Discovering and Characterizing Mobility Patterns in Urban Spaces: A Study of Manhattan Taxi Data

Lisette Espín-Noboa
GESIS - Leibniz Institute for the Social Sciences
Lisette.Espin@gesis.org

Florian Lemmerich
GESIS - Leibniz Institute for the Social Sciences
Florian.Lemmerich@gesis.org

Philipp Singer
GESIS - Leibniz Institute for the Social Sciences
Philipp.Singer@gesis.org

Markus Strohmaier
GESIS - Leibniz Institute for the Social Sciences
& University of Koblenz-Landau
Markus.Strohmaier@gesis.org

ABSTRACT

Nowadays, human movement in urban spaces can be traced digitally in many cases. It can be observed that movement patterns are not constant, but vary across time and space. In this work, we characterize such spatio-temporal patterns with an innovative combination of two separate approaches that have been utilized for studying human mobility in the past. First, by using non-negative tensor factorization (NTF), we are able to cluster human behavior based on spatio-temporal dimensions. Second, for characterizing these clusters, we propose to use HypTrails, a Bayesian approach for expressing and comparing hypotheses about human trails. To formalize hypotheses, we utilize publicly available Web data (i.e., Foursquare and census data). By studying taxi data in Manhattan, we can discover and characterize human mobility patterns that cannot be identified in a collective analysis. As one example, we find a group of taxi rides that end at locations with a high number of party venues on weekend nights. Our findings argue for a more fine-grained analysis of human mobility in order to make informed decisions for e.g., enhancing urban structures, tailored traffic control and location-based recommender systems.

Keywords: Human Mobility; Tensor Factorization; HypTrails

1. INTRODUCTION

Human mobility can be studied from several perspectives utilizing different kinds of data from the online (e.g., Twitter) and the offline (e.g., taxi rides) world. A large body of work has focused on identifying general mechanisms that guide and explain human mobility behavior on an individual [9] or collective level [17]. For example, previous research has shown that human mobility is highly predictable [24] and shows temporal and spatial regularity [9]. At the same time, it exhibits also spatio-temporal heterogeneity, see for example [3, 15]. For instance, daily routines such as going from home to work (space) in the morning (time) can be observed. This argues for a more fine-grained analysis, that goes beyond the universal mobility patterns which tend to ignore several aspects of human mobility such as time, weather or race of people. Towards that end, we propose to *discover and characterize mobility patterns* in human behavior in urban space.

Material and approach. We expand existing research [3, 15] on studying human mobility with a case study using taxi data of Manhattan. For identifying behavioral differences in terms of time and space, previous research [20, 26] has suggested to utilize tensor decomposition [4]. However, interpreting results from tensor decomposition has mostly been based on personal intuitions. On the other hand, recent research [1, 23] proposed methods that allow to understand human sequences by comparing hypotheses about the production of trails at interest. Yet, this approach is limited in the sense that it can only explain global behavior without being able to provide more detailed insights. To circumvent these limitations, we propose a unique and original combination of tensor decomposition and HypTrails to discover and characterize human behavior on a spatio-temporal level. We first utilize *non-negative tensor factorization* (NTF) [4] for automatically identifying clusters of mobility behavior. Second, for characterizing these clusters, we utilize *HypTrails* [23]—a Bayesian approach for expressing and comparing transitional hypotheses about human trails.

Findings and contributions. Our main contributions are three-fold: First, we present an innovative combination of two methodologies, i.e., NTF and HypTrails, in order to characterize heterogeneous human mobility behavior. Second, we incorporate existing human mobility patterns into a hypothesis-based schema built upon human beliefs. Third, we demonstrate the benefits of online data for characterizing human behavior on a spatio-temporal level. As one example, we discover a group of taxi rides that have drop-off locations with a high number of party venues on weekend nights. Results of this study can improve e.g., planning of future events or reconstructions, traffic control or location-based recommender systems.

2. DATASETS

In this work, we study *taxi rides* in Manhattan—one of the most densely populated areas in the world. However, the presented methodology can be also applied to other kinds of mobility data. We represent human mobility as *user trails* representing single *transitions* between taxi pick-up and drop-off locations. To construct a rich set of hypotheses for characterizing the movement of users, we additionally retrieved information on local venues (such as parks or churches) from Foursquare and public census data, i.e., data on demographics and land-use.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890468 .

An extended version of this work is available on arXiv: http://arxiv.org/abs/1601.05274.

