Transportation Research Part C 58 (2015) 240–250

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Origin–destination trips by purpose and time of day inferred from mobile phone data

Lauren Alexander [a,*], Shan Jiang [b], Mikel Murga [a], Marta C. González [a]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States
[b] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, United States

ARTICLE INFO

Article history:
Received 1 June 2014
Received in revised form 18 January 2015
Accepted 17 February 2015
Available online 9 March 2015

Keywords:
Mobile phone data
Data mining
Human mobility
Trip production and attraction
Trip distribution
Travel surveys

ABSTRACT

In this work, we present methods to estimate average daily origin–destination trips from triangulated mobile phone records of millions of anonymized users. These records are first converted into clustered locations at which users engage in activities for an observed duration. These locations are inferred to be *home*, *work*, or *other* depending on observation frequency, day of week, and time of day, and represent a user's origins and destinations. Since the arrival time and duration at these locations reflect the *observed* (based on phone usage) rather than *true* arrival time and duration of a user, we probabilistically infer departure time using survey data on trips in major US cities. Trips are then constructed for each user between two consecutive observations in a day. These trips are multiplied by expansion factors based on the population of a user's *home* Census Tract and divided by the number of days on which we observed the user, distilling average daily trips. Aggregating individuals' daily trips by Census Tract pair, hour of the day, and trip purpose results in trip matrices that form the basis for much of the analysis and modeling that inform transportation planning and investments. The applicability of the proposed methodology is supported by validation against the temporal and spatial distributions of trips reported in local and national surveys.

1. Introduction

The ubiquity of cell phones, along with rapid advancement in mobile technology, has made them increasingly effective sensors of our daily whereabouts (Lane et al., 2010). Call detail records (CDRs) from mobile phones contain time-stamped coordinates of anonymized customers, thereby providing rich spatiotemporal information about human mobility patterns. Since CDRs are automatically collected by cell phone carriers for billing purposes, this data can be gathered more frequently and economically than travel survey data collected once (or twice) a decade for transportation planning purposes. Additionally, mobile phone data offers digital footprints at a scale and resolution that may not be captured by surveys that typically record one day of travel diaries per household.

Despite these advantages, mobile phone data lacks information typically available from travel surveys about a respondent (e.g. age or income) or his/her trip (e.g. purpose or mode) (Richardson et al., 1995; Stopher and Greaves, 2007; Hu and Reuscher, 2004). Furthermore, CDRs contain traces of a user at approximated locations when his/her phone communicates

* Corresponding author.

http://dx.doi.org/10.1016/j.trc.2015.02.018
0968-090X/

with a cell phone tower, providing an inexact and incomplete picture of daily trip-making. Accordingly, much research has focused on developing methods to extract meaningful information about human mobility from mobile phone traces as well as understanding its limitations.

It has been demonstrated that CDR data can be used to infer origin–destination (OD) trips using microsimulation and limited traffic count data (Iqbal et al., 2014). At the level of the individual, daily trip chains/trajectories constructed from mobile phone data are consistent with household surveys (Jiang et al., 2013; Schneider et al., 2013). Further, road usage inferred from the CDR data has been validated against GPS speed data (Wang et al., 2012) and highway assignment results from a travel demand model (Huntsinger and Donnelly, 2014).

There is still work to be done to explore the usage of phone data to generate trip distributions of different modes, purposes, and times of day. As a step in that direction, this research proposes a methodology to extract OD trips by purpose and time of day from CDR data. This segmentation captures distinct trip-making patterns pertinent for transportation planning applications. Moreover, other than CDR data, the techniques presented in this paper rely only upon nationally-available survey data to allow transferability of the methodology to other study areas in the US.

Extensive research has been conducted into OD estimation, as these trips provide the basis for transportation feasibility and impact studies. Conventional OD estimation approaches rely on surveys and/or travel demand models to provide trip matrices. Often, such trip matrices are calibrated or updated using traffic counts and estimation techniques such as maximum likelihood, generalized least squares, and optimization (Spiess, 1987; Cascetta, 1984; Bell, 1991; Yang et al., 1992). This research provides a realistic, cost-effective alternative to these traditional OD data sources and estimation approaches. By presenting a systematic and replicable procedure to extract data relevant to the transportation community, we hope this work will help to facilitate the use of mobile phone data in practice.

In this paper, we demonstrate methods to analyze mobile phone records for the Boston metropolitan area. In Section 2, we present an overview of the data and the methods developed to produce OD trips by purpose and time of day. In Section 3, we summarize and validate our results against independent data sources for the study area, including the US Census and household travel surveys. Based on these findings, we conclude with a discussion of the limitations and applications of CDR data in the context of transportation planning and modeling.

2. Data and methods

2.1. CDR data

The studied dataset contains more than 8 billion anonymized mobile phone records (from several carriers) from roughly 2 million users in the Boston metropolitan area over a period of two months in the Spring of 2010. Although the CDR data spans 60 days, the data provider reindexed the anonymous user IDs for most of the users after the 17th day of the dataset. Effectively, we observe some users for at most 17 days, some users for at most 43 days, and still others for up to 60 days.

Each record contains an anonymous user ID, longitude, latitude, and timestamp at the instance of a phone call or other types of phone communication (such as sending SMS, etc.). The coordinates of the records are estimated by service providers based on a standard triangulation algorithm, with an accuracy of about 200–300 m. In typical mobile phone data sets, locations are represented by cell towers rather than triangulated coordinates and therefore have a lower spatial resolution; however, the method proposed here is expected to hold for such cases (Song et al., 2010a; Wang et al., 2012).

2.2. Stay extraction

The first step to reliably infer activities and trips from CDR data is to filter out noise resulting from (1) tower-to-tower call balancing performed by the mobile service provider, creating the appearance of false movements, and (2) inexact signal triangulation. Furthermore, we wish to distinguish users' stationary stay locations (when/where users engage in an activity) from their moving pass-by locations (when/where users are en-route to activities). To do so, we develop a method based in the work of Hariharan and Toyama (2004) for processing GPS traces. The spatial and temporal filtering methods are discussed below and illustrated in Fig. 1.

Let sequence $D_i = (d_i(1), d_i(2), d_i(3), \ldots, d_i(n_i))$ be the observed data for a given anonymous user i, where $d_i(k) = (t(k), x(k), y(k))'$ for $k = 1, \ldots, n_i$, and $t(k)$, $x(k)$, and $y(k)$ are the time, longitude, and latitude of the k-th observation of user i. First, we extract points $d_i(k)$ that are spatially close (i.e. within roaming distance of 300 m) to their subsequent observations, say, $d_i(k+1), d_i(k+2), \ldots, d_i(k+m)$. To reduce the jumps in the location sequence of the mobile phone data, we assume that $d_i(k), \ldots, d_i(k+m)$ are observed when user i is at a specific location, i.e., the medoid of the set of locations $(x_i(k), y_i(k))'$, $\ldots, (x_i(k+m), y_i(k+m))'$, which is denoted by

$$Med((x_i(k), y_i(k))', \ldots, (x_i(k+m), y_i(k+m))').$$

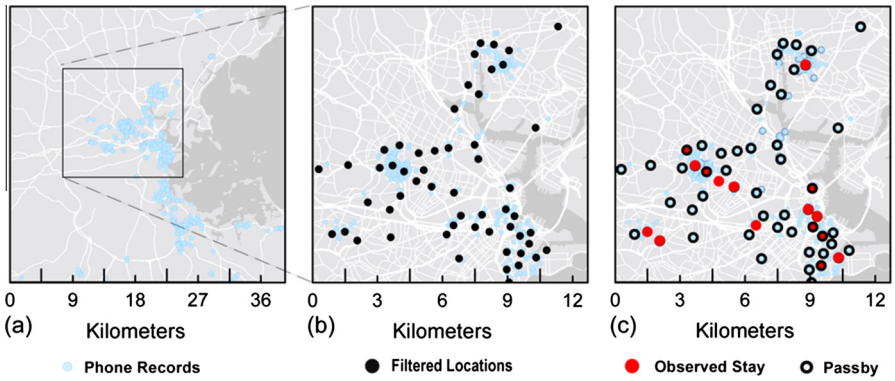

Fig. 1. Extracting stay and pass-by areas from the phone data for an anonymous user in the 2-month period.

This treatment respects the time order at first, to ignore noisy jumps in estimated location, but then disregards time ordering to apply the *agglomerative clustering algorithm* (Hariharan and Toyama, 2004) to consolidate points that are close in space but may be far apart in time. The points to be consolidated together form a cluster whose diameter is required to be no more than a certain threshold (set as 500 m). Again we modify the observation locations to the corresponding medoids of the clusters (see Fig. 1(a) and (b)).

Next, we impose the time duration criterion on the clean data, and extract the stay locations whose durations exceed a certain threshold (set as 10 min). In the example presented in the figure we extract 31 distinct stay locations from the 1776 phone records in the two-month period of the exhibited anonymous user (see Fig. 1(c)). The rest of the points are called pass-by points, at which we do not observe any lengthy stays. Note that it is possible that the user stays in some of these pass-by locations as well as locations that we do not observe. In these cases, information about time and location is totally or partially latent to us as we do not observe it from the phone records. However, all the stay locations frequently visited by the user ought to be extracted from the mobile phone data, if the observation period is long enough. As such, the pass-bys are filtered out and the stays are assumed to be trip origins or destinations, between which trips are made. Analysis of the pass-by points is out of the scope of the present work, in which we focus on simple trip chains with origins and destinations labeled as: home, work, or other.

2.3. Activity inference

Trips are induced by the need or desire to engage in activities (Pinjari and Bhat, 2011) and therefore understanding patterns and types of activities is crucial in estimating travel demand. It has been demonstrated that human mobility patterns are characterized by regularity with frequent returns to previously visited locations (Song et al., 2010b; Song et al., 2010a; Hasan et al., 2013). Due to this predictability, we are able to reasonably infer stay activities for users' most visited locations (i.e. home and work).

Accordingly, our first task is to label the stay regions in order to assign trip purpose. For each user, the stay extraction process detailed above results in a timestamp and duration for each observed visit to a stay location. For this study, we assign an activity type of either *home*, *work*, or *other* to each users' stay locations. Future research can expand the *other* designation to activity types such as school, shopping, recreation and social, using land use information.

Each user's *home* location is identified as the stay with the most visits on (i) weekends and (ii) weekdays between 7 pm and 8 am, representing the time windows in which we expect users to spend substantial amounts of their time at home. In addition to inferring trip purpose, the *home* stay location of each user is used to filter out users with too few data points and expand the data from phone users to study area population, as summarized in Section 2.4.

A *work* location is identified as the stay (not previously labeled as *home*) to which the user travels the maximum total distance from *home*, $max(d*n)$, where n is the total number of visits to a given stay on weekdays between 8 am and 7 pm and d is the distance between the latitude-longitude coordinates of the *home* stay and the given stay using plane approximation. This assumption is based on the rationale and historical evidence (Levinson and Kumar, 1994; Schafer, 2000) that for a given frequency of visits, longer distance trips are more likely to be work trips than shorter distance trips, which are more likely to be for non-work purposes (i.e. to the nearby grocery store).

If the user visits the identified *work* stay less than 8 times ($n < 8$; once a week, on average) or the distance is less than 0.5 km ($d < 0.5$), then the activity of the stay region is identified as *other* rather than *work*. In effect, not all users are assigned a *work* stay, accounting for the fact that not all users commute to a job. Subsequently, all the remaining stay locations not identified as *home* or *work* are designated as *other*. These classification assumptions serve to avoid falsely identifying a location as work that is either not visited frequently enough or close enough to a user's home that it could reflect signal noise rather than a distinct location.

We acknowledge that under these simple assumptions we may misidentify users' *true* home and work locations and, by extension, their trip purposes. However, based on comparisons with census data (presented below) this procedure give us very good estimates of the distribution of home and work locations and home-work flows in our study region. Note that these assumptions are related to the duration and spatial resolution of this dataset, and it may be necessary to adjust them for applications of other datasets.

2.4. Data filtering and expansion

For users with too few stay locations, the CDR data may not fully represent their travel patterns. Accordingly, users with fewer than 8 (one per week, on average) visits to designated *home* stays are filtered out. This filter serves the additional purpose of ensuring with a reasonable degree of certainty that the designated stay is the user's home, a key assumption in our method of upscaling users to population. Note that this filtering process necessarily excludes visitors, for whom a home location is not observed in the studied dataset. Future research could look at extracting visitor trips from CDR data using an assumption other than home location to upscale these trips.

After this filtering, 335,795 users remain in the Boston CDR dataset. This sample size is an order of magnitude larger than in most household travel surveys, and should increase given longer periods of observation. To upscale these users to total population of the study region, the number of *home* stays were aggregated to the 974 Census Tracts in the study area. An expansion factor was then calculated for each Tract as the ratio of the 2010 Census population and the number of residents identified in the CDR data. For the 10 Census Tracts with fewer than 10 CDR residents, the scaling factor is set to 0 to ensure that we do not overweight users that are not representative of a given Census Tract. The 1st, 2nd, and 3rd quartiles of the expansion factors are 9.4, 14.2, and 25.1, respectively, as illustrated by the tight probability distribution of expansion factors in Fig. 2a. The spatial distribution illustrated in Fig. 2b suggests that the Tracts in the western portion of the study area tend to be more heavily weighted. CDR data for a period greater than 60 days would likely have lower expansion factors and an improved spatial distribution of users, however, we show that already this limited data set gives reasonable results.

2.5. Trip estimation

With stays for each user designated by activity type and expansion factors to upscale users to population, average daily origin–destination trips can be constructed by time of day and purpose—home-based work (HBW), home-based other (HBO), and non-home based (NHB). This segmentation allows us to capture distinct trip-making patterns and is consistent with segmentation in the trip distribution stage of trip-based travel demand models.

Since the timestamp and duration associated with each stay reflect the *observed* (based on phone usage) rather than *true* arrival time and duration of a user, we infer trip departure time using probability density functions to account for this uncertainty. The publicly-available 2009 National Household Travel Survey (NHTS) (U.S. Department of Transportation Federal Highway Administration, 2011), filtered for respondents residing in a consolidated metropolitan statistical area (CMSA) or MSA with populations greater than or equal to 3 million, is a reasonable source as it approximates temporal travel patterns of major US cities comparable to Boston, while allowing for transferability of this methodology to other US cities. Using this departure time data, we generate six hourly distributions for weekdays and weekends and the following trip purposes: HBW, HBO, and NHB.

For each user, it is assumed that a trip is made between two consecutive stays $(i,\ i+1)$ occurring within a 24 h period beginning and ending at 3 am. The trip occurs at a point in time spanned by the range $[s_i+\delta_i, s_{i+1}]$, where s is the observed arrival time and δ is the observed duration of a stay. The departure hour is randomly generated in this time window using the NHTS distribution that corresponds to the day (weekday, weekend) and the trip purpose identified from the origin and destination stay activities (HBW, HBO, NHB).

Furthermore, it is presumed that a user starts and ends each 24 h period at home such that if a user is not recorded at his/her *home* stay for the first (last) record of the 24 h period, his/her first (last) trip begins (ends) at his/her *home* stay. The first (last) trips are assumed to occur at point in time spanned by the range $[3AM, s_{i+1}]$ ($[s_i+\delta_i, 3AM]$), where s is the observed arrival time and δ is the observed duration of a stay. As before, the departure hour is randomly generated in this window using the NHTS distribution that corresponds to the day (weekday, weekend) and the trip purpose based on the destination (origin) stay activity (HBW, HBO).

Through this process, we construct trips on all days we observe each user. The frequency of weekday observations per user is illustrated in Fig. 3. The distribution of total weekday trips per user is shown in Fig. 3a, with first, second, and third quartiles of 33, 58, and 96 trips, respectively. The reindexing of anonymous user IDs mentioned previously in Section 2.1 is evident in the two peaks of the distribution of the number of weekday days we observe each user, as seen in Fig. 3b. Despite this reindexing, we achieve a sufficiently large number of observation days per person, with first, second, and third quartiles of 11, 17, and 21 days, respectively. Dividing each user's total weekday trips by his/her total weekday days, we get the distribution of average weekday trips shown in Fig. 3c. The distribution has a long tail, however, the first, second, and third quartiles are 2.6, 3.2, and 4.3 average trips per weekday, respectively, demonstrating that the vast majority of users have a reasonably small number of daily trips.

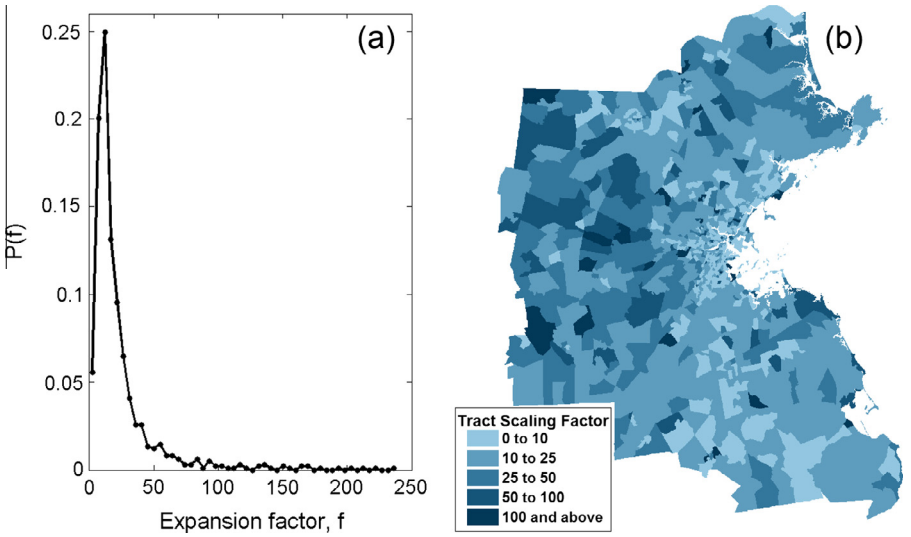

Fig. 2. (a) Probability distribution of Census Tract expansion factors. (b) Thematic map showing the spatial distribution of Census Tract expansion factors.

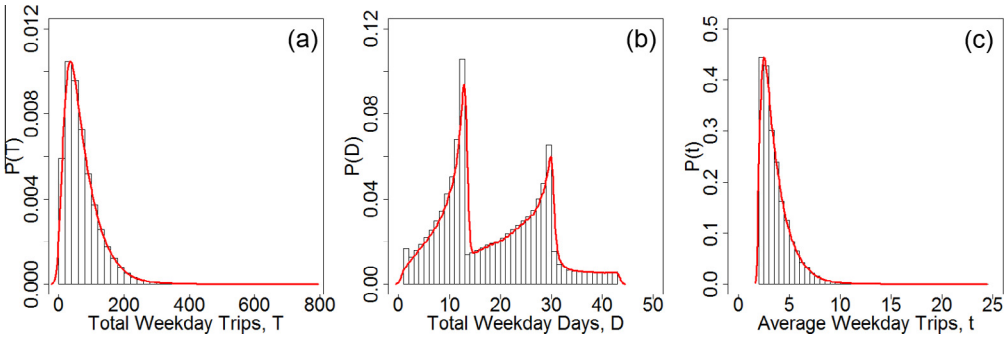

Fig. 3. Frequency of weekday observations per user. (a) Probability distribution of total weekday trips per user. (b) Probability distribution of total weekday days per user. (c) Probability distribution of average weekday trips per user.

In order to obtain average daily OD trips, each users trips are multiplied by the expansion factors described in Section 2.4 for the user's *home* Census Tract and divided by the number of days from which we constructed the user's trips. For users assigned a *work* stay, weekday trips are only constructed on days in which the user is observed at his/her *work* stay to ensure we capture representative weekdays of commuters. Unlike traditional travel surveys which ask a respondent details about one or a few recent days, this method has the advantage of capturing many days per user and thus variations in his/her daily travel behavior. Lastly, each user's average daily trips are aggregated into Census Tract pair trip matrices by day type (weekday, weekend), purpose (HBW, HBO, NHB), and hour of departure.

3. Results and validation

3.1. Productions and attractions

Accurately extracting and upscaling users' stays is crucial to trip generation. Due to the regularity of human behavior (Song et al., 2010a; Song et al., 2010b; Hasan et al., 2013), we are able to infer users' *home* and (if applicable) *work* stay locations from CDR data. For this dataset, we find that we can reasonably represent the spatial distribution of home and work locations when aggregated to the 164 study area cities and towns (MassGIS, 2014). Refer to Section 3.2 below for more information on the impact of aggregation level on accuracy. Fig. 4a shows a comparison of home locations by town from 2010 Census data and the raw and upscaled CDR data.

As we would expect since Tract population was used to upscale the data, the number of residents in each town is almost identical to that of the upscaled CDR data. However, the slope of a best-fit line through the raw CDR data is close to 1, which speaks to the fact that the overall distribution of raw CDR users is fairly representative and a simple factoring

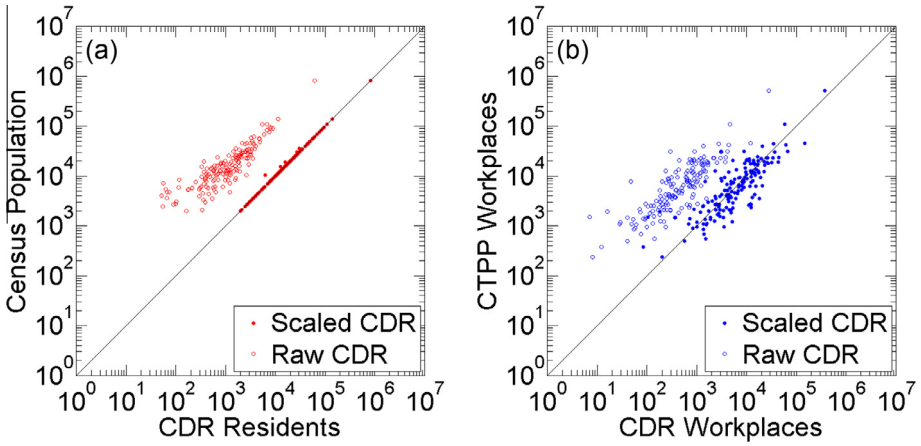

Fig. 4. (a) CDR residents vs. 2010 Census population by town before and after population expansion. (b) CDR vs. Census Transportation Planning Products (CTPP) (U.S. Department of Transportation Federal Highway Administration, 2013) workers by town before and after population expansion.

method is in fact appropriate to expand the phone users to population. Similarly, Fig. 4b shows a comparison of work locations aggregated by town. As with the raw CDR data on the home-end, the distribution of raw workplaces is fairly consistent with the 2006–2010 Census Transportation Planning Products (CTPP) (U.S. Department of Transportation Federal Highway Administration, 2013) data (slope approximately 1), and the upscaling method adjusts well for the difference in magnitude. This strong correlation is noteworthy considering that each users' *home* and *work* locations were scaled based on their *home* location only.

3.2. Trip distribution

With the establishment of reasonable distributions of trip productions and attractions, we next validate the distribution of trips using two local surveys. The 1991 Boston Household Travel Survey (BHTS) contains information on 39,300 trips made by 3737 households (Boston Metropolitan Planning Organization, 1991), while the 2010/2011 Massachusetts Travel Survey (MHTS) contains data on 153,099 trips made by 32,739 people (NUSTATS, 2012). We find that the CDR trips compare well with trips from these data sources by time of day and purpose. Fig. 5 illustrates the distributions of hourly departure times for (a) HBW, (b) HBO, (c) NHB, and (d) total average weekday trips. Note that we also benchmark against the NHTS departure time distributions, which were used to infer departure time for the CDR trips. Accordingly, differences between each of the hourly NHTS and CDR distributions reflect the observed arrival and duration times of CDR stays.

Most notably, there are consistently more CDR trips in the late night hours than that of the surveys. While this may be due to a slight mismatch between the frequency of calling and trip-making throughout the day, it may also highlight an advantage of CDR data to capture late night trips not typically reported in survey responses of an average day. Regardless, most transportation planning applications focus on trips in the morning and evening peak periods, when congestion is most prevalent, and for which we compare well. Similar trends are evident for average weekday trip shares segmented by key time periods, as presented in Table 1.

Furthermore, the relative share of average weekday trips for each trip purpose is comparable for the CDR and survey data. Table 1 shows that the shares of HBW, HBO, and NHB CDR trips are within the ranges of trip purpose shares across all three surveys. This again suggests that our inferences of *home*, *work*, and *other* activities, as well as their relative prevalence in the data set, seem reasonable.

To draw comparisons on the magnitude of daily CDR trips, we MHTS data, which includes weights to expand respondents to population estimated from the 2006–2010 American Community Survey (NUSTATS, 2012). Table 2 shows a comparison of average weekday trips by purpose and period of the day for the CDR trips and weighted MHTS trips. The survey reports more daily trips than we observe in the CDR data, with most of the difference coming from the NHB trip segment. Still, the total CDR and MHTS trips imply reasonable numbers of average weekday trips per person – 3.50 and 4.24, respectively.

Lastly, Table 2 presents a comparison of the spatial distribution of daily CDR and MHTS trips at the Tract-pair and town-pair level. The correlation coefficients of the trip matrices improve significantly with aggregation to the 164 study area cities and towns. In particular, the HBW and AM correlations at the Tract-pair level see the largest improvement. This may be indicative of the role of the size of Tracts, which are considerably smaller in downtown Boston where many of the morning commute trips end. We discuss the relationship between aggregation level and correlation in more detail in Section 3.3 below.

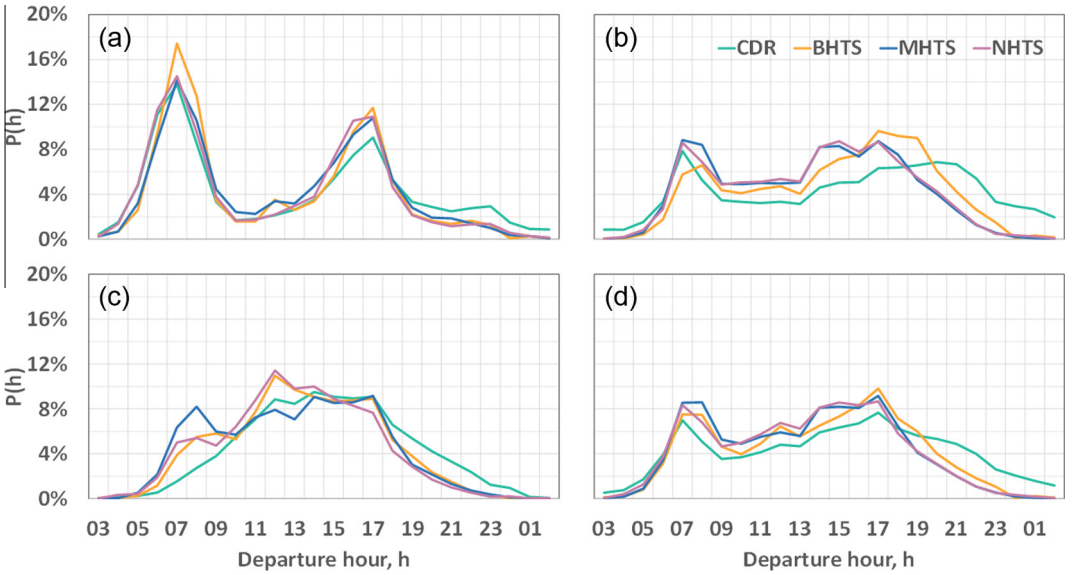

Fig. 5. Distribution of average weekday hourly departure time from CDR data, 1991 Boston Household Travel Survey (BHTS) (Boston Metropolitan Planning Organization, 1991), the 2010/2011 Massachusetts Travel Survey (MHTS) (NUSTATS, 2012), and 2009 National Household Travel Survey (NHTS) (U.S. Department of Transportation Federal Highway Administration, 2011) for (a) home-based work trips, (b) home-based other trips, (c) non-home based trips, and (d) all trips.

Table 1
Average weekday trip shares by purpose and period from CDR data, 1991 Boston Household Travel Survey (BHTS) (Boston Metropolitan Planning Organization, 1991), the 2010/2011 Massachusetts Travel Survey (MHTS) (NUSTATS, 2012) and the 2009 National Household Travel Survey (NHTS) (U.S. Department of Transportation Federal Highway Administration, 2011).

Source (%)	HBW (%)	HBO (%)	NHB (%)	Morning 6a–9a (%)	Mid-day 9a–3p (%)	Evening 3p–7p (%)	Rest-of-day 7p–6a (%)
CDR	18	51	31	16	27	27	30
BHTS	20	48	32	18	32	33	17
MHTS	12	49	39	21	34	33	12
NHTS	14	55	30	19	37	31	13

Table 2
Average daily trips by purpose and period from CDR data and the 2010/2011 Massachusetts Travel Survey (MHTS) (NUSTATS, 2012), as well as the correlation coefficients of CDR and MHTS Tract-pair and Town-pair trips.

	HBW	HBO	NHB	AM 6a–9a	MD 9a–3p	PM 3p–7p	RD 7p–6a	Total
CDR Trips (in millions)	2.81	7.84	4.73	2.46	4.12	4.15	4.65	15.37
MHTS Trips (in millions)	2.14	8.99	7.18	3.99	6.24	6.06	2.31	18.61
Tract-pair correlation	0.30	0.64	0.58	0.42	0.65	0.54	0.40	0.58
Town-pair correlation	0.96	0.97	0.98	0.97	0.98	0.97	0.96	0.98

3.3. Home-work flows

Commuting trips represent a key travel market and source of daily roadway congestion, and accurately representing these trips is an important step in validating trips estimated from CDR data. Accordingly, we next compare with flows between people's home and work locations, as reported by the 2006–2010 Census Transportation Planning Products (CTPP) (U.S. Department of Transportation Federal Highway Administration, 2013). Distinct from the average daily HBW trips compared in Section 3.2, these flows simply link home and work, ignoring that people's daily trip chains may in fact include work trips to/from locations other than home.

Table 3 summarizes statistics that support the comparison of CDR and CTPP home-work (HW) flows. In addition to the total magnitude of trips, the similarities between the percentages of inter-tract and inter-town flows and average trip length give a high-level indication that the distributions of HW flows are similar.

At the flow level, we find that the correlation between CDR and CTPP HW Tract-to-Tract and town-to-town flows is 0.45 and 0.99, respectively, indicating that the level of aggregation of trips has a significant impact on accuracy. We demonstrate that as we gradually increase average aggregation size using variably-sized buffers around each origin and destination Tract

Table 3
Comparison of average weekday HW CDR and 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows.

Source	Daily HBW trips (millions)	Inter-Tract share (%)	Inter-town share (%)	Average trip length (miles)
CDR	2.11	94	68	9.67
Census	2.10	90	68	10.72

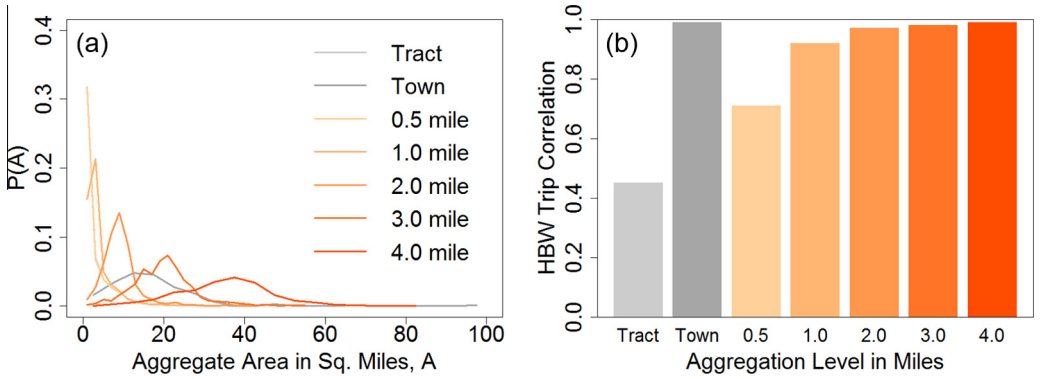

Fig. 6. (a) Probability density distributions of aggregation area size by designated areas (Tract or towns) and variable buffers. (b) Correlation between HBW CDR and 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows corresponding to the aggregation levels in (a).

