Estimating the presence of people in buildings using Call Detail Records

by

Siddharth Gupta

Submitted to the Department of Civil and Environmental Engineering
in partial fulfilment of the requirements for the degree of

Master of Science in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2017

Author .... Signature redacted ....
Department of Civil and Environmental Engineering
May 12, 2017

Certified by .... Signature redacted ....
Marta C. González
Associate Professor
Thesis Supervisor

Accepted by .... Signature redacted ....
Jesse Kroll
Chairperson, Department Committee on Graduate Theses

Estimating the presence of people within blocks using Call Detail Records

by

Siddharth Gupta

Submitted to the Department of Civil Engineering
on May 12, 2017 in partial fulfillment of
the requirements for the degree of
Master of Science in Transportation

ABSTRACT

As geographic data about individual movement become increasingly available, they open up the possibility of understanding and modeling urban mobility patterns. While no all-encompassing dataset regarding mobility is available, this study explores how Call Detail Records (CDRs), a highly ubiquitous dataset, can be leveraged to create models that can reproduce mobility patterns observed from time consuming, capital-intensive and infrequent travel surveys. While mechanisms have been proposed for reproducing particular characteristics of individual mobility, this is the first attempt to generate all mobility patterns at fine spatial and temporal scales at the level of individual buildings.

Two shortcomings of any dataset include spatial uncertainty at very high resolution and the presence of high-fidelity traces for only a fraction of the population. While the proposed model addressed the former to some extent by providing high accuracy counts at the level of census tracts, a separate method has been explored to address this along with the latter phenomenon. To achieve this, the study leverages hyper-local datasets such as building footprints and places of interest. In the absence of primary datasets, the study is able to provide a model to estimate of the presence of people at the level of individual buildings.

Hence, this study provides a pipeline to proceed from high fidelity location traces from a fraction of the population to building level occupancy profiles using fairly ubiquitous data sources.

Thesis Supervisor: Marta Gonzalez
Title: Associate Professor of Civil Engineering

Acknowledgements

I would like to thank Prof. Marta González for giving me the chance to work at the Human Mobility and Networks Lab. Not only did she guide me on performing research for the group, but also prepared me for life as a professional outside academia.

My team members at the HuMNet Lab have been instrumental to the learning over the course of my stay at MIT. Special thanks to Yingxiang and Serdar for your advice and support outside the lab. Shan, Ed, Riccardo, Andreá, Ricky, Phil, Steve, Luis, Yanyan, Fahad, Sharon, Adham, Marco and Adrian- you all have been a great team and your comments and our discussions have been invaluable to this work.

To the Department of CEE and co-habitants of 1-139 and 1-151, thank you for providing a great working environment. The classes at MIT at CEE, Media Lab, DUSP and the Sloan School, the talks across MIT and conferences and ecosystem in the Greater Boston Area have been instrumental in my growth during my stay at MIT. Having an empathetic advisor in Prof. Saurabh Amin played an important role in helping me adjust to life at MIT.

Special thanks to my alma mater, IIT-Madras, Prof. Karthik K. Srinivasan and Prof. Mark Hickman for shaping me as an individual and preparing me for life at MIT and beyond.

Last, but not the least, I would like to thank my family for their unconditional support over the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The proliferation of Information Technology and the internet has led to the amassing of an array of datasets by a range of service providers. Internet companies analyze web interaction, banks collect and analyze spending patterns from credit cards, roads collect flow information using inductive loops, video cameras and other vehicle counters, WiFi hotspots record MAC addresses and cellular service providers collect Call Detail Records (CDRs) to list only a few. Each of these applications clearly warrants the need for collecting data and the data provide sufficient detail to serve the purpose for which they were collected. For example, observations made in call detail records might include information about the location of the call for efficient routing through cell phone towers, the nature and duration/magnitude of transmission (call/text/data) and the user ID for billing purposes. As with CDRs and credit card transactions, the data usually serve a very particular functional purpose. Many times, however, the data can be used to study a variety of other phenomena as well. For example, the National Household Travel survey is used to analyze the travel patterns of sample of households that are likely to be representative of the population but is often also used in a variety of other studies.

Each of these data sources provide their own perspective of human behavior and are accompanied by inherent biases. These biases and specific design objectives have to be carefully considered when using a data source for a purpose disparate from what it was originally designated to perform. Since a variety of the data streams record individual human transactions, they serve as proxy for individual mobility patterns albeit with periods without observation.

It is therefore possible to glean insights into mobility behaviors of people from a variety of datasets emanating from the urban landscape. The portability and scalability of the analysis framework however would be significantly dependent on the ubiquity of the data sources and the extent to which the framework serves to represent the urban population.

This work builds on TimeGeo (13), an attempt to simulate high-resolution population wide mobility patterns using ubiquitous datasets. The mobility model is developed in Boston and is primarily dependent on Call Detail Records. The model is made scalable by incorporating detailed characteristics of individual trip and used to simulate the entire urban population. The fidelity of the scaled results has been demonstrated by a comparison to current state of the art models. While the current models depend on tedious and expensive surveys, the proposed model circumvents the need for this process. Further, the portability of the model is demonstrated by training the model on smartphone users in a controlled tracking experiment in Denmark and comparing the results to the observed behavior.

The utilization of results at a micro scale is limited by two factors. Firstly, the resolution of the training data. In case of Call Detail Records in Boston, the resolution was insufficient to resolve user locations to individual buildings. The location traces in Denmark were accurate enough to discuss mobility at the level of buildings. Secondly, the availability of seed data for the entire population. The simulation process depends on the availability of some information about each user that is being simulated. Ideally, this would consist of detailed movement histories captured in the training data. However, there is no single dataset that provides the mobility history for all individuals in an area. Hence assumptions are made on these based on available aggregate metrics. The resolution of these aggregate metrics also dictates the resolution to which high-fidelity results can be obtained. In case of Boston, this was the census-tract level.

In order for the results to be useful at a micro level, we need to reconcile the presence of people from the results of the simulation with the engagement opportunities available in the landscape of the city. Buildings and points of interest are the chief attractors of human activity in an urban form. With the advent of detailed digital maps, this information is widely available. All the user stays that are found within unit resolution area of the simulation can thereafter be distributed to individual hyperlocal attractions. This provides last-mile detail to individual stays and completes the process of simulating high-resolution mobility tariffs for the entire urban population.

The process of simulation explored in this study is shown to be able to replicate results from the present state-of-the-art. But, like any abstraction of reality, there are certain elements that remain unaccounted for and some biases that are implicit with the data used for modeling. Going ahead, as data become bigger and more accurate, we discuss the possibilities of improving the analysis framework and bringing the results closer to reality.

Population-wide details of movement of individuals have been of interest to several traffic-control, energy consumption and city planning applications (1,2). Transportation OD matrices have been widely studied and are usually constructed at a more aggregate level. The simulation is shown to provide accurate estimates of these matrices when compared to the existing surveys. One of the added benefits of the microscopic details in the simulation lies in the ability to estimate occupancy profiles of individual buildings. The presence of people can act as a proxy for the energy load demands for buildings. These profiles can in turn be used to identify building archetypes for appropriate interventions and planning. The study provides an analysis of building occupancy profiles and comments on the likely use-case scenarios of buildings in the Boston area. High-resolution mobility patterns hold immense potential to facilitate informed decision making for planning the urban landscape and to facilitate more meaningful and efficient interaction between city dweller. This study provides the data required to explore possible interventions and hopes to spark interest in exploring the results and improving the framework. The analysis just begins to scratch the surface of possible applications of high-resolution simulations.

1.2Literature Review

1.2.1 Works on Mobility Simulation

Several studies that focus on human mobility at high resolution (3,4) are available in transportation literature. These forecast the whereabouts of individuals at the scales of hundreds of meters and minutes for an entire city. Traditionally, the input for such models is 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 this information put severe limits on survey sample size and frequency. 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 the choices of the entire urban population (3, 5–7). Another attempt at mobility modeling has been in the ILUTE framework (8) at Toronto.

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 (9–11). The problem, however, is that the geotagged data available from communication technologies, in its massive and low cost form, cannot inform us about the detailed activity choices of their users, making most of the data useless for such purposes. In order 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 (12).

1.2.2 TimeGeo and Background Studies

Identifying this, the TimeGeo framework (13) was developed to extract individual features and key mechanisms needed to effectively generate complete urban mobility profiles from the sparse and incomplete information available in telecommunication activities. The TimeGeo framework serves as the base on top of which this study is built. The details of the TimeGeo model are provided in Section 2.1. Its scalability and portability are studied separately in Chapter 2. TimeGeo builds on existing work on Call Detail Records (CDRs) (26,27).

Mobile phones are the prevalent communication tools of the twenty-first century, with the worldwide coverage up to 96% of the population (14). The call detailed records (CDRs), managed by mobile phone service providers for billing purposes, contain information in the form of geo-located traces of users across the globe. Mobile phone data have been useful so far to improve our knowledge on human mobility at unprecedented scale, informing us about the frequency and the number of visited locations over long term observations (15–20), daily mobility networks of individuals (17, 21), and the distribution of trip distances (15, 17, 19, 22–24). Due to the sparse nature of mobile phone usage, these data sources have sampling biases and do not provide complete journeys in space and time for each individual (11). 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 timing of their visits (25). By labeling the visited locations of 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 (26, 27). They are aggregated estimates of person-trips 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 (26, 27).

A fundamental question still remains on how to perform a spatiotemporal mapping of raw mobile phone data to establish survey-less models of travel demand with high spatiotemporal resolution, through which individuals' disaggregated daily journeys can be generated. This question is answered by TimeGeo (13) which generates a coherent framework to generate these spatiotemporal patterns of individual daily from passive data.

1.2.3 Occupancy Studies

Globally, more people live in urban areas than in rural areas, with 54 per cent of the world's population residing in urban areas in 2014. In 1950, 30 per cent of the world's population was urban, and by 2050, 66 per cent of the world's population is projected to be urban. Today, the most urbanized regions include Northern America (82 per cent living in urban areas in 2014), Latin America and the Caribbean (80 per cent), and Europe (73 per cent) (28). Buildings are the basic units of urban development. Office buildings alone represent approximately 17% of the energy used in the US commercial building sector (29).

Building occupancy is a paramount factor in building energy simulations (30). Most of the studies in this area adopt a bottom-up approach, calculating occupancies for individual buildings using elaborate data collection techniques such as occupancy sensors and are performed on some small set/ few categories of buildings (31-33). Some stochastic models of the occupancy level of single offices have been proposed in the last decade within the scientific community (34-36). However, these do not scale well to large buildings where the number of agents, rooms and interactions lead to non-trivial choices.

This study focus on a fundamentally different top-down approach wherein the results from a high fidelity urban-scale simulation, TimeGeo (13), which provides an accurate estimate of the presence of people in an aggregate region over 10-minute time intervals, are leveraged to approximate the occupancy profiles for all the buildings in city. The proposed method uses hyper-local information and tries to incorporates user perception and urban dynamics. However, ground truth data are rare since occupancy data are difficult to acquire as most buildings do not have the infrastructure required to accurately sense people throughout the building.

1.2.4 Energy Studies

According to the U.S. Department of Energy, energy for heating and cooling accounts for approximately 35 - 45% (37) of the total expenditure within a building. Forecasting the occupancy of buildings can lead to significant improvement of heating and cooling systems. For example, an HVAC system can pre-heat or pre-cool a room just prior to its use, instead of always keeping the room at a set temperature. Or, an HVAC system could take advantage of times when electricity cost is lower, to chill a cold-water storage tank, in anticipation of needed cooling (38). While such systems can be useful in smart home applications, another form of intervention can be in the form of retrofitting buildings with tailor made solutions based on the occupancy archetype that the building follows. For this, there is a need to identify the archetypes based on the predicted occupancy profiles. A technique to do so has been explored in this study. A future avenue of the work of this thesis is to look at the possibility of reconciling the results with those from building archetypes using material and construction characteristics (39).

1.3Thesis Outline

The remainder of the study is divided into four chapters. Chapter 2 discusses how models are trained on location traces. It demonstrates the scalability of the models and the portability of the results. It also discusses the limitations of the model and motivates the need for hyperlocal analysis. Chapter 3 goes into the details of obtaining hyperlocal data from a variety of sources and putting them together to serve as objects that attract trips. Chapter 4 discusses a novel method of distributing trips within the highest unit of resolution of the models- the census tracts. The method attempts to incorporate individual perception and urban dynamics into the distribution process. Chapter 5 presents an analysis of the results at the level of individual buildings and census tracts. It is able to identify archetypes that can be used for planning and energy interventions, particularly when more data become available for individual buildings.

Chapter 6 provides a reflection on the contributions of the study, its limitations and the possible directions of future work.

Chapter 2

The Mobility Model: Theory, Scalability and Portability

2.1Details of the mobility models

The TimeGeo framework starts off with the CDRs of 1.92 million in the Greater Boston Area. For maintaining privacy, the user IDs have been encrypted and each record has the following format: [anonymized user ID, longitude, latitude, time stamp (in seconds)]. The location coordinates are estimated by the data provider using standard triangulation algorithms and have higher resolution (accuracy of 200 to 300 meters) compared to those in the traditional tower-based CDRs (40-42). This finer granularity enables the application of data mining methods previously tested for GPS records (43-45).

The processing of CDRs is done by individual. The process involves the identification of stay locations for the individual followed by the determination of the location type (home, work or other) and validation using the census population data.

The processing of the CDRs is followed by separate models for commuters, estimation of mobility parameters, simulation for the entire population and validation of the simulated results.

2.1.1 Processing of CDRs

Stay-points are identified from a sequence of consecutive mobile phone records based on spatial and temporal thresholds. The spatial threshold is a roaming distance when a user is staying at a location, related to the accuracy of the underlying location positioning technology. In this study, we set the roaming distance of a stay-point as 300 meters. The temporal threshold is the minimum stay time (e.g., 10 minutes) at a location, measured as the duration between the first and the last record observed at a stay-point. Once a stay-point is identified, its location is set as the centroid of all records belonging to that stay-point.

The second step is to cluster stay-points into stay-regions, since stay-points identified from a user's different trajectories over time may refer to the same location although the triangulated coordinates may not be exactly the same. We use a grid-based clustering method to cluster stay-points into stay- regions. The advantage of the grid-based clustering method is that it sets the output cluster sizes- which is desirable when we know that each location has a bounded size and the accuracy of the records is within a threshold. In this study, the maximum stay-region size is set as d = 300 meters. The procedure to perform grid-based clustering is as follows: First, divide the entire region into rectangular cells of size d/3. Next, map all the stay-points to each cell. Then, iteratively merge the unlabeled cell with the maximum stay-points and its unlabeled neighbors to a new stay-region. Once a cell is assigned to a stay-region, it's marked as labeled. (46) discuss details of this method.

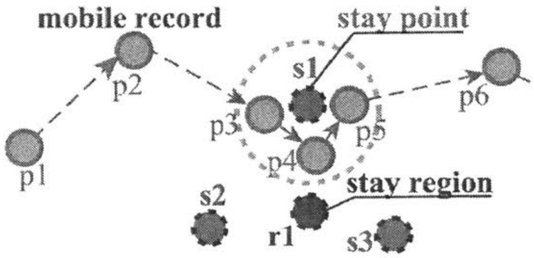


Fig. 2.1. Extraction of stays. Gray dots are raw mobile phone observations. Green dots are extracted stay-points, and blue point represents the stay-region

The next step is the identification of the location type for each stay region- home, work or other. We label the most frequently visited stay-region during weekday nights (between 7pm of first day and 8am of second day) and weekend as the home stay-region. 1.44 million users (75% of the 1.92 million) are identified with home.

For a non-home stay, if its start time is during weekday daytime (between 8am and 7pm), it is defined as a potential work stay. The assumption to label a potential work stay into a work stay is based on the rationale and historical evidence (47, 48) that for a given visitation frequency, trips with longer distance are more likely to be work trips than those with shorter distance, which are more likely to be for non-work purposes (e.g., grocery shopping near home). For a user who has an identified home location, we label her potential work stay-region i as a work place if its distance from home (d'_i) times its visitation frequency (n'_i) is the maximum among those of all potential work stay-regions. We also restrict that a user's visitation to her work stay-region should not be less than 3 times in the observation period (i.e., $n'_i >= 3$), and its distance from home should not be too short (e.g., d'_i>500 meter). When a potential work stay-region fulfills the above criteria, it is labeled as work. Otherwise, it will be labeled as other. For validation purpose, we focus on users whose home is within the Metro Boston area defined by the Boston MPO (49). Among the 1.44 million users with home, 0.78 million have home within Metro Boston, and 0.66 million have home outside Metro Boston. Among the 0.78 million users whose home are identified within Metro Boston, 0.42 million have been identified with work stay-regions. If users have few records, it will be difficult to estimate their mobility parameters. We filter users who have more than 50 identified stays and at least 10 home stays in the observation period as active users, and derive a set of 0.177 million such users in the study area. The process of estimating individual trajectories from sparse location data is now complete.

The purpose of using a data source from a ubiquitous technology that enjoys high adoption rates across the world was to be able to gain insights scalable to the entire population. A fundamental check that the results from the data mining process need to pass is their ability to represent the population from which they are derived.

By using the 2010 census population data and the 2006-2010 Census Transportation Planning Products (CTPP) data, we validate the identified home and work locations of the 0.78 million users within Metro Boston at the city and town level. To expand these 0.78 million users to population of the Metro Boston, the number of home stay- regions are aggregated to Census tracts of the Metro Boston. An expansion factor is calculated for each tract as the ratio of the 2010 Census population and the number of residents identified in the CDR data set. For Census tracts with fewer than 10 CDR residents (around 10 in the study area), the expansion factor is set to 0 to ensure that we do not overweight users that are not representative for a given Census tract.

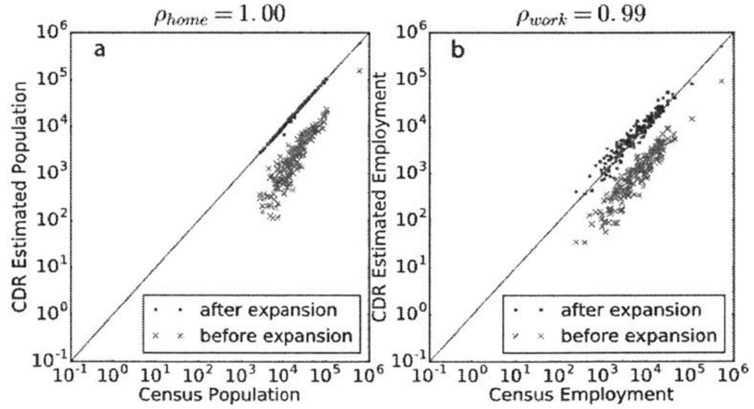


Fig. 2.2: Validation of the home and work labeling for the 0.78 million CDR users with detected home in the Metro Boston at the city and town level. (a) 2010 Census population vs. CDR estimated residents at town-level before and after population expansion. (b) 2006-2010 CTPP (50) workers vs. CDR estimated workers at town-level before and after population expansion.

Fig. 2.2 shows respectively the comparison of (a) residential and (b) employment population at town-level between 2010 Census data and the CDR estimates, and between the 2006-2010 Census Transportation Planning Products (CTPP) (50) data and the CDR estimates, for both before and after expanding the CDR data. The town-level correlation between the CTPP employment data and the estimated CDR employment is 0.99, and the sample expansion method adjusts well for the difference in magnitude. The total expanded CDR users with workplace is 2.3 million, while the CTPP reports a total of 2.1 million. This strong correlation is noteworthy, considering that each user's home and work locations were expanded based on their home location only. It also serves to validate the viability of using the chosen dataset to estimate urban scale mobility.

2.1.2 Modeling Commuters

It is common knowledge that the majority commuters usually go to work in the morning and finish work in the late afternoon with some breaks in between. However, the CDR data do not seem to capture the temporal features of the work activities well- namely, the work start time, duration and the presence of work-breaks. Fig. 2.3 compares the distribution of the start time and duration for work activity for CDR active users and the 2009 National Household Travel Survey (NHTS). It is seen that the NHTS has two peaks (around 8am and 12pm, respectively) in the work start time and two peaks (around 4 hours and 8 hours, respectively) in the work duration. The peaks in start time and duration are caused by breaks during working hours. In fact, both the 2009 NHTS and the 2010 American Time User Survey (ATUS) (51) show that around 20% workers have 1 work-break outside their workplace (which generates trips from workplace during the break.)

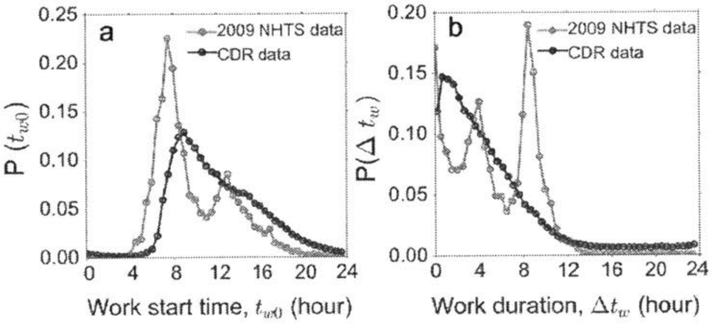


Fig 2.3: Distribution of observed work start time and duration Marginal distribution of (a) work start time and (b) work duration. Note: The CDR data does not compare well with the survey data, not capturing well the beginning and end of the work trips.

To overcome the shortcoming of the CDR data, we sample and fix the work time for commuters prior to simulation as detailed in the following subsection. If large scale mobile phone data with high frequency are available in the future, work time could be observed from mobile phone data. To model the fixed work activity for a commuter, we sample from a joint distribution of work start time and duration, and introduce stochastically a work-break to allow for flexible activities during the break period. To be more specific, the detailed steps are as follows:

- We generate a pair of work start time (t_{w0}) and work duration (Δt_w) from a 2-D Gaussian mixture distribution for each commuter. Note that this t_{w0} is simply the beginning of work in a day (on a weekday), and Δt_w simply measures work duration from beginning to end in the day.
- We introduce a parameter (θ) to stochastically determine if a commuter will take a work-break. We randomly generate a number $x \sim U(0,1)$. If $x<\theta$, the commuter will have a work-break outside workplace. We fix a proportion of commuters to have work-break outside workplace, $\theta = 0.20$.
- If the commuter takes a work-break, we then generate a work-break duration (Δt_b) and a work-break start time tb0 from two distributions, respectively. With the generated t_{b0} and Δt_b, we then update the work time-slots which are now split by the work-break.

On a weekday, for a commuter with predetermined t_{w0} and Δt_w, and t_{b0} and Δt_b (if she takes a work-break), the TimeGeo model is applied to fill the rest time slots that are not occupied by work activities. Fig.2.4 demonstrates how the model works for commuters.