Taxi rides. We use publicly available data[1] of about 173M NYC Taxi rides in 2013. It consists of anonymized records registering when and where a taxi ride started and ended and various features such as the number of passengers. From all records, we removed taxi rides outside the area of Manhattan and some inconsistencies such as records with trip_distance ≤ 0, trip_time_in_secs ≤ 0 and passenger_count ≤ 0; our final dataset consists of 143M rides. A description of all attributes and datatypes can be found in the *TLC Taxi Data - API Documentation*[2].

Census data. Our methodology (see Section 3) operates on a discrete tract state space. We extracted the specifications of all 288 tracts in Manhattan used in the 2010 census from the *NYC Planning portal*[3] and mapped the GPS coordinates of all taxi rides to their respective tracts. We also queried the *NYC OpenData*[4] and the *American FactFinder*[5] databases for accessing relevant data such as census and land-use. Also, we calculated the spatial overlap between land-use types such as residential and commercial zones and each tract to obtain information on a tract level.

Foursquare venues. We gathered additional information about physical places such as churches and parks situated in Manhattan by querying the Foursquare Search API[6] to extract places in each tract for 10 different categories[7] (e.g., Residence, Nightlife Spot). Overall, we collected 153K unique places within Manhattan. Every venue was mapped to its respective tract.

Centroids. Typically, popular places are most likely to be visited at any time. For this reason, we considered three candidate places as centroids to study whether people visit them or not: (a) City center —approximated as the geographical center of Manhattan— (40.79090, -73.96640, 018100), (b) Flatiron Building (40.74111, -73.98972, 005600), and (c) Times Square (40.75773, -73.98570, 011900). Tuples represent (latitude, longitude, tract id).

3. METHODOLOGY

The goal of this work is to discover and characterize mobility patterns in taxi data for better understanding people's travel behavior within Manhattan. To that end, we propose an innovative combination of two methodologies. First, we suggest to use *non-negative tensor factorization* (NTF) [4] for automatically clustering human mobility behavior. Research has shown that NTF can detect latent features of human mobility in different dimensions such as space and time, cf. [26]. Second, to characterize these clusters, we utilize *HypTrails* [23]—a Bayesian approach for expressing and comparing hypotheses about human trails. We outline in the following both methodological components of this work, but refer to the original publications for details.

Clustering mobility patterns. For clustering the data, we utilize NTF which decomposes a given n-way tensor $\underline{\mathbf{X}}$ into n components (matrices) that approximate the original tensor when multiplied with each other. Each matrix contains information on r factors (clusters). In this paper, we define a three-way tensor of taxi rides whose dimensions capture human transitions from one place to another at a certain time: pick-up tract, drop-off tract and pickup time (hour of week); thus, clustering in terms of both space and time. Each element of every component determines the scale of mobility flow (weight) with respect to the corresponding factor. In other words, the higher the weight, the more dominant that instance is in that cluster. Similar to other clustering methods, defining the number of clusters is arbitrary. However, there exist some guidelines to determine a good value of r, see e.g., [8]. In this work, we are not focused on finding the most appropriate number of behavioral components, but being able to characterize different behaviors, which will be detailed in the following section.

Characterizing clustered mobility patterns. For characterizing the clustered human mobility behavior, we utilize HypTrails [23], a Bayesian approach for expressing and comparing hypotheses about human trails. HypTrails models the data with first-order Markov chains where the state space contains all 288 tracts of Manhattan. Fundamentally, hypotheses are represented as matrices Q expressing beliefs in Markov transitions; Section 4 describes the hypotheses about human mobility used in this work. Elements $q_{i,j}$ indicate the belief in the corresponding transition probability between states i and j; higher values refer to higher belief. The main idea of HypTrails is to incorporate these hypotheses as Dirichlet priors into the Bayesian inference process. HypTrails automatically elicits these priors from expressed hypotheses; an additional parameter k (weighting factor) reflects the overall strength of belief in a hypothesis. For comparing the relative plausibility of hypotheses, HypTrails utilizes the *marginal likelihood* (*evidence*) of the Bayesian framework which describes the probability of the data given a hypothesis. We can infer the partial ordering based on the plausibility of a given set of hypotheses by ranking their evidences from the largest to the smallest for a specific value of k.

4. HYPOTHESES

In this section, we describe the hypotheses we used to characterize the individual clusters with HypTrails. These are expressed as hypothesis matrices Q, in which the elements $q_{i,j}$ capture a belief in the likelihood of people transitioning from tract i to tract j (see Section 3).