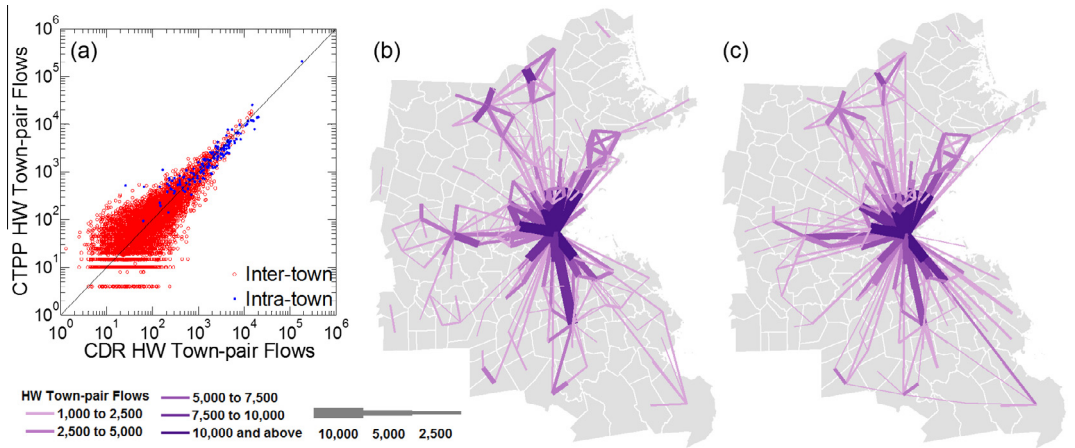

Fig. 7. (a) Intra-town and inter-town pair daily HW CDR flows and 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows. (b) Spatial distribution of daily inter-town HW CDR flows (>1000). (c) Spatial distribution of daily inter-town HW 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows (>1000).

(Fig. 6a), the correlation between CDR and CTPP HW trips increases as well (Fig. 6b). We find that using small aggregation buffers has the most significant impacts on correlation, while having minimal influence on average aggregation size (as illustrated by the fact that the distribution for the 0.5 mile buffer obscures that of the Tract-level aggregation in Fig. 6b). In effect, using a 0.5 mile buffer aggregates the small, dense Tracts (i.e. in the city center) and results in a notable improvement in accuracy. In the absence of meaningful districts or communities to which to aggregate, this can inform suitable distance thresholds for trip clustering to overcome limitations of sparse data and/or spatial inaccuracy.

We further investigate comparisons of the data sets using town-pairs flows. Fig. 7a shows the CDR and CTPP HW flows for all of the intra-town and inter-town pairs, which have correlations of 0.99 and 0.95, respectively. It is evident from Fig. 7a that town pairs with many trips validate better than those pairs with few trips, especially those with fewer than about 500 daily trips. This trend is likely due to sparsity in data for these smaller markets. Fig. 7b and c illustrate spatially the HW flow distribution for key markets (inter-tract pairs with greater than 1000 daily trips) for the CDR and Census data, respectively. Inspecting the figure, it is evident that the CDR data captures very similar patterns to that of the CTPP commuting data, with the majority of flows directed in and out of Boston as well as a few shorter distance markets in the suburban towns.

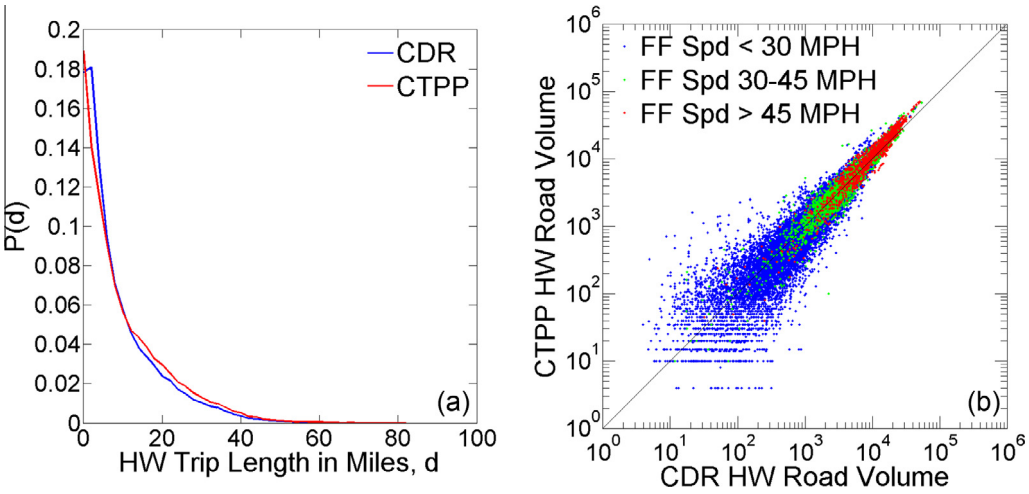

Fig. 8. (a) Trip length distribution of daily HW CDR and 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows. (b) Road segment volumes for daily HW CDR and 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows by free flow speed.

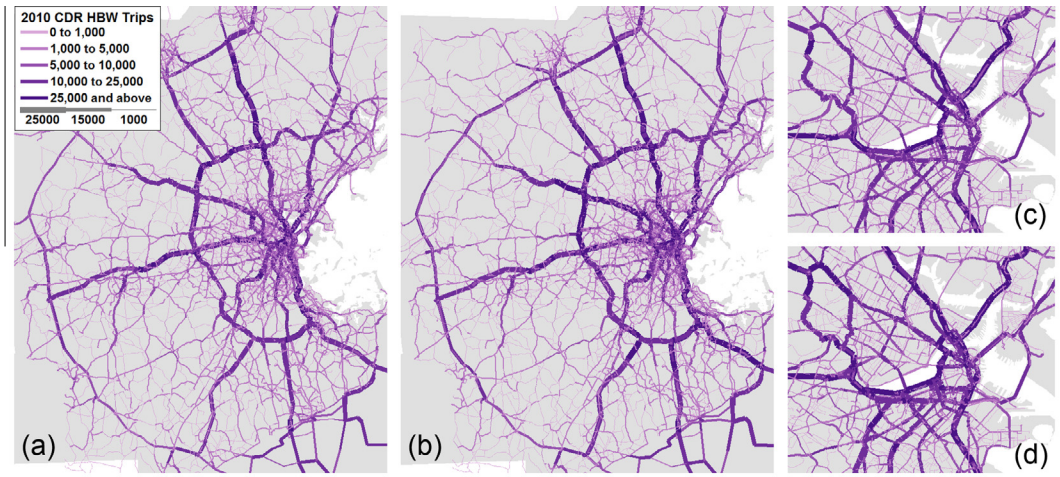

Fig. 9. Road segment volumes of HBW trips (a) from CDR data in the Boston metro area and (c) downtown Boston, and (b) from 2006–2010 CTPP data (U.S. Department of Transportation Federal Highway Administration, 2013) in the Boston metro area, and (d) downtown Boston. Flows are generated from Tract-to-Tract ODs in TransCAD (Caliper, 2009) using all-or-nothing highway assignment minimizing travel time.

Assigning the Tract-to-Tract trips to roads offers another valuable spatial comparison as it considers potential paths of OD trips and has important implications for planning applications. Although it is not representative of a meaningful traffic scenario, we assign all daily HW flows (irrespective of time of day or mode) to a road network for this comparison. Traffic assignment also allows us to estimate and compare trip length distributions across the two datasets. Fig. 8a illustrates that the trip length distributions are indeed very similar, consistent with our findings of comparable trip distributions.

Fig. 8b illustrates strong correlation between CDR and CTPP road segment volumes by free-flow speed, which serves as a proxy for major and minor arterials. With correlations of 0.97, 0.98, and 0.95 for all segments with free flow speeds greater than 45 MPH, between 30 and 45 MPH, and less than 30 MPH, respectively, it is evident that the roadway volumes estimated from CDR and CTPP data are very similar, especially for major roads. The lower correlation on more minor road segments follows the finding observed in Fig. 7a, in which Tract-pairs having few daily trips have the lowest correlation, since minor roads typically serve these smaller markets. Spatially, Fig. 9 illustrates these correlations for the greater metropolitan area and downtown. Although differences in road segment volumes are virtually indistinguishable visually (in Fig. 9), the CTTP

commuting trips result in slightly higher road segment volumes on major roads experiencing the highest volumes (in Fig. 8b).

Despite the lower correlations observed for Tract-pairs and road segments with lower flows and volumes, respectively, these markets have minimal impact on the network as a whole. Further, surveys are susceptible to inaccurate sampling and/or upscaling due to infrequency or scarcity of trips in these minor markets. Accordingly, the *ground truth* number of trips for Tract-pairs with few trips is unknown and comparisons between survey and CDR trips reflect this uncertainty and noise.

4. Conclusions

In this paper, we detailed steps necessary to extract average daily origin–destination trips by purpose and time of day from mobile phone call detail records (CDRs). The proposed techniques were applied to CDRs in the Boston metropolitan area and validated against local and national surveys. The methods are transferable to other study areas and could be reproducible by researchers and practitioners using mobile phone and census data.

Emphasizing the importance of data preprocessing, much of the methods serve to filter out noise and extract accurate travel patterns representative of the study area. While this processing reduces the immensity of the CDR data, we are left with a sample size that is an order of magnitude larger than most household travel surveys. Further, we observe many days per user, allowing us to capture variation in daily behavior, including weekends, not typically reported household travel surveys.

We find that the size of the areas used to aggregate trips is a very important factor in how well the CDR and survey data compare. We observe significantly higher trip correlation when aggregating origins and destinations to 164 cities and towns rather than the 974 Census Tracts in the study area. This improvement in accuracy is seemingly an effect of aggregating small Census Tracts (i.e. in the city center), for which CDR data may not have a sufficiently-large sample size or the necessary spatial accuracy. In general, aggregating trip origins and destinations to areas greater than 1 square mile produces agreement with survey data. As mobile phone providers collect more dense data such as GPS traces or wifi access points, spatial and temporal data sparsity will decrease, and accordingly, aggregation size can decrease relative to a given level of precision. Although we can reasonably represent average daily activity and trip patterns with CDRs, data limitations preclude its use in applications requiring richer data such as real-time, dynamic OD estimation.

Aggregating to towns results in similar distributions of upscaled home and work locations inferred from the CDR data and the home- and workplace-based tabulations from the 2006–2010 US Census Transportation Planning Package (CTPP) (U.S. Department of Transportation Federal Highway Administration, 2013). Additionally, our inferred distributions of trips by hour of the day and purpose are comparable with the 1991 Boston Household Travel Survey (Boston Metropolitan Planning Organization, 1991), 2010/2011 Massachusetts Travel Survey (NUSTATS, 2012), and the 2009 National Household Travel Survey (U.S. Department of Transportation Federal Highway Administration, 2011) (filtered for trips in MSAs and CSAs with populations greater than 3 million). Finally, the spatial distribution of home-work flows is highly correlated with that of the CTPP, a well-established nation-wide source for Tract-to-Tract commuting data.

In validating OD trips by purpose and time of day, we demonstrate that CDR data can be effectively used to represent distinct mobility patterns across market segments typically relevant to transportation planning applications. In particular, CDR data can be used to augment or complement traditional survey data, which provides detailed information about a respondent and his/her trip but is more costly and onerous to collect. Transportation models rely heavily on survey data for inputs, calibration, and validation, and CDR data can be a valuable new resource. Furthermore, the outputs of our proposed methodology are analogous to the outputs of the trip generation and distribution steps of traditional four-step travel demand models. In areas where public transportation is significant, OD matrices developed from CDRs can be post-processed to obtain mode-specific trip tables, equivalent to the mode split step. As such, CDR data can be very useful for planning applications and/or study areas where running such a model is either not feasible or not necessary.

In addition to average daily origin–destination trips, mobile phone data captures individuals' daily trip chains and is therefore well-suited for activity-based models, especially if land use information can be used to infer activity types beyond home, work, and other. Future steps for analyzing this data include traffic assignment of vehicle trips inferred from CDRs by time of day, allowing us to explore how these data sets help to improve existing urban trip models and applications related to mitigating congestion.

Acknowledgements

This work was partially funded by the MIT-Accenture alliance, the BMW-MIT collaboration under the supervision of PI Mark Leach,[1] the Austrian Institute-HuMNet collaboration agreement under the supervision of PI Dietmar Bauer[2] and the Center for Complex Engineering Systems (CCES) at KACST under the co-direction of Anas Alfaris.[3] We thank Yingxiang Yang and Peter Widhalm for technical support.

[1] mark.leach@bmw.de.
[2] Dietmar.Bauer@ait.ac.at.
[3] anas@mit.edu.

References

Bell, M.G.H., 1991. The estimation of origin-destination matrices by constrained generalised least square. Transport. Res. Part B: Methodol. 25, 13–22.

Boston Metropolitan Planning Organization, 1991. 1991 Boston Household Travel Survey. <http://www.surveyarchive.org/Boston/Boston_91.zip>.

Caliper, 2009. TransCAD Transportation Planning Software. <http://www.caliper.com/tcovu.htm>.

Cascetta, E., 1984. Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator. Transport. Res. Part B: Methodol. 18, 289–299.

Hariharan, R., Toyama, K., 2004. Project lachesis: parsing and modeling location histories. Geogr. Inform. Sci., 106–124

Hasan, S., Schneider, C., Ukkusuri, S.V., González, M.C., 2013. Spatiotemporal patterns of urban human mobility. J. Stat. Phys. 151.

Hu, P.S., Reuscher, T.R., 2004. Summary of Travel Trends: 2001 National Household Travel Survey. Technical Report. U.S. Department of Transportation Federal Highway Administration. <http://nhts.ornl.gov/2001/pub/stt.pdf>.

Huntsinger, L.F., Donnelly, R., 2014. Reconciliation of regional travel model and passive device tracking data. In: Proceedings of the 93rd Annual Meeting of the Transportation Research Board.

Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin-destination matrices using mobile phone call data. Transport. Res. C 40, 63–74.

Jiang, S., Yang, Y., Fiore, G., Jr., J.F., Frazzoli, E., González, M., 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing.

Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T., 2010. A survey of mobile phone sensing. IEEE Commun. Mag. 48, 140–150.

Levinson, D.M., Kumar, A., 1994. The rational locator: why travel times have remained stable. J. Am. Plan. Assoc. 60, 319–332.

MassGIS, 2014. Community Boundaries. <http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/towns.html>.

NUSTATS, 2012. Massachusetts Department of Transportation: 2010/2011 Massachusetts Travel Survey.

Pinjari, A.R., Bhat, C.R., 2011. Activity-based travel demand analysis. Handbook Transp. Econ. 1, 1–36.

Richardson, A., Ampt, E.S., Meyburg, A.H., 1995. Survey Methods for Transport Planning. Eucalyptus Press, Melbourne.

Schafer, A., 2000. Regularities in travel demand: an international perspective. J. Transport. Stat. 3, 1–31.

Schneider, C., Belik, V., Couronné, T., Smoreda, Z., González, M.C., 2013. Unraveling daily human mobility motifs. J. Roy. Soc. Interface 10.

Song, C., Koren, T., Wang, P., Barabási, A.L., 2010a. Modelling the scaling properties of human mobility. Nature Phys. 6, 818–823.

Song, C., Qu, Z., Blumm, N., Barabási, A.L., 2010b. Limits of predictability in human mobility. Science 327, 1018–1021.

Spiess, H., 1987. A maximum likelihood model for estimating origin-destination matrices. Transport. Res. Part B: Methodol. 21, 395–412.

Stopher, P., Greaves, S., 2007. Household travel surveys: where are we going? Transport. Res. Part A: Policy Practice 41, 367–381.

U.S. Department of Transportation Federal Highway Administration, 2011. 2009 National Household Travel Survey. <http://nhts.ornl.gov/download.shtml>.

U.S. Department of Transportation Federal Highway Administration, 2013. CTPP 2006-2010 Census Tract Flows. <http://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/2006-2010_tract_flows/index.cfm>.

Wang, P., Hunter, T., Bayan, A.M., Schectner, K., González, M.C., 2012. Understanding road usage patterns in urban areas. Sci. Rep. 2.

Yang, H., Sasaki, T., Iida, Y., Asakura, Y., 1992. Estimation of origin-destination matrices from link traffic counts on congested networks. Transport. Res. Part B: Methodol. 26, 417–434.

The TimeGeo modeling framework for urban mobility without travel surveys

Shan Jiang[a,1], Yingxiang Yang[a,1], Siddharth Gupta[a], Daniele Veneziano[a], Shounak Athavale[b], and Marta C. González[a,c,2]

[a]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Research & Innovation Center, Ford Motor Company, Palo Alto, CA 9304; and [c]Center 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.

Author contributions: S.J., Y.Y., and M.C.G. designed research; S.J., Y.Y., S.G., D.V., S.A., and M.C.G. performed research; S.J., Y.Y., and S.G. analyzed data; and S.J., Y.Y., D.V., and M.C.G. wrote the paper.

Conflict of interest statement: The authors declare a conflict of interest. The presented work is part of a patent pending: Massachusetts Institute of Technology Case 18887, "UrbanFlows: Improving Urban Traffic Without Surveys," by M.C.G.

This article is a PNAS Direct Submission.

[1]S.J. and Y.Y. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: martag@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1524261113/-/DCSupplemental.

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

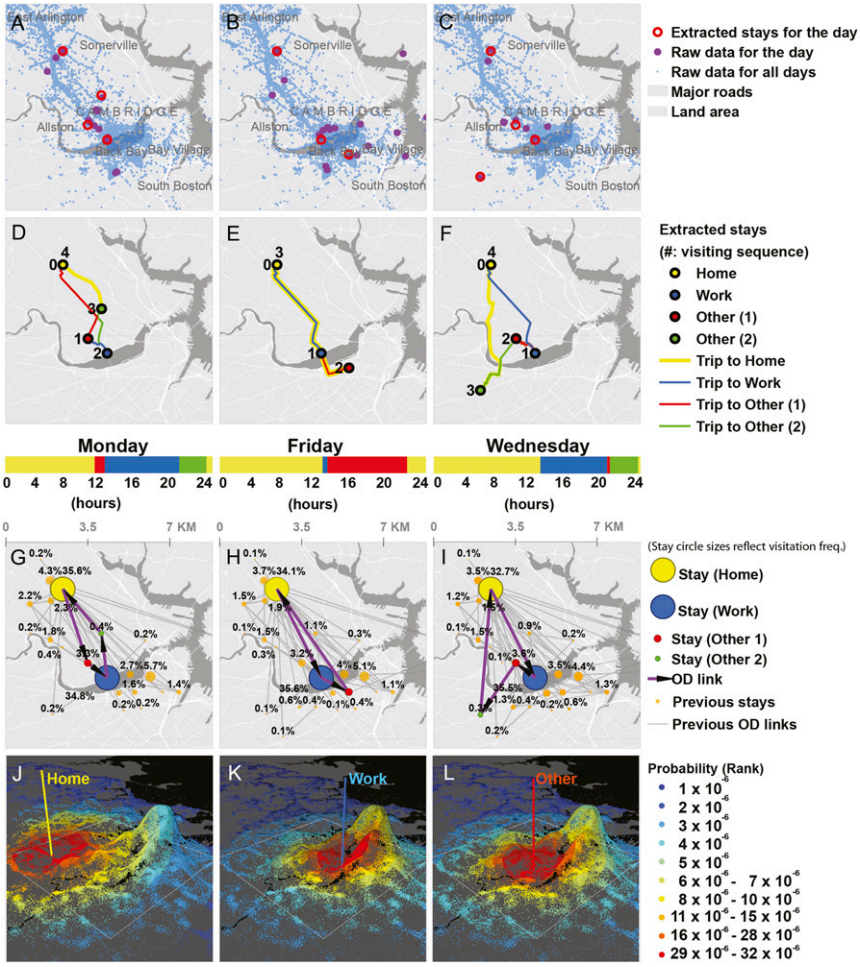

Fig. 1. Extraction of stays and daily journeys from raw cell phone data. (*A–C*) Stay locations extracted from the self-collected cell phone records of a student in three sample days. (*D–F*) Illustration of trips between consecutive stays in each day. (*G–I*) Visitation frequency of all locations, counting 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*A*. 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 (β_1), and burst rate (β_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

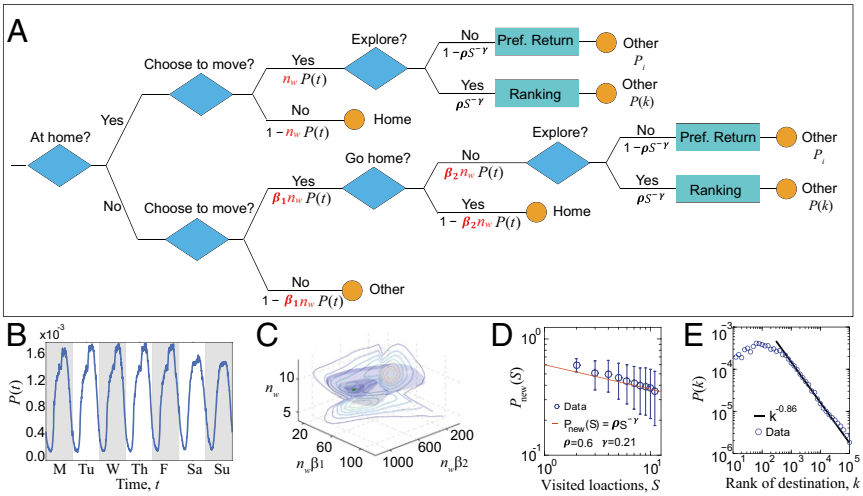

Fig. 2. Flowchart of TimeGeo and input features extracted from active CDR users. (*A*) Spatial and temporal choices per time step. Three individual specific parameters control temporal patterns, including the weekly home-based tours (n_w), dwell rate (β_1), and the burst rate (β_2). n_w influences the 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 $\beta_1 n_w$, $\beta_2 n_w$, 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 $P_{new} = 0.6S^{-0.21}$. (*E*) Empirical probability of choosing the rank k location as a trip destination follows $P(k) \sim k^{-0.86}$.

expected variation of travel in different time of the week (shown in Fig. 2*B*). For commuters, because work is modeled as a fixed activity, $P(t)$ does not include trips to or from work. The product of the two, $n_w P(t)$, less than 1, defines the individual travel probability at a specific time interval (t) while she is at home.

To model an individual's propensity to travel from an other (out-of-home) state, we introduce a dwell rate β_1 which measures how much more active (or likely to travel) the person is at an other state compared with home. The probability of traveling when an individual is at an other state is defined as $\beta_1 n_w P(t)$. By capturing individual propensity to move from an other state, $\beta_1 n_w$ controls the stay duration Δt for flexible activities. The higher the product $\beta_1 n_w$, the more likely the person will choose to move and thus the shorter duration Δt she will stay at other locations.

Next, if an individual is already out of home and chooses to move at time t, we then model her decision to either go home or go to an additional other location by introducing a burst rate β_2. We define the probability that the individual travels from an other location O_1 to an additional other location O_2 as $P(O_1 \to O_2) = \beta_2 n_w P(t) \beta_1 n_w P(t)$. It is assumed that for an individual who has decided to move, the probability of visiting an additional other location is proportional to $\beta_2 n_w$. The ratio between the two choices of going to an additional other location or going home can be presented as follows:

$$\frac{P(O_1 \to O_2)}{P(O_1 \to H)} = \frac{\beta_2 n_w P(t)}{1 - \beta_2 n_w P(t)}. \quad [1]$$

For a given value of $\beta_2 n_w$, when $P(t)$ is high (e.g., in the afternoon), people are more likely to visit additional other locations; when $P(t)$ is low, people are more likely to return home. For a given $P(t)$, the higher the value of $\beta_2 n_w$, the higher probability the individual will keep visiting flexible (other) locations, and thus the greater number of daily locations N she will visit.

Compared with previous models that randomly draw the stay duration (or waiting time Δt) or the number of visited locations (N) from aggregated empirical distributions (14, 15, 29), by introducing three individual-specific parameters including weekly home-based tour number n_w, dwell rate β_1, and burst rate β_2, we explicitly model the temporal dynamics of individual mobility. The Markov model framework allows it to be analytically tractable and to derive explicit effects in the resulting stay-duration and daily-location distributions $P(\Delta t)$ and $P(N)$ (*SI Appendix*, section 6).

Spatial Choices. To model the spatial choices of individual mobility, we propose a rank-based exploration and preferential return (r-EPR) model by incorporating a rank-based selection of new locations to the original EPR model (14). The EPR model explains well the differences in the frequency of visits of each location (13–18, 32). For each movement, an individual decides either to explore a new location with probability P_{new}, or return to a previously visited location with probability $1 - P_{new}$. The exploration probability $P_{new} = \rho S^{-\gamma}$ captures a decreasing propensity to visit new locations as the number of previously visited locations (S) increases with time, and effectively captures individual mobility choices between explorations and returns. If the individual decides to return to previously visited locations, she chooses a specific location i with probability P_i defined as the visitation frequency of location i (14). Fig. 1 *G–I* illustrates P_i with different circle sizes, using the volunteered student's location records as an example. In each subfigure, we label the visitation frequency of each location up to the current day. We highlight locations visited in the current day in the foreground and show the previously visited ones in the background.

If the individual decides to explore a new location, she needs to choose a destination from a large number of possible alternatives. One limitation of the original EPR model proposed in ref. 14 is its lack of a mechanism for the new-location selection. To select a new location, the original EPR model randomly draws the exploration jump-size (Δr) from a global empirical distribution. To model the exploration mechanism more sensible to the urban structure, in this study, we incorporate a rank-based selection mechanism for newly explored locations (i.e., r-EPR model).

Our selection mechanism gives a rank k to each alternative destination based on their distances to the trip origin (33–36). Among all potential new destinations, the one closest to the current location is of $k = 1$, the second closest $k = 2$, etc. The empirical probability of selecting the kth location as a destination is quantified as $P(k) \sim k^{-\alpha}$; 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 other activity extracted from the CDR data. The height of the dot on the z axis represents the dot 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 an 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. 3*E*). We progressively partition Ω_0 into 4^1, 4^2,..., 4^n square tiles with side length $2^{-1}, 2^{-2}, \ldots, 2^{-n}$. Each time a mother tile Ω_{i-1} (at resolution level $i - 1$) is partitioned into four daughter tiles Ω_i (at resolution level

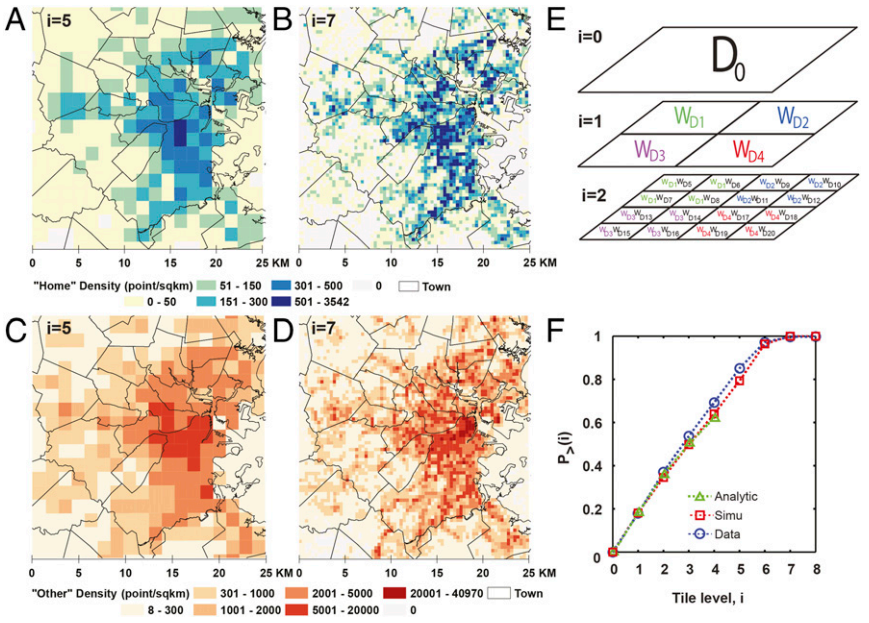

Fig. 3. Multiplicative cascade analysis framework. (*A* and *B*) The distribution of home locations in the Boston area at two different resolutions. (*C* and *D*) The distribution of other locations at two different resolutions. The variance of both distributions and their correlations depend on the resolution of the grids, or the cascade level *i*. At the scale with larger grid cells, the number of nonresidential (other) locations has higher correlation with the distribution of home locations, whereas at the scale with smaller grid cells separation between residential and other land-use types are observed. (*E*) Illustration of the hierarchical cascade process generating trip demand *D*. Each tile is repetitively divided into four smaller tiles. The density of locations in each tile is controlled by the cascade generator *W* at each tile level. (*F*) $P_>(i)$ is the probability of an exploration trip going outside their origin tiles at level *i* at eight tile levels with tile side length from 24 km to 187 m (the entire Boston Metro area, larger than the area shown in the maps, is set as a 48-km square). Results show the calculation with the multiplicative cascade framework, in the simulation and measured by the mobile phone data.

i). Then, the probability that a trip goes outside its origin tile at resolution level *i*, $P_>(i)$, can be expressed as

$$P_>(i) = \int_1^M P_>(k) f_{S_{i,trip}}(k) dk, \qquad [2]$$

where M is the total number of supplies in the entire region Ω_0; $P_>(k)$ is the probability that the k supplies in the origin tile are not chosen; $f_{S_{i,trip}}(k)$ is the probability of finding k supplies within the origin tile. The tile exceeding probability $P_>(i)$ at different tile resolutions generates the resulting distribution of trip distances. Eq. **2** ties together the rank-based selection mechanism $P_>(k)$ and the geographic distribution of locations $f_{S_{i,trip}}(k)$, which can be calculated as

$$f_{S_{i,trip}}(k) = \int_0^Q f_{D_{i,trip}}(D) f_{S_i|D_i=D}(k) dD, \qquad [3]$$

where $f_{D_{i,trip}}(D)$ is the conditional probability that a trip originates in a tile at level *i* given D demands are in that tile. $f_{S_i|D_i}$ 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 $P_>(i)$, 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 β_1 and β_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 $P_{new} = \rho S^{-\gamma}$. These three parameters are extracted from the aggregated data of the entire population (Fig. 2 *D* and *E*).