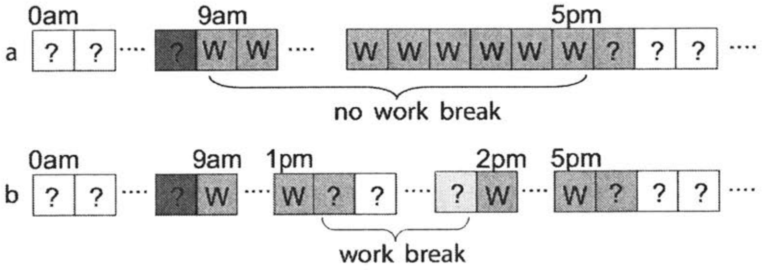

0am
9am
5pm
a
? ? ···· ? W W ···· W W W W W W ? ? ? ····
no work break
0am
9am
1pm
2pm
5pm
b
? ? ···· ? W ···· W ? ? ···· ? W ···· W ? ? ? ····
work break

Fig.2.4: Modeling commuters. For commuters with fixed work activity and (a) with no work-break, (b) with work-break. Work start time and duration, and break start time and duration are predetermined from distributions discussed later.

As illustrated in Fig. 2.4 (a), on a sample weekday, a user is predetermined to stay at workplace from 9 am to 5 pm with no work-break. The proposed TimeGeo model will fill the rest time slots. The user will make a trip to work at 9 am, independent of her activity before 9 am (the blue slot). She will move from work after 5pm (the green slot), but whether to go home or to other location is simulated by the TimeGeo model. Similarly, in Fig. 2.4 (b), a user is first predetermined to work from 9 am to 5pm and take a break outside workplace between 1 pm and 2 pm. The model works the same as in (a) for the time slots before 9 am and after 5pm. For the work-break, the user will definitely make a trip from work after 1pm (the cyan slot) and then move back to work at 2 pm. She decides where to go in the break based on the TimeGeo model. The user could visit multiple other locations during the break, simulated by the TimeGeo model.

To characterize the statistic of work start time (t_{w0}) and work duration (Δt_w) from alternative data sources, we use NHTS and ATUS to estimate the joint distributions of t_{w0} and Δt_w. We fit the data with a mixture of multivariate normal distributions, and use a Gaussian Mixture Model (GMM) (52) to estimate parameters of the following joint distribution $f(x|\mu_1, \Sigma_1, \dots, \mu_K, \Sigma_K) = \sum_{k=1}^{K} \pi_k \left(\frac{1}{2\pi\sqrt{|\Sigma_k|}}\right) \exp\left(-\frac{1}{2}(x-\mu_k)^{\mathrm{T}} \Sigma_k^{-1} (x-\mu_k)\right)$ where K(> 0) is the number of modes (distinct local maxima), x = $(A,B)^{\mathrm{T}}$, $\mu_k = (\mu_{kB}, \mu_{kB})^{\mathrm{T}}$, $\Sigma_k = \begin{pmatrix} \sigma_{kAA}^2 & cov_{xAB} \\ cov_{kAB} & \sigma_{BB}^2 \end{pmatrix}$, A stands for t_{w0}, B stands for Δt_w, π_k is the mixing coefficient, $\Sigma_{k=1}^{K} \pi_k = 1$.

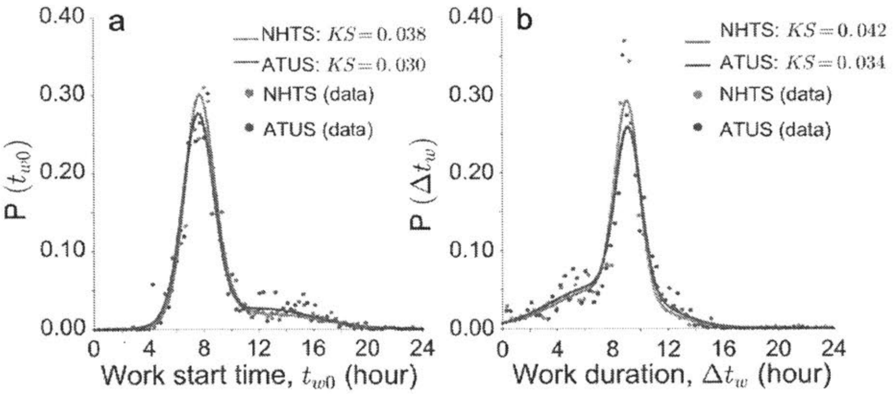


Fig.2.5: Marginal distribution of work start time (t_{w0}) and duration (Δt_w). Note: t_{w0} represents the beginning of work in a day. Δt_w simply measures work duration from beginning to end in a day. Parameters are estimated from a 2-dimensional GMM.

We find that three clusters (K = 3) best fit the empirical data. It is in good agreement with existing studies on the temporal behavior of workers' daily activity patterns found in (53) (e.g., early workers, regular workers, and late workers). Fig. 2.5 presents the marginal distributions of t_{w0} and Δt_w estimated from the 2010 NHTS and 2009 ATUS data, respectively. The results are quite similar, and we use the set of parameters from NHTS to jointly generate t_{w0} and Δt_w: K = 3, $\pi_1 = 0.17$, $\pi_2 = 0.29$, $\pi_3 = 0.53$, $\mu_{1A} = 12.8$, $\mu_{2A} = 7.9$, $\mu_{3A} = 7.6$, $\mu_{1B} = 6.6$, $\mu_{2B} = 7.5$, $\mu_{3B} = 9.0$, $\sigma_{1A} = 3.7$, $\sigma_{2A} = 1.5$, $\sigma_{3A} = 1.0$, $\sigma_{1B} = 4.4$, $\sigma_{2B} = 3.2$, $\sigma_{3B} = 0.9$, $cov_{1AB} = -4.3$, $cov_{2AB} = -2.6$, $cov_{3AB} = -0.3$ (Time units are in hours).

The final step in modeling commuters is to simulate the characteristics of their work breaks. To characterize work- breaks from data, we first estimate the distribution of their duration (Δtb) from the NHTS and AUTS, shown in Fig. 2.6(a). The probability density of Δt_b follows a log-normal distribution, i.e., $P(\Delta t_b) = \left(\frac{1}{sqrt(2\pi)\Delta t_b}\right) exp\left(-\frac{(\ln \Delta t_b - \mu)^2}{2\sigma^2}\right)$, where $\mu = 3.9$, $\sigma = 0.9$ (time unit in minute). Note, in the simulation, when generating Δt_b, we make sure that $\Delta t_b < \Delta t_w$.

From common knowledge, work-break start time (t_{b0}) peaks in the middle of work, although it may occur anytime during work. We measure the distribution of the normalized deviation of work-break midpoint (t_{bm}) from the work midpoint (t_{wm}), noted as $D_{cw} = \frac{t_{bm} - t_{wm}}{\Delta t_w - \Delta t_b}$, from the 2009 NHTS data, shown in Fig. 2.6(b). We find that D_{bw} follows the following truncated Cauchy distribution $P(D_{bw}) = \left(\frac{1}{\pi\gamma}\right) \cdot \left(\frac{\gamma^2}{(D_{bw} - x_0)^2 + \gamma^2}\right)$, where $x_0 = 0.0$, $\gamma = 0.1$, $-0.5 < D_{bw} < 0.5$ (since a work break has to start after work starts and end before work ends). For the simulation of work-breaks, we randomly draw a D_{bw} from the truncated Cauchy distribution, and then determine the work-break start time according to $t_{b0} = t_{bm} - 0.5\Delta t_b = t_{w0} + (\Delta t_w - \Delta t_b)(0.5 + D_{bw})$. With the generated work-break start time and duration, we then update the work time slots, splitting them by the work-break.

Although the above parameters are estimated using NHTS data for consistency check, we do not think that expensive travel surveys are necessary for the purpose of modeling typical work activity (i.e., work start time, work duration, work-break duration, work-break start time).

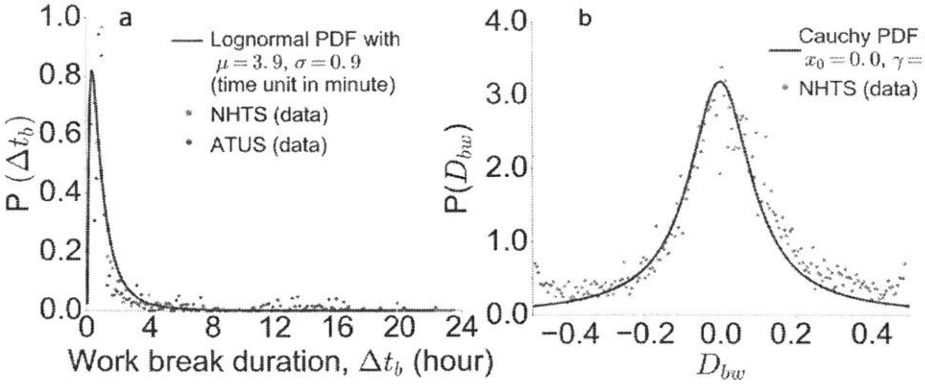


Fig. 2.6. Model for work-break. (a). Empirical distribution of work-break duration Δt_b. (b) Empirical distribution of the normalized deviation of work-break midpoint from the work midpoint, D_{bw}, ($-0.5 < D_{bw} < 0.5$).

2.1.3 Estimation of Mobility Parameters

The modeling framework for the mobility decisions made by an individual in TimeGeo (13) is shown in Fig.2.7 and detailed below.

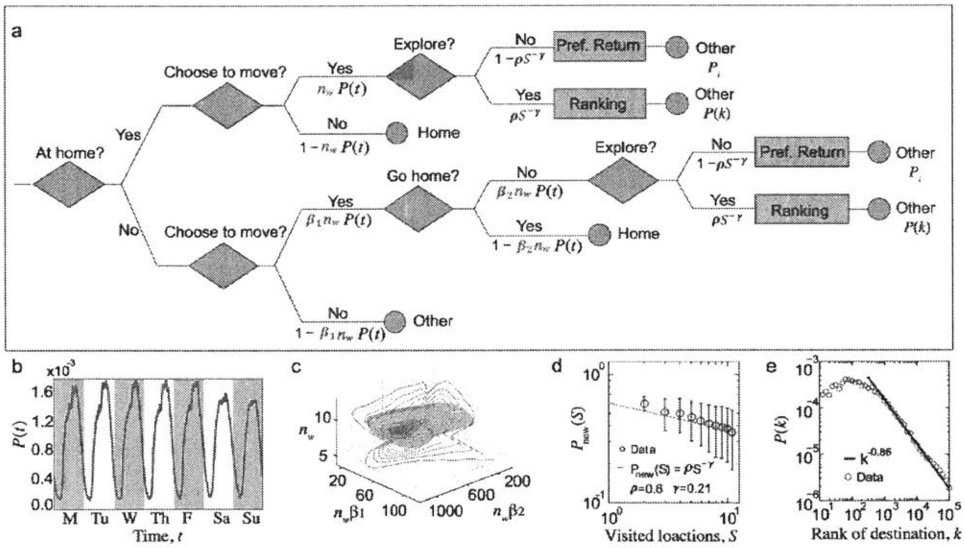


Fig.2.7: Modeling framework for the decision to move made by an individual in TimeGeo (13)

The travel decisions of an individual are determined using parameters for both the population and the individual. The population parameters include the travel circadian rhythms and exploration and preferential return (EPR) parameters (55). The travel circadian rhythms measure the population travel circadian rhythm as the probability of traveling to and from flexible activities in every 10-

minute time slot t, which is denoted as P(t). Since commuters' work activities are modeled as fixed choices, the probability P(t) for commuters does not include trips to and from the work activity. In other words, in a time slot t, only if a commuter travels both to and from either home or other (not work), the trip counts towards the probability P (t). The circadian rhythms are shown in Fig. 2.8.

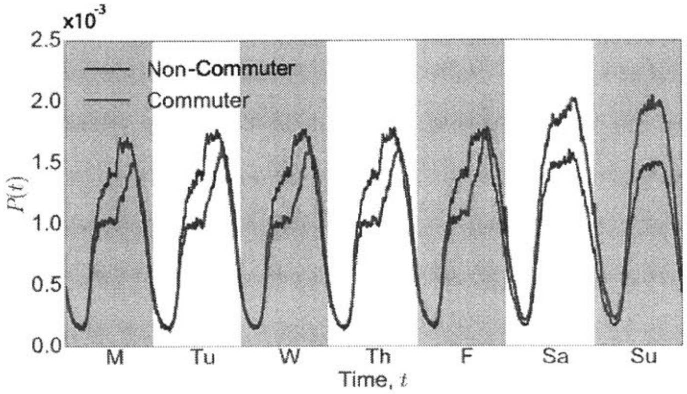


Fig.2.8: Circadian travel rhythms for commuters and non-commuters. Note: Because the fixed activity—work—is not determined by the Markov model, travels to and from work for commuters are excluded from the measure of P (t).

While estimating global parameters from the CDR data to simulate the EPR mechanisms differences among individuals are found. This leads to the introduction of individual parameters captured by the newly introduced weekly trips (n_w) and the two individual parameters of the Markov model (β_1, β_2). Fig. 2.9 (a) shows that for different S groups, the number of visited distinct locations S(t) versus time follows S(t) ~ t^μ. For S group: 5-10, $\mu = 0.54$; S group: 10-20, $\mu = 0.68$; S group: 20-30, $\mu = 0.76$; S group: 30-40, $\mu = 0.80$.

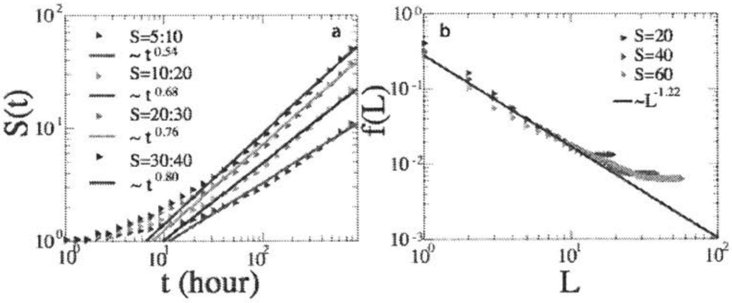


Fig.2.9: Empirical results on the exploration and preferential return (EPR) parameters estimated from the CDR data of active users. (a) the number of visited distinct locations S(t) versus time for different S groups, (b) the visitation frequency to the Lth most visited locations fL follows: fL ~ L−ξ, with ξ = 1.2 ± 0.1

Fig. 2.9 (b) shows that for users with different distinct locations (S), their visitation frequency to the Lth most visited locations f_L follows: $f_L \sim L{-}\xi$, with $\xi = 1.2 \pm 0.1$, similar to the finding in (41). We estimated global EPR parameters for model $P_{new} = \rho S{-}\gamma$, with $\rho = 0.6$ and $\gamma = 0.21$. Our finding is consistent with (41), which showed that $\xi = 1 + \gamma$, if $\gamma > 0$.

Fig. 2.10 illustrates the Markov model of transition probabilities at Home (H) or Other (O) in a 10-minute time-slot t for a non-commuter. We show in this example two other states—Other 1 (O_1) and Other 2 (O_2)—to distinguish the current other state, and a consecutive new other state. In time-slot t, when an individual is at home, her probability of staying home is $P_1 = 1 - n_wP(t)$. Her probability of traveling to either O_1 or O_2 is $0.5(1\text{-}P_1)$. When she is not at home, but at another location (either O_1 or O_2), the individual is in an active state—her probability of staying at the current other location is $P_2 = 1 - \beta_1 n_w P(t)$, and her probability of visiting a consecutive other location is $P_3 = \beta_1 n_w P(t)\beta_2 n_w P(t)$. When she moves from another state, she can either choose to go to an additional other location with probability P_3, or go home with a probability $P(O_{1(2)} \rightarrow H) = 1 - P_2 - P_3 = \beta_1 n_w P(t)(1 - \beta_2 n_w P(t))$. To ensure that a person will go home at the end of the day, we add a condition that after certain hour in the late afternoon (e.g., 5 pm) the individual's returning home probability is the maximum value of $P\ (O_{1(2)} \rightarrow H)$ and $(1 - P\ (t))$. $\max(P\ (t))$

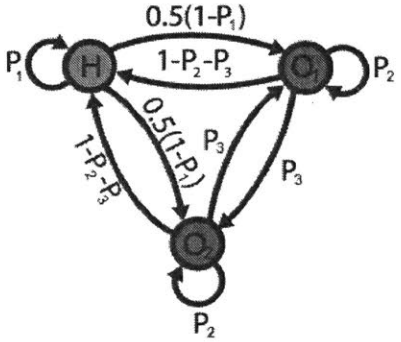


Fig.2.10: Illustration of the Markov model of transition probabilities between Home and Other state in a 10-minute time-slot t for a non-commuter.

We now show the empirical individual parameters measured for active commuters (133,448 individuals) and non-commuters (43,606 individuals). The joint distribution of parameters n_w, $n_w\beta_1$ and $n_w\beta_2$ is in Fig. 2.11. The median values of n_w , $n_w\beta_1$ and $n_w\ \beta_2$ for non-commuters are 7.4, 34.2, and 355.6, while the values for commuters are 5.7, 21.2, and 286.7. The two dimensional marginal distributions are shown by the contour plots. The marginal distribution of each parameter is in Fig. 2.12. The distribution of n_w and $n_w\beta_1$ could be approximated by log-normal distributions while the distribution of $n_w\beta_2$ is approximated by Weibull distribution. The corresponding estimation results are also shown in Fig. 2.12.

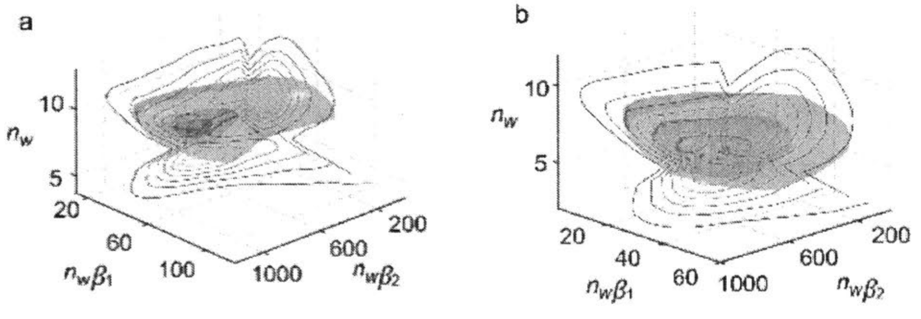


Fig.2.11: Joint distribution of parameters n_w, $n_w\beta_1$ and $n_w\beta_2$. Parameter distribution for (a) non-commuters, and (b) commuters.

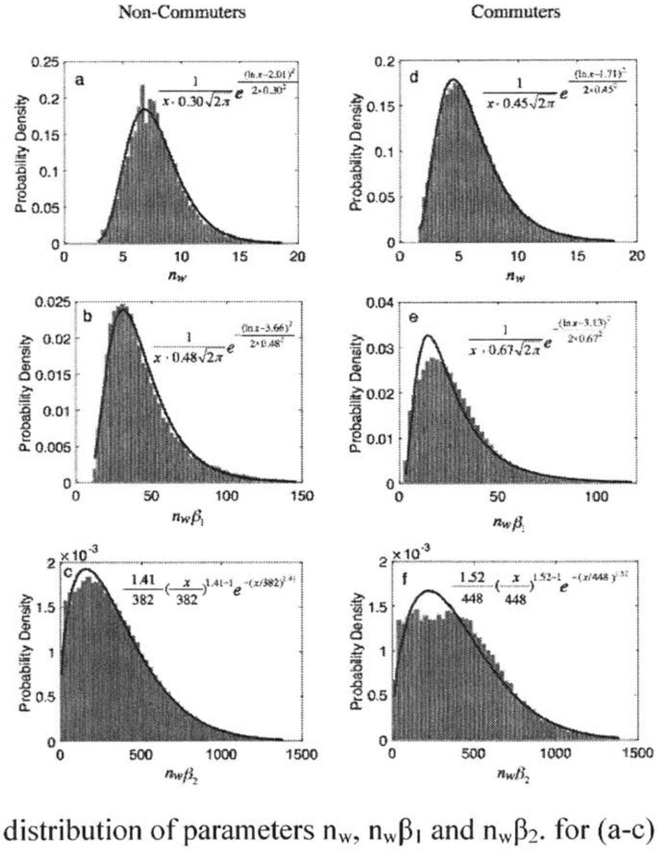


Fig.2.12: Marginal distribution of parameters n_w, $n_w\beta_1$ and $n_w\beta_2$. for (a-c) non-commuters, and (d-f) commuters.

2.1.4 Simulation and Validation

Based on the models and assumption outlined above, the simulation is able to capture the mobility patterns both at the individual and the metropolitan level. At the level of the individual, in addition to the number of daily visited locations, the model is able to capture similar daily mobility motif distribution revealed from the CDR data. Fig. 2.13 (a) is the aggregated motif distributions for all active users. The empirical data are in blue and the simulation results are in green. As a guide to the eye, two dashed lines at 1% and 5% are shown in the figure. The most popular trip motif is traveling between two locations. Fig. 2.13 (b) shows the motif distribution for commuters and non-commuters separately. To show the less popular motifs clearer, we plot the distribution in log scale in the inset of each figure. In general, the more complex a motif is, the lower the percentage is.

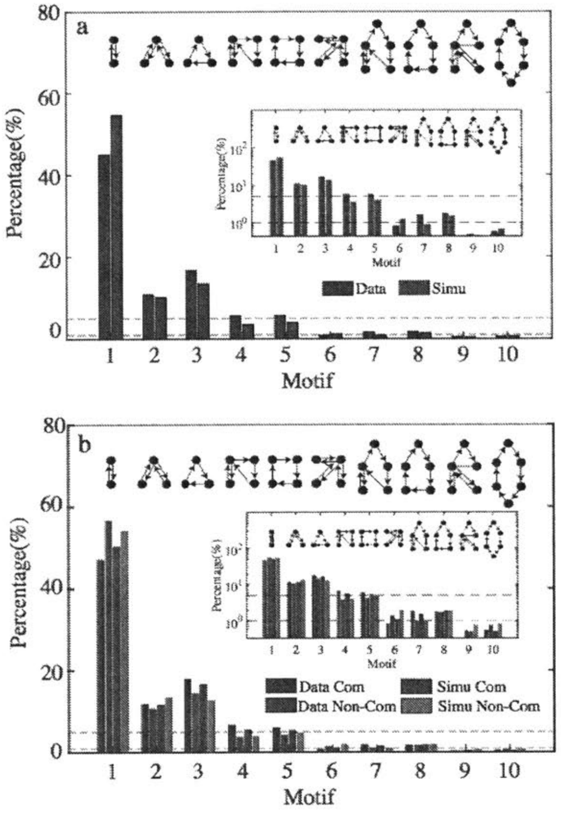


Fig.2.13: Motif distributions. (a) Motif distribution for all active users, shown in linear and log scale (inset figure). The two dashed horizontal lines are respectively 1% and 5%. (b) Motif distribution for commuters and non-commuters (inset figure is in log scale).

At the metropolitan level, the we simulate the mobility trajectories in Metro Boston for population aged 16 years and over (to be consistent with the census data, e.g., American Community Survey) and compare the results to the 2009 NHTS (54), and 2010-2011 Massachusetts Travel Survey (MTS) (55) for validation.

We expand the active phone users to the population aged 16 and over (i.e., 3.54 million people) in Metro Boston. We derive two sets of expansion factors, to expand active commuters to 2.10 million workers and expand active non-commuters to the rest 1.44 million non-workers, at the census tract level respectively (data avail- able from American Community Survey). Fig.2.14 shows the distribution of expansion factors for (a) commuters, and (b) non-commuters.