Our hypotheses are mostly based on existing theories. In this regard, the most prevailing human mobility model is the *gravitational law* [30] $q_{i,j} = \frac{venues(i).venues(j)}{dist(i,j)}$, which explains mobility by an attraction force between places i and j (e.g., number of venues) and the inverted shortest distance between them. However, due to some limitations found in this model [22], several alternatives have emerged to circumvent these issues. In that direction, the *rank model* [17] $q_{i,j} = \frac{1}{|\{w:dist(i,w)<dist(i,j)\}|}$ indicates that the number of people traveling to a given location is inversely proportional to the number of places w surrounding the source location. Similarly, the so called *intervening opportunities* model [25] $q_{i,j} = \frac{|\{w_1:dist(i,w_1)=dist(i,j)\}|}{|\{w_2:dist(i,w_2)<dist(i,j)\}|}$ additionally includes the number of opportunities (i.e. venues) at a given distance. In order to express the assumption that people prefer to visit places that are similar to the departure place (e.g., vectors V containing the categorical distribution of places), also the *Cosine similarity* $q_{i,j} = \frac{V_i.V_j}{||V_i||||V_j||}$ can be employed.

These theories require additional data, e.g., to determine the weight of places for the gravitational law. We categorize our hypotheses with respect to the utilized additional datasets into three types: *Distance-based*, *Foursquare* and *Census* hypotheses, see Table 1. Additionally, we use the uniform hypothesis as a baseline to express the belief that all tracts are equally likely to be visited. For all hypotheses, we set the diagonal of Q to 0 to avoid self-loop transitions—accounting for only 1.5% of all taxi rides—which (in this work) do not contribute on mobility.

[1] http://www.andresmh.com/nyctaxitrips/

[2] https://dev.socrata.com/foundry/data.cityofnewyork.us/gkne-dk5s

[3] http://www.nyc.gov/dcp

[4] https://nycopendata.socrata.com/

[5] http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

[6] https://developer.foursquare.com/docs/venues/search

[7] https://developer.foursquare.com/categorytree

Table 1: Tract properties. These three categories contain all tract indicators, i.e., statistics about tracts, used in combination with universal theories to construct hypotheses. For instance, the hypothesis *church* expresses that the probability of visiting a certain location is proportional to its number of churches (using for example the density theory).

Category	Properties per tract
Distance-based	Fixed points of interest: *Geographical Center*, *Flatiron Building*, *Times Square*.
Foursquare	Number of venues: *Arts & Entertainment, Education (i.e., colleges, universities, elementary schools and high schools), Food, Nightlife Spot, Outdoors & Recreation, Work (i.e., auditoriums, buildings, convention centers, event space, factories, government buildings, libraries, medical centers, military base, non-profit, office, post office, prison, radio station, recruiting agency, TV station, and ware house), Residence, Shop & Service, Travel & Transport, and Church*.
Census	*Population size*, and *Tract area*. Percentage of: *White people, Black people, People in labor force, Unemployed people, People below poverty level, People above poverty level*. Number of places: *Libraries, Art Galleries, Theaters, Museums, WiFi Hotspots, and Places of Interest*. Moreover, the occupied area of: *Residential Zoning, Commercial Zoning, Manufacturing Zoning, Park properties, Historic Districts and Empower zones*.

Distance-based hypotheses. Based on the *geographic distance* [10] (likewise its inverse: $q_{i,j} = \frac{1}{dist(i,j)}$), we can construct very intuitive hypotheses: *proximity* and *centroid*. The proximity hypothesis assumes that places near the current location are more likely to be visited next, the centroid hypothesis suggests that locations near to the city center (specified by a fixed geo-coordinate, see Section 2) are more likely for the next stop. According to these hypotheses, the location of the next visited place follows a two dimensional *Gaussian distribution* $q_{i,j} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{dist(i,j)^2}{2\sigma^2}}$ that is centered at the current location i (or a city center), see [1] for a more detailed description. For the parametrization of the distribution, we included 7 different values for the standard deviation σ, i.e., $\sigma \in \{0.01, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0\} km$.

Foursquare hypotheses. We leverage Foursquare venues to measure the *density* $q_{i,j} = |(venues(j)|$ of a place, which consists of the number of all venues in a given tract (e.g., destination). Similarly, the cumulative check-ins in an area can be used to measure the *popularity* $q_{i,j} = \sum_{V \in venues(j)} checkins(V)$ of a given tract. Accordingly, these can be combined with the universal mobility theories (i.e., gravitational, rank-distance and intervening opportunities). Additionally, we group places according to their category (e.g., Residence, Church). In other words, we use categories as filters. Thus, every Foursquare category induces a subset of all venues per state. Table 1 shows all 10 Foursquare categories included in this study. To avoid an abundant amount of hypotheses, we only use the gravitational theory in combination with these category-based hypotheses. Furthermore, we construct a similarity-based hypothesis that suggests that transitions are more likely between two states that have a similar category distribution of venues based on *Cosine similarity*.

