The TimeGeo modeling framework for urban motility without travel surveys

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.
The TimeGeo modeling framework for urban motility without travel surveys

Shan Jiang1,2, Yingxiang Yang1, Siddharth Gupta3, Daniele Veneziano4, Shounak Athavale5, and Marta C. González6,7,8,9

1Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; 2Research & Innovation Center, Ford Motor Company, Palo Alto, CA 93040; and 3Center for Advanced Urbanism, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved July 13, 2016 (received for review December 9, 2015)

Well-established fine-scale urban mobility models today depend on detailed but cumbersome and expensive travel surveys for their calibration. Not much is known, however, about the set of mechanisms needed to generate complete mobility profiles if only using passive datasets with mostly sparse traces of individuals. In this study, we present a mechanistic modeling framework (TimeGeo) that effectively generates urban mobility patterns with resolution of 10 min and hundreds of meters. It ties together the inference of home and work activity locations from data, with the modeling of flexible activities (e.g., other) in space and time. The temporal choices are captured by only three features: the weekly home-based tour number, the dwell rate, and the burst rate. These combined generate for each individual: (i) stay duration of activities, (ii) number of visited locations per day, and (iii) daily mobility networks. These parameters capture how an individual deviates from the circadian rhythm of the population, and generate the wide spectrum of empirically observed mobility behaviors. The spatial choices of visited locations are modeled by a rank-based exploration and preferential return (r-EPR) mechanism that incorporates space in the EPR model. Finally, we show that a hierarchical multiplicative cascade method can measure the interaction between land use and generation of trips. In this way, urban structure is directly related to the observed distance of travels. This framework allows us to fully embrace the massive amount of individual data generated by information and communication technologies (ICTs) worldwide to comprehensively model urban mobility without travel surveys.

human mobility | urban model | mobile phone data | networks | urban planning

Our ability to correctly model urban daily activities for traffic control, energy consumption, and urban planning (1, 2) has critical impacts on people’s quality of life and the everyday functioning of our cities. To inform policy making of important projects such as planning a new metro line and managing the traffic demand during big events, or to prepare for emergencies, we need reliable models of urban travel demand. These are models with high resolution that simulate individual mobility for an entire region (3, 4). Traditionally, inputs for such models are based on census and household travel surveys. These surveys collect information about individuals (socioeconomic, demographic, etc.), their household (size, structure, relationships), and their journeys on a given day. Nonetheless, the high costs of gathering the surveys put severe limits on their sample sizes and frequencies. In most cases, they capture only 1% of the urban household population once in a decade with information of only one or few days per individual. The low sampling rate has made it very costly to infer choices of the entire urban population (3, 5–7).

More recent studies try to learn about human behavior in cities by using data collected from location-aware technologies, instead of manual surveys, to infer the preferences in travel decisions that are needed to calibrate existing choice modeling frameworks (8–10). The problem, however, is that the geotagged data available from communication technologies, in the massive and low-cost form, cannot inform us about the detailed activity choices of their users, making most of the data useless for meaningful urban-scale mobility models. To make the best use of the massive and passive data, a fundamental paradigm shift is needed to model urban mobility and enhance new opportunities emerging through urban computing (11). This is our goal with TimeGeo, a modeling framework that extracts individual features and key mechanisms needed to effectively generate complete urban mobility profiles from the sparse and incomplete information available in telecommunication activities.

Mobile phones are the prevalent communication tools of the 21st century, with the worldwide coverage up to 96% of the population (12). The call detailed records (CDRs), managed by mobile phone service providers for billing purposes, contain information in the form of geolocated traces of users across the globe. Mobile phone data have been useful so far to improve our knowledge on human mobility at an unprecedented scale, informing us about the frequency and the number of visited locations over long-term observations (13–18), daily mobility networks of individuals (15, 19), and the distribution of trip distances (13, 15, 17, 20–22). Due to the sparse nature of mobile phone use, these data sources have sampling biases and do not provide complete journeys in space and time for each individual (9). Nonetheless, it has been possible to extract and characterize from phone data where each individual may stay or pass by, and then infer the types of activities that they engage in at various urban locations depending on the time of their visits (23). By labeling visited location types for individual users as home, work, or other, representative traffic origin–destination (OD) matrices for an average day and by time of day can be generated (24, 25). They are aggregated estimates of person-trips

