Clustering Daily Patterns of Human Activities in the City

Shan Jiang

Department of Urban Studies and Planning, Massachusetts Institute of Technology 77 Massachusetts Ave. E55-19E Cambridge, MA 02142

Email: shanjang@mit.edu

Joseph Ferreira

Department of Urban Studies and Planning, Massachusetts Institute of Technology 77 Massachusetts Ave. 9-532 Cambridge, MA 02139

Email: jf@mit.edu

Marta C. González

Department of Civil and Environmental Engineering and Engineering Systems Division, Massachusetts Institute of Technology 77 Massachusetts Ave. Room 1-153 Cambridge, MA 02139

Email: martag@mit.edu

Abstract Data mining and statistical learning techniques are powerful analysis tools yet to be incorporated in the domain of urban studies and transportation research. In this work, we analyze an activity-based travel survey conducted in the Chicago metropolitan area over a demographic representative sample of its population. Detailed data on activities by time of day were collected from more than 30,000 individuals (and 10,552 households) who participated in a 1-day or 2-day survey implemented from January 2007 to February 2008. We examine this large-scale data in order to explore three critical issues: (1) the inherent daily activity structure of individuals in a metropolitan area, (2) the variation of individual daily activities—how they grow and fade over time, and (3) clusters of individual behaviors and the revelation of their related socio-demographic information. We find that the population can be clustered into 8 and 7 representative groups according to their activities during weekdays and weekends, respectively. Our results enrich the traditional divisions consisting of only three groups (workers, students and non-workers) and provide clusters based on activities of different time of day. The generated clusters combined with social demographic information provide a new perspective for urban and transportation planning as well as for emergency response and spreading dynamics, by addressing when, where, and how individuals interact with places in metropolitan areas.

Keywords: *Human Activity, Eigen Decomposition, Daily Activity Clustering, Metropolitan Area, Statistical Learning*

Preliminary 2011 version of paper ultimately published in Data Mining and Knowledge Discovery: Volume 25, Issue 3, pages 478-510, Article DOI: 10.1007/s10618-012-0264-z, (2012)

1 Introduction

Considerable efforts have been put into understanding the dynamics and the complexity of cities (Reggiani and Nijkamp 2009; Batty 2005). To our advantage, in general, individuals exhibit regular yet rich dynamics in their social and physical lives. This field of study 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 the advantages of interdisciplinary approaches and rapid technology innovations (Foth et al. 2011; Portugali et al. 2012). Emerging urban sensing data such as massive mobile phone data, and online user-generated social media data, both in the physical 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 human allocate time to different activities as part of a spatial, temporal socio-economic system*," becomes treatable within an interdisciplinary domain. By clustering individuals according to their daily activities, our ultimate goal is to provide a clear picture of how groups of individuals interact with different places at different time of day in the city.

The advances of our study lie in two folds. First, we do not superimpose any predefined social demographic classification on the observations, but use the presented methodology to cluster the individuals. This provides an advantage over traditional human activity studies, which tend to treat metropolitan residents either as more homogeneous groups or pre-specified subgroups differentiated by social characteristics (Shen 1998; Sang et al. 2011; Kwan 1999). We let the inherent activity structure inform us of the patterns in order to generate the clusters of daily activities in a metropolitan area. Second, compared with recent studies on human mobility and dynamics employing large-scale objective data such as mobile phone or GPS traces of individual trajectories (Wang et al. 2011a; Song et al. 2010; Gonzalez et al. 2008; Candia et al. 2008), we linked in the usually absent rich information regarding activity categories and social demographics of individuals. By summarizing the socio-demographic characteristics of each cluster, we try to reveal the social connections and differences within and among each activity cluster. The scope of our results can be applied to inform diverse areas that are concerned by models of human activity such as: time-use studies, human dynamics and mobility analysis, emergency response or epidemic spreading. We hope that this work connects with researchers in urban studies, computer sciences and complex systems, as a case of study of how interdisciplinary research across these fields can produce useful pieces of information to understand city dynamics.

The rest of the paper is organized as follows. In Section 2 we survey the literature of related studies. Section 3 describes the data that we are using in this

Preliminary 2011 version of paper ultimately published in Data Mining and Knowledge Discovery: Volume 25, Issue 3, pages 478-510, Article DOI: 10.1007/s10618-012-0264-z, (2012)

study, and our data processing methodology. In Section 4, we provide the mathematical framework and justify the selected methods of analysis, including the principle component analysis (PCA) to extract the primary eigen activities, the *K*-means clustering algorithm, and the cluster validity measurement that we propose to use to identify the number of clusters. We present our findings on the eigen activities, clustering of daily activity patterns, and their associated socio-demographic characteristics in Section 5, and conclude our study and summarize its significance and applications for future work in Section 6.

2 Background and Related Work

Different facets of spatiotemporal characteristics of human activities have long been studied by researchers in sociology (Geerken and Gove 1983), social ecology (Chapin 1974; Taylor and Parkes 1975; Goodchild and Janelle 1984), psychology (Freud 1953; Maslow and Frager 1987), geography (Hägerstrand 1989; Yu and Shaw 2008; Harvey and Taylor 2000; Hanson and Hanson 1980; Hanson and Kwan 2008), economics (Becker 1991, 1965, 1977), and urban and transportation studies (Ben-Akiva and Bowman 1998; Bhat and Koppelman 1999; Axhausen et al. 2002). Nowadays, studies in these fields can benefit from recent innovation in both data sources and analytical approaches, which have inspired a new generation of studies about the dynamics of human activities. For example, Gonzalez et al. (2008) studied the trajectories of 100,000 anonymized mobile phone users, and showed a high degree of spatial regulatirty of human travels. Eagle and Pentland (2009) analyzed continous mobile phone logging locations collected from an experiment at MIT, studied the behavioral structure of the daily routine of the students, and explored individual community affiliations based on some apriori information of the subjects. Song et al.(2010) measured the entropy of individuals' trajectory using mobile phone data, and found high predictability and regularity of users daily mobility. Wang et al. (2011a) tracked trajectories and communiction records of 6 million mobile phone users, and examined how individual mobility patterns shape and impact their social network connections.

Due to privacy and legal constraints, these kinds of studies generally face challenges in depicting a whole picture that connects behavior with social, demographic and economic characteristics of the studied subjects. While the new datasets allow us to study massive aggregated travel behavior and social interactions, they have limited capacity in revealing the underlying reasons driving human behavior (Nature Editorial 2008). In order to have details, usually we must limit group sizes. For example, Eagle et al.(2009) used the Reality Mining data to infer friendship network structure. The data mining technique of this study is very promising but, without socioeconomic information, it is hard for researchers to further explore the determinig factors beneath the network, especially when the constraint imposed on a specific community (such as

university campus), and the scale are enlarged to include entire metropolitan area and beyond.

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 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 socioeconomic and demographic information, survey data provides much richer 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 regional transportation planning purposes, it is free for public access, reliable, and representative of the total regional population. Daily activities of groups of individuals in cities 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). To those means, in this work we show that the PCA/eigen decomposition method (Turk and Pentland 1991) and *K*-means clustering algorithm (Ding and He 2004) are appropriate to analyze urban survey data. These techniques are successfully applied to reconstruct the original data sets and obtain meaningful clusters of individuals. We provide a rich, yet simple enough, set of activity clusters, with additional time-of-day information, which go beyond the traditional simply defined groups and can be adopted by current urban simulators (Waddell 2002; Balmer et al. 1985; Bekhor et al. 2011). The kind of analyses presented here is also useful to compare and understand the dynamics of different cities.

3 Data

In this section, we describe the activity survey data in the Chicago metropolitan region and our techniques for processing the data. From the survey data, we derive two separated sample sets (i.e., for an average weekday and weekend). For each of the sets we know detailed information about individuals' daily activity sequences, and their social demographics. For simplicity reasons, we aggregate the 23 self-reported primary activities into 9 major activities. We divide the 24 hours into 288 five-minute intervals for further data analysis.

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.

3.1 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. Based on similarities between some of the 23 primary purposes in the original survey data, we aggregate them into fewer activity types that are widely adopted in urban studies and transportation planning (Bowman and Ben-Akiva 2001; Axhausen et al. 2002) as shown in Table 1. We also use a specific color for each activity throughout the entire paper.

Table 1 Aggregated 9 activity types v.s. the original 23 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

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".

3.2 Human Daily Activities on Weekdays and Weekends

We generate a separate animation visualizing the movement and activities (differentiated by nine colors demonstrated in Table 1) of the surveyed individuals in the Chicago metropolitan area for an average weekday and weekend (please see supplemental multimedia materials). Since the public location data for each destination that an individual visited is anonymized by the centroid of the census tract, for visualization purposes, we differentiate destinations by adding a very small random factor (see Figures 1 and 2).

An Average Weekday

We use the first-day sample of the 1-day survey distributed from Monday to Thursday, plus the second-day sample of the 2-day survey distributed on Sunday 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, with Friday approaching to the weekend, patterns of human activities on that day 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.

An Average Weekend

For an average weekend (Saturday or Sunday), we get a smaller sample compared to that of weekday, totaling of 5,481 distinct individuals. We can see that the activity patterns of a weekend are very different from those during weekdays (see Figure 2). During the early morning, majority of the people are at home while a few are out for recreation or still at work. At noon time, many people have been

out for recreation/entertainment, shopping or civic (religious) activities, and some are staying at home and a small proportion people are at work. In the early evening, the majority people who are not at home are doing recreation or entertainment, while some are doing shopping. At midnight, while most people are at home, a few are out for recreation/entertainment, mostly concentrated in the downtown area.

