Understanding the Link between Urban Activity Destinations and Human Travel Patterns

Shan JIANG
PhD Student
Department of Urban Studies and Planning
Massachusetts Institute of Technology
77 Massachusetts Ave. E55-19D
Cambridge, MA 02142
Email: shanjang@mit.edu

Joseph FERREIRA
Professor of Urban Planning and Operations Research
Massachusetts Institute of Technology 77 Massachusetts Ave. 9-532 Cambridge, MA 02139
Email: jf@mit.edu

Marta GONZÁLEZ
Assistant Professor
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology
77 Massachusetts Ave. Room 1-153
Cambridge, MA 02139
Email: martag@mit.edu

Abstract In metropolitan areas, individuals exhibit regular yet rich temporal dynamics in their daily activities during weekdays. In this study we employ data mining and statistical learning techniques (viz., principal component analysis and K-means clustering algorithms) to analyze self-reported activity patterns. We explore three critical issues: (1) the inherent daily activity structure of individuals in a metropolitan area, (2) the temporal variation of individual activities—how their numbers grow and fade over time, and (3) the classification of individual behavior based on the daily signatures and related social demographic characteristics. We use urban activity-based travel survey data for the Chicago metropolitan area, including a total of 10,552 households (and more than 30,000 individuals) who participated in a 1-day or 2-day survey implemented from January 2007 to February 2008. The generated classifications, combined with spatial information about activities, provide a useful framework for urban and transportation planning by addressing when, where, and how individuals interact with places in metropolitan areas.

Keywords: Human activity pattern, Eigen decomposition, Time series clustering, Statistical learning

1 INTRODUCTION

Many efforts have been put into understanding the dynamics and the complexity of cities (Batty 2003; Reggiani and Nijkamp 2009) where individuals exhibit regular yet rich dynamics in their social and physical lives. This field was mostly the territory of urban planners and social scientists alone, but has recently attracted a more diverse body of researchers from computer science and complex systems as a result of development in interdisciplinary approaches and rapid technology innovation. Emerging urban sensing data (such as massive mobile phone data, and online user-generated social media data), both in the physical (Kim et al. 2006) and virtual world (Crane and Sornette 2008; Kim et al. 2006), has been accompanied by the development of data mining and statistical learning techniques (Kargupta and Han 2009) and an increasing and more affordable computational power. As a consequence, one of the fundamental and traditional questions in the social sciences, "how humans allocate time to different activities as part of a spatial, temporal socio-economic system," becomes treatable within an interdisciplinary domain.

Different facets of this question have long been studied by researchers in sociology (Geerken et al. 1983), psychology (Freud 1953; Maslow and Frager 1987), geography (Hägerstrand 1989; Hanson and Hanson 1980; Harvey and Taylor 2000; Yu and Shaw 2008), economics (Becker 1991), and urban and transportation planning (Bhat and Koppelman 1999; Chapin 1974). Nevertheless, recent innovation in both data sources and analytical approaches, have inspired new studies about the dynamics of human activities. For example, Gonzalez et al. (2008) by using the mobile phone data, studied the trajectory of 100,000 anonymized mobile phone users and showed a high degree of temporal and spatial regulatirty of human trajectories. Eagle and Pentland (2009) by analyzing the Reality Mining data set from continous mobile phone logging locations collected from an experiment at MIT, studied the behavioral structure of the daily routine of the students and inferred individual community affiliations based on some apriori information of the subjects. On the other hand, researchers are also facing significant challenges when deploying these new sources of data and technology (Nature Editorial 2008). Due to privacy and legal constraints, it is hard to have a whole picture of the social context, as we have no or very limited information about the socioeconomics and demographics of the human subjects being studied in these circumstances. Despite the fact that these new datasets may allow us to study the social relationship and networks (Eagle et al. 2009), they still have limited capacity in revealing underlying reasons driving human behavior.

Meanwhile, technology development in geographic information systems (GIS) such as automated address matching, and in computer-aided self-interview (CASI) enable us to have higher spatial and temporal resolution than in the past, which leads to the improvements in the accuracy, quality and reliability of the self-reported survey data (Axhausen et al. 2002; Greaves 2004). Compared with urban sensing data (such as mobile phone data), survey data is disadvantaged by high cost, low frequency, and small sample size. However, in terms of the richness of the socioeconomic and demographic information, survey data provides much richer social demographic information for exploring social differences underlying the human activity dynamics, and thus enables us to develop more nuanced models for explaining and predicting human activity patterns.

Inspired by many of the aforementioned issues and studies, in this paper, we exploit the richness of survey data using data mining techniques, which have not been applied in this context before. Since the survey collected over the metropolitan area is conducted by the metropolitan planning organization (MPO) for the regional transportation planning purposes, it is free for public access, reliable, and representative of the total regional population. The differences between our study and the traditional time-use studies on human activities lie in our methods. We do not superimpose any predefined social classification on the observations, but employ data mining and statistical learning techniques. We let the inherent activity structure inform us of the patterns and clusters of individual temporal activities in the metropolitan area. By summarizing the social demographic characteristics of each cluster, we try to reveal the social connections and differences within and among the human temporal activity clusters. Daily temporal activity of groups of individuals in a metropolitan area should have underlying structures which can be extracted using data mining techniques similar to the ones applied nowadays to clustering users' on-line behavior (Yang and Leskovec 2011). We use K-means clustering and eigen decompositions which provide a low dimensional characterization of complex phenomena. By classifying individuals according to their activities in combination with spatial information, our ultimate goal is to provide a clear picture of where, when and how groups of individuals interact with different places in the city.

2 DATA

The data used in this study are from a publicly available "Travel Tracker Survey" —a comprehensive travel and activity survey for Northeastern Illinois designed and conducted for regional travel demand modeling (Chicago Metropolitan Agency for Planning 2008). Due to its purpose, the sampling framework of the survey is a stratification and distribution of surveyed household population in the 8 counties of the Northeastern Illinois Region. It closely matches the 2000 US Census data for the region at the county level). The data collection was implemented between January 2007 and February 2008, including a total of 10,552 households (32,366 individuals). Every member of these households participated in either a 1-day or 2-day survey, reporting their detailed travel and activity information starting from 3:00 a.m. in the early morning on the assigned travel day(s). The survey was distributed during 6 days per week (from Sunday to Friday) in the data collection period. Among panels of the publicly available data, in this study, we focus on those containing information about households (e.g., household size, income level), personal social demographics (e.g., age, gender, employment status, work schedule flexibility), trip details (travel day, travel purpose, arrival and departure times, unique place identifiers), and location.

2.1 Data Processing

In the original trip data, location is anonymized by moving the latitude and longitude of each location to the centroid of the associated census tracts. By assuming that people move from point A to point B in a straight line with constant moving speed, we are able to fill in the latitude and longitude locations of the movement between two consecutive destinations. Using this method, we reconstruct the data at a 1-minute interval, providing a time stamp (in minutes), a location with paired latitude and longitude, an activity type,

and a unique person-day ID. In order to reduce the high dimension of the 23 primary purposes in the original survey data, we aggregate them into 9 activity types as shown in Table 1. We also use a specific color for each activity throughout the entire paper. We label the activity type of individuals while traveling to be that of their destination activity type. For example, if an individual starts her morning trip from home to work at 7:00 a.m., arrives at her work place at 7:30 a.m., and begins work from 7:31 a.m. and finishes work at 11:30 a.m., we label her activity type during the time period [7:00 a.m., 11:30 a.m.] as "work".