Significance

Individual mobility models are important in a wide range of application areas. Current mainstream urban mobility models require sociodemographic information from costly manual surveys, which are in small sample sizes and updated in low frequency. In this study, we propose an individual mobility modeling framework, TimeGeo, that extracts required features from ubiquitous, passive, and sparse digital traces in the information and communication technology era. The model is able to generate individual trajectories in high spatial–temporal resolutions, with interpretable mechanisms and parameters capturing heterogeneous individual travel choices. The modeling framework can flexibly adapt to input data with different resolutions, and be further extended for various modeling purposes.
between pairs of ODs within few hours, and these results have been successfully validated in various cities against existing travel demand models that required expensive surveys for calibration (24, 25).

A fundamental question still remains on how to perform a spatiotemporal mapping of raw mobile phone data to establish models of travel demand with high spatiotemporal resolution, through which individuals’ disaggregated daily journeys can be generated. In the current literature that analyzes sparse geotagged data, the daily temporal behavior of human mobility is either not modeled or oversimplified (13, 16). For example, previous studies on human dynamics do not explicitly model individual temporal choices, but randomly draw parameters such as waiting time or the number of activities in each active period from aggregated distributions measured from data (14, 15). The model in ref. 19 introduces time dependency in travel and tendency to arrange short out-of-home activities in consecutive sequences (i.e., bursts of activities) (26–30), but the stay duration at flexible (other) locations is fixed. Furthermore, it does not incorporate spatial choices or the heterogeneity of individual behavior.

To realistically model individual mobility in cities at both micro- and macrolevel, it is necessary to understand the essential features of a population distribution in space at different times. Here we show that these features can be extracted from big data sources. Instead of using social-demographic information to calibrate the set of detailed decisions involved in activity choices—as required by mainstream transportation modeling approaches—the framework consists of directly measurable parameters discovered from passive data. It represents a needed paradigm shift to model individual daily trajectories in cities, adapted to ubiquitously available sparse digital traces of individuals. The results are high-resolution travel diaries for a large sample of users based on their information and communication technology (ICT) data in the urban context. The presented set of parameters can be further refined as more information becomes available at the individual level.

**Activity Extraction from Mobile Phone Data**

To demonstrate the mechanistic modeling framework, we analyze a CDR data set of 1.92 million anonymous mobile phone users for a period of 6 wk in 2010 in the Greater Boston area. To have a control experiment, we also examine a donated set of self-collected mobile phone traces of a graduate student in the same region over a course of 14 mo in 2013 and 2014, recorded by a

---

**Fig. 1.** Extraction of stays and daily journeys from raw cell phone data. (A–C) Stay locations extracted from the self-collected cell phone records of a student in three sample days. (D–F) Illustration of trips between consecutive stays in each day. (G–I) Visitation frequency of all locations, counting from the first day of the observation period to the current day. For this individual, home and work stays dominate all visits. Highlighted arrows mark the trips on that day. The time bar above each subfigure is color-coded by activity type based on each stay’s duration. (J–L) Illustration of the rank-based EPR model. To illustrate different cases we use the individual’s home, work, and one other location as trip origins. The potential trip destinations are color-coded by different chosen probabilities based on their rank. The closer a location is to the origin, the higher the probability it has to be chosen. The height of the dots represents the density of destinations in the surrounding region. The most dense place for other type of activities is in downtown Boston.
smartphone application. When an individual anchors at a location to conduct an activity, it is defined as a stay. We apply the stay extraction method discussed in the literature (23) to both data sets. We filter out signal jumps as well as pass-by records when mobile phone users were traveling. For each user, based on the start time and frequency of visits to each stay location, we infer the stay location type as home (H), work (W), or other (O).

We are able to identify home locations for 1.44 million users, which is 75% of our initial user base. Next, we filter users who have more than 50 total stays and at least 10 home stays in the observation period. These are identified as active users and are used to extract the various parameters of TimeGeo (as explained in detail in the next sections). These active users can be labeled as commuters (133,448 individuals) who have journey-to-work trips, and noncommuters (43,606 individuals) who have no journey-to-work trips.