Census hypotheses. The Census hypotheses are based on tract-level information on demographics or land-use, which replace the density of a place used in the Foursquare hypotheses. Table 1 shows all 20 indicators used to formulate this kind of hypotheses. Cosine similarity measures were obtained under three different categories: Race Group (i.e., white, black, american indian, asian, hawaiian and other pacific islander, other race, two races), Poverty Level (i.e., below, above) and Employment status (i.e., employed, unemployed, in labor force).

Overall we defined 70 hypotheses: (a) 1 uniform, (b) 29 distance-based (i.e., the proximity and 3 centroid hypotheses with 7 σ settings each and the inverse geographical distance), (c) 17 from Foursquare (i.e., density, popularity, gravitational, rank-distance and intervening opportunities for all venues only, gravitational for each Foursquare category and 1 similarity), and (d) 23 from Census data (i.e., 20 from the gravitational with each census indicator and 3 from similarities).

5. EXPERIMENTS

This section reports on experimental results obtained by applying the presented methods to the Manhattan taxi data.

5.1 Configuration

As described in Section 3, we started by using NTF in order to identify different clusters in the taxi data. We worked with a three-way tensor whose dimensions represent the pickup hour of week, pickup and drop-off tracts of the taxi ride. For the hour of week, the first state corresponds to Monday from 12:00 a.m. to 12:59 a.m. and the last state to Sunday from 11:00 p.m. to 11:59 p.m. We did not include the time of the drop-off because most of the rides last less than an hour. Therefore, the tensor has size $168 \times 288 \times 288$. After experimenting with different parameters, we set the number of clusters to $r = 7$ to find seven different groups, as this number subjectively captured all behavioral components best.

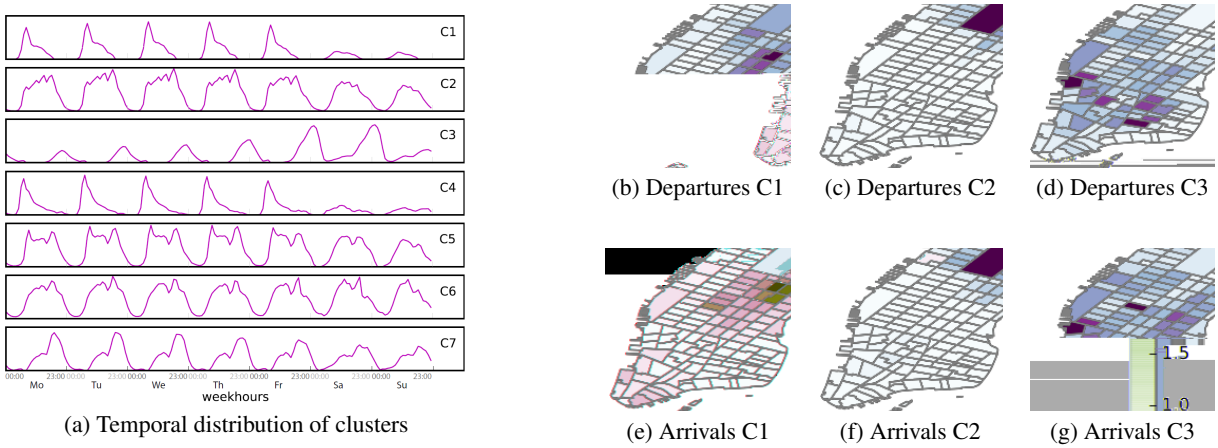
5.2 NTF: Mobility patterns

Since our data tensor has three dimensions, the decomposition returned three different components which determine the scale of mobility flow in each dimension for every cluster, i.e., time (hour of week), departure track, and arrival tract. Thus, individual clusters represent groups of taxi rides in different places at different periods of time in a weekday-hour scale.

The time component is shown in Fig. 1a, whereas location components are shown on the right-hand side of Fig. 1. Due to space limitations we show spatial components only for the first three clusters. From Fig. 1a, it can be observed that all clusters show strong daily regularities, which can be assumed as daily routines in human mobility. Clusters $C1$, $C2$, $C4$, $C5$ and $C7$ capture all behaviors on workdays (almost all peaks are between Monday and Friday) whereas cluster $C3$ is strongly dominated by weekend nights. Cluster $C6$ on the other hand, shows a more periodical behavior across the entire week, however its peaks are around $6pm$ from Monday to Saturday and $2pm$ on Saturdays and Sundays.

