Master's Thesis of Engineering

# Daily Activity Pattern Recognition considering Spatial Distance based on National Household Travel Survey Data

# 가구통행실태조사 자료를 기반으로 공간적 거리를 고려한 일일 활동 패턴 분석

August 2021

Department of Civil & Environmental Engineering
Seoul National University
Civil and Environmental Engineering Major

Jeongwook Seo

# Daily Activity Pattern Recognition considering Spatial Distance based on National Household Travel Survey Data

Dong-Kyu Kim

Submitting a master's thesis of Public Administration

August 2021

Department of Civil & Environmental Engineering
Seoul National University
Civil & Environmental Engineering Major

Jeongwook Seo

Confirming the master's thesis written by
Jeongwook Seo
August 2021

Chair Seung-Young Kho

Vice Chair Dong-Kyu Kim

Examiner Chungwon Lee

# Abstract

As disaggregated travel demand forecasting models are developing in recent years, prediction accuracy of individual activity patterns depends on actual information drawn from activity generation modules. Thus, producing more accurate and homogeneous individual activity patterns from this module will result in increasing prediction accuracy in activity-based travel demand modeling. Even though travel distance plays an important rule to travel decision, most studies did not consider spatial information to recognize activity pattern. In this study, I recognized a new daily activity pattern that considers activity, spatial, and socio-demographic information. To generate representative activity patterns, I considered daily activity episode as multidimensional trajectory. Activity and spatial information is used to generate distance matrix of multidimensional trajectories by Multidimensional Similarity Measure (MSM) method. K-means algorithm is adopted based on the distance matrix to generate representative activity patterns. Multivariate analysis of variance (MANOVA) test is conducted to analyze homogeneity within group and heterogeneity with group of socio-demographic information. The proposed method is applied in Seoul based on K-NHTS. As a result, three cluster are generated, which are worker, student, non-worker cluster and each cluster is divided into four, two, and three groups based on the travel distance and activity sequences. The proposed method enriches the traditional methods such as using socio-demographic attributes for classifying the population and is more straightforward, and easy to implement in practical activity-based model.

**Keyword :**, Travel behaviors, Activity based model, Daily activity pattern, Spatial distance, Multidimensional trajectory similarity measure, K-means clustering analysis, Multivariate analysis of variance

**Student Number :** 2019-26508

# Table of Contents

# Chapter 1. Introduction

In recent years, the goals of urban transportation planning and policies have shifted from the meeting of long-term, supply-oriented mobility needs to the facilitation of short-term, demand-oriented accessibility needs. Disaggregate travel demand models have begun to be employed for meeting those shifted goals. These models improve upon the traditional four stage modeling method, since they are able to more accurately capture the effects of elements that influence travel behavior and time allocation, such as socio-demographic attributes. Latterly, the activity-based modeling approach, along with other disaggregate travel demand modeling methods such as trip-based modeling, has become more popular and commonly used in both the academic and practitioner sectors (Castiglione et al. 2014)

For many years, different approaches are employed for development of activity-based model(ABM). Based on the approaches, it can be classified into three concept categories: (i) constraint-based models, (ii), discrete choice models, and, (iii) computational process models (Rasouli et al. 2014). The constraint-based models consider possible travel patterns with respect to a set of space–time constraints. Lenntorp, Kwan, and Button are some examples of ABM developed through the constraint-based modeling approach (Lenntrop et al. 1977, Kwan et al. 1997, Button et al. 1985). The second category, discrete choice models consider activity pattern consequences from utility maximizing decisions. Bowman and Ben-Akiva, Vovsha et al., and CEMDEP are some examples of activity-based models developed through the econometric approach (Bowman et al. 2001, Vovsha et al. 2002, Bhat et al. 2004). Finally, the computational process models simulate and model activity patterns through computational processes. Garling et al., Arentze and Timmermans, and Miller and Roorda are some examples of ABM developed through the rule-based modeling approach (Garling et al. 1989, Arentze et al. 2000,

Miller et al. 2003). Researchers recently have used machine learning techniques to develop different procedure of ABM. However, there have been very limited applications of such techniques in activity-based modeling. For instance, the K-means clustering technique has been used in a pattern-recognition modeling framework (Jiang et al. 2012, Allahviranloo et al. 2016) and support vector machine has been used in a daily activity sequence recognition process (Allahviranloo et al. 2013).

Given that it is difficult to capture the full complexities of activity–travel patterns at once, two general approaches have been employed to capture the complexity of activity–travel patterns. One is decomposition of an individual's activity pattern into the numerous dimensions and the apply separate measures for each dimension. The other is treatment of the pattern as a multidimensional holistic entity.

At present, the first approach is dominant in activity pattern research, in part because most existing operational activity-based travel forecasting systems are implemented on a microsimulation framework that consists of a series of calibrated econometric models that address the multiple dimensions either individually or jointly. Discrete choice models, continuous choice models, and hazard-based duration models are widely used as the basis of the microsimulation implementation framework.

The other approach, which is called holistic approach, to the measurement of activity patterns was popular in the early stage of activity-based travel behavior analysis. Many transportation researchers found the complexity of activity patterns and recognized the necessity of the analyzing activity patterns to understand travel behaviors. This approach categorizes people's daily or weekly activity patterns into homogeneous groups and identifies of the group determinants or constraints. The holistic approach is divided into two representative categories. For the first category, each activity pattern was described by numerous measures which were used for principal component analysis to identify its notable features, The latter information was often used

to classify the whole set of activity patterns into a small number of similar groups. For the second category, time slice variable was used to compare individuals' activity patterns with each other. The comparison produces a matrix of pairwise dissimilarities between the patterns, which is subsequently used for cluster analysis.

The holistic approach has been paid attention recently since the introduction of sequence alignment methods into time use and transportation research. Most activity pattern recognition studies adopted sequence alignment methods are combined with cluster analysis.

