Unveiling Human Attributes through Life Pattern Clustering using GPS Data Only

Kazuyuki Shoji
shoji@ucl.nuee.nagoya-u.ac.jp
Nagoya University, Japan

Haru Terashima
haru@ucl.nuee.nagoya-u.ac.jp
Nagoya University, Japan

Naoki Tamura
tam@ucl.nuee.nagoya-u.ac.jp
Nagoya University, Japan

Shin Katayama
shin@nagoya-u.jp
Nagoya University, Japan

Kenta Urano
urano@nagoya-u.jp
Nagoya University, Japan

Takuro Yonezawa
takuro@nagoya-u.jp
Nagoya University, Japan

Nobuo Kawaguchi
kawaguti@nagoya-u.jp
Nagoya University, Japan

ABSTRACT

Clustering people by their life patterns is valuable in government and business fields. Existing studies often rely on semantic data such as Point of Interest or stay purpose. However, they have the problem that obtaining large datasets is difficult due to the need for annotation work. Some studies try to use only location data. However, they do not reveal the semantics of the area where visitors stay because they only label visited areas by significance according to duration and frequency of stay. In this paper, we propose a framework, LPSeL, for clustering people's Life Patterns at a Semantic Level using only raw GPS location data. LPSeL is based on the idea that analyzing human mobility first requires understanding urban space. Therefore, it begins with area modeling, which models areas in a city based on people's activities. Then, treating human mobility as a sequence of area representations makes it possible to model individuals by semantic-level characteristics of their life patterns. We showed that LPSeL is capable of estimating people's attributes from their life patterns using a real-world dataset consisting of GPS data collected from tens of thousands of smartphone users.

CCS CONCEPTS

• **Computing methodologies** → **Modeling methodologies**.

KEYWORDS

Urban Computing, Life Pattern Clustering, GPS data, Area Modeling, Mobility Behavior Modeling, Human Modeling

ACM Reference Format:
Kazuyuki Shoji, Haru Terashima, Naoki Tamura, Shin Katayama, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. 2024. Unveiling Human Attributes through Life Pattern Clustering using GPS Data Only. In *The 32nd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '24), October 29-November 1, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3678717.3691309

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGSPATIAL '24, October 29-November 1, 2024, Atlanta, GA, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1107-7/24/10
https://doi.org/10.1145/3678717.3691309

1 INTRODUCTION

Understanding people's life patterns should lead to estimating personal attributes and preferences. If this becomes possible with the large location dataset collected daily from smartphones, it will lead to a detailed analysis of human flow dynamics on an urban scale.

Existing studies on human life patterns often use datasets with semantics such as Points of Interest (POI) or stay purpose [2, 7, 9]. Semantic data is easier to interpret than GPS location data, making it useful for a deeper understanding of human behavior. However, the need to annotate mobility data makes it challenging to obtain large datasets. There are, of course, studies that use location data alone to mine life patterns [1, 3, 8]. However, they only label areas by significance according to the duration and frequency of their stay, ignoring what kind of place the areas are. This makes it impossible to distinguish, for example, between people who work in a store and an office. Therefore, analyzing human mobility at the semantic level in a dataset containing many users, such as GPS location data, is a valuable research topic that can reflect people's more detailed characteristics in urban analysis.

This paper proposes LPSeL, a framework for modeling and clustering people by semantic life pattern characteristics using only GPS location data. LPSeL consists of three modules: "area modeling" to model each area of a city, "behavioral modeling" to model the mobility behavior of people living in the city, and "human modeling" to model individuals in terms of their behavioral schedules. The key point of LPSeL is the series of processing steps from area modeling to behavior modeling and from behavior modeling to human modeling, which enables interpretation at the semantic level of who has what kind of life patterns. We evaluate the effectiveness of LPSeL by estimating people's attributes using real-world GPS location data collected from tens of thousands of smartphone users and by demonstrating the certainty of the estimation results.

2 PRELIMINARIES

Location Data. Our location data is GPS positioning data collected from apps installed in users' smartphones with prior consent. A point of location (p) is represented as $p = (lat, lon, t, acc)$, where lat,

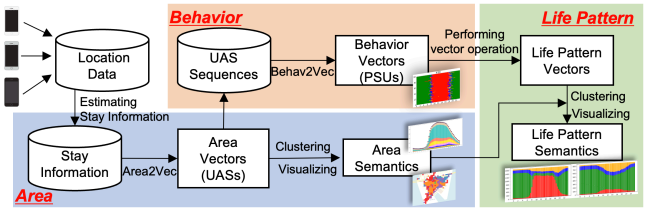


Figure 1: Overview of LPSeL.