The location components shown in figures (1b-1g), together with their respective time periods, provide us initial context about *when* and *where* people move within the city. For instance, cluster $C1$ represents taxi rides around $9am$ (see Fig. 1a) which go to the south-east of Manhattan (see Fig. 1e). Cluster $C2$ concentrates its taxis rides around $6pm$ near Central Park (see Fig. 1f). Finally, cluster $C3$ includes all taxi rides on Fridays and Saturdays around $1am$ in the lower and middle part of Manhattan (see Fig. 1g).

The following section presents characterizations for such behaviors by comparing a list of 70 hypotheses using HypTrails.



(a) Temporal distribution of clusters

(b) Departures C1 (c) Departures C2 (d) Departures C3

(e) Arrivals C1 (f) Arrivals C2 (g) Arrivals C3

Figure 1: Spatio-temporal patterns. This figure illustrates the results obtained when applying NTF on a three-way tensor. (a) Each row represents a behavioral component (cluster) respect to pickup time (hour of week). Maps shown at the right of this figure depicts representative tracts for three clusters at departure time (top) and arrival time (bottom); the darker, the more dominant. (b,e) cluster $C1$ stands for taxis rides at $9am$ around the south-east of Manhattan; (c,f) cluster $C2$ contains all short taxi rides at $6pm$ around Central Park; and (d,g) cluster $C3$ represents all weekend night taxi rides around the lower and middle part of Manhattan.

5.3 HypTrails: Ranking of hypotheses

Since NTF does not explicitly partition the transition input, we first need to identify all transitions for each cluster in order to run HypTrails. Generally, interpreting clusters returned by NTF requires to extract the top-N weights from each factor to determine the most influential instances of every component [26]. In our setting, we extracted the top-10 pickup weekday-hours and drop-off tracts, and query all taxi rides fulfilling these conditions. Note that we did not include the top pickup tracts, because we are interested on places where people go to, rather than places where they come from. We then applied HypTrails and computed the rankings for the weighting factor $k = 10$—see Section 6 for a discussion.

Exemplary results of the characterization step for a hand selected subset of hypotheses are displayed in Table 2. It shows for each hypothesis (i.e., rows) the respective rank in the cluster (i.e., columns); lower numbers imply a higher rank and therefore a better explanation. Thus, for instance, the hypothesis *Gravitational (% White people)* expresses a belief in people going to *nearby* tracts with a *high % of white people* living there. The column *Overall* shows the ranking of hypotheses evaluated over the whole dataset, to compare with the results obtained for the individual clusters. The uniform hypothesis is a baseline which allows us to verify whether a hypothesis can be a good explanation of human mobility or not. Green cells in the table indicate that a hypothesis performed better than the uniform hypothesis in that cluster.

To characterize the different patterns in human mobility, we inspect the obtained rankings for each cluster. In particular, we are interested in which hypotheses perform exceptionally well (i.e., have a top rank indicated by a small number) in the cluster and in comparison to the overall dataset. For example, consider cluster $C3$ that captures taxi rides at weekend nights. For this cluster, we can observe very high ranks for the gravitational hypotheses *Party*, *Popularity* and *Food*.

In summary, clusters $C1$,$C2$ and $C3$ (shown in Fig. 1) can be characterized as follows. Cluster $C1$ predominantly represents taxi rides around $9am$ on workdays. People in this cluster prefer to go to nearby tracts containing popular places such as work places and restaurants near Times Square in a radius of $0.5km$. Cluster $C2$ groups taxi rides on workdays around $6pm$ going to big tracts containing art galleries, museums and parks. Transitions in this group usually leave tracts with few venues around. Finally, cluster $C3$ identifies all taxi rides on weekends around $1am$. People in this cluster usually go to nearby tracts containing very popular places such as nightlife spots and restaurants. Below, we discuss the characteristic properties of all clusters summarized by the types of hypotheses.

Distance-based. As mentioned in Section 4, these hypotheses require the standard deviation (σ) of a two dimensional Gaussian distribution. In Table 2, we show the best result for parametrized hypotheses and their respective σ in parenthesis. In the overall data as well as clusters $C2$ and $C3$, taxi rides are more likely to visit proximate places in a radius of $3km$, $0.5km$ and $1km$ respectively. Clusters $C1$, $C4$, $C6$ and $C7$ show preference on visiting the surroundings of Times Square in a radius of $0.5km$, $0.5km$, $0.01Km$ and $0.01km$ respectively. Finally, taxi rides in cluster $C5$ tend to visit places near the Flatiron building in a radius of $0.5km$.