The activity patterns component provides explicit details on activity type, and the frequency and sequence of activities engaged in. Daily activity patterns of individuals are crucial components in any activity-based travel demand model, as an individual's travel demand originates from their need to engage in particular activities (Hafezi et al. 2019). Activity generation modules can play an important role in every activity-based modeling framework. Prediction accuracy of individual travel behavior depends on actual information drawn from activity generation modules. Thus, producing more accurate and homogeneous information from this module will result in increasing prediction accuracy in ABM.

Even though many empirical approaches have been employed for the activity module, such as decision tree, random forest, and sequential alignment technique, most studies did not consider spatial information to recognize activity pattern, they only considered socio-demographic information. In this study, I recognized a new daily activity pattern that is based on the semantic trajectory to consider both spatial and socio-demographic information. Table 1.1 shows which attributes are used to study previous daily activity pattern in previous. A semantic trajectory A is a sequence of stops $< a_1, \dots, a_n >$ with each stop in the form of a tuple$\big((x, y), [t_1, t_2], attributes\big)$, where (x, y) is the centroid of all points of the subtrajectory identified as the stop, representing the space dimension, $t_1$ and $t_2$ are the start and end time of the stop, respectively, corresponding to the time dimension, and attributes

are the other spatial and socio-demographic information such as category of activity, the distance from previous activity destination, characterizing the semantics (Furtado et al. 2016). I consider an individual record from travel survey as one trajectory, measure the similarity between those trajectories, and cluster based on the measured similarity. The first section presents the specific design and methodology of our study. The second section provides an empirical results and discussion of the results. The third section presents our conclusion and recommendations for future research.

Table 1.1 Summary of daily activity pattern studies

| Study | Spatial attribute | Socio-demographic attribute | Activity attribute | | Activity pattern classification |
|---|---|---|---|---|---|
| | | | Tour-based | Trip-based | |
| Wilfred et al. (1985) | X | O | O | - | O |
| Timmermans et al. (2003) | O | O | - | O | X |
| Ronald et al. (2004) | O | X | - | O | X |
| Kees & Harry (2009) | O | O | - | O | O |
| Jiang et al. (2012) | X | O | O | - | O |
| Mao et al. (2013) | O | O | O | - | X |
| Santi et al. (2014) | O | X | - | O | O |
| Kim (2014) | X | O | O | - | O |
| Plevka et al. (2016) | O | O | O | - | X |
| Jianan & Tao (2016) | O | O | O | - | X |
| Qunying & David (2016) | O | X | - | O | O |
| Allahviranloo (2016) | X | O | O | - | O |
| Fei & Donggen (2017) | X | O | - | O | X |
| Hafezi et al. (2019) | X | O | O | - | O |
| Wang et al. (2019) | X | O | O | - | O |
| **This study** | **O** | **O** | **O** | **-** | **O** |

# Chapter 2. Methodology

## 2.1. Multidimensional Similarity Measure (MSM)

This study adopted the multidimensional similarity measure(MSM) for semantic trajectory by Furtado et al. 2016. MSM computes the similarity between two multidimensional sequences and finds the highest matching score of each element of a sequence A in relation to all the elements of a sequence B, and vice-versa, with respect to l-dimensions. Then, based on the composition of the highest matching scores of all elements in both sequences, MSM obtains the similarity score. I recognized a new daily activity pattern that is based on the semantic trajectory to consider both spatial and activity information. Initially it is necessary to define a multidimensional sequence.

Definition 1: A semantic trajectory T is a sequence of points $< p_1, \dots, p_q >$, where each point p has a set of attributes $D = \{d_1, .., d_k, .., d_l\}$ containing spatial and activity attribute according to l-dimensions. Let S = {A, B} be a set of two semantic trajectories $A =< a_1, \dots, a_n >$ and $B =< b_1, \dots, b_m >$ .

Definition 2: Given any two points a $\in A$ and b $\in B$, the distance between them on an attribute $d_k$ is calculated by the function $dist_k(a, b)$. The two points a $\in A$ and b $\in B$, and a set of attributes D, the matching score between the points is calculated by the function score: $A \times B \rightarrow [0,1]$

$$score(a, b) = \sum_{k=1}^{D}(match_k\,(a, b) * w_k) \quad (2.1)$$

where $w_k$ is the weight on attribute $d_k$, $match_k(a, b)$ is given by the function match: $A \times B \rightarrow [0,1]$

$$match_k(a, b) = \begin{cases} 1 - dist_k(a, b) & if\ dist_k(a,b) \le \sigma_k \\ 0 & otherwise \end{cases} \quad (2.2)$$

where $\sigma_k$ is the threshold value on attribute $d_k$.

Definition 1 states that each point of a trajectory can be characterized by spatial, and activity attributes and Definition 2

states that each attribute should calculated by its own distance function and threshold value to quantify the distance between points. I adopted euclidean distance to calculate numerical data type and discrete distance to calculate categorical data type. All numerical data are normalized to reflect equal influence.

$$dist_1(a,b) = |a-b| \tag{2.3}$$

$$dist_2(a,b) = \begin{cases} 0 & if\ a.type = b.type \\ 1 & otherwise \end{cases} \tag{2.4}$$

An element $a \in A$ can have some degree of similarity with multiple elements of sequence B according to different dimensions. Because of that I aim at finding only the best matching scores of each element a in relation to B. The sum of the highest score of all elements $a \in A$ with any element of B is called parity of A with B.

Definition 3: Given two multidimensional trajectories A and B, the parity of A with B is the sum of the highest score of all the points $a \in A$ when compared with all the points of B, as stated by the function parity: $S^2 \rightarrow [0, |A|]$

$$parity(A,B) = \sum_{a \in A} max\{score(a,b): b \in B\} \tag{2.5}$$

The multidimensional similarity measure MSM(A, B) is given by the average parity(A, B) and parity(B, A).

$$MSM(A,B) = \begin{cases} 0 & if\ |A|=0\ or\ |B|=0 \\ \frac{parity(A,B)+pari\quad(B,A)}{|A|+|B|} & otherwise \end{cases} \tag{2.6}$$

where, $|A|$ is the number of points in trajectory A.