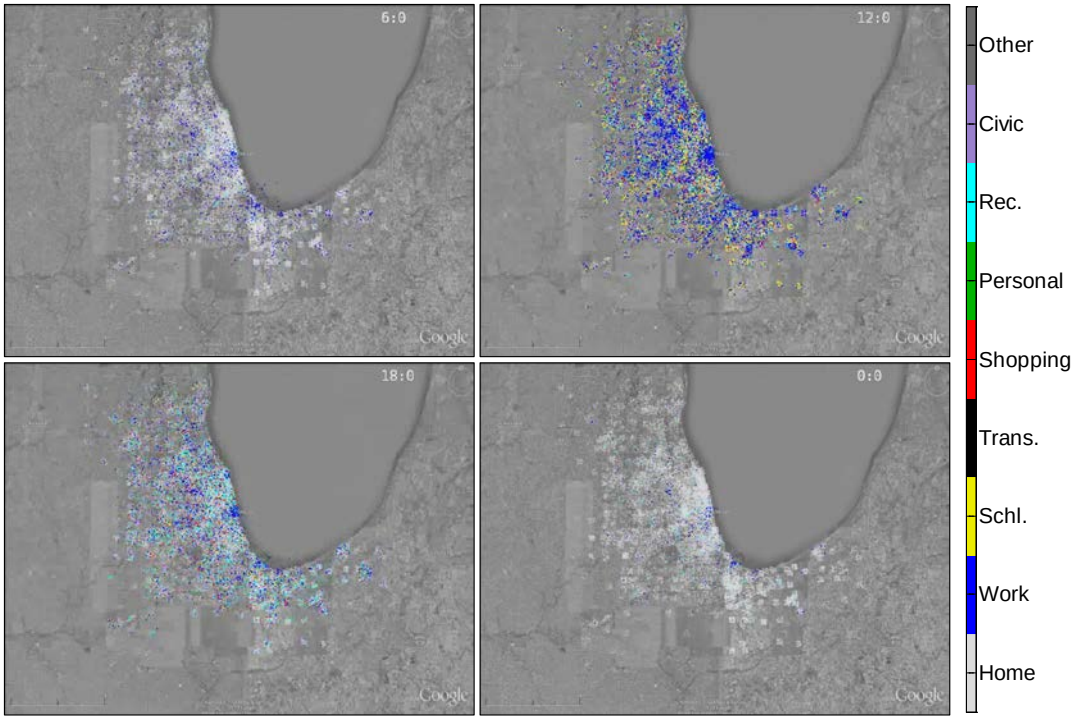

Figure 1 Snapshots of human activities at different times-of-day on a weekday in Chicago

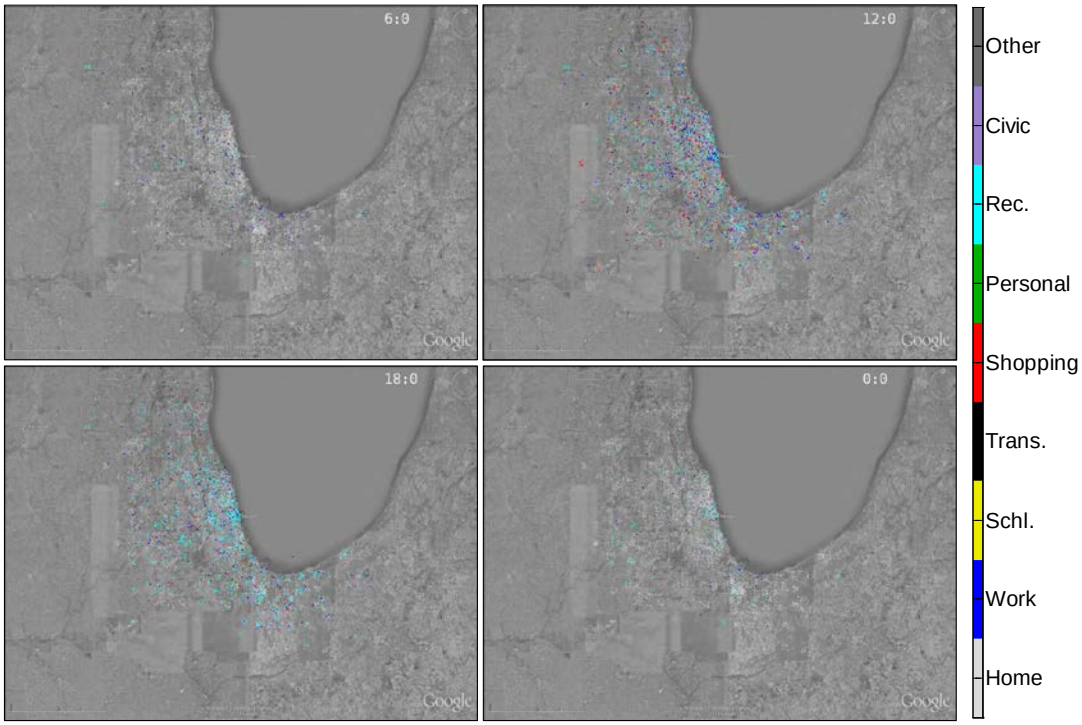

Figure 2 Snapshots of human activities at different times-of-day on a weekend in Chicago

Individual and Aggregated Daily Activity Variations

Figures 1 and 2 provide us with a sensible landscape about individual's daily activities in the metropolitan area. Nevertheless, we need additional tools to analyze the composition of individuals conducting different activities over time. By exhibiting the activity-type change along the time axis for every individual in the sample, we are able to retain rich information about individual activity variation at different time of day. In Figure 3, we depict respectively, for an average weekday and weekend, the 24-hour human activity variations (using the corresponding colors defined in Table 1) in Chicago. 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 all samples (i.e., each line parallel to the *x* axis represents an individual sample). By summing up the total number of individuals conducting different types of activities along the 24-hours of the weekday and weekend, we are able to generate Figure 4, which reveals the aggregated temporal variation of human activities in Chicago. In addition, each inset figure zooms in on the detailed information of the less-major activities (i.e., those with a smaller share of total volume) over time.

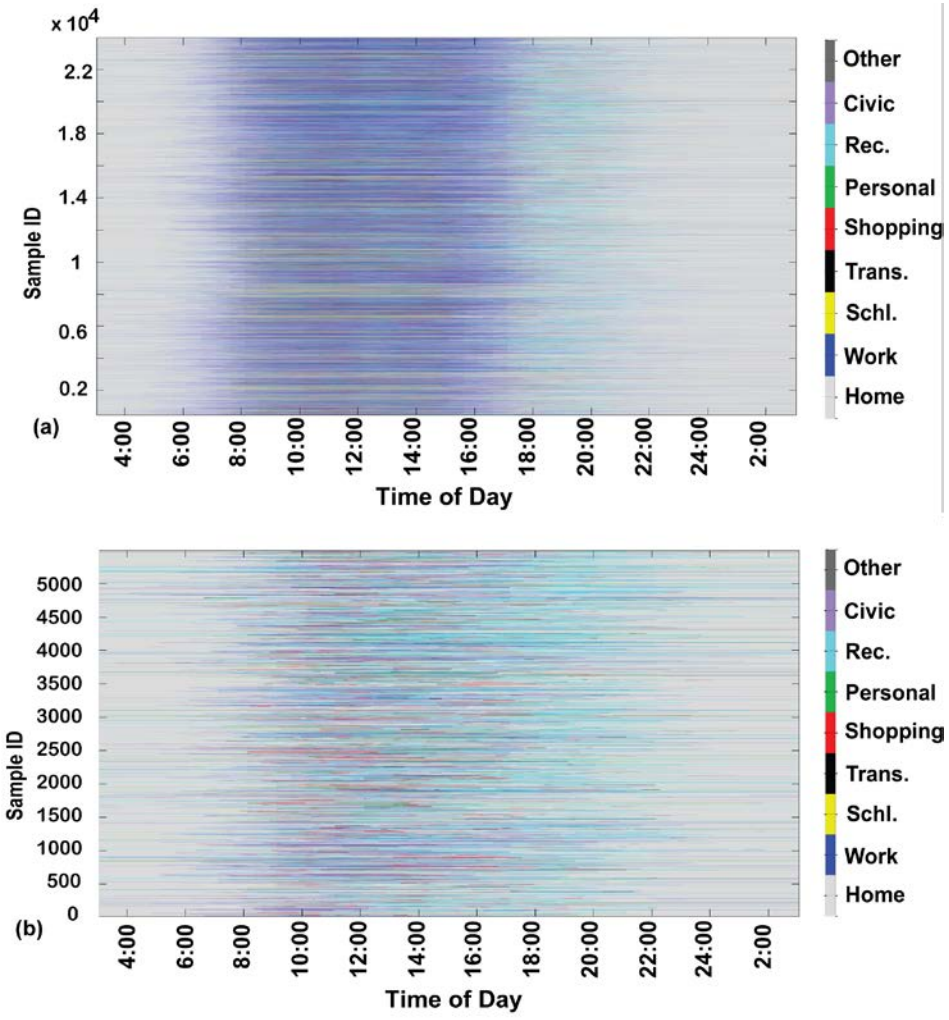

Figure 3 Individual daily activities on a (a) weekday and (b) weekend in Chicago

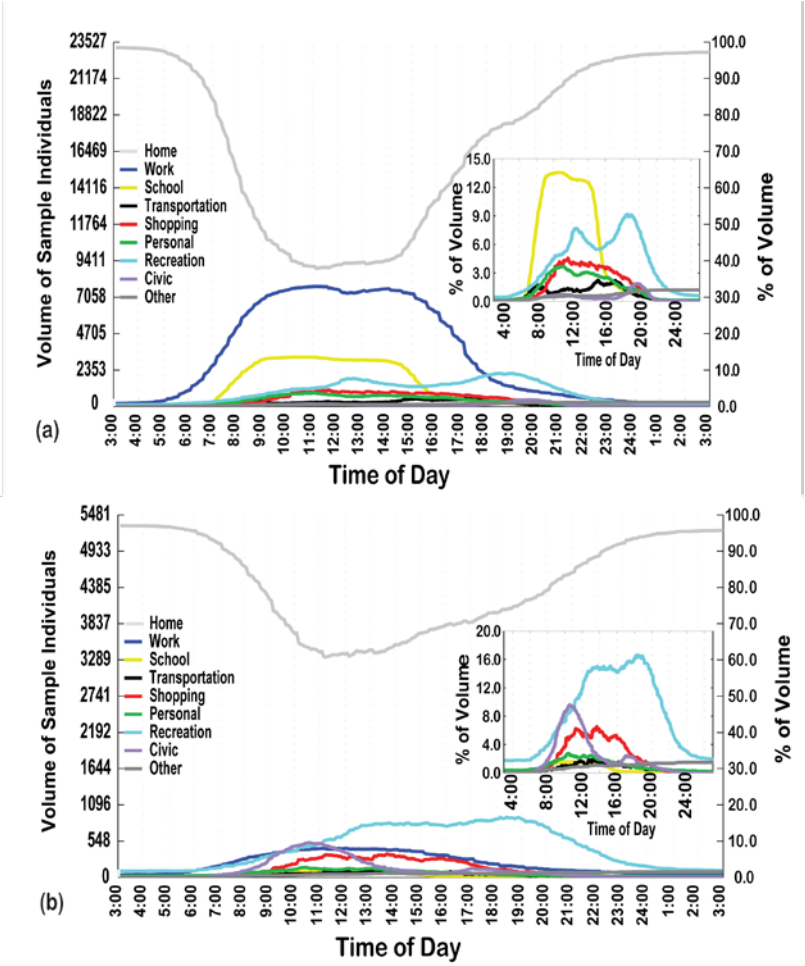

Figure 4 Temporal rhythm of human activities on a (a) weekday and (b) weekend in Chicago

3.3 Data Transformation

We divide the 24 hours in a day into five-minute intervals and use the activity in the first minute of every time 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 interval. In Figure 5, 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.

