Data-Driven Energy and Population Estimation for Real-Time City-Wide Energy Footprinting

Peter Wei
wei.peter@columbia.edu
Columbia University

Xiaofan Jiang
jiang@ee.columbia.edu
Columbia University

ABSTRACT

Energy footprinting has the potential to raise awareness of energy consumption and lead to energy saving behavior. However, current methods are largely restricted to single buildings; these methods require energy and occupancy monitoring sensor deployments, which can be expensive and difficult to deploy at scale. Further, current methods for estimating energy consumption and population cannot provide fine enough temporal or spatial granularity for a reasonable personal energy footprint estimate. In this work, we present *CityEnergy*, a data-driven system for city-wide estimation of personal energy footprints. *CityEnergy* takes advantage of existing sensing infrastructure and data sources in urban cities to provide energy and population estimates at the building level, even in built environments that do not have existing or accessible energy or population data.

CCS CONCEPTS

• **Hardware** → **Power and energy**; • **Computer systems organization** → *Sensor networks*; *Real-time systems.*

KEYWORDS

Energy Footprinting, Real-Time System, Population Estimation

ACM Reference Format:
Peter Wei and Xiaofan Jiang. 2019. Data-Driven Energy and Population Estimation for Real-Time City-Wide Energy Footprinting. In *The 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '19), November 13–14, 2019, New York, NY, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3360322.3360847

1 INTRODUCTION

In urban cities such as New York City, buildings are responsible for up to 75% of total greenhouse gas emissions, and up to 94% of total energy consumption [28]. A significant portion of the energy consumed is to directly service humans such as in retail, commercial, and residential buildings. In addition to buildings, transportation is also responsible for large amounts of energy consumption. As sustainability increasingly becomes an important factor in modern society, energy consumption in the built environment is one area where reduction can have a major impact.

BuildSys '19, November 13–14, 2019, New York, NY, USA

ACM ISBN 978-1-4503-7005-9/19/11…$15.00
https://doi.org/10.1145/3360322.3360847

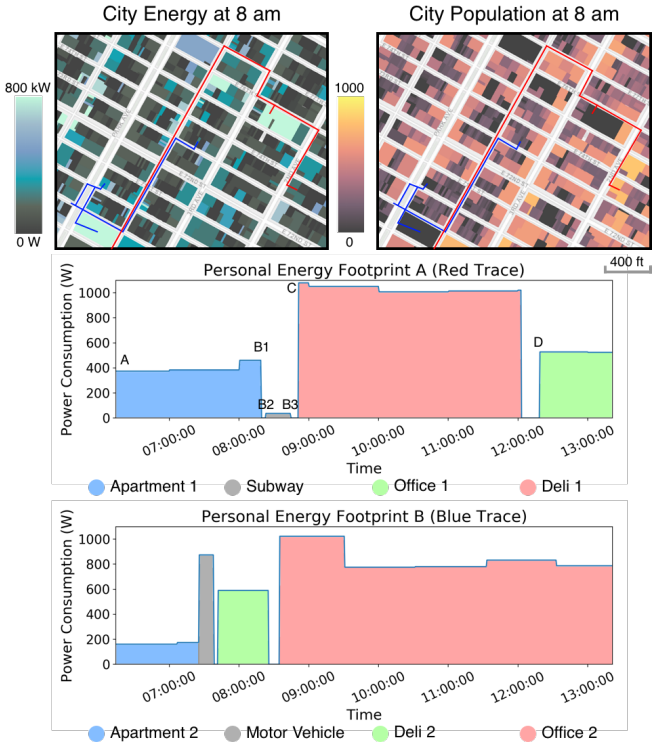


Figure 1: Top: Energy and population are estimated at the building level; personal energy footprints are provided to each person depending on their location traces (red and blue traces). Bottom: Two real city-wide energy footprints. Labels below denote energy source.

In [6], the authors show that certain feedback mechanisms given to occupants raises awareness and can lead to energy saving behavior. One such mechanism is by notifying the occupant in real-time of their numerical energy consumption, or energy footprint. However, many areas of the built environment do not have the capabilities to measure personal energy consumption, much less notify people of their personal energy responsibility. Although companies such as Nest and recent research studies have begun to address this challenge through *energy footprinting* in single residential and commercial buildings, there are still many difficulties to scaling these solutions to the city-scale. Most notably, there is a lack of energy and population data with high temporal and spatial granularity. Without sensors to measure this data, or models to estimate this data, energy footprinting is not possible.