After calculating MSM for all trajectories, I generated distance matrix M.

## 2.2. K−means Clustering algorithm

With calculated distance matrix M, I calculated clustering algorithm with K−means clustering. The K−means algorithm is a simple and most extensively used method for iterative clustering methods to partition datasets into the number of clusters. It requires no prior information of the datasets, and it can be applied to various types of data; it performs especially well for numerical variables. One problem that must be solved in the clustering process is to determine the optimal number of clusters in advance to accurately represent the partitions of the dataset (Ma et al., 2013; Jiang et al., 2012; Zhao et al., 2017). The results of cluster number and cluster centroids identified are used as inputs for the K−means algorithm, which determines the final cluster through a reallocation method. The mathematical expression of two−level hierarchical K−means algorithm in this study is as follows:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{X^{\alpha} s_i} \| X^{\alpha s_i} - \mu_i \|^2 \qquad (2.7)$$

where $S_i$ is $i$th cluster, $X^{\alpha}$ is an element of $S_i$, $\alpha$ is trajectories, and $\mu_i$ is a center of $S_i$.

To determine optimal number of optimal clusters, I adopted silhouette, which refers to a method of interpretation and validation of consistency within clusters of data. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

## 2.3. Multivariate Analysis of Variance (MANOVA)

MANOVA is an extension of analysis of variance (ANOVA) model in which dependent variables are evaluated on a combination of dependent variables. It is used for comparing the means of several groups with many variables. The MANOVA model for g groups is given by $Y_{lj} = \mu_l + e_{lj}$ where, Y is the dependent variable. The null hypothesis for the MANOVA model with g groups (or population) can be written as $H_0: \mu_1 = \mu_2 = \cdots = \mu_g$ . The assumptions required in using MANOVA are as follows:

1) The random samples from different populations are different.
2) All populations have a common covariance matrix Σ.
3) Each group is multivariate normal

The test statistic used for testing the null hypothesis of a MANOVA model is Wilk's lambda (Λ) given by $\Lambda = \frac{W}{B+T'}$ where, W is the within sum of square error and B is between (groups) sum of the square error. Λ is used as a statistic to test if there is any difference between the means of the given groups on a combination of dependent variables. In multivariate analysis, Λ plays the same role as F−test does in one−way ANOVA. It measures the proportion of variance across groups in terms of a combination of variables. If Λ is small, the null hypothesis is rejected concluding that the mean (dependent variable) of at least one of the variables in the groups is significantly different. Distribution of $\Lambda$ is not straightforward but can be approximated. For large sample sizes, a modification of Λ due to Bartlett is used for testing Ho (Johnson and Wichern, 2007). P−value for the tests can be obtained from the approximated distribution.

# Chapter 3. Case Study

## 3.1. Data description

The Korean National Household Travel Survey (K-NHTS) is conducted every five years to solve traffic problems, in accordance with the National Integrated Transport System Efficiency Act. The survey aims to track an individual's single-day travel diaries to identify the travel behaviors and ultimately establish the OD pairs. The data is consist of three categories that include 61 types of information, i.e., 1) socio-demographic (e,g., household size, income level, age, gender, employment status, the flexibility of work schedule), and 2) activity episode information. I focused on the individual-specific and activity episode information to analyze the activity pattern in Seoul, and the data were collected as part of the fourth K-NHTS in 2016. For this study, only 18 sets of information are used, as listed in Table 3.1.

The purpose of data processing is to extract the normal activity data of the residents whose first trip starts at home and last trip ends at home. Firstly, the abnormal data is cleared, including all the data of the person whose activity duration is less than or equal to 0, whose travel end time is less than the start time. Secondly, invalid records are deleted, such non-homebased travel. Lastly, resident travel data with only one travel record is removed. Individuals whose house is located in Seoul are selected. More than two number of activity sequences are selected. The coordination of house, departure, and arrival locations are aggregated with administrative dong, which is traffic analysis zone (TAZ), centroid coordination. As a result of preprocessing, the total number of people is 19,934, and the number of trips is 51,426.

Table 3.1 Definition of Selected Variables

| Name of Variable | Attribute type | Data Type |
|---|---|---|
| Personal ID | − | Numerical |
| Income | Socio−demographic | Categorical |
| Ownership of vehicle | Socio−demographic | Categorical |
| Ownership of driver's license | Socio−demographic | Categorical |
| Age | Socio−demographic | Numerical |
| Gender | Socio−demographic | Categorical |
| Residence type | Socio−demographic | Categorical |
| Occupation type | Socio−demographic | Categorical |
| Working days per week | Socio−demographic | Categorical |
| Working times per day | Socio−demographic | Categorical |
| Residence TAZ code | Spatial | Categorical |
| Activity location TAZ code | Spatial | Categorical |
| Distance from previous activity location [km] | Spatial | Numerical |
| Start time [min] | Activity | Numerical |
| Arrival time [min] | Activity | Numerical |
| Travel time [min] | Activity | Numerical |
| Duration time [min] | Activity | Numerical |
| Activity type | Activity | Categorical |

## 3.2. Calculating Similarity

For the similarity measure, all numerical variables are normalized to compare equally and the threshold for numerical variables is determined to 0.5. The weight for all variables is equally divided. Based on similarities between some of the 12 Trip Purposes in the original survey data, I aggregate them into 7 activity types as shown Table 3.2. All work-related trip purposes are aggregated into work activity type. 9, 10, and 11 trip purposes are aggregated into other activities because each trip purpose has less than 1% number of trips.