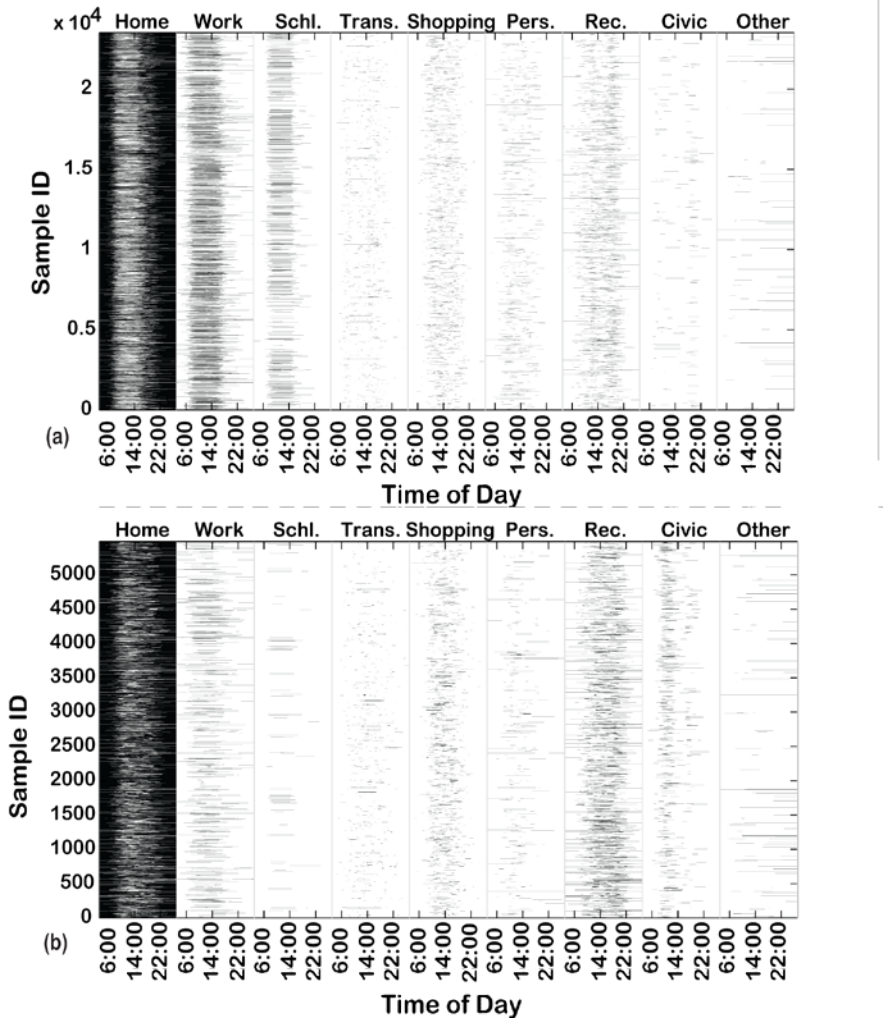

Figure 5 Data transformation of individual activities on a (a) weekday and (b) weekend in Chicago

4 Mathematical Framework and Methods

We employ two methods, namely, the principal component analysis/eigen decomposition and the *K*-means clustering algorithm, to answer the two questions raised earlier in this paper: (1) discovering the inherent daily activity structure of individuals in the metropolitan area; and (2) clustering individuals in the metropolitan area based on dissimilarity of their daily activities.

4.1 The Setting

During any of the 288 five-minute time intervals, an individual must conduct one of the nine activities defined in Table 1. For , we say that satisfies the *compatibility condition*, if for any t = 1, 2, ..., 288, . We define the *space of individuals' daily activity sequence*, , as follows:

.

In this study, the population is the set of individuals in the Chicago metropolitan area. For simplicity, we identify the sample space as the

population. As we study the average weekday and weekend separately, we have two cases. For the weekday case, an individual's daily activity sequence can be described by the following random vector:

.

where for , and , or 1, depending on if the individual is conducting activity *l* in time interval *t* on the weekday. We can define the random vector for the weekend case similarly.

From the survey data, we get a random sample of *n* observations , *i* = 1, …, *n*, where stands for individual *i*'s social demographic information such as age, gender, employment status, work schedule, etc. Note that for a sample individual *i*, we may only observe (i.e. we do not have information on his/her weekend activity) as explained in the data description section. Let O_D and O_E denote the sets of samples where and are observed, respectively. For the weekday (weekend) case, we focus on set O_D (O_E), and renumber the samples in O_D (O_E) from 1 to n_D (n_E), where n_D = 23,527 (n_E = 5,481). As the analytical approaches for the weekday and weekend are the same, henceforth we use the weekday case as an illustration, in which we have observations , *i* = 1, …, *n*, where . We omit the subscript "D" in notations for simplicity when there is no ambiguity.

4.2 Principal Component Analysis/Eigen Decomposition

Principal component analysis (PCA) and eigen decomposition are closely related as 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 denote the deviation from the mean, i.e., , where – is the sample mean. Therefore the sample covariance matrix is given by — — **,** where .

Eigenactivities

We know that is a positive semi-definite matrix, which is diagonalizable. So all the eigenvalues of are nonnegative. Let , where , and is an orthogonal matrix whose *j*-th column is the eigenvector corresponding to . For convenience, we arrange the eigenvalues in descending order, i.e., . We call eigenvector the *j*-th *eigenactivity*.

Projection onto Eigenactivities

As forms an orthonormal basis for , becomes the corresponding change of coordinate matrix. Namely, given a vector whose coordinate with respect to the natural basis is , the corresponding -coordinate will be given by . When for a sample individual *i*, we call ($j = 1, \dots, m$) the *projection of onto the j-th eigenactivity*, which is the projection of onto the *j*-th eigenactivity.[1]

Activity Reconstruction

Having known the eigenactivities and the corresponding projections of an individual *i*'s daily activity deviation from the mean, we can reconstruct the individual's daily activity sequence , by using a subset of eigenactivities. Suppose the projection of onto the first *h* eigenactivities are **,** then we obtain a vector according to formula . We use the following algorithm to reconstruct an individual's daily activity sequence as .

- Given any , let .
- Define so that if and only if .[2]
- So we get a 9-dimensional vector that has one entry of 1, and we let . It turns out that the reconstructed satisfies the desirable relation .[3]

The Appropriate Number of Eigenactivities

To answer the question "*how many eigenactivities sufficient to rebuild the original daily activity structure"*, we define the reconstruction error for as the ratio of the number of incorrectly reconstructed entries to the total number of entries, i.e., ———. 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 is no greater than ε. Let $\varepsilon_0 > 0$ be the

---

[1] We can also consider the *projection* of the random vector onto the *j*-th eigenactivity. Namely, let , then ($j = 1, \dots, m$) is called the *j-th principal component (of the population)*, which is the *projection* of onto the *j*-th eigenactivity. By the Strong Law of Large Numbers (SLLN), we can show that almost surely as . For detailed discussion about SLLN, readers may refer to Durrett (2005). In this study, the sample sizes are large (n_D = 23,527 and n_E = 5,481), so the principal components are uncorrelated with each other. Note that in this study we do not use the principal components ( ), but the projections of onto the *j*-th eigenactivity ( ).

[2] In the generic case, has exactly one entry of value 1. When has more than one entries that are equal to 1, it must be the case that there are more than one such that . In such a case, which is extremely rare or never happens, we keep the first entry 1 and change the others into 0.

[3] This relation can be proved by a discussion of the relative positions of with respect to 0 and 1. This property justifies our reconstruction algorithm and can also be used to derive an equivalent alternative reconstruction algorithm.

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*.

Validity of Applying PCA in this Study

When PCA is used, the distribution of the original data is usually assumed to be multivariate Gaussian. The advantage of multivariate Gaussian assumption lies in that the principal components are not only uncorrelated but also mutually independent. When the principal components are independent of each other, it ensures that different components are measuring separate things, and the high dimensional distribution of the original data can be easily revealed by the distribution of each component, as the product measure can be easily constructed from the distribution measure of each component.

In this study, the original data is a Cartesian product of binary random variables, whose distribution is clearly non-Gaussian. However, as our purpose is not to run a regression or to reveal the high dimensional probability distribution of the original data using those principal components, the independence between components is not necessary.[4] For this reason, PCA/eigen decomposition is widely employed for dimension reduction in similar types of studies (Eagle and Pentland 2009; Turk and Pentland 1991; Calabrese et al. 2010).

The special binary property of the original data in this study strengthens the power of eigen decomposition. Each entry of the random vector only takes values 0 and 1, and satisfies the compatibility condition mentioned previously. The reconstruction algorithm introduced above takes full advantage of it. In order for the reconstructed equal to the original observation , there is no need for to be very close to , which usually requires a large number of eigenactivities to be employed in the reconstruction. Instead, we only need that for any , is the largest in , when . This property greatly lowers the threshold for accurate reconstruction, which ensures low reconstruction errors by using just a small number of eigenactivities.

4.3 Daily Activity Clustering

To answer the second question raised in the beginning of Section 4, we propose to use the *K*-means clustering algorithm, one of the most popular iterative clustering methods, to partition individuals in the metropolitan area into clusters based on their daily activity dissimilarity. In our study, although each of the observations bears a "time stamp", they are not repeated observations of the same phenomenon. In other words, the data is not time series data intrinsically, and therefore we do

[4] In fact, Gaussian assumption is not necessary for PCA. Readers may refer to Jolliffe (2002) (page 396) for discussion about Gaussian assumption and the relationship between PCA and *independent component analysis* (ICA).

not employ time series clustering method here.[5] However, time series clustering method could be appropriate for other related research in clustering human motions (Li and Prakash 2011) when repeated observations are available.

K-Means Clustering and Categorical/Binary Data

The *K*-means algorithm has been widely applied to partition datasets into a number of clusters (Wu et al. 2008). It performs well for many problems, particularly for numerical variables that are normal mixtures (Duda et al. 2001; Bishop 2009). While its definition of "means" in some cases limits the *K*-means application and leaves categorical variables not easy to treat (Xu and Wunsch 2008), a few studies have explored various ways to tackle this issue (Huang 1998; Ordonez 2003; Gupta et al. 1999). Ralambondrainy (1995) proposes to convert multiple categorical data into binary data (indicating if an observation is in the specified category) and treat the binary attributes as numeric in the *K*-means algorithm to cluster categorical data. Huang (1998) criticizes that the drawback of Ralambondrainy's approach is the tremendous computational cost, since it needs to handle a large number of binary attributes, especially when the number of categories are large. Huang (1998) presents two variations of the *K*-means algorithm (i.e., *k*-modes, and *k*-prototypes) for clustering categorical data. For similar motivations, Ordonez (2003) presents three variations of the *K*-means.

K-Means Clustering via PCA

For our study, as discussed previously, we assume that within each of the 288 five-minute intervals of the entire day, an individual conducts one of the 9 types of activities. We then convert the 288 entries of the categorical attributes of an individual's daily activity into a 2592-dimensional binary vector. Our data transformation process is similar to what Ralambondrainy (1995) proposes, which allows us to apply the *K*-means clustering algorithm.