Table 1: Stay information used in area modeling.

Day of Week	Weekday, Weekend (including holidays)
Arrival Time	0:00-1:59, 2:00-3:59, …, 20:00-21:59, 22:00-23:59
Stay Time	-29, 30-59, 60-119, 120-239, 240-359, 360-719, 720-1079, 1080-1439, 1440- (unit:min)

lon, *t*, and *acc* denote latitude, longitude, timestamp, and positioning accuracy, respectively.

Area2Vec and UAS. Area2Vec (Area to Vector) [5] is a representation learning model for modeling areas based on people's stay information (Section 3.1), and UAS (Usage of Area with Stay information) refers to the vector representation of an area generated by Area2Vec. An area refers to each grid of meshed a city, which in this paper is 50m square. Area2Vec makes it possible to represent human mobility described in geocoordinates as a UAS sequence.

Behav2Vec and PSU. Behav2Vec (Behavior to Vector) is a representation learning model for modeling UAS sequences on a daily basis (Section 3.2), and PSU (Pattern of Sequence with UAS) is a vector representation of a UAS sequence generated by Behav2Vec. Vector operations of PSUs create human representation vectors according to life pattern characteristics (Section 3.3).

Overview of LPSeL. Our goal is to model individuals according to their life patterns at the semantic level, i.e., what characteristics of places they stayed, from location data alone, without semantic data such as POIs. The overview of the LPSeL framework that makes it possible is shown in Figure 1. LPSeL starts with estimating people's stay information. Then, using the stay information, we create a UAS of each area that reflects its usage characteristics by Area2Vec. Once UASs have been created, human mobility can be represented by them. The next step is to create a PSU of each UAS sequence that reflects the sequence pattern characteristics by Behav2Vec. Finally, individuals are modeled in terms of their life patterns through the vector operations of PSUs. Applying clustering and visualization allows life patterns to be interpreted at the semantic level.

3 LPSEL

3.1 Area Modeling

Area2Vec is inspired by Word2Vec [4] and consists of input, hidden, and output layers. Area2Vec converts an area into a representation vector, UAS, embedded with "usage" revealed from people's stay information. The training data is a pair of area and stay information, and given an area a, learning proceeds to predict the stay information s. The conditional probability model is defined as follows: $P(s|a) = \frac{\exp(\mathbf{v}_s \cdot \mathbf{u}_a)}{\sum_{s' \in S} \exp(\mathbf{v}_{s'} \cdot \mathbf{u}_a)}$, where $\mathbf{v}_s$ is the weight vector of stay information s in output layer, $\mathbf{u}_a$ is the vector of area s, and S is the combination of all stay information. We used the stay information shown in Table 1; therefore, $|S| = 2 \times 10 \times 12 = 240$. As for loss function, it takes the negative log-likelihood in $P(s|a)$ and is defined as follows: $L = -\frac{1}{|D|} \sum_{(a,s) \in D} \log P(s|a)$, where D is the set of training data. In training, the parameters of the entire model are adjusted to minimize L. Finally, each row of weights between the input and hidden layers is treated as a UAS for each area.

3.2 Behavior Modeling

Behav2Vec is an LSTM Autoencoder that converts a UAS sequence into a representation vector, PSU, embedded with mobility pattern characteristics. First, the encoder is given a UAS sequence, e.g., $\{UAS_1^{10}, UAS_2^{54}, ..., UAS_T^{22}\}$, where UAS_t^i means that the area ID at time t is i, and $T = 48$ because a daily mobility is divided into 30-minute segments. Then, the hidden state h_t and cell state c_t at time t are updated by UAS_t and h_{t-1}, c_{t-1} as follows: $h_t, c_t = f_{LSTM}(h_{t-1}, c_{t-1}, UAS_t)$. The output h_T after the last UAS_T is processed becomes a PSU for the entire input sequence.

The decoder's goal is to reconstruct the input sequence from the PSU of the input sequence. Initially, the hidden and cell states are initialized as follows: $h_1^{dec} = h_T$ and $c_1^{dec} = 0$. The hidden state h_t^{dec} and cell state c_t^{dec} at time t are updated by h_{t-1}^{dec}, c_{t-1}^{dec} and zero vector [6] as follows: $h_t^{dec}, c_t^{dec} = f_{LSTM}(h_{t-1}^{dec}, c_{t-1}^{dec}, 0)$. And h_t^{dec} is fed to a linear layer to generate $\hat{UAS}_t$ which is a reconstructed UAS_t. The loss function is defined to minimize the error between the input sequence and the reconstructed sequence, using the mean squared error as follows: $L = \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{UAS}_t - \hat{\mathbf{UAS}}_t\|^2$. The encoder and decoder parameters are adjusted to minimize the loss function L in training. After training, only the encoder is used, and when the encoder is fed a UAS sequence, its PSU is generated.