The individual values of β_1 and β_2 values are obtained by calibrating the Markov model to minimize the following statistic:

$$A(\beta_1, \beta_2) = \int |P_D(\Delta t) - P_M(\Delta t|\beta_1, \beta_2)| d\Delta t + \eta |\overline{N}_D - \overline{N}_M(\beta_1, \beta_2)|, \qquad [4]$$

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 $\overline{N}_D$ and $\overline{N}_M(\beta_1, \beta_2)$ are the average daily number of visited locations measured from the individual's empirical data and from the model simulation, respectively. The difference between $\overline{N}_D$ and n_w is that $\overline{N}_D$ 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 β_1 and β_2 values are used ($\beta_1 = 1,2,3, \ldots, 20$, $\beta_2 = 1,6,11, \ldots, 101$) to estimate the (β_1, β_2) pair that minimizes $A(\beta_1, \beta_2)$ for each person. The empirical results of $n_w\beta_1$, $n_w\beta_2$, and n_w for all of the individuals are presented in Fig. 2*C*. 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. 4 *A–C* shows three simulated days for the student. The days are predominated 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. 4*D* 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. 4 *E–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 β_1, β_2, and n_w from the joint distribution shown in Fig. 2*C* to generate two synthetic realizations of the user. Fig. 4 *F* and *G* shows the two resulting profiles of simulated journeys of the same sparse user and Fig. 4*H* shows the distinct motif distributions.

The importance of the individual features extracted from data (Fig. 2*C*) lies in their ability to capture diverse travel behaviors observed in the population. Fig. 5 *A* 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, β_1, and β_2. Fig. 5 *A* 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. 5 *C–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

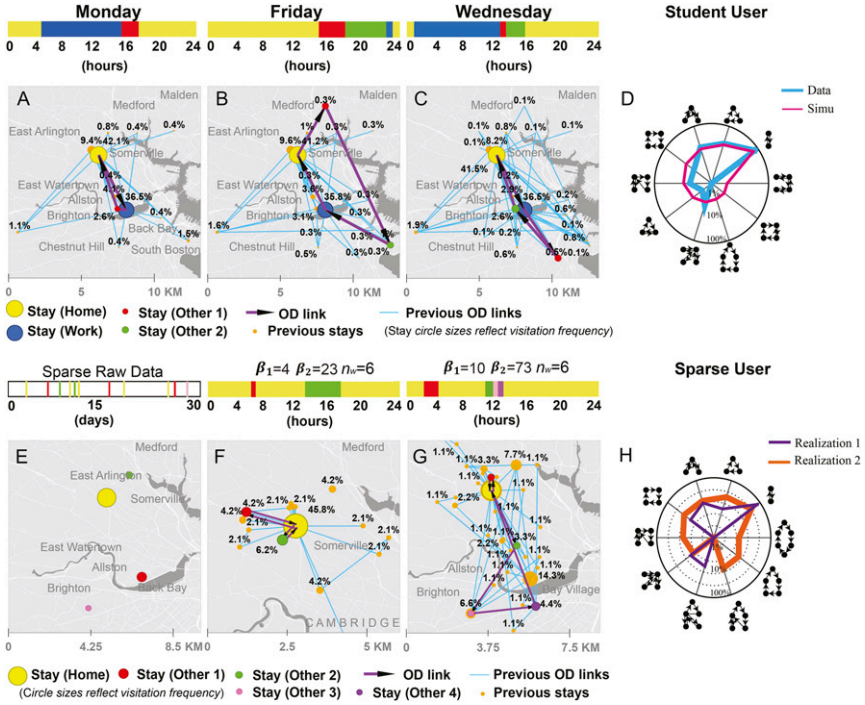

Fig. 4. Simulation of daily trajectories of one active commuter and one sparse user. (*A–C*) Simulated trajectories of the student with self-collected cell phone records. Three sample days are shown here. The trips for each sample day are in purple, and the visitation frequency of each location is calculated until the sample day and represented by the circle sizes. (*D*) Distributions of daily mobility motifs for the active commuter's data vs. simulation. The model captures well the higher propensity of motifs with node sizes 2 and 3 as well as some other occurrences. (*E*) A sample sparse user with 10 stays at 4 distinct locations in an observation period of 30 d. (*F* and *G*) Two different realizations for simulating the same sparse user with different parameter values. The first realization uses $n_w = 6$, $\beta_1 = 4$, $\beta_2 = 23$. The second realization uses $n_w = 6$, $\beta_1 = 10$, $\beta_2 = 73$. Larger values of β_1 and β_2 generate more consecutive out-of-home activities and more daily visited locations. (*H*) Distributions of daily mobility motifs for the two realizations of the same sparse user using different parameter values. With small n_w, $\beta_1 n_w$, and $\beta_2 n_w$ values the person is likely to have simple motifs, whereas large parameter values lead to more complex daily activity chains.

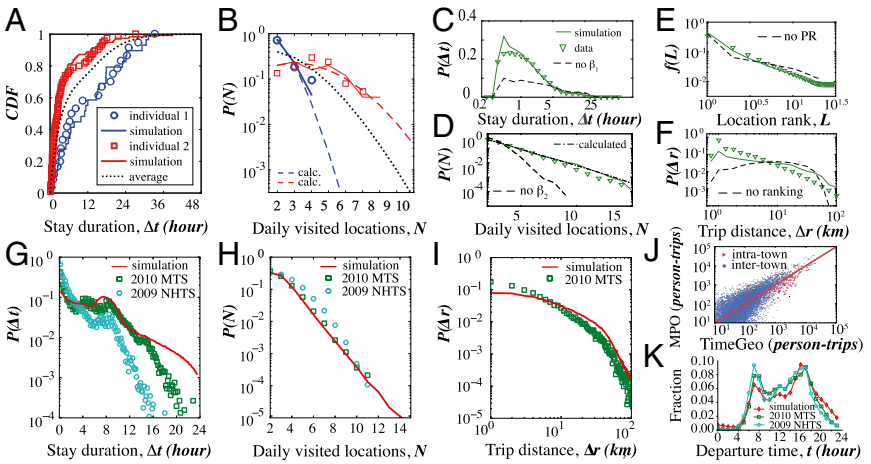

Fig. 5. Mobility patterns for different individuals and population distributions. The top panels (*A*–*F*) show the comparison of the simulation results with the phone data for noncommuters. (*A* and *B*) Comparison of mobility patterns for three representative noncommuters. Individuals 1 and 2 represent two extreme cases, one has shorter stays (shown in squares, $n_w = 10.86$, $\beta_1 = 6$, $\beta_2 = 41$) and the other travels less frequently (shown in circles, $n_w = 5.51$, $\beta_1 = 1$, $\beta_2 = 36$). The third case represents an average noncommuter and is simulated using the median parameter values of n_w, β_1, and β_2. (*A*) Stay duration distribution. (*B*) Daily visited location distribution. The Markov modeling framework allows the calculations of the number of visited locations per day, as shown in dashed lines, and is discussed in *SI Appendix*. (*C*) Activity duration distribution $P(\Delta t)$. The model without differentiating home and other states (setting $\beta_1 = 1$) is considered as a benchmark here. In this case, stays with short duration are underestimated. (*D*) The distribution of the number of daily visited location $P(N)$. Both the model's calculation and simulation results are shown. It shows the need for the β_2 parameter in the model. (*E*) Visitation frequency $f(L)$ to the Lth most visited location follows the form $f(L) \sim L^{-1.2\pm0.1}$. 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 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. S18 shows comparisons for various trip purposes.

proposed in the current modeling framework, namely, the weekly home-based tour number, dwell rate, 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 (β_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 β_2 is 0.03 and 0.22, respectively. The bursts of flexible activities, captured by the dwell and burst rates β_1 and β_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 Lth most visited location follows $f(L) \sim L^{-1.2\pm0.1}$. In Fig. 3 *D*, $P_>(i)$ 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 3 km), the cascade is a pure log-normal cascade, $P_>(i)$ 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, including the 2010–2011 Massachusetts Travel Survey (MTS) and the 2009 National Household Travel Survey (NHTS). When comparing the simulation results with the MTS and NHTS, respectively, the KS statistic for $P(\Delta t)$ is 0.23 and 0.59 (Fig. 5*G*). Note that these stay duration distributions are significantly different among the surveys and our simulation. It is mainly because in the 1-d surveys people rarely report duration of stays longer than 12 h, whereas the active mobile phone users' data records informed our simulation. This range of stays can add up to 20% of the data, as seen in the cumulative distribution of Fig. 5*A*. Besides, the distribution of the daily visited locations $P(N)$ compares well among the simulation and the surveys, as presented in Fig. 5*H*, with the KS statistic of 0.07 and 0.23, respectively. For $P(\Delta r)$, comparing the simulation with the MTS, the KS statistic is 0.24 (Fig. 5*I*). Here the model, which does not consider trip distances in the selection of return locations, overestimates long distance trips. We do not compare with travel distances from the national survey (NHTS), because spatial aspects of travel depend directly on the specific extension of the urban form, which varies across the nation (39).

Fig. 5*J* compares the total number of trips from home to work in our simulation with the estimates of the model developed by the Boston Region Metropolitan Planning Organization (MPO) for 2010 (40). The comparisons of the number of commuting trips are presented both for those between the 164 cities and towns in the metropolitan area (intertown) and for trips within them (intratown). The results for commuting trips are excellent, with a Pearson correlation coefficient of 0.90 and 0.99, respectively. More differences are present in the trips from home to other locations and between other types of locations. Finally, Fig. 5*K* compares the fraction of trips being initiated at different times of the day among our simulation, the 2009 NHTS, and the 2010 MTS. Although the total estimates compare well, we estimate more trips between nonhome destinations in the evening

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 home-based tour number, dwell rate, and burst rate to model the bursts of short flexible activities in activity chains. This mechanism can reproduce individual distributions of stay duration, number of daily visited locations, and daily mobility motif distribution. Spatially, an improved rank-based EPR model is introduced to explain individual activity location selection choices. Compared with the original EPR model, the ranking mechanism quantifies the likelihood of selecting new destinations in space based on the distribution of facilities around trip origins. Moreover, the covariance of the distributions of 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, ρ, 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, as 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 $\omega \subset \Omega_0$, $D(\omega)$ is the number of trip origins in ω and $S(\omega)$ 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 ω, 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. 3*C*. The demand and supply in a generic i-tile Ω_i are D_i and S_i and the associated measure densities are $D_i = D_i/|\Omega_i|$ and $S_i = S_i/|\Omega_i|$. One starts with uniform measure densities D_0' and S_0' in Ω_0, then progressively partitions Ω_0 into 4^1, 4^2, …, 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_{D_i} and W_{S_i}, with mean value 1. The random vectors $W_i = [W_{D_i}, W_{S_i}]$, $i = 1,2, \ldots, n$ 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 $i-1$ and i are related as

$$\begin{bmatrix} D_i' \\ S_i' \end{bmatrix} \overset{d}{=} \begin{bmatrix} W_{D_i} & 0 \\ 0 & W_{S_i} \end{bmatrix} \begin{bmatrix} D_{i-1}' \\ S_{i-1}' \end{bmatrix}. \quad [5]$$

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 $\sigma^2_{W_{D_i}}$ and $\sigma^2_{W_{S_i}}$, mean values $-1/2\sigma^2_{W_{D_i}}$ and $-1/2\sigma^2_{W_{S_i}}$, and correlation coefficient ρ_{LN_i}, then $\ln(D_i)$ and $\ln(S_i)$ have joint normal distribution with mean values m_{D_i} and m_{S_i}, variances $\sigma^2_{D_i}$ and $\sigma^2_{S_i}$, and correlation coefficient ρ_i given by

$$\sigma^2_{D_i} = \sum_{j=1}^{i} \sigma^2_{W_{D_j}}, m_{D_i} = \ln\left(D_0 4^{-i}\right) - 1/2\sigma^2_{D_i}, \quad [6]$$

$$\sigma^2_{S_i} = \sum_{j=1}^{i} \sigma^2_{W_{S_j}}, m_{S_i} = \ln\left(S_0 4^{-i}\right) - 1/2\sigma^2_{S_i}, \quad [7]$$

$$\rho_i = \sum_{j=1}^{i} \frac{\rho_{LN_j} \sigma_{W_{D_j}} \sigma_{W_{S_j}}}{\sigma_{D_i} \sigma_{S_i}}. \quad [8]$$

Therefore, once we can estimate $\sigma_{W_{D_i}}$, $\sigma_{W_{S_i}}$, and ρ_{LN_j}, the rest of the variables can be calculated.

At smaller tile sizes, empty tiles cannot be ignored and extreme forms of dependence like mutual exclusion may occur. In this case the generator is better modeled as a β-cascade, in which a tile is either filled or empty. The generators $W(i) = [W_{D(i)}, W_{S(i)}]$ of a bivariate β-cascade have a discrete distribution with probability masses concentrated at four (w_D, w_S) points: mass P_{00} at (0,0), mass P_{D0} at $(1/P_D, 0)$, mass P_{0S} at $(0, 1/P_S)$, and mass P_{DS} at $(1/P_D, 1/P_S)$. $P_D = P_{D0} + P_{DS}$, $P_S = P_{0S} + P_{DS}$, and $P_{D0} + P_{DS} + P_{0S} + P_{00} = 1$. Thus, a tile is either filled or empty. The correlation between the supply and demand is ρ_{β_i}.

ACKNOWLEDGMENTS. We thank Chaoming Song for enlightening discussions during the design of this work. The research reported herein was funded in part by the MIT–Ford Alliance, MIT–Philips Alliance, the MIT-Brazil Program, the MIT-Portugal Program, the Samuel Tak Lee Real Estate Entrepreneurship Laboratory at MIT, US Department of Transportation via the program New England University Transportation Center (UTC) Year 25, and the Center for Complex Engineering Systems at King Abdulaziz City for Science and Technology (KACST).

1. Goodchild MF (2007) Citizens as sensors: The world of volunteered geography. *GeoJournal* 69(4):211–221.
2. Batty M (2013) *The New Science of Cities* (MIT Press, Cambridge, MA).
3. Nagel K, Beckman RJ, Barrett CL (1999) Transims for urban planning. *6th International Conference on Computers in Urban Planning and Urban Management, Venice, Italy* (Los Alamos National Laboratory, Los Alamos, NM). Available at https://www.researchgate.net/publication/243768002. Accessed August 6, 2016.
4. Ben-Akiva M, Bierlaire M (1999) Discrete choice methods and their applications to short term travel decisions. *Handbook of Transportation Science* (Springer, New York), pp 5–33.
5. Balmer M, et al. (2008) Agent-based simulation of travel demand: Structure and computational performance of MATSim-T. *2nd TRB Conference on Innovations in Travel Modeling, Portland, Oregon* (Eidgenössische Technische Hochschule Zürich, Zurich).
6. Arentze T, Timmermans H (2000) *Albatross: A Learning Based Transportation Oriented Simulation System* (EIRASS, Eindhoven, The Netherlands).
7. Bowman JL, Ben-Akiva ME (2001) Activity-based disaggregate travel demand model system with activity schedules. *Transp Res Part A Policy Pract* 35(1):1–28.
8. Danalet A, Tinguely L, de Lapparent M, Bierlaire M (2016) *Location Choice with Longitudinal WiFi Data* Location choice with longitudinal WiFi data. *Journal of Choice Modelling* 18:1–17.
9. Zilske M, Nagel K (2014) Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. *Procedia Comput Sci* 32:802–807.
10. Zilske M, Nagel K (2015) A simulation-based approach for constructing all-day travel chains from mobile phone data. *Proc Comput Sci* 52:468–475.
11. Zheng Y, Capra L, Wolfson O, Yang H (2014) Urban computing: Concepts, methodologies, and applications. *ACM Trans Intell Syst Technol* 5(3):38.
12. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. arXiv:1502.03406.
13. González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. *Nature* 453(7196):779–782.
14. Song C, Koren T, Wang P, Barabási AL (2010) Modelling the scaling properties of human mobility. *Nat Phys* 6(10):818–823.
15. Perkins TA, et al. (2014) Theory and data for simulating fine-scale human movement in an urban environment. *J R Soc Interface* 11(99):20140642.
16. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. *Science* 327(5968):1018–1021.
17. Hasan S, Schneider CM, Ukkusuri SV, González MC (2013) Spatiotemporal patterns of urban human mobility. *J Stat Phys* 151(1-2):304–318.
18. Toole JL, Herrera-Yaqüe C, Schneider CM, González MC (2015) Coupling human mobility and social ties. *J R Soc Interface* 12(105):20141128.
19. Schneider CM, Belik V, Couronné T, Smoreda Z, González MC (2013) Unravelling daily human mobility motifs. *J R Soc Interface* 10(84):20130246.
20. Kölbl R, Helbing D (2003) Energy laws in human travel behaviour. *New J Phys* 5(1):48.
21. Balcan D, et al. (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc Natl Acad Sci USA* 106(51):21484–21489.
22. Viswanathan G, et al. (1996) Lévy flight search patterns of wandering albatrosses. *Nature* 381(6581):413–415.
23. Jiang S, et al. (2013) A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13 (ACM, New York), pp 2:1–2:9.
24. Toole JL, et al. (2015) The path most traveled: Travel demand estimation using big data resources. *Transp Res, Part C Emerg Technol* 58(B):162–177.
25. Alexander L, Jiang S, Murga M, González MC (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transp Res, Part C Emerg Technol* 58:240–250.
26. Vázquez A, et al. (2006) Modeling bursts and heavy tails in human dynamics. *Phys Rev E Stat Nonlin Soft Matter Phys* 73(3 Pt 2):036127.
27. Barabási AL (2005) The origin of bursts and heavy tails in human dynamics. *Nature* 435(7039):207–211.
28. Hidalgo R, César A (2006) Conditions for the emergence of scaling in the inter-event time of uncorrelated and seasonal systems. *Physica A* 369(2):877–883.
29. Malmgren RD, Stouffer DB, Motter AE, Amaral LA (2008) A Poissonian explanation for heavy tails in e-mail communication. *Proc Natl Acad Sci USA* 105(47):18153–18158.
30. Karsai M, Kaski K, Barabási AL, Kertész J (2012) Universal features of correlated bursty behaviour. *Sci Rep* 2:397.
31. Jo HH, Karsai M, Kertész J, Kaski K (2012) Circadian pattern and burstiness in mobile phone communication. *New J Phys* 14(1):013055.
32. Pappalardo L, et al. (2015) Returners and explorers dichotomy in human mobility. *Nat Commun* 6:8166.
33. Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. *Nature* 484(7392):96–100.
34. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: Universal patterns in human urban mobility. *PLoS One* 7(5):e37027.
35. Yang Y, Herrera C, Eagle N, González MC (2014) Limits of predictability in commuting flows in the absence of data for calibration. *Sci Rep* 4:5662.
36. Noulas A, Shaw B, Lambiotte R, Mascolo C (2015) Topological properties and temporal dynamics of place networks in urban environments. *Proceedings of the 24th International Conference on World Wide Web Companion* (International World Wide Web Conferences Steering Committee, New York), pp 431–441.
37. Batty M (2008) The size, scale, and shape of cities. *Science* 319(5864):769–771.
38. Veneziano D, Gonzalez MC (2010) Trip length distribution under multiplicative spatial models of supply and demand: Theory and sensitivity analysis. arXiv:1101.3719.
39. Newman PG, Kenworthy JR (1989) *Cities and Automobile Dependence: An International Sourcebook* (Gower, Aldershot, UK), pp 1–388.
40. CTPS (2013) Methodology and assumptions of central transportation planning staff regional travel demand modeling. Available at www.ctps.org/Drupal/data/pdf/about/mpo/recert_2014/CTPS_GLE_Modeling_Method_20130416.pdf. Accessed March 17, 2016.

KDD '17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Pages 1357-1366

Using Convolutional Networks and Satellite Imagery to Identify Patterns in Urban Environments at a Large Scale

Adrian Albert*
Massachusetts Institute of Technology
Civil and Environmental Engineering
77 Massachusetts Ave
Cambridge, MA 02139
adalbert@mit.edu

Jasleen Kaur
Philips Lighting Research North America
2 Canal Park
Cambridge, MA 02141
jasleen.kaur1@philips.com

Marta C. González
Massachusetts Institute of Technology
Civil and Environmental Engineering
77 Massachusetts Ave
Cambridge, MA 02139
martag@mit.edu

ABSTRACT

Urban planning applications (energy audits, investment, etc.) require an understanding of built infrastructure and its environment, i.e., both low-level, physical features (amount of vegetation, building area and geometry etc.), as well as higher-level concepts such as land use classes (which encode expert understanding of socio-economic end uses). This kind of data is expensive and labor-intensive to obtain, which limits its availability (particularly in developing countries). We analyze patterns in land use in urban neighborhoods using large-scale satellite imagery data (which is available worldwide from third-party providers) and state-of-the-art computer vision techniques based on deep convolutional neural networks. For supervision, given the limited availability of standard benchmarks for remote-sensing data, we obtain ground truth land use class labels carefully sampled from open-source surveys, in particular the Urban Atlas land classification dataset of 20 land use classes across 300 European cities. We use this data to train and compare deep architectures which have recently shown good performance on standard computer vision tasks (image classification and segmentation), including on geospatial data. Furthermore, we show that the deep representations extracted from satellite imagery of urban environments can be used to compare neighborhoods across several cities. We make our dataset available for other machine learning researchers to use for remote-sensing applications.

CCS CONCEPTS

•Computing methodologies →Computer vision; Neural networks; •Applied computing →Environmental sciences;

KEYWORDS

Satellite imagery, land use classification, convolutional networks

1 INTRODUCTION

Land use classification is an important input for applications ranging from urban planning, zoning and the issuing of business permits, to real-estate construction and evaluation to infrastructure development. Urban land use classification is typically based on surveys performed by trained professionals. As such, this task is labor-intensive, infrequent, slow, and costly. As a result, such data are mostly available in developed countries and big cities that have the resources and the vision necessary to collect and curate it; this information is usually not available in many poorer regions, including many developing countries [9] where it is mostly needed.

This paper builds on two recent trends that promise to make the analysis of urban environments a more democratic and inclusive task. On the one hand, recent years have seen significant improvements in satellite technology and its deployment (primarily through commercial operators), which allows to obtain high and medium-resolution imagery of most urbanized areas of the Earth with an almost daily revisit rate. On the other hand, the recent breakthroughs in computer vision methods, in particular deep learning models for image classification and object detection, now make possible to obtain a much more accurate representation of the composition built infrastructure and its environments.

Our contributions are to both the applied deep learning literature, and to the incipient study of "smart cities" using remote sensing data. We contrast state-of-the-art convolutional architectures (the VGG-16 [19] and ResNet [7] networks) to train classifiers that recognize broad land use classes from satellite imagery. We then use the features extracted from the model to perform a large-scale comparison of urban environments. For this, we construct a novel dataset for land use classification, pairing carefully sampled locations with ground truth land use class labels obtained from the Urban Atlas survey [22] with satellite imagery obtained from Google Maps's static API. Our dataset - which we have made available publicly for other researchers - covers, for now, 10 cities in Europe (chosen out of the original 300) with 10 land use classes (from the original 20). As the Urban Atlas is a widely-used, standardized dataset for land use classification, we hope that making this dataset available will encourage the development analyses and algorithms for analyzing the built infrastructure in urban environments. Moreover, given that satellite imagery is available virtually everywhere on the globe, the methods presented here allow for automated, rapid classification of urban environments that can potentially be applied to locations where survey and zoning data is not available.

Land use classification refers to the combination of physical land attributes and what cultural and socio-economic function land serves (which is a subjective judgement by experts) [2]. In this paper, we take the view that land use classes are just a useful discretization of a more continuous spectrum of patterns in the organization of urban environments. This viewpoint is illustrated in Figure 2: while

*Corresponding author.

KDD'17, August 13–17, 2017, Halifax, NS, Canada.

DOI: http://dx.doi.org/10.1145/3097983.3098070

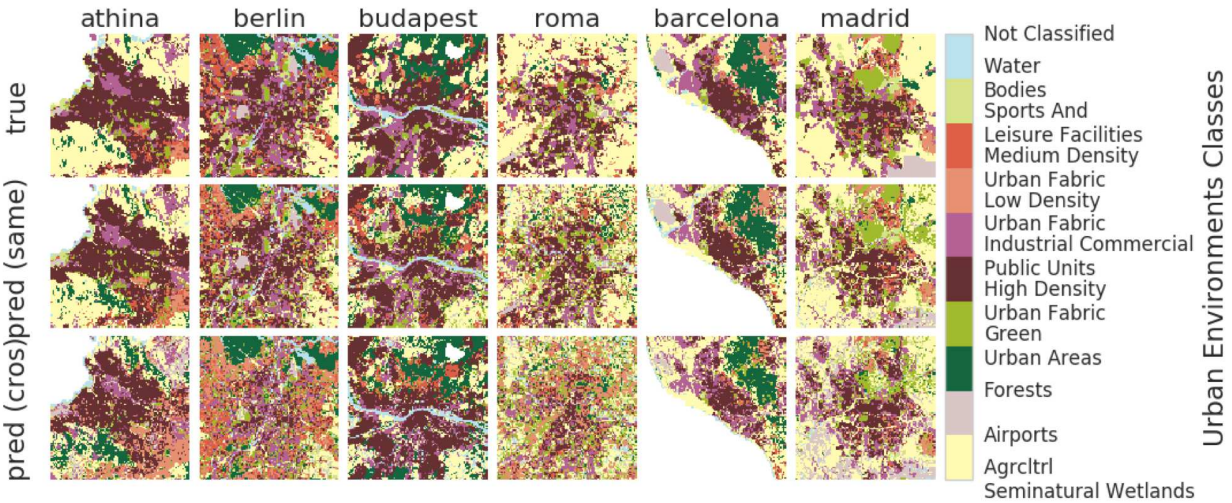

Figure 1: Urban land use maps for six example cities. We compare the ground truth (*top row*) with the predicted land use maps, either from using separate data collected from the same city (*middle row*), or using data from all other cities (*bottom row*).

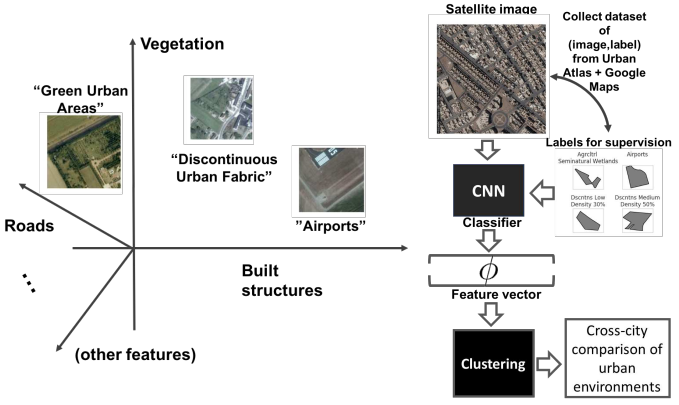

Figure 2: *Left:* Comparing urban environments via deep hierarchical representations of satellite image samples. *Right:* approach outline - data collection, classification, feature extraction, clustering, validation.

some attributes (e.g., amount of built structures or vegetation) are directly interpretable, some others may not be. Nevertheless, these patterns influence, and are influenced by, socio-economic factors (e.g., economic activity), resource use (energy), and dynamic human behavior (e.g., mobility, building occupancy). We see the work on cheaply curating a large-scale land use classification dataset and comparing neighborhoods using deep representations that this paper puts forth as a necessary first step towards a granular understanding of urban environments in data-poor regions.

Subsequently, in Section 2 we review related studies that apply deep learning methods and other machine learning techniques to problems of land use classification, object detection, and image segmentation in aerial imagery. In Section 3 we describe the dataset we curated based on the Urban Atlas survey. Section 4 reviews the deep learning architectures we used. Section 5 describes model validation and analysis results. We conclude in Section 6.

All the code used to acquire, process, and analyze the data, as well as to train the models discussed in this paper is available at http://www.github.com/adrianalbert/urban-environments.

2 LITERATURE

The literature on the use of remote sensing data for applications in land use cover, urban planning, environmental science, and others, has a long and rich history. This paper however is concerned more narrowly with newer work that employs widely-available data and machine learning models - and in particular deep learning architectures - to study urban environments.

Deep learning methods have only recently started to be deployed to the analysis of satellite imagery. As such, land use classification using these tools is still a very incipient literature. Probably the first studies (yet currently only 1-2 years old) include the application of convolutional neural networks to land use classification [2] using the UC Merced land use dataset [25] (of 2100 images spanning 21 classes) and the classification of agricultural images of coffee plantations [17]. Similar early studies on land use classification that employ deep learning techniques are [21], [18], and [15]. In [11], a spatial pyramid pooling technique is employed for land use classification using satellite imagery. The authors of these studies adapted architectures pre-trained to recognize natural images from the ImageNet dataset (such as the VGG16 [19], which we also use), and fine-tuned them on their (much smaller) land use data. More recent studies use the DeepSat land use benchmark dataset [1], which we also use and describe in more detail in Section 2.1. Another topic that is closely related to ours is remote-sensing image segmentation and object detection, where modern deep learning models have also started to be applied. Some of the earliest work that develops and applies deep neural networks for this tasks is that

of [13]. Examples of recent studies include [26] and [12], where the authors propose a semantic image segmentation technique combining texture features and boundary detection in an end-to-end trainable architecture.

Remote-sensing data and deep learning methods have been put to use to other related ends, e.g., geo-localization of ground-level photos via satellite images [3, 24] or predicting ground-level scene images from corresponding aerial imagery [27]. Other applications have included predicting survey estimates on poverty levels in several countries in Africa by first learning to predict levels of night lights (considered as proxies of economic activity and measured by satellites) from day-time, visual-range imagery from Google Maps, then transferring the learning from this latter task to the former [9]. Our work takes a similar approach, in that we aim to use remote-sensing data (which is widely-available for most parts of the world) to infer land use types in those locations where ground truth surveys are not available.

Urban environments have been analyzed using other types of imagery data that have become recently available. In [4, 14], the authors propose to use the same type of imagery from Google Street View to measure the relationship between urban appearance and quality of life measures such as perceived safety. For this, they hand-craft standard image features widely used in the computer vision community, and train a shallow machine learning classifier (a support vector machine). In a similar fashion, [5] trained a convolutional neural network on ground-level Street View imagery paired with a crowd-sourced mechanism for collecting ground truth labels to predict subjective perceptions of urban environments such as "beauty", "wealth", and "liveliness".

Land use classification has been studied with other new data sources in recent years. For example, ground-level imagery has been employed to accurately predict land use classes on an university campus [28]. Another related literature strand is work that uses mobile phone call records to extract spatial and temporal mobility patterns, which are then used to infer land use classes for several cities [6, 10, 20]. Our work builds on some of the ideas for sampling geospatial data presented there.

2.1 Existing land use benchmark datasets

Public benchmark data for land use classification using aerial imagery are still in relatively short supply. Presently there are two such datasets that we are aware of, discussed below.

UC Merced. This dataset was published in 2010 [25] and contains 2100 256×256, $1m/px$ aerial RGB images over 21 land use classes. It is considered a "solved problem", as modern neural network based classifiers [2] have achieved > 95% accuracy on it.

DeepSat. The DeepSat [1] dataset[1] was released in 2015. It contains two benchmarks: the *Sat-4* data of 500,000 images over 4 land use classes (*barren land, trees, grassland, other*), and the *Sat-6* data of 405,000 images over 6 land use classes (*barren land, trees, grassland, roads, buildings, water bodies*). All the samples are 28×28 in size at a $1m/px$ spatial resolution and contain 4 channels (red, green, blue, and NIR - near infrared). Currently less than two years old, this dataset is already a "solved problem", with previous studies [15] (and our own experiments) achieving classification accuracies of over 99% using convolutional architectures. While useful as input for pre-training more complex models, (e.g., image segmentation), this dataset does not allow to take the further steps for detailed land use analysis and comparison of urban environments across cities, which gap we hope our dataset will address.