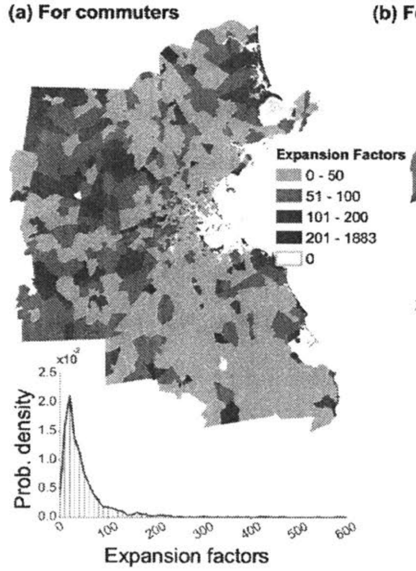


Fig.2.14: Expansion factor distributions. Top figures show the spatial distribution of expansion factors to expand the active CDR users whose home are in the census tracts: (a) to expand commuters to the total employment population, and (b) to expand non-commuters to non-employment population (above 16 years old) in the census tract. Bottom figures show the probability density distributions of the expansion factors for (a) commuters and (b) non-commuters.

We keep home and work locations and stay location records of active users to simulate the expanded 3.54 million individuals' daily trajectories based on the model discussed in the paper.

Fig.S16 shows the population distribution on (a) stay duration (Δt), (b) daily visited location (N), (c) trip distance (Δr) for simulation as well as survey data. Note that we do not include the travel distance distribution from NHTS for comparison, given that spatial aspects of travel are affected more directly (than the temporal aspects) by urban form, which varies across the nation [57].

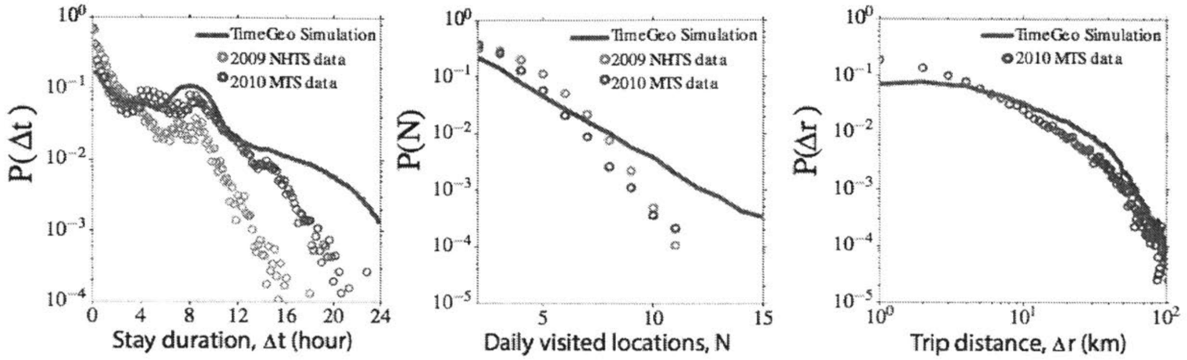


Fig.2.15: Population distribution comparison: simulation and travel survey data—including 2009 National Household Travel Survey (NHTS), and 2010-2011 Massachusetts Travel Survey (MTS). (a) Stay duration distribution. (b) Daily visited location distribution. (c) Trip distance distribution.

Fig.2.16 compares the simulated person-trips per day by trip purpose and by time period with the Boston MPO travel demand model for years 2010 (49) and 2007 (57), both of which follow the traditional four-step travel-modeling of trip generation, trip distribution, mode choice, and trip assignment. The trip purposes include (1) home-based work (HBW), (2) home-based other (HBO), and (3) non-home-based (NHB) trips. The time periods include AM peak (6am-9am), midday (9am-3pm), PM peak (3pm-6pm), and nighttime (6pm-6am). We also include the 2010-2011 MTS data for comparison. Note that Boston MPO 2010 model is for all aged population including 4.46 million persons (in year 2010) but excluding school trips, and the MPO 2007 model is for all aged population in year 2007 including all trip types (school trips are included in the HBO category). In general, the simulated HBW trips, and trips in three of the four time-periods (AM, PM, RD) are in good agreement with the MPO 2010 model. The simulated HBO and NHB trips are a little less than those estimated by the MPO 2010 model.

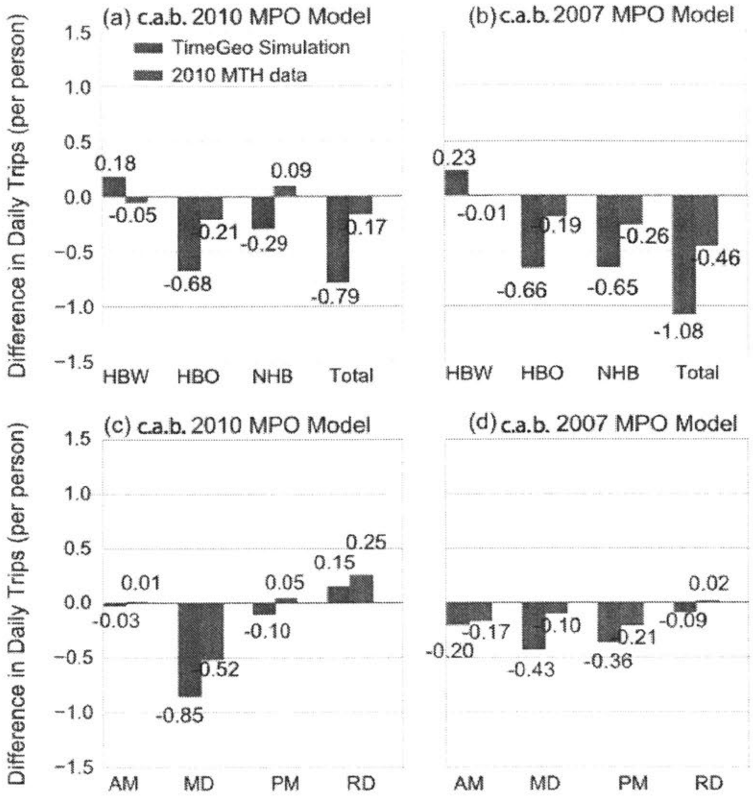


Fig.2.16: Comparison against baseline (c.a.b.) —2010 and 2007 Boston MPO travel demand models. (Note: Values closer to zero mean simulation results are with small differences from the base line) Note: Simulation is only for population older than 16 years old (3.54 million persons in 2010). Boston MPO 2010 model is for total population (4.46 million persons in 2010) excluding school trips (categorized as the HBO trips). Boston MPO 2007 model is for total population (in 2007) including all trip types. The 2010 MHT data presented here only include trips made by individuals aged 16 and over.

Fig. 2.17 shows the departure time of travel by trip purpose. The comparisons are among the TimeGeo simulation, 2009 NHTS, and 2010-2011 MTS (extracted for Metro Boston). The patterns for HBW, and all trip purposes are similar among the three sources. The simulation doesn't have a morning peak for HBO trips (potential reason might be that we are not simulating school trips, while travel surveys include those trips.)

Fig. 2.18 shows the comparison of origin-destination (OD) trips by trip purpose and time periods between the TimeGeo simulation and Boston MPO 2010 model at the city and town level (with inter-town and intra-town trips separated).

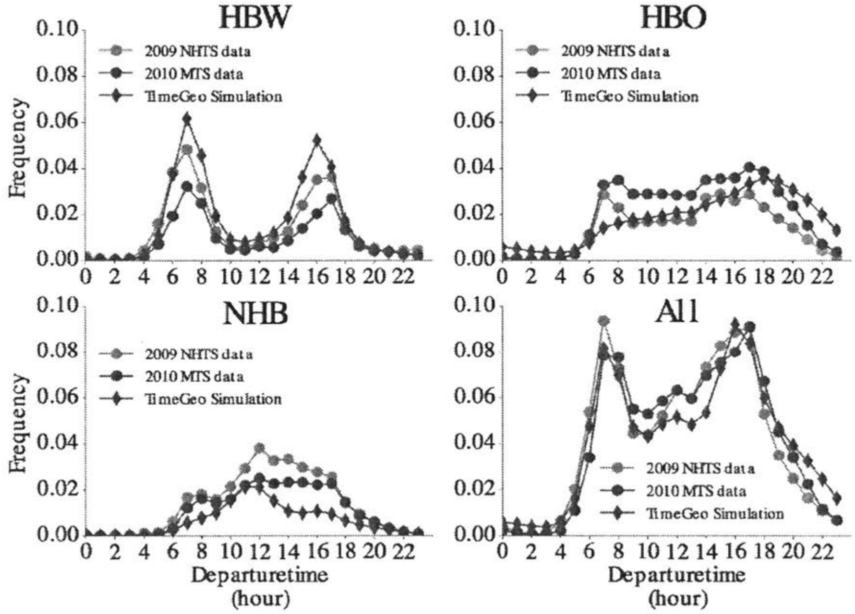


Fig.2.17: Comparison of travel departure time among TimeGeo simulation, 2009 NHTS data, and 2010 MTS data, for trips by purpose of home-based work (HBW), home-based other (HBO), non-home-based (NHB), and all types (All).

Correlation of the intra-town estimation between the two models are very good (the Pearson correlation coefficients are between 0.99 and 1.0), and those for the inter-town estimation are relatively lower. Since the employed population for the two sources are the same, the correlation between simulation and the MPO model is very good for HBW and AM peak period. Due to the difference in population for non-workers (the TimeGeo simulation only includes population aged 16 years and above), simulation trips for HBO and NHB trips are systematically lower than the MPO model estimates. However, the correlation between the two models are still very high (above 0.8 even for inter-town trips).

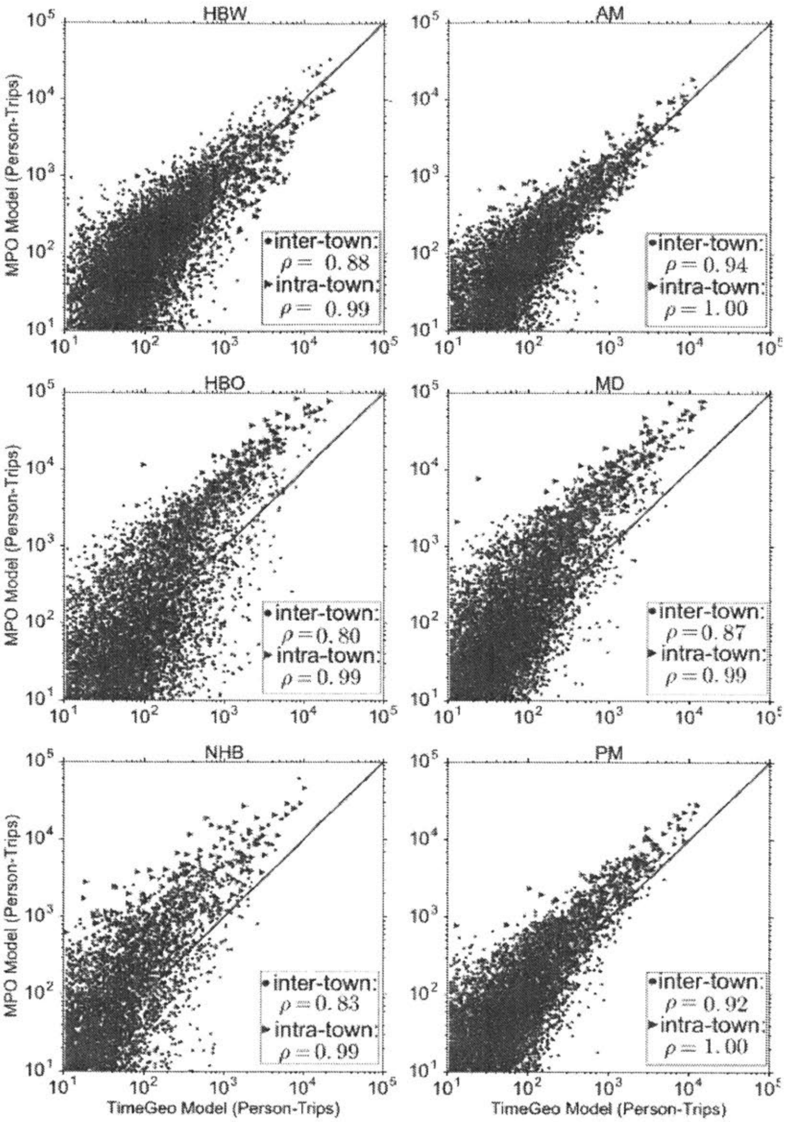


Fig.2.18: Comparison of Origin-Destination(OD) trip tables at city and town level: TimeGeo simulation (x-axis) and Boston MPO 2010 model (y-axis). Note: TimeGeo simulation is only for population aged 16 and over (i.e., 3.54 million persons), while the Boston MPO 2010 model is for population of all age groups (i.e., 4.46 million persons) within Metro Boston.

Hence, the results from the simulation are in conformity with existing studies both at the individual and metropolitan level. As a final step, we need to demonstrate the portability of the system to a new location. This is done in a study performed in conjugation with the Denmark Technical University (DTU) and is illustrated in the following section.

2.2Portability of the model, a study in Denmark

The CDR data in Boston provided intermittent location information for a large number of users. TimeGeo was able to simulate high-resolution individual trajectories for the entire population with these intermittent traces from frequent users. Since the model made no assumptions on the geography of the study, it should be portable to other regions with similar datasets. An aspect of TimeGeo that still needs to be demonstrated is its portability to different data conditions. In this section, we will explore how the model can be adapted to work with regular high resolution location datasets for a small set of users. Since the dataset is small and not representative of the metropolitan population, results are only verified at the level of the individual.

2.2.1 Introduction to the Dataset

The data used in this study are derived from the Copenhagen Network Study (58). The project has collected mobile sensing data from smartphones for more than 800 students at the Technical University of Denmark (DTU). The data sources include GPS location, Bluetooth, SMS, phone contacts, WiFi, and Facebook friendships.

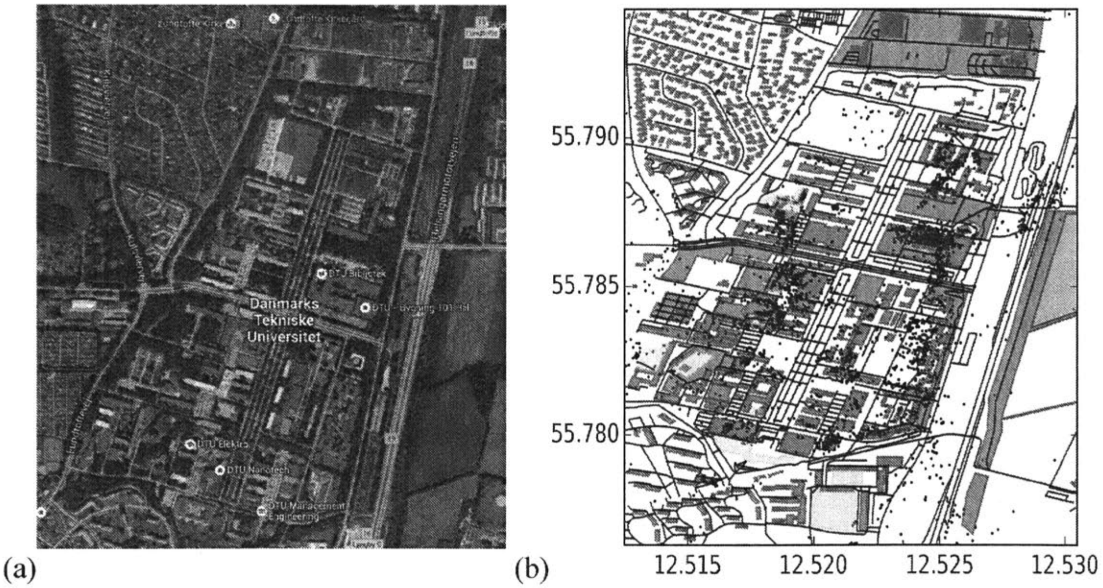


Fig.2.19: Panel (a) shows the satellite image of the DTU campus, panel (b) shows the boundaries on Open Street Maps of the features in the area along with the stay locations

For this study, we focus on the location data, which is collected by the smartphone with frequency of one sample every 15 minutes. Each location sample contains a timestamp, a latitude and longitude, and an accuracy value. The location is determined by the best available provider, either GPS or WiFi, with a median accuracy of $\approx$ 20 meters; more than 90% of the samples are reported to have an accuracy better than 40 meters. For individual participants, there may be periods missing data. These periods can occur for various reasons, for example due to a drained battery, the phone being switched off, the location probe being disabled, or due to software issues. Since we are interested in predicting mobility patterns, using data immediately preceding the window of prediction, we select the longest period which has at least one sample in 90% of the 15-minutes time bins for each participant.

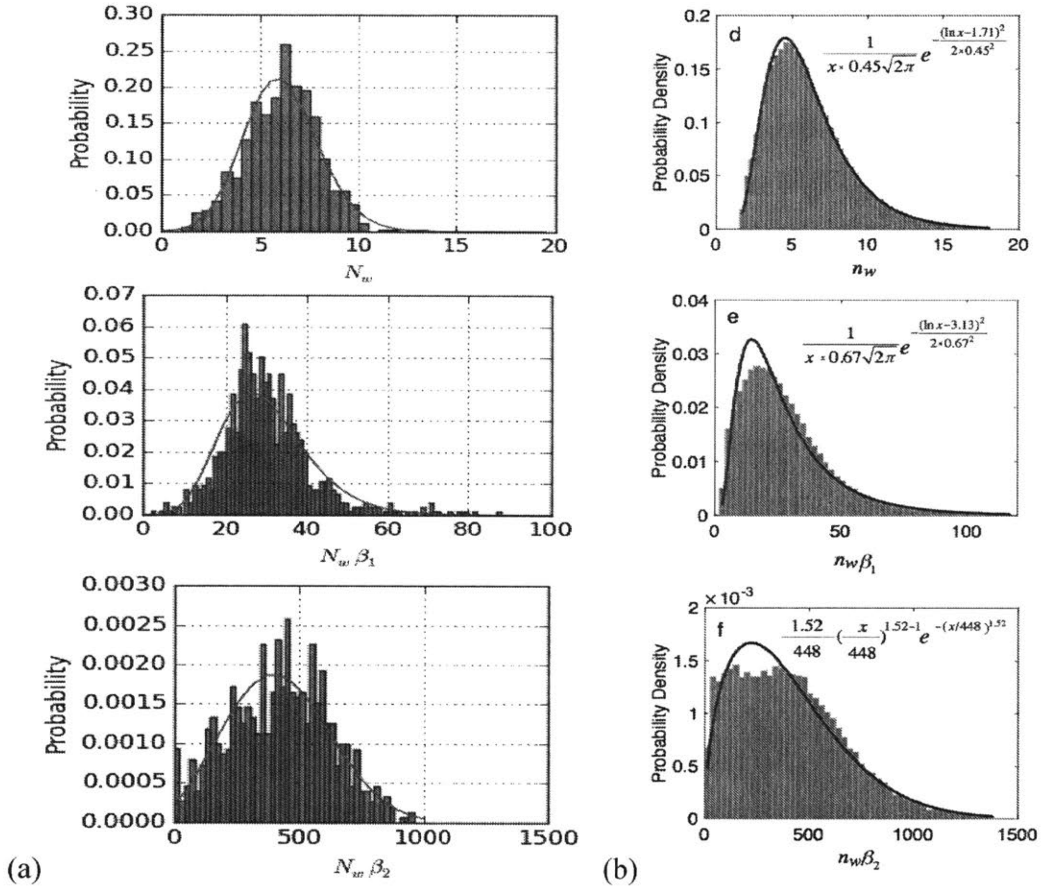
The data are mainly concentrated in Denmark where the study takes place, but because students use the phones during travel, the dataset spans several other countries as well. For this study, since we are interested in predictions within a metropolitan area, we discard data points outside the Copenhagen area. Fig.2.19 (a) shows the satellite image and 2.19(b) shows the feature boundaries and stay locations at the DTU campus, which serves as the center of the study.

2.2.2 Data Processing

Unlike the Call Detail Records, the location traces provided in the Copenhagen Network Study are pre-processed and accurate enough to constitute stay points by themselves. The starting point of this analysis is therefore the identification of stay regions. Comparing this to Fig.2.1, the raw data are able to provide us with individual stay points (green circles) and the output after running the algorithm are the stay regions shown by the blue circles. After the filtering process described in 2.2.1, we end up with 774 students.

In the next step, we estimate the mobility parameters for these users and compare them with those from the CDR data in Boston. Using a population of university students as the study group introduces systematic biases into the analysis. One of these is that their lifestyle dictates that nearly all of them should be commuters. The results of the parameter estimation provide evidence supporting this- 736 students were estimated to be demonstrating the behaviors of commuters. Another aspect of a student sample is that they can be expected to be more active in general and also more willing to explore new places. Hence, we expect students to have higher values of the mobility parameters. The observations of the mobility parameters are in keeping with the expectations. While the median values of n_w , $n_w\beta_1$ and n_w β_2 for commuters in the Greater Boston Area were 5.7, 21.2, and 286.7 respectively, for the students at DTU these have been determined to be 5.8, 25.2 and 396.8. A comparison of the probability distributions for these two groups is provided in Fig.2.20.

Having estimated the mobility parameters for our study population, the next section focuses on simulating its mobility patterns. Since the population is relatively small, different time periods of simulation have been explored.



(a)

(b)

Fig.2.20: The probability distributions of mobility parameters for (a) students at DTU, and (b) commuters in Boston. There is a clear right skew in the distributions for the students when compared to the commuters in Boston. This is expected since students tend to be more active than the average commuter. The parameter pairs (μ,σ) for the log-normal distributions for N_w and $N_w\beta_1$ for the DTU students are (3.09,0.09), (3.77,0.24). For the Weibull distribution of $N_w\beta_2$, the (k, λ) pair has the value (2.72,577)

2.2.3 Simulation Results, Validation and Observations

Now that we have the mobility parameters, we proceed to simulating the mobility patterns. As mentioned, since the population sample is not representative of the urban population, we analyze the results at the level of the individual and the urban population. In Section 2.1.2, we asserted that the use of expansive surveys is not necessary for modeling the work activities for commuters. In this analysis, we test the hypothesis by using the modeling parameters that we estimated using the NHTS in Boston for modeling students in Copenhagen.

We simulate the mobility patterns of the students for 100 days. The results are evaluated using four key metrics- stay duration, trip distance, number of locations visited in a day and the frequency of choosing the L^{th} ranked location. These metrics provide insights into both the stay characteristics as well as the destination choice and are shown in Fig.2.21.