3.3 Human Modeling

Individuals are modeled using PSU based on their life patterns in human modeling. We create each individual's representation vector from behavioral characteristics on weekday and weekend (including holidays) as follows:

$$\begin{aligned} \mathbf{v}_u &= concat(PSU_{u,Weekday}, PSU_{u,Weekend}) \\ &= concat(\frac{1}{|M|} \sum_{m \in M} PSU_{u,m}, \frac{1}{|N|} \sum_{n \in N} PSU_{u,n}) \end{aligned}$$

$PSU_{u,x}$ denotes the PSU corresponding to the behavior x taken by human u. M and N denote the set of behaviors generated by human u on weekdays and weekends, respectively. That is, the vector formed by concatenating the averages of PSUs for each weekday and weekend of a human u is the representation vector of that person. It should be noted that this is just one example. Human vectors can be constructed flexibly by utilizing PSUs. For example, by concatenating the PSUs corresponding to each day of the week from Monday to Sunday, i.e., $\mathbf{v}_u = concat(PSU_{u,Mon.}, PSU_{u,Tue.}, ..., PSU_{u,Sun.})$, you can model individuals on a weekly basis.

4 EXPERIMENT

In this paper, Nagoya City, Aichi Prefecture, Japan, is the experiment's target. Nagoya is one of the three largest metropolitan cities

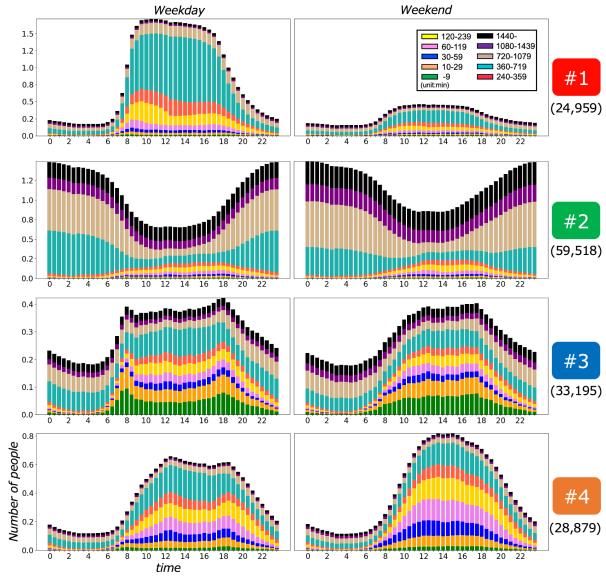


Figure 2: Graphing of UAS clustering result.

in Japan. With its diverse urban functions such as residential, office, downtown, and transportation areas, Nagoya stands out as a prime city for human mobility research. The dataset used in the experiment is GPS data provided by Blogwatcher Inc.[1] The location data was collected from apps installed on users' smartphones with prior consent. We used the data from April and May of 2019 in Nagoya.

4.1 Interpreting Areas Semantically

In LPSeL, it is first necessary to understand characteristics embedded in UASs to analyze human life patterns at the semantic level. Therefore, the first step is to interpret UASs. Figure 2 shows the result of clustering all UASs in Nagoya into four clusters. These stacked bar graphs represent the number of people by their length of stay inside the areas included in each cluster. The color of each layer represents the length of stay time. The left is for weekdays, and the right is for weekends, including holidays. The number of areas in each cluster is written in parentheses under the cluster number. The horizontal axis shows the time, and the bin width is 30 minutes. The vertical axis represents the number of people. By looking at this graph, we can interpret each area's characteristics. The interpretations of Cluster #1 to #4 and the reasons are as follows.

#1 : "Office areas" • Long-term stays from about 8:00. • Low number of people at night and on weekends.

#2 : "Residential areas" • Long-term stays from night to morning on both weekdays and weekends. • Decrease people during the daytime, probably due to people going to work.

#3 : "Traffic areas" • A large proportion of short-term stays. • Short--term stays occur during commuting between 7:00 and 18:00.

#4 : "Commercial areas" • More people on weekends than weekdays. • A large percentage of short- and medium-term stays during the daytime. • Increase people at lunch and dinner on weekdays.

[1]https://www.blogwatcher.co.jp/

It is important to note that the number of clusters depends on the detail required for the analysis. As the number of clusters is increased, you perform more detailed urban analysis.