Other open-source efforts. There are several other projects that we are aware of related to land use classification using open-source data. The TerraPattern[2] project uses satellite imagery from Google Maps (just like we do) paired with truth labels over a large number (450) of detailed classes obtained using the Open Street Map API[3]. (Open Street Maps is a comprehensive, open-access, crowd-sourced mapping system.) The project's intended use is as a search tool for satellite imagery, and as such, the classes they employ are very specific, e.g., baseball diamonds, churches, or roundabouts. The authors use a ResNet architecture [7] to train a classification model, which they use to embed images in a high-dimensional feature space, where "similar" images to an input image can be identified. A second open-source project related to ours is the DeepOSM[4], in which the authors take the same approach of pairing OpenStreetMap labels with satellite imagery obtained from Google Maps, and use a convolutional architecture for classification. These are excellent starting points from a practical standpoint, allowing interested researchers to quickly familiarize themselves with programming aspects of data collection, API calls, etc.

3 THE URBAN ENVIRONMENTS DATASET

3.1 Urban Atlas: a standard in land use analysis

The Urban Atlas [22] is an open-source, standardized land use dataset that covers ~ 300 European cities of 100,000 inhabitants or more, distributed relatively evenly across major geographical and geopolitical regions. The dataset was created between 2005-2011 as part of a major effort by the European Union to provide a uniform framework for the geospatial analysis of urban areas in Europe. Land use classification is encoded via detailed polygons organized in commonly-used GIS/ESRI shape files. The dataset covers 20 standardized land use classes. In this work we selected classes of interest and consolidated them into 10 final classes used for analysis (see Figure 3). Producing the original Urban Atlas dataset required fusing several data sources: high and medium-resolution satellite imagery, topographic maps, navigation and road layout data, and local zoning (cadastral) databases. More information on the methodology used by the Urban Atlas researchers can be obtained from the European Environment Agency[5]. We chose expressly to use the Urban Atlas dataset over other sources (described in Section 2.1 because *i)* it is a comprehensive and consistent survey at a large scale, which has been extensively curated by experts and used in research, planning, and socio-economic work over the past decade, and *ii)* the land use classes reflect higher-level (socio-economic, cultural) functions of the land as used in applications.

We note that there is a wide variance in the distribution of land use classes across and within the 300 cities. Figure 3 illustrates the differences in the distribution in ground truth polygon areas

[1] Available at http://csc.lsu.edu/~saikat/deepsat/.

[2] http://www.terrapattern.com/

[3] http://www.openstreetmap.org

[4] https://github.com/trailbehind/DeepOSM

[5] http://www.eea.europa.eu/data-and-maps/data/urban-atlas/

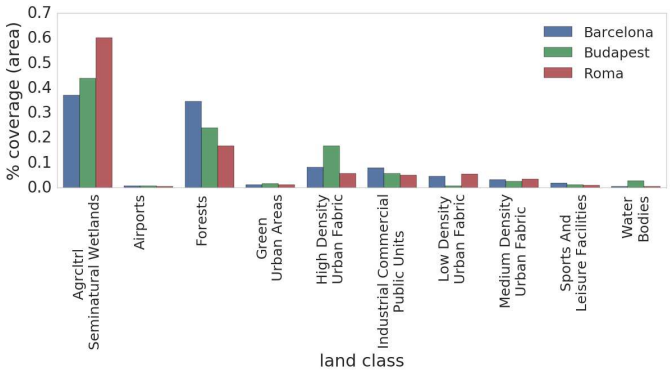

Figure 3: Ground truth land use distribution (by area) for three example cities in the Urban Environments dataset.

for each of the classes for three example cities (Budapest, Rome, Barcelona) from the dataset (from Eastern, Central, and Western Europe, respectively). This wide disparity in the spatial distribution patterns of different land use classes and across different cities motivates us to design a careful sampling procedure for collecting training data, described in detail below.

3.2 Data sampling and acquisition

We set out to develop a strategy to obtain high-quality samples of the type (satellite image, ground truth label) to use in training convolutional architectures for image classification. Our first requirement is to do this solely with freely-available data sources, as to keep costs very low or close to zero. For this, we chose to use the Google Maps Static API[6] as a source of satellite imagery. This service allows for 25,000 API requests/day free of charge. For a given sampling location given by (latitude, longitude), we obtained 224×224 images at a zoom level 17 (around $1.20m/px$ spatial resolution, or $\sim 250m \times 250m$ coverage for an image).

The goals of our sampling strategy are twofold. First, we want to ensure that the resulting dataset is as much as possible balanced with respect to the land use classes. The challenge is that the classes are highly imbalanced among the ground truth polygons in the dataset (e.g., many more polygons are agricultural land and isolated structures than airports). Second, the satellite images should be representative of the ground truth class associated to them. To this end, we require that the image contain at least 25% (by area) of the associated ground truth polygon. Thus, our strategy to obtain training samples is as follows (for a given city):

- Sort ground truth polygons in decreasing order according to their size, and retain only those polygons with areas larger than $\frac{1}{4}(224 \times 1.2m)^2 = 0.06km^2$;
- From each decile of the distribution of areas, sample a proportionally larger number of polygons, such that some of the smaller polygons also are picked, and more of the larger ones;
- For each picked polygon, sample a number of images proportional to the area of the polygon, and assign each image the polygon class as ground truth label;

[6] https://developers.google.com/maps/documentation/static-maps/

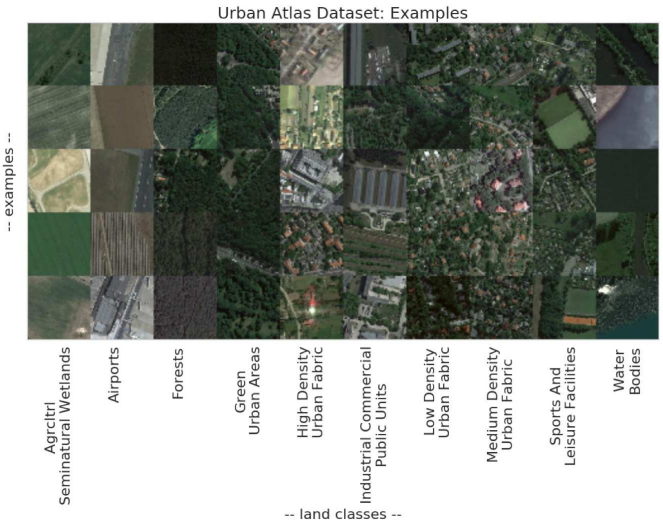

Figure 4: Example satellite images for the original land use classes in the Urban Atlas dataset.

Example satellite images for each of the 10 land use classes in the Urban Environments dataset are given in Figure 4. Note the significant variety (in color schemes, textures, etc) in environments denoted as having the same land use class. This is because of several factors, including the time of the year when the image was acquired (e.g., agricultural lands appear different in the spring than in the fall), the different physical form and appearance of environments that serve the same socioeconomic or cultural function (e.g., green urban areas may look very different in different cities or in even in different parts of the same city; what counts as "dense urban fabric" in one city may not be dense at all in other cities), and change in the landscape during the several years that have passed since the compilation of the Urban Atlas dataset and the time of acquisition of the satellite image (e.g., construction sites may not reflect accurately anymore the reality on the ground).

Apart from these training images, we constructed ground truth rasters to validate model output for each city. For that, we defined uniform validation grids of 100×100 ($25km \times 25km$) around the (geographical) center of a given city of interest. We take a satellite image sample in each grid cell, and assign to it as label the class of the polygon that has the maximum intersection area with that cell. Examples of land use maps for the six cities we analyze here are given in Figure 1 (top row). There, each grid cell is assigned the class of the ground truth polygon whose intersection with the cell has maximum coverage fraction by area. Classes are color-coded following the original Urban Atlas documentation.

In Table 1 we present summaries of the training (left) and validation (right) datasets we used for the analysis in this paper. The validation dataset consists of the images sampled at the centers of each cell in the $25km \times 25km$ grid as discussed above. This dataset consists of $\sim 140,000$ images distributed across 10 urban environment classes from 6 cities: Roma (Rome), Madrid, Berlin, Budapest, Barcelona, and Athina (Athens). Because of the high variation in appearance upon visual inspection, we chose to consolidate several

classes from the original dataset, in particular classes that indicated urban fabric into "High Density Urban Fabric", "Medium Density Urban Fabric, and "Low Density Urban Fabric". As mentioned above and illustrated in Figure 3, we did notice a great disparity in the numbers and distribution of ground truth polygons for other example cities that we investigated in the Urban Atlas dataset. As such,for the analysis in this paper, we have chosen cities where enough ground truth polygons were available for each class (that is, at least 50 samples) to allow for statistical comparisons.

4 EXPERIMENTAL SETUP

4.1 Neural network architectures and training

For all experiments in this paper we compared the VGG-16 [19] and ResNet [7, 8] architectures.

VGG-16. This architecture [19] has become one of the most popular models in computer vision for classification and segmentation tasks. It consists of 16 trainable layers organized in blocks. It starts with a 5-block convolutional base of neurons with 3×3 receptive fields (alternated with max-pooling layers that effectively increase the receptive field of neurons further downstream). Following each convolutional layer is a ReLU activation function [19]. The feature maps thus obtained are fed into a set of fully-connected layers (a deep neural network classifier). See Table 2 for a summary.

ResNet. This architecture [7, 8] has achieved state-of-the-art performance on image classification on several popular natural image benchmark datasets. It consists of blocks of convolutional layers, each of which is followed by a ReLU non-linearity. As before, each block in the convolutional base is followed by a max-pooling operation. Finally, the output of the last convolutional layer serves as input feature map for a fully-connected layer with a softmax activation function. The key difference in this architecture is that *shortcut* connections are implemented that skip blocks of convolutional layers, allowing the network to learn residual mappings between layer input and output. Here we used an implementation with 50 trainable layers per [7]. See Table 3 for a summary.

Transfer learning. As it is common practice in the literature, we have experimented with training our models on the problem of interest (urban environment classification) starting from architectures pre-trained on datasets from other domains (*transfer learning*). This procedure has been shown to yield both better performance and faster training times, as the network already has learned to recognize basic shapes and patterns that are characteristic of images across many domains (e.g., [9, 12, 15]). We have implemented the following approaches: *1)* we used models pre-trained on the ImageNet dataset, then further fine-tuned them on the Urban Atlas dataset; and *2)* we pre-trained on the DeepSat dataset (See Section 2), then further refined on the Urban Atlas dataset. As expected, the latter strategy - first training a model (itself pre-trained on ImageNet data) on the DeepSat benchmark, and the further refining on the Urban Atlas dataset - yielded the best results, achieving increases of around 5% accuracy for a given training time.

Given the large amount of variation in the visual appearance of urban environments across different cities (because of different climates, different architecture styles, various other socio-economic factors), it is of interest to study to what extent a model learned on one geographical location can be applied to a different geographical location. As such, we perform experiments in which we train a model for one (or more) cities, then apply the model to a different set of cities. Intuitively, one would expect that, the more neighborhoods and other urban features at one location are similar to those at a different location, the better learning would transfer, and the higher the classification accuracy obtained would be. Results for these experiments are summarized in Figure 6.

4.2 Comparing urban environments

We next used the convolutional architectures to extract features for validation images. As in other recent studies (e.g., [9]), we use the last layer of a network as feature extractor. This amounts to feature vectors of $D = 4096$ dimensions for the VGG16 architecture and $D = 2048$ dimensions for the ResNet-50 architecture. The codes $x \in \mathbb{R}^D$ are the image representations that either network derives as most representative to discriminate the high-level land use concepts it is trained to predict.

We would like to study how "similar" different classes of urban environments are across two example cities (here we picked Berlin and Barcelona, which are fairly different from a cultural and architectural standpoint). For this, we focus only on the $25km \times 25km$, 100×100-cell grids around the city center as in Figure 1. To be able to quantify similarity in local urban environments, we construct a KD-tree $\mathcal{T}$ (using a high-performance implementation available in the Python package `scikit-learn` [16]) using all the gridded samples. This data structure allows to find k-nearest neighbors of a query image in an efficient way. In this way, the feature space can be probed in an efficient way.

5 RESULTS AND DISCUSSION

In Figure 1 we show model performance on the 100×100 ($25km \times 25km$) raster grids we used for testing. The top row shows ground truth grids, where the class in each cell was assigned as the most prevalent land use class by area (see also Section 3). The bottom row shows model predictions, where each cell in a raster is painted in the color corresponding to the maximum probability class estimated by the model (here ResNet-50). Columns in the figure show results for each of the 6 cities we used in our dataset. Even at a first visual inspection, the model is able to recreate from satellite imagery qualitatively the urban land use classification map.

Further, looking at the individual classes separately and the confidence of the model in its predictions (the probability distribution over classes computed by the model), the picture is again qualitatively very encouraging. In Figure 5 we show grayscale raster maps encoding the spatial layout of the class probability distribution for one example city, Barcelona. Particularly good qualitative agreement is observed for agricultural lands, water bodies, industrial, public, and commercial land, forests, green urban areas, low density urban fabric, airports, and sports and leisure facilities. The model appears to struggle with reconstructing the spatial distribution of roads, which is not unexpected, given that roads typically appear in many other scenes that have a different functional classification for urban planning purposes.

Table 1: Urban Environments dataset: sample size summary.

(a) Dataset used for training & validation (80% and 20%, respectively)

class/city	athina	barcelona	berlin	budapest	madrid	roma	class total
Agricultural + Semi-natural areas + Wetlands	4347	2987	7602	2211	4662	4043	25852
Airports	382	452	232	138	124	142	1470
Forests	1806	2438	7397	1550	2685	2057	17933
Green urban areas	990	722	1840	1342	1243	1401	7538
High Density Urban Fabric	967	996	8975	6993	2533	3103	23567
Industrial, commercial, public, military and pr...	1887	2116	4761	1850	3203	2334	16151
Low Density Urban Fabric	1424	1520	2144	575	2794	3689	12146
Medium Density Urban Fabric	2144	1128	6124	1661	1833	2100	14990
Sports and leisure facilities	750	1185	2268	1305	1397	1336	8241
Water bodies	537	408	1919	807	805	619	5095
city total	15234	13952	43262	18432	21279	20824	132983

(b) $25km \times 25km$ ground truth test grids (fractions of city total)

class / city	athina	barcelona	berlin	budapest	madrid	roma
Agricultural + Semi-natural areas + Wetlands	0.350	0.261	0.106	0.181	0.395	0.473
Airports	0.003	0.030	0.013	0.000	0.044	0.006
Forests	0.031	0.192	0.087	0.211	0.013	0.019
Green urban areas	0.038	0.030	0.072	0.027	0.125	0.054
High Density Urban Fabric	0.389	0.217	0.284	0.365	0.170	0.215
Industrial, commercial, public, military and pr...	0.109	0.160	0.190	0.096	0.138	0.129
Low Density Urban Fabric	0.016	0.044	0.012	0.006	0.036	0.029
Medium Density Urban Fabric	0.041	0.025	0.129	0.045	0.042	0.047
Sports and leisure facilities	0.017	0.034	0.080	0.025	0.036	0.025
Water bodies	0.005	0.006	0.026	0.044	<0.001	0.004

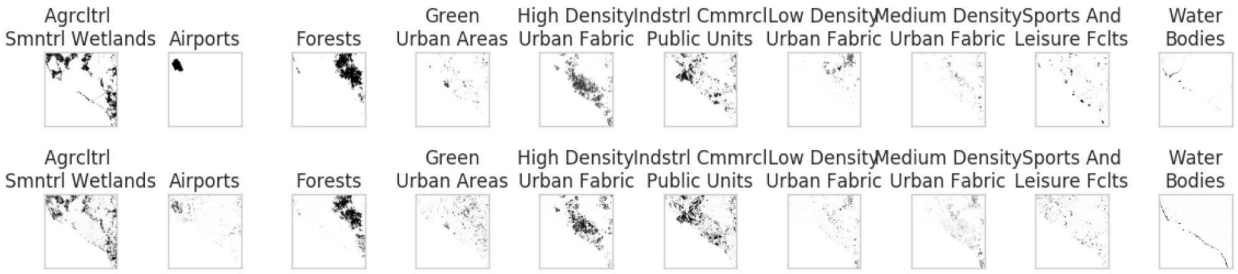

Figure 5: Barcelona: ground truth (*top*) and predicted probabilities (*bottom*).

Table 2: The VGG16 architecture [19].

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Conv(3,64)	Conv(3,128)	Conv(3,256)	Conv(3,512)	Conv(3,512)	FC(4096)
Conv(3,64)	Conv(3,128)	Conv(3,256)	Conv(3,512)	Conv(3,512)	FC(4096)
Max-Pool(2,2)	Max-Pool(2,2)	Conv(3,256) Max-Pool(2,2)	Conv(3,512) Max-Pool(2,2)	Conv(3,512) Max-Pool(2,2)	FC(N_{classes}) SoftMax

Table 3: The ResNet-50 architecture [7].

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Conv(7,64) Max-Pool(3,2)	3x[Conv(1,64) Conv(3,64) Conv(3,256)]	4x[Conv(1,128) Conv(3,128) Conv(1,512)]	6x[Conv(1,256) Conv(3,256) Conv(1,1024)]	3x[Conv(1,512) Conv(3,512) Conv(1,2048)]	FC(N_{classes}) SoftMax

5.1 Classification results

We performed experiments training the two architectures described in Section 4 on datasets for each of the 6 cities considered, and for a combined dataset (`all`) of all the cities. The diagonal in Figure 6 summarizes the (validation set) classification performance for each model. All figures are averages computed over balanced subsets of 2000 samples each. While accuracies or $\sim 0.70 - 0.80$ may not look as impressive as those obtained by convolutional architectures on well-studied benchmarks and other classification tasks (e.g., natural images from ImageNet or small aerial patches from DeepSat), this only attests to the difficulty of the task of understanding high-level, subjective concepts of urban planning in complex urban environments. First, satellite imagery typically contains much more semantic variation than natural images (as also noted, e.g., in [2, 13]), i.e., there is no "central" concept that the image is of (unlike the image of a cat or a flower). Second, the type of labels we use for supervision are higher-level concepts (such as "low density urban fabric", or "sports and leisure facilities"), which are much less specific than more physical land features e.g., "buildings" or "trees" (which are classes used in the DeepSat dataset). Moreover, top-down imagery poses specific challenges to convolutional architectures, as these models are inherently not rotationally-symmetric. Urban environments, especially from from a top-down point of view, come in many complex layouts, for which rotations are irrelevant. Nevertheless, these results are encouraging, especially since this is a harder problem by focusing on wider-area images and on higher-level, subjective concepts used in urban planning rather than on the standard, lower-level physical features such as in [1] or [17]. This suggests that such models may be useful feature extractors. Moreover, as more researchers tackle problems with the aid of satellite imagery (which is still a relatively under-researched source of data in the machine learning community), more open-source

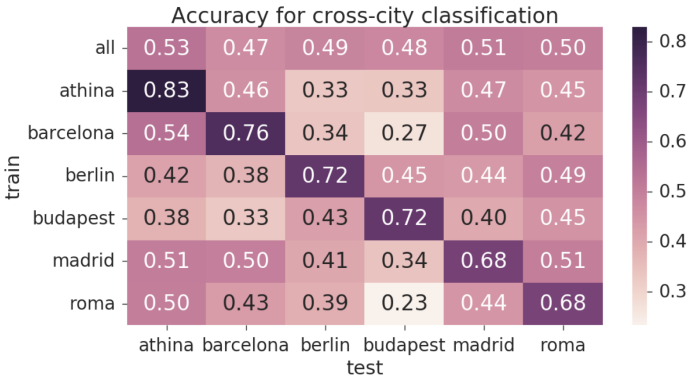

Figure 6: Transferability (classification accuracy) of models learned at one location and applied at another. Training on a more diverse set of cities (`all`) yields encouraging results compared with just pairwise training/testing.

datasets (like this one) are released, performance will certainly improve. For the remainder of this section we report results using the ResNet-50 architecture [7], as it consistently yielded (if only slightly) better classification results in our experiments than the VGG-16 architecture.

Transfer learning and classification performance. Next, we investigated how models trained in one setting (city or set of cities) perform when applied to other geographical locations. Figure 6 summarizes these experiments. In general, performance is poor when training on samples from a given city and testing on samples from a different city (the off-diagonal terms). This is expected, as these environments can be very different in appearance for cities as different as e.g., Budapest and Barcelona. However, we notice that a more diverse set (`all`) yields better performance when applied at different locations than models trained on individual cities. This is encouraging for our purpose of analyzing the high level "similarity" of urban neighborhoods via satellite imagery.

We next looked at per-class model performance to understand what types of environments are harder for the model to distinguish. Figure 7 shows such an example analysis for three example cities, of which a pair is "similar" according to Figure 6 (Rome and Barcelona), and another dissimilar (Rome and Budapest). The left panel shows model performance when training on samples from Barcelona, and predicting on test samples from Barcelona (intra-city). The middle panel shows training on Rome, and predicting on test samples in Barcelona, which can be assumed to be "similar" to Rome from a cultural and architectural standpoint (both Latin cities in warm climates). The right figure shows training on Barcelona, and predicting on test samples in Budapest, which can be assumed a rather different city from a cultural and architectural standpoint. For all cases, the classes that the model most struggles with are "High Density Urban Fabric", "Low Density Urban Fabric, and "Medium Density Urban Fabric". Considerable overlap can be noticed between these classes - which is not surprising given the highly subjective nature of these concepts. Other examples where the model performance is lower is forests and low-density urban areas being sometimes misclassifed as "green urban areas", which, again, is not surprising. This is especially apparent in the cross-city case, where the model struggles with telling apart these classes. For both the case of training and testing on "different cities" (Budapest and Barcelona) and on "similar" cities (Rome and Barcelona), we note that airports and forests are relatively easier to distinguish. However, more subjective, high-level urban-planning concepts such as "high density urban fabric" are harder to infer (and more easily confused with "medium density" or "low density" urban fabric) in the case of more similar cities (Rome and Barcelona) rather than dissimilar cities (Budapest and Barcelona). Urban environments containing sports and leisure facilities and green areas are under this view more similar between Rome and Barcelona than they are between Budapest and Barcelona.

Choosing the spatial scale: sensitivity analysis. So far, we have presented results assuming that tiles of $250m$ is an appropriate spatial scale for this analysis. Our intuition suggested that tiles of this size have enough variation and information to be recognized (even by humans) as belonging to one of the high-level concepts of land use classes that we study in this paper. However, one can find arguments in favor of smaller tile sizes, e.g., in many cities the size of a typical city block is $100m$. Thus, we trained models at different spatial scales and computed test-set accuracy values for three example cities, Barcelona, Roma, and Budapest - see Figure 8. It is apparent that, for all example cities, smaller spatial scales ($50m$, $100m$, $150m$) that we analyzed yield poorer performance than the scale we chose for the analysis in this paper ($250m$). This is likely because images at smaller scales do not capture enough variation in urban form (number and type of buildings, relative amount of vegetation, roads etc.) to allow for discriminating between concepts that are fairly high-level. This is in contrast with a benchmark such as DeepSat [1] that focuses on lower-level, physical concepts ("trees", "buildings", etc.). There, a good spatial scale is by necessity smaller ($28m$ for DeepSat), as variation in appearance and compositional elements is unwanted.

5.2 Comparing urban environments

Finally, we set to understand, at least on an initial qualitative level, how "similar" urban environments are to one another, across formal land use classes and geographies. Our first experiment was to project sample images for each class and city in this analysis to lower-dimensional manifolds, using the `t-SNE` algorithm [23]. This serves the purpose of both visualization (as `t-SNE` is widely used for visualizing high-dimensional data), as well as for providing an initial, coarse *continuous representation* of urban land use classes. In our experiments, we used balanced samples of size $N = 6000$, or 100 samples for each of the 10 classes for each city. We extracted features for each of these samples using the `all` models (trained on a train set with samples across all cities except for the test one).

Figure 9 visualizes such `t-SNE` embeddings for the six cities in our analysis. For most cities, classes such as low density urban fabric, forests, and water bodies are well-resolved, while sports and leisure facilities seem to consistently blend into other types of environments (which is not surprising, given that these types of facilities can be found within many types of locations that have a different formal urban planning class assigned). Intriguing differences emerge in this picture among the cities. For example, green

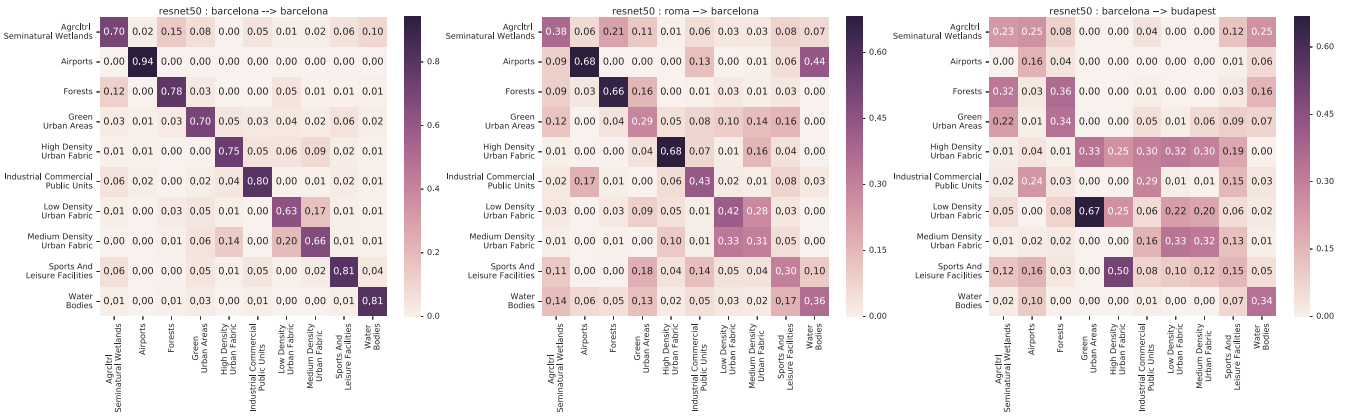

Figure 7: Example classification confusion matrix for land use inference. *Left*: training and testing on Barcelona; *Middle*: training on Rome, testing on Barcelona; *Right:* training on Rome, predicting on Budapest.

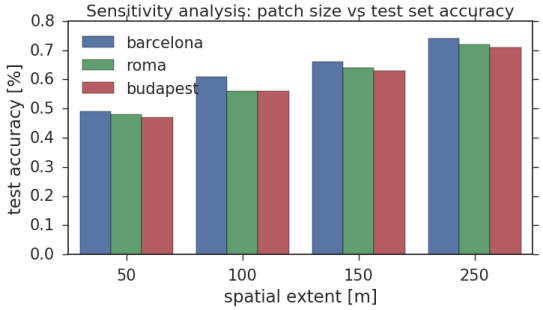

Figure 8: Sensitivity of training patch size vs test accuracy.

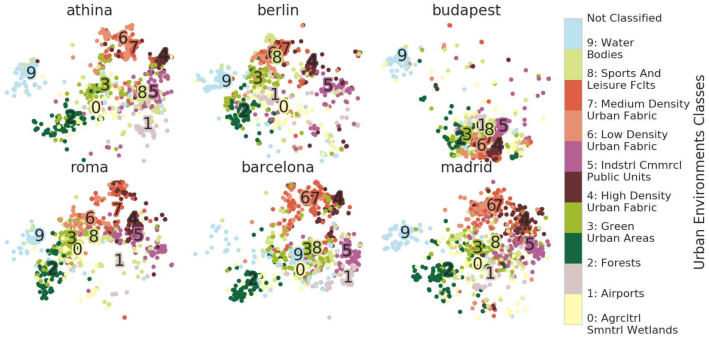

Figure 9: `t-SNE` visualization (the first 2 dimensions) of urban environments (satellite image samples) across six cities.

urban spaces seem fairly well resolved for most cities. Commercial neighborhoods in Barcelona seem more integrated with the other types of environments in the city, whereas for Berlin they appear more distinct. Urban water bodies are more embedded with urban parks for Barcelona than for other cities. Such reasoning (with more rigorous quantitative analysis) can serve as coarse way to benchmark and compare neighborhoods as input to further analysis about e.g., energy use, livelihood, or traffic in urban environments.

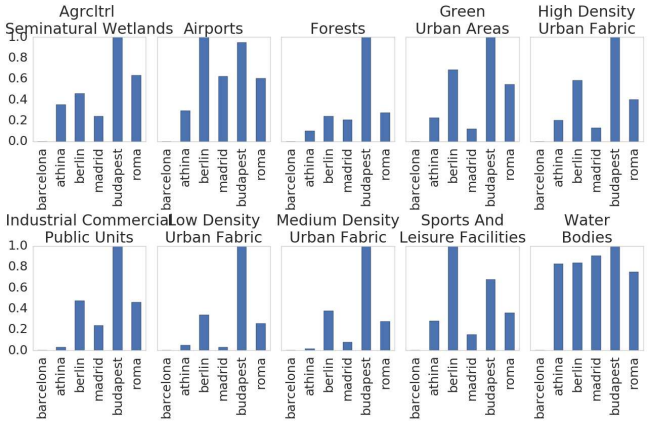

Figure 10: Comparing urban environments across cities (with reference to Barcelona) We show relative inter-city similarity measures computed as the sum of squares across the clusters in Figure 9.

We further illustrate how "similar" the six cities we used throughout this analysis are starting off the embeddings plots in Figure 9. For each land use class, we compute intra-city sum of squares in the 2-d t-SNE embedding, and display the results in Figure 10. Note that the distances are always shown with Barcelona as a reference point (chosen arbitrarily). For each panel, the normalization is with respect to the largest inter-city distance for that land use class. This visualization aids quick understanding of similarity between urban environments. For example, agricultural lands in Barcelona are most dissimilar to those in Budapest. Airports in Barcelona are most similar to those in Athens, and most dissimilar to those in Berlin and Budapest. Barcelona's forests and parks are most dissimilar to Budapest's. Water bodies in Barcelona are very dissimilar to all other cities. This point is enforced by Figure 11 below, which suggests that areas marked as water bodies in Barcelona are ocean waterfronts, whereas this class for all other cities represents rivers or lakes.

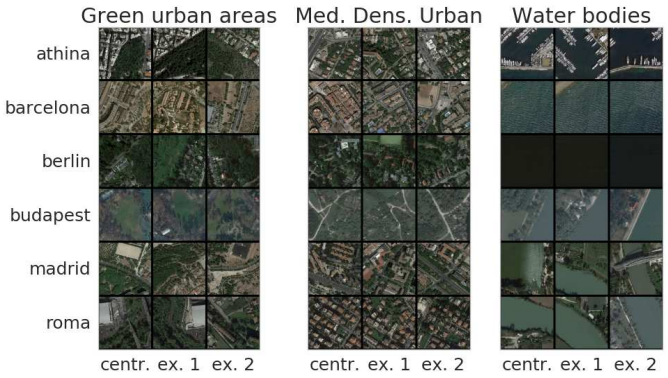

Figure 11: Samples from three urban environments across our 6 example cities. We sampled the 2-d t-SNE embedding of Figure 9 and queried for the closest real sample to the centroid using an efficient KD-tree search.