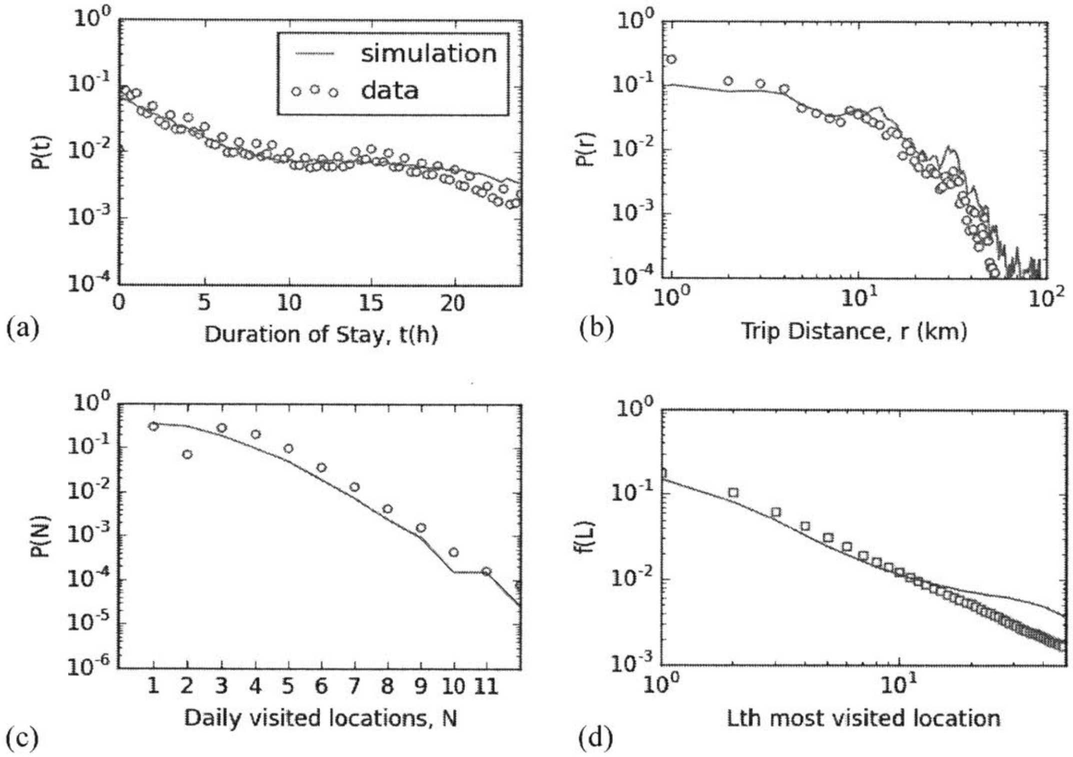
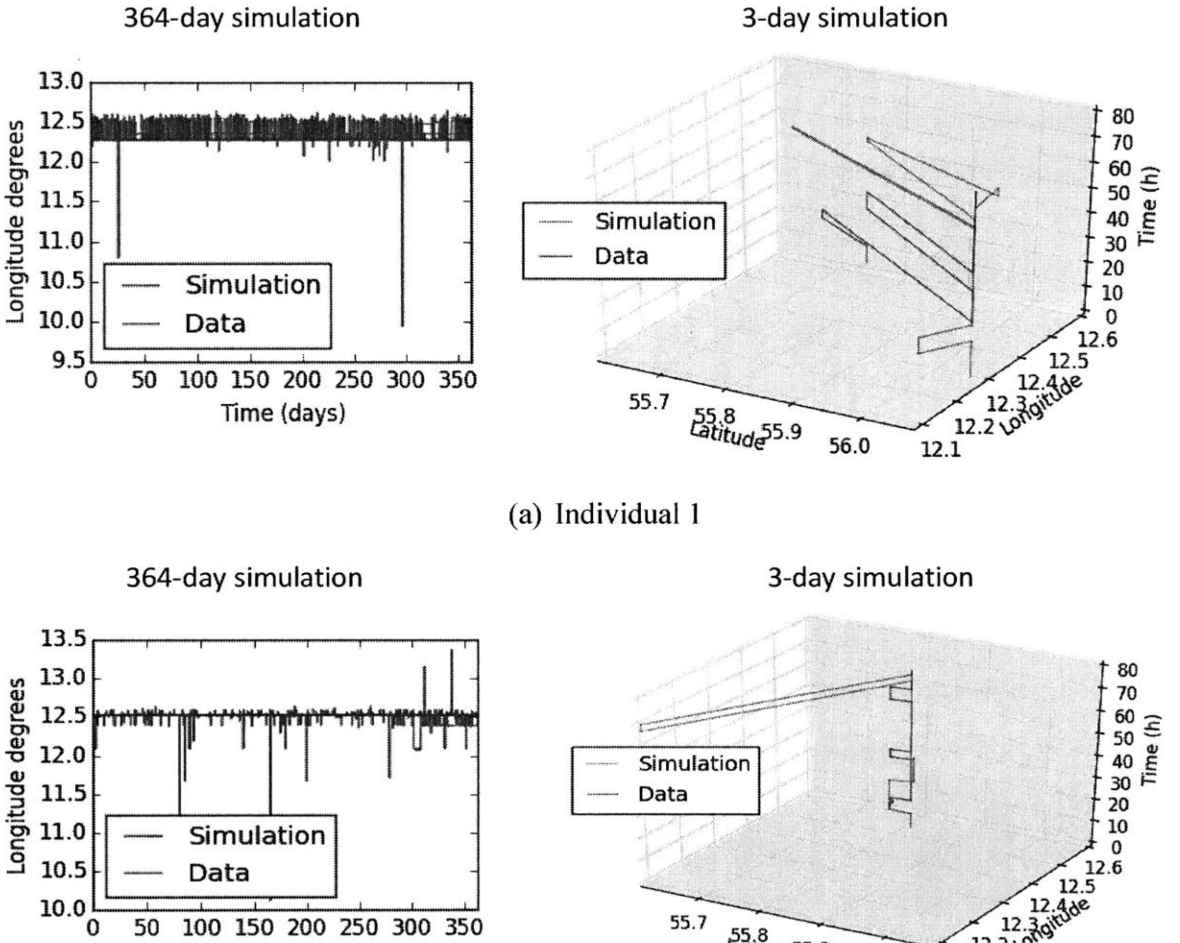


Fig.2.21: Evaluating the results for a 100-day simulation of the student population. (a) analyzes the stay durations, (b) shows the distribution of trip distance, (c) shows the number of locations visited by individuals in a day, and (d) shows the rank of the visited location. It is observed that the results of the simulation closely match the observations from the data for all four metrics.

The conformity between the data and the results of the simulation validates the quality of results from simulation.

Since the data from the Copenhagen Network Study have very high resolution, we are also interested in evaluating the quality of the resultant trajectories at the level of the individuals. For this, we randomly select two individuals from the data and simulate their mobility patterns over the short term, as well as long-term. The trajectories for these individuals are shown in Fig.2.22.



(a) Individual 1

(b) Individual 2

Fig.2.22: Simulation of two individuals over the short-term (3 days) as well as the long-term (364 days). The accuracy in terms of fraction of time-slots for which location predictions were correct is observed to be 46.9% for Individual 1 and 53.7% for Individual 2. For the 3-day simulation, the daily accuracies are (58,70.8,79.8)% and (28.5,54,65.3)% for individuals 1 and 2

At 46.9% for Individual 1 and 53.7% for Individual 2, the simulation seems to be performing moderately well at the level of the individual. And the results from both Boston and Copenhagen show that the results provide high fidelity patterns at the aggregate level. The purpose of the simulation was to model the basic mechanisms involved in mobility decisions so that the aggregate measures of urban dynamics could be captured. And these are indeed being captured well by the model in people's tendency to return to previously visited locations, their circadian rhythm and a model for choosing new destinations from a set of available options. Further improvements can definitely be made regarding individual level predictions. Some such attempts can be seen in (60-71). An important feature of this framework is that it is sufficiently modular to incorporate improvements in individual mobility while maintaining its framework to provide accurate urban scale measures.

2.3Summary

This chapter has been able to demonstrate the scalability and portability of the proposed modeling framework. In the remaining chapters, we discuss novel ways of making the simulated results from low-resolution data conditions usable at the microscale- the level of individual buildings and points of interest.

Chapter 3

Hyperlocal mobility: Setting up the analysis

3.1The need for Hyperlocal data

CDRs constitute the most ubiquitous large scale mobility traces present across the world. The accuracy of the locations indicated in these traces is often coarse as observed in the data from Boston. When expansion from the sample to the population is performed, location biases are introduced in favor of the sample population. In other words, when the results from the expanded population are analyzed, the set of locations frequented by the sample population will be visited more frequently. This is induced by construction since in the absence of rich location histories for the entire population, the location histories of the sampled users need to be replicated for users with poor sampling rates. This will be a challenge even if the data are of higher resolution. Since the sample population is large enough with CDR records, the replication of travel histories is able to maintain the accuracy of flow measures at an aggregate level such as at the level of census tracts in the case of Boston. This has been illustrated in the section 2.1.4.

In order to talk about microscopic mobility patterns, we therefore need to come up with an alternate method for analysis at a high resolution. Given that we have demonstrated the accuracy of stay counts and details at an aggregate level, this information can serve as the starting point of estimating high-resolution (within-tracts in case of Boston) mobility patterns.

In order to distribute people within tracts, we need to gather information about engagement opportunities available within the tracts. Once these data are available, we can proceed to estimating the likely location where a stay with particular characteristics should occur. Since no exhaustive dataset detailing all characteristics of hyperlocal features is available, there is a need to construct such a dataset using a variety of disparate data sources. This chapter focuses on constructing such a dataset. Later, in Chapter 4, we discuss the distribution of stays within tracts based on stay characteristics and hyperlocal data.

3.2Required hyperlocal characteristics

Buildings and Points of Interest (POIs) serve as the center of engagement opportunities in a region.

The task at hand requires a behavioral matching between the stays in a region and the engagement opportunities within that region. The information available for each stay includes-

1. Purpose of stay
2. Starting time of the stay
3. The duration of the stay
4. ID of the person performing the stay

Data about buildings within a tract provides the basic layer which can be enriched to perform stay allocation. There are two key characteristics for a dataset to be a valid input to our model- it should be ubiquitously available and accessible across the world and, secondly, the information contained in it should be orthogonal to the objective of the analysis. For example, the dataset of building outlines is admissible because building outlines are available and accessible for most places in the world. Using taxi pickup and drop-off locations is however not valid because these are not ubiquitous. They are also not available in a consistent format which is publicly accessible. It is also not appropriate to use popular times from Google Places because the ability to talk about popular times is the very objective of this analysis. Using these to distribute stays will cause a bias for distributed profiles to match up to them whereas coming up with an alternate method was the objective in the first place.

To be able to match a stay to a building, we need information about the functional use of the buildings and also its hours of operation. In addition, the surroundings of a building can also play a role in dictating the number of stays attracted to it. Keeping this in mind, the following sources of data have been leveraged to distribute stays within tracts-

1. Building outlines and heights
2. Building classification
3. Per capita floor area required for different functional uses of buildings
4. Locations, classes and hours of operation of points of interest from Google Places

In the following section, we will show how these data can be combined to obtain building *objects* suitable for matching to stays. The actual process of matching is discussed in depth in Chapter 4.

3.3Hyperlocal datasets

In this section, we describe how each of the datasets listed above can be used to create building *objects* that can act as attractors for stays.

3.3.1 Building Outlines, Heights and Classification

The Boston Department of Innovation and Technology published a dataset of buildings in the City of Boston in January 2012 (72). This dataset provides a variety of attributes for each building including building name, type, footprints, height (in feet), shape/outline, zip code and address. There is a total of 82,542 buildings in the City of Boston. The geographic extent is shown in Fig.3.1.

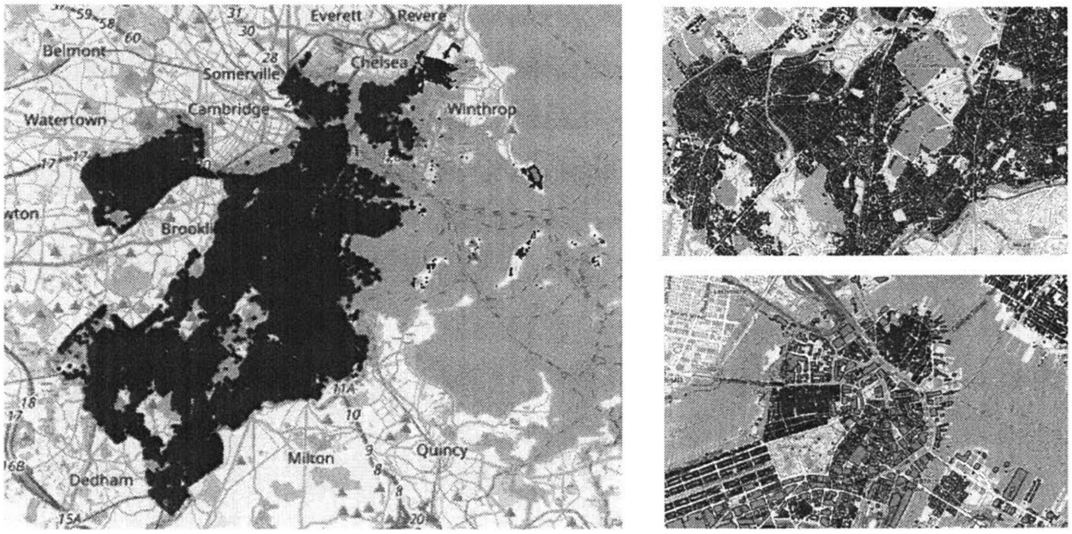


Fig.3.1: The Building outlines available for the City of Boston. The left panel shows the extent of the City while the panels on the right show building details in (top) Roslindale, the southernmost extent of the city, and (bottom) Downtown (73)

The building type provides insights into the functional use of the building and is therefore critical to matching buildings to stays. The building type is divided into the following classes in the dataset-

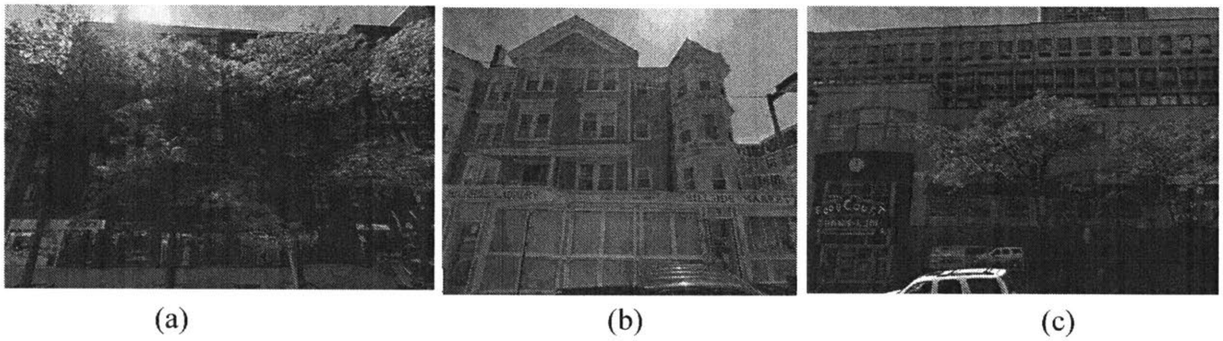
1. Apartments (A)
2. Condo Main (CM)
3. Residential Land (RL)
4. Single Family dwellings (R1)
5. Double Family dwellings (R2)
6. Triple Family dwellings (R3)
7. 4-6 Apartment units (R4)

8. Commercial (C)
9. Exempt (E)
10. Industrial (I)
11. 121-A Property (EA)
12. Commercial Parking (CP)
13. Multipurpose Residential (RC)

For the purpose of the analysis in the paper, these classes are grouped into 3 buckets- Residential, Commercial and Other uses. The building classes in each bucket are shown below-

1. Residential: A, CM, RL, R1, R2, R3, R4, RC
2. Commercial: C, E, I, EA, RC
3. Others: CP

Processing 'RC' buildings: Buildings in the use class *RC* have both a residential and a commercial component to them. Mostly they exist as the first floor being dedicated to the Commercial component of the building and the second floor onwards being residential (Figure A2).



(a) (b) (c)

Fig. 3.2: Examples of buildings in the RC use case. Panel A- A line of shops on the first floor of an apartment building on Newbury Street; Panel B- A Laundry/grocery shop under apartments in Mission Hill; Panel C- A food court under an apartment building in Longwood

When such buildings are analyzed, they have been broken down into 2 components- a residential part and a commercial part. If the building has more than 1 floor, the commercial part has been assumed to span only the first floor (hence the gross area for it is equal to the building footprint) and the remaining floors are assumed to be used for residential purposes.

If the building has only 1 floor, the area is split equally between the residential and commercial components.

Since our analysis begins at the level of census tracts, the tract within which each building falls is determined by using the building location. In addition to the building outline, information about the building location (latitude and longitude) is available in the dataset and these are the columns used for determining the census tract within which the building lies.

Further, the number of floors in a building is determined by assuming the height of each floor to be 10 feet. The area of the building divided by 10 floored to the nearest integer is used to create a column for the number of floors in each building.

Though the dataset does have columns for the census tract of a building as well as the number of floors in the building, these are sparsely populated and the approach detailed above is used to compute these details.

The building footprint and the number of floors are used to compute the gross floor area of the buildings. The gross floor area is taken to be the product of the building footprint and the number of floors in the building.

One other aspect of the dataset is that the coordinates are provided with reference to the North American Datum. This original shapefile was converted into the WGS 84 reference system using QGIS.

In summary, after processing this shapefile, we get the building type and are able to compute the gross floor area for each building. Two important characteristics of buildings, namely the maximum occupancy and the hours of operation still need to be estimated for each building.

3.3.2 Per-capita Floor Area

In section 3.3.1, we were able to estimate the gross floor area for buildings. The number of people that a building can accommodate, however, depends on more than just the gross floor area. It also depends on the functional use of the building. For example, the area required per person at home is significantly higher than that required in a classroom.

To translate the gross floor area into building capacity (or maximum building occupancy), we rely on a mapping between the functional use of buildings and the per capita space required for that use case. One such mapping is provided by values used to compute the human sensible and latent heat loads as indicated in the Engineering Toolbox (74). The values provided in this data source are intuitive and can be changed at any point if more detailed data become available.

Table 3.1: The area per person (per capita area or PCA, in square meters) mapped to the functional use of buildings as indicated by the Engineering Toolbox

S.No.	Toolbox Use	PCA
1	Retail Store	4.2
2	Shops- supermarkets	2
3	Cafe	2.8
4	Banks	9.3
5	Bars	3
6	Restaurant with Service	1.5
7	Motels/Dorms	14
8	Assembly Building	0.6
9	School	2
10	Tavern	3
11	Offices- single	10
12	Medical centers- clinics and offices	9.3
13	Municipal buildings	9.3
14	Hospitals	9.3
15	Restaurant without service	1
16	Museums	6
17	Sports- gymnasium	1.5
18	None	0
19	Library	5
20	Offices- meeting room	1.5
21	Malls	7
22	Other-Laundry	4
23	Other-Park	1
24	Police Stations	25
25	Post Offices	25
26	Nightclubs	3

The functional use of buildings is thus far available in 3 broad categories- Residential (74875), Commercial (6241) and Industrial (1303). More details about functional uses is obtained along with the POI details in the next section.

3.3.3 Points of interest and Estimating Building Characteristics

The Google Places API Web Service is a service that returns information about places- defined within this API as establishments, geographic locations, or prominent points of interest- using HTTP requests. Each of the services is accessed as an HTTP request, and returns either an JSON or XML response. All requests to a Places service must use the https:// protocol, and include an API key. The Google Places API Web Service uses a place ID to uniquely identify a place

We use the Google Places API to determine the points of interest located throughout the City of Boston.

The process of querying POI details in Google Places consists of 2 main steps:

1. Determining the set of POI ids in the proximity of a given location
2. Individual query for each POI id to obtain details about that POI

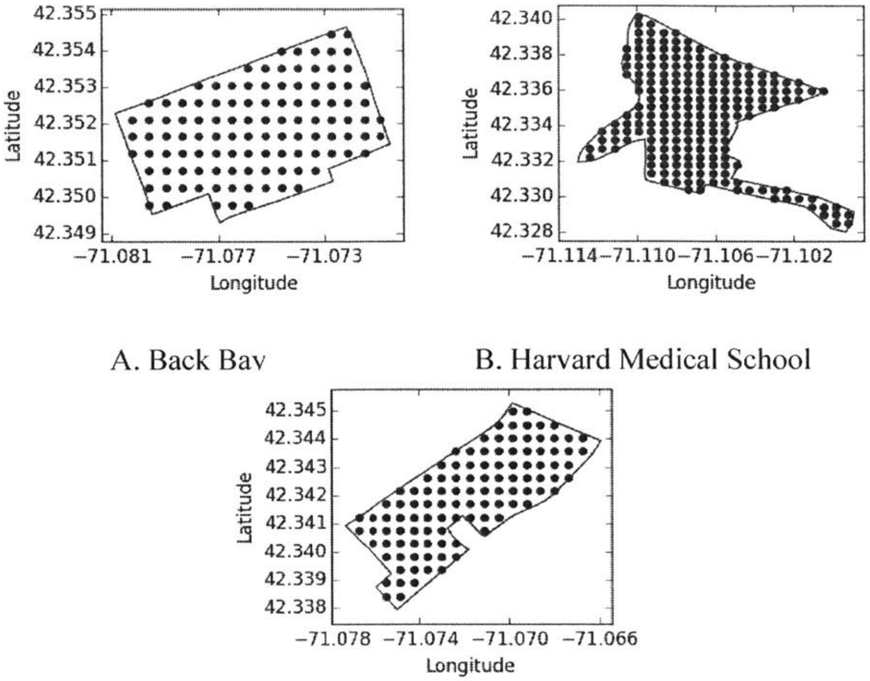
The definition of 'proximity' or the radius in which we scan for POIs around a location is important since it influences the type of results that we obtain from the query. For example, if the query radius is several hundred kilometers, we will end up with names of cities in this radius while if the radius is of the order of tens of meters, we get richer local information.

Keeping this in mind, we decide to use a radius of 50m for the queries within tracts. We define a grid of points with a spacing of 70 (50*√2) meters. This is shown in Figure 3.3.

For each grid point, the query determines the set of POI ids within 50 meters of the point. The set of POI IDs is then aggregated and repeated values, if any, are discarded. After this, in the second phase, individual queries are used to obtain details about each POI ID.

The details about a POI include the place name, location, categories, address, hours of operation, ratings, reviews etc. Some of these details might not be returned by queries based on their availability and relevance for the particular POI. For example, reviews are not always available for restaurants and hours of operation are irrelevant for street names.

To illustrate the results and the working of the mining process, we select 3 census tracts- Back Bay, South End and Harvard Medical School that offer variability in land use patterns. However, the analysis has been performed, and results are available for all 173 tracts in the City of Boston.



A. Back Bav

B. Harvard Medical School

A. South End

Fig.3.3: The grid of POI query points over the 3 tracts under analysis

Hours of Operation of buildings

The details of the POIs obtained above include the location coordinates for each POI. These locations are used to determine the set of POIs within each building. Not all POIs lie within buildings (places of interest such as statues, memorials, lakes and parks are located outside buildings) and not all buildings (especially residential) have associated POIs. In Back Bay, a total of 933 POIs were found out of which 698 were inside buildings. However, only 172 out of 372 buildings had POIs. For Harvard Medical School, 622 out of 974 POIs were found inside 76 buildings (the number of buildings is very small since some hospital/university buildings in this tract are very large). For South End 277 of the 555 POIs were found inside 110 of the 736 buildings (since most of the buildings are residential).