4.2 Mining Representative Life Patterns

We cluster the human modeling results for 59,179 people and interpret their semantics by visualizing the life pattern characteristics. We picked up a few clusters with representative life pattern characteristics, shown in Figure 3a. This figure shows the probability distribution of which areas people stayed in at which times of the day. The left and right sides represent weekdays and weekends, respectively. The horizontal axis represents time, where the bins are 30 minutes. The vertical axis represents the percentage. The colors correspond to the area modeling results shown in Figure 2.

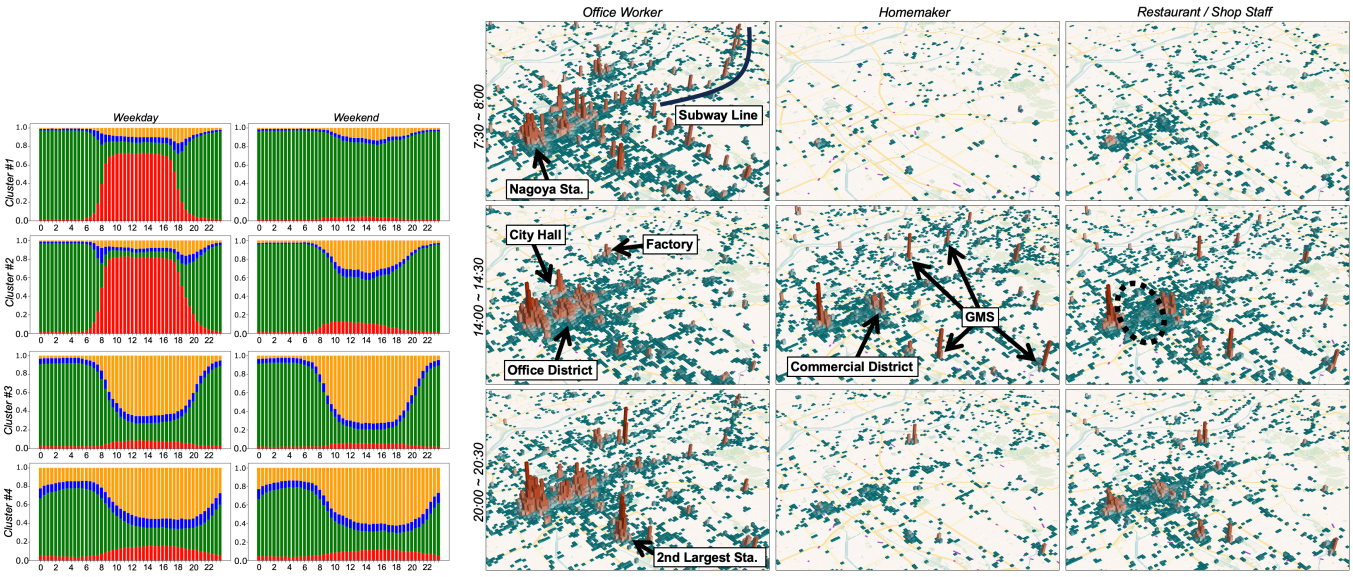
It can be observed that a high percentage of people in #1 spent time in the office area during the day and in the residential area from nighttime to morning on weekdays. On weekends, they spent most of their time in residential areas. Accordingly, we can estimate that the people in #1 are "office workers." The people in #2 can be estimated to be "office workers," as those in #1, but they spent much of their time in commercial areas on weekends. This means that they are "outdoor people on weekends" in addition to being "office workers." From #3, it can be seen that they spent a high percentage of their time in the residential area from night to morning on both weekdays and weekends, and a high percentage of their time is spent in the commercial area roughly from 9:00 to 21:00. We can estimate that they are "staff working in stores and entertainment facilities." People in #4 are more likely than those in #3 to stay in the commercial area during the late evening and even into the early morning. This means we can estimate that these people are "staff at restaurants or bars" open later than the stores where the people in #3 work. Thus, even when the same attribute labels are assigned, more detailed personas emerge by looking at life patterns by LPSeL.

4.3 Verifying Attribute Estimation Result

We prove the validity of the estimated attributes by looking at their mobility in real space. Figure 3b shows the population distribution of office workers, homemakers, and store staff at 7:30-8:00, 14:00-14:30, and 20:00-20:30 on weekdays. The grid is 100m square. Note that grids with less than five unique individuals are not displayed.

First, we look at 7:30-8:00. Regarding office workers, it can be observed that many areas had large populations. The areas with large populations were arranged at regular intervals because of the presence of subway stations. Office district had a large population as well. In contrast, homemakers and store staff had few populated areas. This period is commuting time, consistent with our intuition that office workers gather at the station. Meanwhile, since the hours are early for homemakers and store staff, it is natural that there are few places for them to gather in large numbers.