Finally, we explore the feature maps extracted by the convolutional model in order to illustrate how "similar" the six cities we used throughout this analysis are across three example environments, green urban areas, water bodies, and medium density urban fabric. For each city and land use class, we start off the centroid of the point cloud in the 2-d space of Figure 9, and find the nearest several samples using the KD-tree method described in Section 4. We present the results in Figure 11. Visual inspection indicates that the model has learned useful feature maps about urban environments: the sample image patches show a very good qualitative agreement with the region of the space where they're sampled from, indicated by the land use class of neighboring points. Qualitatively, it is clear that the features extracted from the top layer of the convolutional model allow a comparison between urban environments by high-level concepts used in urban planning.

6 CONCLUSIONS

This paper has investigated the use of convolutional neural networks for analyzing urban environments through satellite imagery at the scale of entire cities. Given the current relative dearth of *labeled* satellite imagery in the machine learning community, we have constructed an open dataset of over 140,000 samples over 10 consistent land use classes from 6 cities in Europe. As we continue to improve, curate, and expand this dataset, we hope that it can help other researchers in machine learning, smart cities, urban planning, and related fields in their work on understanding cities.

We set out to study similarity and variability across urban environments, as being able to quantify such patterns will enable richer applications in topics such as urban energy analysis, infrastructure benchmarking, and socio-economic composition of communities. We formulated this as a two-step task: first predicting urban land use classes from satellite imagery, then turning this (rigid) classification into a continuous spectrum by embedding the features extracted from the convolutional classifier into a lower-dimensional manifold. We show that the classification task achieves encouraging results, given the large variety in physical appearance of urban environments having the same functional class. Moreover, we exemplify how the features extracted from the convolutional network allow for identifying "neighbors" of any given query image, allowing a rich comparison analysis of urban environments by their visual composition.

The analysis in this paper shows that some types urban environments are easier to infer than others, both in the intra- and inter-city cases. For example, in all our experiments, the models had most trouble telling apart "high", "medium", and "low" density urban environments, attesting to the subjectivity of such a high-level classification for urban planning purposes. However, agricultural lands, forests, and airports tend to be visually similar across different cities - and the amount of relative dissimilarity can be quantified using the methods in this paper. Green urban areas (parks) are generally similar to forests or to leisure facilities, and the models do better in the intra-city case than predicting across cities. How industrial areas look is again less geography-specific: inter-city similarity is consistently larger than intra-city similarity. As such, for several classes we can expect learning to transfer from one geography to another. Thus, while it is not news that some cities are more "similar" than others (Barcelona is visually closer to Athens than it is to Berlin), the methodology in this paper allows for a more quantitative and practical comparison of similarity.

By leveraging satellite data (available virtually world-wide), this approach may allow for a low-cost way to analyze urban environments in locations where ground truth information on urban planning is not available. As future directions of this work, we plan to *i)* continue to develop more rigorous ways to compare and benchmark urban neighborhoods, going deeper to physical elements (vegetation, buildings, roads etc.); *ii)* improve and further curate the open Urban Environments dataset; and *iii)* extend this type of analysis to more cities across other geographical locations.

A PRACTICAL TRAINING DETAILS.

We split our training data into a training set (80% of the data) and a validation set (the remaining 20%). This is separate from the data sampled for the ground truth raster grids for each city, which we only used at test time. We implemented the architectures in the open-source deep learning framework Keras[7](with a TensorFlow[8] backend). In all our experiments, we used popular data augmentation techniques, including random horizontal and vertical flipping of the input images, random shearing (up to 0.1 radians), random scaling (up to 120%), random rotations (by at most 15 degrees either direction). Input images were $224 \times 224 \times 3$ pixels in size (RGB bands). For all experiments, we used stochastic gradient descent (with its Adadelta variant) to optimize the network loss function (a standard multi-class cross-entropy), starting with a learning rate of 0.1, and halving the rate each 10 epochs. We trained our networks for at most 100 epochs, with 2000 samples in each epoch, stopping the learning process when the accuracy on the validation set did not improve for more than 10 epochs. Given the inherent imbalance of the classes, we explicitly enforced that the minibatches used for training were relatively balanced by a weighted sampling procedure. For training, we used a cluster of 4 NVIDIA K80 GPUs, and tested our models on a cluster of 48 CPUs.

[7]https://github.com/fchollet/keras
[8]www.tensorflow.org

REFERENCES

[1] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna R. Nemani. 2015. DeepSat - A Learning framework for Satellite Imagery. *CoRR* abs/1509.03602 (2015).

[2] Marco Castelluccio, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. 2015. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *CoRR* abs/1508.00092 (2015). http://arxiv.org/abs/1508.00092

[3] Dragos Costea and Marius Leordeanu. 2016. Aerial image geolocalization from recognition and matching of roads and intersections. *CoRR* abs/1605.08323 (2016). http://arxiv.org/abs/1605.08323

[4] Marco De Nadai, Radu Laurentiu Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, Cesar Augusto Hidalgo, Nicu Sebe, and Bruno Lepri. 2016. Are Safer Looking Neighborhoods More Lively?: A Multimodal Investigation into Urban Life. In *Proceedings of the 2016 ACM on Multimedia Conference (MM '16)*. ACM, New York, NY, USA, 1127–1135. DOI: http://dx.doi.org/10.1145/2964284.2964312

[5] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A. Hidalgo. 2016. *Deep Learning the City: Quantifying Urban Perception at a Global Scale*. Springer International Publishing, Cham, 196–212. DOI: http://dx.doi.org/10.1007/978-3-319-46448-0_12

[6] Sebastian Grauwin, Stanislav Sobolevsky, Simon Moritz, István Gódor, and Carlo Ratti. 2015. *Towards a Comparative Science of Cities: Using Mobile Traffic Records in New York, London, and Hong Kong*. Springer International Publishing, Cham, 363–387. DOI: http://dx.doi.org/10.1007/978-3-319-11469-9_15

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 770–778. DOI: http://dx.doi.org/10.1109/CVPR.2016.90

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity Mappings in Deep Residual Networks. In *ECCV (4) (Lecture Notes in Computer Science)*, Vol. 9908. Springer, 630–645.

[9] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.

[10] Maxime Lenormand, Miguel Picornell, Oliva G. Cantú-Ros, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, Maxi San Miguel, and José J. Ramasco. 2015. Comparing and modelling land use organization in cities. *Royal Society Open Science* 2, 12 (2015). DOI: http://dx.doi.org/10.1098/rsos.150449 arXiv:http://rsos.royalsocietypublishing.org/content/2/12/150449.full.pdf

[11] Qingshan Liu, Renlong Hang, Huihui Song, and Zhi Li. 2016. Learning Multi-Scale Deep Features for High-Resolution Satellite Image Classification. *CoRR* abs/1611.03591 (2016). http://arxiv.org/abs/1611.03591

[12] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. 2016. Classification With an Edge: Improving Semantic Image Segmentation with Boundary Detection. *ArXiv e-prints* (Dec. 2016). arXiv:cs.CV/1612.01337

[13] Volodymyr Mnih. 2013. *Machine learning for aerial image labeling*. Ph.D. Dissertation. University of Toronto.

[14] Nikhil Naik, Ramesh Raskar, and Csar A. Hidalgo. 2016. Cities Are Physical Too: Using Computer Vision to Measure the Quality and Impact of Urban Appearance. *American Economic Review* 106, 5 (May 2016), 128–32. DOI: http://dx.doi.org/10.1257/aer.p20161030

[15] M. Papadomanolaki, M. Vakalopoulou, S. Zagoruyko, and K. Karantzalos. 2016. Benchmarking Deep Learning Frameworks for the Classification of Very High Resolution Satellite Multispectral Data. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* (June 2016), 83–88. DOI: http://dx.doi.org/10.5194/isprs-annals-III-7-83-2016

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[17] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos. 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 44–51. DOI: http://dx.doi.org/10.1109/CVPRW.2015.7301382

[18] A. Romero, C. Gatta, and G. Camps-Valls. 2016. Unsupervised Deep Feature Extraction for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* 54, 3 (March 2016), 1349–1362. DOI: http://dx.doi.org/10.1109/TGRS.2015.2478379

[19] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556

[20] Jameson L. Toole, Michael Ulm, Marta C. González, and Dietmar Bauer. 2012. Inferring Land Use from Mobile Phone Activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing (UrbComp '12)*. ACM, New York, NY, USA, 1–8. DOI: http://dx.doi.org/10.1145/2346496.2346498

[21] Nagesh Kumar Uba. 2016. *Land Use and Land Cover Classification Using Deep Learning Techniques*. Master's thesis. Arizona State University.

[22] European Union. 2011. Urban Atlas. Urban Atlas is a product commissioned by DG REGIO and provided by the Copernicus programme. http://www.eea.europa.eu/data-and-maps/data/urban-atlas/. (2011).

[23] L.J.P van der Maaten and G.E. Hinton. 2008. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9: 2579-2605 (2008).

[24] Scott Workman, Richard Souvenir, and Nathan Jacobs. 2015. Wide-Area Image Geolocalization with Aerial Reference Imagery. *CoRR* abs/1510.03743 (2015).

[25] Yi Yang and Shawn Newsam. 2010. Bag-of-visual-words and Spatial Extensions for Land-use Classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '10)*. ACM, New York, NY, USA, 270–279. DOI: http://dx.doi.org/10.1145/1869790.1869829

[26] Jun Yue, Wenzhi Zhao, Shanjun Mao, and Hui Liu. 2015. Spectral-spatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sensing Letters* 6, 6 (2015), 468–477. DOI: http://dx.doi.org/10.1080/2150704X.2015.1047045 arXiv:http://dx.doi.org/10.1080/2150704X.2015.1047045

[27] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. 2016. Predicting Ground-Level Scene Layout from Aerial Imagery. *CoRR* abs/1612.02709 (2016).

[28] Yi Zhu and Shawn Newsam. 2015. Land Use Classification Using Convolutional Neural Networks Applied to Ground-level Images. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15)*. ACM, New York, NY, USA, Article 61, 4 pages. DOI: http://dx.doi.org/10.1145/2820783.2820851

ARTICLES
https://doi.org/10.1038/s41560-018-0136-x
nature energy

Planning for electric vehicle needs by coupling charging profiles with urban mobility

Yanyan Xu[1,6], Serdar Çolak[1,2,6], Emre C. Kara[3], Scott J. Moura[4] and Marta C. González[1,2,5*]

The rising adoption of plug-in electric vehicles (PEVs) leads to the temporal alignment of their electricity and mobility demands. However, mobility demand has not yet been considered in electricity planning and management. Here, we present a method to estimate individual mobility of PEV drivers at fine temporal and spatial resolution, by integrating three unique datasets of mobile phone activity of 1.39 million Bay Area residents, census data and the PEV drivers survey data. Through coupling the uncovered patterns of PEV mobility with the charging activity of PEVs in 580,000 session profiles obtained in the same region, we recommend changes in PEV charging times of commuters at their work stations and shave the pronounced peak in power demand. Informed by the tariff of electricity, we calculate the monetary gains to incentivize the adoption of the recommendations. These results open avenues for planning for the future of coupled transportation and electricity needs using personalized data.

The excessive exploitation of petroleum and coal affect not only the security of energy supply but also air quality and climate change. These shortcomings have triggered the search for cleaner alternative fuels for transportation[1–3]. Today's PEV technology is one of the most promising candidates to date[4,5]. The main issues that have hindered their adoption are: range anxiety, charger unavailability and high prices[3]. However, improvements in battery technology, tax breaks and subsidized charging programmes[6,7] have somewhat mitigated these limitations. As a result, PEVs are becoming a more viable means to move and are being adopted by drivers at steadily increasing rates[4]. According to the US Energy Information Administration, the number of PEVs in the United States doubled between 2013 and 2015 and is expected to reach 20 million by 2020[8].

Planning for the mobility needs of PEVs is particularly important in the context of the vulnerability of the power grid to outages that can cascade drastically[9]. Large-scale failures signified a need to reexamine the balance between power demand and the electricity infrastructure, opening the need for interdisciplinary approaches to study this complex system[10]. A body of literature has focused on the nature of network reliability of power grids, the role of network topology on the spread of cascading failures[11–17]. On the subject of PEVs and their impact on the grid, methods of optimization and control of PEV electricity consumption have a rich set of avenues[18–20]. Research topics on this front include measuring impact on the grid[21–27], developing accurate PEV energy consumption models[28], energy management[29,30], smart charging strategies that probe centralized and decentralized approaches[31,32], scheduling[33,34], peak shaving, emissions, pricing models[35,36] and joint optimization of power and transportation networks[37]. A common shortcoming in these works is the narrow scope in incorporating individual mobility needs into the analyses, often limited to the estimation of arrival or departure hours. Up-to-date data on individual mobility demand at metropolitan scale have not yet been incorporated into the planning schemes to manage electricity demand.

In this work, we target these gaps in the literature to extend the current knowledge of transportation-based electricity. For this purpose, we bring together three independent data sources: (i) mobile phone activity of a large sample of San Francisco Bay Area residents, (ii) charging sessions obtained from the commercial PEV supply equipment in the same region and (iii) surveys on the use of conventional and electric vehicles, together with census data for income information at the ZIP code level (see Methods). In the first part of the work, we estimate individual vehicular mobility per week day in the Bay Area using the mobile phone activity of a large sample of residents. We then present a Bayesian methodology to sample the PEV drivers from all travellers by utilizing information obtained from surveys regarding the household income and daily travel distances of PEV drivers. In the second part, with the charging session data, we analyse the various aspects of charging activity to characterize the nature of electricity demand at charging stations. We observe that PEV charging patterns are highly regular with morning and evening peaks following the traffic peaks. These peaks of demand are undesired because they can cause instabilities in the power grid. To tackle this problem, we explore the relationship between the electricity consumption of simulated PEV commuters working in the selected ZIP codes and the observed energy demand at individual commercial charging stations in the same region. We calibrate the charging behaviours of PEV drivers to match the observed demand. As an application, we lay out a charging scheme that minimizes the peak power by changing the start and end of the charging sessions, while also taking into account the constraints in changing departures and arrivals. We show how not knowing the mobility constraints decreases the potential of the peak minimization schemes. In contrast, introducing the awareness of individual mobility increases the feasibility of their adoption, affecting less the benefits of peak minimization. The resulting effects on the commuting travel times and the monetary benefits from the changes in charging times support the viability of the charging time shifts. Figure 1 depicts a summary of the proposed framework.

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [3]SLAC National Accelerator Laboratory, Menlo Park, CA, USA. [4]Department of Civil and Environmental Engineering, University of California, Berkeley, CA, USA. [5]Department of City and Regional Planning, University of California, Berkeley, CA, USA. [6]These authors contributed equally: Yanyan Xu and Serdar Çolak. *e-mail: martag@mit.edu

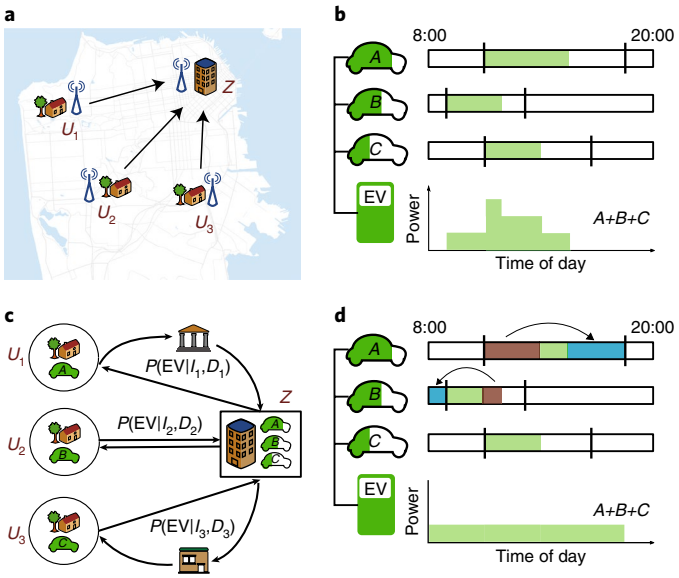

Fig. 1 | Coupling PEV charging with urban mobility. a, Mobile phone data are used to model individual mobility. Three users, U_1, U_2 and U_3, interacted with their mobile phone at their home and then at workplace Z. **b**, Charging sessions data are used to characterize individual and total electricity demand curves. The charging power per timeslot of vehicle A, B and C are 1 kW, and their charging sessions are overlapped. The green bar plot shows the total electricity demand of vehicle A, B and C at the charging station during one day. The peak electricity demand reaches 3 kW. **c**, The Bayesian inference method is proposed to find the probability that the vehicular trip is made by a PEV. **d**, Charging activity is shifted to create a recommendation scheme that relieves peaks in demand. The peak electricity demand at the charging station is reduced to 1 kW via shifting the charging sessions of vehicle A and B.

Estimating individual mobility of PEV drivers

We simulate the individual mobility of the entire population of the Bay Area using a fine-scale urban mobility model, TimeGeo[38]. This process begins with the extraction of stay locations in the trajectories of each individual[39–41]. Each stay is then accordingly labelled as home, work or other, based on temporal properties of the call activities. According to whether the workplaces are detected or not, we model the trips of commuters and non-commuters respectively (see Methods). Figure 2a represents the simulated trajectory and the labelled activities of a mobile phone user. The simulations of individual mobility based on mobile phone data compare very well with the results using two travel surveys, the 2010–2012 California Household Travel Survey (CHTS)[42] and the 2009 National Household Travel Survey (NHTS)[43]. As shown in Fig. 2b,d, the daily visited locations and fraction of departures per time of the day simulated by our model based on phone data agree well with the travel surveys. Further comparisons are presented in the Supplementary Figs. 1 and 2.

Mobility motifs[44] describe the individual daily travel networks, where nodes are visited locations and directed edges are trips from one location to another. For example, the motif of an individual whose only trips in a day are to and from work will consist of two nodes with two directed edges (one in both directions). On average, individuals visit three different places per day. When constructing all possible directed networks with six or fewer nodes, there exist over a million ways for an individual to travel between. However, 90% of people use one of just 17 networks, called motifs[44]. While nearly half of the population follow the simple two-locations motif. These results can be modelled with a probabilistic Markov model[38] that assigns particular rates to each individual informed by their trip behaviour. The top ten motifs of nearly six million simulated drivers in the Bay Area are summarized in Fig. 2c, which implies that the distribution of our simulated motifs agrees well with the information gathered from mobile phone users. The comparison of motifs of commuters and non-commuters are shown in Supplementary Fig. 1c.

After simulated individual mobility overall, we can probabilistically estimate the individual mobility of PEVs. To that end, we utilize the vehicle usage rate from the US census data and the California Plug-in Electric Vehicle Driver Survey[45]. According to this survey, PEV drivers' income distribution is skewed towards higher income segments. In particular, the percentage of those with average annual incomes above US$150,000 among conventional vehicle drivers is 15%, compared with the 47% observed among PEV drivers. The survey also highlights the typical distances PEV drivers travel: 64% of PEV drivers travel less than 30 miles per day (Table 1). This information is used to subsample PEV trips from total vehicular trips by implementing the Bayesian sampling procedure. Namely, we use the individual income estimated from the US census data at the census tract level and daily route distance from TimeGeo to estimate the probability of that the driver travels with a PEV, both for commuters and non-commuters (see Methods).

Figure 3a depicts the number of PEVs estimated from the Bayesian method at each ZIP code and the number obtained from the dataset on PEVs collected by the California Air Resources Board's Clean Vehicle Rebate Project[46], referred as the CVRP dataset. Figure 3b shows a good agreement between the number of PEVs obtained via the Bayesian estimates and the mobility model versus the ground truth of PEV usage. Figure 3c,d compares the distributions of the morning route distance, D, made by all commuters versus PEV drivers, as well as the commuting travel time, T, under free flow conditions. There are fewer PEV trips shorter than 5 km and longer than 25 km, in agreement with the findings of the survey. Figure 3e depicts the four mobility motifs from PEV commuters, showing that approximately 66% of PEV commuters mostly travel between home and work during weekdays. The simple motif (with ID = 1) is more prevalent among PEV drivers than among commuters using conventional vehicles, this may be a sign of the driver's concerns on the range of PEVs.

Electric vehicle charging session data profiles

In this section, we analyse PEV charging in non-residential regions by examining: visitation patterns and adoption rates, temporal features of arrivals and departures, and typical energy and power consumption levels. PEV drivers display varying degrees of regularity in terms of how often they visit charging stations. Figure 4a reveals that for the majority of PEV drivers the average number of charging sessions per day, N_{day}, is less than one. The bottom left inset in Fig. 4a displays the logarithmic distribution of the number unique PEV charging stations (EVSEs) visited by each PEV driver, N_{EVSE}. Noticeably, the great majority of PEV drivers (95.6%) is observed in less than 20 distinct EVSEs. The top right inset of Fig. 4a depicts the rate of PEV adoption observed throughout the year. The 3,000 drivers observed in January 2013 increases by an average of 1,000 per month, doubling twice over the course of 2013.

We look at the arrival and departure hours of charging sessions, h_a and h_d, in Fig. 4b. Approximately 50% of all arrivals take place in the 6:00–11:00 morning period, and as expected, the morning and the evening peaks are highly pronounced. This points to the parallels between the temporal component of overall travel demand to electricity demand. We compare the distribution of departure time in the morning of commuters with the arrival time of charging sessions and find notable delay between these two distributions (see Supplementary Fig. 6a). Such delay represents the driving time

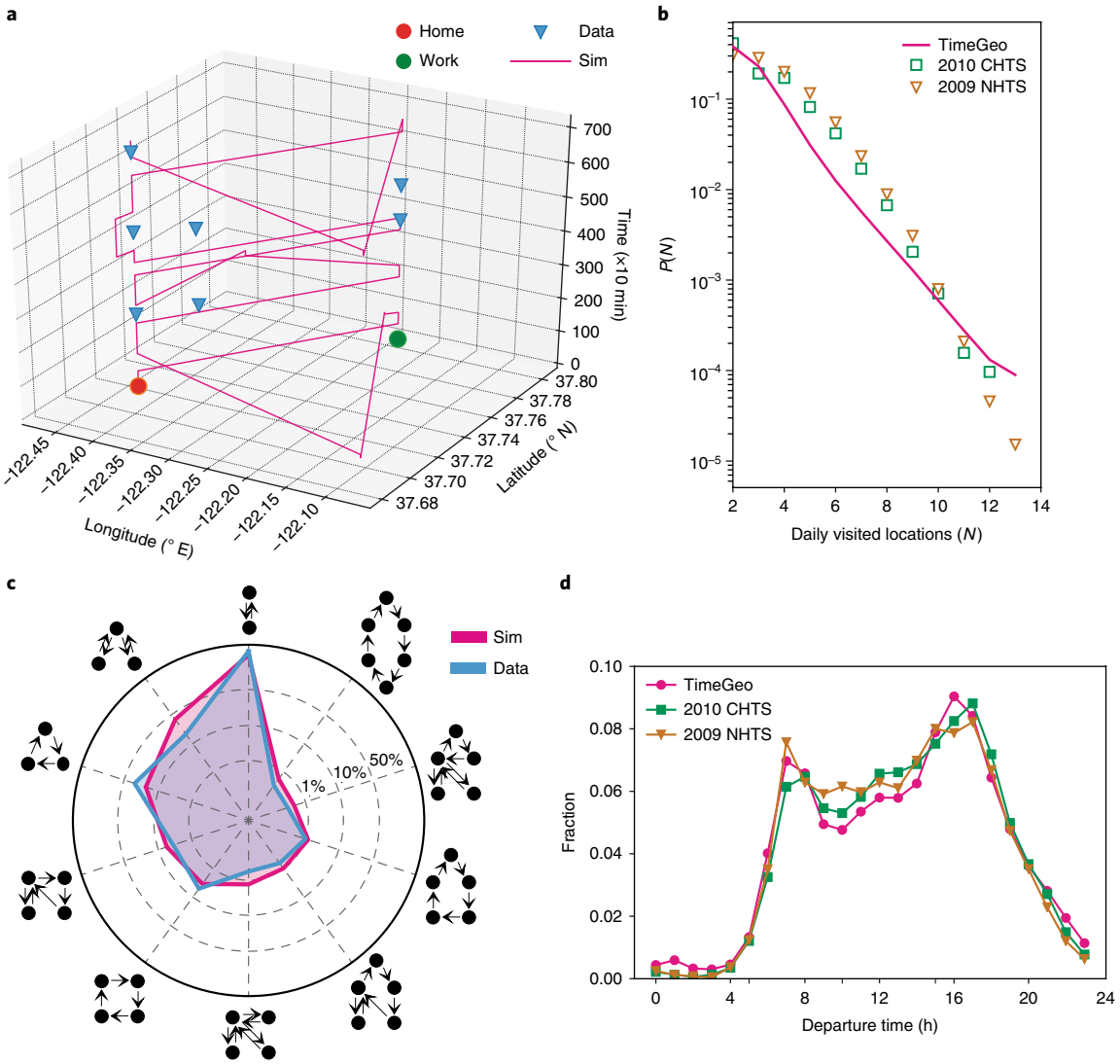

Fig. 2 | Validation of individual mobility simulation in the Bay Area. a, Simulated trajectory of a mobile phone user who is labelled as commuter. The blue triangles represent the actually recorded activities of the selected user from 22–25 October 2012. The red and green circles are the recognized home and work from the mobile phone data respectively. **b**, Population distribution comparison of daily visited locations between simulation and two travel survey datasets: 2010 CHTS and 2009 NHTS. Supplementary Fig. 1a,b presents comparisons of two more properties of mobility: stay duration (panel a) and trip distance (panel b). **c**, Validation of individual mobility motifs for a sample of mobile phone users in simulation and CDRs data. Nodes and directed edges in the label denote the visited locations in sequence during one day. 1%, 10% and 50% denote the values of the three dashed grid lines in the radar plot. The red-violet and blue shaded areas denote the fraction of the motifs in all users' travel activities observed from the CDRs data and the simulation, respectively. Around 65% of users observed only travel between two locations from the mobile phone data. The distribution of the ten primary motifs shows a high degree of similarity to the CDRs data. **d**, Fraction of trip departures by time of the day, comparing the simulation, the 2010 CHTS and the 2009 NHTS. Supplementary Fig. 2 shows the comparisons for various trip purposes.

of commuters from home to work in the Bay Area, which is around 30 min on average[47]. In the inset of Fig. 4b we look at the distribution of inter-arrival and inter-departure times, Δh_a and Δh_d, that is, the time between two consecutive charging sessions for the same driver ID. These distributions are peaked at multiples of 24 h, pointing to the diurnal periodicity of PEV drivers' charging behaviour. These findings reinforce the notion that commuting and charging behaviour in the non-residential regions are highly related.

Next, we shift our focus to measures per session, such as energy, duration and power. Figure 4c exhibits the average energy consumption per session, E_S. The battery sizes of Nissan Leaf (24 kWh) and Chevrolet Volt (16 kWh), two of the most commonly used PEVs in the region, are marked (see also Supplementary Fig. 4)[46]. Typically, E_S are well below these capacities, indicating that PEV drivers typically stay within the range of their PEVs. PEV drivers can charge at home or not necessarily start their commute at full capacity. On the other hand, the distribution of session durations reveals that 98.4% of all charging sessions last less than a day (Fig. 4c), in line with the ubiquitous charging at work places. Given the flexibility in terms of battery capacity and mobility patterns, here we assume that the session energy E_S represents not a single commuting trip, but rather a number of them.

The actual charging activity does not last as long as the session duration, δ_S, as seen in Fig. 4d. We divide sessions into four categories based on their session duration and plot the average power consumption for each segment at various percentages of the total duration. We observed three levels of power rate that are most common, denoted here as levels 1–3 (L1, L2 and L3). The first

Table 1 | Characteristics of PEV drivers

	Category	Conventional	PEV
Income (US$1,000)	Unknown	20%	17%
	<50	20%	2%
	50–100	30%	13%
	100–150	14%	20%
	>150	15%	47%
Distance (miles)	<15	-	14%
	15–30	-	50%
	30–45	-	28%
	>45	-	8%

Distribution of household income and the daily travelled miles of PEV drivers in California, USA[45].

two deliver 120 V and 240 V, typically corresponding to 3.3 kW and 6.6 kW, respectively. L3 chargers are mainly for fast charging at 480 V and are relatively uncommon. As faster charging technology becomes more abundant, the peak load yielded by PEVs will be even higher as the charging sessions start intensively in the morning. In the charging dataset, L1 and L2 chargers make up 99.9% of all the sessions in the dataset (see Supplementary Fig. 6d). This composition of power ratings explains the 4 kW upper limit to average power consumption observed in Fig. 4d. For sessions lasting less than 4 h, average power stays above 3 kW up to 80% of the duration of the session. Conversely, for sessions that last longer than 12 h but less than a day, only in the starting 25% of the session duration is there active charging. This corresponds approximately to 3–6 h, and the power remains zero thereafter. This is consistent with constant-current, constant-voltage battery charging behaviour and it suggests that currently there is no strategy to charging involved: PEVs are charged immediately on arrival.

Coupling PEVs energy demand and mobility patterns

This section presents the coupling of the individual mobility of PEV drivers with the energy demand at each destination. First, we measure the distribution of electricity demand at a ZIP code from the charging sessions, and connect that to the distribution of estimated energy demand of simulated PEVs commuting to that ZIP code. The charging sessions data is provided by a private company with a partial coverage of the market with reasonable agreement of the most popular destinations for charging (see Methods, and the comparison of estimated PEVs versus the charging data in Supplementary Fig. 5). To estimate the energy demand of each PEV trip, we first assign each PEV a mode from the four most popular modes (see Supplementary Fig. 4). Each PEV mode is associated an energy consumption model. Specifically, we use the drivetrain model for the two battery electric vehicle (BEV) modes, Nissan Leaf and Tesla Model S[48], and the charge-depleting model for the two plug-in hybrid electric vehicle (PHEV) modes, Chevrolet Volt and Toyota Prius[49]. For each PEV trip, we estimate the energy demand using its average speed and route distance (see Methods).

Due to the limited information from charging sessions data, we can not infer the state of charge of each vehicle. Therefore, we assign different shares of charging states: morning consumption, daily consumption and two-days consumption. This corresponds to different charging behaviours respectively: charging both at home and work every day, charging at work once per day or charging at work once every other day, indicating that the energy consumption at the arrival equals to the consumption of the trips in the last two days. In Fig. 3f, we show the comparison of probability distributions of the energy consumption of the three scenarios together with the actual charge, E_S, in a selected ZIP code. The peaks in the distribution of E_S demonstrate the heterogeneity in the electricity demand, which is mainly caused by the travel distance, the battery capacities and the charging behaviours of various PEVs. The 3–4 kWh peak, which can be observed in both actual and estimated consumptions, is a combination of low energy demand as a consequence of short commuting trips and PHEVs that typically have a battery capacity around 4 kWh (ref. [50]). Further comparisons between daily consumption estimates and the charging station data in selected ZIP codes are shown in Fig. 3g. The charging data have more pronounced peaks than our daily curves, this may be because our charging behaviours are simplified, leaving room for further improvements.

To estimate the charging behaviour in the given ZIP code, we limit the amount of charging sessions to 30 kWh. The charging behaviour is distributed to match the demand of the ZIP code with most charging sessions. Corresponding to average charging schemes that result in: 10%, 35% and 55% of the PEVs drivers, respectively (see also Supplementary Fig. 6c). Note that only 10% of the drivers charging at home is in agreement with a recent report by the Department of Energy, which states that 80% of partners in their Workplace Charging Challenge programme provide free PEV charging[51]. We randomly assign the charging speed from the charging session data to the simulated PEVs to make the distribution of charging speed match with the ground truth (see Supplementary Fig. 6d).