In this work, we present *CityEnergy*, a scalable, real-time system for computing personal energy footprint estimates. As shown in

Figure 1, *CityEnergy* estimates the energy consumption and population for each building in real-time; these estimates are used to calculate the per capita energy footprint at the building level. *CityEnergy* is able to provide personal energy footprints to people who provide their location data through a mobile application.

An example city-wide personal energy footprint generated by *CityEnergy* is shown in Figure 1. In this example, a *CityEnergy* user, Stephen, (A) begins the day in a residential building. *CityEnergy* uses a detailed model of the particular built environment Stephen is in, and computes the energy and occupancy level to estimate the individual energy that Stephen is responsible for.

Stephen's commute consists of (B1) walking to the subway station, (B2) riding the subway downtown, and (B3) walking from the subway station to the office building; for each of the modes of transportation, *CityEnergy* associates the relevant energy consumption to Stephen's personal energy footprint. After arriving at the office, (C) Stephen works until the lunch break. During this period, *CityEnergy* may interface with the local energy footprinting system to determine Stephen's personal footprint, or rely on energy and occupancy models to estimate Stephen's footprint. Finally, Stephen leaves the office and (D) walks to the local deli, where he spends an hour to eat lunch.

In this work, we present the following contributions:

- We create an energy and occupancy *digital twin* of the city, with a focus on major aspects of the built environment including buildings and transportation.
- We present the design and implementation of *CityEnergy*, a **city-scale** energy footprinting system that utilizes the city's digital twin to provide **real-time energy footprints** with a focus on 100% coverage.
- We deploy *CityEnergy* in New York City, utilize local data sources to develop energy and population models, and collect and evaluate the accuracy of real world personal energy footprint data.
- We have developed a number of tools and applications such as mobile and web applications to provide citizens with insights into their everyday energy consumption, and city planners with important information at the city-scale.

CityEnergy is a tool for *estimating* an individual's energy footprint at any location in the city, at any time. Due to the data sources available, the energy footprint estimate may not be accurate for any one person. However, we believe that *CityEnergy* is an important step towards realizing city-wide energy footprinting.

2 RELATED WORKS

There are a number of recent works addressing topics such as building energy consumption estimation, population estimation, transportation detection, and energy footprinting.

Predicting energy consumption of a building normally constitutes one of three approaches: building level regression, software modeling, or city-wide energy estimation. For building level regression, historical energy consumption data is typically collected with fine temporal granularity (minute to hourly frequency). Once the data is collected, a regression model [12, 14, 38, 40] or neural network [12, 15, 38] is trained on the data. After training, the model is validated on a different dataset from the same building. These models tend to produce low error rates (< 10 MAE), but require a large amount of data. At a city-scale level, these methods are impractical without data already collected from a major entity, such as a government project.

Another method for predicting energy consumption in a building is through software modeling. EnergyPlus [5] is a popular program for simulating energy consumption in custom buildings under various internal and external conditions. [23, 31] are two recent works which utilize EnergyPlus to model specific buildings. However, EnergyPlus is a complex program which requires careful modeling of the building to provide an accurate estimate. For larger number of buildings, this becomes increasingly difficult to scale.

Recently, a number of studies have explored energy estimation from city-wide datasets. In Rotterdam, [21] utilized a Geographical Information System (GIS) to "downscale" the energy consumption in the city to individual residential buildings. A multiple linear regression model was able to achieve a Mean Absolute Percentage Error (MAPE) of 9% for electricity prediction. Similarly, [17] and [11] utilize city-level energy consumption data in New York City, along with regression models to predict energy consumption at the block level and building level. *CityEnergy* builds on these works by incorporating population estimates to compute personal energy footprints.

There are additionally a number of methods for estimating population in a city. A recent study [20] proposed a model for estimating dynamic populations of communities from subway smart card data. In [30], the authors study flow from multiple transportation modalities using different machine learning models. *CityEnergy* utilizes similar broad ideas to estimate dynamic populations from subway and vehicle transportation modalities, but further develops the ideas towards a more granular population estimate.