Table 3.2 Aggregated 7 activity types vs. the original 12 Trip Purposes

| Aggregated Activity Types | Original Trip Purpose | Number of Trips (%) |
|---|---|---|
| Home | 1. Home activities | 19,312 (46.73%) |
| Work | 2. Work/Job; 3. Work/Business related; 4. All other activities at work | 10,723 (25.95%) |
| School | 5. School | 3,126 (7.56%) |
| Education | 6. Education related activities | 1,166 (2.82%) |
| Shopping | 7. Shopping | 1,718 (4.16%) |
| Recreation | 8. Leisure/Exercise/ Recreation/Entertainment | 1,841 (4.45%) |
| Other activities | 9. Pick up and Drop off; 10. Eat meal outside of home; 11. Visit relatives; 12. Other (personal errands/Religious activities) | 3,442 (8.33%) |

To identify the difference activity types of shopping, recreation and other activities, Kolmogorov−Smirnov Test is conducted. The Kolmogorov−Smirnov (K−S) test is a typical technique for testing goodness of fit. The K−S is nonparametric or distribution free, not assuming for the shape of the population from which the samples are drawn. The K−S test measures the maximum vertical difference between the two cumulative probability distributions. The estimated maximum difference is compared with the tabulated value of the K−S statistic. The K−S statistic provides the threshold values for sample sizes and levels of significance. The K−S test is especially valuable in cases where the number of observations is small.

$$D_{n,m} = \sup_x |F_{i,n}(x) - F_{j,m}(x)| \quad (3.1)$$

$$D_{c,a} = c(a)\sqrt{\frac{n+m}{nm}} \quad (3.2)$$

where, $F_{\mathrm{i},n}(x)$ is the cumulative distribution function which has n sample size, $D_{c,a}$ is the critical value at a% p−value. When $D_{n,m}$ is greater than $D_{c,a}$, the null hypothesis is rejected that sample i and sample j are distributed same.

I conducted K−S test for start time, duration time, and distance from previous activity location with 0.01 p−value. As a result, only distance from previous activity location of shopping and recreation are not rejected the null hypothesis as shown Table 3.3. Therefore, I determined that these three activities are different.

Table 3.3 K−S test results for Shopping, Recreation and Other activities

| Activity/ Activity | $D_{n,m}$ | | | $D_{c,0.01}$ |
|---|---|---|---|---|
| | Start time | Duration time | Distance from previous activity location | |
| Shopping/ Recreation | 0.560 | 0.364 | **0.152** | 0.174 |
| Shopping/ Other activities | 0.323 | 0.450 | 0.240 | 0.200 |
| Recreation/ Other activities | 0.460 | 0.455 | 0.254 | 0.129 |

The distance from previous activity location is calculated based on the centroid coordination of departure and arrival TAZ. If the departure TAZ and arrival TAZ are same, the distance from previous activity location is 0km. The descriptive statistics of distance from previous activity location is as shown table 3.5.

Table 3.4 The descriptive statistics of distance from previous activity location

| | |
|---|---|
| Average [km] | 4.16 |
| Std.ev [km] | 3.93 |
| Minimum value [km] | 0 |
| Maximum value [km] | 100.92 |

As the average of distance is 4.16km, I categorized the distance into 0km, more than 0km and less than 4.2km, and more than 4.2km to avoid bias. I named each category as short distance, middle distance, and long distance respectively. The number of trips is shown for each distance category Table 3.5.

Table 3.5 The number of trips for each distance category

| Category | Number of Trips (%) |
|---|---|
| Short distance (0km) | 13,178 (25.7%) |
| Middle distance (0~4.2km) | 22,198 (43.3%) |
| Long distance (Longer than 4.2km) | 15,870 (31.0%) |

## 3.3. Clustering Results

I measured similarity of trajectories using four attributes which are activity type, start time, duration time, and distance from previous activity location. The time interval of start time and duration time is 1 minute. Based on the calculated distance matrix, the k-means algorithm is adopted to analyze heterogeneity and homogeneity of user's travel behavior. To determine the optimal k, silhouette score is calculated. The result of silhouette score for k is shown in Figure 3.1. The highest silhouette score is obtained when k is 6, which value is 0.501. Therefore, I determined the number of clusters as 6.

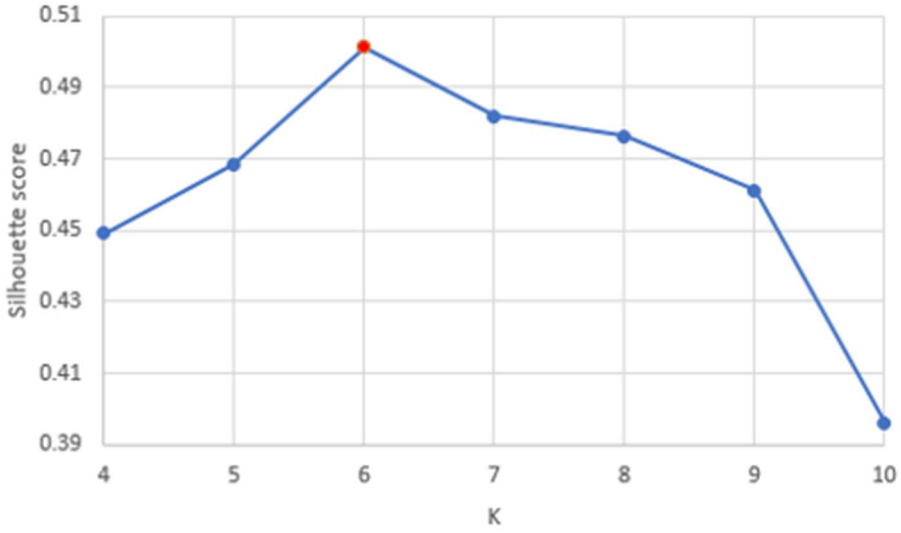


Figure 3.1 Silhouette score for each K

As a result of k-means clustering algorithm, I divided total sample data into six groups based on the activity pattern and travel distance. Table 3.6 shows the number of users for each group and

the number of users for each group according to the number of activity sequences.

The temporal pattern of individual activities for the six identified groups is shown in Figure 3.2, and the distribution of distance from previous activity location according to start time for the six identified groups is shown in Figure 3.3. We clustered six groups into three clusters based on representative activity type of each group. Group 1, 2, and 3 are clustered into working groups whose representative activity type is work. Group 4 is clustered into educational that group school activity represents it. Group 5 and 6 are clustered into non-working groups whose representative activity types are shopping, recreation and other activities.