Strategies to mitigate peak demand with mobility needs

Our goal in this section is to transform the load curve into one that is more uniformly distributed across the day. To that end, we propose changes in the start and end of the charging sessions such that the peak power load is minimized. We cast the problem as a mixed-integer linear program with discrete shifts in arrival times and charging end times as inputs (see Methods). The program modifies the total power P_t measured through the day resulting from the overlapping charging activities of a population of PEVs in a way that minimizes the peak power while keeping the total energy consumed constantly. In this context, we test two different strategies. The first fixes the departure times for PEVs and shifts the arrival time in advance by d^i, an amount specific to session i within the interval $[0,d]$, to minimize the peak power load P_{peak}. We refer to this strategy as end bound. The second strategy, referred to as flexible, offers modifications to both of the arrival and departure times. In this approach, charging activity is shifted in the interval $[-d,d]$. In both of the two strategies, the PEVs are charged once they are plugged into the charging station and PEV driver could depart at their scheduled time if the charging session has finished. As a future scenario, we show the peak load saving when the charging station is able to control the start of the charging session independent of the arrival time. The current infrastructure does not allow the start of charging at an optimized time and it is coupled to the PEV arrival. With a smarter charging-shift scenario, the charging could be freely shifted between the plug-in/arrival and plug-out/departure time. We test the time shifting scenarios on the 448 PEVs travelling to our ZIP code with the largest number of incoming users.

Figure 5a illustrates an instance where 40 charging sessions in one day are shifted flexibly. Some users are recommended to charge earlier and others later than their actual request. Figure 5b depicts how the power curves are modified under the flexible strategy with varying value of d (Supplementary Fig. 7 presents the results for the end bound strategy). The flexible strategy reduces P_{peak} down 47% from 1,019 kW to approximately 479 kW for $d=4$, or 1 h. In contrast, the charging-shift strategy reduces the peak load by 66% as we have more room to operate the PEV charging.

We also evaluate the effects of introducing the constraints of the individual mobility motifs of each PEV. Mobility constraints are introduced via the four motifs depicted in Fig. 3e. More than 30% of PEV drivers have other activities before or after work,

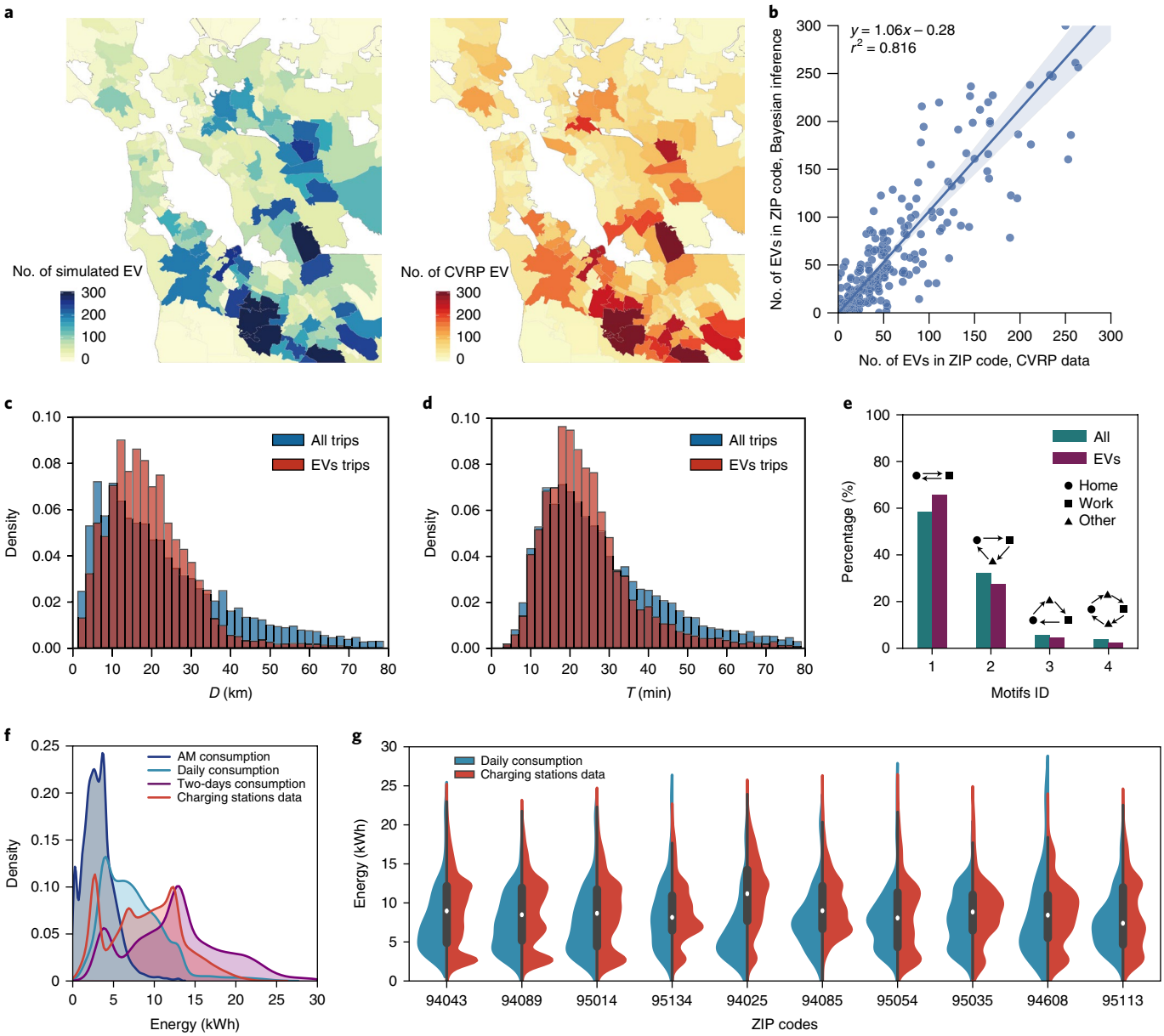

Fig. 3 | Validation of PEV mobility estimation and calibration of PEV charging behaviour. a, Number of PEVs in the residential ZIP codes of the Bay Area from the simulation and the CVRP datasets in the end of 2013. The total number of PEVs is 15,963 from our simulation, which is close to the actual number from CVRP datasets, 16,103. **b**, Correlation between the simulated PEV and the PEV from CVRP data. **c**,**d**, Probability distributions of commuting distances, *D*, and commuting travel times, *T*, of all vehicle trips and EV trips estimated through income information and trip distances. **e**, Fractions of four types of mobility motif of all commuters and PEV users. 58% of commuters only travel between home and work on weekdays, while the rest 42% have other activities before or after work. Similarly, 66% of commuters using PEVs only travel between home and work, and the rest 34% have other activities before or after work. **f**, Probability distributions of charging energy E_s obtained from charging sessions compared to those of the energy demand estimated by energy consumption model on three charging behaviour scenarios, morning, daily and two-days. **g**, Probability distributions of charging energy E_s and those of the daily energy demand of simulated PEVs for the ZIP codes that have the most charging session records.

therefore, they are limited to accept the recommendations of a time-shift strategy if the recommendation falls before their usual arrival and departure times. Namely, we impose the following restrictions per motif ID. (1) Home-work-home, can change both of the arrival and departure time; (2) home-work-other-home, indicates activities after work, and they can not delay their departure time; (3) home-other-work-home, which indicates activities before work, and they can not change their arrival time; (4) home-other-work-other-home, which indicates the PEV drivers can change neither the arrival nor the departure time.

Figure 5c contrasts the estimates of peak saving with and without the consideration of individual mobility constraints in the optimization strategy, looking at the percentage of peak loads of three schemes with variant *d*. The three schemes are (i) the optimal strategy, where all PEV drivers can follow the time shifts; (ii) the motif blind subtracts to the optimal estimates, all the PEV drivers that will not accept the recommendations due to the individual mobility constraints; (iii) the motif aware is a customized strategy informed by the individual mobility constraints to distribute the time shifts. These three strategies are evaluated under the

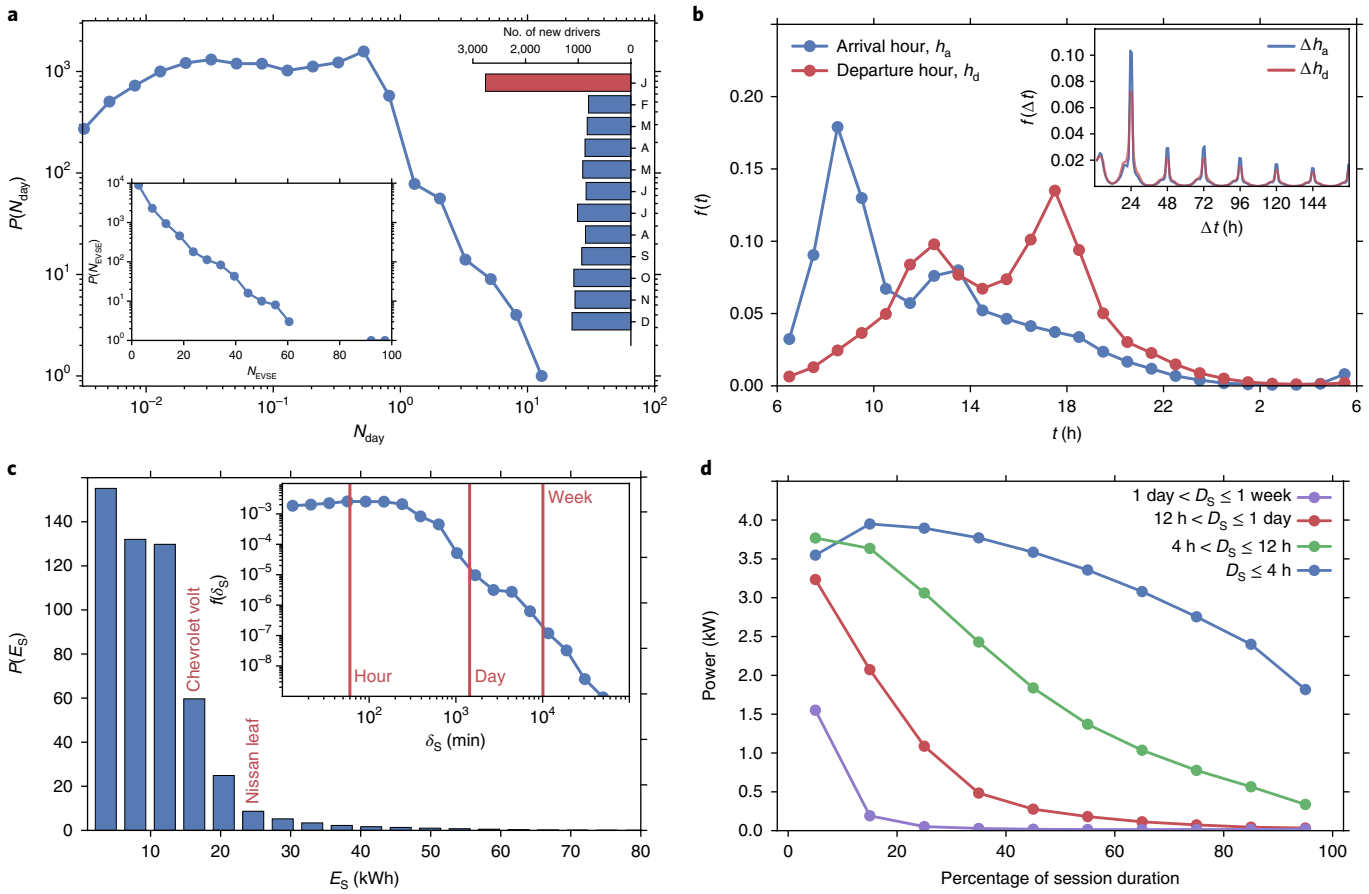

Fig. 4 | PEV charging session profiles. a, Distribution of N_{day}, the number of sessions per day for each driver ID starting from the day of first record (right inset: number of new driver IDs added every month; left inset, distribution of N_{EVSE}, the number of unique EVSEs visited by every driver ID). **b**, Distributions of h_a and h_d, the arrival and departure hours to and from an EVSE. (Inset, the distributions of Δh_a and Δh_d, the inter-arrival and inter-departure times for a driver ID visiting a specific EVSE.) **c**, Distribution of E_S, the total energy withdrawn per session. The battery capacities of the two most commonly used PEVs, Nissan Leaf and Chevrolet Volt, are labelled. (Inset, the distribution of δ_S, session durations. The sessions lasting one hour, one day, and one week are highlighted with red vertical lines.) **d**, Power consumption as a function of the normalized session duration segmented by total duration groups.

end bound and flexible schemes. The comparisons of the power curves of the three schemes and the two strategies are shown in Supplementary Figs. 7 and 8.

The optimal results are the best case scenario because all sessions can be shifted. The motif aware scheme represents the more feasible gains by coupling the charging strategies with the constraints of the drivers. The motif blind scheme is added to show how optimization strategies based on charging data only overestimate the benefits of the savings from 3% to 15%, while this loss can be overcome with the mobility information.

For the motif aware scheme, we examine the peak load saving versus the PEV driver's adoption rate (also known as flexibility) of the time shifts. Figure 5d shows a linear relationship between the acceptance rate and the savings for both of the motif blind and motif aware schemes. With the increase of adoption rate, motif aware contributes more on peak load saving than motif blind. For instance, when 80% of PEV drivers accept the time-shift recommendations, the motif aware scheme reduces the peak load by 424 kW on average, while the motif blind scheme reduces 279 kW on average.

In Fig. 5e, we present the peak load saving of the three flexible schemes versus the number of PEVs travelling to the selected ZIP code. The inset of Fig. 5e shows the estimated peak load without energy demand management. Both the peak load and the three saving powers grow linearly with the numbers of PEVs. The gap between optimal and motif aware is negligible in comparison with optimal versus motif blind.

This framework allows us to evaluate how the time shifts in their departures affect the commuting travel times of the PEV trips into the subject ZIP code. Figure 5f shows that the peak load reductions can be achieved without causing major discomfort to commuters in terms of travel times. The most negatively influenced drivers end up losing a maximum of 20 min in the case of $d=4$ (1 h), and are far less than those who are unaffected by the proposed changes. There is a number of drivers that even achieve travel time savings.

Finally, we examine the monetary outcomes of the proposed strategies, end bound and flexible. We use the max part-peak demand summer rates in the E-19 rate structure for the region to calculate the change in demand charge as a proxy of the cost in terms of dollars[32,52]. When implemented, the possible benefits of the schemes we proposed are displayed in Fig. 5g: monthly potential savings in the demand charge can reach up to US$2,500 for the motif customized and flexible strategy, corresponding to roughly US$5.6 per month per session. Without managing charging, these savings remain unrealized, and are paid by PEV drivers or the companies that subsidize the charging activity. As a sum the savings are substantial, yet for the number of sessions on a typical weekday considered here, the amount saved per individual is relatively small, making the uniform distribution of savings a relatively unexciting reward for cooperation. However, as the

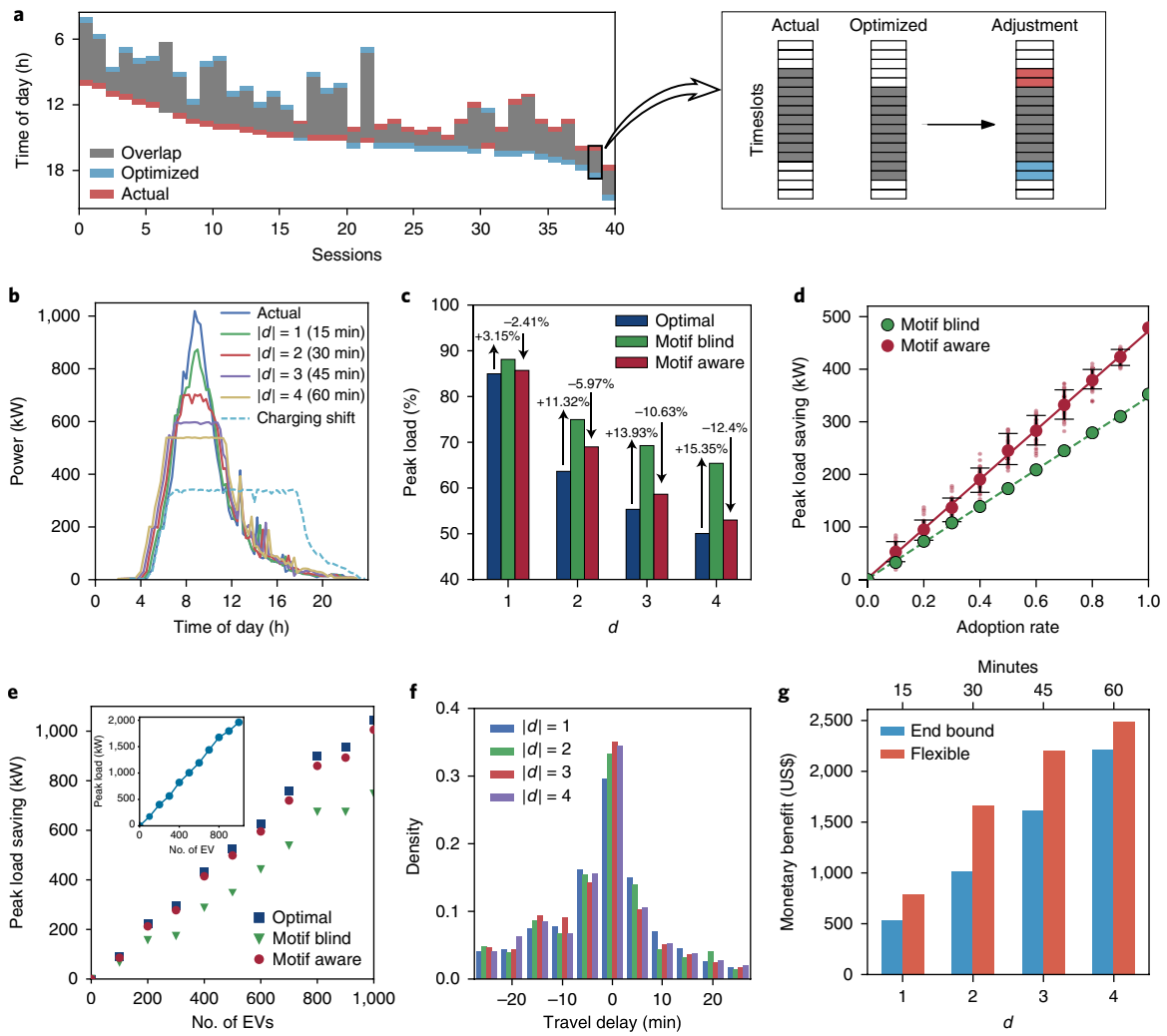

Fig. 5 | Assessing the benefits of minimizing peak power. a, An illustration of 40 sessions during one day in the flexible strategy. The sessions are shifted earlier or later such that the overall peak power is minimized. **b**, Decrease in peak power measurements for the varying d of the flexible and charging-shift strategy. The flexible time shifts are constrained by mobility motifs of PEV users, and the peak shaving of 47% can be achieved. The charging-shift strategy reduces the peak load by 66%. **c**, Percentage peak load of the three schemes 'optimal', 'motif blind' and 'motif aware'. The positive value on the bar implies the gap in the estimates between the optimal versus motif blind, showing the effects of lacking mobility information. The negative value on the bars implies the improvements using motif-aware information versus motif blind, customizing the optimization strategy to the mobility constraints. **d**, Peak-load saving versus PEV driver's adoption rate of the time-shift recommendations, for the motif blind and motif aware schemes when $d=4$. For each adoption rate, we randomly choose PEV drivers who accept the recommendations and average over 50 realizations. The error bar shows the 10th and 90th percentile of the savings of motif aware for the given acceptance rate. The red solid (green dash) line presents the linear fit between the saving of motif aware (blind) and adoption rate. **e**, Projected peak load saving of the flexible strategy when $d=4$ versus the number of PEV users working in the selected ZIP code. The inset shows the projected peak load versus the number of PEV users. **f**, Changes on travel times of morning and evening trips. The majority of the drivers are not influenced, the worst case is a few individuals suffering from additional 20 min delays for $d=4$. **g**, Monthly savings for varying values of d with flexible and end bound strategies.

biggest beneficiary of the PEV traveling management, the power grid operator could pay the PEV drivers to encourage their initiatives on the travel schedule shift. In addition, recent studies have suggested that gamified systems are successful in promoting behaviour that helps achieve social good[53]. More specifically, these systems encourage engagement by building raffles in which each participant has a chance to win a bigger reward with a probability proportional to their cooperation level. This type of mechanism may make incentivization more attractive in the context of PEVs and their electricity demand management.

Moreover, we analyse the potential impact of the PEV travel demand management strategy on the non-EV charging electricity demand in both of the residential regions and commercial buildings. Among the PEV drivers involved in the time-shift strategy, 60% of them are recommended to depart from home earlier and 40% are recommended to depart from home later. Considering most of the household electricity usages have two peaks, morning and evening peak[54], we argue that the time-shift strategy is more likely to relieve the morning peak of electricity usage in the residential area as a part of PEV drivers are leaving home earlier than usual. Similarly, as the PEV drivers arrive home at variant time, the time-shift strategy is also likely to relieve the evening peak load. For the commercial buildings, the power load curve is more stable during the working hours[55]. The change of arrival time to or departure time from the commercial building will have no significant influence on the power load.

Discussion

This work presents an exploratory analysis that couples two unique large datasets on urban mobility and energy consumption of electric vehicles. We address a gap in existing PEV management literature, namely the simplistic modelling of urban mobility, by generating a model of individual mobility informed by large-scale mobile phone data. Moreover, we extend the proposed methodology by demonstrating a charging management scheme and assess its applicability by using the information on individual mobility constraints of the drivers.

We evaluate recommendation schemes of time shifts in the charging sessions constrained by the individual mobility motif of PEV drivers. That is, about 30% of PEV drivers could be limited to change their travel plan due to their schedule in other activities before or after work. Following these, peak power values can be shed by up to 47%. To assess the feasibility of the recommendations, we estimate the possible monetary benefits and the travel time losses resulting from the proposed schemes. The resulting daily savings, while modest at the individual level, are certainly substantial enough to fund game-based prizes that induce cooperation and raise awareness. On the other hand, the travel time losses are almost imperceptible to the majority of the drivers, and a substantial number of drivers, in fact, benefit from the adjustment of their arrival times as it aids them to escape morning traffic.

The presented framework relies on individual location data from mobile phones and a survey of PEV drivers. We designed a Bayesian inference framework to estimate the PEV usage probability of each vehicle driver. The Bayesian inference framework relies on three properties of the PEV users: the distribution of their household income, the distribution of their daily driving distances, and the adoption rate of PEV in the city. While the profiles of PEV adopters may change due to the rebate policy, new battery technique and so on, the surveys can be updated based on sales. Thus, via changing one or more properties, it's easy to estimate the use of PEV in each ZIP code under different scenarios. Associated with the mobility information, the proposed model is adaptable and can be used to evaluate different scenario analysis of future energy demand of PEVs in time and space. In contrast, the prevalent data-driven energy demand methods mainly predict the future demand in a given region with the historical data, which can not estimate long-term consumption under different scenarios[56].

There are various avenues in which this work can be extended. Better understanding of the charging behaviour of PEV drivers and the energy demand in residential regions would complete and enrich our planning estimates. As the PEVs are spreading widely at the moment, a stronger comprehension of the tie among mobility, socioeconomic characteristics of PEV owners and the PEV incentive policies is necessary to accurately grasp the future energy demand as well as the pressure on the power grid. Another interesting avenue is to investigate PEV management when the future energy structure changes, such as the rise of the wind and solar power.

Methods

Datasets. Mobile phone activity data, also known as call detail records (CDRs), have been widely popular in the past decade, especially in the context of mobility modelling[39–41,57–59]. For this work, we make use of the CDRs for the Bay Area including approximately 1.39 million users and more than 200 million calls they made over 6 weeks. Each record contains the anonymized user ID, timestamp, duration and the geographic location of the associated cell tower. The spatial resolution is discretized to the service areas of 892 distinct cell towers. This information is used to build the TimeGeo mobility model for the Bay Area for a typical weekday. More details of the CDRs can be found in ref. [59].

PEV charging profiles provided by ChargePoint (https://www.chargepoint.com), a charging station construction company, contain information on 580,000 PEV charging sessions at commercial PEV supply equipment (EVSE) locations across the Bay Area in 2013. For each charging session, the following information is available: (i) one-time information on the EVSE location type, unique driver ID, total energy transferred and plug-in/plug-out times; and (ii) charging power readings obtained every 15 min. The locations of the charging stations are anonymized to ZIP code level. As a preprocessing step, we filter out those records lasting less than 1 min, not occuring in 2013 or with erroneous power measurements exceeding typical cable capacity and maximum charging rates.

Census and survey information used in this study consists of shapefiles describing census tracts, their population and income information[60]. The survey information is obtained from the California Plug-in Electric Vehicle Driver Survey carried out in 2013[45]. This survey contains information on various sociodemographic characteristics and travel behaviour of PEV drivers in California. We utilize information regarding income and average daily vehicle miles traveled in the estimation of PEV mobility.

Individual mobility model. From the CDRs data, we are able to extract the visited places and time during the period of the dataset for each user. With that information, TimeGeo models and integrates the flexible temporal and spatial mobility choice of the individual. In the model, each day of a week is divided into 144 discrete intervals. For each interval, the individual decides to stay or move, and then where to go if she chooses to move. To represent the movement mechanisms, TimeGeo introduces a time-inhomogeneous Markov chain model with three individual-specific mobility parameters: a weekly home-based tour number (n_w), a dwell rate (β_1) and a burst rate (β_2). $P(t)$ is defined as the global travel circadian rhythm of the population in an average week and it is different for commuters and non-commuters.

For the temporal movement choices, TimeGeo begins with determining if the individual is at home. If true, she will move with probability $n_wP(t)$, which represents her likelihood of making a trip originated from home in a time-interval t of a week. If false, she will move with probability $\beta_1 n_wP(t)$. Then, if she decides to move, she goes to *other* places with probability $\beta_2 n_wP(t)$ and goes back to home with probability $1-\beta_2 n_wP(t)$. The $P(t)$, distribution of n_w, $\beta_1 n_w$ and $\beta_2 n_w$ are illustrated in Supplementary Fig. 3.

For the spatial movement choices, TimeGeo uses a rank-based exploration and preferential return (r-EPR) to determine the next place of the individual. In detail, when the individual chooses to move to another place, she could return to a visited place or explore a new place. The model assumes that the individual explores a new place with probability $P_{new}=\rho S^{-\gamma}$, which captures a decreasing propensity to visit new locations as the number of previously visited locations (S) increases with time. The two parameters, $0<\rho\leq 1$ and $\gamma\geq 0$, are used to control the user's tendency to explore a new location and are calibrated with empirical data. If the individual decides to return, the return location is selected from the visited locations according to her visiting frequency. If she decides to explore a new location, the alternative destinations are selected according to the distance to her origin with probability $P(k)\sim k^{-\alpha}$, where k is the rank of alternative destinations, the one closest to the current location is $k=1$, the second closest $k=2$ and so on, and α is calibrated with the empirical data. More details of the TimeGeo model can be found in ref. [38]. To assess the simulation of individual mobility in Bay Area, we compare the aggregate performance of TimeGeo with NHTS and CHTS and show the results in Fig. 2 and Supplementary Figs. 1 and 2.

Electric vehicle mobility estimation. With the purpose of sampling PEV users from all vehicular drivers in Bay Area, we first extract the vehicular drivers from the entire population with the vehicle usage rate at census tract scale. Then, each vehicular driver is associated with a probability of using PEV, $P(\mathrm{EV}\,|\,I_u, D_u)$ on the basis of the driver's household income I_u and daily driving distance D_u. I_u is the random variable that denotes the income of the trip maker and follows a standard normal distribution centred at the median income of the residential tract. The median income information at tract scale is from census data[60]. $P(I_u)$ is the probability density of the household income of all trip makers in the Bay Area. Similarly, D_u is the random variable that denotes the daily travel distance of the trip maker. The visited locations of the trip maker are obtained from the TimeGeo model and the routing distance is calculated by using a publicly available online API service for routing. $P(D_u)$ is the probability density of the daily travel distance of all travelers in the Bay Area. We assume that for a given trip maker, his or her income I_u and daily travel distance D_u are independent, thereby, $P(I_u, D_u\,|\,\mathrm{EV})=P(I_u\,|\,\mathrm{EV})P(D_u\,|\,\mathrm{EV})$, that is, I_u and D_u are also conditionally independent given a PEV driver.

To estimate the probability of using PEV, $P(\mathrm{EV}\,|\,I_u, D_u)$, we begin by expressing the Bayesian relation:

$$P(\mathrm{EV}|I_u, D_u)=\frac{P(I_u, D_u|\mathrm{EV})P(\mathrm{EV})}{P(I_u, D_u)} \quad (1)$$

By imposing our aforementioned assumptions on equation (1), we have

$$P(\mathrm{EV}|I_u, D_u)=\frac{P(I_u|\mathrm{EV})P(D_u|\mathrm{EV})P(\mathrm{EV})}{P(I_u)P(D_u)} \quad (2)$$

In estimating this value, the share of PEVs within all cars in the Bay Area in 2013 is 0.62% according to the CVRP data[46], that is, $P(\mathrm{EV})=0.62\%$. We make use of the PEV driver survey information regarding income and daily travel distance,

namely $P(I_u|\mathrm{EV})$ and $P(D_u|\mathrm{EV})$, respectively. Once $P(\mathrm{EV}|I_u, D_u)$ is estimated, the probabilities are used to select the PEV drivers from all vehicular drivers. Figure 3c represents the distribution of travel distance in the morning of all vehicular and PEV commuters.

Energy consumption models. We design different energy consumption models for the four popular PEV modes in the Bay Area. In detail, we estimate the power demand of Nissan Leaf with a drivetrain model and the trip information. This drivetrain model builds the relationship between the energy consumption and two aggregate properties of the trip, the average travel speed and the route distance, which we estimate from a publicly available online API service for each PEV trip. That is,

$$E_{\mathrm{trip}}^{\mathrm{Nissan}} = f(V_{\mathrm{trip}})D_{\mathrm{trip}} \tag{3}$$

where V_{trip} and D_{trip} are the average speed and route distance of the trip respectively. $f(V_{\mathrm{trip}})$ implies the consumed power per mile (kWh mile^{-1}) when the PEV is traveling at speed V_{trip} (mile h^{-1}). However, $f(V_{\mathrm{trip}})$ depends on the battery used by the PEV model, meaning that different PEV models show different shapes of $f(V_{\mathrm{trip}})$. In this work, we fit $f(V_{\mathrm{trip}})$ with a piecewise linear function using the data observed from Nissan Leaf[5]. The curve of $f(V_{\mathrm{trip}})$ is given in Supplementary Fig. 6b and the formulation is given as follows:

$$f(V_{\mathrm{trip}}) = \begin{cases} -23.12\times10^{-3}\times V_{\mathrm{trip}} + 0.439 & V_{\mathrm{trip}} \le 6.70 \\ -8.14\times10^{-3}\times V_{\mathrm{trip}} + 0.338 & 6.70 \le V_{\mathrm{trip}} \le 12.71 \\ -0.38\times10^{-3}\times V_{\mathrm{trip}} + 0.240 & 12.71 \le V_{\mathrm{trip}} \le 21.75 \\ 2.11\times10^{-3}\times V_{\mathrm{trip}} + 0.185 & 21.75 \le V_{\mathrm{trip}} \le 60.00 \end{cases} \tag{4}$$

The consumption of Tesla Model S is the estimated by scaling the consumption of Nissan Leaf in the same trip by 1.229, as the Tesla model S consumes 22.9% more energy on average than the Nissan Leaf[51]. That is, $E_{\mathrm{trip}}^{\mathrm{Tesla}} = 1.229 f(V_{\mathrm{trip}})D_{\mathrm{trip}}$.