Foursquare. Taxi rides in the overall dataset are more likely to visit nearby dense areas containing work places, restaurants and discos. Similarly, preferred targets in clusters $C1$ $C4$ and $C5$, which capture morning rides around $7-9am$, contain work places. Taxi rides in cluster $C6$ go to tracts dominated by parks around $6pm$. Cluster $C3$, which features weekend night trips, can be best characterized by the party hypothesis which means that taxi rides tend to visit tracts containing nightlife spots. Likewise, in cluster $C7$ people tend to visit tracts containing popular places such as restaurants and nightlife spots. Note that all hypotheses in this group perform better than the uniform hypothesis, demonstrating the overall high explanatory power of such data sources with respect to human mobility.

Census. From the overall data, we can infer that people tend to visit nearby tracts with high % of white people living in them. We can also observe that in general taxi rides are not going to residential areas but to commercial zones, which is also the case of clusters $C1$ $C3$ and $C7$, opposite to clusters $C2$ $C4$ and $C5$ where it is more likely to visit tracts containing art galleries and museums. In cluster $C6$ we can deduce that people tend to visit tracts containing parks rather than residential zones.

Table 2: Ranking of Hypotheses. This table shows the ranking of 21 out of 70 hypotheses evaluated with HypTrails over 3 different groups. *Overall* represents all 143*M* taxi rides in Manhattan, clusters C_i are identified by NTF. Numeric cells represent the ranks of the hypotheses in respective clusters. For the distance-based hypotheses, we only show results for the best parameter of the standard deviation σ (parameter in parentheses). Green cells highlight all hypotheses that outperform the uniform hypothesis.

HYPOTHESES	Overall *2013*	C1 *Workdays 9am*	C2 *Workdays 6pm*	C3 *Weekends 1am*	C4 *Workdays 7am*	C5 *Workdays 9am, 6pm*	C6 *Mo-Sa 6pm Sa-Su 2pm*	C7 *Workdays 6pm*
Baseline								
Uniform	42	56	56	56	56	55	62	59
Distance-based (σ)								
Proximity	14 (3.0)	7 (1.0)	2 (0.5)	14 (1.0)	10 (1.0)	19 (1.0)	10 (0.01)	13 (1.0)
Centroid (Geographical Center)	38 (5.0)	50 (5.0)	25 (1.0)	58 (5.0)	52 (5.0)	58 (5.0)	51 (3.0)	51 (5.0)
Centroid (Flatiron Building)	29 (5.0)	32 (2.0)	51 (5.0)	17 (1.0)	2 (0.01)	4 (0.5)	44 (3.0)	20 (0.5)
Centroid (Times Square)	22 (3.0)	1 (0.5)	43 (3.0)	46 (3.0)	1 (0.5)	43 (2.0)	2 (0.01)	1 (0.01)
Foursquare								
Gravitational (All venues)	1	12	14	10	14	10	14	9
Gravitational (Check-ins)	9	3	30	2	11	5	4	3
Gravitational (Work)	2	5	12	24	8	6	13	11
Gravitational (Food)	5	4	31	4	12	15	11	4
Gravitational (Party)	7	17	37	1	19	9	20	5
Gravitational (Recreation)	15	21	10	9	17	13	7	33
Venue Similarity	39	53	53	53	53	52	58	53
Census								
Gravitational (Population)	21	61	28	20	59	46	42	25
Gravitational (Tract Area)	23	34	8	26	24	20	24	38
Gravitational (%White people)	6	24	13	28	28	27	35	27
Gravitational (Residential zoning)	50	65	19	35	65	61	67	49
Gravitational (Commercial zoning)	13	8	32	22	9	24	19	15
Gravitational (Art Galleries)	46	23	1	38	5	2	52	54
Gravitational (Museums)	54	13	3	40	6	7	26	58
Gravitational (Parks)	63	62	4	44	63	59	6	64
Race Similarity	32	48	50	52	50	50	59	50

6. DISCUSSION

We have shown that spatio-temporal patterns in human behavior can be characterized by considering parts of data separately. We identified clusters of taxi rides and utilized openly available Web data to characterize them. However, there are some aspects that need to be taken into account for the current approach.

Concentration parameter k. HypTrails requires a parameter k to elicit Dirichlet priors from hypotheses [23]. Higher values of k express stronger beliefs in respective hypotheses. Technically, larger values of k imply higher values of the hyperparameters (pseudo counts) of the Dirichlet distributions. In our experiments, we tried several values of k from 0 to 100; overall very similar results. The reported results in this paper use an intermediate value of $k = 10$.