When considering the options of dissimilarity measurements between individuals' daily activity sequences, if following the most natural approach, one could calculate the Euclidean distance between the two 2592-dimensional vectors. However, as Huang (1998) points out, when the number of binary attributes is large (in our case, 2592 dimensions), the computation cost is very high. Alternatively, as PCA/eigen decomposition can reduce the dimension of the problem significantly, a better approach is to measure the Euclidean distance between the $h(\varepsilon_0)$-dimensional vectors and , where and are the projection of and onto the first $h(\varepsilon_0)$ eigenactivities. Since the change of orthonormal bases does not affect the Euclidean distance no matter whether the original or the new coordinates are used, the Euclidean distance

[5] For instance, the data about an individual's activity in the time interval 6:55-7:00 a.m. and that in 7:00-7:05 a.m. can't be viewed as two consecutive observations of one phenomenon. Instead, they should be viewed as one observation of two phenomena that happen consecutively in time.

obtained from the coordinates of reduced dimension via PCA is very close to the original one, and can be used as the dissimilarity measurement between individuals' daily activity sequences.

In the latter approach, the original 2592 dimensions will be reduced to a much smaller dimension $h(\varepsilon_0)$, and the computational cost is significantly lowered, while the accuracy of clustering results is still maintained. Many studies have demonstrated the successfulness of applying the *K*-means algorithm via PCA (Ding and He 2004; Zha et al. 2001), and our study illustrates the effectiveness of applying *K*-means via PCA when having categorical/binary data. The readers can also see from later sections of the paper that our clustering results are very significant and intuitively meaningful.

Cluster Validity

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 (Halkidi et al. 2001). 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 do not have pre-specified cluster 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) which maximizes inter-cluster distances while minimizing the intra-cluster distances**,** and Silhouette index (Rousseeuw 1987) which reflects the compactness and separation of clusters to help us select the optimal number of clusters. A higher value of Dunn or Silhouette index indicates a better clustering result.

5 Findings: Patterns of Human Daily Activity

In this section, we present our findings of the human activity patterns on an average weekday and weekend in the Chicago metropolitan area. (1) We compute the eigenactivities of all the sample individuals on an average weekday and weekend to indentify the inherent daily activity structure of individuals in a metropolitan area. (2) By using the *K*-means clustering algorithm, we cluster the individual daily activities patterns for an average weekday and weekend, and their variation of daily activity types. We summarize the social demographic characteristics for each group of individuals, and find distinct patterns among the individuals within each group.

5.1 Eigenactivities

By employing the principal component analysis method discussed in the previous session, we derived the eigenactivities for an average weekday and weekend in the Chicago metropolitan area. Due to limited space, in this section we only display the first three eigenactivities for both the weekday and weekend cases.

Weekday

Figure 6 shows the first three eigenactivities of individuals in Chicago on an average weekday. We see that the first weekday eigenactivity (the 1st column of Figure 6) mainly describes the high probability of working (and low probability of staying at home) from 7:00 a.m. till 5:00 p.m. compared to the sample mean. The direction of the first eigenactivity on a weekday accounts for the largest variance of individuals' daily activities of the weekday data, which means that the major difference of individuals' daily activities on a weekday is if 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 6) reveals a high probability of schooling from 8:00 a.m.to 3:00 p.m. combined with a low probability of either staying at home during the same time period or working from 8:00 m to 5:00 p.m. (when compared to the sample mean in the data). The second eigenactivity direction accounts for the largest variance that is orthogonal to the first eigenactivity. The third weekday eigenactivity (the 3rd column of Figure 6) portrays a high probability of staying at home from 3:00 p.m. to 11:00 p.m., and a relatively high probability of working from 7:00 a.m.to 12:00 p.m., together with low probabilities 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 direction of the third eigenactivity accounts for the largest variance whose direction is orthogonal to the 1st and 2nd eigenactivities.

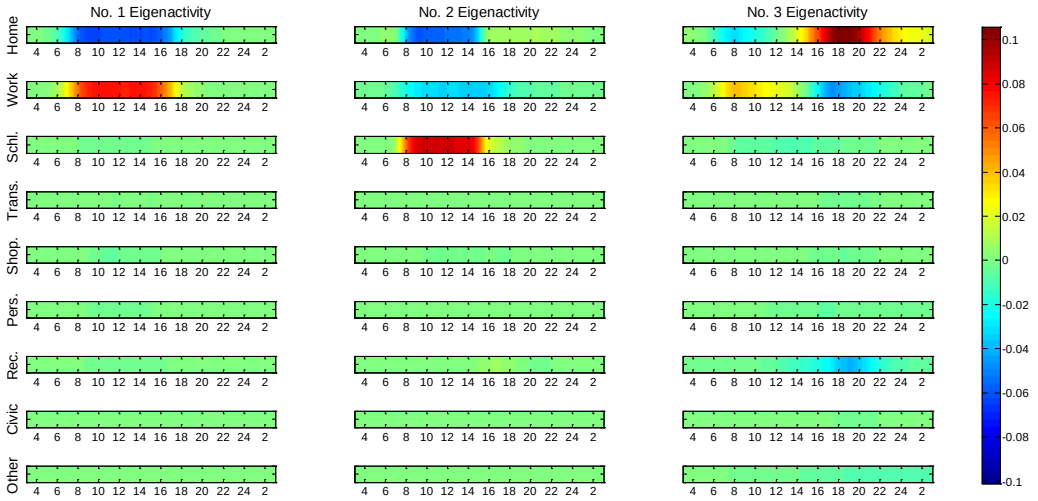

Figure 6 The first three eigenactivities of a weekday in Chicago

Weekend

Figure 7 illustrates the first three eigenactivities of individuals in Chicago on an average weekend. The first weekend eigenactivity (the 1st column of Figure 7) includes a high probability of recreating or visiting friends between 10:00 a.m. and 11:00 p.m., combined with a very low probability of staying at home from 8:00 a.m.to 9:00 p.m., and a somewhat high probability of working between 8:00 a.m. and 5:00 p.m., compared to the sample mean. The first eigenactivity of the weekend indicates that the largest discriminator of individuals' activities on a weekend is if they leave home from late morning to late evening, either for work or for recreational activities. The second weekend eigenactivity (the 2nd column of Figure 7) has a high probability of working from 7:00 a.m.to 4:00 p.m. and staying at home from 4:00 p.m. to the next early morning, combined with low probabilities of staying at home from 7:00 a.m. to 2:00 p.m. or recreating from 2:00 p.m. to 11:00 p.m. There is also some increased probability of engaging in civic or religious activities from 9:00 a.m. to 12:00 p.m. The second weekend eigenactivity (orthogonal to the first weekend eigenactivity), reveals that the 2nd largest clustering of individuals' activities during the weekend comes from either (a) working during the day time or conducting civic or religious activity in the morning, and then staying at home from late afternoon to the next early morning, or (b) staying at home during the morning and early afternoon and going out for recreation and entertainment in the early afternoon till late evening. The third weekend eigenactivity (the 3rd column of Figure 7) portrays high probability of staying at home from 3:00 p.m. to 11:00 p.m., and relatively high probability of working from 7:00 a.m. to 12:00 p.m.; combined with low probability of staying at home from 7:00 a.m. to 11:00 a.m., or working from 3:00 p.m. to 11:00 p.m., or recreation from 4:00 p.m. to 9:00 p.m. (all compared to the sample mean). The direction of the third eigenactivity accounts for the largest variance whose direction is orthogonal to the 1st and 2nd eigenactivities.

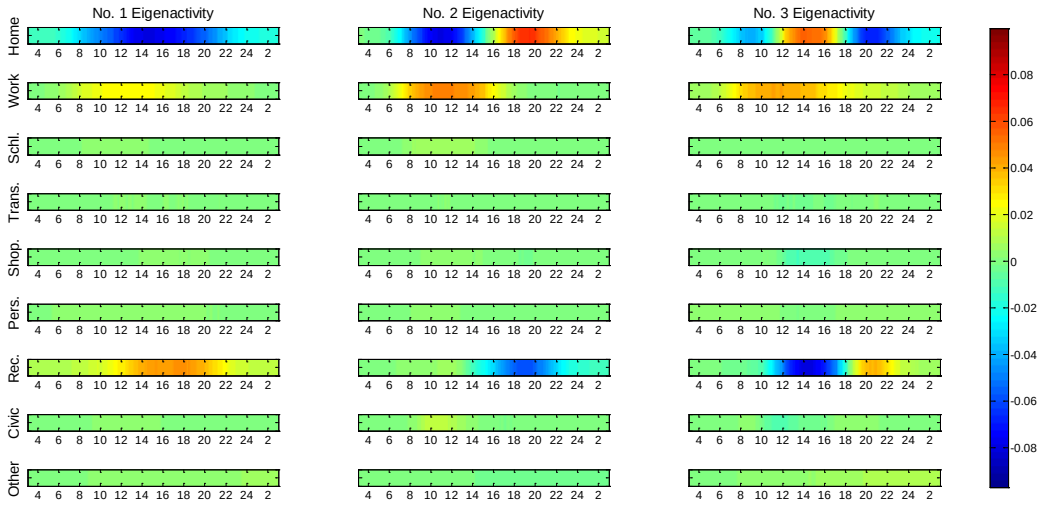

Figure 7 The first three eigenactivities of a weekend in Chicago

Selecting Eigenactivities for Daily Activity Reconstruction

We employ the reconstruction error measure defined in Section 4.2 to select the appropriate number of eigenactivities that are good enough to represent accurately the elements in . In Figure 8 and Figure 9, the left panels show the relationship between eigenvalues and the rank of eigenactivities, and the right panels display 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 for an average weekday, and 18 eigenactivities for an average weekend, will allow us to reconstruct a weekday and weekend daily 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 daily 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 daily activity sequence when using 21 eigenactivities for a weekday, or 18 eigenactivities for a weekend. 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.