In the simplest case, a building has a single POI with hours of operation available for it. In this case, the hours of operation for the building are taken as those of the POI. When hours of operation are available for multiple POIs within the building, the hours of operation for the building are assumed to be the superset of the hours of operation of the POIs.

Table 3.2: Matching between the uses indicated in the POI information and the Engineering Toolbox uses. This in turn enables us to match buildings with POIs to a PCA value

S.No.	Toolbox Use	POI Uses
1	Retail Store	Store, book store, jewelry store, liquor store, electronics store, shoe store, clothing store, furniture store, pharmacy, home goods store, florist, bakery
2	Shops- supermarkets	Grocery or supermarket, convenience store, department store
3	Cafe	cafe
4	Banks	bank
5	Bars	bar
6	Restaurant with Service	restaurant
7	Motels/Dorms	lodging
8	Assembly Building	Church, synagogue, place of worship, Hindu temple
9	School	School, university
10	Tavern	Food
11	Offices- single	Accounting, lawyer, real estate agency, general contractor, painter, spa, car repair, roofing contractor, moving company
12	Medical centers- clinics and offices	Dentist, doctor, physiotherapist
13	Municipal buildings	Local government office, embassy
14	Hospitals	Hospital
15	Restaurant without service	Meal takeaway
16	Museums	Art gallery
17	Sports- gymnasium	Gym, health
18	None	ATM, street address, parking, premise, route
19	Library	Library
20	Offices- meeting room	Finance, travel agency, beauty salon, hair care, insurance agency
21	Malls	Shopping mall
22	Other-Laundry	Laundry
23	Other-Park	park
24	Police Stations	police
25	Post Offices	Post office

Most residential buildings do not have POIs within them and are always assumed to have 24-hour operation. If a non-residential building does not have POIs with operating hours associated with it, the hours of operation for the building have been assumed to be 7am to 10pm. This is a fairly conservative estimate and should enable these buildings to be matched to most non-residential stays.

Per-Capita Area for buildings

As described above, we obtain the set of POIs that lie within each building and POIs have classes attached to them. If there is a single POI within the building, the building is assumed to have the functional use corresponding to that POI and the PCA is chosen accordingly from Table 3.2. If there are multiple POIs within a building and they correspond to multiple possibilities of functional use and PCAs, the highest value of the PCA has been assumed as the PCA for the building.

For residential buildings, the PCA is assumed to be 250square feet (23.2 m^2).

For non-residential buildings, if no POIs are found within the building the PCA is chosen randomly from the distribution of PCAs of other non-residential buildings in the tract. Hence, if there are several restaurants in a tract and a particular non-residential establishment does not have a POI associated with it, there would be a high probability of it being allocated the PCA of a restaurant.

Capacity of a building

The PCA is used to estimate the capacity of a building. The gross floor area of the building is divided by the per-capita area based on the functional use to derive the occupancy of the building.

$$Capacity = \frac{Gross\ Floor\ Area}{PCA}$$

3.4Summary and Conclusion

We started off this chapter with a basic building shapefile, and an aim of constructing building objects with sufficient detail so that individual stays could be mapped onto them. Along the way, we added further meat to the basic bones of the building object using a variety of datasets.

Basic building classification needed to be elaborated so that the precise functional use of the buildings could be determined in order to estimate the per-capita area and the capacity of each building. In the end, the building objects are detailed enough to proceed to matching them to stays obtained from the simulation. The matching process can be used to come up with occupancy profiles for a building and is discussed in the next chapter.

Chapter 4

Estimating Building Occupancy

4.1The Task of Matching Stays

In the previous chapter, we created building objects and enriched them by adding information from various data sources. The stay objects are of the type (User ID, Starting Time, Ending Time, Type-H/W/O) and the building objects are now of the form (Building ID, type, Capacity, Operating Hours, Location, Shape). There is sufficient information in the building objects to proceed to matching stays to buildings.

Given that there are likely to be multiple buildings to which a stay could be mapped, the task of matching should be smart enough to incorporate the likely mechanism of decision making at the individual level. In this chapter, we outline the process of allocating stays to buildings. Since no primary data are available to verify the results, we will start off with a simple model and refine it along the way to incorporate different aspects of urban mobility.

We present the results from three tracts as before and the analysis has been performed for the entire City of Boston (173 tracts). The next section provides the details of the tracts, after which we dive into the process of matching.

4.2The Test Tracts

The different building classes and their frequency distribution across the City of Boston are shown in Figure 4.1.

The choice of the three tracts, Back Bay, South End and Harvard Medical School, was not arbitrary. They were systematically chosen since they provide disparate distribution of building types. The building outlines and categories are shown in Figure 4.2.

Before providing details about each of the three tracts it is important to refresh the genesis of these tracts. Tracts are defined as part of the census. Each tract contains about 5,000 residents and the shapes are roughly adjusted as per local geography. Key pieces of information including total residential and working population are provided for each tract.

Back Bay is mostly a mixture of mid to high end commercial establishments and condominiums. The total area of the tract is 3.24M square feet with about 1.26M square feet or about 40% of the area being under its 372 buildings. The gross floor area of the buildings is 9.97M square feet, which is about 3 times the total area of the tract and 8 times the total area under buildings.

South End presents a predominantly residential setting. At 3.21M square feet, it has roughly the same area as the tract in Back Bay. The total building footprint is lower at 0.94M square feet constituting about 29% of the total floor area. The 736 buildings have a gross floor area of 4.19M square feet which is only 1.3 times the area of the tract and about 4.46 times the total area under buildings. South End is known for its residential parks which would explain the lower fraction of area under buildings. In such a setting, we could also expect a larger fraction of total stays to occur outdoors.

The tract covering the Harvard Medical School presents a very different use-case. Tax-exempt properties constitute the major fraction of buildings. These are buildings owned by the university. Being a medical school and having several hospitals associated with it, several of these buildings are used like a home location would be used. The total area of the tract is 6.04M square feet. About 34% of this area or 2.09M square feet is covered by buildings whose gross area is 11.62M square feet which is 1.92 times the area of the tract and 5.56 times the total building footprint.

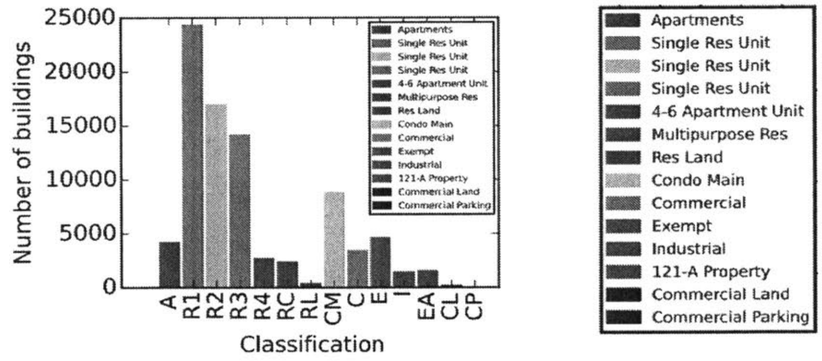
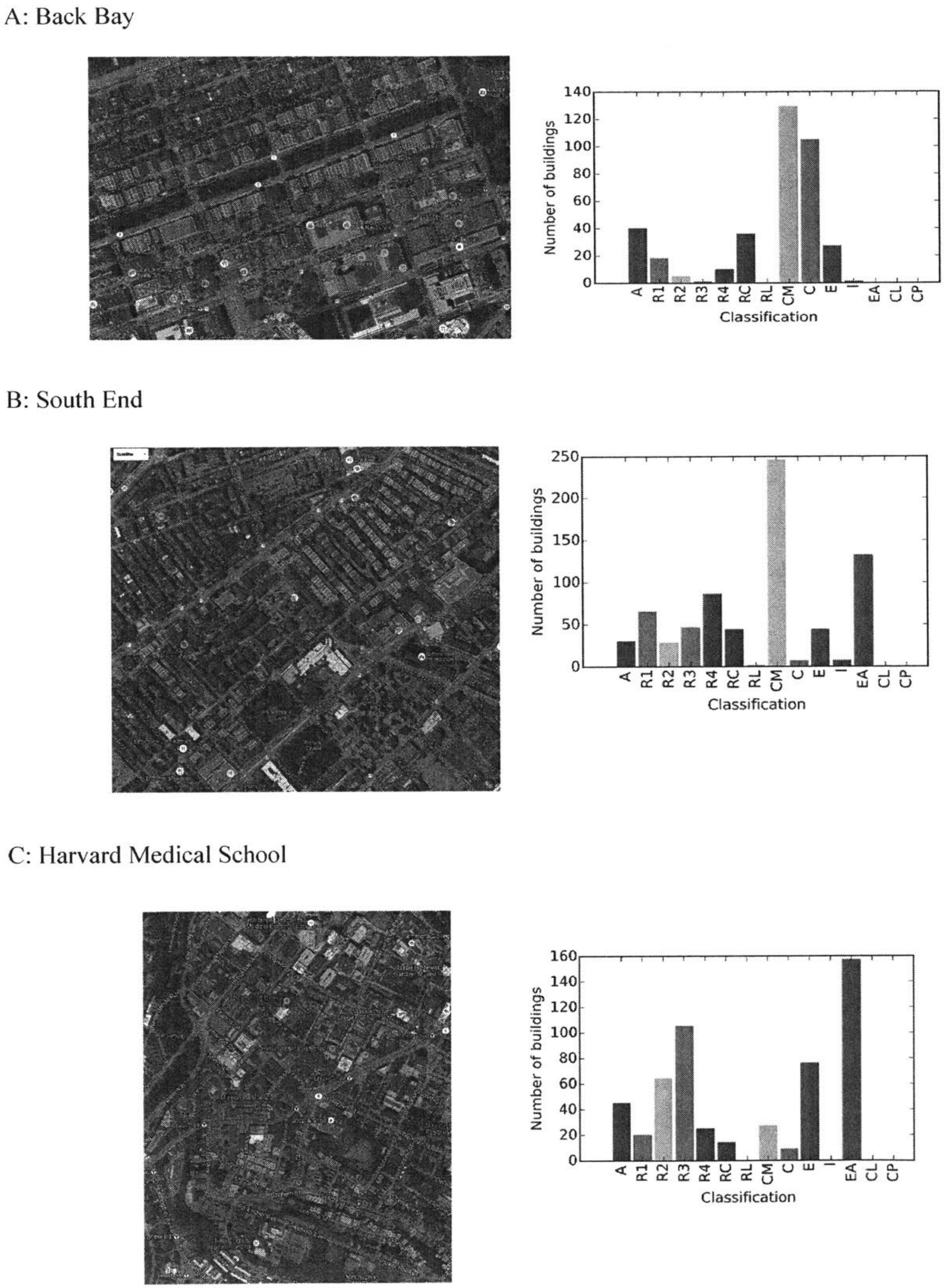


Fig. 4.1: Distribution of different building classes in the City of Boston

Even within the residential building categories, the tracts feature different categories of units. Back Bay and South End mostly consist of residential condos, with South End having a number of 2, 3 and multi-family residential units spread across. HMS on the other hand features very few condos and has a lot more double and triple family homes and apartments.



A: Back Bay

B: South End

C: Harvard Medical School

Fig. 4.2: Distribution of building types in the chosen tracts (75)

4.3Model 1

As described in Chapter 3, we obtain a range of local data for smarter distribution of people into buildings. Once we have all the hyper-local data processed for each building, we move on to allocating stays to individual buildings. The process can be outlined in the following steps:

1. Determination of the set of buildings compatible with the type, timestamp and duration of the stay
2. Probabilistically choosing a building to allocate the stay based on the gross available floor area in the building

Both of these steps are explained in more detail below.

All stays are not attracted to all buildings. We define the following 3 rules to establish compatibility between stay type and building class-

1. Home based stays can be attracted to residential buildings alone
2. Work stays occur only in non-residential buildings
3. Other stays can take place in either residential or non-residential buildings

Once the set of compatible building has been determined for the type of stay to be distributed, we need to determine the set of buildings that are compatible with individual characteristics of the particular stay under consideration. The following checks are performed on the set of compatible buildings to compile a list of feasible matches-

1. Operating Hours: the operating hours of the building should overlap completely with the duration of the stay.
2. Available capacity: The capacity for each building has been computed as described in Chapter 3. The magnitude of attraction of a building is proportional to the capacity of a building.

The process of distribution begins by segregating stays into three classes- Home, Work and Other. The stays in each of these buckets are then arranged in chronological order by starting time.

For the purpose of distribution, the home stays are distributed first, followed by work stays and finally Other stays. The sequence of distributing Home and Work stays is interchangeable but both of these should be distributed before Other stays. The rationale behind this is that Other stays can be allocated to both residential and non-residential buildings while Home and Work stays can be associated with only residential and non-residential buildings respectively. By allocating Other stays prior to (say) Home stays, there is a possibility for all residential locations to be occupied by

Other stays rendering the set of feasible matches for Home stays to a null set while leaving alternate options for Other stays untapped.

With each stay allocation, the available capacity in a building is updated and future calculations are based on this residual capacity. The allocation of a building to a stay is done probabilistically on the set of feasible matches. The probability of a stay being allocated to a building is proportional to the residual space in the building (the maximum occupancy minus the number of agents already in the building at the given timestamp). This prevents overflow in buildings and forces empty buildings to be occupied once other buildings start filling up.

The results from this distribution are shown in Figure 4.4. The process is summarized in Figure 4.3.

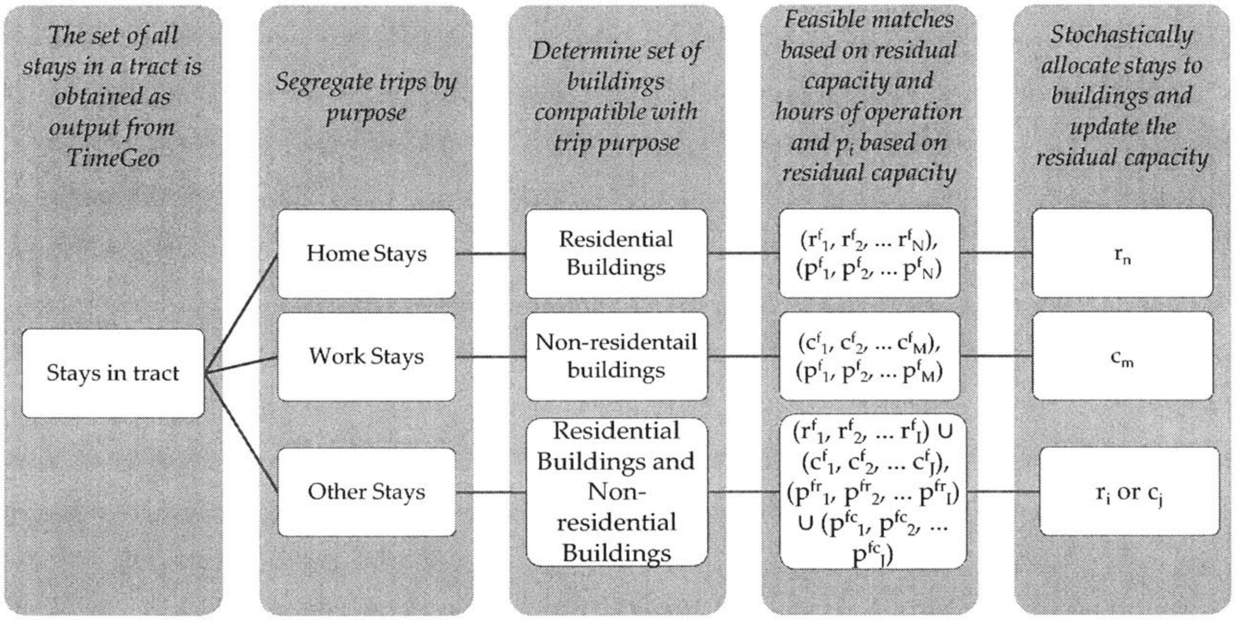
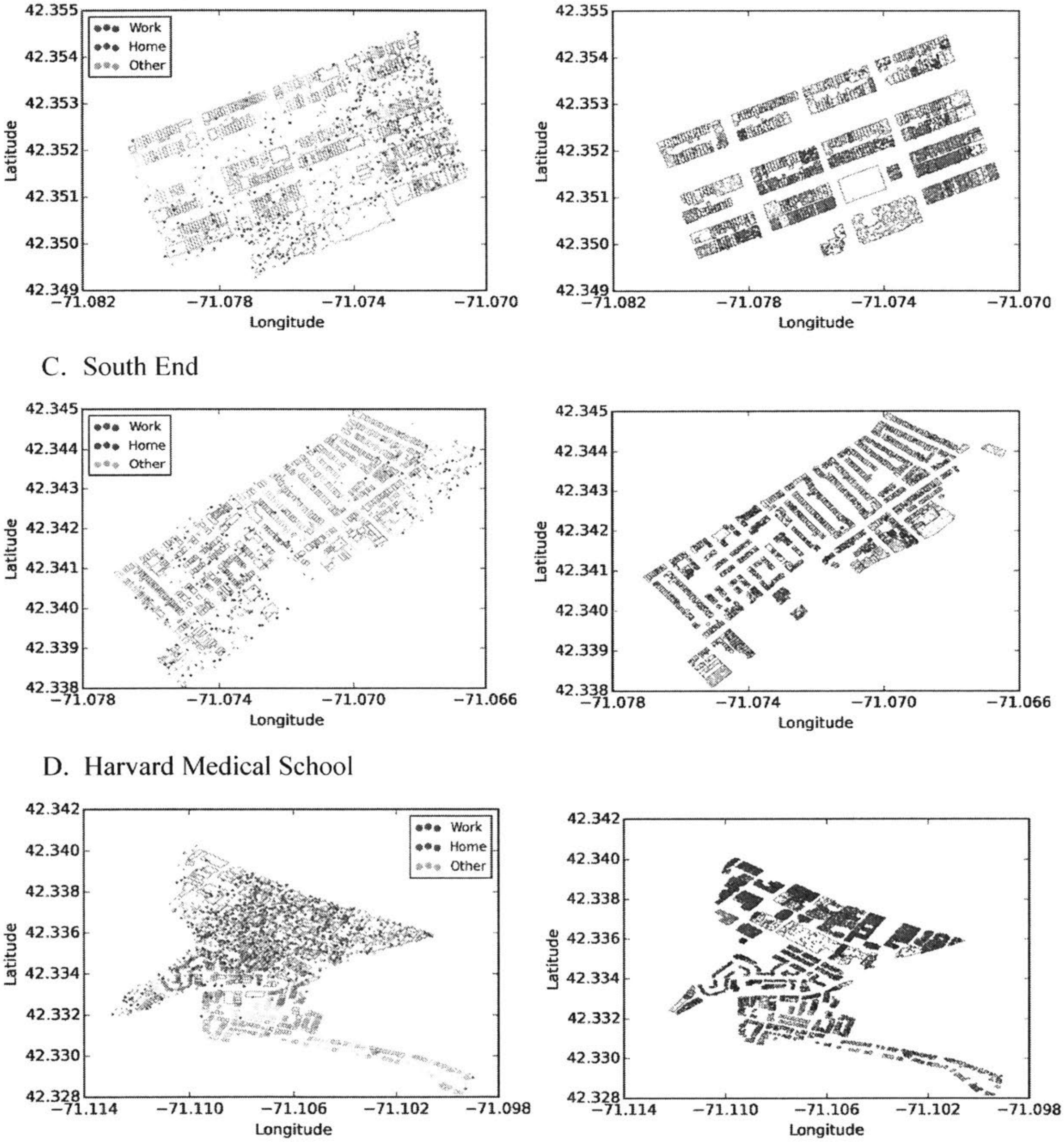

The set of all stays in a tract is obtained as output from TimeGeo
Segregate trips by purpose
Determine set of buildings compatible with trip purpose
Feasible matches based on residual capacity and hours of operation and p_i based on residual capacity
Stochastically allocate stays to buildings and update the residual capacity
Stays in tract
Home Stays
Work Stays
Other Stays
Residential Buildings
Non-residentail buildings
Residential Buildings and Non-residential Buildings
$(r^f_1, r^f_2, \dots r^f_N)$, $(p^f_1, p^f_2, \dots p^f_N)$
$(c^f_1, c^f_2, \dots c^f_M)$, $(p^f_1, p^f_2, \dots p^f_M)$
$(r^f_1, r^f_2, \dots r^f_I) \cup (c^f_1, c^f_2, \dots c^f_J)$, $(p^{fr}_1, p^{fr}_2, \dots p^{fr}_I) \cup (p^{fc}_1, p^{fc}_2, \dots p^{fc}_J)$
r_n
c_m
r_i or c_j

Fig. 4.3: Outline of the process of building allocation in Model 1

The initial distribution of trip locations and final allocation for each of the tracts is shown in Fig. 4.4. As was hypothesized, the (for example) home locations from the CDR data are hardly ever observed in residential buildings, which asserts the need for the process of building allocation.



B. Back Bay

C. South End

D. Harvard Medical School

Fig.4.4: Results of assignment of stays to buildings within the three tracts. The left panel in each figure shows the locations indicated in the CRDs. The right panel maps each of those stays to buildings.

4.4Model 2

The model described above incorporates local information while distributing stays within tracts. It however fails to account for two important phenomena- urban dynamics and individual behavior. Census tracts are defined as urban regions with about 5,000 residents. As can be seen from the three tracts under consideration, tracts are big enough for there to be significant variability in land utilization within them. The ability of a location (such as restaurants, pubs, gyms etc.) to attract people/stays is not solely a function of floor area as assumed in the previous model. Several other factors including quality of service, accessibility of location and popularity of region also contribute to the magnitude of attraction of a destination. For example, a restaurant located on Newbury Street in Back Bay if likely to attract more customers than one on Marlborough Street. This is mainly because a larger number of activities are likely to be clustered together in the hub of activity and engagement opportunities provided on and around Newbury Street. In the previous model, however, both these locations would be inferred as having equal attraction to stays.

The second factor that the model is unable to account for is how agents make choices regarding their destinations. When an agent performs consecutive activities within the same tract, there is a higher likelihood of choosing the subsequent stay location which is closer to the previous stay location.

Both these factors are accounted for in this second model. The process of allocating a building to a stay in this model has been elaborated to span the steps detailed in Figure 4.5.

The transition from layer 1 to layer 2 encompasses stay segregation, the determination of compatible buildings and the determination of feasible matches as covered in the transition from layers 1 through 4 in model 1. In the transition from layer 2 to layer 3, the feasible matches are clustered based on spatial density to capture the likely perception of the agent performing the trip. Once the clusters are determined, the probability of the agent to choose one of the determined clusters is computed based on a super-linear relation with the gross building area in the cluster. After a cluster has been chosen, a building is chosen within the cluster based on the probabilities proportional to the total capacity of the buildings.

Detailed explanation of these steps is provided below. Section 4.6 lists some limitations of this model and explores possibility of improvement. Section 4.7 provides some concluding remarks on the distribution of trips within tracts.