For the PHEVs, we introduce the charge-depleting models to estimate their energy consumptions,

$$E_{\mathrm{trip}}^{\mathrm{PHEV}} = \min\{D_{\mathrm{trip}}\times r, C_{\mathrm{PHEV}}\} \tag{5}$$

where r and C_{PHEV} are the electricity consumption rate and the battery capacity of the PHEV, respectively. It has been previously calibrated that $r = 0.288$ kWh mile^{-1} for PHEVs with 10 mi electric range; $r = 0.337$ kWh mile^{-1} for PHEVs with 20 mile electric range; and $r = 0.342$ kWh mile^{-1} for PHEVs with 40 mile electric range[49]. In the Bay Area, the two most popular modes of PHEVs are Chevrolet Volt and Toyota Prius, and their electric ranges are 40 and 10 mile, respectively. The energy consumption models used here also match with the EV efficiency ratings released by the US Department of Energy (see Supplementary Table 1).

Optimization model. We begin by discretizing a day into 15-min intervals such that each day starts at $t = 0$ and ends at $t = 95$ (ref. [32]). For each charging session i among N in a day happen in a selected ZIP code, we define t_a^i as the arrival time index, t_c^i as the time index where charging is complete, and t_d^i as the departure time index. We represent the time indices by the vector $\boldsymbol{\tau}^i$, and the power consumption by vectors $\mathbf{P}^i$ and $\mathbf{Q}^i$, all defined as follows:

$$\begin{aligned} \boldsymbol{\tau}^i &= [t_a^i, \ldots, t_c^i]^\top \\ \mathbf{P}^i &= [P_0^i, \ldots, P_{95}^i]^\top \\ \mathbf{Q}^i &= [P_{t_a^i}^i, \ldots, P_{t_c^i}^i]^\top \end{aligned} \tag{6}$$

By shifting $\mathbf{Q}^i$ within $\mathbf{P}^i$ by an amount d^i for all sessions, we can modify the overall power demand curve. We define $M^i = (t_c^i - t_a^i) + 1$ as the total number of non-zero power measurements in this charging session (that is, total number of elements in $\mathbf{Q}^i$), given that charging sessions start immediately on arrival. We enforce continuity of the charging process, the non-violation of departure times and amounts of session energy.

To capture the constraints proposed above, we introduce the following formal constraints:

$$\left.\begin{aligned} &\tau_j^i \ge 0 \\ &\tau_j^i \le 95 \\ &\tau_j^i \ge t_a^i + d^i \\ &\tau_j^i \le t_c^i + d^i \\ &t_d^i \ge t_c^i + d^i \\ &\tau_j^i < \tau_{j+1}^i \end{aligned}\right| \begin{aligned} &\forall i \in [1, N] \\ &\forall j \in [1, M^i] \end{aligned} \tag{7}$$

where d^i is the delay of the charging session of the ith PEV driver. As the mobility motif of the PEV driver limits the acceptability of the recommendations, we customize d^i for the PEV drivers with different mobility motifs as shown in Fig. 3e. Assuming that the delay of strategy is d, we introduce the following constraints:

$$d^i = \begin{cases} d & \mathrm{H-W-H} \\ \min\{0, d\} & \mathrm{H-W-O-H} \\ 0 & \mathrm{H-O-W-H\ \&\ H-O-W-O-H} \end{cases} \tag{8}$$

In this customization of delay, the drivers with mobility motif 'home-work-home' (H-W-H) could accept any change of arrival and departure time; the drivers with mobility motif 'home-work-other-home' (H-W-O-H) can not delay their departure time, that is, the delay d must be non-positive; the drivers with mobility motif 'home-other-work-home' (H-O-W-H) can not change their arrival time; the drivers with mobility motif 'home-other-work-other-home' (H-O-W-O-H) can change neither their arrival time nor departure time.

We construct the proposed constraints using a binary decision matrix to represent charging or non-charging time slots within the optimization duration. To represent the candidate time slot at which Q_j^i can be positioned, we create binary row vectors $\boldsymbol{x}_j^i$ each consisting of 95 binary decision variables: $x_{j,k}^i \in \{0,1\}$, $\forall j \in [1, M^i]$, $\forall i \in [1, N]$, $\forall k \in [0, 95]$.

$$\mathbf{X}^i = \begin{bmatrix} \mathbf{x}_1^i \\ \vdots \\ \mathbf{x}_{\mathbf{M}^i}^i \end{bmatrix} = \begin{bmatrix} x_{1,0}^i & & x_{1,95}^i \\ \vdots & \ddots & \vdots \\ x_{M^i,0}^i & & x_{M^i,95}^i \end{bmatrix} \tag{9}$$

Finally, we write the variables in the constraints given in equation (7) using the binary decision variable as follows:

$$\boldsymbol{\tau}^i = \mathbf{X}^i \begin{bmatrix} 0 \\ \vdots \\ 95 \end{bmatrix} \tag{10}$$

The aggregate power vector $\mathbf{AP}$ is given as follows:

$$\mathbf{AP} = \sum_{i=0}^{N} \mathbf{P}^i = \begin{bmatrix} \mathbf{Q}^1 \\ \vdots \\ \mathbf{Q}^N \end{bmatrix}^\top \begin{bmatrix} \mathbf{X}^1 \\ \vdots \\ \mathbf{X}^N \end{bmatrix} \tag{11}$$

The resulting formulation is a mixed-integer linear program, with decision variables $\boldsymbol{X}^i$, P_{peak} and d^i of which the latter two are integers. The problem can be proposed to minimize the daily peak load P_{peak} for a group of PEVs arriving to the same ZIP code location, subject to equation (7) and the following additional constraints:

$$\mathrm{AP}_t^i \le P_{\mathrm{peak}}, \forall i \in [1, N], \forall t \in [0, 95] \tag{12}$$

Data availability. All data needed to evaluate the conclusions in the paper are present in the paper. Additional data related to this paper may be requested from the authors.

Received: 24 May 2016; Accepted: 19 March 2018;
Published online: 30 April 2018

References

1. Michalek, J. J. et al. Valuation of plug-in vehicle life-cycle air emissions and oil displacement benefits. *Proc. Natl Acad. Sci. USA* **108**, 16554–16558 (2011).
2. Atia, R. & Yamada, N. More accurate sizing of renewable energy sources under high levels of electric vehicle integration. *Renew. Energy* **81**, 918–925 (2015).
3. Needell, Z. A., McNerney, J., Chang, M. T. & Trancik, J. E. Potential for widespread electrification of personal vehicle travel in the united states. *Nat. Energy* **1**, 16112 (2016).
4. Nykvist, B. & Nilsson, M. Rapidly falling costs of battery packs for electric vehicles. *Nat. Clim. Change* **5**, 329–332 (2015).
5. Melton, N., Axsen, J. & Sperling, D. Moving beyond alternative fuel hype to decarbonize transportation. *Nat. Energy* **1**, 16013 (2016).
6. Hu, X., Moura, S. J., Murgovski, N., Egardt, B. & Cao, D. Integrated optimization of battery sizing, charging, and power management in plug-in hybrid electric vehicles. *IEEE Trans. Control Syst. Technol.* **24**, 1036–1043 (2016).
7. DeShazo, J. Improving incentives for clean vehicle purchases in the united states: challenges and opportunities. *Rev. Environ. Econ. Policy* **10**, 149–165 (2016).
8. *Global EV Outlook: Understanding the Electric Vehicle Landscape to 2020* (International Energy Agency, 2013).
9. Hines, P., Apt, J. & Talukdar, S. Large blackouts in North America: historical trends and policy implications. *Energy Policy* **37**, 5249–5259 (2009).

10. Brummitt, C. D., Hines, P. D., Dobson, I., Moore, C. & D'Souza, R. M. Transdisciplinary electric power grid science. *Proc. Natl Acad. Sci. USA* **110**, 12159–12159 (2013).
11. Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. Catastrophic cascade of failures in interdependent networks. *Nature* **464**, 1025–1028 (2010).
12. Brummitt, C. D., D'Souza, R. M. & Leicht, E. Suppressing cascades of load in interdependent networks. *Proc. Natl Acad. Sci. USA* **109**, E680–E689 (2012).
13. Pahwa, S., Scoglio, C. & Scala, A. Abruptness of cascade failures in power grids. *Sci. Rep.* **4**, 3694 (2014).
14. McAndrew, T. C., Danforth, C. M. & Bagrow, J. P. Robustness of spatial micronetworks. *Phys. Rev. E* **91**, 042813 (2015).
15. Mwasilu, F., Justo, J. J., Kim, E.-K., Do, T. D. & Jung, J.-W. Electric vehicles and smart grid interaction: A review on vehicle to grid and renewable energy sources integration. *Renew. Sustain. Energy Rev.* **34**, 501–516 (2014).
16. Halu, A., Scala, A., Khiyami, A. & González, M. C. Data-driven modeling of solar-powered urban microgrids. *Sci. Adv.* **2**, e1500700 (2016).
17. Mureddu, M., Caldarelli, G., Chessa, A., Scala, A. & Damiano, A. Green power grids: how energy from renewable sources affects networks and markets. *PLoS ONE* **10**, e0135312 (2015).
18. Bayram, I. S., Michailidis, G., Devetsikiotis, M., Granelli, F. & Bhattacharya, S. *Control and Optimization Methods for Electric Smart Grids* 133–145 (Springer, New York, 2012).
19. Callaway, D. S. & Hiskens, I. A. Achieving controllability of electric loads. *Proc. IEEE* **99**, 184–199 (2011).
20. Moura, S. J., Fathy, H. K., Callaway, D. S. & Stein, J. L. A stochastic optimal control approach for power management in plug-in hybrid electric vehicles. *IEEE Trans. Control Syst. Technol.* **19**, 545–555 (2011).
21. Clement-Nyns, K., Haesen, E. & Driesen, J. The impact of charging plug-in hybrid electric vehicles on a residential distribution grid. *IEEE Trans. Power Syst.* **25**, 371–380 (2010).
22. Tal, G., Nicholas, M., Davies, J. & Woodjack, J. Charging behavior impacts on electric vehicle miles traveled: who is not plugging in? *Transp. Res. Rec.* **2454**, 53–60 (2014).
23. Harris, C. B. & Webber, M. E. An empirically-validated methodology to simulate electricity demand for electric vehicle charging. *Appl. Energy* **126**, 172–181 (2014).
24. Lin, Z. Optimizing and diversifying electric vehicle driving range for US drivers. *Transp. Sci.* **48**, 635–650 (2014).
25. Rajakaruna, S., Shahnia, F. & Ghosh, A. *Plug In Electric Vehicles in Smart Grids* (Springer, Singapore, 2015).
26. Tamor, M. A., Moraal, P. E., Reprogle, B. & Milačić, M. Rapid estimation of electric vehicle acceptance using a general description of driving patterns. *Transp. Res. C* **51**, 136–148 (2015).
27. Hines, P. et al. *Understanding and Managing the Impacts of Electric Vehicles on Electric Power Distribution Systems* (Univ. Vermont, 2014).
28. Yuksel, T. & Michalek, J. J. Effects of regional temperature on electric vehicle efficiency, range, and emissions in the united states. *Environ. Sci. Technol.* **49**, 3974–3980 (2015).
29. Rezaei, P., Frolik, J. & Hines, P. D. Packetized plug-in electric vehicle charge management. *IEEE Trans. Smart Grid* **5**, 642–650 (2014).
30. Valogianni, K., Ketter, W., Collins, J. & Zhdanov, D. Effective management of electric vehicle storage using smart charging in *Proc. 28th AAAI Conf. Artif. Intel.* 472–478 (2014).
31. Ma, Z., Callaway, D. S. & Hiskens, I. A. Decentralized charging control of large populations of plug-in electric vehicles. *IEEE Trans. Control Syst. Technol.* **21**, 67–78 (2013).
32. Kara, E. C. et al. Estimating the benefits of electric vehicle smart charging at non-residential locations: a data-driven approach. *Appl. Energy* **155**, 515–525 (2015).
33. Subramanian, A., Garcia, M. J., Callaway, D. S., Poolla, K. & Varaiya, P. Real-time scheduling of distributed resources. *IEEE Trans. Smart Grid* **4**, 2122–2130 (2013).
34. Yang, L., Zhang, J. & Poor, H. V. Risk-aware day-ahead scheduling and real-time dispatch for electric vehicle charging. *IEEE Trans. Smart Grid* **5**, 693–702 (2014).
35. Zakariazadeh, A., Jadid, S. & Siano, P. Multi-objective scheduling of electric vehicles in smart distribution system. *Energy Convers. Manag.* **79**, 43–53 (2014).
36. Garca-Villalobos, J., Zamora, I., San Martn, J., Asensio, F. & Aperribay, V. Plug-in electric vehicles in electric distribution networks: A review of smart charging approaches. *Renew. Sustain. Energy Rev.* **38**, 717–731 (2014).
37. Alizadeh, M. et al. Optimal pricing to manage electric vehicles in coupled power and transportation networks. *IEEE Trans. Control Netw. Syst.* **4**, 863–875 (2016).
38. Jiang, S. et al. The TimeGeo modeling framework for urban mobility without travel surveys. *Proc. Natl Acad. Sci. USA* **113**, E5370–E5378 (2016).
39. Jiang, S. et al. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proc. 2nd ACM SIGKDD Int. Worksh. Urban Computing* 2 (ACM, 2013).
40. Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiratta, S. R. & González, M. C. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transp. Res. Rec.* **2526**, 126–135 (2015).
41. Toole, J. L. et al. The path most traveled: travel demand estimation using big data resources. *Transp. Res. Part C* **58**, 162–177 (2015).
42. *Transportation Secure Data Center* (National Renewable Energy Laboratory, accessed 15 January 2015); http://www.nrel.gov/tsdc
43. *National Household Travel Survey* (US Department of Transportation, Federal Highway Administration, accessed 1 October 2016); http://nhts.ornl.gov
44. Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z. & González, M. C. Unravelling daily human mobility motifs. *J. R. Soc. Interface* **10**, 20130246 (2013).
45. *California Plug-in Electric Vehicle Driver Survey Results: May 2013* (California Center for Sustainable Energy, 2013).
46. *California Air Resources Board Clean Vehicle Rebate Project, Rebate Statistics* (Center for Sustainable Energy, accessed 5 April 2017); https://cleanvehiclerebate.org/rebate-statistics
47. *Commute Time* (Vital Signs, accessed 16 May 2017); http://www.vitalsigns.mtc.ca.gov/commute-time
48. Saxena, S., Floch, C. L., MacDonald, J. & Moura, S. Quantifying EV battery end-of-life through analysis of travel needs with vehicle powertrain models. *J. Power Sources* **282**, 265–276 (2015).
49. Wu, X., Dong, J. & Lin, Z. Cost analysis of plug-in hybrid electric vehicles using GPS-based longitudinal travel data. *Energy Policy* **68**, 206–217 (2014).
50. Yilmaz, M. & Krein, P. T. Review of battery charger topologies, charging power levels, and infrastructure for plug-in electric and hybrid vehicles. *IEEE Trans. Power Electron.* **28**, 2151–2169 (2013).
51. *Workplace Charging Challenge, Mid-program Review: Employees Plug in* (US Department of Energy, 2015).
52. *Electric Schedule e-19: Medium General Demand-metered TOU Service* (Pacific Gas and Electric Company, 2010).
53. Merugu, D., Prabhakar, B. S. & Rama, N. An incentive mechanism for decongesting the roads: A pilot program in bangalore. *Proc. ACM NetEcon Worksh.* (ACM, 2009).
54. Xu, S., Barbour, E. & González, M. C. Household segmentation by load shape and daily consumption. *Proc. 6th ACM SIGKDD Int. Worksh. Urban Computing,* 2 (ACM, 2017).
55. Luo, X., Hong, T., Chen, Y. & Piette, M. A. Electric load shape benchmarking for small-and medium-sized commercial buildings. *Appl. Energy* **204**, 715–725 (2017).
56. Xydas, E. et al. A data-driven approach for characterising the charging demand of electric vehicles: a UK case study. *Appl. Energy* **162**, 763–771 (2016).
57. Blondel, V. D., Decuyper, A. & Krings, G. A survey of results on mobile phone datasets analysis. *EPJ Data Sci.* **4**, 10 (2015).
58. Alexander, L., Jiang, S., Murga, M. & González, M. C. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transp. Res. Part C* **58**, 240–250 (2015).
59. Çolak, S., Lima, A. & González, M. C. Understanding congested travel in urban areas. *Nat. Commun.* **7**, 10793 (2016).
60. *Census Data* (United States Census Bureau, accessed 15 October 2016); https://www.census.gov/data.html
61. Fiori, C., Ahn, K. & Rakha, H. A. Power-based electric vehicle energy consumption model: Model development and validation. *Appl. Energy* **168**, 257–268 (2016).

Acknowledgements

We would like to thank ChargePoint for providing the electric vehicle charging data and Airsage for providing the call detail records used in this study. We also would like to thank S. Kiliccote and M. Tabone for their valuable feedback. This work was supported by the Siebel Energy Institute and MIT Energy Initiative.

Author contributions

Y.X., S.C. and E.C.K. conceived the research and designed the analyses. Y.X., S.C. and M.C.G. performed the analyses and wrote the paper. S.J.M. and M.C.G. provided general advice and supervised the research.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41560-018-0136-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.C.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-05690-8 OPEN

Sequences of purchases in credit card data reveal lifestyles in urban populations

Riccardo Di Clemente[1,2], Miguel Luengo-Oroz[3], Matias Travizano[4], Sharon Xu[1], Bapu Vaitla[5] & Marta C. González[1,6,7]

Zipf-like distributions characterize a wide set of phenomena in physics, biology, economics, and social sciences. In human activities, Zipf's law describes, for example, the frequency of appearance of words in a text or the purchase types in shopping patterns. In the latter, the uneven distribution of transaction types is bound with the temporal sequences of purchases of individual choices. In this work, we define a framework using a text compression technique on the sequences of credit card purchases to detect ubiquitous patterns of collective behavior. Clustering the consumers by their similarity in purchase sequences, we detect five consumer groups. Remarkably, post checking, individuals in each group are also similar in their age, total expenditure, gender, and the diversity of their social and mobility networks extracted from their mobile phone records. By properly deconstructing transaction data with Zipf-like distributions, this method uncovers sets of significant sequences that reveal insights on collective human behavior.

[1] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2] The Bartlett Centre for Advanced Spatial Analysis, University College London, London WC1E 6BT, UK. [3] United Nations Global Pulse, 46th Street and 1st Avenue, New York, NY 10017, USA. [4] GranData, 550 15th Street Suite 36C, San Francisco, CA 94103, USA. [5] Department of Environmental Health, Harvard University, 677 Huntington Avenue, Boston, MA 02115, USA. [6] Department of City and Regional Planning, Berkeley, CA 94720-1820, USA. [7] Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720-1820, USA. Correspondence and requests for materials should be addressed to M.C.Gál. (email: martag@mit.edu)

In the age of information, we leave digital traces of our everyday activities: the people we call, the places we visit, the things we eat, and the products we buy. Each of these human activities generates data that when analyzed over long periods yield a comprehensive portrait of human behavior[1–6].

In the past decade, call detailed records (CDRs) have been of paramount importance to understand the daily rhythms of human mobility[7–11]. By properly analyzing billions of digital traces, our modern society has a whole framework to analyze wealth[12], socio-demographic characteristics[13], and to better tackle the origins of urban traffic[14,15]. By contrast, we still need to better exploit the credit card records (CCRs) to uncover the behavioral information they may hide. Main uses of CCRs have been to measure similarity in purchases via affinity algorithms[16,17]. Recent research has also shown that credit card data can be used analogously to mobile phone data to detect human mobility. Namely, the CCRs inform us about the preferred transitions between business categories, identifying the unevenness of the spatial distributions of people's most preferred shopping activities[18], and to enrich urban activity models. Consumers' habits are shown to be highly predictable[19], and groups that share work places have similar purchase behavior[20]. These results allowed defining the spatial–temporal features to improve the estimates of the individual's financial well-being[21].

It has been measured by individual surveys and confirmed by credit card and cash data that the vast majority of daily purchases is dominated by food and then followed by mobility and communication–social activities[13,22]. Their frequency seems to follow Zipf distribution, meaning that the most frequent category of purchases will occur approximately twice as often as the second most frequent category, three times as often as the third, etc. Grouping the consumers depending on their socio-demographic attributes preserves the Zipf-like behavior and dominant purchase (food). For each group, there is a peculiar order in the abundance of less frequent category. As pointed out by Lenormand et al.[13] and Sobolevsky et al.[23] this depends on the socio-demographic features such as income, gender, and age.

Hence, the challenge at hand is to obtain meaningful information within these highly uneven spending frequencies to capture a comprehensive picture of their shopping styles related to socio-economic dynamics within the city.

A similar challenge appears in the sequence of diseases in the medical records[24] or phenotype associations with diseases[25]. Existing approaches cluster patients based on their historical medical records described by the International Classification of Diseases. In this case, the frequency-inverse document frequency (TF-IDF) ranking is used to eliminate redundant information.

In the matter of uneven word frequency in the text corpora[26], Bayesian inference methods have been used to detect the hidden semantic structure. In particular, the latent Dirichlet allocation (LDA)[27] is a widely used method for the detection of topics (ensemble of words) from a collection of documents (corpus) that best represent the information in data sets.

However, both of the above-mentioned approaches do not take into account the temporal order in the occurrence of the elements. Our goal is to eliminate redundancy while detecting habits and keeping the temporal information of the elements, which in the case of purchases are an important signature of an individual's routine and connect them to their mobility needs. In this work, we identify significantly ordered sequences of transactions and group the users based on their similarity. This allows offering deeper description of consumer behavior, unraveling their routines.

In this work, we are interested in uncovering diverse patterns of collective behavior extracted from this data. Specifically, how the digital footprint of CCRs can be used to detect spending habits, reflecting interpretable lifestyles of the population at large. By integrating credit card data with demographic information and mobile phone records, we have a unique opportunity to tackle this question.

The presented method is able to deconstruct Zipf-like distribution into its constituent's distributions, separating behavioral groups. Paralleling motifs in network science[28], which represent significant subnetworks, the uncovered sets of significant sequences are extracted from the labeled data with Zipf-type distribution. Applied to CCRs, this framework captures the semantic of spending activities to unravel types of consumers. The resulting groups are further interpreted by coupling together their mobile phone data and their demographic information. Consistently, individuals within the five detected groups are also similar in age, gender, expenditure, and their mobility and social network diversity. We show that the selection of significant sequences is a critical step in the process; it improves the TF-IDF method that is not able to discern the spending habits within the data. Remarkably, our results are comparable with the ones obtained by LDA, with the added advantage that it takes into account the temporal sequence in the activities.

Results

Data analysis. We analyze individual CCR transactions over 10 weeks in 150,000 users who live in one of the most populated cities in Latin America (Mexico City, Mexico). The data set contains age, gender, and residential zipcode of the users (Supplementary Figure 1A–C). For each user, we analyze the chronological sequence of their transactions and the associated expenditure labeled with the transaction type via a Merchant Category Code (MCC)[29]. The purchase entries are aggregated by the user and are temporally ordered with respect to each day. For one-tenth of the analyzed users, we also have their CDR data over a period of 6 months (overlapping the CCR time period), including time, duration, location of the calls, and identification of the receiver. While payment with cards and electronic payment terminals are being promoted in the region to improve financial inclusion, credit card adoption rates remain relatively low at 18% for the population[30]. First, we check how representative the CCR users are within the city. We observe the correlation between the median CCR expenditure in the data set at the district level and the average monthly wage in the same district, according to the census (Fig. 1a) (Source: INEGI, National Survey of Occupation and Employment (ENOE) and population aged 15 years and older.). The monthly expenditure of card users is high in relation to their monthly wages, indicating that the adoption of credit cards predominantly occurs among users with higher wages in each district. However, our users' sample spans over all the city districts with different income levels. We observe that wider adoptions of credit card are across male and young adults (aged 35–50 years) in each district (Supplementary Figure 1B–F). The spending patterns in the CCRs reveal that the frequency of the purchase types follows Zipf's law (Supplementary Figure 2A). The majority of shoppers use more frequently the top 20 transactions codes presented in Fig. 1b, among hundreds of possible MCCs. Moreover, slight variations emerge in this trend when dividing the population by wealth, age, and gender (Fig. 1c). In general, transaction codes related to food, mobility, and communication, in that order, dominate the number of top transactions in all groups and the number of transactions per day; for each user is not affected by any socio-demographic category (Supplementary Figure 2B, C).

Credit card transaction codes as sequence of words. Our main goal is to amplify the signal in the data to identify the individuals'

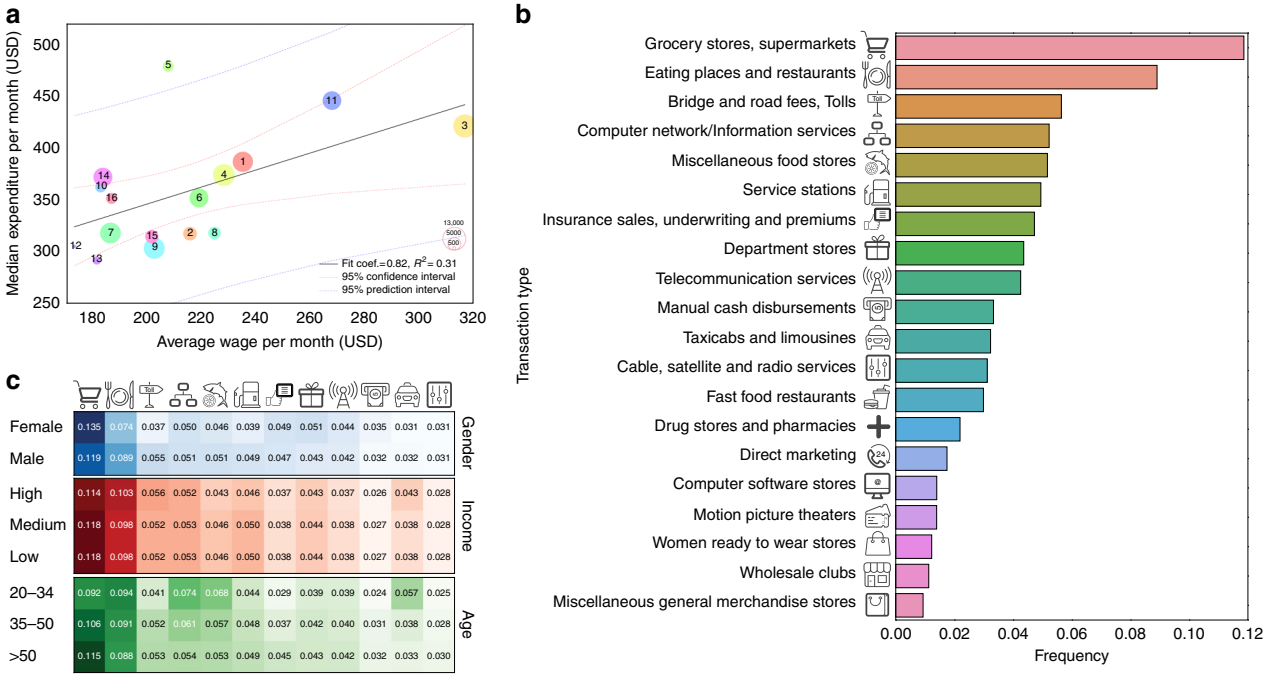

Fig. 1 Transaction frequency by type and their demographics. **a** User median expenditure per month in CCR transactions vs. the average monthly wage in their district of residence. The color and the number represent different districts of Mexico City (see Supplementary Figure 1), and the size of the circles is proportional to the number of users in the district. **b** Transactions by type as defined by MCC[29]. **c** Comparison of frequencies by transaction types (same as in **b**) separating users in groups according to their gender, income, and age. The share of transaction frequency is distributed similarly among different groups. The icons used in this figure are work of Azaze11o/Shutterstock.com

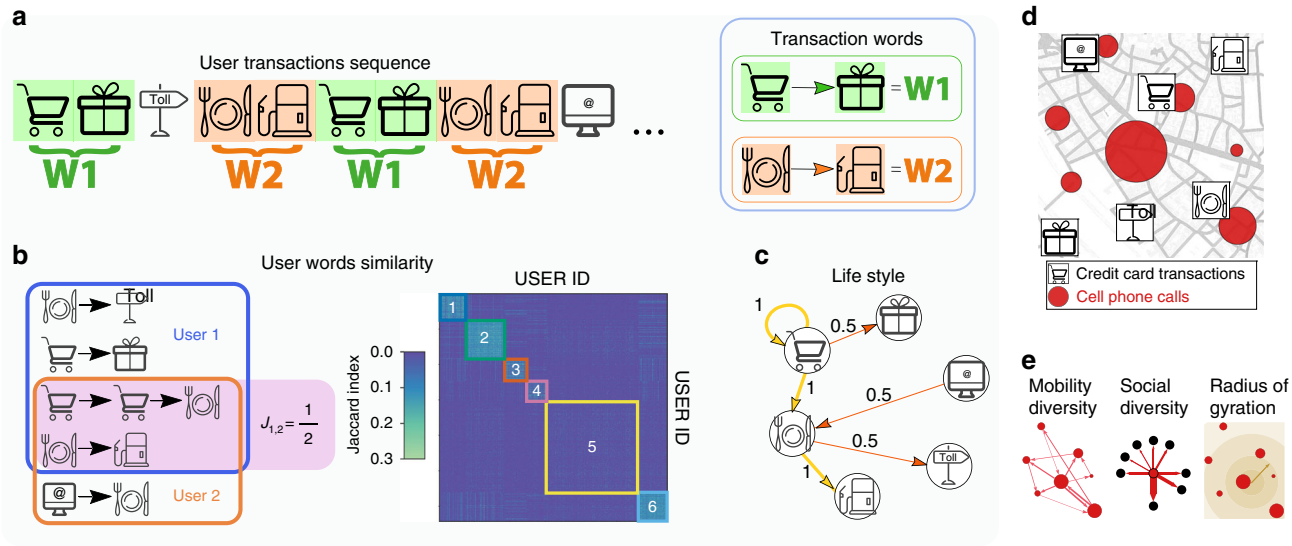

Fig. 2 Methods and metrics. **a** Schematic representation of the Sequitur's algorithm applied to a sequence of transactions of one user to detect words and identify significant transaction sequences in the data set. **b** Calculation of the similarity between two users (left) based on the Jaccard index of their significant sequences to define the matrix of users' similarity (right). Group of users are detected based on similar sequences of transactions. **c** Lifestyle representation based on sample users 1 and 2 of **b**. **d** Example of traces of CDR and CCR data for the user. **e** Metrics adopted for the analysis of CDR data. The icons used in this figure are work of Azaze11o/Shutterstock.com

expenditure habits hidden in the non-uniform distribution of transaction types present in a Zipf's type of distribution. The first step in this direction is to transform the chronological sequence of user MCC codes into a sequence of symbols given by the transaction codes (Fig. 2a). We apply the Sequitur algorithm[31] to infer a grammatical rule that generate words, defined as MCC symbols that repeat in sequence. The result of this process applied recursively is a compression of the original sequence with new symbols called words, which offer insights into the repeated sequences of transactions. We take each word as a routine in shopping, as they are a chronological sequence of two or more MCCs that appear frequently. We detect more than 10,000 different words also following a Zipf-type distribution, as presented in Fig. 3. We noticed that the inter-time transactions between