Table 3.6 Number of users for each group according to the number of sequences

| | | Number of users | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Working groups | | | Educational group | Non-working grouops | | |
| | | #1 | #2 | #3 | #4 | #5 | #6 | |
| Number of activity Sequences | 2 | 1,221 (92%) | 3,471 (93.0%) | 3,824 (97.8%) | 15 (2.2%) | 872 (74.2%) | 1,073 (78.0%) | |
| | 3 | 35 (2.6%) | 160 (4.3%) | 44 (1.1%) | 334 (49.9%) | 117 (10.0%) | 130 (9.4%) | |
| | 4 | 66 (5.0%) | 93 (2.5%) | 39 (1.0%) | 275 (41.4%) | 177 (15.1%) | 166 (12.1%) | |
| | 5 | 4 (0.3%) | 6 (0.2%) | 2 (0.1%) | 36 (5.4%) | 7 (0.6%) | 6 (0.4%) | |
| | 6 | 2 (0.2%) | 2 (0.1%) | 1 (0.0%) | 9 (1.3%) | 2 (0.2%) | 1 (0.1%) | |
| Total cluster membership | | 1,466 | 4,797 | 5,374 | 3,181 | 2,258 | 2,858 | |
| Percentage in total | | 7.38 | 24.10 | 26.96 | 15.96 | 11.33 | 14.34 | |

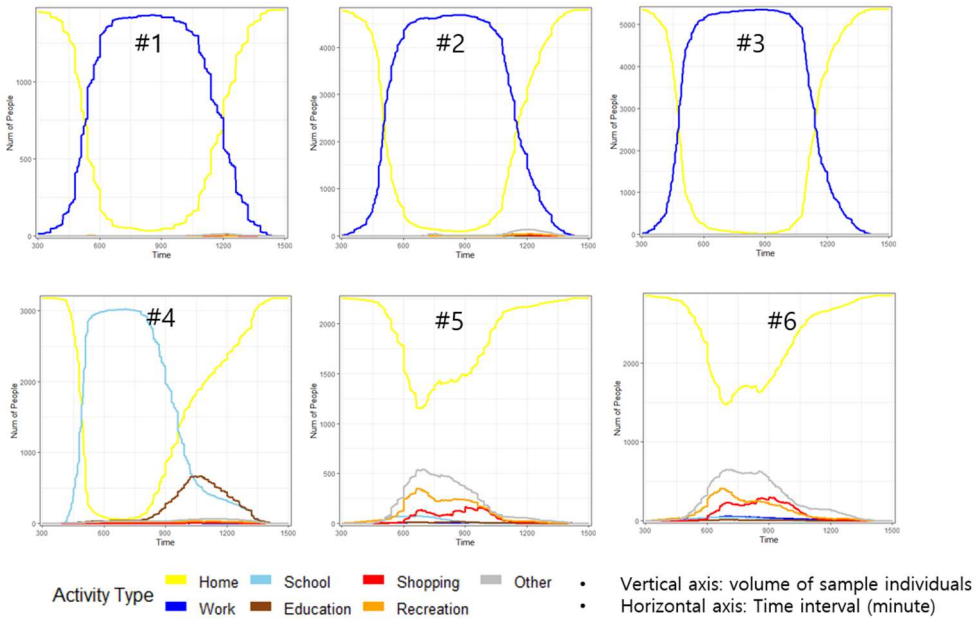


Figure 3.2 Temporal pattern of person-day activities for six identified groups

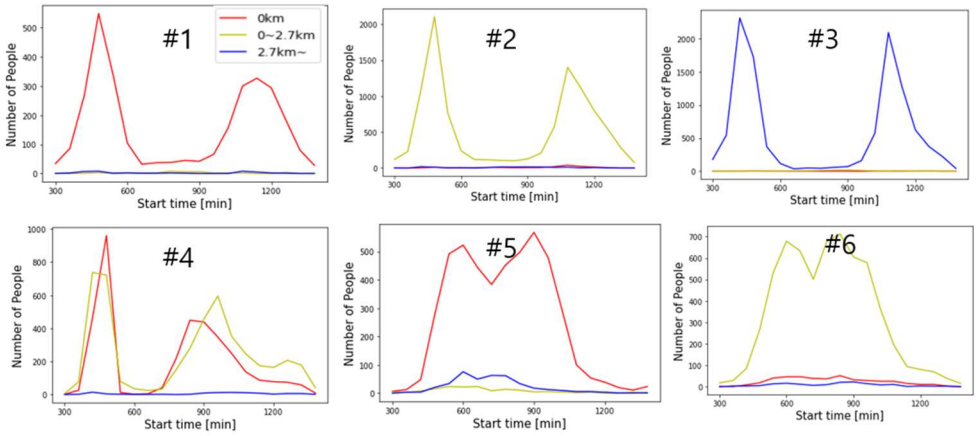


Figure 3.3 The distribution of distance from previous activity location according to start time for six identified groups

Table 3.7 Analysis of grouped data: share of different socio-demographic variables, membership analysis and representative patterns

| | | Sample mean (%) | Mean of Groups | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Working groups | | | Educational | Non-working groups | |
| | | | #1 | #2 | #3 | #4 | #5 | #6 |
| Income | Low | 14.59 | 15.69 | 11.44 | 6.51 | 6.76 | 33.08 | 28.38 |
| | Middle | 59.54 | 63.23 | 63.77 | 61.20 | 64.01 | 51.20 | 53.71 |
| | High | 25.87 | 21.08 | 24.79 | 32.29 | 29.24 | 15.72 | 17.91 |
| Residence type | Apartment | 52.27 | 35.61 | 46.20 | 56.42 | 56.46 | 54.69 | 54.90 |
| | Townhouse | 13.99 | 17.26 | 14.01 | 14.38 | 13.80 | 12.53 | 12.11 |
| | Multifamily Housing | 11.40 | 16.98 | 12.97 | 9.68 | 11.19 | 9.83 | 10.74 |
| | Detached house | 20.55 | 27.42 | 23.97 | 18.09 | 17.23 | 22.01 | 21.03 |
| | Studio | 1.31 | 1.30 | 2.115 | 1.19 | 0.88 | 0.53 | 0.91 |
| | Others | 0.48 | 1.43 | 0.71 | 0.24 | 0.44 | 0.40 | 0.31 |
| Ownership of vehicle | | 73.43 | 67.87 | 72.55 | 78.97 | 83.87 | 59.96 | 64.70 |
| Ownership of driver's license | | 58.93 | 66.03 | 76.49 | 85.58 | 6.19 | 38.84 | 47.83 |
| Male | | 49.77 | 49.86 | 55.33 | 68.74 | 51.81 | 23.03 | 21.80 |