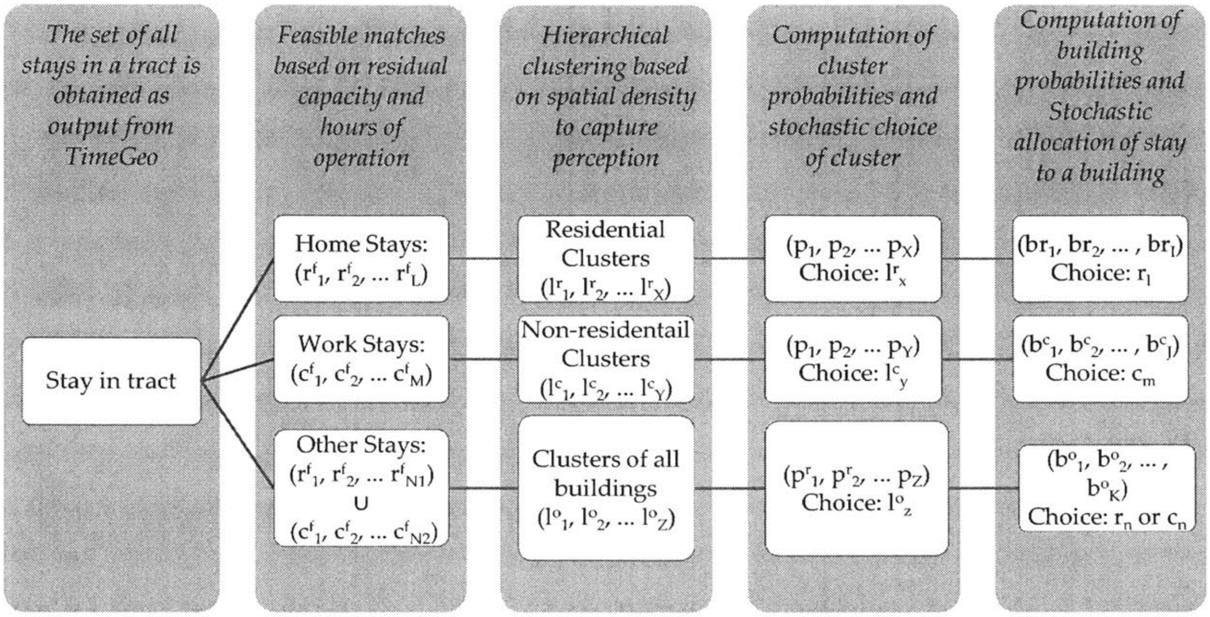

The set of all stays in a tract is obtained as output from TimeGeo
Feasible matches based on residual capacity and hours of operation
Hierarchical clustering based on spatial density to capture perception
Computation of cluster probabilities and stochastic choice of cluster
Computation of building probabilities and Stochastic allocation of stay to a building
Stay in tract
Home Stays: $(r^f_1, r^f_2, \ldots r^f_L)$
Work Stays: $(c^f_1, c^f_2, \ldots c^f_M)$
Other Stays: $(r^f_1, r^f_2, \ldots r^f_{N1})$ U $(c^f_1, c^f_2, \ldots c^f_{N2})$
Residential Clusters $(l^r_1, l^r_2, \ldots l^r_X)$
Non-residentail Clusters $(l^c_1, l^c_2, \ldots l^c_Y)$
Clusters of all buildings $(l^o_1, l^o_2, \ldots l^o_Z)$
$(p_1, p_2, \ldots p_X)$ Choice: l^r_x
$(p_1, p_2, \ldots p_Y)$ Choice: l^c_y
$(p^r_1, p^r_2, \ldots p_Z)$ Choice: l^o_z
$(br_1, br_2, \ldots, br_I)$ Choice: r_l
$(b^c_1, b^c_2, \ldots, b^c_J)$ Choice: c_m
$(b^o_1, b^o_2, \ldots, b^o_K)$ Choice: r_n or c_n

Fig. 4.5: Outline of the process of building allocation in Model 2

4.5.1. Determination of building clusters

This step accounts for interaction amongst buildings/places within a block. It is easily observed that regions with a large number of engagement opportunities tend to be more popular than those with fewer opportunities. This 'neighborhood interaction' increases the attractiveness of each building within a popular cluster of buildings.

It should be noted that these clusters are not formally defined and depend on the perception of the individual and the purpose of the stay. We identify clusters for each of the 3 classes of stays- H, W and O. The identified clusters need to account for spatial proximity and density. The method chosen for this is Hierarchical Agglomerative Clustering (HAC) with Ward's minimum-variance method (76). Figure 4.6 shows the outline of the buildings colored by class in the three tracts under consideration without the underlying map so that building distribution and subsequent clustering become more prominent.

Clustering for 'Home', 'Work' and 'Other' stays

As mentioned in Section 3.3.1, residential units are associated with building uses 'A', 'R1', 'R2', 'R3', 'R4', 'RL', 'CM', 'EA' and 'RC'. Focusing just on these residential use cases, we run HAC to determine the residential clusters. The classification, along with the corresponding dendrogram, is shown in figure 4.7. These clusters should act as the major attractors of 'Home' stays.

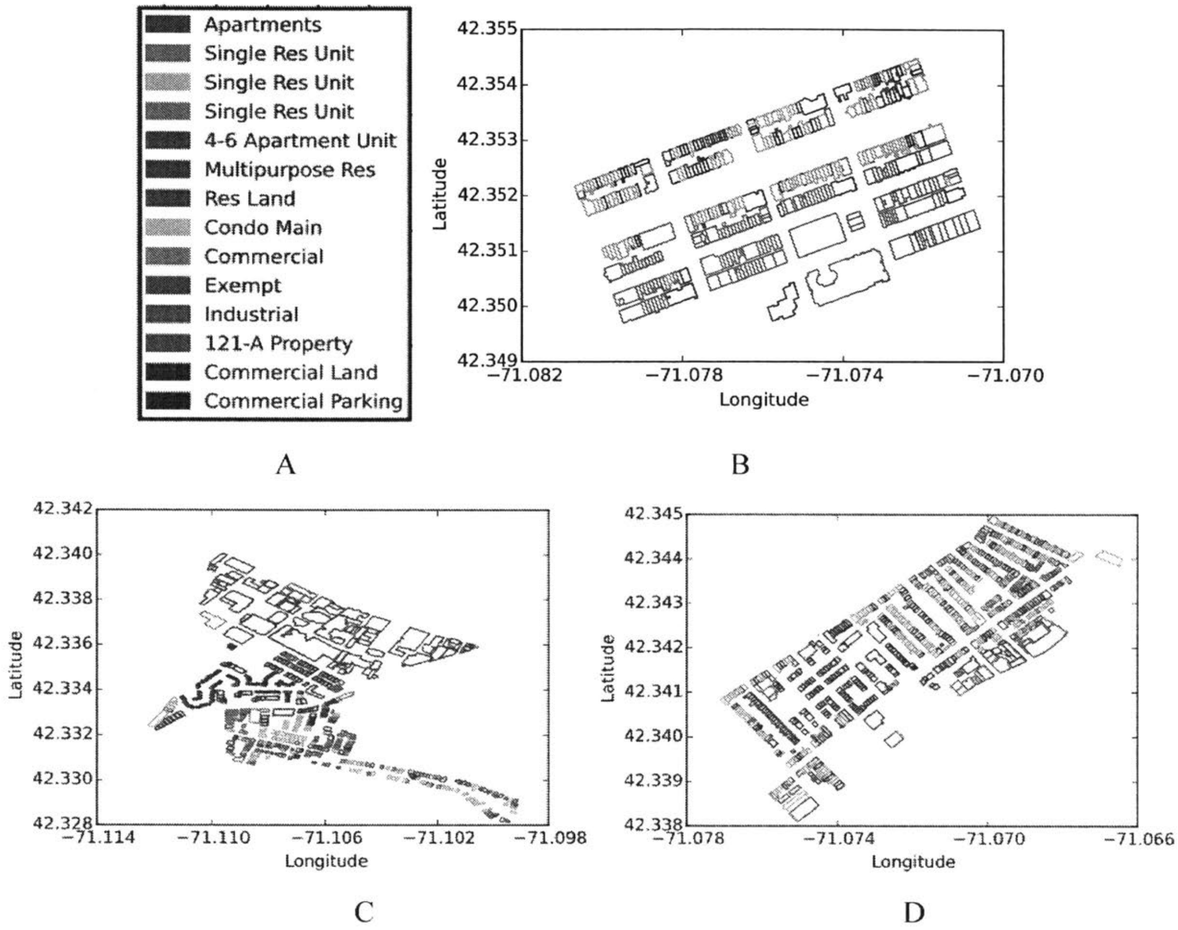


Fig. 4.6: Building outlines (without footprint) and classification for the tracts under analysis. Panel A shows the color-coded classes into which the buildings are classified (same as Figure 4.1), Panel B shows the footprint of buildings in Back Bay, Panel C shows building footprints at the Harvard Medical School and Panel D shows those for buildings in South End

Similar analysis has been performed in Figures 4.8 and 4.9 for buildings attracting 'work' and 'other' stays. For work stays, we consider buildings with categories of use 'C', 'E', 'I' and 'RC'. For 'other' stays, all buildings are assumed to be valid candidates. The clusters shown in these figures have been estimated under the assumption that all buildings are operational.

We now incorporate operating hours and demonstrate the variance of these clusters with time.

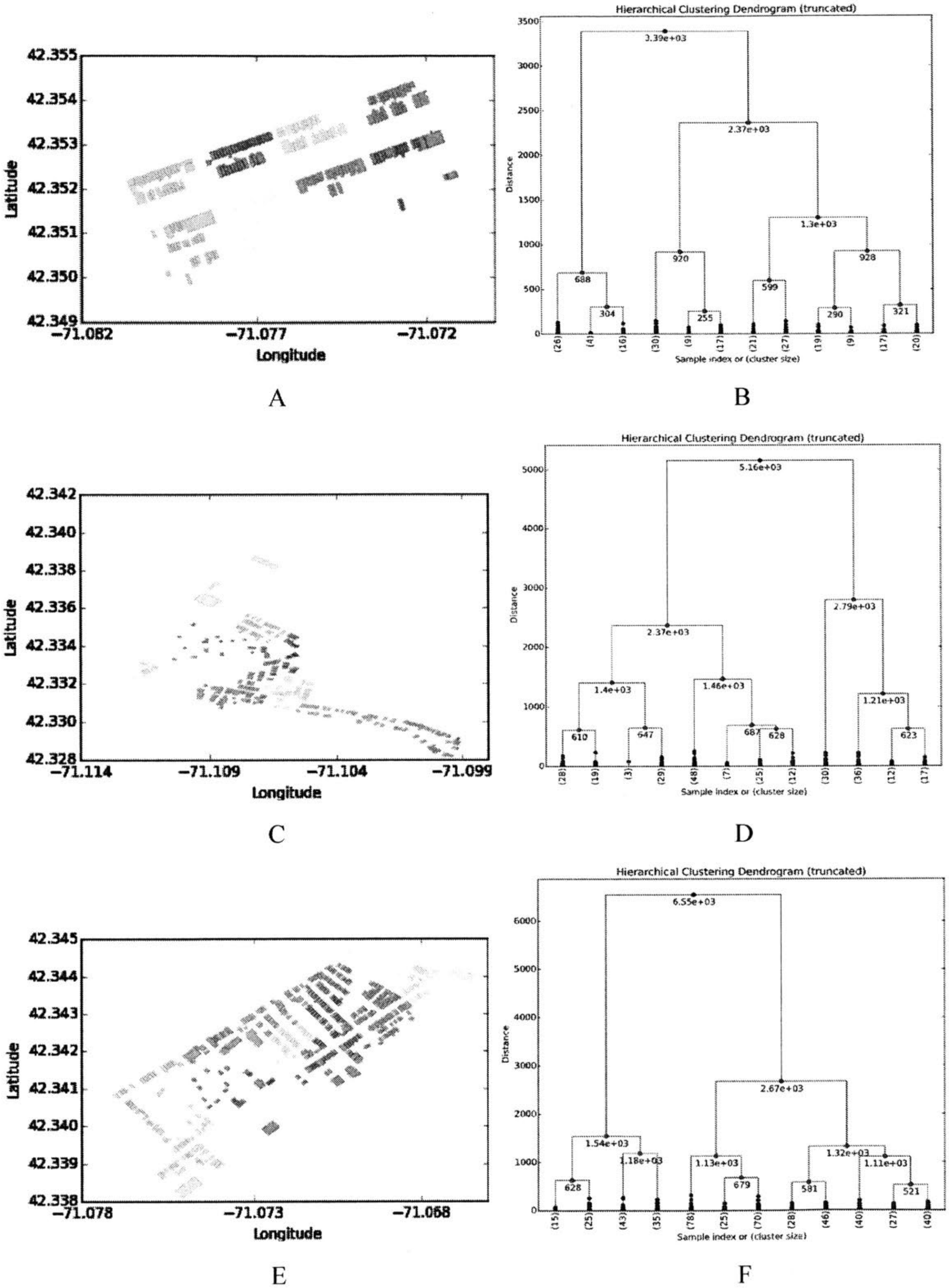
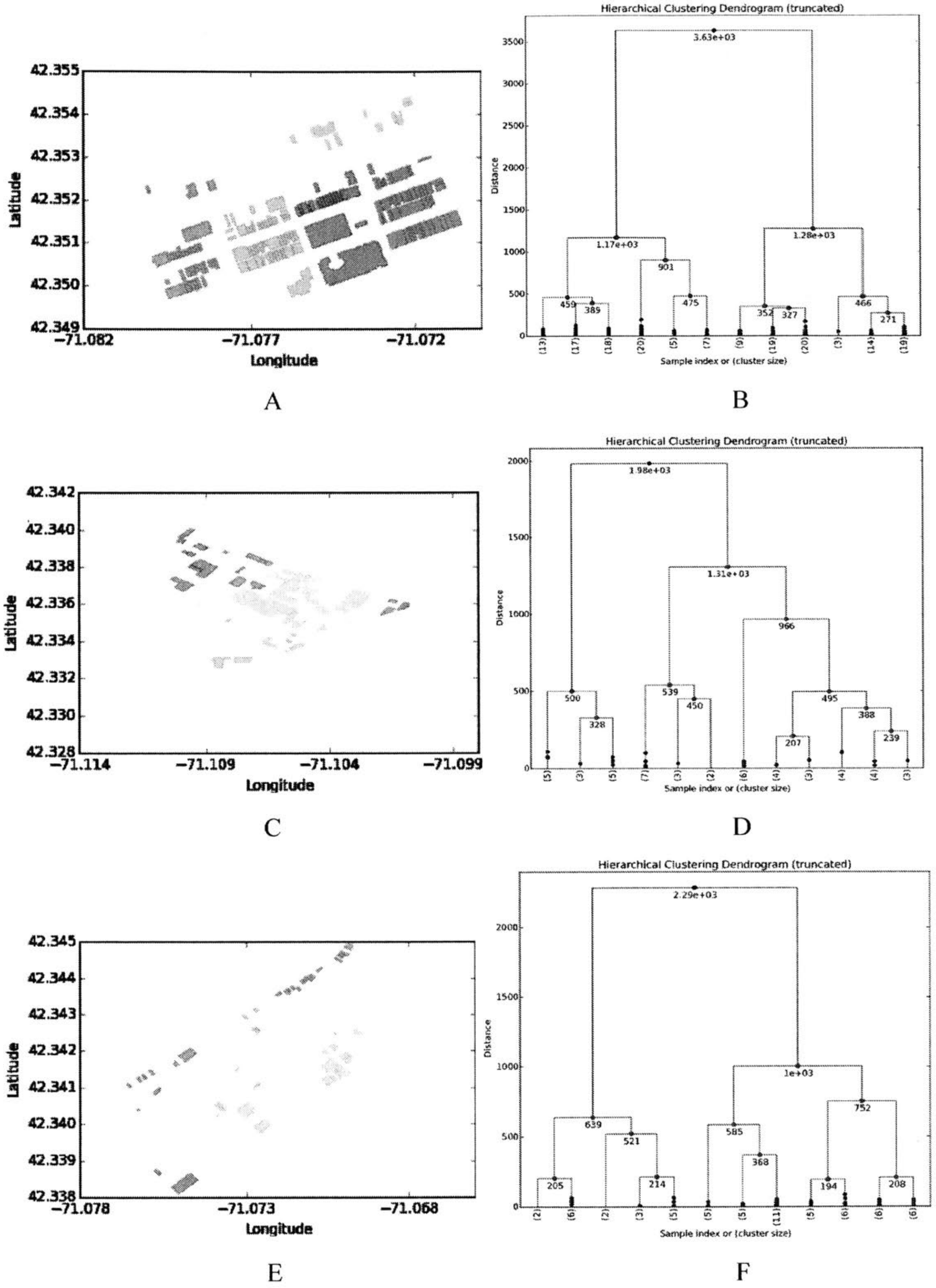


Fig. 4.7: Residential clusters (A, C, E) and dendrograms (B, D, F) identified in Back Bay, Harvard Medical School and South End



A

B

C

D

E

F

Fig. 4.8: Non-residential clusters (A, C, E) and dendrograms (B, D, F) identified in Back Bay, Harvard Medical School and South End

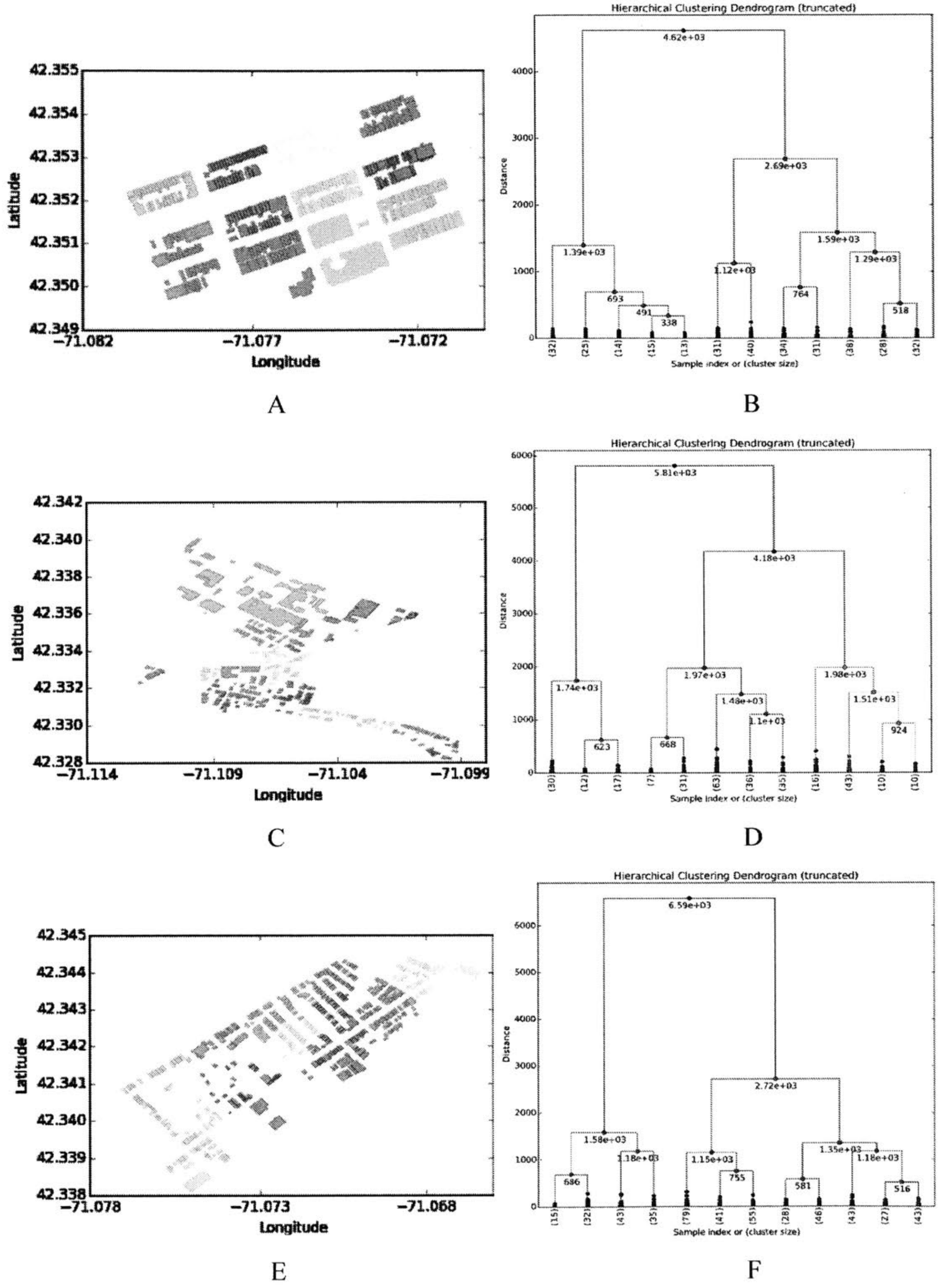


Fig. 4.9: Clusters (A, C, E) and dendrograms (B, D, F) identified for buildings attracting 'other' stays in Back Bay, Harvard Medical School and South End

Time dependent clusters

The analysis above assumes a hypothetical snapshot wherein all buildings are operational concurrently. Most realistic situations would involve a subset of buildings being non-operational and another subset having hours of operation incompatible with the duration of the trip (compatibility checks). To account for such occurrences, the set of feasible building matches is determined for each stay. HAC is then performed on the set of compatible buildings.

Figure 4.10 shows the set of buildings compatible with 'work' stays starting at 3pm and midnight. For a work stay starting at 3pm and going till midnight, the set of compatible buildings would be the intersection of the two sets of buildings.

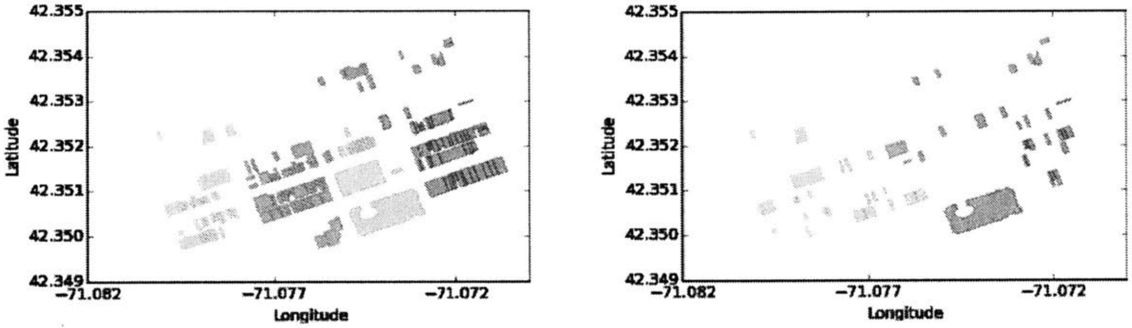


Fig.4.10: The set of buildings compatible with work stays in Back Bay (left) at 3pm and (right) at midnight

Rich Getting Richer- the cluster choice

As mentioned in the previous sections, it is assumed that the magnitude of attraction to a cluster is a function of the gross operational floor area of the buildings in the cluster. Also, it is calculated for each 10-min time interval.

The naïve distribution assumes a simple linear relationship between the gross building footprint and the fraction of trips attracted to a building.