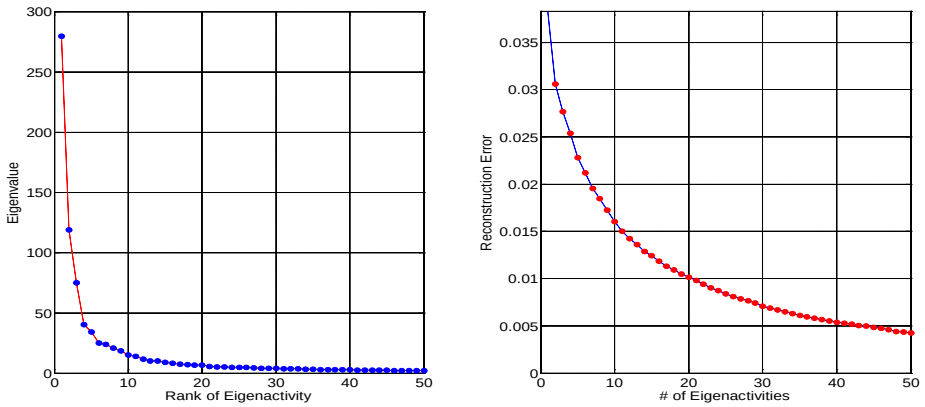

Figure 8 The eigenvalue and the reconstruction error w.r.t. the rank of eigenactivity of a weekday

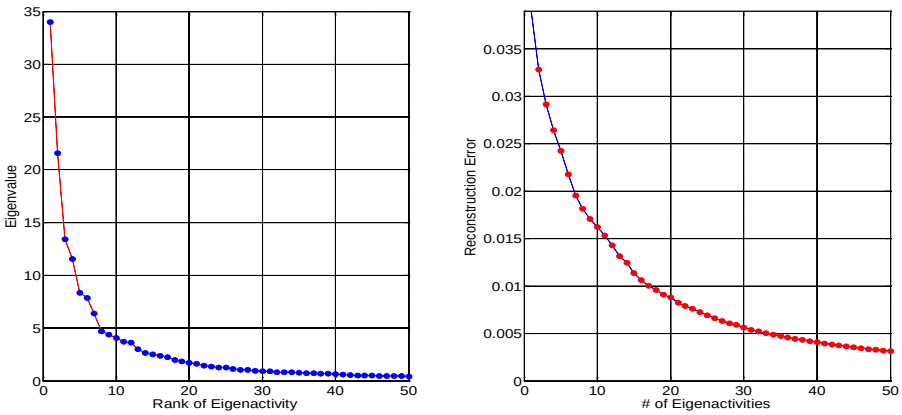

Figure 9 The eigenvalue and the reconstruction error w.r.t. the rank of eigenactivity of a weekend

Figure 10 exhibits our reconstructed individuals' daily activity sequence during the weekday and weekend, using the 21 eigenactivities for the weekday, and the 18 eigenactivities for the weekend, respectively. Comparing this figure with Figure 5, we can see that, in general, our reconstructed daily activities match the original sample data very well, except that at the 1% error level it does not allow us to reconstruct the activities in the "*Transportation Transitions*" category very accurately. Recall that this category involves not very common activities such as, "changing type of transportation/transfer; dropping off passenger from car; picking up passenger; service private vehicle; and loop trips" as described in Table 1.

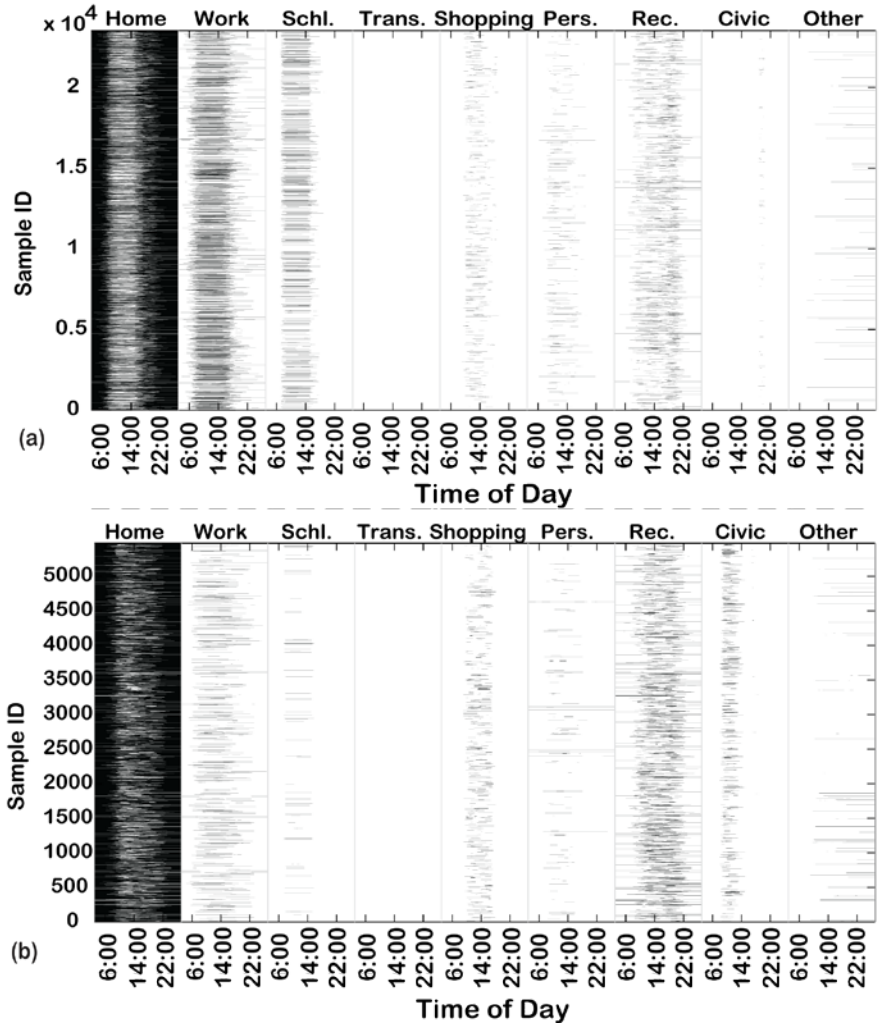

Figure 10 Reconstructed individual activities for samples on a (a) weekday and (b) weekend in Chicago

5.2 Clustering Individuals' Daily Activities and Social Demographics

In this section, we employ the *K*-means clustering via PCA method discussed in Section 4.3 to identify groups of individuals in the metropolitan area based on their daily activity sequences during the weekday and the weekend. We use two major cluster validity indices to determine the optimal number of clusters for the weekday and weekend case. After clustering individuals in the Chicago metropolitan area based on their daily activity sequence, we also summarize the

social demographic statistics of the different groups, and find interesting and suggestive signatures among clusters.

The Average Weekday

We use the Dunn's index (Dunn 1973) and the average Silhouette index (Rousseeuw 1987)—for both of which the higher the value the better the clustering—to indentify the appropriate number of clusters for the *K*-means clustering (Brun et al. 2007). Figure 11 shows the value of the indices with respect to the number of clusters.

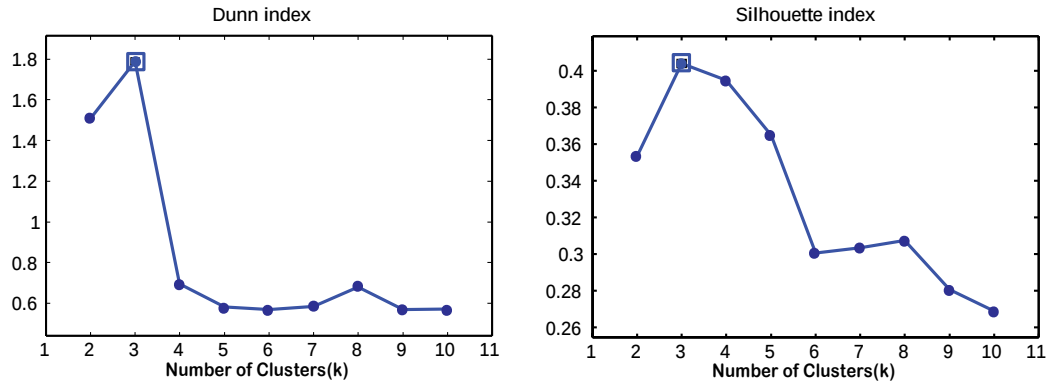

Figure 11 Cluster validity indices for the weekday case

Both the Dunn's index and the Silhouette index suggest that when the cluster number is 3, it gives the best clustering results for the average weekday case. This corresponds to three commonly identified groups of the population: (I) students (13%), (II) workers (33%), and (III) people who spend most of their time at home (54%). However, we want to further explore the temporal activity patterns of individuals that are beyond the three commonly known groups in the metropolitan area. From the Dunn's index and the Silhouette index (in Figure 11), we can see that the cluster number of eight is the second best alternative, which satisfies both the study purpose and provides relatively stable clusters.

Figure 12 exhibits the K-means clustering via PCA results (with cluster number=8), revealing individuals' activity patterns on an average weekday and their social demographic characteristics. Each row of the figure 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, presented in the last column. 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 12 displays individuals' daily activity sequences for each cluster. The second column shows the aggregated volume of different types of activities in the metropolitan area during a specific time interval over 24 hours, and the third column is a zoomed-in view of the previous column. The fourth column presents the social demographic statistics of the cluster in that row. We use star diagrams to represent proportions of people with various social

demographic characteristics in the sample and in each cluster. Figure 13 demonstrates the star diagrams and Table 2 lists the social demographic statistics in detail.