Table 1 Newly aggregated nine activity types v.s. the original twenty-three primary trip purposes

Aggregated Activity Types	Original Primary Trip Purposes
Home	1. Working at home (for pay); 2. All other home activities
Work	3. Work/Job; 4. All other activities at work; 11. Work/Business related
School	5. Attending class; 6. All other activities at school
Transportation Transitions	7. Change type of transportation/transfer; 8. Dropped off passenger from car; 9. Picked up passenger; 10. Other, specify- transportation; 12. Service private vehicle; 24. Loop trip
Shopping/Errands	13. Routine shopping ; 14. Shopping for major purchases; 15. household errands
Personal Business	16. Personal Business; 18. Health Care
Recreation/Entertainment	17. Eat meal outside of home; 20. Recreation/Entertainment; 21. Visit friends/Relatives
Civic/Religious	19. Civic/Religious activities
Other	97. Other

2.2 Human Activities on an Average Weekday

By using the processed data (with location and activity type information for each individual at a 1-minute interval), we generate a separate animation visualizing the movement and activities (differentiated by nine colors demonstrated in Table 1) of the surveyed individuals in Chicago. We use the 1-day survey distributed from Monday to Thursday as an average weekday sample. We get a total of 23,527 distinct individuals who recorded their travel and activities during any day (starting from 3:00 a.m. on Day 1, and ending at 2:59 a.m. on Day 2) between Monday and Thursday. We exclude surveys on Fridays on purpose, because as confirmed from our analysis, temporal patterns of human activities on Friday usually differ from those during the rest of the weekdays. Figure 1 shows four snapshots of the animation of movement and human activities in the Chicago metropolitan area that we generated for an average weekday. The top row shows snapshots at 6:00 a.m. and 12:00 p.m., and the bottom pair are those at 6:00 p.m. and 12:00 a.m. We can see that in the early morning, the majority of people are at home while some have already started work. At noon time, a large percent of people are at work or at school, with some groups of people doing shopping, recreation, and personal businesses. In the early evening, some people are out for recreation or entertainment and some are already at home. At midnight, most people are at home, and only a few are out for recreation, or still at work place.

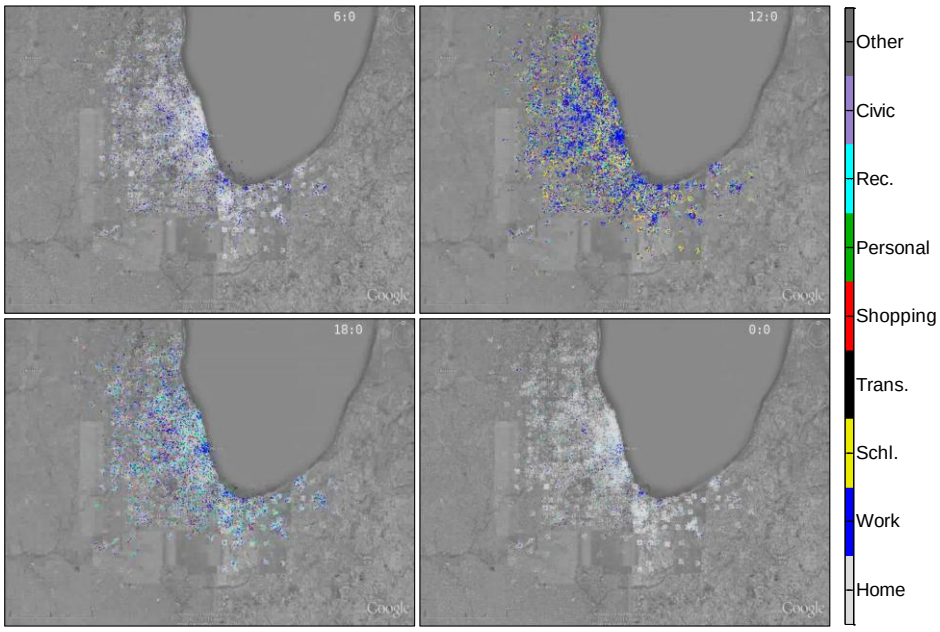

Figure 1 Snapshots of human activities at different times on a weekday in Chicago

Individual and Aggregated Temporal Activity Variations

In Figure 2, we depict the 24-hour human activity variations for an average Chicago weekday using the corresponding colors defined in Table 1. The x axis represents time-of-day (starting from 3:00 a.m. of Day 1 and ending at 2:59 a.m. on Day 2); and the y axis displays one sampled individual (i.e., each line parallel to x axis represents a sample individual). By summing up the total number of individuals conducting different types of activities at one-minute resolution along the 24-hours of the weekday, we are able to generate Figure 3, which reveals the aggregated temporal variation of human activities in the Chicago metropolitan area. In addition, the inset figure zooms in on the detailed information of the less-major activities (i.e., those with a smaller share of the trip volume) over time.

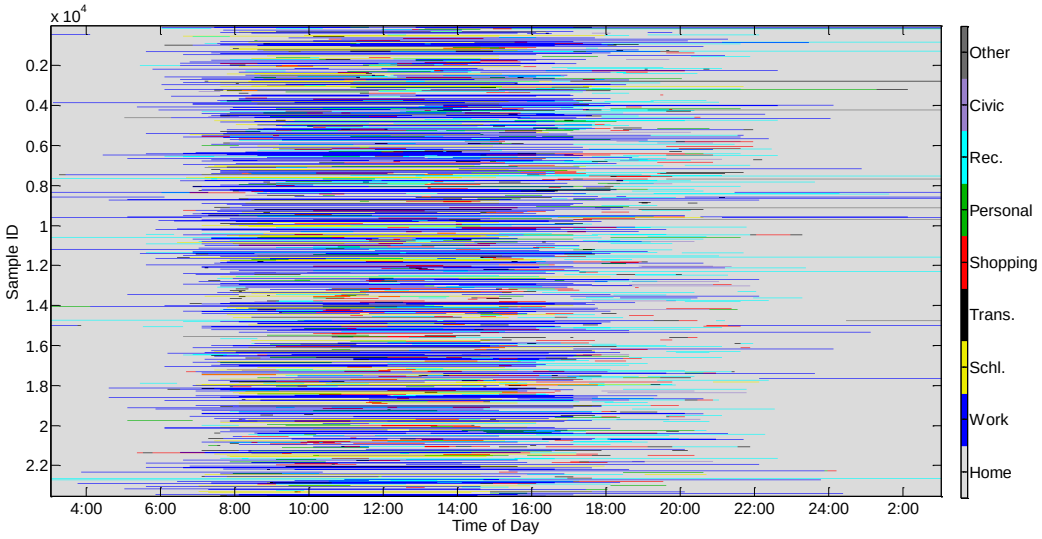

Figure 2 Individual temporal activities for samples on a weekday in Chicago

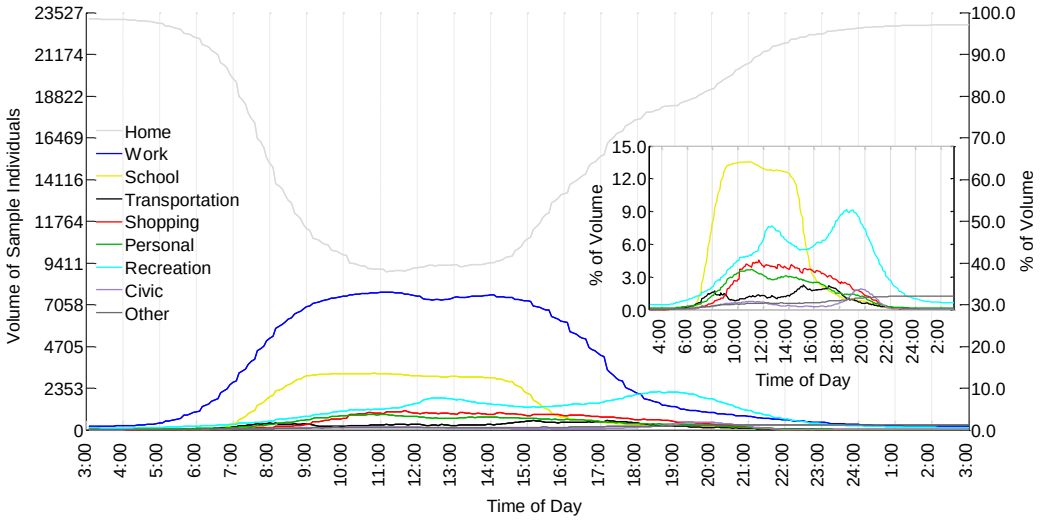

Figure 3 Temporal rhythm of human activities on a weekday in Chicago

2.3 Data Transformation

In order to lower the calculation cost incurred by the large dimension of the data while still keeping a relatively high resolution for analysis in later sections, we divide the 24 hours in a day into five-minute intervals and use the activity in the first minute of every five-minute interval to represent an individual's activity during that five-minute period. During each five-minute interval, an individual is labeled with one of the nine activities (defined as in Table 1). We then use a sequence of 288 zeros or ones (=24 hours x 12 five-minute intervals per hour) to indicate whether the individual is engaged in each particular activity during each 5-minute interval. In Figure 4, a "one" (meaning 'yes') is marked black while "zero" is white. For each sampled individual, the 9 activities and 288 time steps result in a sequence of 2,592 black/white dots along one row. Each of the 23,527 sampled individuals generates a row that is stacked along the y-axis in Figure 4.