The determined clusters allow us to account for more complex urban dynamics. Here, we outline a simple technique to achieve this-

1. Compute the total gross building area for all buildings in the tract (A)
2. Compute the relative cluster size, i.e. the fraction (f_i) of gross building area in each cluster (G_i) ($f_i = G_i/A$)

3. The average fraction (a) = 1/n, where n is the number of clusters. This is used to define the relative fractional area ($r_i = f_i/a$) for each cluster
4. The product of the square root of the relative fractional area (r_i) and the relative cluster size (f_i) is assumed to be the relative attraction of the cluster (a_i), $a_i = f_i * \sqrt{r_i}$
5. The relative attractions of clusters are normalized to obtain the cluster probabilities (p_i).

Sample calculations for a hypothetical case are shown in the tables 4.1 and 4.2.

Table 4.1: Hypothetical example of computation of probabilities for a stay to occur in 5 building clusters. One of the clusters is larger than any of the others by a factor of 2

i	f_i	r_i	$\sqrt{r_i}$	a_i	p_i
1	0.5	2.5	1.58	0.79	0.64
2	0.25	1.25	1.12	0.28	0.23
3	0.1	0.5	0.71	0.07	0.06
4	0.1	0.5	0.71	0.07	0.06
5	0.05	0.25	0.50	0.03	0.02

Average fraction, a = 1/5 = 0.2

Table 4.2: Example computation of probabilities for a stay to occur in 8 building clusters with one being significantly larger (5X) than any of the others

i	f_i	r_i	$\sqrt{r_i}$	a_i	p_i
1	0.5	4	2.00	1.00	0.72
2	0.05	0.4	0.63	0.03	0.02
3	0.05	0.4	0.63	0.03	0.02
4	0.1	0.8	0.89	0.09	0.06
5	0.05	0.4	0.63	0.03	0.02
6	0.05	0.4	0.63	0.03	0.02
7	0.1	0.8	0.89	0.09	0.06
8	0.1	0.8	0.89	0.09	0.06

Average fraction, a = 1/8 = 0.125

In the example in Table 4.1, the largest cluster accounts for half of the floor area in the tract. The next largest cluster is only about half its size. In the previous formulation, the probability of a stay to be attracted to this cluster would have been 50%. In this case, however, it is identified that the cluster accounts for a large fraction of the engagement opportunities and the probability of an agent to be attracted to the largest cluster is a bit more than 50%. The model estimates this to be about 64%. Since the second largest cluster, accounting for 25% of the gross floor area, also has 3 clusters smaller than itself, the probability of a trip being attracted to it is only slightly diminished at 23%. For the third largest cluster, accounting for 10% of the area, it is significantly diminished to 6%.

In the second example in Table 4.2, the largest cluster again accounts for 50% of the area. The next largest cluster, however, is significantly smaller- accounting for only 10% of the area. In this case, the largest cluster gains a lot more prominence, since the next option is significantly smaller, and is able to attract 72% of the stays to itself. This would lead to a high amount of activity in this cluster and this is often observed in popular activity centers in cities.
The second largest cluster is only able to attract 6% of the trips.

Stay allocation for agents with consecutive stays in the same tract

If an agent has consecutive stays in the same tract, the destination choice of the 'O' stay is likely to be influenced by the home and work locations. If both the stays are 'O' stays, the destination choice of the latter stay should be influenced by that of the preceding one.
To account for this, we propose a ranking based scheme which draws learning from TimeGeo. To decide the subsequent location, the distance of the current location from each cluster center is used to generate a proximity based rank (R) for each of the clusters. This rank is used to compute a rank based probability ($k_i = R^{-0.86}$). The rank-probability (k_i) is multiplied by the overall cluster probability (p_i) to determine the contextual attraction ($a^k{}_i$) for each cluster for the running sequence of trips. This contextual attraction is normalized to obtain the contextual probability ($p^k{}_i$).

Once a cluster has been chosen for a stay, the stay is probabilistically distributed to one of the compatible buildings in the cluster probabilistically based on the building capacity.

Table 4.3: Example of the process of determining the destination of a stay when the preceding stay occurred in the same tract

i	d_i	R	k_i	p_i	a^k_i	p^k_i
1	500	3	0.39	0.64	0.25	0.55
2	450	2	0.55	0.23	0.12	0.28
3	60	1	1.00	0.06	0.06	0.13
4	900	5	0.25	0.06	0.01	0.03
5	600	4	0.30	0.02	0.01	0.01

In Table 4.3, we analyze the clusters from table 4.1 for a subsequent trip starting from cluster 3 and having the proximity based ranking to different clusters denoted by R. Since such a stay will have a higher affinity to be attracted to a nearby cluster, the probability of the trip being attracted to cluster 3 itself rises from 6% to 13%. It also rises for the second closest cluster, Cluster 2, from 23% to 28% and falls for all the others. It should be noted that Cluster 1 still has the highest attraction for the trip, but the probability of attraction is reduced from 64% to 55%.

This concludes the process of distributing trips within tracts.

4.5Summary and Conclusion

Starting off with individual stays and building objects, this chapter has focused on incorporating numerous phenomena into the allocation of stays to buildings. It started off with a simple formulation where the attraction of stays to a building was only a function of the capacity of the building. This was followed by the development of a method to cluster buildings within a tract. This allowed us to identify regions within a tract and subsequently rank them in likely order of popularity. This was also intended to capture the likely perception of regions within the tract by individual agents.

Once a region was selected, a building was stochastically allocated to a stay based on the total capacity of the building. Therefore, different iterations of the model will yield different (though not statistically different) occupancy profile.

Another aspect that has been captured in the formulation is decision making with regards to short trips, identified by consecutive trips that occur within tracts. For such trips, cluster identification has been done as before but probability allocation further incorporated the location rank of a cluster from the previous stay location.

A limitation of this method is that it relies on boundaries as defined by the shape of the census tracts. Due to the boundary conditions of the census tracts, some regions within a tract might be part of or influenced by popular regions in adjoining tracts. Such occurrences are not captured by this model. Also, the method assumes that all stays need to be matched to building. Some stays do occur outdoors. Such stays are generally shorter and should be sampled from 'other' stays.

The model, however, provides a good first-attempt to distributing individual stays within a tract. In the next chapter, we analyze the results of matching stays to buildings and identify major occupancy typologies that might be used for mobility and energy interventions.

Chapter 5

Results and Analysis

5.1Results of Building-Stay matching

After the allocation of stays to individual buildings, we obtain the occupancy profiles for each building in the City of Boston. The progression of the analysis is summarized in Fig.5.1.

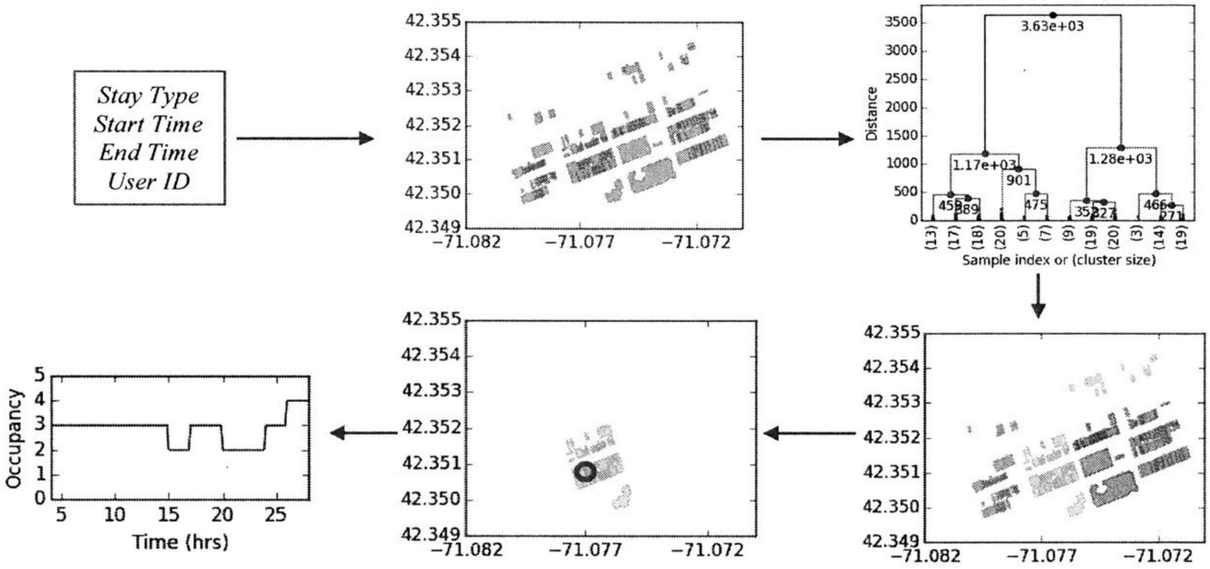


Fig.5.1: Outline of the process of obtaining occupancy profiles for buildings starting off from individual stay information. Clockwise from top left- start off with the information for an individual stay, determine the set of feasible building matches for the stay within tract in which the stay occurs, using hierarchical agglomerative clustering determine the regions and their relative attractions for the stay within the tract, choose a particular cluster, stochastically chose a building within the tract and repeat the process for all stays to obtain occupancy profiles for buildings

Having the occupancy profiles for buildings provides us an opportunity to analyze buildings at an aggregate level and identify different typologies of building occupancy. It also opens up the possibility of analyzing occupancy trends in regions rather than in individual buildings. Studying individual buildings would be useful since buildings with similar occupancy profiles can be potential candidates for similar interventions. These interventions can range from energy retrofits to smarter lighting and mobility decisions to name a few. Similarly, at an aggregate level, the ability to better understand utilization can be used for urban design, setting up new facilities and energy forecasting and storage.

In the following sections, we provide analyses at both these levels.

5.2Analyzing Occupancy Profiles of Individual Buildings

As shown in Fig.5.1, the occupancy profiles for individual buildings have been estimated for a 24-hour period from 4am on a given day to 4am on the day to follow. This has been done since there are several activities that stretch beyond midnight but conclude before 4am.

The occupancy profiles are available as 144-dimensional vectors with each element representing the occupancy of the building in one 10-minute time slot of the day. The objective of this analysis is to identify buildings with similar utilization patterns. A small single-family home and a large residential complex can have the same pattern of utilization with people leaving in the morning and getting back in the evening. Hence, it is not the magnitude of the curve that is critical to the utilization pattern of a building but rather the shape that governs the use case. Therefore, in order to compare buildings, we normalize the occupancy profiles by dividing each element by the maximum observed occupancy for the building. This reduces the value in each slot to a number between 0 and 1.

The difference between two buildings is quantified by the Euclidean distance between their normalized occupancy profiles. The task at hand is to identify the key utilization typologies for 75.5k building that were operational on the day of the analysis, simulated as a Monday.

To accomplish this, we choose to use k-means clustering. The first step is to identify an appropriate k-value. This has been done using the elbow method. The plots for this method are shown in Fig.5.2. The fraction of explained variance stagnates at about 80% and choosing to have 5 clusters is able to explain about 65% of the variance. We therefore proceed with 5 clusters for the analysis. Since there are more than 75k buildings, it is not suitable to plot the normalized occupancy curves for each of them on a plot. Instead, we focus on the centroids of the clusters which provide us a sense of the shape of the curves that fall into that cluster. These are shown in Fig.5.3.

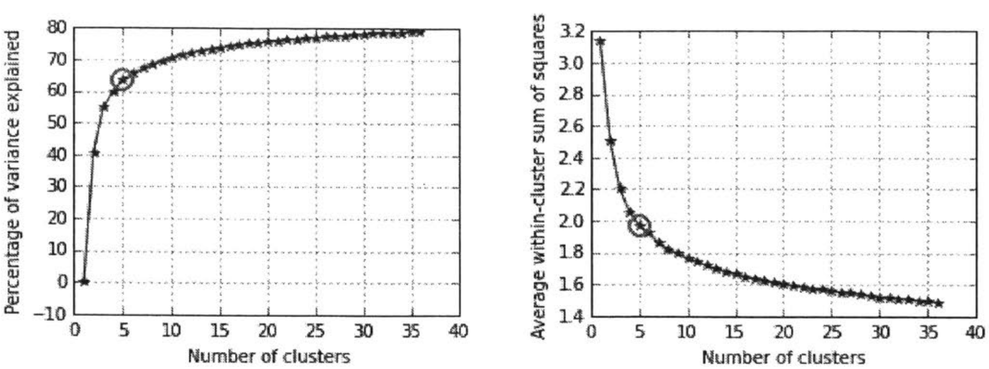


Fig.5.2: (left) The percentage of variance explained by different cluster counts for buildings, (right) The trend of the within cluster sum of squares for different number of clusters

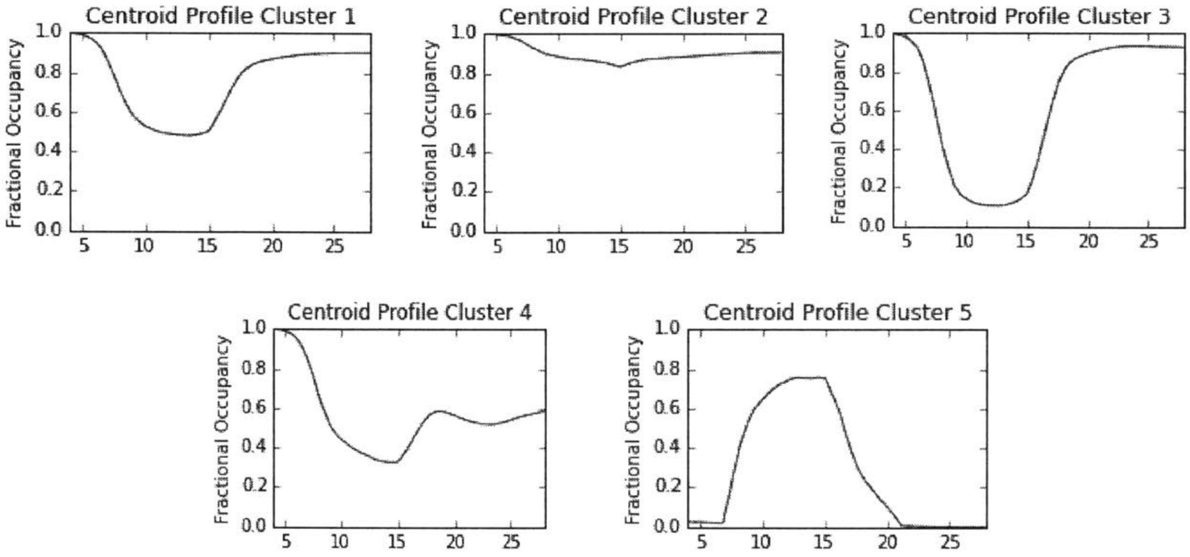


Fig.5.3: Normalized Occupancy Profile for the centroids of the five clusters

While the centroid profiles provide meaningful insights into the basic occupancy trends for the buildings in a cluster, before making inferences regarding building use, we should account for the categories of buildings that constitute the clusters. This is shown in Fig.5.4. Clusters 1 through 4 are exclusively residential clusters while Cluster 5 incorporates all the non-residential buildings of the city.

Now we proceed to the interpretation of building archetypes revealed by each cluster.

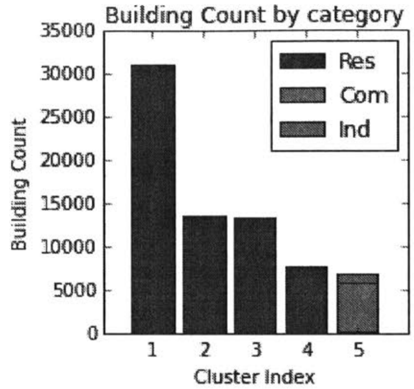


Fig5.4: Distribution of building count and categories in the five clusters

Cluster 1 comprises of residential buildings with full occupancy in the morning which reduces to about 50% in the afternoon and early evening and springs back to capacity later in the day. These are likely to be households with a fraction of working members or buildings that house a mixture of working and non-working households.

Cluster 2 also comprises of residential buildings. The characteristic building in this cluster appears to have a stable occupancy profile throughout the day. These buildings might be residences of people working from home, unemployed people, people taking the day off or perhaps old-age homes.

Cluster 3, also completely residential, is characterized by a sharp dip in occupancy in the morning period (6am-9am) and a more gradual recovery from 3pm-7pm. The occupancy levels approach near-zero in the mid-day and recovery almost completely at night. These buildings are likely to be home to households with all workers, young professionals or families with children at school during the day.

Cluster 4 comprises of fewer building, is again residential, and has a unique fluctuating occupancy curve. This cluster observes the usual drop in occupancy levels in the morning period till about 50% between 8-9am. A more gradual drop is observed till 3pm when the occupancy reaches its lowest value of about 30%. This is followed by a brief period of recovery expected from the trickling back of morning workers till about 6pm with the occupancy replenished to about 60%. There is a slight dip observed after this which recovers back to the 60% mark at the end of the day. This profile most likely belongs to residences of a mixture of morning and evening workers, individuals spending the day at a friend's place or people leaving the city for business/personal reasons.

Cluster 5 comprises of all the commercial and industrial buildings of the city. As is expected of commercial establishments, the occupancy picks up early in the morning, peaks in the afternoon and early evening and tapers down at night.

As mentioned, segregating buildings into these five classes enables us to capture a large fraction of the variance inherent in the occupancy profiles of the buildings. The ability to forecast occupancy can also help us to predict energy loads and find smarter ways to distribute storage units and manage capacity. These provisions often require retrofits. The buildings lying in the same cluster can potentially be candidates for similar retrofits and pricing plans to promote predictability and therefore prevent shocks to the system.

Figure 5.5 shows the geographical distribution of buildings in the five clusters

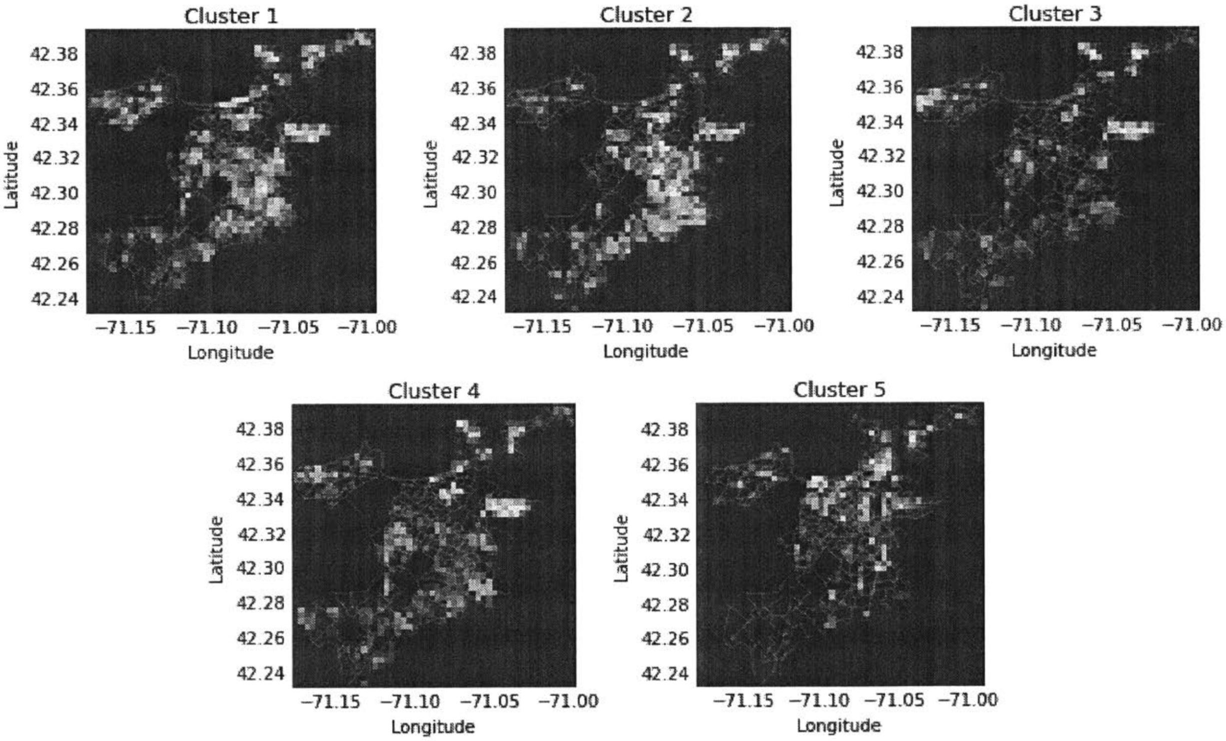


Fig.5.5: The geographical distribution of buildings belonging to the five occupancy clusters

Some of the differences in the geographical distributions of the buildings are interesting to note. As expected, Cluster 1, the largest and most characteristic residential cluster, is the most widespread. Cluster 2 is also quite widespread but has lower prominence in the west in Allston and Jamaica Plain. Cluster 3, on the other hand appears most prominently in South Boston, Allston and Jamaica Plain. Cluster 5, as expected, is almost a complement to the residential clusters. Thus, analyzing the geographical distribution of the different building clusters can shed light on the utilization of different regions of the city. This calls for analysis at a more aggregate level than individual buildings and paves way for the tract-level analysis explored in the next section.

5.3Generating and Analysis of Occupancy Profiles at the Tract Level

While analysis at the level of individual buildings was useful to identify similar utilization archetypes across the city, regional utilization can be studied by aggregating the building profiles from an area and determining a single profile to represent the entire region. We perform this analysis at the level of census tracts. We first explain the method used to generate the profile for a tract and then determine tracts with similar occupancy patterns. We conclude by making some observations from the results.

A census tract is an area comprising of several buildings belonging to different classes. Each of these buildings has its own normalized occupancy for each time slot of the day. For one randomly chosen tract the distribution of the occupancy for each time slots for all buildings put together is shown in Fig.5.6.

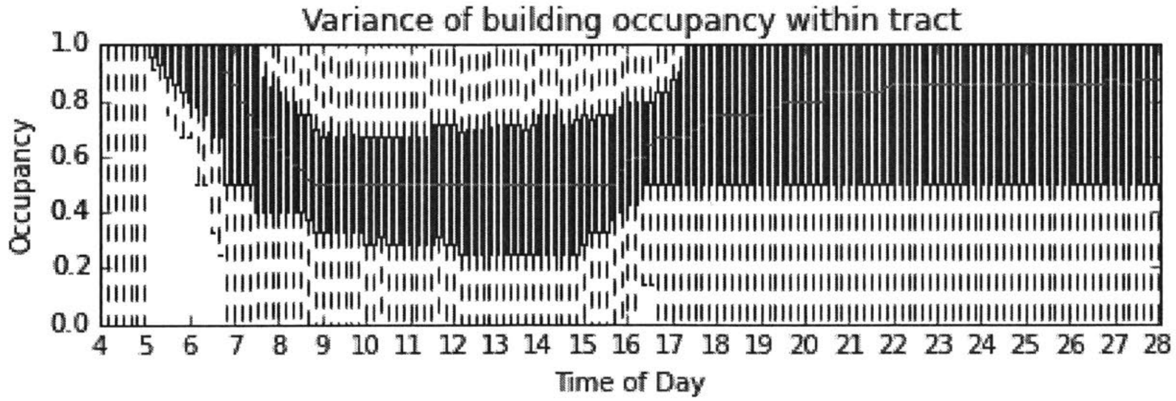


Fig.5.6: Box plot for the distribution of occupancy profiles for all buildings in a tract

For the purpose of the analysis, we trace the median value of the occupancy (shown in red in Fig.5.6) for each time slot and use this as the single curve that represents the occupancy for the tract. More complex vectors might also be used to characterize a tract such as those also accounting for variance in occupancy. For this analysis, we have stuck to a simple function constructed from the median value.