Fig. 1 illustrates the pipeline of extracting stays, labeling activity types, and deriving individual mobility features from raw mobile phone data for each of three demonstrated days. Fig. 1 A–C shows the raw cell phone records (in blue for 14 mo, and in purple for each day), and the extracted stay locations of the individual (in red). Fig. 1 D–F shows that for active users the extracted stays in each day define a daily journey (usually starting and ending at home). A trip is made when a user changes stay locations. The time bar shows the start time and duration for each stay, and activity types are color-coded.

Generating Mechanisms of Individual Mobility

The modeling framework of TimeGeo is presented in Fig. 2. It integrates the temporal and spatial choice mechanisms of human mobility. We assume that for an individual agent, her work activity has a fixed location, start time, and duration; her home activity is fixed in terms of location but flexible with start time and duration; her other activity is flexible with regard to location, start time, and duration. The presented framework aims to model the flexible spatial and temporal mobility choices, whereas the schedule of the fixed activity (i.e., work) is assumed as predetermined (see SI Appendix, section 2 for details). We divide each day of a week into 144 discrete intervals of 10 min (i.e., 1,008 time intervals in a week). For each time interval \( t \) within a week, an individual first decides to stay or move. If she chooses to move, she then decides where to go. We improve from previous human mobility models (14, 19) by generating spatiotemporal patterns while introducing individual-specific mobility parameters, namely: a weekly home-based tour number, a dwell rate, and a burst rate (explicitly defined later). These parameters capture the heterogeneity of individual daily mobility observed in the passive digital traces. Nevertheless, due to the limited observation period of the CDR data used in this study, some parameters cannot be extracted at the individual level. These global parameters measure the preferential return and exploration rates, and the rank selection probability. As large-scale data with higher frequency (e.g., GPS traces) and longer observation periods (e.g., many months) become available, these global parameters could be measured at the individual level as well.

Temporal Choices. To uncover the key generating mechanisms needed to reproduce individual daily trajectories, we propose a time-inhomogeneous Markov chain model with three individual-specific parameters—weekly home-based tour number \( (n_w) \), dwell rate \( (\beta_1) \), and burst rate \( (\beta_2) \)—to capture individual circadian propensity to travel (16, 19, 31) and likelihood of arranging short activities in consecutive sequences (26–30). As work activity is assumed to have fixed start time and duration, we consider two Markov states: home and other. Home is considered as a less-active state, because the average stay duration at home is significantly longer than that at other states where people are more active (i.e., likely to travel).

When an individual \( l \) is at home, her individual travel circadian rhythm is defined as \( n_w P(t) \), representing her likelihood of making a trip originated from home in a time-interval \( t \) of a week. The weekly home-based tour number \( n_w \) counts the total number of trips that an individual \( l \) initiated from home to other places. \( P(t) \) is the global travel circadian rhythm of the population in an average week. We differentiate \( P(t) \) for commuters and noncommuters (SI Appendix, section 3.1). For noncommuters, \( P(t) \) is measured as the fraction of all user-trips in the time interval \( t \) of the week for the population (i.e., \( \sum_{t=1}^{1,008} P(t) = 1, t = 1,2, \ldots, 1,008 \)), capturing the

---

**Fig. 2.** Flowchart of TimeGeo and input features extracted from active CDR users. (A) Spatial and temporal choices per time step. Three individual specific parameters control temporal patterns, including the weekly home-based tours \( (n_w) \), dwell rate \( (\beta_1) \), and the burst rate \( (\beta_2) \). \( n_w \) influences the travel likelihood when a person is at home, \( \beta_1 n_w \) influences the travel likelihood when a person is out of home, whereas \( \beta_2 n_w \) influences the likelihood of performing consecutive out-of-home activities. (B) \( P(t) \) shown here is the empirical travel circadian rhythm in an average week measured from data for active noncommuters (who have no journey-to-work trips). (C) Joint distribution of \( \beta_1 n_w, \beta_1 n_w, \text{and } n_w \) for active noncommuters in the CDR data set. The 2D marginal distributions are shown by the contour plots. The green dot is the most probable parameter value combination with \( n_w = 6, 1, \beta_1 n_w = 22.4, \beta_2 n_w = 508.0 \). (D) Empirical probability to visit a new location \( P_{new} \) as a function of distinct visited locations \( S \); it follows \( P_{new} = 0.65^{-0.21} \). (E) Empirical probability of choosing the rank \( k \) location as a trip destination follows \( P(k) = k^{-0.86} \).
expected variation of travel in different time of the week (shown in Fig. 2B). For commuters, because work is modeled as a fixed activity, the probability of return (r-EPR) model by incorporating a rank-based selection of newly explored locations (i.e., r-EPR model).