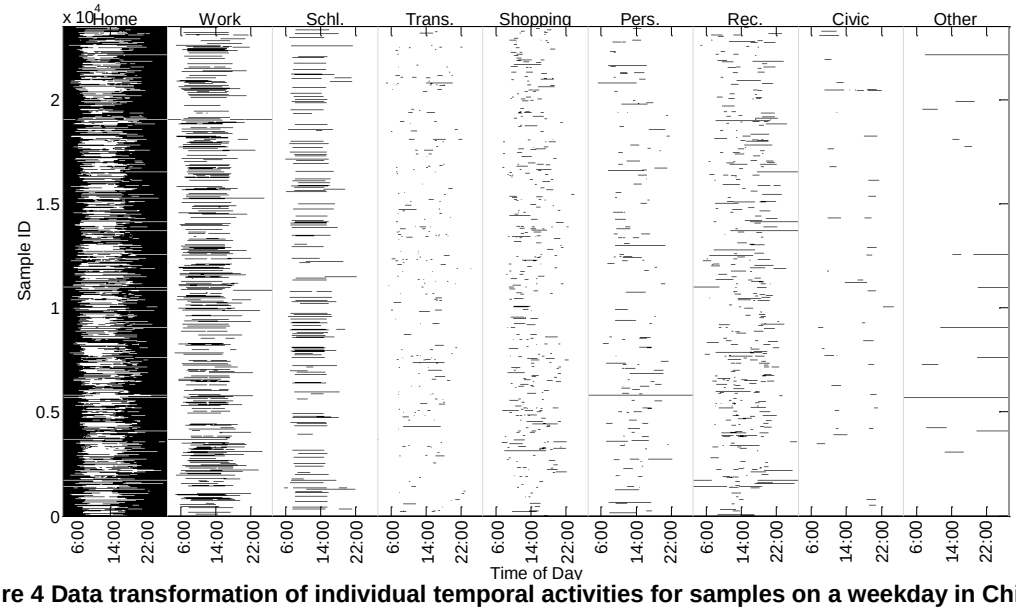

Figure 4 Data transformation of individual temporal activities for samples on a weekday in Chicago

3 MATHEMATICAL FRAMEWORK AND METHODS

In this section, we explain our mathematical notation. By employing two methods—the principal component analysis and the K-means clustering algorithm, we are able to address the two issues that we raised earlier in this paper: (1) discovering the inherent daily temporal activity structure of individuals in the metropolitan area; and (2) the classification of individuals in the metropolitan area based on their temporal activity dissimilarity.

3.1 The Setting

As mentioned above, each individual conducts only one of the nine activities defined in Table 1 during each of the 288 five-minute time intervals during one 24-hour day. For the 2,592 element vector defining an individual's activity pattern, we have $(a_1, \dots, a_m)' \in \{0,1\}^m \subset \mathbb{R}^m, m = 2{,}592$, and we say that $(a_1, \dots, a_m)'$ satisfies the compatibility condition, if for any *t* = 1, 2, ..., 288, $\sum_{l=1}^{9} a_{t+288\times(l-1)} = 1$. We define the *space of individuals' daily temporal activity sequence*, $\boldsymbol{S}$, as follows:

$$\boldsymbol{S} \triangleq \left\{ \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \in \{0,1\}^m \subset \mathbb{R}^m \middle| \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \text{ satisfies compatibility condition} \right\}.$$

In this study, the population is the set of individuals in the Chicago metropolitan area. For simplicity, we identify the sample space $\boldsymbol{\Omega}$ as the population. An individual's average weekday daily temporal activity sequence can be described by the following random vector:

$$\begin{aligned} \boldsymbol{A}: \boldsymbol{\Omega} &\longrightarrow \boldsymbol{S} \\ \omega &\mapsto \begin{pmatrix} A_1(\omega) \\ \vdots \\ A_m(\omega) \end{pmatrix}. \end{aligned}$$

Where for $j = t + 288 \times (l-1)$, $t \in \{1, \dots, 288\}$ and $l \in \{1, \dots, 9\}$, $A_j(\omega) = 0$ or 1, depending on if the individual ω is conducting activity *l* in time interval *t* on the weekday. From the survey data, we get *n* (=23,527) random weekday samples $(\boldsymbol{a}_i, \boldsymbol{b}_i)$, *i* = 1, …, *n*, where $\boldsymbol{b}_i$ stands for individual *i*'s social demographic information such as age, gender, employment status, work schedule, etc.

3.2 Principal Component Analysis/Eigen Decomposition

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to give a set of orthogonal directions in which the random vector/data have maximal variance possible. Principal components are obtained from the eigen decomposition of the population/sample covariance matrix (Hastie et al. 2009). We present the sample version here, and the population version is similar. For each sample individual *i*, let $\boldsymbol{d}_i$ denote the deviation from the mean, i.e. $\boldsymbol{d}_i = \boldsymbol{a}_i - \bar{\boldsymbol{a}}$, where $\bar{\boldsymbol{a}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{a}_i$ is the sample mean. Therefore the sample covariance matrix is given by

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n} \boldsymbol{d}_i \boldsymbol{d}_i{}' = \frac{1}{n-1}\mathbf{D}\mathbf{D}',$$

where $\mathbf{D} = [\boldsymbol{d}_1, \dots, \boldsymbol{d}_n]$.

Eigenactivities

We know that $\mathbf{C}$ is a positive semi-definite matrix, which is diagonalizable. So all the eigenvalues of $\mathbf{C}$ are nonnegative. Let $\mathbf{V'CV} = \mathbf{\Lambda}$, where $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \dots, \lambda_m\}$, and $\mathbf{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_m]$ is an orthogonal matrix whose *j*-th column $\boldsymbol{v}_j$ is the eigenvector corresponding to λ_j. For convenience, we arrange the eigenvalues in descending order, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. We call eigenvector $\boldsymbol{v}_j$ the *j*-th *eigenactivity*.

As $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_m\}$ forms an orthonormal basis for $\mathbb{R}^m$, $\mathbf{V}$ becomes the corresponding change of coordinate matrix. Namely, given a vector $\boldsymbol{v} \in \mathbb{R}^m$ whose coordinate with respect to the natural basis is $\boldsymbol{x} = (x_1, \dots, x_m)'$, the corresponding $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_m\}$-coordinate $\boldsymbol{y} = (y_1, \dots, y_m)'$ will be given by $\boldsymbol{y} = \mathbf{V'}\boldsymbol{x}$. When $\boldsymbol{v} = \boldsymbol{d}_i$ for a sample individual *i,* we call y_j (*j* = 1, … , *m*) the *projection* of individual *i*'s daily temporal activity deviation (from the mean) onto the *j*-th eigenactivity (or projection onto the *j*-th eigenactivity, for short).