Table 3.7 Continued

| | | Sample mean (%) | Means of Groups | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Working groups | | | Educational | Non-working groups | |
| | | | #1 | #2 | #3 | #4 | #5 | #6 |
| Age | 11-19 | 11.75 | 0.00 | 0.02 | 0.00 | 71.83 | 7.26 | 4.06 |
| | 20-29 | 10.71 | 3.34 | 5.27 | 6.81 | 25.09 | 2.61 | 3.60 |
| | 30-49 | 36.11 | 31.17 | 47.40 | 58.49 | 1.60 | 25.47 | 29.15 |
| | 50-64 | 24.81 | 42.09 | 36.15 | 29.14 | 0.50 | 19.18 | 24.70 |
| | More than 64 | 16.61 | 23.40 | 11.15 | 5.56 | 0.97 | 45.44 | 38.49 |
| Occupation Type | Profession | 3.75 | 4.02 | 5.80 | 7.13 | 0.03 | 0.31 | 0.63 |
| | Service | 11.79 | 27.15 | 23.56 | 15.28 | 0.57 | 1.02 | 2.03 |
| | Sales | 12.51 | 39.43 | 24.66 | 13.88 | 0.19 | 0.84 | 1.78 |
| | Office worker | 22.90 | 15.28 | 32.06 | 52.44 | 0.22 | 0.66 | 1.05 |
| | Agriculture | 0.25 | 0.20 | 0.48 | 0.39 | 0.00 | 0.09 | 0.10 |
| | Mechanic operator | 6.11 | 10.50 | 11.72 | 9.43 | 0.09 | 0.71 | 1.08 |
| | Housewife | 15.78 | 1.23 | 0.54 | 0.35 | 1.07 | 58.81 | 62.42 |
| | Unemployed/ Student | 25.46 | 0.62 | 0.50 | 0.28 | 97.52 | 32.73 | 27.57 |
| | Others | 0.50 | 1.5 | 0.67 | 0.78 | 0.03 | 0.22 | 0.10 |

# Chapter 4. Discussion

## 4.1. Working Groups

Table 4.1 Work activity variable description for Working groups

| Variable | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| Average start time | 9:16 | 9:03 | 8:18 |
| Average duration time [hr] | 9:39 | 9:24 | 9:49 |
| Average travel distance [km] | 0.22 | 2.21 | 10.85 |

Working groups consist of three groups which are divided into distance and number of activity sequences. Table 4.1 shows the average start time, average duration time, and average travel distance of work activity for groups of work cluster. Individuals in Group 1 travel only short distance. The number of users of group 1 is 1,328. 99.8% of group 1 users' first activity is work. The average start time of work activity is 9:16 and the average duration time of work activity is 9:39. 92.77% of group 1 users' number of activity sequences is two. Users in Group 2 travel only middle distance. The number of users of group 2 is 4,797, which is 24.10% of total sample. 99.9% of group 2 users' first activity is work. The average start time of work activity is 9:03 and the average duration time is 9:24. 87.45% of group 2 users' number of activity sequences is two. Users in Group 3 travel only long distance. The number of users of group 3 is 5,374, which is 26.96% of total sample. All group 3 users' first activity is work. The average start time of work activity is 8:18 and the average duration time is 9:49. 97.04% of group 3 users' number of activity sequences is two.

To analyze the socio-demographic homogeneity and heterogeneity for each group, I analyzed MANOVA test with 10 variables. Table 4.2 shows the result of MANOVA test. All independent socio-demographic variables are significantly different between groups.

Table 4.2 MANOVA test results for groups of Working groups

| Variable | Variable Type | F value | P-value |
|---|---|---|---|
| Income | 3 Categories | 50.394 | 0.000*** |
| Ownership of vehicle | Indicator | 31.950 | 0.000*** |
| Ownership of driver's license | Indicator | 51.752 | 0.000*** |
| Age | Numerical | 247.910 | 0.000*** |
| Male | Indicator | 37.879 | 0.000*** |
| Residence type | 6 Categories | 50.970 | 0.000*** |
| Occupation type | 9 Categories | 50.767 | 0.000*** |
| Students | Indicator | 146.540 | 0.000*** |
| Working days per week | 6 Categories | 143.750 | 0.000*** |
| Working hours per day | 4 Categories | 39.142 | 0.000*** |

Group 1 has a high proportion of over 50 years old and a high proportion of service workers and simple labor workers. This is because it is a group that includes a large number of senior citizens and enters a job with low barriers to employment due to its low skill level. Accordingly, the ratio of low-income user is high, and the ratio of multifamily and detached houses is high. The older the people, the more sensitive they are to the distance traveled, so they belonged to Group 1. Group 2 consisted of a high proportion of service workers and managers, with a high proportion of 30-60 years old and a high proportion of middle income. Group 3 has a high proportion of office worker. The ratio of middle-income and high-income is high because the higher the income level, the greater the distance to travel. The proportion of people aged 30-50 is high, and the proportion of living in apartments is high.