Our selection mechanism gives a rank $k$ to each alternative destination based on the empirical distribution of travel across trips (33–36). Among all potential new destinations, the one closest to the current location is of $k = 1$, the second closest is $k = 2$, etc. The empirical probability of selecting the $k$th location as a destination is quantified as $P(k) \sim k^{-α}$; the same form has been measured in various studies that analyze aggregated trips between locations for both commuting and noncommuting trips (33–36). For an individual to select an exploration destination, we measure $P(k)$ aggregating all users’ destinations. Fig. 1 J–L illustrates probabilities of selecting different destinations (with higher ranks in red and lower ranks in blue). Each dot represents a location for an activity outing that is extracted from the CDR data. The height of the dot on the z axis represents the density at the location.

Because the observation period of the empirical data in this study is 6 wk, most users have a limited number of exploration trips, making it difficult to estimate the spatial parameters of $P(k)$ at the individual level. Given more abundant data, this distribution could be estimated at the individual level as well.

Role of Land Use on Travel Distance

Different spatial patterns of cities imply different geographical advantages to urban functioning (37). TimeGeo takes the spatial distribution of locations (e.g., observed from the CDR data) as input. To explain and quantify the influence of land use on travel, we propose a hierarchical multiplicative cascade framework of analysis. It allows scenario tests on how changes in land-use patterns will affect individual travel. It can generate different scenarios of urban structure (i.e., spatial distribution of home and other activities).

Fig. 3 A–D shows the distribution of different types of locations (home and other) extracted from the mobile phone data set at two scales: At a scale with larger grids, home and other locations are mixed spatially, showing high spatial correlations. At a scale with smaller grids, the separation between home and other types of land use becomes clear (35). The intuition behind this phenomenon is that at a scale with smaller grids (e.g., similar to the census block level), land use is often separated—meaning that residential land use is separated from nonresidential one, whereas at a scale with larger grids (e.g., at the district, town, or regional level), residential and nonresidential land uses mix together. A hierarchical multiplicative cascade divides an area of interest into grids with different granularity and quantifies the spatial correlation of each type of land use at different scales.

The current framework integrates the two features that influence the spatial choices of exploration to other locations. These are (i) the spatial distribution of activity locations, and (ii) the rank-based location-selection mechanism (illustrated in Fig. 1 J–L). By characterizing the spatial distributions of population and facilities at various scales, here we formalize how these two features influence the observed trip–distance distribution.

To quantitatively represent home to other ($H-O$) trip distance, we denote home locations as the demand side $D$, and other locations as the supply side $S$. The entire region of interest is $Ω_0$ (taken as a unit square, shown in Fig. 3E). We progressively partition $Ω_0$ into $4^2$, $4^3$, ..., $4^8$ square tiles with side length $2^{-1}, 2^{-2}, \ldots, 2^{-8}$. Each time a mother tile $Ω_i$ at (resolution level $i-1$) is partitioned into four daughter tiles $Ω_j$ (at resolution level $i$).
i). Then, the probability that a trip goes outside its origin tile at resolution level \( i \), \( P_x(i) \), can be expressed as

\[
P_x(i) = \int_0^M P_x(k) f_{S_{x\rightarrow y}}(k) dk,
\]

where \( M \) is the total number of supplies in the entire region \( \Omega_0 \); \( P_x(k) \) is the probability that the \( k \) supplies in the origin tile are not chosen; \( f_{S_{x\rightarrow y}}(k) \) is the probability of finding \( k \) supplies within the origin tile. The tile exceeding probability \( P_x(i) \) at different tile resolutions generates the resulting distribution of trip distances. Eq. 2 ties together the rank-based selection mechanism \( P_x(k) \) and the geographic distribution of locations \( f_{S_{x\rightarrow y}}(k) \), which can be calculated as