Transportation mode recognition using mobile devices is a well studied problem, and primarily relies on GPS, accelerometer, and gyroscope sensor data. Examples of studies using mobile sensor data include [27, 41], which only utilize GPS data and classifiers such as decision trees to determine the mode of transportation; [10, 13], which utilize accelerometer and gyroscope data along with support vector machines; and [37] which utilizes GPS, accelerometer, and gyroscope data. While *CityEnergy* incorporates transportation mode detection, we do not claim any novelty in this area.

Recently, the field of personal energy footprinting has begun to grow, with a number of studies combining energy monitoring, occupant localization, and energy apportionment to estimate real-time personal energy footprints. In residential homes and apartment buildings, [18] and [29] deployed energy monitoring sensors to detect energy consumption of each occupant. For commercial buildings, [32, 33] presented a scalable system to apportion energy consumption using different policies; [36] expanded on this system to additionally deliver energy saving recommendations. However, the main difficulties in scaling these works to multiple buildings is the *cost of deployment*. Most buildings require retrofitting of energy and occupant monitoring systems to enable energy footprinting. Building on our previous work [34], *CityEnergy* is the first city-scale energy footprinting system which does not require such monitoring systems.

3 CHALLENGES

The goal of *CityEnergy* is to provide "full coverage" for a person's energy footprint. Coverage in this application broadly refers to the percentage of time and locations for which our system can provide a reasonable personal energy footprint estimate. We noted three desirable characteristics to maximize coverage: high spatial granularity, high temporal granularity, and high accuracy. In the design of *CityEnergy*, the two most critical components which must adhere to these characteristics are energy estimation and population estimation.

3.1 Energy Estimation

To enable a real-time energy footprint estimate, a system must be capable of producing an energy consumption estimate for any building at any time of day. The main challenge is in achieving *both high temporal granularity as well as high spatial granularity, without sacrificing* accuracy. We demonstrate this challenge by comparing two types of energy estimation methods frequently seen in literature.

Building level regression models, which use historical data of a building to predict energy consumption, can achieve high accuracy for different levels of temporal granularity [12, 14, 38, 40**?**]. However, energy data at sub-daily or sub-hourly time scales is difficult to obtain for even a single building, much less at the city-scale. There are two main obstacles to extending this method to the city-scale. Firstly, fine temporal energy data is often proprietary, and cannot be easily accessed through public datasets. Secondly, this data is often obtained for specific purposes, meaning that most buildings will not have accessible historical energy data. Thus, while building level regression models are a good option for localized energy consumption studies, they are difficult to generalize to the city-scale.

City-wide energy estimation, which uses data containing energy consumption of a sample of buildings, has been used to estimate energy consumption at the city-scale [11, 17, 21]. However, the data is often collected at the yearly or monthly scale, and thus does not achieve sufficient temporal granularity. The main obstacle to "downscaling" the monthly or yearly energy consumption data to a finer temporal granularity, is the lack of knowledge about the specific behavior of the building for different environmental conditions (weather, time of day, day of week, etc.). City-wide energy estimation is able to achieve the necessary spatial granularity, but requires additional information to achieve temporal granularity.

3.2 Population Estimation

Similarly to energy estimation, a system must be capable of producing a population estimate with high spatial and temporal granularity. The most common form of population estimation is from mobility models; however, these models typically require GPS traces or cellular information from base stations, which are often not accessible by the public due to privacy concerns. Another method of population estimation is by studying the different types of transportation and inferring dynamic population.

According to [4], of the 4.8 million commuters in New York City, 38.7% ride the subway and 26.9% drive in vehicles. Thus, by estimating these two modalities of transportation, 67% of the dynamic population due to commuting can be accounted for. The remaining challenges are how to estimate each of these modes of transportation, and how these estimates translate to dynamic population.

Each of these modes presents unique challenges in both spatial and temporal granularity. For subways in New York City, there exists historical data at four hour intervals describing the number of people exiting and entering the station through the turnstiles; however, finer temporal granularity, as well as the changes in population of the surrounding area, needs to be modeled.