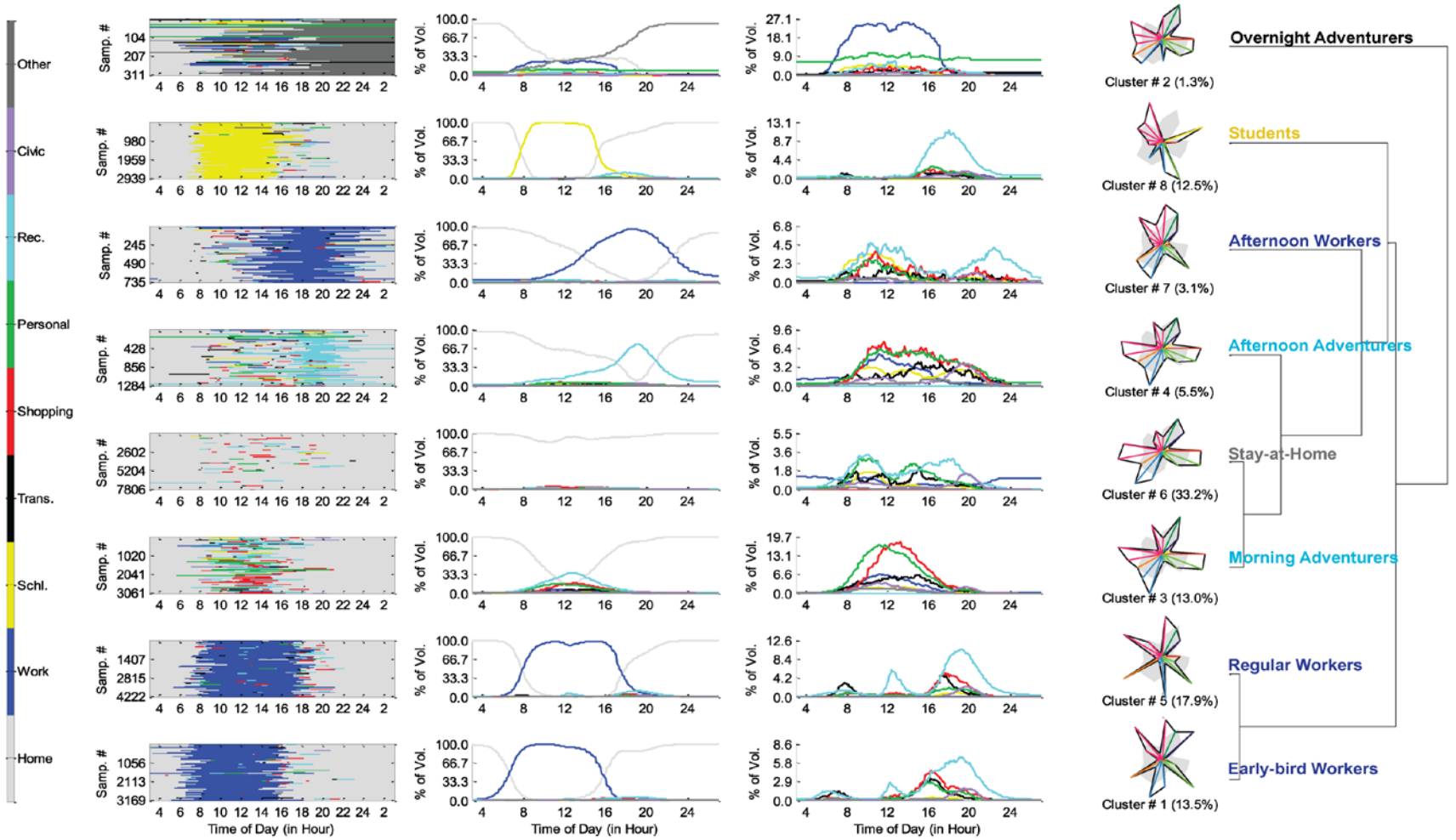

Figure 12 Clustering of individuals' weekday activity patterns and their social demographic characteristics in Chicago (cluster number=8).

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

Social Demographic Variables		Sample Mean	Mean of Cluster							
			#2	#8	#7	#4	#6	#3	#5	#1
1	Female	53.30%	55.0% (0.61)	49.0%** (-4.69)	46.7%** (-3.6)	57.3%** (2.9)	57.7%** (7.85)	58.9%** (6.22)	50.1%** (-4.11)	44.9%** (-9.47)
2	Student	22.10%	16.4%* (-2.43)	91.1%** (90.11)	13.2%** (-5.83)	22.5% (0.33)	13.2%** (-18.98)	15.7%** (-8.5)	8.3%** (-21.65)	7.2%** (-20.29)
3	Homemaker	13.40%	12.9% (-0.17)	3.5%** (-6.42)	11.1% (-0.29)	12.8% (-0.44)	15.1%** (3.06)	12.0% (-1.6)	13.4% (-0.02)	21.3%* (2.23)
4	Retired	59.70%	66.3% (1.36)	8.9%** (-22.82)	61.1% (0.12)	58.3% (-0.68)	62.5%** (3.55)	70.7%** (8.32)	54.9% (-1.16)	53.2% (-1.29)
5	Work	53.40%	51.1% (-0.8)	12.1%** (-44.85)	95.4%** (22.82)	42.8%** (-7.58)	33.7%** (-34.78)	38.4%** (-16.63)	94.1%** (53.01)	95.0%** (47.02)
6	Part Time	20.40%	15.9% (-1.34)	52.3%** (14.27)	28.6%** (5.32)	36.2%** (8.41)	29.7%** (10.88)	39.5%** (14.17)	10.9%** (-14.8)	12.7%** (-10.32)
7	No Flexibility	34.40%	24.2%** (-2.69)	36.4% (0.8)	43.0%** (4.74)	26.2%** (-4.02)	27.8%** (-7.14)	22.8%** (-8.34)	32.8%* (-2.16)	46.8%** (14.15)
8	Some Flexibility	42.30%	47.1% (1.21)	44.4% (0.76)	39.9% (-1.28)	36.0%** (-3.02)	35.8%** (-6.71)	37.2%** (-3.55)	50.3%** (10.04)	40.8%+ (-1.67)
9	Much Flexibility	23.20%	28.7% (1.61)	19.2%+ (-1.79)	17.1%** (-3.83)	37.8%** (8.05)	36.4%** (15.88)	40.0%** (13.54)	17.0%** (-9.32)	12.4%** (-13.97)
10	Work at Home	8.10%	11.3% (1.5)	5.9% (-1.51)	3.1%** (-4.8)	17.5%** (8.07)	19.8%** (22.05)	15.4%** (9.22)	2.5%** (-12.79)	1.8%** (-12.65)
11	Edu.>Tech School	45.90%	55.3%** (3.29)	6.5%** (-42.75)	48.3% (1.31)	45.8% (-0.02)	40.3%** (-9.75)	46.5% (0.72)	72.5%** (34.46)	58.9%** (14.63)
12	Low HH Income	16.90%	19.6% (1.23)	14.4%** (-3.41)	19.4%+ (1.78)	17.1% (0.25)	24.3%** (16.57)	19.3%** (3.42)	7.8%** (-14.99)	10.0%** (-9.88)
13	Middle HH Income	32.50%	28.0% (-1.58)	28.1%** (-4.84)	34.4% (1.08)	35.1%+ (1.93)	33.4%+ (1.75)	35.6%** (3.51)	28.8%** (-4.87)	35.0%** (2.94)
14	High HH Income	50.70%	52.4% (0.56)	57.4%** (7.08)	46.2%* (-2.35)	47.8%* (-1.99)	42.3%** (-14.04)	45.1%** (-5.84)	63.4%** (15.79)	55.0%** (4.64)
15	The Young (age<35)	34.60%	30.4% (-1.54)	92.4%** (65.55)	30.8%* (-2.15)	32.5% (-1.51)	27.8%** (-12.47)	26.1%** (-9.7)	24.9%** (-13.05)	20.3%** (-16.71)
16	The Middle-aged	39.70%	38.0% (-0.61)	5.0%** (-38.15)	55.7%** (8.83)	32.5%** (-5.17)	32.7%** (-12.55)	31.4%** (-9.21)	62.3%** (29.8)	66.4%** (30.43)
17	The Older (age>60)	25.80%	31.7%* (2.36)	2.5%** (-28.6)	13.5%** (-7.54)	34.9%** (7.43)	39.6%** (27.59)	42.4%** (20.85)	12.8%** (-19.14)	13.3%** (-15.87)

Note: Numbers in parentheses are the corresponding *t*-statistics which measure the departure of the proportions in each cluster (i.e., mean of cluster) from the proportions in the sample (i.e., sample mean). Cluster mean values are marked correspondingly if statistically significant in two-sided *t*-tests at the + 10%, * 5%, and ** 1% level.

Preliminary 2011 version of paper ultimately published in Data Mining and Knowledge Discovery: Volume 25, Issue 3, pages 478-510, Article DOI: 10.1007/s10618-012-0264-z, (2012)

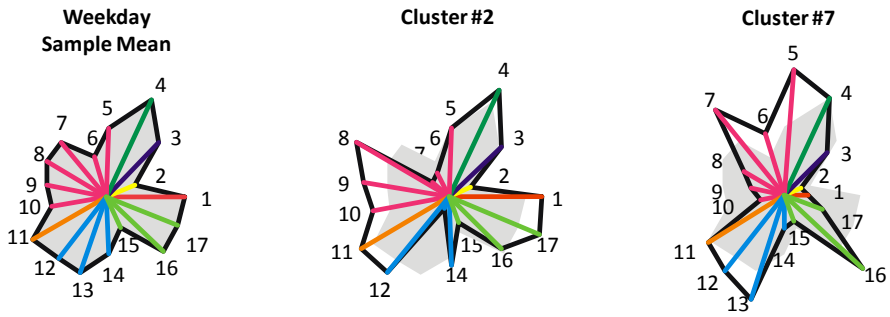

Figure 13 A Demonstration of the Star Diagram of the cluster's social demographics on a weekday
Note: In this Star Diagram, we use different colors to represent different vectors of each cluster (corresponding to the numbered social demographic variables in Table 2), and also set the sample mean as the gray background for each cluster for comparison convenience.

By mapping the relationship between the two clustering alternatives (clusters of three and of eight), we are able to explore the subdivisions within the previously identified three clusters (see Table 3).

Table 3 Cluster mapping: clusters of individuals based on their weekday activities

Clusters of Eight			Share in Cluster (size=three)			Clusters of Three		
Clusters		% in Total	I	II	III	Clusters		% in Total
#8	Students	12.50%	**99.80%**	0.00%	0.20%	I	Students	12.90%
#5	Regular workers	17.90%	0.00%	**99.90%**	0.10%	II	Traditional Workers	33.20%
#1	Early-bird workers	13.50%	0.00%	**99.20%**	0.80%			
#7	Afternoon workers	3.10%	2.70%	32.40%	**64.90%**	III	Adventurers + Stay-at-home	53.90%
#2	Overnight adventurers	1.30%	4.80%	26.40%	**68.80%**			
#4	Afternoon adventurers	5.50%	1.90%	3.20%	**94.90%**			
#6	Stay-at-home	33.20%	0.00%	0.00%	**100.00%**			
#3	Morning adventurers	13.00%	1.30%	2.90%	**95.80%**			

In the following paragraphs, we discuss specifically the clustered individual activity patterns of a weekday in each of the eight clusters and their social demographic characteristics (shown in Figure 12).

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. This group shares 12.5% of the total sample. The average annual household income for the cluster members is higher than the average weekday sample mean, and over 92% of cluster members are under 35 years old.

Regular Workers Cluster #5 is the group of workers who have a relatively regular schedule. They leave home for work at around seven to eight in the morning, and finish work at around five to six 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 and late afternoon.

Early-Bird Workers Individuals in Cluster #1 have similar daily 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 activities of the two clusters are similar, Cluster #1 seems to have a lower share of peak volume in other activities compared to those of Cluster #5. When we compare the social demographic characteristics of the two clusters, 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 (between the age of 35 to 60), 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 daily activity rhythm are quite different from those in Clusters #1 and #5. The majority of them (64.6%) spend their morning at home, and 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 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 lower household income. The middle-aged population share of this cluster 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 higher share of older population, a lower average educational level, and a lower household income level, compared to the weekday sample mean. It also has the greatest share of members who work at home (19.8%) compared to the other 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 and older population, higher shares of people who work at home or whose schedule is flexible, lower share of workers and lower household income level. In total, there are 13% of total weekday samples in Cluster #3, the "morning adventurers ", and 5.5% in Cluster #4, the "afternoon adventurers".

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 unexpectedly higher than the weekday sample mean.

The Average Weekend

We conduct similar analysis on the average weekend samples, and discuss the findings in this section. Based on the Dunn's index and the average Silhouette index (see Figure 14), we find that the optimal number of cluster size is three, which divides the total weekend samples into groups of (I) "adventurers" (26.3%), (II)"weekend workers"(7.7%), and (III) the "stay-at-home" (66.0%).