Next, we look at 14:00-14:30. Office workers were mainly concentrated in office areas. On the other hand, we observed scattered areas with large populations of homemakers, most of which are identified as general merchandise stores (GMS). As for the staff, it can be observed that the same areas as the homemakers were heavily populated. During this period, office workers generally work in their offices, homemakers go shopping or work at home, and



(a) Representative life patterns. **(b) Population distribution of office workers, homemakers, and shop staff.**

Figure 3: Clustering and attribute estimation based on life patterns.

staff work in the stores. Namely, the population distribution of each attribute during this period is consistent with our intuition. Also, we confirmed that store staff did not gather in the office district surrounded by the dotted line in the figure. This is a clear result because store staff work in different locations than office workers, indicating that their life patterns are appropriately differentiated.

Finally, we look at 20:00-20:30. Many office workers were still in the office district and along the subway line as in the morning. For homemakers, they had few areas with large numbers of people. As for store staff, they gathered in commercial areas and around large stations. We assumed that most office workers and homemakers went home and that few areas were where people gathered. The results for homemakers matched this assumption. However, it is worth noting that many office workers remained in the city, particularly at large stations. This could be due to their dedication to working overtime or socializing on Fridays. As for store staff, since restaurants and bars are still open, we expected that people would gather in commercial areas and at train stations, and indeed, we could confirm this situation, a result that fits our intuition.

The above results show that the LPSeL helps estimate people's attributes since the population distribution in the city for each of the estimated attributes is explainable and consistent.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose a new framework, LPSeL, which enables the estimation of people's attributes by modeling and clustering people based on their semantic-level life patterns using only raw GPS location data. We showed that LPSeL is effective in estimating individuals' attributes using a real-world dataset consisting of GPS data collected from tens of thousands of smartphone users. Future work includes making LPSeL robust to missing data, allowing travel time to be reflected in modeling results as a characteristic, and additional evaluation of LPSeL using multiple datasets.

ACKNOWLEDGMENTS

This research was supported in part by JST CREST (JPMJCR21F2, JPMJCR22M4), NICT (222C01, 22609), NEDO (JPNP23003, JPJ012495), JSPS KAKENHI (22H03580, 22K18422), and JST ACT-X. We would like to thank Blogwatcher Inc. for providing the valuable data.

REFERENCES

[1] Irad Ben-Gal, Shahar Weinstock, Gonen Singer, and Nicholas Bambos. 2019. Clustering Users by Their Mobility Behavioral Patterns. *ACM Trans. Knowl. Discov. Data* 13, 4, Article 45 (aug 2019), 28 pages.

[2] Hancheng Cao, Fengli Xu, Jagan Sankaranarayanan, Yong Li, and Hanan Samet. 2020. Habit2vec: Trajectory Semantic Embedding for Living Pattern Recognition in Population. *IEEE Transactions on Mobile Computing* 19, 5 (2020), 1096–1108.

[3] Wenjing Li, Haoran Zhang, Jinyu Chen, Peiran Li, Yuhao Yao, Xiaodan Shi, Mariko Shibasaki, Hill Hiroki Kobayashi, Xuan Song, and Ryosuke Shibasaki. 2023. Metagraph-Based Life Pattern Clustering With Big Human Mobility Data. *IEEE Transactions on Big Data* 9, 1 (2023), 227–240.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems* 26 (Oct 2013).

[5] Kazuyuki Shoji, Shunsuke Aoki, Takuro Yonezawa, and Nobuo Kawaguchi. 2024. Area Modeling using Stay Information for Large-Scale Users and Analysis for Influence of COVID-19. arXiv:2401.10648

[6] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. 2015. Unsupervised Learning of Video Representations Using LSTMs. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*. 843–852.

[7] You Wan, Chenghu Zhou, and Tao Pei. 2017. Semantic-Geographic Trajectory Pattern Mining Based on a New Similarity Measurement. *ISPRS International Journal of Geo-Information* 6, 7 (2017).

[8] Fengli Xu, Tong Xia, Hancheng Cao, Yong Li, Funing Sun, and Fanchao Meng. 2018. Detecting Popular Temporal Modes in Population-scale Unlabelled Trajectory Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1 (mar 2018).

[9] Chao Zhang, Jiawei Han, Lidan Shou, Jiajun Lu, and Thomas La Porta. 2014. Splitter: Mining Fine-Grained Sequential Patterns in Semantic Trajectories. *Proc. VLDB Endow.* 7, 9 (may 2014), 769–780.