For motor vehicles, New York City traffic cameras can be utilized to determine the density of vehicles. Three challenges are how to determine density of vehicles from traffic cameras, how to infer vehicle density of unobserved roads, and translating vehicle density to dynamic population.

4 CITY-WIDE ENERGY ESTIMATION

4.1 Data Sources

As discussed in Section 3, one major challenge to estimating energy consumption is a lack of data at a sufficiently granular level. In New York City, there are no public datasets of energy consumption at the building level. This poses a problem, as many energy estimation techniques rely on such data.

However, in [17], the authors demonstrate a method for estimating the energy use intensity of any building in New York City by training a machine learning model on the energy benchmarking dataset from New York City's Local Law 84 [16]; this type of dataset is also collected in other cities with benchmarking laws such as Los Angeles and San Francisco. This dataset, however, only includes buildings greater than 50, 000 square feet. One disadvantage of this dataset is that due to the size of the buildings benchmarked, most of the included buildings are commercial; this excludes residential buildings.

We chose to further incorporate an energy dataset of residential buildings from the New York City Housing Authority (NYCHA) [24]. This dataset contains energy consumption data of over 2, 400 residential buildings at a monthly scale, and is home to 1 in 14 New Yorkers. We reasoned that including a residential dataset would increase the overall model accuracy for energy estimation.

In addition to the Local Law 84 and NYCHA datasets, we also incorporated hourly energy traces from the Department of Energy's (DOE) reference buildings [7]. This dataset contains baseline energy consumption values for 16 building types covering commercial and residential buildings. We utilize this dataset drill down from monthly to hourly energy estimation.

Finally, to account for out-of-sample buildings, we utilize features derived from the Primary Land Use Tax Lot Output (PLUTO) datsaet [25]. This dataset contains numerous features of every building in New York City, including building age, gross floor area, and floor area for different use types. We derive many of the input parameters of our energy consumption model from the PLUTO dataset.

4.2 Energy Consumption Model

As discussed in Section 3.1, an energy consumption model should have high spatial and temporal granularity. To achieve this, we

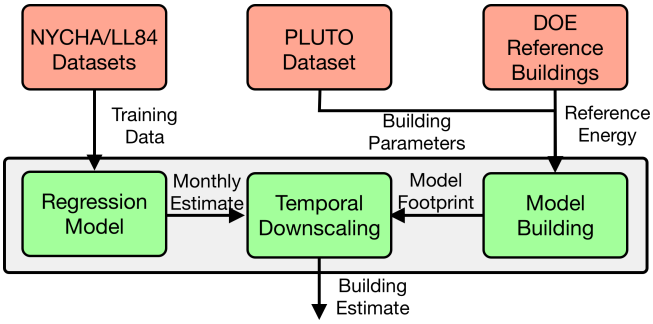


Figure 2: Diagram of the energy estimation pipeline.

developed a two stage model, as shown in Figure 2. In the first stage, a regression model is trained on monthly energy consumption data; this enables prediction of energy consumption of individual buildings. The second stage is a fitting model to downscale monthly energy consumption to hourly energy consumption.

To enable monthly energy estimation of any building in New York City, we trained a machine learning model on energy consumption data from the Local Law 84 and NYCHA datasets. As inputs to the model, we extracted parameters from the PLUTO dataset that represented important features of each building. As described in [17], features such as gross floor area, year built, office floor area, residential floor area, and borough are the best predictors of energy consumption. In addition to these parameters, we also included weather data, office floor area, retail floor area, garage floor area, and factory floor area.

The first stage outputs a monthly energy consumption estimate for each building; however, *CityEnergy* also requires fine time granularity. To transform monthly energy consumption to hourly energy consumption, we construct a model building based on the Department of Energy Commercial Reference Buildings [7]. The Reference Buildings consist of 16 building types, and provide hourly energy simulations in EnergyPlus.

For each building in New York City, the PLUTO dataset provides the floor areas by usage type; these roughly correspond to common building types in the DOE Reference Buildings. The model for each building is constructed via a linear combination of the Reference Building energy traces, with the weights corresponding to the percentage floor area of each type of building. An example is provided in Figure 3; a New York City building which is 50% apartment floor area, 30% office floor area, and 20% retail floor area is modeled by combining the respective percentages of the DOE Reference Buildings energy traces.