\[
f_{S_{x\rightarrow y}}(k) = \int_0^\Omega f_{D_{x\rightarrow y}}(D)f_{S_{x\rightarrow y}}(D=k) dD,
\]

where \( f_{D_{x\rightarrow y}}(D) \) is the conditional probability that a trip originates in a tile at level \( i \) given \( D \) demands are in that tile; \( f_{S_{x\rightarrow y}}(k) \) is the conditional probability of supply given demand. \( Q \) is the number of demand in the entire study area. In summary, to quantify trip distance through \( P_x(i) \), we not only need the distribution of each type (home and other) of location, but also the correlation between them at different scales. The detailed introduction to the cascade method of analysis can be found in ref. 38 and in Materials and Methods; the derivation of the resulting trip distance distribution is presented in SI Appendix, section 5.

**Results**

**Extracted Mobility Features from Mobile Phone Data.** In this section we show the results for noncommuters. For each individual, the weekly home-based tour number \( n_w \) is directly extracted from the data, whereas the \( \beta_1 \) and \( \beta_2 \) parameters are calibrated using the temporal Markov model. The rest of the parameters needed are \( \alpha = 0.86 \) for the rank selection probability \( P(k) \sim k^{-\alpha} \), and \( \rho = 0.6 \) and \( \gamma = 0.21 \) for the preferential return mechanism \( P_{x\rightarrow y} = \rho S^\gamma \). These three parameters are extracted from the aggregated data of the entire population (Fig. 2 D and E).

The individual values of \( \beta_1 \) and \( \beta_2 \) values are obtained by calibrating the Markov model to minimize the following statistic:

\[
A(\beta_1, \beta_2) = \int |P_D(\Delta t) - P_M(\Delta t; \beta_1, \beta_2)| d\Delta t + \eta |\bar{N}_D - \bar{N}_M(\beta_1, \beta_2)|,
\]

where \( P_D(\Delta t) \) and \( P_M(\Delta t; \beta_1, \beta_2) \) are the distributions of the individual empirical and modeled stay duration, respectively. Scalar values \( \bar{N}_D \) and \( \bar{N}_M(\beta_1, \beta_2) \) are the average daily number of visited locations measured from the individual’s empirical data and from the model simulation, respectively. The difference between \( \bar{N}_D \) and \( n_w \) is that \( \bar{N}_D \) counts all trips whereas \( n_w \) only counts trips starting at home. Metaparameter \( \eta = 0.035 \) controls the weight between the two components. Because \( A(\beta_1, \beta_2) \) is a nonconvex function, discrete \( \beta_1 \) and \( \beta_2 \) values are used (\( \beta_1 = 1, 2, 3, \ldots, 20 \) and \( \beta_2 = 1, 6, 11, \ldots, 101 \)) to estimate the \((\beta_1, \beta_2)\) pair that minimizes \( A(\beta_1, \beta_2) \) for each person. The empirical results of \( n_w, \beta_1, \beta_2 \), and \( n_w \) for all of the individuals are presented in Fig. 2C. The median values of \( n_w, n_w/\beta_1 \), and \( n_w/\beta_2 \) for noncommuters are 7.4, 34.2, and 355.6, respectively. Median dwell rate \( \beta_1 = 4.6 \), suggesting that when people are not at home, they are on average 4.6 times more likely to travel.

**Simulated Mobility Features.** Taking the featured parameters measured directly from active users of the mobile phone data set, TimeGeo can generate realistic individual daily trajectories over a long time period at the urban scale.
We first use the student volunteer’s 14-mo mobile phone records as an example to explain the simulation and interpret the results of TimeGeo. We fix the locations of home and work (in this case school is identified as work) and apply the proposed modeling framework to simulate the spatiotemporal choices of flexible other activities and temporal choices of home activities. For the student, we computed that his dwell rate $\beta_1 = 4$, burst rate $\beta_2 = 36$, and weekly home-based tour number $n_w = 7$. His burst rate is lower than the population average, reflecting smaller likelihood to conduct consecutive short activities. Fig. 4A–C shows three simulated days for the student. The days are predominantly by home–work trips, with a few trips to other locations. The model is able to capture not only the number of locations visited each day, but also more detailed configuration of daily trip chains. Fig. 4D shows the distribution of the most frequent daily mobility networks, i.e., daily motifs, of the student. We represent unique locations as nodes and trips between locations as edges and count the motif distribution for days start and end at home. The dominating motif is traveling just between two locations in a day. To show the infrequent motifs clearer, we present the percentage in log scale.