Correlations in characterizations. Using HypTrails to characterize clusters cannot identify *causes* of movement patterns, but only *correlations*. As an example, for cluster $C2$ the *Art Galleries* hypothesis performs best. This does not mean that taxi rides in that cluster prominently have art galleries as destinations, but that people go to nearby places containing art galleries. Thus, we also intend to integrate correlations between the used hypotheses in future work.

Clustering method. In this paper, we used HypTrails to characterize clusters obtained by NTF. While NTF is a reliable and established method in this line of research, the clustering approach is exchangeable and could be replaced by any other clustering technique.

State space. As our approach requires a discrete state space, we aggregate pick-up and drop-off locations of taxi rides in an area. The choice of aggregated units (i.e., states) can potentially influence the results which is known as the *Modifiable areal unit problem* [19]. In this paper, we chose tracts for the level of our analysis as it allowed for the direct integration of information from census data.

Multiple dimensions. In this paper, we clustered taxi rides with respect to time and space. However, our approach allows to extend these by additional information, e.g., # of passengers. In this case, a higher-way tensor would be used by NTF, but the resulting clusters could also be characterized with HypTrails. The scale of these dimensions can also suggest more fine-grained patterns. For instance, in this work we defined the time dimension as all 168 hours of a week in order to distinguish patterns on workdays and weekends.

7. RELATED WORK

Human mobility research. Human mobility is a phenomenon that has attracted the attention of governments and researchers from different fields. Studying the movement of people from a social science perspective has helped us to understand who, where and why people move [2, 16, 25, 28] as well as what consequences such movement carries, by means of e.g., demographic, socio-economic and land-use factors. In literature, they are also referred to as activity-based analysis. Natural sciences, on the other hand, have shown us that *universal patterns* exist and are modelled by movement-based techniques which can predict human dynamics [17, 22].

Ubiquitous data. Due to the lack of open and updated information at global scale (e.g., surveys or census data), and thanks to the rise of ubiquitous technologies such as mobile phone data and GPS, researchers can get access to human trails facilitating the study of human mobility. Several studies have revealed spatio-temporal patterns in different cities based on mobile phone call detail records [12, 27], taxi trips [3, 5, 15] and bike rides [14, 21]. The rapid emerge of social networks has also benefited the study of human mobility based on geo-tagged data. For instance, the work by Jurdak et al. [13] studies Twitter as a proxy of human movement by using universal

indicators such as *displacement distribution* and *gyration radius distribution* measuring how far individuals typically move based on geo-located tweets. Similarly, the authors in [18] proposed a network of places built upon Foursquare's venues and model human mobility by considering temporal and network dynamics inferred from user's check-ins. Gabrielli et al. [6] proposed a technique to analyze human trajectories of residents and tourists by semantically labeling source and destination spots. Based on time-evolving networks, the work in [7] identifies and ranks collective features for epidemic spread, by tracking human movements with wearable sensors.

Activity-based human behavior. Some previous works have identified and explained periodical movement-based patterns by activity-based human behavior. In [29], the authors proposed a model representing transition probabilities of travel demands during a time interval and suggested that travel demands can be associated with fixed locations under some circumstances. Jiang et al. [11] explained when, where and how individuals interact with places in metropolitan areas based on activity survey data in Chicago. The work shows daily patterns as eigenvectors and employs K-means clustering to identify groups of individuals based on their daily activities on weekdays and weekends. From taxi trips in Shanghai, the work in [20] shows how to detect basis patterns for collective traffic flow and correlates them with trip categories and temporal activities such as commuting to/from work in the mornings and evenings. Linear combinations are used to describe macro patterns and non-negative matrix factorization for detecting how many different patterns exist in a day.

Differentiation of our work. The novelty of our approach relies on: (1) a multidimensional pattern recognition process using NTF [4] to identify different mobility behaviors in taxi data, (2) the expansion of activity-based human mobility behavior into a hypothesis-based schema built upon human beliefs and (3) quantifying the plausibility of beliefs for mobility behavior using HypTrails [23].

8. CONCLUSIONS

In this paper, we have presented an innovative approach for discovering and characterizing patterns in human mobility behavior. It (i) clusters transition data using non-negative tensor factorization (NTF) and (ii) characterizes these clusters using the Bayesian HypTrails method. Our experiments on taxi data from Manhattan identified several patterns of human mobility and characterized them using Foursquare and census data. As one example, we discovered a group of taxi rides that end at locations with a high density of party venues on weekend nights. The strength of this approach relies on the fact that the interpretation of the clustering results can be easily characterized with high level hypotheses using HypTrails.