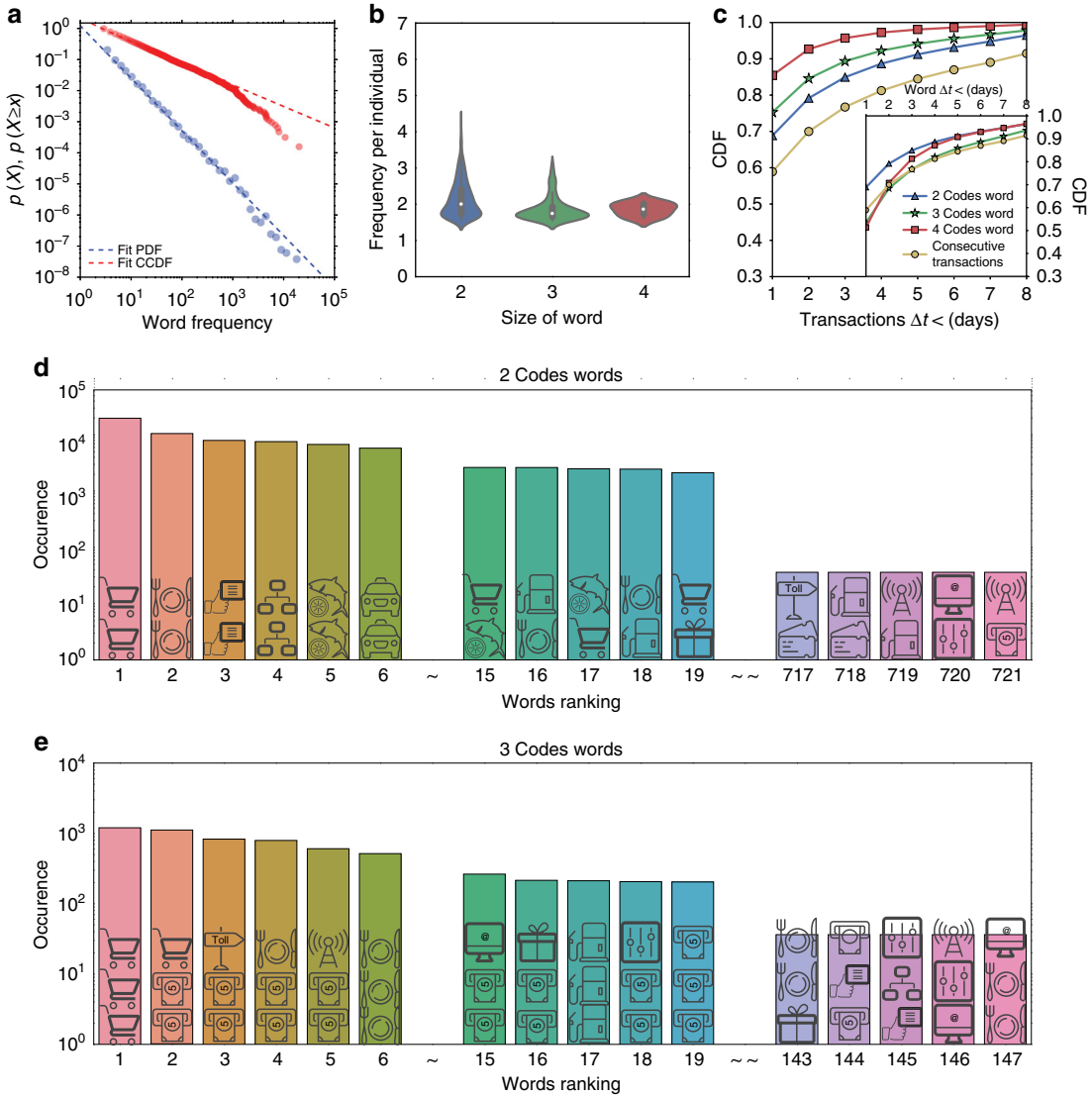

Fig. 3 Semantic analysis of transaction sequences. **a** Probability density function plot of the occurrence of words $\{w_i\}$ and its complementary cumulative distribution; the probability distribution words manifest a power-law behavior $p(w_i) \propto x_i^{(-1.70)}$, with x_i frequency of the $\{w_i\}$ and Kolmogorov–Smirnov distance $D_n = 0.014$. **b** Distribution of the occurrence of words in the transaction sequences by the word length. **c** Inter-time transactions between purchases. The purchases within each word are more likely to occur within a day with respect to two random consecutive transactions. **c** (inside) Moreover, the purchase time to accomplish a word completely is less with respect to two random consecutive transactions. **d**, **e** Examples of words composed by two and three codes, respectively, ordered by the number of occurrences. The icons used in this figure are work of Azaze11o/Shutterstock.com

word purchases are smaller with respect to two random consecutive transactions. Moreover, the time to perform an n-transaction word, defined as the time between the first and the last purchase of the word, is smaller than the time of two consecutive transactions picked randomly (Fig. 3c). The set of words $\{w_i\}$ for user i are significant only if their occurrence differs from the outcome of a random process with the same number of transactions per type. To detect the words that are significant, we generate 1000 randomized code sequences for each user. For each realization, we apply the Sequitur algorithm to define the words in the randomized sequences and evaluate the significance level of the user's words by computing the z-score of the occurrence of the real words with respect to the randomized ones. Z-score test needs to be performed on a Gaussian distribution of word occurrence. The word-occurrence distribution of simulated samples has in general a normal shape. But in several cases, the frequency of the generated words has a small number of occurrences; in Supplementary Figures 3, 4, we show the robustness of a z-score benchmark to assess the word significance for non-Gaussian distributions. We extract for each user, the set of significant words with z-score >2, defined as $\{w_i\}$. The selected words represent the shopping routines that indicate informative choices in the user's spending behavior (see Supplementary Figure 5), given that their occurrence vary from the mean by two standard deviations. In the Supplementary Figure 5D, E, we analyze the number of valid users with at least a significant word depending on the z-score threshold.

The lifestyles. With these meaningful samples, we can now measure the similarity between shopping behaviors among users. To that end, we decompose each significant word as direct links between its transaction codes. Each user is represented by a directed network, in the space of MCC, that collects all the links present in the user's words. We then calculate the Jaccard

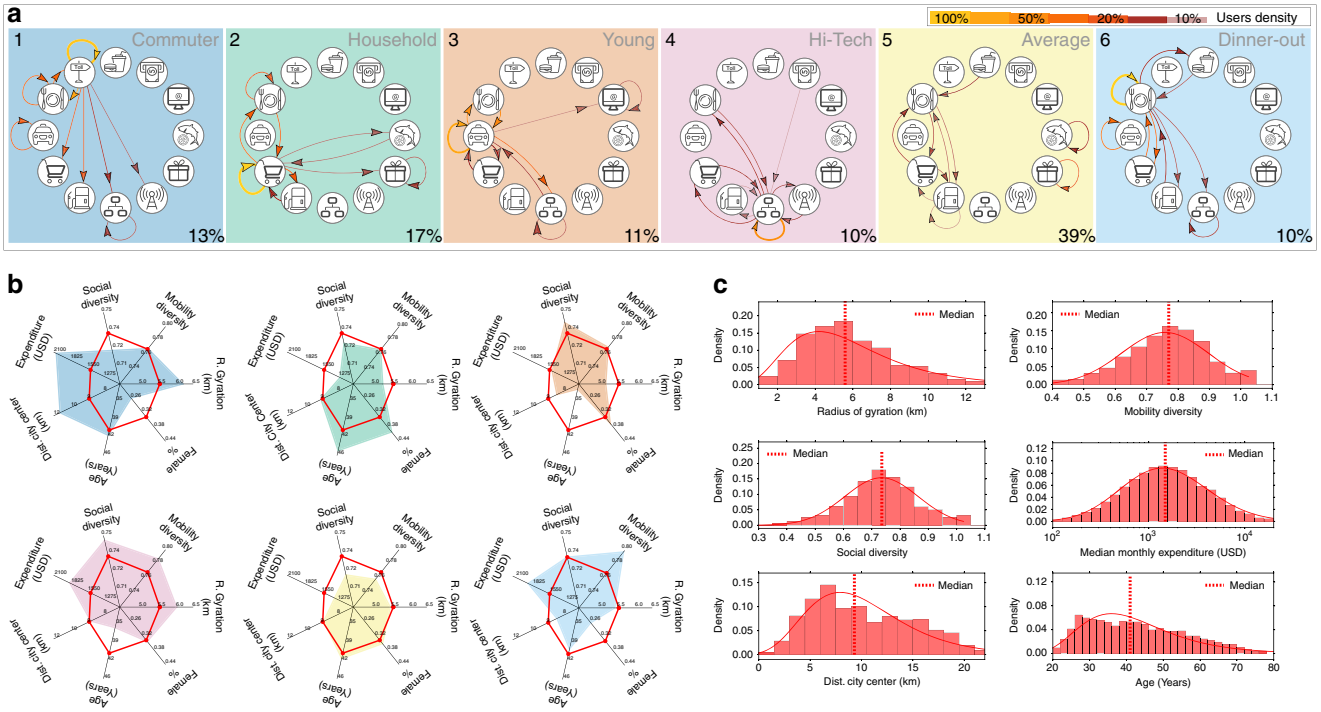

Fig. 4 Identified lifestyles I. **a** Groups based on their spending habits. We show the top 10 most frequent spending sequences of the users in each group, representing more than 30% of users' shopping routines. The percentage of the total users in each group is shown in the bottom-right corner. **b** Comparison of the median of socio-demographic variables within each group with respect to the median of all users is in red. (The color of the radar plot identifies the spending habits in **a**.) **c** Distribution of individual characteristics among users: gender, radius of gyration, mobility diversity, social diversity, median expenditure by month, average distance traveled from the center of residence zipcode to the city center, and age (see Supplementary Figures 11–16, 21 for further information). The icons used in this figure are work of Azaze11o/Shutterstock.com

similarity coefficient between all the users to compare the set of links in their networks (see the illustration of the method in Fig. 2b). Since user networks have a low degree, our similarity measure is not sensitive to the sets' size (Supplementary Figure 6C). Moreover, our results are in agreement with the ones that use the turnover component of Jaccard dissimilarity index[32], which is less susceptible to the sets' size (see Supplementary Figure 6). Owing to the Jaccard index, we obtain the matrix M of users' similarity in shopping sequences.

Finally, we identify the groups in this matrix by applying a parallel Louvain algorithm for faster unfolding of communities in M[33,34]. The same clusters appear with Leading Eigenvector[35] and Walking Trap[36] (Supplementary Figures 7, 8). We detect six clusters or groups of users who share similarities in their spending habits; one of the six encloses unlabeled users who are close to the average behavior, while the other five present interesting behavioral preferences as confirmed later by their demographics and their mobile phone records.

Figure 2c shows the group's shopping habits. The weight of the arrows between two codes represents the fraction of users of a given cluster that have the given transaction sequence. This schematic representation of the group's routines is possible because our method firstly, detects the most significant sequences of transactions and secondly preserves the temporal information embedded in word as the ordered sequence of transaction.

Coupling credit card data with mobile phone data. In order to gather a more comprehensive portrait of the users' behavior, we couple the information of the CCR users with their CDR data (Fig. 2d, e). From the mobile phone data, we analyze the basic characteristics of an individual's social contacts and their mobility network with well-established metrics, namely, social diversity, homophily, mobility diversity, radius of gyration[8,37], tower residual activity[38], and mobility behavioral pattern. Social network diversity is the entropy associated with the number of individual i's communication events with their reciprocal contacts divided by the number of contacts[1]. Homophily, in the call graph from the mobile phone data, is a metric that investigates whether or not two users in the same cluster have a higher probability of contacting each other. Mobility diversity is measured via entropy in the number of trips between locations normalized by the number of visited locations[37]. Ego networks are defined by a focal node (ego) and the users to whom the ego is directly connected. High diversity score in the ego network implies that individuals split three times evenly among their social ties. High diversity in the network of trips among locations means that individuals distribute their number of trips evenly among their visited urban locations. Radius of gyration, in turn, defines the radius of the circle within which they are more likely to be found, it is centered in all the visited locations of i and weighted by the number of mobile phone records in each location[8]. From the urban science perspective, we investigate the cell towers' residual activity as defined by Toole et al.[38] to determine whether users who belong to the same cluster tend to aggregate in a specific area of the city. Residual activity can be interpreted as the amount of mobile phone activity in a region relative to the expected mobile phone activity in the whole city. Finally, to assess the mobility behavioral pattern, we analyze the portion of explorers and returners among the users[39]. Returners are the users who limit much of their mobility to a few locations; in contrast, the explorers have a tendency to wander between a larger number of different locations.

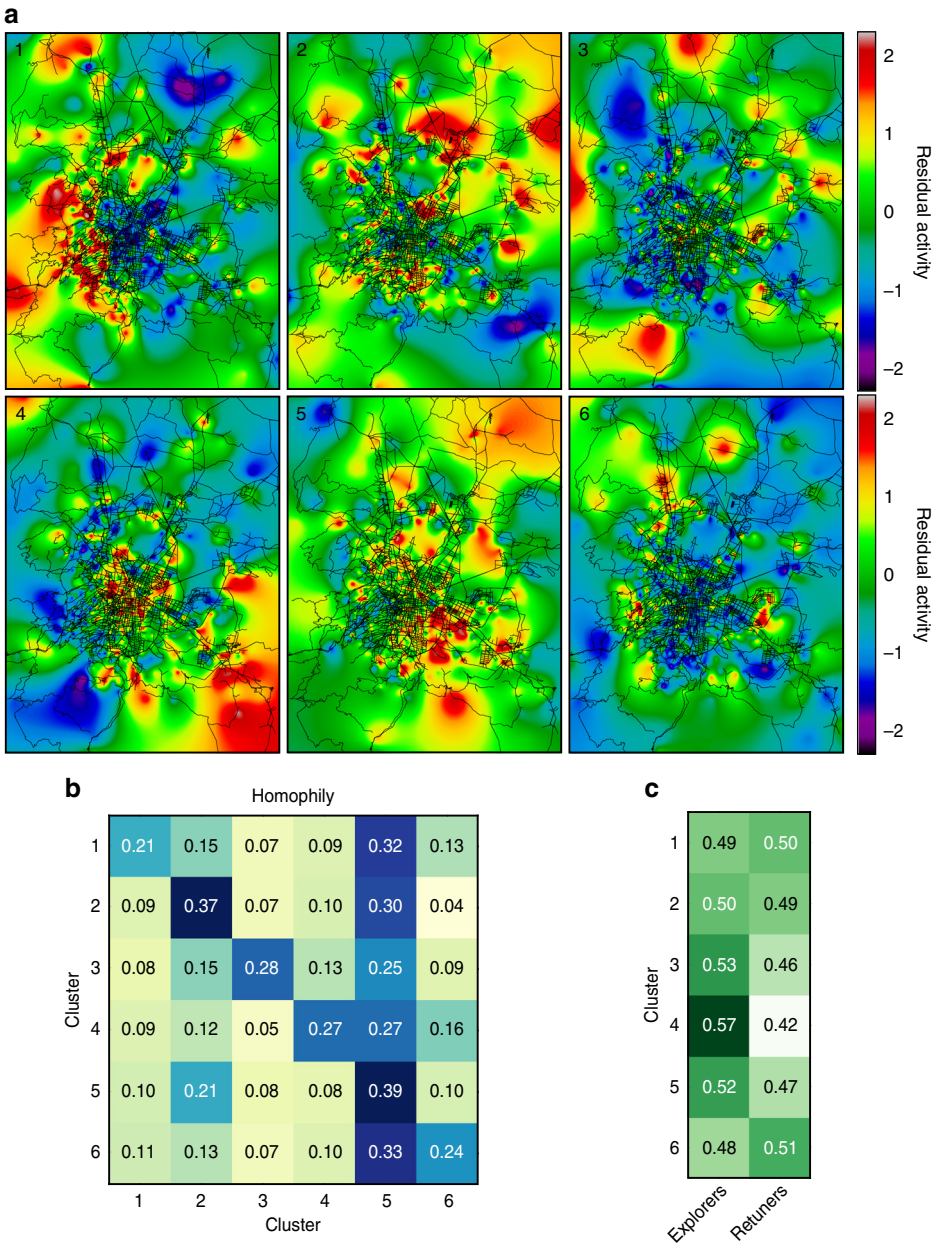

Fig. 5 Identified lifestyles II. **a** Cell towers residual activity by clusters. **b** Clusters' homophily. As expected, each user tends to contact the users that belong to the same clusters or cluster 5 "uncategorized," which is the cluster with the highest number of users. Remarkably, there is a slight preference to contact cluster 2, the homemakers, which represent the oldest group. **c** Distribution of returners and explorers across the clusters (see Supplementary Figure 11–16, 21 for further information). Maps in this figure were created using the software QGIS using OpenStreetMap data

Discussion

Five of the six clusters detected depict a particular lifestyle on how individuals spend their money, move, and contact other individuals. One transaction type is at the core of the spending activities in each group, and 90% of the users within the cluster have it repeated as a sequence (or significant word, represented by yellow loop in Fig. 4a). This transaction also appears in more than 45% as starting or ending transaction of the sequences of other types of transactions within the group (Fig. 4a). The users clustered by using our approach have relatively high Shannon entropy in their transactions and a Sequitur compression ratio of 1.5 or larger (Supplementary Figure 10). Cluster 5 aggregates the uncategorized users. In particular, users who belong to this cluster have less than five significant sequences and less variation in their expenditure types (Supplementary Figures 7–9).

Figures 4b, 5 show that each cluster reveals consistent relations between expenditure patterns and age, mobility, and social networks of their members, hinting that the method actually unravels behavioral groups in the data or actual lifestyles. Cluster 1 aggregates users whose core transaction is toll fees, and accordingly we label them as Commuters. They live furthest from the city center, expend the most, travel longest distances, and are majority male, as confirmed from the analysis of the radius of gyration and the residual activity in Fig. 5a. Conversely, users in the cluster 2 or homemakers have grocery stores as a core transaction. They represent the oldest group with least expenditure, mobility, and a larger share of women. Although the social network of this cluster manifests a lower diversity, there is a slight preference in the homophily matrix in this cluster, suggesting that the few connections are cluster transversal (Fig. 5b). Younger

users are split into two groups (clusters 3 and 4) with different values in their expenditure, and social and mobility diversity. Cluster 3 is labeled as Youths because it has the youngest individuals with taxis as their core transaction. Cluster 4 is close in age to cluster 3, but has computer networks and information services as a core transaction. They are labeled as Tech users and have higher than average expenditure and higher diversity in their social contacts and mobility networks. The residual activity (Fig. 5a) suggests that their movements are within the city center. Moreover, clusters 3 and 4 are the only ones with a majority of explorers within their users, supporting the lifestyle fingerprint (Fig. 5c). Finally, cluster 6, labeled as Diners, aggregates middle-aged users who have restaurants as their core transaction with high mobility diversity and higher expenditures (see Supplementary Figures 11–16, 21 for further information).

We compare the detected groups with the ones extracted via the patients' stratification technique to analyze the health records[24]. Instead of applying the Sequitur algorithm to assess the likelihood of a given sequence of codes, we compute, for each user's code, the TF-IDF frequency measure[40], which rewards high code frequency in the individual records and penalizes high prevalence across the all user's history. The similarity matrix among users is based on the cosine similarity in the space of the code frequency TF-IDF. The clusters extracted via this method (Supplementary Figure 17) do not have socio-demographic similarities, and the characteristics of the members within each group average similarly to the population. Moreover, TF-IDF does not disentangle the Zipf distribution (Supplementary Figure 17c), meaning each cluster keeps the same overall transaction frequency.

Furthermore, we compare our clusters with the LDA[27,41]. This method first identifies five topics represented by an ensemble of MCCs. Each user is identified by a vector v_i weighting the mixture of those five topics. We compute the users' similarity matrix using Jensen–Shannon divergence[42] among v_i. Finally, we perform the Louvain algorithm over the matrix. Four of the seven identified clusters (1, 2, 3, and 7), in the Supplementary Figure 18, are similar to our clusters (1, 2, 3, and 6). Furthermore, the LDA is able to untangle the similar variance from the Zipf distribution (Supplementary Figure 18C) compared with our method (Supplementary Figure 13B).

With respect to the above-mentioned methods (TD-IDF and LDA), our approach deconstructs the Zipf distribution into the constituents' behavior (see Supplementary Figure 13B). The resulting clusters of the latter are comparable with our method. Furthermore, our framework is able to capture the routines of each cluster as ordered sequence of transaction; this temporal information is lost using the above-mentioned approaches. These tests stress the effectiveness of our method.

Finally, we apply our framework to another minor city of Mexico: Puebla (Supplementary Figure 19–21). As already shown by Sobolevsky et al.[23], different cites manifest a general behavior in terms of spending patterns, maintaining some unique characteristics. In Puebla, we detect six clusters; four of them share similar routines and attributes to the main city (Mexico City clusters (2, 3, 5, and 6)). Comparing the median absolute deviation of each cluster, it is possible to assess the diversity of every socio-demographic attribute (Supplementary Figure 21). In particular, the routines of Commuters' clusters are identifiable in both of the cities, with some difference in the mobility attributes. Finally, in Puebla, the Youth cluster is replaced with one with different core transactions in the miscellaneous food store and insurance instead of taxi and restaurants. This result stresses how our framework can capture cities' differences in terms of spending patterns, providing a tool to enrich the urban activity models.

Taken together, we present a method to detect behavioral groups in chronologically labeled data. It could be applied also to similar data sets with Zipf-like distributions, such as disease codes in patients' visits[24,25] or law-breaking codes in police databases[43]. Given the ubiquitous nature of the CCR transaction distribution by type[23], similar groups could be detected and compared among cities worldwide. Analogous to the price index that uses online information to improve survey-based approaches to measure inflation[44], the meaningful information of groups extracted from the CCR data can be used to compare consumers worldwide[4]. Interesting avenues for the application of this method are policy evaluation of macroeconomic events such as inflation and employment and their effects on the spending habits of various groups[45].

Methods

Credit card data sets. Credit card data sets, also referred to as CCRs, used in this study consists of 10 weeks of records, starting from the 1st week of May 2015, of all the credit card users of a particular bank across each subject city. Each individual CCR consists of a hashed user identification string, the time stamp of the transaction, the associated expenditure labeled with the transaction type via an MCC[29], and the transaction amount. For each user, the data set contains age, gender, and residential zipcode of the user (Supplementary Figure 1A–C).The purchase entries are aggregated by user and are temporally ordered with respect to each day.

Mobile phone data sets. Mobile phone data sets, also referred to as CDRs, used in this study consist of 6 months of records, starting form March 2015, of all mobile phone users of a particular carrier across each subject city. Each individual CDR consists of a hashed user identification string, a time stamp, and location of the activity. The spatial granularity of the data varies between cell tower levels.

Census data. The census data used in this work were download from the Instituto Nacional de Estadística Geografía e Informática, México (http://www.inegi.org.mx/ last checked 13/Jun/2018). In particular, the data regarding the population distribution among the districts are from "Source: INEGI, Intercensal Survey 2015" and the data on the district income are from "Source: INEGI, National Survey of Occupation and Employment (ENOE). Population aged 15 years and older."

Data availability. For contractual and privacy reasons, the raw data is not available. Upon request, the authors can provide the data of the matrix of user similarity along with appropriate documentation for replication.

Received: 7 August 2017 Accepted: 6 July 2018
Published online: 20 August 2018

References

1. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science* **328**, 1029–1031 (2010).
2. Giles, J. et al. Making the links. *Nature* **488**, 448–450 (2012).
3. Lazer, D. et al. Life in the network: the coming age of computational social science. *Science* **323**, 721 (2009).
4. Mervis, J. Agencies rally to tackle big data. *Science* **336**, 22–22 (2012).
5. "Sandy" Pentland, A. The data-driven society. *Sci. Am.* **309**, 78–83 (2013).
6. Vespignani, A. Modelling dynamical processes in complex socio-technical systems. *Nat. Phys.* **8**, 32 (2012).
7. Blondel, V. D., Decuyper, A. & Krings, G. A survey of results on mobile phone datasets analysis. *EPJ Data Sci.* **4**, 10 (2015).
8. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779 (2008).
9. Jiang, S. et al. The timegeo modeling framework for urban motility without travel surveys. *Proc. Natl Acad. Sci. USA* **113**, E5370–E5378 (2016).
10. Song, C., Qu, Z., Blumm, N. & Barabasi, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).
11. Toole, J. L., Herrera-Yaqüe, C., Schneider, C. M. & González, M. C. Coupling human mobility and social ties. *J. R. Soc. Interface* **12**, 20141128 (2015).
12. Blumenstock, J., Cadamuro, G. & On, R. Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).
13. Lenormand, M. et al. Influence of sociodemographic characteristics on human mobility. *Scientific Rep.* **5**, https://doi.org/10.1038/srep10075 (2015).
14. Çolak, S., Lima, A. & González, M. C. Understanding congested travel in urban areas. *Nat. Commun.* **7**, 10793 (2016).

15. Louail, T. et al. From mobile phone data to the spatial structure of cities. *Scientific Rep.* **4**, https://doi.org/10.1038/srep05276 (2014).
16. Pennacchioli, D., Coscia, M., Rinzivillo, S., Giannotti, F. & Pedreschi, D. The retail market as a complex system. *EPJ Data Sci.* **3**, https://doi.org/10.1140/epjds/s13688-014-0033-x (2014).
17. Solomon, M. R., Dahl, D. W., White, K., Zaichkowsky, J. L. & Polegato, R. *Consumer Behavior: Buying, Having, and Being*, Vol. **10** (Pearson, Upper Saddle River, 2014).
18. Yoshimura, Y., Sobolevsky, S., Bautista Hobin, J. N., Ratti, C. & Blat, J. Urban association rules: uncovering linked trips for shopping behavior. *Environ. Plan. B* **45**, 367–385 (2016).
19. Krumme, C., Llorente, A., Cebrian, M., Pentland, A. & Moro, E. The predictability of consumer visitation patterns. *Scientific Rep.* **3**, https://doi.org/10.1038/srep01645 (2013).
20. Dong, X. et al. Social bridges in urban purchase behavior. *ACM Trans. Intell. Syst. Technol.* **9**, 1–29 (2017).
21. Singh, V. K., Bozkaya, B. & Pentland, A. Money walks: Implicit mobility behavior and financial well-being. *PLoS ONE* **10**, e0136628 (2015).
22. Matheny, W., O'Brien, S. & Wang, C. The state of cash: preliminary findings from the 2015 diary of consumer payment choice. *FedNote* **3**, http://www.frbsf.org/cash/files/FedNotes-The-State-of-Cash-Preliminary-Findings-2015-Diary-of-Consumer-Payment-Choice.pdf (2016).
23. Sobolevsky, S. et al. Cities through the prism of people's spending behavior. *PLoS ONE* **11**, e0146291 (2016).
24. Roque, F. S. et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput. Biol.* **7**, e1002141 (2011).
25. Hidalgo, C. A., Blumm, N., Barabási, A.-L. & Christakis, N. A. A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* **5**, e1000353 (2009).
26. Piantadosi, S. T. Zipf's word frequency law in natural language: a critical review and future directions. *Psychon. Bull. Rev.* **21**, 1112–1130 (2014).
27. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
28. Milo, R. et al. Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
29. Visa Commercial Solution, Merchant Category Codes for IRS Form 1099-MISC Reporting Visa U.S.A. Inc (USA 2004)
30. PYMNTS.com. Global Cash Index Mexico Analysis. Technical Report, pymnts http://pymnts.fetchapp.com/files/442f09 (2017).
31. Nevill-Manning, C. G. & Witten, I. H. Identifying hierarchical structure in sequences: a linear-time algorithm. *J. Artif. Intell. Res.* **7**, 67–82 (1997).
32. Baselga, A. The relationship between species replacement, dissimilarity derived from nestedness, and nestedness. *Glob. Ecol. Biogeogr.* **21**, 1223–1232 (2012).
33. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
34. Staudt, C. L. & Meyerhenke, H. Engineering parallel algorithms for community detection in massive networks. *IEEE Trans. Parallel Distrib. Syst.* **27**, 171–184 (2016).
35. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, https://doi.org/10.1103/PhysRevE.74.036104 (2006).
36. Pons, P. & Latapy, M. in *Computer and Information Sciences—ISCIS 2005* (eds Yolum, P. et al.) 284–293 (Springer, Berlin, Heidelberg, 2005).
37. Pappalardo, L., Pedreschi, D., Smoreda, Z. & Giannotti, F. Using big data to study the link between human mobility and socio-economic development. In *2015 IEEE International Conference on Big Data (Big Data)* 10.1109/BigData.2015.7363835, 871–878 (2015).
38. Toole, J. L., Ulm, M., González, M. C. & Bauer, D. Inferring land use from mobile phone activity. In *Proc. ACM SIGKDD International Workshop on Urban Computing—UrbComp'12*, https://doi.org/10.1145/2346496.2346498 (2012).
39. Pappalardo, L. et al. Returners and explorers dichotomy in human mobility. *Nat. Commun.* **6**, https://doi.org/10.1038/ncomms9166 (2015).
40. Robertson, S. E. & Jones, K. S. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.* **27**, 129–146 (1976).
41. Krestel, R., Fankhauser, P. & Nejdl, W. Latent dirichlet allocation for tag recommendation. In *Proc. 3rd ACM Conference on Recommender Systems—RecSys '09*, https://doi.org/10.1145/1639714.1639726 (2009).
42. Lin, J. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).
43. Schuerman, L. & Kobrin, S. Community careers in crime. *Crime Justice* **8**, 67–100 (1986).
44. Cavallo, A. Scraped data and sticky prices. *Rev. Econ. Stat.* https://doi.org/10.3386/w21490 (2016).
45. Vaitla, B. et al. *Big Data and the Well-being of Women and Girls: Applications on the Social Scientific Frontier*. Technical Report, Data2x http://data2x.org/wp-content/uploads/2017/03/Big-Data-and-the-Well-Being-of-Women-and-Girls.pdf (2017).

Acknowledgements

This work was supported by the Gates Foundation (grant OPP1141325) and United Nations Foundation (grant UNF-15-738). We acknowledge Rebecca Furst-Nichols and Jake Kendall for planning the study. We also thank Edward Barbour, Philip Chodrow, and Balazs Lengyel for the helpful discussions. Views and conclusions in this document are those of the authors and should not be interpreted as representing the policies, either expressed or implied, of the sponsors. Riccardo Di Clemente as Newton International Fellow of the Royal Society acknowledges support from the Royal Society, the British Academy, and the Academy of Medical Sciences (Newton International Fellowship, NF170505). The icons used in this paper are work of Azaze11o/Shutterstock.com.

Author contributions

R.D.C analyzed the data, performed the research, and created the maps, S.X. developed and tested the machine learning algorithm; R.D.C., M.T., M.L.-O., B.V., and M.C.G. planned the study; R.D.C. and M.C.G. designed the study and wrote the paper; and M.C.G. coordinated the study. All authors gave their final approval for publication.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05690-8.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.