A key value of TimeGeo is to use ICT records to generate individual trajectories from discovered mobility features at the urban scale. In Fig. 4E–H, we illustrate a user with very sparse data. She only had four distinct locations in 30 d and we simulate her complete daily trajectories in space and time. We select two different sets of $\beta_1$, $\beta_2$, and $n_w$ from the joint distribution shown in Fig. 2C to generate two synthetic realizations of the user. Fig. 4F and G shows the two resulting profiles of simulated journeys of the same sparse user and Fig. 4H shows the distinct motif distributions.

The importance of the individual features extracted from data (Fig. 2C) lies in their ability to capture diverse travel behaviors observed in the population. Fig. 5A and B compares mobility patterns for different individual profiles. The individual 1 and 2 represent two extreme cases: one travels more frequently (shown in squares, $n_w = 10.86$, $\beta_1 = 6$, $\beta_2 = 41$) and the other travels less frequently (shown in circles, $n_w = 5.51$, $\beta_1 = 1$, $\beta_2 = 36$). As a comparison we also present the average case—a simulation using median values of the parameters $n_w$, $\beta_1$, and $\beta_2$. Fig. 5A and B shows that these three individuals have distinct $P(\Delta t)$ and $P(N)$ distributions. The less-frequent traveler has significantly longer stay duration and visits fewer locations per day. To quantify the differences between empirical distributions of data and the model simulation, we use the Kolmogorov–Smirnov (KS) test. The KS statistic between empirical and simulated $P(\Delta t)$ for the two extreme individuals is 0.12 and 0.11, respectively. If we compare their empirical data with the average case, the KS statistic increases to 0.25 and 0.20, respectively. Similarly, for these two individuals, the KS statistic for $P(N)$ is 0.05 and 0.12. When comparing with the average case, the KS statistic increases to 0.40 and 0.50, respectively. It confirms the importance of including individual-specific parameters to model temporal choices. With data of high frequency and longer observation period available in future studies, machine learning methods can be applied to better learn from choices at individual level when choosing return trips for improvement of our proposed modeling framework.

Fig. 5C–F compares aggregated mobility features extracted from data and simulation for all of the active noncommuters. These results show that to reproduce individual mobility patterns realistically, it is critical to incorporate each of the mechanisms.
proposed in the current modeling framework, namely, the weekly home-based tour number, dwell time, burst rate, and the rank-based EPR, over the land-use profile of the city under consideration. The results on the aggregated daily mobility motif distribution are presented in SI Appendix, section 4.2. For the dwell rate ($\beta_1$), if $\beta_1 = 1$, i.e., the model does not differentiate the mobility circadian rhythms of home or other activities. The resulting $P(\Delta t)$ distribution will underestimate trips with short duration, and the KS statistic increases from 0.04 to 0.27. For the $P(N)$ distribution, the KS statistic for the model with and without the burst rate $\beta_2$ is 0.03 and 0.22, respectively. The bursts of flexible activities, captured by the dwell and burst rates $\beta_1$ and $\beta_2$, ensure realistic distributions of the stay duration $P(\Delta t)$ and the number of daily visited locations $P(N)$. The improved rank-based EPR mechanism models the selection of locations. It improves the KS statistic of the trip distance distribution from 0.52 to 0.39. The visitation frequency to the $L$th most visited location follows $f(L) \sim L^{-1.2 \pm 0.1}$. In Fig. 3 D, $P_s(i)$ measures the probability that a generic exploration trip goes outside its origin tile at resolution level $i$. At the largest four tile sizes (24, 12, 6, and 5 km), the cascade is a pure log-normal cascade. $P_s(i)$ can be analytically calculated, and the result compares very well with the data. The empirical data, simulation, and analytical calculation all show that 10% of the trips cross the tile with a size of 24 km, and over 60% cross the tile with a size of 3 km. Taken together, we now use the extracted features from active mobile phone users with the presented modeling framework to estimate the daily mobility for the entire metropolitan area. To do so, we expand the users (commuters and noncommuters) to the population (aged 16 and over), and generate 1-weekday mobility trajectories using TimeGeo for the population (see SI Appendix, section 4.3 for more details). In the Fig. 5 (Bottom), we compare our simulated daily mobility patterns for the population in Metro Boston (3.54 million individuals aged 16 and over) with traditional travel survey data. $\text{NHTS}$ and the 2010 MTS in our simulation with the estimates of the model developed by the $\text{BRMPO}$ (40). The third case represents an average noncommuter and is simulated using the median parameter values of $\beta_1$, $\beta_2$, and $\beta_3$.