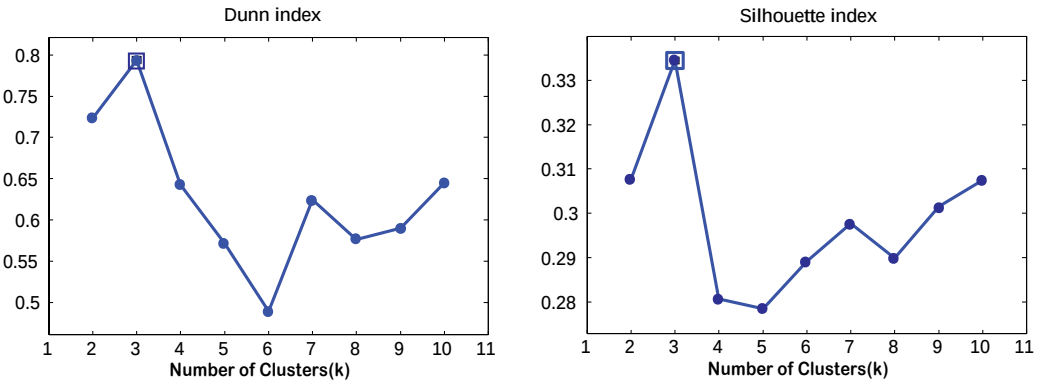

Figure 14 Cluster validity indices for the weekend case

In order to further explore detailed submarkets of the individuals’ weekend activity patterns, we choose the cluster number (=7) which gives the second best clustering result. The seven clusters are described in Table 4 and associated with the broader categories from the earlier three-clusters. We can see that there are four submarkets of the "adventurers" group (including overnight adventurers with known and unknown activities, afternoon adventurers, and evening adventurers),

two submarkets for the "stay-at-home" group (including afternoon stay-at-home, and all-day stay-at-home) and one single market for the "weekend workers".

Table 4 Clusters mapping: clusters of individuals based on their weekend activities

Clusters of Seven			Share in Cluster (size=three)			Clusters of Three		
Clusters		% in Total	I	II	III	Clusters		% in Total
#1	Overnight adventurers (unspecified)	1.6%	**91.0%**	0.0%	9.0%	I	Adventurers	26.3%
#3	Overnight adventurers	2.5%	**98.6%**	1.4%	0.0%			
#7	Afternoon adventurers	12.4%	**87.9%**	0.7%	11.4%			
#5	Evening adventurers	11.7%	**81.9%**	0.2%	17.9%			
#2	Workers	7.4%	0.0%	**100.0%**	0.0%	II	Weekend workers	7.7.%
#4	Afternoon stay-at-home	44.1%	0.1%	0.0%	**99.9%**	III	Stay-at-home	66.0%
#6	All day stay-at-home	20.3%	9.1%	1.2%	**89.8%**			

Similar to our handling of the weekday case, we use the "star diagram" to describe the group profiles and summarize their social demographic characteristics in Figure 15. In order to differentiate the activities on different days (Saturday or Sunday) we add a new variable (the share of Saturday samples) as the first vector, and retain the other factors described in the weekday star diagram. Since we have a slightly large number of samples on Saturday than on Sunday, the total weekend population mean has a higher share of Saturday activities. We summarize signature profiles of social demographic characteristics of the seven weekend activity clusters in Table 5 based on the same statistical tests employed in Table 2.

Even though the majority of people are either stay-at-home (66%) or enjoying their social life (26.3%), there are still a group of them working during the weekend, consisting of 7.4% of the total weekend sample. These people have a schedule that is similar to the regular workers on the weekdays— they leave home for work around seven to eight in the morning, and finish work around five to six in the afternoon. A few of them go out for recreation/entertainment after work in the early evening, with a peak around 12% at 8:00 p.m. There is a greater share of people working on Saturday than on Sunday, and there are more males than females who work during the weekends, compared to the weekend sample mean.

Table 5 Social demographic characteristics of the weekend activity clusters

Clusters		% in Total	Social Demographic Characteristics, Compared to Weekend Total Sample Mean
#1	Overnight adventurers (unspecified)	1.6%	More on Sunday; greater share of female; higher share with much flexibility; higher share of low household income; lower average education level; greater than average elderly people.
#3	Overnight adventurers	2.5%	More on Saturday, greater share of female; greater share of students; greater share with not flexible work schedule; lower average educational level; higher share of either low or high

			household income; higher share of young population.
#7	Afternoon adventurers	12.4%	Mix of Sunday and Saturday; mix of female and male; higher share of students; mix in work flexibility; relatively high education level; mixed income level; high share of young population.
#5	Evening adventurers	11.7%	More on Saturday; high share with some work flexibility; higher share with high income households; fewer elderly population.
#2	Workers	7.4%	A bit more on Saturday; lower share of female; lower share of students; higher share of people who work; higher share with no work flexibility; relatively high education level; high share of middle and high household income; higher share of middle-aged population.
#6	Afternoon stay-at-home	44.1%	Much more on Sunday; higher share of female; higher share of students; higher share with some or much work flexibility; higher share of people who work at home; higher than average educational level; higher share of low or high income household income; higher share of young population.
#4	All-day stay-at-home	20.3%	More on Saturday; higher share of female; lower share of students; lower share of people who work; higher share with no work flexibility; lower than average educational level; higher share of low household income; more elderly population.

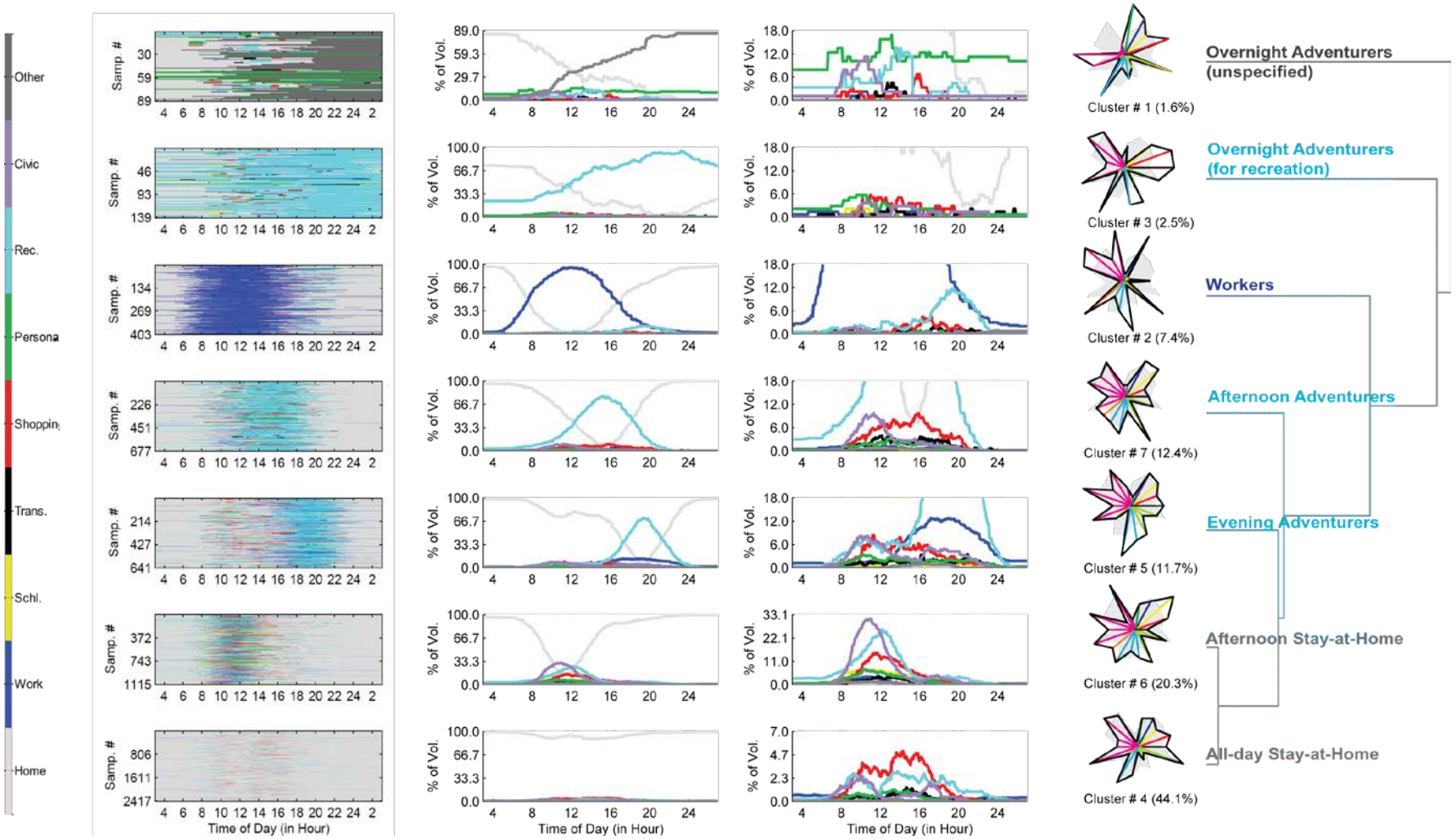

Figure 15 Clustering of individuals' weekend activity patterns and their social demographic characteristics in Chicago (cluster number=7).

6 Conclusions and Discussion

In this section, we summarize our methods and major findings and discuss potential implications of this study and future work.

6.1 Summary

In this paper, we analyze the activity patterns for 23,527 individuals on an average weekday and 5,481 individuals on an average weekend in the Chicago Metropolitan Area, by dividing the entire day into 288 five-minute intervals. We define the eigenvectors of the covariance matrix of the activity data as Eigenactivities, which are a set of vectors that span an 'activity space' and characterize the differences between individuals' daily 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 and 18 primary eigenactivities for the weekday and weekend case respectively to represent individuals' daily activities in the metropolitan area. We perform a *K*-means clustering algorithm on the obtained eigen decomposition projections to partition the samples into *k* clusters. By reducing the dimension of the problem with a small number of eigenactivites, we lower the computational cost of the method.

We successfully cluster individuals in the metropolitan area into groups within which they have relatively homogeneous daily activity patterns, and across which they have heterogeneous diversity. We cluster all the weekday samples into eight detailed categories including students (12.50%), regular workers (17.90%), early-bird workers (13.50%), afternoon workers (3.10%), the stay-at-home (33.20%), the morning adventurers (13.00%), the afternoon adventurers (5.50%) and the overnight adventurers (1.30%). For the weekend case, we cluster people into seven categories including the weekend workers (7.4%), the afternoon stay-at-home (44.1%) ,the all-day stay-at-home (20.3%),the afternoon adventurers (12.4%), the evening adventurers (11.7%), the overnight adventurers (for recreation—2.5%, and with unspecified activities—1.6%).

