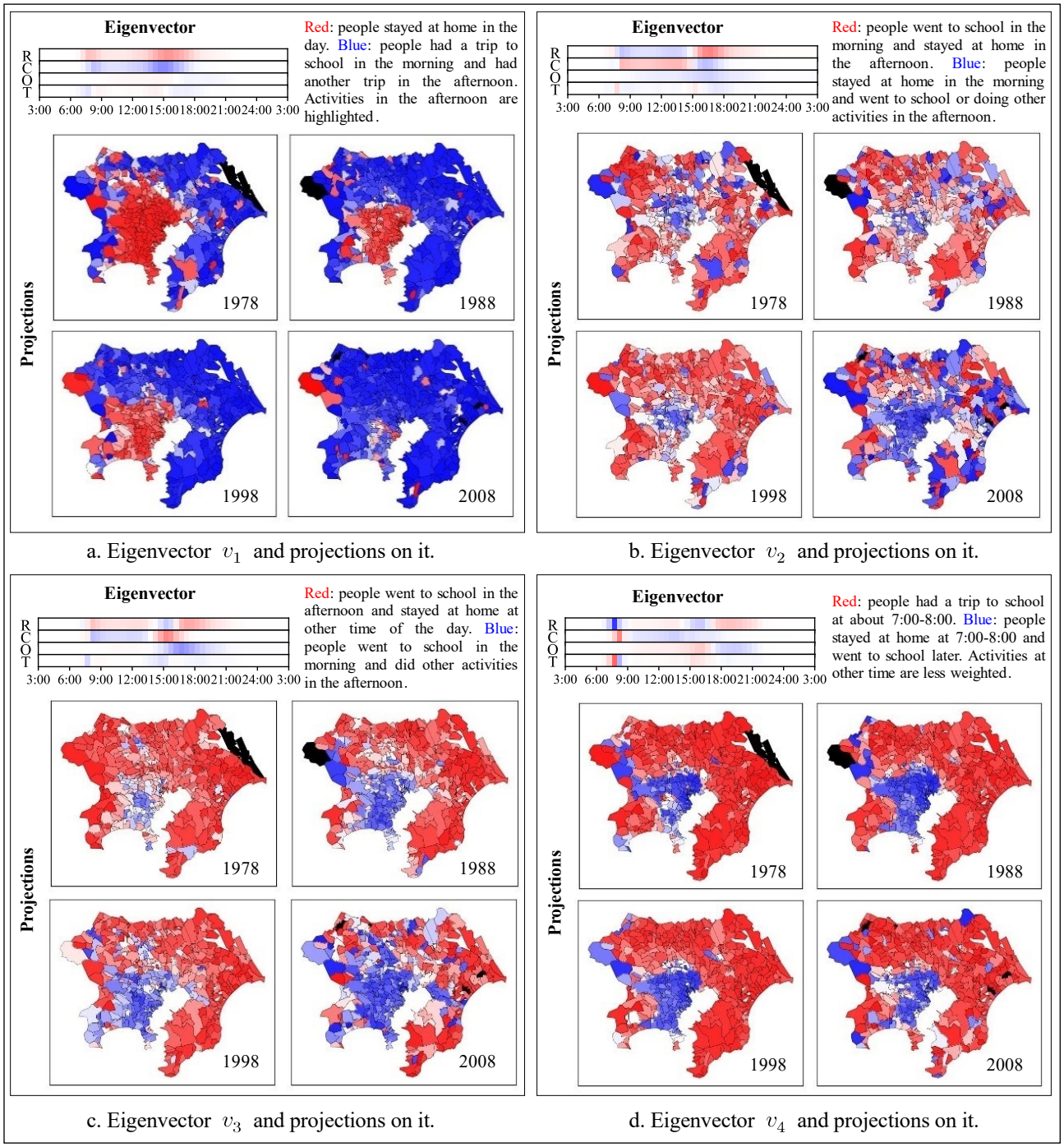
Fig. 7. Eigenvectors and the average value of projections of each area over years for Group 2 (household wives/ husbands).



a. Eigenvector v_1 and projections on it. b. Eigenvector v_2 and projections on it.