As a result of comparing the working groups, it can be seen that the length of the travel distance and the ratios of high-income, ownership of vehicle, ownership of driver's license, male, 20-40 years old, and residence in apartment correlate,

## 4.2. Educational group

Table 4.3 School activity variable description for Educational group

| Variable | Group 4 |
|---|---|
| Average start time | 8:10 |
| Average duration time [hr] | 7:50 |
| Average travel distance [km] | 0.52 |

Educational group consists of one group. Table 4.3 shows the average start time, average duration time, and average travel distance of school activity for groups of student cluster. Users in Group 4 travel short and middle distance. The number of users of group 4 is 3,181. 61.27% of group 4 users' first activity is school. The average start time of school activity is 8:10 and the average duration time of school activity is 7:50. 57.19% of group 4 users' number of activity sequences is two. 87.39% of group 4 users do education activity after school activity.

Table 4.4 shows the result of MANOVA test for student cluster. Age, residence type, ownership of driver's license, and student are significantly different between groups. Group 4 has a high percentage of teens and a high percentage of apartment dwellers. Apartments are generally close to commercial areas and are therefore short distances.

Table 4.4 MANOVA test results for groups of Student cluster

| Variable | Variable Type | F value | P−value |
|---|---|---|---|
| Income | 3 Categories | 0.882 | 0.008*** |
| Ownership of vehicle | Indicator | 5.890 | 0.03** |
| Ownership of driver's license | Indicator | 49.962 | 0.000*** |
| Age | Numerical | 24.017 | 0.000*** |
| Male | Indicator | 0.151 | 0.698 |
| Residence type | 6 Categories | 18.338 | 0.54 |
| Occupation type | 9 Categories | 5.299 | 0.021*** |
| Students | Indicator | 43.420 | 0.000*** |
| Working days per week | 6 Categories | 28.613 | 0.000*** |
| Working hours per day | 4 Categories | 22.581 | 0.000*** |

## 4.3. Non-working groups

Table 4.5 First activity variable description for Non-working groups

| Activity Type | Group 5 | | Group 6 | |
|---|---|---|---|---|
| | Average Start time | Average Duration time | Average Start time | Average Duration time |
| All | 13:41 | 4:05 | 13:24 | 4:43 |
| Shopping | 13:49 | 1:15 | 13:05 | 1:40 |
| Recreation | 11:42 | 2:41 | 11:41 | 2:24 |
| Other Activities | 12:06 | 3:41 | 11:21 | 2:27 |

Non-working groups consists of two groups which are divided into distance. Table 4.5 shows the average start time and average duration time of first activities for groups of non-working groups. Users in Group 5 travel only short distance. The number of users of group 5 is 2,258. The first activity of group 7 is 28.23%, 33.73%, and 36.58% of shopping, recreation, and other activities respectively. The average start time of the school activity in Group 5 is 13:41. For shopping, recreation, and other activities, since it is a non-working group with less time constraints, it has a longer duration time than Group 6 has an opportunity cost for long travel distance but has the latest start time. Users in Group 6 travel only middle distance. The number of users of group 6 is 2,858. The first activity of group 6 is 28.49%, 26.84%, and 41.83% of shopping, recreation, and other activities respectively.

Table 4.5 shows the result of MANOVA test for non-workers cluster. Income, ownership of vehicle, ownership of driver's license, age, residence type and student are significantly different between groups.

Group 5 has a high proportion of teenagers and more than60s, and a high proportion of high incomes. Due to the high income, there are many destinations with long distances. In the case

of unemployed people in their 60s, they are not sensitive to travel over long distances because public transportation costs decrease and physical and time allowances increase. Group 6 has a high percentage of people in their 70s or older. Because of the elderly, the travel distance decreases due to physical limitations.

Table 4.5 MANOVA test results for groups of Non-workers cluster

| Variable | Variable Type | F value | P-value |
|---|---|---|---|
| Income | 3 Categories | 12.929 | 0.000*** |
| Ownership of vehicle | Indicator | 9.266 | 0.002*** |
| Ownership of driver's license | Indicator | 19.752 | 0.000*** |
| Age | Numerical | 20.416 | 0.000*** |
| Male | Indicator | 3.017 | 0.082 |
| Residence type | 6 Categories | 3.741 | 0.003** |
| Occupation type | 9 Categories | 0.007 | 0.933 |
| Students | Indicator | 6.878 | 0.009** |
| Working days per week | 6 Categories | 0.136 | 0.712 |
| Working hours per day | 4 Categories | 0.086 | 0.769 |

# Chapter 5. Conclusion

Because most studies did not consider spatial information to recognize activity pattern, I recognized a new daily activity pattern that is based on the semantic trajectory to consider both spatial and socio-demographic information. Based on Household travel surveys, this study obtained the representative activity patterns of residents in Seoul by implementing semantic trajectory clustering method. I considered an individual record from travel survey as one trajectory, measure the similarity between those trajectories, and cluster based on the measured similarity. The clusters are evaluated by silhouette score. As a result, six activity patterns are generated. Each activity pattern is analyzed based on activity type, distance from previous activity location, and social demographic information. The proposed method enriches the traditional methods such as using socio-demographic attributes for classifying the population and is more straightforward, and easy to implement in practical ABM. Because semantic trajectory includes both socio-demographic and spatial information, ABM could be constructed more precise and approached diverse ways. However, the household travel survey data used in this study only reflect the aggregated spatial information. Further studies could be analyzed the activity pattern considering disaggregated spatial information and adapt spatial attribute (distance from home, distance from previous activity) to calculate MSM. 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 matrices, and/or simulate travel patterns under different circumstances

# Bibliography

Allahviranloo, Mahdieh, De Castaing, Ludovic Chastanet, Rehmann, Jakob, (2018), Mobility knowledge discovery to generate activity pattern trajectories, IEEE Conference on Intelligent Transportation Systems, Proceedings, pp.1-8

Andre Salvaro Furtado, Despina Kopanaki, Luis Otavio Alvares, Vania Bogorny. 2016), Multidiminsional Similarity Measuring for Semantic Trajectories, Transactions in GIS, vol. 20(2), pp.280-298

Anil NP Koushik, M. Manoj, N. Nezamuddin, (2020), Machine learning applications in activity-travel behaviour research: a review, Transport Reviews, pp. 1-24

Bhat, Chandra R., Guo, Jessica Y., Srinivasan, Sivaramakrishnan, Sivakumar, Aruna, (2004), Comprehensive econometric microsimulator for daily activity-travel patterns, Transportation Research Record, vol 1894, pp.57-66

Castiglione, J., Bradley, M., Gliebe, J., (2015). Activity-based travel demand models: A primer. No. SHRP 2 Report S2-C46-RR-1.