Activity Reconstruction

Having known the eigenactivities and corresponding projection, we can reconstruct an individual's daily temporal activity sequence by using a subset of eigenactivities. Given a sample individual *i*'s daily temporal activity sequence $\boldsymbol{a}_i$, suppose the projection of $\boldsymbol{d}_i$ onto the first *h* eigenactivities are $(y_1, \dots, y_h)'$, then we obtain a vector $\boldsymbol{w} = (w_1, \dots, w_m)' \in \mathbb{R}^m$ according to formula

$$\boldsymbol{w} = \bar{\boldsymbol{a}} + (\boldsymbol{v}_1, \dots, \boldsymbol{v}_h)\begin{pmatrix} y_1 \\ \vdots \\ y_h \end{pmatrix}.$$

We use the following algorithm to reconstruct an individual's daily temporal activity sequence as $\hat{\boldsymbol{a}}_i = (\hat{a}_1, \dots, \hat{a}_m)' \in \boldsymbol{S}$.

- Given any $t \in \{1,2,\dots,288\}$, let $M_t = \max\{w_t, w_{t+288}, \dots, w_{t+288\times 8}\}$.
- Define $\boldsymbol{u}_t = (u_{t1}, \dots, u_{t9})' \in \{0,1\}^9$ so that $u_{tl} = 1$ if and only if $w_{t+288\times(l-1)} = M_t$.
- So we get a 9-dimensional vector $\boldsymbol{u}_t \in \{0,1\}^9$ that has one component of 1, and we let $(\hat{a}_t, \hat{a}_{t+288}, \dots, \hat{a}_{t+288\times 8}) = \boldsymbol{u}_t'$. It turns out that the reconstructed $\hat{\boldsymbol{a}}_i$ satisfies the desirable relation $\hat{\boldsymbol{a}}_i = \arg\min_{\boldsymbol{s}\in\boldsymbol{S}} \|\boldsymbol{s} - \boldsymbol{w}\|$.

The Appropriate Number of Eigenactivities

To answer the question "*how many eigenactivities are sufficient to rebuild the original daily temporal activity structure"* we first define the reconstruction error as the ratio of the number of incorrectly reconstructed entries to the total number of entries, i.e.,

$$\text{Reconstruction Error for } \boldsymbol{a}_i = \frac{\|\boldsymbol{a}_i - \hat{\boldsymbol{a}}_i\|^2}{2592}.$$

Given any $\varepsilon > 0$, it is clear that we can find some $h > 0$, so that the average reconstruction error caused by ignoring the projections onto the ignored eigenactivities $\{\boldsymbol{v}_{h+1}, \ldots, \boldsymbol{v}_m\}$ is no greater than ε. Let $\varepsilon_0 > 0$ be the acceptable error level, and define $h(\varepsilon_0)$ to be the smallest h such that the average reconstruction error induced by using the first h eigenactivities is no greater than ε_0. We then call $h(\varepsilon_0)$ the *appropriate number of eigenactivities*.

3.3 Time Series Clustering

To answer the second question raised in the beginning of Section 3, we use the K-means clustering algorithm to classify individuals in the metropolitan area based on their temporal activity dissimilarity. As K-means algorithm is one of the most popular iterative clustering methods, the readers can refer to Duda, Hart & Stork (2001) for further detailed discussion. We use the Euclidean distance to measure the dissimilarity between individuals' temporal activity sequences. Given two samples $\boldsymbol{a}$ and $\boldsymbol{a'}$ in $\boldsymbol{S}$, we have two alternative approaches to measure their dissimilarity: (1) in the most natural approach, we calculate the Euclidean distance $\|\boldsymbol{a} - \boldsymbol{a'}\|$ between the original 2592-dimensional vectors $\boldsymbol{a}$ and $\boldsymbol{a'}$; (2) since we have introduced the principal component analysis which can reduce the dimension of the problem significantly, the second approach is to measure the Euclidean distance $\|\boldsymbol{y} - \boldsymbol{y'}\|$ between the $h(\varepsilon_0)$-dimensional vectors $\boldsymbol{y}$ and $\boldsymbol{y'}$, where $\boldsymbol{y}$ and $\boldsymbol{y'}$ are the projection of $(\boldsymbol{a} - \bar{\boldsymbol{a}})$ and $(\boldsymbol{a'} - \bar{\boldsymbol{a}})$ onto the first $h(\varepsilon_0)$ eigenactivities. As the principal component analysis compresses the 2592 dimensions into a relatively small $h(\varepsilon_0)$, the second approach will significantly reduce the computational cost while still maintaining the accuracy of the clustering results, we therefore present the second approach for clustering in this study.

One problem that needs to be solved in the clustering process is to determine the optimal number of clusters that best fits the inherent partition of the data set. In other words, we need to evaluate the clustering results given different cluster numbers, which is the main problem of cluster validity. There are mainly three approaches to validate the clustering results, based on (1) external criteria, (2) internal criteria and (3) relative criteria, and various indices under each criteria (Brun et al. 2007). For our study, since we don't have pre-specified clustering structure, we use internal validation indices whose fundamental assumption is to search for clusters whose members are close to each other and far from members of other clusters. More specifically, we propose to use Dunn's index (Dunn 1973), and Silhouette index (Rousseeuw 1987) to help us select the optimal number of clusters.

4 FINDINGS: TEMPORAL PATTERNS OF HUMAN ACTIVITIES

4.1 Eigenactivities

By employing the principal component analysis method discussed in Section 3.2, we derived the eigenactivities on an average weekday in the Chicago metropolitan area. Due to limited space, we only display the first three eigenactivities (in Figure 5). Each horizontal bar color-codes 288 components of one-activity in the corresponding

eigenvector (eigenactivity). The darkest red is near +0.1 and the darkest blue is near -0.1. We see that the first eigenactivity (the 1st column of Figure 5) captures the activity signatures of those who tend to be away from home and at work during the day (from 7:00 a.m. till 5:00 p.m.). Most components of each eigenactivity are close to the sample mean (i.e., the component values are near zero) except the reds for 'work' and blues for 'home' during the 7-5 period. This first weekday eigenactivity accounts for the largest variation in temporal activities, which means that the major difference in individuals' temporal activities on a weekday is whether they are working or staying at home from 7:00 a.m.to 5:00 p.m. The second weekday eigenactivity (the 2nd column of Figure 5) reveals a high likelihood of schooling from 8:00 a.m.to 3:00 p.m. combined with a low likelihood of either staying at home during the same time period or working from 8:00 m to 5:00 p.m. (compared to the sample mean in the data). The second eigenactivity accounts for the largest variation in travel patterns that is orthogonal to the first eigenactivity. The third weekday eigenactivity portrays a high likelihood of staying at home from 3:00 p.m. to 11:00 p.m., plus a relatively high likelihood of working from 7:00 a.m.to 12:00 p.m., together with low likelihoods of staying at home from 7:00 a.m. to 11:00 am, working from 3:00 p.m. to 11:00 p.m., and recreation from 4:00 p.m. to 9:00 p.m. (all compared to the sample mean). The third eigenactivity accounts for the largest variance whose direction is orthogonal to the 1st and 2nd eigenactivities.