The burst rate $P(\Delta t)$ compares the fraction of trips being initiated at different times of the day among our simulation, the 2009 NHTS, and the 2010 MTS. The KS statistic of the trip distance distribution from 0.52 to 0.39. The benchmark shows the result without the preferential return mechanism. (f) Trip distance distribution $P(\Delta r)$ extracted from data, and simulation results using an $r$-EPR mechanism, compared with the random selection of exploration locations (not using the rank-based selection mechanism). The bottom panels (G–K) show the comparison of the simulated daily mobility patterns for the population (aged 16 and over, for both commuters and noncommuters) in Metro Boston (3.54 million individuals) with traditional travel survey data, including the 2009 NHTS, and the 2010–2011 MTS. (G) Stay duration distribution. (H) Daily visited location distribution. (I) Trip distance distribution. (J) Comparison of total commuting trips between and within the 164 cities and towns (i.e., inter- and intratown) estimated by our simulation and the model of the $\text{Boston Region MPO}$ (40). (K) Fraction of trip departures by time of the day, comparing the simulation, the 2009 NHTS, and the 2010 MTS. SI Appendix, Fig. 5.18 shows comparisons for various trip purposes.
than those reported in the surveys (see SI Appendix, section 4.3 for detailed comparisons). Overall, the results show good agreement with existing MPO models which needed expensive travel survey for their calibration.

Conclusion

We present a mechanistic modeling framework to generate individual daily mobility with fine resolution at urban scale. Temporally, we introduce the weekly observation of two mobile phone carriers. The student, who donated the hierarchical multiplicative cascade model is to have population and facilities in a given region is characterized using a hierarchical multiplicative cascade framework of analysis. In this way, we take account of the influence of region-specific spatial structure on individual travel distances. This enables us to perform scenario tests on how changing land use in the city would affect microlevel individual travel behavior and macrolevel OD flows.

TimeGeo serves as a general modeling framework of urban trajectories that can be flexibly adapted to different application scenarios using population density and the distributions of facilities in any city. It can be coupled with sparse location data from ICTs that sample the visitation preferences of actual individuals and can complement or, for some applications, substitute the need for expensive travel surveys for modeling urban travel. The framework is flexible to generate trajectories with various data conditions. The minimum requirement is to have population and facility distributions. In the current results, the parameters to model exploration and returns (α, β, and γ) are assumed to be the same across population, whereas the temporal mobility rates of an individual are assumed to be independent of the actual location. In future studies, more data of higher frequency and over longer periods become available, it is possible to further learn from the individual variations of the proposed parameters. It is also interesting to explore the variations of the model parameters across urban areas, and across population groups with different demographics and lifestyles.

Materials and Methods

All study procedures were carried out with Institutional Review Board approval from Massachusetts Institute of Technology (MIT) Committee on the Use of Humans as Experimental Subjects (COUHES) (Protocol 1405006399) approved on June 10, 2014. CDR data were collected by AirSage for operational purposes of two mobile phone carriers. The student, who donated his 14-mo self-collected mobile phone traces through a smartphone application (OpenPaths), provided informed consent for the research.