We identify the signatures of the social demographic profile of each cluster. In general, we find that, for the weekday case, when compared with the sample mean (1) the weekday workers have a lower share of females; but higher share of the middle-aged population, people who have less flexibility in work schedule and who have higher education, and/or higher household income level. (2) The weekday "adventurers" have a higher share of female, older population, people who work at home, and/or whose schedule is more flexible; and a lower share of students, working people, and people who have lower educational level, and/or lower household income level. For the weekend case, we find that (1) the profile of the weekend workers is very similar to those of the weekday workers, and (2) while comparing the "stay-at-home" with the "adventurers", the former

have a higher share of female, older population, low income households and people who work at home; and lower share of students and people who work.

6.2 Research Implications for Future Work

By applying a public available travel survey data, and employing the *K*-means clustering via PCA methods, this study reveals the regular patterns of human daily activities which reinforce previous findings on high predictability of human mobility using large-scale data sets (Song et al. 2010; Wang et al. 2011a; Gonzalez et al. 2008). The activity groups found here constitute a basic piece of information on urban activities that can be used to enrich models extracted from other sources of large-scare urban sensing data, i.e., mobile phones or Wi-Fi access points. To these means, sensing data must be combined with spatially detailed GIS data (such as land use data and points-of-interests data). 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 could impute their location information. This information combined with data mining and statistical learning methods could represent advances in urban transportation questions, which are essential but lacked enough information in the past.

Taken together, the significance of clustering people based on their daily activity patterns sheds lights on potential future applications in urban and transportation planning, emergency response and spreading dynamics. For example, without heavy-burdened computational costs, urban and transportation researchers may understand activity-based signature of daily travel patterns for different types of individuals, and/or construct individuals' mobility networks. Knowing more about the links between land use and activity patterns could facilitate congestion management, and improve models that try to predict human mobility, estimate origin-destination (OD) matrices, and/or simulate travel patterns under different circumstances. As demonstrated by Wang et al. (2009), Balcan et al. (2009), and Wang et al. (2011b), improvement in human mobility prediction models also has important implications in modeling and estimating the spreading of mobile phone viruses, infectious diseases, and information dissemination.

Acknowledgments

This research was funded in part by the MIT Urban Studies and Planning Department, by the US Department of Transportation Region One University Transportation Center, and by the Singapore National Research Foundation (NRF) through the Singapore-MIT Alliance for Research and Technology (SMART) Center for Future Mobility (FM). The authors acknowledge the comments and feedback from our colleagues and participants in the SC/OM seminar at MIT, at the 2011 CUPUM conference, Canada, and by Dr. Donald G. Janelle at the University of California, Santa Barbara. We are especially grateful for the comments by the anonymous reviewers and to Dr. Dietmar Bauer from Austrian Institute of Technology.

References

Axhausen KW, Zimmermann A, Schönfelder S, Rindsfüser G, Haupt T (2002) Observing the rhythms of daily life: A six-week travel diary. Transportation 29 (2):95-124. doi:10.1023/a:1014247822322

Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences 106 (51):21484-21489. doi:10.1073/pnas.0906910106

Balmer M, Axhausen KW, Nagel K (1985) Agent-based demand-modeling framework for large-scale microsimulations. vol 1985. National Research Council, Washington, DC, ETATS-UNIS

Batty M (2005) Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals. the MIT press, Cambridge (Mass.); London

Becker GS (1965) A theory of the allocation of time. The Economic Journal 75 (299):493-517

Becker GS (1977) The economic approach to human behavior. University of Chicago Press, Chicago; London

Becker GS (1991) A treatise on the family. Harvard University Press,

Bekhor S, Dobler C, Axhausen KW Integration of Activity-Based with Agent-Based Models: An Example from the Tel Aviv Model and MATSim. In: Transportation Research Board 90th Annual Meeting, Washington DC, 2011.

Ben-Akiva M, Bowman JL (1998) Integration of an Activity-based Model System and a Residential Location Model. Urban Studies 35 (7):1131-1153. doi:10.1080/0042098984529

Bhat CR, Koppelman FS (1999) A retrospective and prospective survey of time-use research. Transportation 26 (2):119-139. doi:10.1023/a:1005196331393

Bishop CM (2009) Pattern recognition and machine learning. Springer,

Bowman JL, Ben-Akiva M (2001) Activity-Based Disaggregate Travel Demand Model System with Activity Schedules. Transportation Research Part A: Policy and Practice 35 (1):1-28

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

Calabrese F, Reades J, Ratti C (2010) Eigenplaces: Segmenting Space through Digital Signatures. vol 9.

Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási A-L (2008) Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical 41 (22):224015

Chapin FS (1974) Human Activity Patterns in the City: Things People Do in Time and in Space. Wiley, New York

Chicago Travel Tracker Household Travel Inventory (2008) http://www.cmap.illinois.gov/travel-tracker-survey.

Crane R, 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. doi:10.1073/pnas.0803685105

Ding C, He X (2004) K-means clustering via principal component analysis. Paper presented at the Proceedings of the twenty-first international conference on Machine learning, Banff, Alberta, Canada,

Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley, New York

Dunn JC (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics 3 (3):32 - 57

Durrett R (2005) Probability: theory and examples. Thomson Brooks/Cole,

Eagle N, Pentland A (2009) Eigenbehaviors: identifying structure in routine. Behavioral Ecology and Sociobiology 63 (7):1057-1066. doi:10.1007/s00265-009-0739-0

Eagle N, Pentland A, Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.0900282106
Foth M, Forlano L, Satchell C, Gibbs M (eds) (2011) From social butterfly to engaged citizen: urban informatics, social media, ubiquitous computing, and mobile technology to support citizen engagement. MIT Press, Cambridge, Mass
Freud S (1953) Collected Papers, vol IV. vol v. 1-5. London: Hogarth Press and The Institute of Psychoanalysis,
Geerken M, Gove WR (1983) At home and at work: the family's allocation of labor. Sage Publications ; Published in cooperation with the National Council on Family Relations, Beverly Hills, CA
Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453 (7196):779-782. doi:http://www.nature.com/nature/journal/v453/n7196/suppinfo/nature06958_S1.html
Goodchild MF, Janelle DG (1984) The city around the clock: space - time patterns of urban ecological structure. Environment and Planning A 16 (6):807-820
Greaves S (2004) GIS and the collection of travel survey data. In: Hensher DA (ed) Handbook of transport geography and spatial systems. Elsevier,
Gupta S, Rao K, Bhatnagar V (1999) K-means Clustering Algorithm for Categorical Attributes. DataWarehousing and Knowledge Discovery 1676:797-797. doi:10.1007/3-540-48298-9_22
Hägerstrand T (1989) Reflections on “what about people in regional science?”. Papers in Regional Science 66 (1):1-6
Halkidi M, Batistakis Y, Vazirgiannis M (2001) On Clustering Validation Techniques. Journal of Intelligent Information Systems 17 (2):107-145. doi:10.1023/a:1012801612483
Hanson S, Hanson P (1980) Gender and Urban Activity Patterns in Uppsala, Sweden. Geographical Review 70 (3):291-299
Hanson S, Kwan M-P (eds) (2008) Transport: Critical Essays in Humman Geography. 1 edn,
Harvey A, Taylor M (2000) Activity settings and travel behaviour: A social contact perspective. Transportation 27 (1):53-73. doi:10.1023/a:1005207320044
Hastie T, Tibshirani R, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer,
Huang Z (1998) Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2 (3):283-304. doi:10.1023/a:1009769707641
Jolliffe IT (2002) Principal component analysis. Springer-Verlag,
Kargupta H, Han J (eds) (2009) Next generation of data mining. CRC Press,
Kim M, Kotz D, Kim S Extracting a mobility model from real user traces. In: IEEE INFOCOM’06, Barcelona, Spain, 2006. doi:citeulike-article-id:903652
Kwan M-P (1999) Gender and Individual Access to Urban Opportunities: A Study Using Space--Time Measures. The Professional Geographer 51 (2):210-227
Li L, Prakash BA Time Series Clustering: Complex is Simpler! In: Proceedings of the 28th International Conference on Machine learning, 2011.
Maslow AH, Frager R (1987) Motivation and personality. Harper and Row,
Nature Editorial (2008) A flood of hard data. Nature 453 (7196):698-698
Ordonez C (2003) Clustering binary data streams with K-means. Paper presented at the Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, San Diego, California,
Portugali J, Meyer H, Stolk E, Tan E (eds) (2012) Complexity Theories of Cities Have Come of Age: An Overview With Implications to Urban Planning and Design. Springer Verlag,
Ralambondrainy H (1995) A conceptual version of the K-means algorithm. Pattern Recognition Letters 16 (11):1147-1157. doi:10.1016/0167-8655(95)00075-r

Reggiani A, Nijkamp P (eds) (2009) Complexity and Spatial Networks: In Search of Simplicity. Springer,
Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53-65
Sang S, O'Kelly M, Kwan M-P (2011) Examining Commuting Patterns. Urban Studies 48 (5):891-909. doi:10.1177/0042098010368576
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
Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of Predictability in Human Mobility. Science 327 (5968):1018-1021. doi:10.1126/science.1177170
Taylor PJ, Parkes DN (1975) A Kantian view of the city: a factorial-ecology experiment in space and time. Environment and Planning A 7 (6):671-688
Turk M, Pentland A (1991) Eigenfaces for Recognition. Journal of Cognitive Neuroscience 3 (1):71-86. doi:doi:10.1162/jocn.1991.3.1.71
Waddell P (2002) UrbanSim: Modeling Urban Development for Land Use, Transportation and Environmental Planning. Journal of the American Planning Association 68 (3):297-314
Wang D, Pedreschi D, Song C, Giannotti F, Barabási A-L (2011a) Human mobility, social ties and link prediction. Paper presented at the 17th ACM SIGKDD Conference on Knoledge Discovery and Data Mining (KDD'11),
Wang D, Wen Z, Tong H, Lin C-Y, Song C, Barabási A-L (2011b) Information spreading in context. Paper presented at the Proceedings of the 20th international conference on World wide web, Hyderabad, India,
Wang P, González MC, Hidalgo CA, Barabási A-L (2009) Understanding the Spreading Patterns of Mobile Phone Viruses. Science 324 (5930):1071-1076. doi:10.1126/science.1167053
Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan G, Ng A, Liu B, Yu P, Zhou Z-H, Steinbach M, Hand D, Steinberg D (2008) Top 10 algorithms in data mining. Knowledge and Information Systems 14 (1):1-37. doi:10.1007/s10115-007-0114-2
Xu R, Wunsch DC (2008) Partitional Clustering. In: Clustering. John Wiley & Sons, Inc., pp 63-110. doi:10.1002/9780470382776.ch4
Yang J, Leskovec J (2011) Patterns of temporal variation in online media. Paper presented at the Proceedings of the fourth ACM international conference on Web search and data mining, Hong Kong, China,
Yu H, 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
Zha H, Ding C, Gu M, He X, Simon H (2001) Spectral relaxation for K-means clustering. Advances in Neural Information Processing Systems 14 (NIPS'01):1057–1064