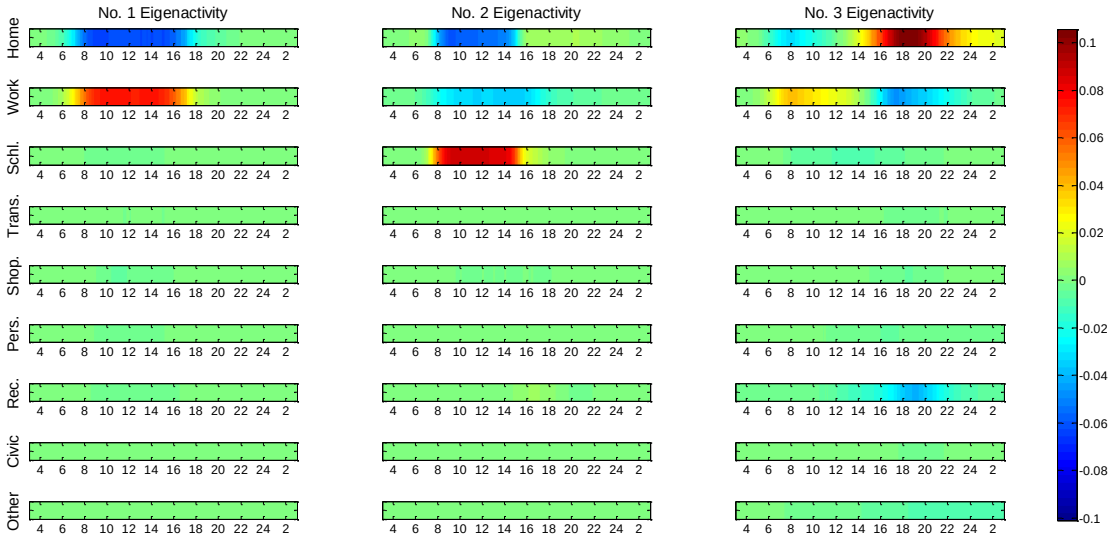

Figure 5 The first three eigenactivities of a weekday in Chicago

4.2 Selected Eigenactivities for Temporal Activity Reconstruction

We employ the reconstruction error measurement defined in Section 3.2 to select the appropriate number of eigenactivities that are sufficient to represent accurately the elements in $\boldsymbol{S}$. In Figure 6, the left panel shows the relationship between eigenvalues and the rank of eigenactivities, and the right panel displays the relationship between the reconstruction error and the number of eigenactivities used in the activity reconstruction. We can see that the eigenvalues decrease very fast with the ascending rank of eigenactivities. We find that 21 eigenactivities will allow us to reconstruct a weekday daily temporal activity sequence for individuals in the metropolitan area with an average

1% error, which means that for an average sample there are about 26 (≈2592×1%) entries (or 13 of the five-minute intervals) of our reconstructed weekday daily temporal activity sequence that are different from the original observed data. It is equivalent to say that we have around one-hour estimation error in recovering an individual's weekday daily temporal activity sequence when using 21 eigenactivities. Considering that a whole day is divided into 288 five-minute intervals and we have 9 activities in total, this reconstruction precision is very satisfactory.

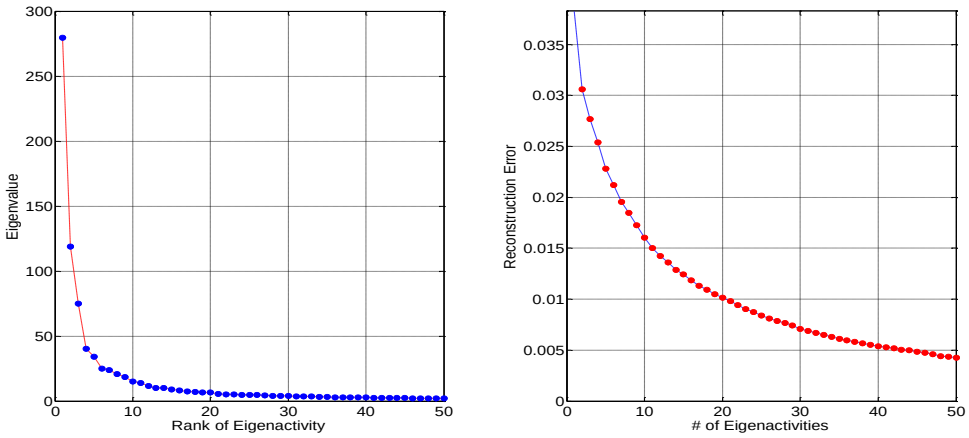

Figure 6 The eigenvalue and the reconstruction error w.r.t. the rank of eigenactivity of a weekday

Figure 7 exhibits our reconstructed individuals' daily temporal activity sequence during an average weekday using the 21 eigenactivities. Comparing it with Figure 4, we can see that, in general, our reconstructed temporal activities match the original sample data very well, except that our method does not allow us to reconstruct the activities in the "*Transportation Transitions*" category very accurately. Recall that this category involves activities such as, "changing type of transportation/transfer; dropping off passenger from car; picking up passenger; service private vehicle, etc." as described in Table 1.

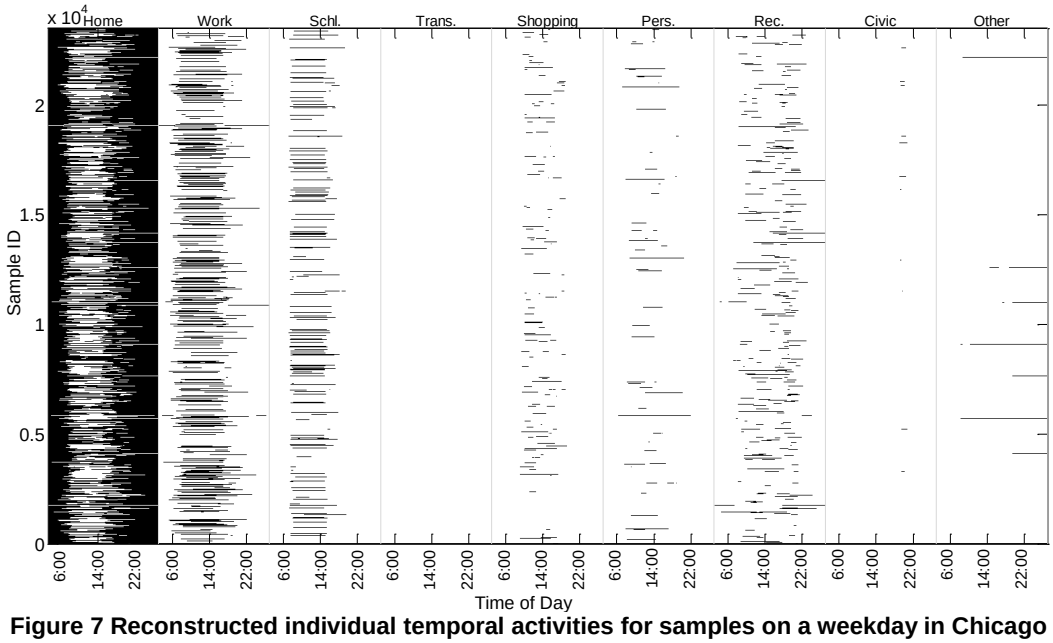

Figure 7 Reconstructed individual temporal activities for samples on a weekday in Chicago

4.3 Clustering Individuals' Temporal Activities and Social Demographics

As introduced in Section 3.3, we use the Dunn's index and the average Silhouette index (for both of which the higher the value, the better the clustering) to indentify the appropriate number of clusters for the K-means clustering. Figure 8 shows the value of the indices with respect to the number of clusters. Both indices suggest that when the cluster number is 3, it gives the best clustering results. However, we want to further explore detailed temporal activity patterns of individuals in the metropolitan area. We find that using eight clusters is satisfactory for this purpose.