Fig. 8. Eigenvectors and the average value of projections of each area over years for Group 3 (the unemployed).

unexpected results as their relative proportions vary over years (Fig. 3). The first eigenvectors in Fig. 7-a and Fig. 8-a differentiate people who stayed at home (blue) from those who conducted shopping, entertainment, or personal activities (red). Household wives/ husbands and the unemployed living in the west of the city center started to do more of these activities in 1988 (red), and this type of lifestyle expands to other areas. The second eigenvectors (Fig. 7-b and Fig. 8-b) differentiate people who stayed at home in the morning and went out in the afternoon (blue) from those who behaved contrariwise. The spatial variation is obvious: those living in the city center conducted activities more often in the afternoon compared with others. For household wives/ husbands, activities in the morning became slightly more often in the city center over years, while such a change has not occurred for the unemployed. On the two eigenvectors, projection values seem to show continuous variations spatially.



a. Eigenvector v_1 and projections on it.

b. Eigenvector v_2 and projections on it.

c. Eigenvector v_3 and projections on it.

d. Eigenvector v_4 and projections on it.

Fig. 9. Eigenvectors and the average value of projections of each area over years for Group 5 (people under fifteen).

4.3. Results for people under fifteen

The first four eigenvectors of Group 4 (students aged fifteen or above) and Group 5 (people under fifteen) are almost identical, and the lifestyle changes are similar. Based on the same reasoning as Section 4.2, they are not combined into one group. To save space, we only show the results for people under fifteen (Group 5), because the number of samples in this group is almost three times as many as the other one. In Fig. 9-a, the first eigenvector captures the variation of the time staying at home. In 1978, students living in the south and in the west spent more time at home (red) compared with others. As time goes, this type of lifestyle shrank. More students spent more time on educational-related activities especially in the afternoon (deeper color). The second eigenvector (v_2 in Fig. 9-b) differentiates people who went to school in the morning and back home at 15:00 (red) from those who stayed at home in the morning and/ or went to school or do other activities in the afternoon (blue). The projections on the vector suggest that people in Group 5 from the city center have more "irregular" daily routines. They are more likely to spend more time at home during the day and were more often having educational-related or other activities in the afternoon. Such "irregularity" seems to have grown slightly for Group 5 (people under fifteen), while the growth is more obvious for Group 4 (aged fifteen or above). The third eigenvector (Fig. 9-c) differentiates people who were at school before 15:00 and had other activities later (blue) from others. This type of lifestyle expanded from the southwest of the city center to other areas, probably due to the growing number of students attending the cram school. The fourth eigenvector (Fig. 9-d) mostly captures the difference in the morning. Blue indicates that students were at home between 7:00 to 8:00, while red implies that students had a trip to school at that time. The city center is almost in blue probably because schools are closer to students' homes and the public transportation is more developed so that the commuting time is shorter. This eigenvector, as well as projections on it, is close to the third eigenvector and projections of workers. For students aged fifteen or above, the pattern is not obvious.

5. Discussions and conclusions

The changes of complicated activity vectors are captured by a limited number of eigenvectors. There are generally three types of variations of patterns: expanding/ shrinking pattern, such as projections on the second vector of workers and the first vector of the unemployed and household wives/ husbands; stable pattern, such as the projections on the third vector of workers; and others. Many projections belong to the first or the second variation pattern. This suggests structured spatial variations of people's lifestyles. The visualized variation patterns on the map are mostly centered on the city center. This agrees with the distribution of many external factors, such as the density of population, shopping centers, job opportunities, the development of public transportation (e.g., the density of railways), etc. We can conclude that these external factors are likely to be correlated with lifestyle changes. However, because of the complexity of daily activity patterns, this work does not draw any conclusion about the explicit relationships between them. On the other hand, some eigenvectors and/ or projections on eigenvectors may not be easily interpreted, especially eigenvectors that capture small variances. We will continue this work and seek better methods for visualization and analysis. In addition, the correlations between the change of lifestyle and external factors will be analyzed in future works.

References

Eagle, N., Pentland, A.S., 2009. Eigenbehaviors: identifying structure in routine. Behav Ecol Sociobiol 63, 1057–1066. https://doi.org/10.1007/s00265-009-0739-0

Osaragi, T., Kudo, R., 2020. Enhancing the Use of Population Statistics Derived from Mobile Phone Users by Considering Building-Use Dependent Purpose of Stay, in: Kyriakidis, P., Hadjimitsis, D., Skarlatos, D., Mansourian, A. (Eds.), Geospatial Technologies for Local and Regional Development. Springer International Publishing, Cham, pp. 185–203.