Our work extends recent research concerned with a better understanding of human mobility. We have demonstrated that human mobility is not one-dimensional but rather contains different facets including (but not limited to) time and space. Future research can benefit from our methodological and experimental concepts presented in this work. A more fine-grained view on human mobility can also facilitate e.g., city planners, traffic control and location-based recommender systems. In the future, we aim to generalize our findings by studying similar data (e.g., bike trips or geo-tagged tweets) available for New York and other cities. In doing so, we could not only unveil novel general patterns of mobility, but also discover similarities and differences between cities.

References

[1] M. Becker, P. Singer, F. Lemmerich, A. Hotho, D. Helic, and M. Strohmaier. Photowalking the city: Comparing hypotheses about urban photo trails on Flickr. In *Int. Conference on Social Informatics*, 2015.

[2] C. Brettell and J. Hollifield. *Migration Theory: Talking Across Disciplines*. Taylor & Francis, 2014.

[3] G. Chen, X. Jin, and J. Yang. Study on spatial and temporal mobility pattern of urban taxi services. In *Int. Conference On Intelligent Systems and Knowledge Engineering*, 2010.

[4] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[5] L. Ding, H. Fan, and L. Meng. Understanding Taxi Driving Behaviors from Movement Data. In *AGILE 2015*, pages 219–234. Springer, 2015.

[6] L. Gabrielli, S. Rinzivillo, F. Ronzano, and D. Villatoro. From tweets to semantic trajectories: mining anomalous urban mobility patterns. In *Citizen in Sensor Networks*, pages 26–35. Springer, 2014.

[7] L. Gauvin, A. Panisson, A. Barrat, and C. Cattuto. Revealing latent factors of temporal networks for mesoscale intervention in epidemic spread. *arXiv:1501.02758*, 2015.

[8] L. Gauvin, A. Panisson, and C. Cattuto. Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *PloS One*, 9(1), 2014.

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

[10] F. Ivis. Calculating geographic distance: concepts and methods. In *Proceedings of the 19th Conference of Northeast SAS User Group*, 2006.

[11] S. Jiang, J. Ferreira, and M. C. González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25(3):478–510, 2012.

[12] S. Jiang, J. Ferreira Jr, and M. C. González. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. In *Int. Workshop on Urban Computing*.

[13] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. Understanding Human Mobility from Twitter. *arXiv:1412.2154*, 2014.

[14] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs. Bicycle cycles and mobility patterns-Exploring and characterizing data from a community bicycle program. *arXiv:0810.4187*, 2008.

[15] X. Liu, L. Gong, Y. Gong, and Y. Liu. Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography*, 43:78–90, 2015.

[16] P. Merriman. *Mobility, Space, and Culture*. Routledge, 2012.

[17] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS One*, 7(5):e37027, 2012.

[18] A. Noulas, B. Shaw, R. Lambiotte, and C. Mascolo. Topological Properties and Temporal Dynamics of Place Networks in Urban Environments. In *Int. Conference on World Wide Web Companion*, 2015.

[19] S. Openshaw. *The modifiable areal unit problem*. Geo Books Norwich, UK, 1983.

[20] C. Peng, X. Jin, K.-C. Wong, M. Shi, and P. Liò. Collective human mobility pattern from taxi trips in urban area. *PloS One*, 7(4):e34487, 2012.

[21] A. Sarkar, N. Lathia, and C. Mascolo. Comparing citiesâĂŹ cycling patterns using online shared bicycle maps. *Transportation*, 42(4):1–19, 2015.

[22] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[23] P. Singer, D. Helic, A. Hotho, and M. Strohmaier. HypTrails: A Bayesian Approach for Comparing Hypotheses About Human Trails on the Web. In *Int. Conference on World Wide Web*, 2015.

[24] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[25] S. A. Stouffer. Intervening opportunities: a theory relating mobility and distance. *American Sociological Review*, 5(6):845–867, 1940.

[26] K. Takeuchi, R. Tomioka, K. Ishiguro, A. Kimura, and H. Sawada. Non-negative multiple tensor factorization. In *Int. Conference on Data Mining*, 2013.

[27] J. L. Toole, M. Ulm, M. C. González, and D. Bauer. Inferring land use from mobile phone activity. In *Int. Workshop on Urban Computing*, 2012.

[28] K. Willis. Introduction: mobility, migration and development. *Int. Development Planning Review*, 32(3-4):i–xiv, 2010.

[29] L. Wu, Y. Zhi, Z. Sui, and Y. Liu. Intra-urban human mobility and activity transition: evidence from social media check-in data. *PloS one*, 9(5):e97010, 2014.

[30] G. K. Zipf. The P1 P2/D hypothesis: On the intercity movement of persons. *American Sociological Review*, 11(6):677–686, 1946.