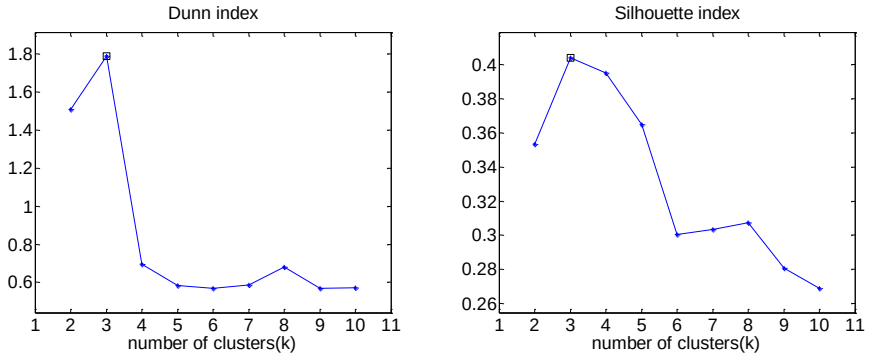

Figure 8 Cluster validity indices for the weekday case

Figure 9 exhibits the K-means clustering results (with cluster number=8) of individuals' temporal activity patterns on an average weekday along with the social demographic characteristics of those individuals grouped into each cluster based on their activity signatures. Each row of Figure 9 describes different information for the same cluster, while each column portrays temporal effects in different ways. The order of the clusters (in the row) is organized by the dendrogram of the hierarchical structure of the clusters, which is presented in the last column of Figure 9. The horizontal length of the hierarchical dendrogram measures the average distance between the two clusters being connected (Duda et al. 2001). The first column of Figure 9 (not counting the vertical color bar legend showing the color of each activity) displays the individuals' temporal activity sequences for each of the eight clusters. The second column shows the aggregated volume of different types of activities in the metropolitan area during a specific time interval over the 24 hours, and the third column is a zoomed-in view of the figures in the previous column. The fourth column presents the social demographic statistics of the cluster in that row. We use a Star Diagram (see Figure 10) to represent the average social demographic characteristics of each cluster (summarized in Table 2), including the average share of cluster members (1) who are female, (2) who are students, (3) who do not work and are homemakers (4) or who do not work and are retired (5) who work, (6) who work part time (less than 30 hours per week), (7) whose working schedule has no flexibility, (8) some flexibility, or (9) much flexibility, (10) who work at home, (11) whose educational level is greater than a technical school degree, (12) whose annual household income is low level (below $35,000), (13) middle level (between $35,000 and $75,000), or (14) high level (above $ 75,000), and (15) who are young (below 35 years old), (16) middle-aged (between 35 and 60 years old), or (17) older (above 60 years old).

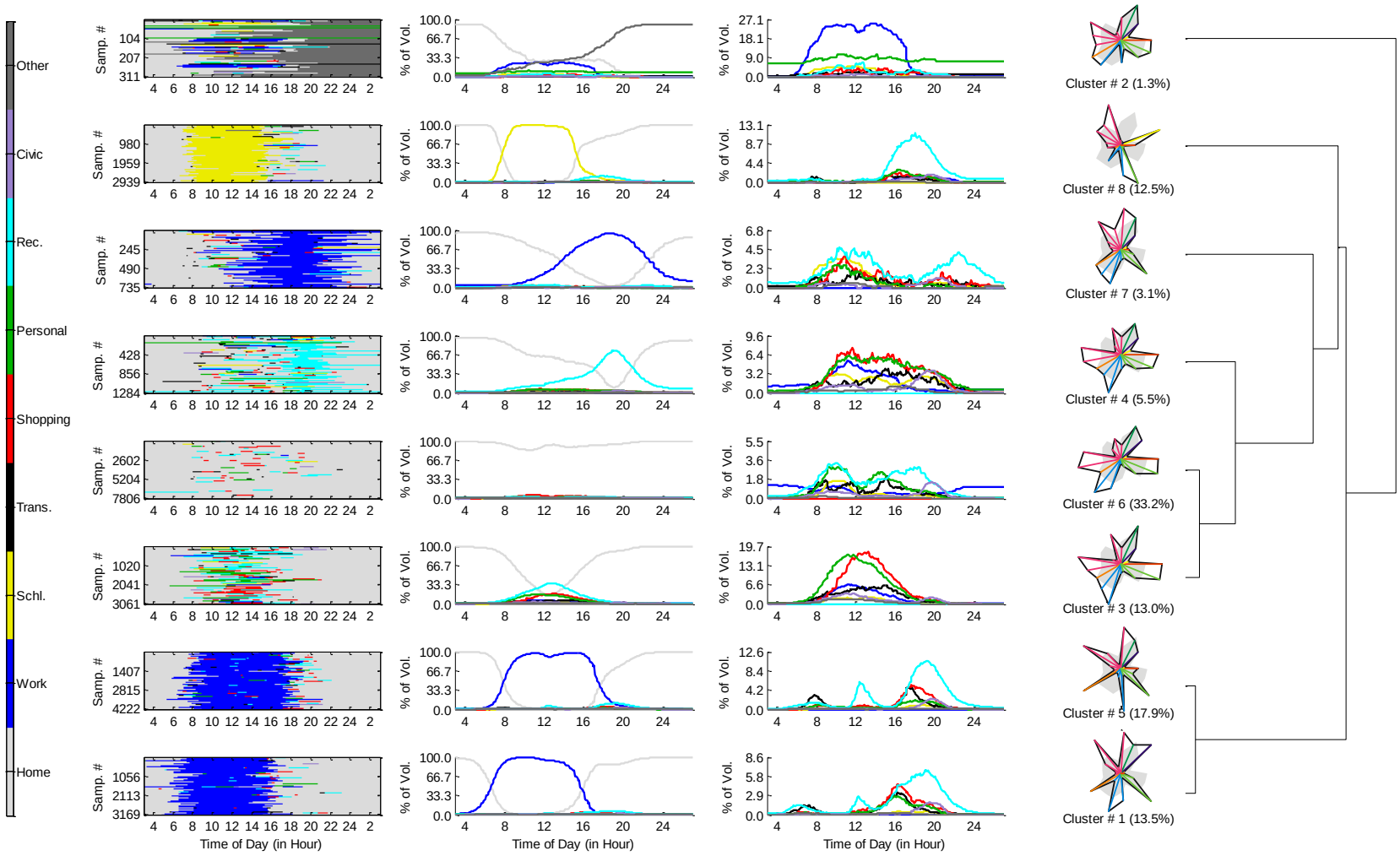

Figure 9 Clustering of individuals' weekday temporal activity patterns and their social demographic characteristics in Chicago (cluster number=8).

In the Star Diagram, we use different colors to represent different vectors of each cluster (corresponding to the numbered social demographic variables in Table 2), and also set the sample mean as the gray background for each cluster for comparison convenience.

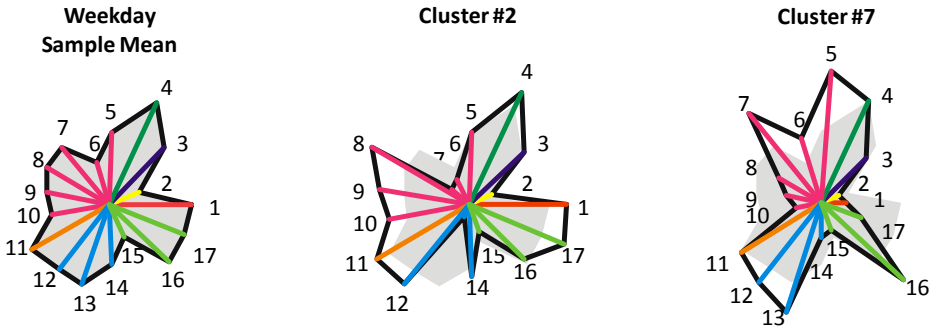

Figure 10 A demonstration of the Star Diagram of a cluster's social demographics on a weekday

Table 2 Statistics on social demographics of the total sample and each cluster on a weekday