Chunmiao Wang, Chao Yang, Wen Ye, (2019), Activity Patterns Identification Based on National Household Travel Survey Data, CICTP 2019, pp. 6094-6108

D'Urso, Pierpaolo, Massari, Riccardo. (2013), Fuzzy clustering of human activity patterns, Fuzzy Sets and Systems, vol. 215, pp.29-54

Huang, Qunying, Wong, David W.S., (2016), Activity patterns, socioeconomic status and urban spatial structure: what can

social media data tell us?, International Journal of Geographical Information Science, vol 30, pp. 1873-1898

Jan Drchal, Michal Certicky, Michal Jakob, (2019), Data-driven activity scheduler for agent-based mobility models, Transportation Research Part C: Emerging Technologies, vol(98), pp.370-390

J.L. Bowman, M.E. Ben-Akiva, (2000), Activity-based disaggregate travel demand model system with activity schedules, Transportation Research Part A: Policy and Practice, vol.35, pp.1-28

Joh, Chang Hyeon, Arentze, Theo, Timmermans, Harry, (2001), Pattern recognition in complex activity travel patterns: Comparison of Euclidean distance, signal-processing theoretical, and multidimensional sequence alignment methods, Transportation Research Record, vol.1752, pp.16-22

Kihong Kim, (2014), Discrepancy analysis of activity sequences what explains the complexity of people's daily activity-travel patterns?, Transportation Research Record, vol 2413, pp.24-33

Kitamura, Ryuichi, Chen, Cynthia ,Pendyala, Ram M. (1997), Generation of synthetic daily activity-travel patterns, Transportation Research Record, vol 1607, pp.154-162

Kitamura, Ryuichi, (1988), An evaluation of activity-based travel analysis, Transportation, vol.15, pp.9-34

Li, Fei, Wang, Donggen, (2017), Measuring urban segregation based on individuals' daily activity patterns: A multidimensional approach, Environment and Planning A, vol.49(2), pp. 467-486

Mahdieh Allahviranloo, Will Recker, (2013), Daily activity pattern recognition by using support vector machines with multiple classes, Transportation Research Part B, vol 58, pp. 16-43

Mohammad Hesam Hafezi, Lei Liu, Hugh Millward, (2018), Learning Daily Activity Sequences of Population Groups using Random Forest Theory, Transportation Research Record, vol. 2672(47), pp.194-207

Mohammad Hesam Hafezi, Lei Liu, Hugh Millward, (2019), A time-use activity-pattern recognition model for activity-based travel demand modeling, Transportation, vol.46(4), pp. 1369-1394

Nyaupane, Gyan P, Graefe, Alan Burns, Robert C, (2003), Does distance matter? Differences in characteristics, behaviors, and attitudes of visitors based on travel distance, Northeastern Recreation Research Symposium, pp. 74-82

Recker, Wilfred W, Mcnally, Michael G, Root, Gregory S, (1985), Travel/activity analysis: pattern recognition, classification and interpretation, Transportation Research, vol.19A(4), pp. 279-296

Shou, Zhenyu, Di, Xuan, (2018), Similarity analysis of frequent sequential activity pattern mining, Transportation Research Part C: Emerging Technologies, vol.96, pp. 122-143

Snellen, Danielle MEGW. (2002), Urban form and activity-travel patterns : an activity-based approach to travel in a spatial context

Sreela, Parambath Koyilerian, Lakshmi, Mathavenkitachala, Anjaneyulu, Ranga, (2018). Modelling the Activity Travel Pattern of Commuters, International Journal for Traffic and Transport Engineering, vol.8(4), pp.481-493

# Abstract

최근 세분화된 활동 수요 기반 예측 모델이 개발됨에 따라 개별 통행 행동의 예측 정확도는 활동 생성 모듈에서 산출된 실제 정보에 따라 달라진다. 따라서 이 모듈에서 보다 정확하고 동질적인 정보를 생성하면 활동 수요 기반 예측 모델링의 예측 정확도가 높아질 수 있다. 통행 거리가 개개인의 통행 결정에 중요한 역할을 하지만 대부분의 연구에서는 공간 정보를 고려하지 않고 활동에 대한 정보, 사회인구 정보만을 활용하여 활동 패턴을 분석하였다. 본 연구는 활동 정보, 공간 정보 및 사회인구 정보를 고려한 새로운 일상 활동 패턴을 생성하여 분석하였다. 대표성을 가지는 활동 패턴을 생성하기 위해 일일 활동 에피소드를 다차원의 궤적으로 고려하였다. 활동 및 공간 정보는 MSM (Multidimensional Similarity Measure) 방법으로 다차원 궤적의 distance matrix 를 계산하였다. 계산된 distance matrix 를 기반으로 K-means algorithm 을 사용하여 대표 활동 패턴 그룹을 생성하였다. 다변량 분산 분석 (MANOVA) 테스트를 통하여 사회인구 정보를 분석하여 그룹 내 동질성, 그룹 간 이질성에 대해 분석하였다. 제안된 방법을 통해 가구통행실태조사 자료를 기반으로 서울 내에 발생한 통행에 대해 분석하였다. 결과적으로 근로자, 학생, 비 근로자 군집의 3 개 군집이 생성되고 각 군집은 통행 거리 및 활동 순서 개수에 따라 각각 3 개, 1 개, 2 개의 그룹으로 나뉘었다.

학번: 2019-26508