There are 173 census tracts in the City of Boston and we compute characteristic occupancy curves for each of these. It is noteworthy that the curve generated by tracing the median occupancy value need not resemble any single building in a tract. One instance when this can happen is when there are multiple competing sets of profiles in a tract.

After obtaining occupancy profiles for individual tracts, we proceed to the identification of tracts with similar occupancy plots. We use an approach similar to that employed for individual buildings- k-means clustering with the distance between curves computed as the Euclidean distance between each of their 144-dimensions.

Figure 5.7 shows the results for the elbow analysis for the clustering for tracts. It is observed that a large amount of variance can be explained by having 2 clusters and three clusters explain about 90% of the variance beyond which the increments are fairly low. Hence, we choose 3 clusters for analysis.

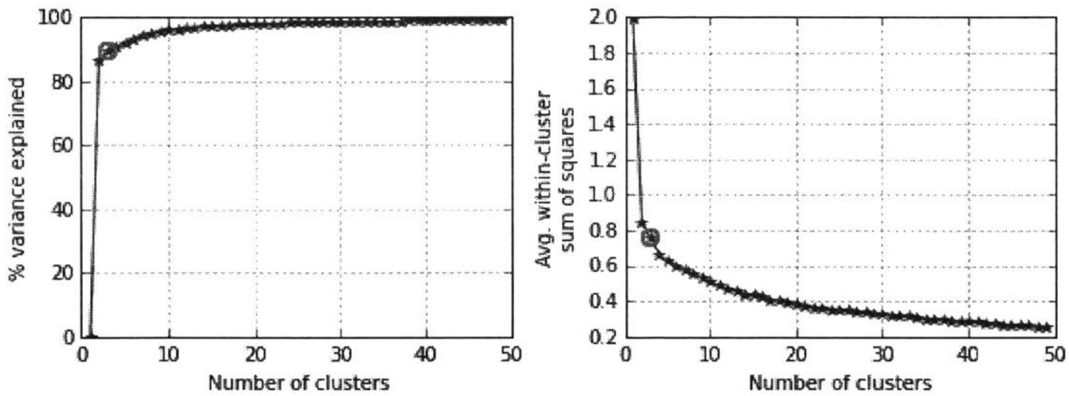


Fig.5.7: (left) The percentage of variance explained by different cluster counts for tracts, (right) The trend of the within cluster sum of squares for different number of clusters

The results for k-means clustering for tracts are shown in the left panel in figure 5.8. The profiles for the cluster centroids are shown in the right panel.

From the profiles of the centroids in figure 5.8, it can be seen that the profile for Cluster 1 (blue) resembles that of a typical residential building. Cluster 2 (green) is similar to Cluster 1 but is slightly more regularized with a lower dip during the day accompanied by a smaller jump at night. The profile for the centroid of Cluster 3 (red) resembles that of commercial establishments.

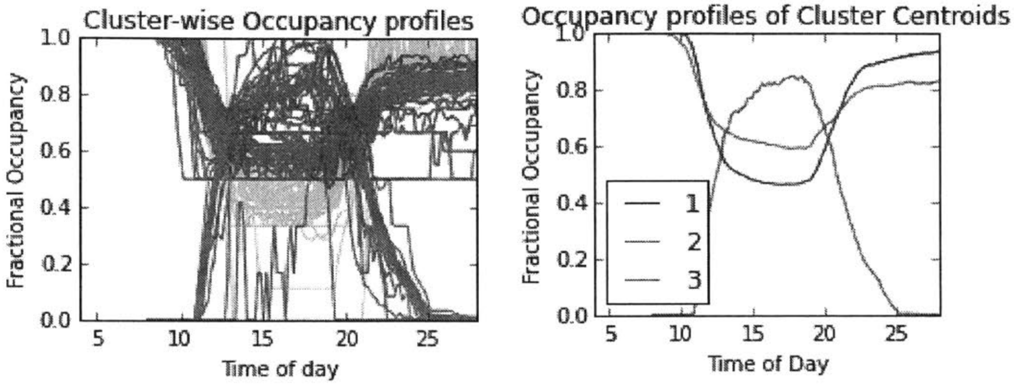


Fig.5.8: (left) Individual characteristic profiles for each tract colored by the cluster that they belong to, (right) The profile of the centroids of the three clusters

In order to explore the characteristics of the clusters further, we need to explore the tract count and the composition of clusters in figure 5.9. As expected, a majority of the tracts are in the cluster the profile of whose centroid resembles that of a residential building. In the right panel of figure 5.9, we explore the composition of each cluster, quantified by the sum of the capacities of the buildings belonging to each of the three classes- residential, commercial and industrial.

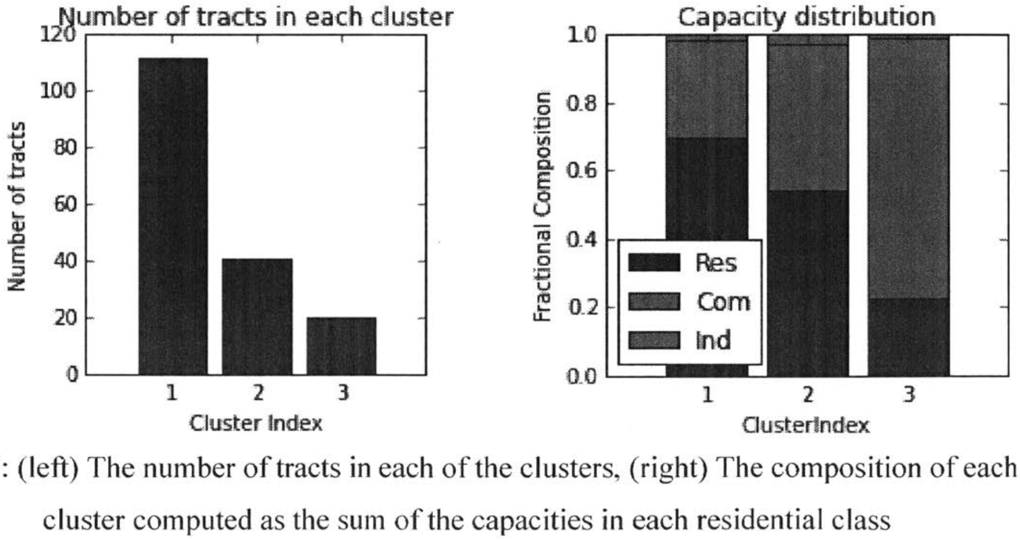


Fig.5.9: (left) The number of tracts in each of the clusters, (right) The composition of each cluster computed as the sum of the capacities in each residential class

In figure 5.9, it is interesting to note that Clusters 1 and 3 are dominated by a single building use-residential for Cluster 1 and commercial for Cluster 3; whereas a slightly more balanced split between residential and commercial capacities leads to a more regularized occupancy profile as revealed in figure 5.8. Since a basic minimum mixing between building classes is expected in any tract (grocery stores, restaurants, laundry services and other basic facilities are almost always present even in predominantly residential settings), we would expect to achieve further flattening of the occupancy curve for a cluster with a slightly higher proportion of commercial capacity than that observed for Cluster 2. This is an important result for planning urban spaces since it makes a case for mixed use planning in cities.

The distribution of the tracts in each of the three clusters is shown in figure 5.10.

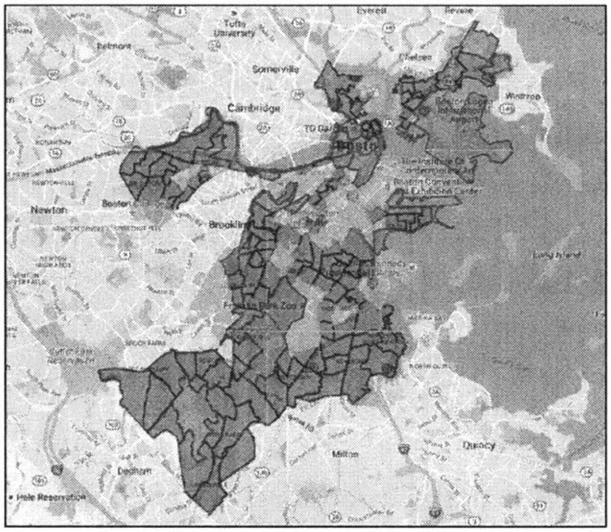


Fig.5.10: The tracts belonging to the 3 clusters- Cluster 1 in blue, Cluster 2 in green and Cluster 3 in red

In figure 5.10, it is seen that Cluster 3, with the occupancy curve of its centroid resembling that of a commercial establishment, covers most of the commercial areas (downtown, airport etc.) and the recreational areas (the Esplanade, river banks and parks). Back Bay is included in Cluster 2 and shows a mixed profile, as would be expected of an area with heterogeneous land use. Residential areas are also recognized quite well (highlighted in blue).

5.4Summary and Conclusion

In this chapter, we started off with the occupancy profiles for individual buildings and wanted to identify recurring patterns and typologies in the city. We organized the analysis at two levels- individual buildings and census tracts. Analyzing the occupancy profiles yielded key typologies that could be used to characterize buildings. Identifying five key typologies provided an explanation of majority of the variance observed in building utilization and can for the basis for further exploration of use based policies, retrofits and incentives.

Analysis of individual buildings, while useful for identifying typologies throughout the city, provides very little information about how smaller regions of the city are utilized. To talk about regional utilization, we aggregated the building within individual regions (chosen to be census tracts) and generated representative occupancy profiles for each region. This provided insights into the overall utilization of individual tracts. We then proceeded to identifying tracts with similar utilization patterns to be able to study how the distribution of building types influences occupancy loads in a region. The results, as expected, reveal that having greater heterogeneity leads to more regularized occupancy loads in a region.

Occupancy profiles for buildings can in themselves provide useful information to individuals and businesses. These have been explored in popular time as displayed on Google Maps. Here, we have shown that aggregated analysis of these profiles can serve to be a useful tool for planning as well.

Chapter 6

Conclusion

This study started off with the goal of estimating building level occupancy using ubiquitous data sources. In our attempt to model mobility using Call Detail Records, we have shown that the resolution of the results derived from simulation is limited by the accuracy of the input data. Since cell phone traces usually have an uncertainty of a few hundred meters, they are insufficient to model mobility at the resolution of individual buildings.

We have used TimeGeo, a state of the art mobility model, and improved it to make it scalable and portable. The scalability has been established by comparing our results with those from elaborate surveys which are the currently accepted standard in industry. To establish the portability of the model, we evaluate its performance on a dataset with a higher level of resolution, in a different part of the world, in a controlled tracking experiment in Denmark. Conformity with the observed ground truth establishes the portability of the model.

We then proceed to model the location of users at a higher resolution. To achieve this, we rely on other widely available data such as building shapefiles and information about Points of Interest. We create a model that combines the different hyperlocal data sources and distributes passenger trips to individual buildings in a behavioral manner while accounting for urban dynamics.

While occupancy profiles by themselves are useful resources for individuals and businesses, we demonstrate that further learning can be derived by analyzing building occupancy profiles as an aggregate. This has enabled us to identify typologies that have the potential to be used to assist policy decisions and to design retrofits. We have taken the analysis one step further and outlined a way to generate occupancy profiles for regions. We have used these profiles to study the impact of different development patterns on the utilization of a region. This has provided a strong case for heterogeneous urban development.

This work is only a first attempt towards creating data-driven models for mobility at a high resolution at an urban scale. The model is, by construction, individualistic. All decisions are made by an individual and do not directly influence the decisions made by others. There are, however, certain decisions that are made by entities other than the individual. For example, the decision to go out on a weekend is often made in groups, rather than individually; mobility decisions might be made by the drivers in a household, rather than all individuals; the decision to change residence location might be taken by the household as a unit. These decisions and entities have not been incorporated in the model.

The matching between stays and buildings has been performed in the absence of user feedback for individual establishments. Reviews for establishments, especially commercial establishments, are widely available can be used to augment or diminish the degree of attraction of places of interest for individuals.

The model currently considers all locations other than home and work to be identical. Analyzing trip chains can be useful to add further detail to individual profiles. In the realm of analysis, even though the occupancy profiles look reasonable, there is scope for further validation by collecting primary data about actual occupancy loads. Also, only the absolute value of the occupancy has been considered for analysis. There is scope for buildings to have different motifs for utilization based on the count of visitors (people moving in and out frequently rather the static occupants of the building) to a building. Similarly, the information about a building can be enriched by analyzing the type of other activities that workers/residents/visitors engage in when they are not in the building. Adding these layers of information and analysis can provide further insights into utilization patterns and archetypes.

While all of these steps will definitely help in improving results, the model, as it stands, has been able to demonstrate that high-resolution mobility can indeed be modeled from currently available ubiquitous data sources. Even though it only just scratches the surface, the study has shown the immense potential of leveraging high-resolution mobility traces to make informed decisions and improve urban spaces.

References

[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).

[3] Nagel K, Beckman RJ, Barrett CL (1999) Transims for urban planning in 6th International Conference on Computers in Urban Planning and Urban Management, Venice, Italy. (Citeseer).

[4] Ben-Akiva M, Bierlaire M (1999) Discrete choice methods and their applications to short term travel decisions in Handbook of transportation science. (Springer), pp. 5–33.

[5] Balmer M et al. (2008) Agent-based simulation of travel demand: Structure and computational performance of MATSim-T. (ETH, Eidgenössische Technische Hochschule Zürich, IVT Institut für Verkehrsplanung und Transportsysteme).

[6] Arentze T, Timmermans H (2000) Albatross: a learning based transportation oriented simulation system. (Eirass Eindhoven).

[7] Bowman JL, Ben-Akiva ME (2001) Activity-based disaggregate travel demand model system with activity schedules. Transportation Research Part A: Policy and Practice 35(1):1–28.

[8] Salvini P, Miller E J, 2005, ``ILUTE: an operational prototype of a comprehensive microsimulation model of urban systems" Networks and Spatial Economics 5 217-234

[9] Danalet A, Tinguely L, Cochon de Lapparent MM, Bierlaire M (2015) Location choice with longitudinal WiFi data, (Lausanne, Switzerland), Technical report.

[10] Zilske M, Nagel K (2014) Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. Procedia Computer Science 32:802–807.

[11] Zilske M, Nagel K (2015) A simulation-based approach for constructing all-day travel chains from mobile phone data. Procedia Computer Science 52:468 – 475. The 6th International Conference on Ambient Systems, Networks and Technologies (ANT-2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015).

[12] Zheng Y, Capra L, Wolfson O, Yang H (2014) Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 5(3):38.

[13] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González. The TimeGeo modeling framework for urban motility without travel surveys. Proceedings of the National Academy of Sciences, page 201524261, 2016.

[14] Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. arXiv preprint arXiv:1502.03406.

[15] Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782.

[16] Song C, Koren T, Wang P, Barabási AL (2010) Modelling the scaling properties of human mobility. Nature Physics 6(10):818–823.

[17] Perkins TA et al. (2014) Theory and data for simulating fine-scale human movement in an urban environment. Journal of The Royal Society Interface 11(99):20140642.

[18] Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. Science 327(5968):1018–1021.

[19] Hasan S, Schneider CM, Ukkusuri SV, González MC (2013) Spatiotemporal patterns of urban human mobility. Journal of Statistical Physics 151(1-2):304–318.

[20] Toole JL, Herrera-Yaqüe C, Schneider CM, González MC (2015) Coupling human mobility and social ties. Journal of The Royal Society Interface 12(105):20141128.

[21] Schneider CM, Belik V, Couronné T, Smoreda Z, González MC (2013) Unravelling daily human mobility motifs. Journal of The Royal Society Interface 10(84):20130246.

[22] Kölbl R, Helbing D (2003) Energy laws in human travel behaviour. New Journal of Physics 5(1):48.

[23] Balcan D et al. (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences 106(51):21484–21489.

[24] Viswanathan G et al. (1996) Lévy flight search patterns of wandering albatrosses. Nature 381(6581):413–415.

[25] Jiang S et al. (2013) A review of urban computing for mobile phone traces: current methods, challenges and opportunities in Proceedings of the 2nd ACM SIGKDD International Work- shop on Urban Computing. (ACM), p. 2.

[26] Toole JL et al. (2015) The path most traveled: Travel demand estimation using big data resources. Transportation Research Part C: Emerging Technologies.

[27] Alexander L, Jiang S, Murga M, González MC (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies.

[28] United Nations, Department of Economic and Social Affairs, Population Division (2014). World Urbanization Prospects: The 2014 Revision, Highlights (ST/ESA/SER.A/352).

[29] J. Laustsen, Energy Efficiency Requirements in Building Codes, Energy Efficiency Policies for New Buildings, 2008, OECD/IEA International Energy Agency

[30] S. D'Oca, T. Hong, Occupancy schedules learning process through a data mining framework Journal of Energy and Buildings, 88 (Feb 2015), pp. 395–408

[31] C. Duartea, K.Van Den Wymelenberga, C. Riegerb, Revealing occupancy patterns in an office building through the use of occupancy sensor data, Energy and Buildings 67 (2013) 587–595

[32] W. Chang., T. Hong, Statistical Analysis and Modeling of Occupancy Patterns in Open-Plan Offices using Measured Lighting-Switch Data. Building Simulation 6 (2013) 23-32

[33] T. Hong T, H. Lin, Occupant Behavior: Impacts on Energy Use of Private Offices. ASim 2012 - 1st Asia conference of International Building Performance Simulation Association, Shanghai, China (2013)

[34] D. Wang, C. C. Federspiel, and F. Rubinstein, Modeling occupancy in single person offices, Energy and Buildings 37 (2005) 121–126

[35] V. Tabak, B. de Vries, Methods for the prediction of intermediate activities by office occupants, Building and Environment 45 (2010) 1366–1372

[36] K. Sun, D. Yana, T. Hong, S. Guo, Stochastic modeling of overtime occupancy and its application in building energy simulation and calibration, Building and Environment 79 (2014) 1-12

[37] U. DOE. Building Energy Databook, 2010.

[38] J. Howard, W. Hoff, Forecasting building occupancy using sensor network data, Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, ACM (2013)

[39] Cerezo C, Sokol J, Reinhart C, AlMumin A: Three methods for characterizing building archetypes in urban energy simulation: A case study in Kuwait city. In proceedings of Building Simulation 2015: Hyderabad, India; 2015.

[40] Candia J et al. (2008) Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical 41(22):224015.

[41] Song C, Koren T, Wang P, Barabási AL (2010) Modelling the scaling properties of human mobility. Nature Physics 6(10):818–823.

[42] Wang P, Hunter T, Bayen AM, Schechtner K, González MC (2012) Understanding road usage patterns in urban areas. Scientific reports 2.

[43] Hariharan R, Toyama K (2004) Project lachesis: parsing and modeling location histories in Geographic Information Science. (Springer), pp. 106–124.

[44] Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with gps history data in Proceedings of the 19th international conference on World wide web. (ACM), pp. 1029–1038.

[45] Zheng Y, Xie X (2011) Learning travel recommendations from user-generated gps traces. ACM Transactions on Intelligent Systems and Technology (TIST) 2(1):2.

[46] Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with gps history data in Proceedings of the 19th international conference on World wide web. (ACM), pp. 1029–1038.

[47] Levinson DM, Kumar A (1994) The rational locator: why travel times have remained stable. Journal of the american planning association 60(3):319–332.

[48] Schafer A (2000) Regularities in travel demand: an international perspective. Journal of transportation and statistics 3(3):1–31.

[49] CTPS (2013) Methodology and assumptions of central transportation planning staff regional travel demand modeling.

[50] 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).

[51] United States Department of Labor Bureau of Labor Statistics (2010) American time use survey (ATUS), 2010 (http://www.bls.gov/tus/datafiles_2010.htm).

[52] Bishop CM (2006) Pattern Recognition and Machine Learning. (Springer-Verlag New York, Inc., Secaucus, NJ, USA).

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

[54] U.S. Department of Transportation Federal Highway Administration (2011) 2009 National Household Travel Survey (http://nhts.ornl.gov/download.shtml).

[55] Massachusetts Department of Transportation (2012) 2010/2011 Massachusetts travel survey. [Online; accessed 17-March-2016].

[56] Newman PG, Kenworthy JR (1989) Cities and automobile dependence: An international sourcebook.

[57] CTPS (2008) Central transportation planning staff regional travel demand modeling methodology and assumptions.

[58] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. 2014. Measuring large-scale social networks with high resolution. PloS one 9, 4 (2014), e95978.

[59] OpenStreetMap contributors. (2015) Planet dump [Data file from April 19, 2016]. Retrieved from http://planet.openstreetmap.org

[60] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 1082–1090.

[61] Trinh Minh Tri Do, Olivier Dousse, Markus Miettinen, and Daniel Gatica-Perez. 2015. A probabilistic kernel method for human mobility prediction with smartphones. Pervasive and Mobile Computing 20 (2015), 13–28.

[62] Trinh Minh Tri Do and Daniel Gatica-Perez. 2012. Contextual conditional models for smartphone-based human mobility prediction. In Proceedings of the 2012 ACM conference on ubiquitous computing. ACM, 163–172.

[63] Nathan Eagle and Alex Pentland. 2006. Reality mining: sensing complex social systems. Personal and ubiquitous computing 10, 4 (2006), 255–268.

[64] Halgurt Bapierre, Georg Groh, and Stefan Theiner. 2011. A variable order markov model approach for mobility prediction. Pervasive Computing (2011), 8–16.

[65] Huiji Gao, Jiliang Tang, and Huan Liu. 2012. Mobile location prediction in spatio-temporal context. In Nokia mobile data challenge workshop, Vol. 41. 44.

[66] Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. 2013. Approaching the limit of predictability in human mobility. Scientific reports 3 (2013).

[67] Adam Sadilek, Henry Kautz, and Jeffrey P Bigham. 2012. Finding your friends and following them to where you are. In Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 723–732.

[68] Adam Sadilek and John Krumm. 2012. Far Out: Predicting Long-Term Human Mobility.. In AAAI.

[69] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. 2011. NextPlace: a spatio-temporal prediction framework for pervasive systems. In Pervasive computing. Springer, 152–169.

[70] Libo Song, David Kotz, Ravi Jain, and Xiaoning He. 2006. Evaluating next-cell predictors with extensive Wi-Fi mobility data. Mobile Computing, IEEE Transactions on 5, 12 (2006), 1633–1649.

[71] Yu Zheng, Xing Xie, and Wei-Ying Ma. 2010. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. IEEE Data Eng. Bull. 33, 2 (2010), 32–39.

[72] Buildings Boston MA 2012, Boston (Mass.). Dept. of Innovation and Technology, Jan 19, 2012

[73] © OpenStreetMap contributors, CC by SA

[74] The Engineering Toolbox, (http://www.engineeringtoolbox.com/number-persons-buildings-d_118.html), accessed Nov 26, 2016

[75] Imagery © 2016 Google, Map data © 2016 Google

[76] Ward, Joe H. (1963). "Hierarchical Grouping to Optimize an Objective Function". Journal of the American Statistical Association. 58 (301): 236–244. doi:10.2307/2282967. JSTOR 2282967. MR 0148188.