	Social Demographic Variables	Weekday Sample Mean	Mean of Cluster							
			#2	#8	#7	#4	#6	#3	#5	#1
1	Female	53.3%	55.0%	49.0%	46.7%	57.3%	57.7%	58.9%	50.1%	44.9%
2	Student	22.1%	16.4%	91.1%	13.2%	22.5%	13.2%	15.7%	8.3%	7.2%
3	Homemaker	13.4%	12.9%	3.5%	11.1%	12.8%	15.1%	12.0%	13.4%	21.3%
4	Retired	59.7%	66.3%	8.9%	61.1%	58.3%	62.5%	70.7%	54.9%	53.2%
5	Work	53.4%	51.1%	12.1%	95.4%	42.8%	33.7%	38.4%	94.1%	95.0%
6	Part Time	20.4%	15.9%	52.3%	28.6%	36.2%	29.7%	39.5%	10.9%	12.7%
7	No Flexibility	34.4%	24.2%	36.4%	43.0%	26.2%	27.8%	22.8%	32.8%	46.8%
8	Some Flexibility	42.3%	47.1%	44.4%	39.9%	36.0%	35.8%	37.2%	50.3%	40.8%
9	Much Flexibility	23.2%	28.7%	19.2%	17.1%	37.8%	36.4%	40.0%	17.0%	12.4%
10	Work at Home	8.1%	11.3%	5.9%	3.1%	17.5%	19.8%	15.4%	2.5%	1.8%
11	Edu.>Tech School	45.9%	55.3%	6.5%	48.3%	45.8%	40.3%	46.5%	72.5%	58.9%
12	Low HH Income	16.9%	19.6%	14.4%	19.4%	17.1%	24.3%	19.3%	7.8%	10.0%
13	Middle HH Income	32.5%	28.0%	28.1%	34.4%	35.1%	33.4%	35.6%	28.8%	35.0%
14	High HH Income	50.7%	52.4%	57.4%	46.2%	47.8%	42.3%	45.1%	63.4%	55.0%
15	The Young	34.6%	30.4%	92.4%	30.8%	32.5%	27.8%	26.1%	24.9%	20.3%
16	The Middle-aged	39.7%	38.0%	5.0%	55.7%	32.5%	32.7%	31.4%	62.3%	66.4%
17	The Older	25.8%	31.7%	2.5%	13.5%	34.9%	39.6%	42.4%	12.8%	13.3%

In the following paragraphs, we discuss specifically the classified temporal activity patterns of individuals in each of the eight clusters and their social demographic characteristics (shown in Figure 9).

Students: Cluster #8 consists of students who go to school during the day time, and go out for meal, recreation or entertainment starting from 3:00 p.m. to around 10:00 p.m., with a peak of 12% of them at around 7:00 p.m. This group shares 12.5% of the total

sample. The average annual household income for the cluster #8 group members is higher than the average weekday sample mean, and over 92% of the cluster #8 population falls within the demographic category of “the young”.

Regular Workers Cluster #5 is the group of workers who have a relatively regular schedule. They leave home for work at around 7:00 a.m. to 8:00 a.m. in the morning, and finish work at around 5:00 p.m. to 6:00 p.m. in the late afternoon. Some of them go out for meal or recreation at lunch break time. Some do similar activities in the late afternoon, with a peak of 5% them doing shopping at 6:00 p.m., and another peak of 12% dining, recreating or entertaining at around 7:00 p.m. There is also a small proportion of the group members engaged in “*transportation transition*” activities in the early morning (with a peak around 2% at roughly 8:00 a.m.) and late afternoon (with a peak around 4% at around 5:30 p.m.).

Early-Bird Workers Individuals in Cluster #1 have similar temporal activity pattern to those in Cluster #5, except for the overall time shift--members in Cluster #1 start their day about one hour earlier than folks in Cluster #5 in general. While the rhythms of other temporal activities of the two clusters are similar (such as recreational activities at noon or early evening, shopping in the late afternoon, or *transportation transition* activities in the early morning and late afternoon), Cluster #1 seems to have a lower share of peak volume in these activities compared to those of Cluster #5.

When we compare the social demographic characteristics of the two clusters (Clusters #1 and #5), the above observations make more sense. Compared with the total weekday sample mean, Cluster #5 has a greater share of males, a higher educational level, higher share with high annual household income, and a greater share of middle-aged population. But, compared with Cluster #5, Cluster #1 has an even higher share of males, a higher share of individuals with no flexible work schedule, a lower share of people with some or much flexible work schedule, a lower share of individuals with high education level (beyond the technical school degree), a lower share of individuals with high annual household income, a higher share of people with middle level household income, and a higher share of middle-aged population. In summary, the early-birds workers in Cluster #1 live generally less flexible lives and tend to have a lower educational level and household income level, and there are a greater proportion of them in the middle-aged group, compared to their counterparts of the regular workers in Cluster #5.

Afternoon Workers For members in Cluster #7, a large proportion of them work but their temporal activity rhythm are quite different from those in Clusters #1 and #5. The majority of them (64.6%) spend their morning at home, for a small proportion they go shopping (with a 3% peak at 11 a.m.) or do personal business (with a 2% peak at 10 a.m.) or do recreation (with 4.5% peak around noon time). Most of them start work around noon to early afternoon (from 12:00 p.m. to 1:00 p.m.) and finish work very late (from 10:00 p.m. in the evening till midnight or early the next morning). Some of them also do recreational activities after work in the evening (with a 4.5% peak at around 11:00 p.m.). Only 3.1% of the total weekday samples belong to this cluster. The social demographic characteristics of this group are somewhat similar to those of members in

Cluster #1 (the early-bird workers), except that Cluster #7 members have lower average educational level and annual household income level. The middle-aged population share of this cluster is higher than the weekday sample mean.

Overnight Adventurers We call Cluster #2 members “overnight adventurers” because only a quarter of them work during the day and the majority of members in this group do "other" activities (that are not specified in their survey report) from early afternoon till midnight. There are only 1.3% of the total weekday sample in this cluster, among which three quarters claim to have some or great schedule flexibility, and 11.3% work at home. Their educational level is higher than the population mean, yet lower than the regular workers and the early-bird workers. The share of the older population in this group is higher than the weekday sample mean.

Stay-at-home We call Cluster #6 members "stay-at-home" because they spend most of their time at home with only a few of them (3%) conducting personal business or recreational activities over the day. This cluster is large in size and constitutes 33.2% of the total weekday sample, and has a higher share of females, a lower average educational level, a lower household income level, and a higher share of the older population, compared to the total sample mean. It also has the greatest share of members who work at home (19.8%) compared to the other seven clusters. Members in this cluster also claim to have very flexible schedules.

Morning & Afternoon Adventurers Members in Clusters #3 and #4, are similar to "stay-at-home" persons in Cluster #6 except that a greater share of them go out for shopping, recreation and personal business either in the morning (the "morning adventurers ") or in the afternoon (the "afternoon adventurers ").

The majority of the Cluster #3 members stay at home most of the time, and only some of them go out in the morning for recreation/entertainment, social activities (with a peak around 30% of them at noon), for shopping and personal business (with a peak around 13% around noon). 6.6% of them do some work in the morning too. While most members of Cluster #4 stay at home during the day time, they start their recreational/ entertainment/ social activities in the late afternoon, with a peak of 66% of them at around 7:00 p.m. in the evening. A smaller proportion of Cluster #4 members do shopping or personal business during the day time (around 6% of peak volume). Cluster #3 and #4 members share similar social demographic characteristics. Compared to the total weekday sample mean, these two clusters have greater shares of females, lower shares of workers, higher shares of people whose schedule is flexible, higher shares of people who work at home, lower household income level, and higher share of the older population. In total, there are 13% of total weekday samples in Cluster #3, the "morning adventurers ", and 5.5% in Cluster #4, the "afternoon adventurers".

5 CONCLUSIONS AND DISCUSSION

In this paper, we analyze the activity patterns for 23,527 individuals in the Chicago Metropolitan Area, by dividing the entire day into 288 five-minute intervals. We define

the eigenvectors of the covariance matrix of activity data as *Eigenactivities*, which are a set of vectors that span an 'activity space' and characterize the differences between individuals' temporal activities in the metropolitan area. A linear combination of the metropolitan area's eigenactivities can accurately reconstruct the activity pattern of each individual. Based on a small activity reconstruction error (1%), we select 21 primary eigenactivities to represent individuals' weekday temporal activities in the metropolitan area. We perform a K-means clustering algorithm on the obtained eigen decomposition projections to partition the 23,537 weekday time series samples into k clusters By reducing the dimension of the problem with a small number of eigenactivites, we lower the computational cost of the algorithm.