Mobile Phone Data. We extracted activity stay locations of 1.92 million cell phone users from their CDRs in the Greater Boston area during an observation period of 6 wk in 2010. A stay means performing an activity at a location. A stay sequence, or an activity sequence, represents consecutive stays a person made in a period (usually a day). A trip is made between consecutive stay locations. These stay locations are also called trip origins and destinations. In the CDR data, a record is made when a user calls, sends text messages, or uses data through the cellular networks. Each record is in the following format: (UserID, longitude, latitude, time). The precision of the location is about 200–300 m in urban areas. For the voluntarily self-collected mobile phone user example, a record is made every time the smartphone application detects a significant spatial movement. The data are in the same format and similar spatial resolution as the CDR data. The detailed methods to extract stay locations and to label location types (as home, work, and other) are presented in SI Appendix, section 1. For the CDR data, the records do not directly correspond to a user’s stays—a stay could not be detected if a user did not use his or her cell phone more than once during a stay. Even for cases when more than one cell phone use was recorded, the stay duration can only be approximated for active phone users. Therefore, not all cell phone users have enough records to be measured for basic mobility patterns presented in this study. Meanwhile, we cannot determine if long stays at one location (for over 2 d) are caused by no cell phone use or actual stay at one location for over 2 d; therefore, these stays were removed from the analysis and not captured by the model.

The Hierarchical Multiplicative Cascade Model. For any given subregion μ = CDR, D(ω) is the number of trip origins in ω and S(ω) is the number of trip destinations in ω. We use bivariate random measures \( X(\omega) = (D(\omega), S(\omega)) \) to represent the number of demand and supply locations in \( \omega \), where \( X \) results from a cascade process in which the fluctuations at different spatial scales combine in a multiplicative way. The generation of bivariate \( D, S \) cascades is illustrated in Fig. 3C. The demand and supply in a generic i-tile \( \Omega_i \) are \( D_i \) and \( S_i \), and the associated measure densities \( D_0 = D_i/|\Omega_i| \) and \( S_0 = S_i/|\Omega_i| \). One starts with uniform measure densities \( D_0 \) and \( S_0 \) in \( \Omega_0 \), then progressively partitions \( \Omega_0 \) into \( 4^1, 4^2, \ldots, 4^n \) square tiles of side length \( 2^{-1}, 2^{-2}, \ldots, 2^{-n} \). The demand and supply densities in the daughter tiles are multiplied by independent realizations of nonnegative random factors \( W_0 \) and \( W_S \) with mean value 1. The random vectors \( W_i = [W_{D,i}, W_{S,i}] \), \( i = 1, \ldots, n \) are the generators of the cascade. Although the generators \( W_i \) have independent values in different i-tiles, their components \( W_{D,i} \) and \( W_{S,i} \) in a given i-tile may be dependent. Moreover, the distribution of \( W_i \) may vary with the resolution level \( i \). These features provide important modeling flexibility. The measured densities at resolution level \( i \) and \( i + 1 \) are related as

\[
\begin{align*}
D_i & = D_{i+1}/|\Omega_i|, & S_i & = S_{i+1}/|\Omega_i|, \\
0 & \leq D_{i+1}, S_{i+1} \leq 1. 
\end{align*}
\]

According to Fig. 3 A–D, at larger tile sizes almost all tiles are nonempty and the supply and demand have positive correlation. Consequently for small i values (large tile sizes) the generator can be described as joint log-normal variables (38). If the log generators \( \ln W_{D,i} \) and \( \ln W_{S,i} \) have joint normal distribution with variances \( \sigma_{D,i} \) and \( \sigma_{S,i} \), mean values \(-1/2\sigma_{D,i}^2\) and \(-1/2\sigma_{S,i}^2\), and correlation coefficient \( \rho_{D,S,i} \), then \( \ln D_i \) and \( \ln S_i \) have joint normal distribution with mean values \( m_D \) and \( m_S \), variances \( \sigma_D^2 \) and \( \sigma_S^2 \), and correlation coefficient \( \rho_D \) given by

\[
\begin{align*}
\sigma_D^2 &= \sum_{j=1}^i \sigma_{D,j}^2 = \ln(D_04^{i-1}) - 1/2\sigma_D^2, \\
\rho_D &= \sum_{j=1}^i \rho_{D,j} W_{D,j} W_{S,j}. \end{align*}
\]