We successfully classify individuals in the metropolitan area into the following groups within which they have relatively homogeneous temporal activity patterns, and across which they have heterogeneous diversity: students (12.50%), regular workers (17.90%), early-bird workers (13.50%), afternoon workers (3.10%), overnight adventurers (1.30%), afternoon adventurers (5.50%), morning adventurers (13.00%), and stay-at-home (33.20%). We identify the signatures of the social demographic profile of each of the clusters. In general, we find that, when compared with the weekday sample mean, the "adventurers" have a higher share of female, a lower share of students, a lower share of people who work, a higher share of working at home, and a higher share with much work flexibility. They also tend to have lower educational level, lower household income level, and higher share of the older population. For the workers, there is a lower share of females, higher share of people who have no flexibility in work schedule, relatively higher education, higher household income level, and higher share of the middle-aged population, compared to the weekday sample mean. A similar approach was used to classify weekend activities, but those results are beyond the scope of this paper.

This paper provides a new approach for studying temporal patterns of human activities in the metropolitan area, and will be useful for urban and transportation planning. For example, traditional studies on measuring individual's accessibility to urban opportunities tend to ignore individuals' temporal activity differences (Hanson and Kwan 2008), and treat metropolitan residents either as more homogeneous groups or pre-specified subgroups differentiated by social characteristics (Handy 1993; Shen 1998). This study provides methods for classifying people based on their temporal activity patterns and allows urban researchers to construct activity-based signature of daily travel patterns for different types of individuals without heavy-burdened computational costs. Our method provides a straight forward approach for better understanding the individuals' temporal activity patterns beyond the limited time periods that the traditional methods permit.

The framework of our study also allows us to link the temporal dimension with the spatial dimension, as we not only transform the traditional travel and activity survey into individuals' activity types at each time interval but also impute their location information (latitude and longitude). With more spatially detailed GIS data (such as land use data, points-of-interests data) and the most recently available and attractive massive urban sensing data (such as high resolution orthophotos, cell phone data, and data from the intelligent transportation systems), combined with data mining and statistical learning methods, we will be able to probe questions which are essential but complicated to

answer. For example, visualizing the travel patterns of our sample indicated that workers with time flexibility tend to live in different neighborhoods, and tend to follow different daily patterns for undertaking non-work activities. Knowing more about the links between land use and activity patterns could facilitate congestion management and improve models that simulate travel patterns under different road capacity, travel cost, and land development circumstances. Further study could also address questions such as *"Can we change the individuals' spatiotemporal activity distributions by changing the distribution of land use (so as to reduce trip length, vehicle miles traveled, congestion, energy consumption and air pollution…)?"* Answering these questions is important, as it helps determine how planners should design more attractive, efficient, equitable, and healthy cities in order to enable sustainable futures for our current and next generations (Hensher and Button 2003; Yoon et al. 2009).

Acknowledgments

The authors appreciate the public release of the anonymized activity survey by the Chicago Metropolitan Agency for Planning, and the partial support of the MIT Urban Studies and Planning Department and the MIT/Portugal Program. We also acknowledge the comments and feedback from our research colleagues and Jameson Lawrence Toole and other participants in the SC/OM seminar at MIT.

References

Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfüser, G., and Haupt, T., 2002. Observing the rhythms of daily life: A six-week travel diary. *Transportation,* 29 (2), 95-124.

Batty, M., 2003. New developments in urban modeling: Simulation, representation, and visualization. *In: Integrated land use and environmental models: A survey of current applications and research*. New York: Springer, 13-46.

Becker, G. S., 1991. *A treatise on the family.* Harvard University Press.

Bhat, C. R., and Koppelman, F. S., 1999. A retrospective and prospective survey of time-use research. *Transportation,* 26 (2), 119-139.

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., and Dougherty, E. R., 2007. Model-based evaluation of clustering validation measures. *Pattern Recognition,* 40 (3), 807-824.

Chapin, F. S., 1974. *Human activity patterns in the city: Things people do in time and in space.* New York: Wiley.

Chicago Metropolitan Agency for Planning. 2008. Chicago travel tracker household travel inventory (2007).

Crane, R., and Sornette, D., 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences,* 105 (41), 15649-15653.

Duda, R. O., Hart, P. E., and Stork, D. G., 2001. *Pattern classification.* New York: Wiley.

Dunn, J. C., 1973. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics,* 3 (3), 32 - 57.

Eagle, N., and Pentland, A., 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology,* 63 (7), 1057-1066.

Eagle, N., Pentland, A., and Lazer, D., 2009. Inferring friendship network structure by

using mobile phone data. *Proceedings of the National Academy of Sciences*.

Freud, S., 1953. *Collected papers.* Vol. IV: London: Hogarth Press and The Institute of Psychoanalysis.

Geerken, M., Gove, W. R., and Relations, N. C. O. F., 1983. *At home and at work: The family's allocation of labor.* Sage Publications ; Published in cooperation with the National Council on Family Relations.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L., 2008. Understanding individual human mobility patterns. *Nature,* 453 (7196), 779-782.

Greaves, S., 2004. Gis and the collection of travel survey data. *In:* Hensher, D. A. ed. *Handbook of transport geography and spatial systems*: Elsevier.

Hägerstrand, T., 1989. Reflections on “what about people in regional science?”. *Papers in Regional Science,* 66 (1), 1-6.

Handy, S., 1993. Regional versus local accessibility: Implications for nonwork travel. *Transportation Research Record: Journal of the Transportation Research Board* (1400), 58-66.

Hanson, S., and Hanson, P., 1980. Gender and urban activity patterns in uppsala, sweden. *Geographical Review,* 70 (3), 291-299.

Hanson, S., and Kwan, M.-P. (Eds.). 2008. *Transport: Critical essays in humman geography* (1 ed.).

Harvey, A., and Taylor, M., 2000. Activity settings and travel behaviour: A social contact perspective. *Transportation,* 27 (1), 53-73.

Hastie, T., Tibshirani, R., and Friedman, J. H., 2009. *The elements of statistical learning: Data mining, inference, and prediction.* Springer.

Hensher, D. A., and Button, K. J., 2003. *Handbook of transport and the environment.* Elsevier.

Kargupta, H., and Han, J. (Eds.). 2009. *Next generation of data mining*: CRC Press.

Kim, M., Kotz, D., and Kim, S., 2006. *Extracting a mobility model from real user traces.* Paper presented at the IEEE INFOCOM’06, Barcelona, Spain.

Maslow, A. H., and Frager, R., 1987. *Motivation and personality.* Harper and Row.

Nature Editorial. 2008. A flood of hard data. *Nature,* 453 (7196), 698-698.

Reggiani, A., and Nijkamp, P. (Eds.). 2009. *Complexity and spatial networks: In search of simplicity*: Springer.

Rousseeuw, P. J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* 20, 53-65.

Shen, Q., 1998. Location characteristics of inner-city neighborhoods and employment accessibility of low-wage workers. *Environment and Planning B: Planning and Design,* 25 (3), 345-365.

Yang, J., and Leskovec, J., 2011. Patterns of temporal variation in online media. *In: Proceedings of the fourth ACM international conference on Web search and data mining,* Hong Kong, China: ACM.

Yoon, S. Y., Golob, T. F., and Goulias, K. G., 2009. *California statewide exploratory analysis correlating land use density, infrastructure supply, and travel behavior.* Paper presented at the TRB 88th Annual Meeting Compendium of Papers DVD

Yu, H., and Shaw, S.-L., 2008. Exploring potential human activities in physical and virtual spaces: A spatio-temporal gis approach. *International Journal of Geographical Information Science,* 22 (4), 409 - 430.