Finally, to incorporate the monthly energy consumption estimate from the first stage, we scale the model trace to equal the monthly energy consumption estimate. When a client requests an energy footprint of the building, the hourly energy footprint of the building is used in the computation of the personal energy footprint. An illustration of the average building energy footprint by block is shown in Figure 4. We evaluate the accuracy of this approach in Section 7.1.

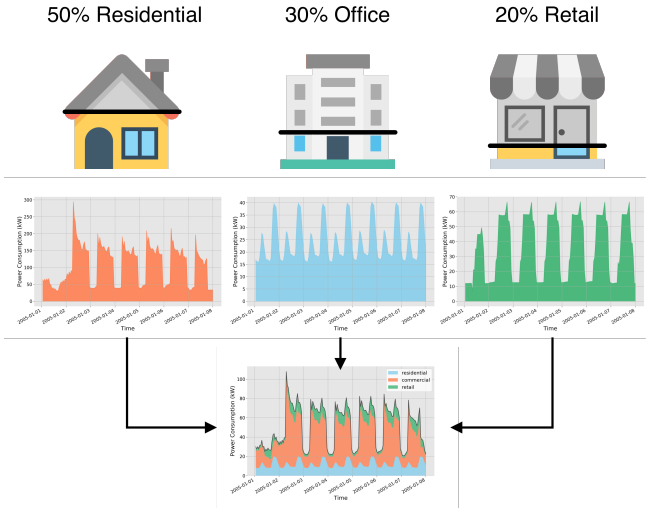


Figure 3: One example model building combination from the DOE Reference Buildings using weights corresponding to floor area for this building.

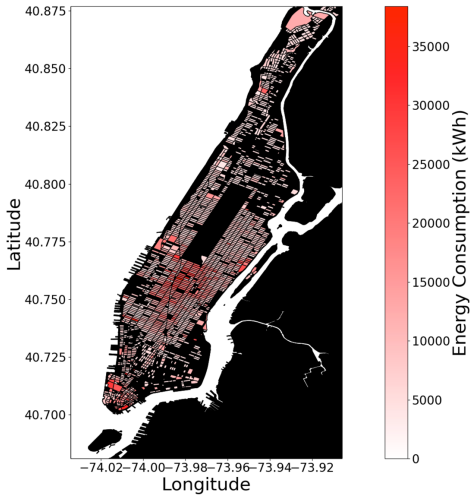


Figure 4: Heatmap of average building energy footprint in Manhattan by block.

5 CITY-WIDE POPULATION ESTIMATION

Another critical source of information for *CityEnergy* is real-time population estimation. A estimate of the number of people in a particular building, coupled with the energy estimate, is enough information to provide a real-time energy footprint. Unfortunately, there are no public datasets detailing population counts at any spatial granularity. In addition, the population varies throughout the day and week due to the large number of commuters; this further complicates modeling of real-time building population.

As stated in Section 3.2, 67% of dynamic population due to commuters can be estimated from subway ridership and driving. By combining dynamic population estimates from different modes of

transportation, a real-time population estimate can be built. *CityEnergy* utilizes data from the U.S. Census as a baseline population model, and uses real-time and historical data from the New York City Metropolitan Transportation Authority (MTA) and NYC traffic cameras to estimate dynamic population.

5.1 Baseline Population

A preliminary baseline model was constructed using the US Census dataset as initial populations for each block. As population data at the building level is not available from the Census, we downscaled the block level population to the building level through a simple model. We assume that people only reside in residential floor area; the residential floor area of each building in the block is collected from the PLUTO dataset. The estimated static population of building x can then be computed by Equation 5.1, as the ratio of the floor area of the building (FA_x) to the aggregate floor area of the census block times the census block population P_{CB}.

$$P_x = P_{CB} \frac{FA_x}{\sum_{B \in CB} FA_B}$$

5.2 Dynamic Population

To estimate real-time population dynamics, we analyze two of the most common modes of transportation in New York City: subway and motor vehicle. Estimates of the two modes of transportation are combined with Google Places data and population models to estimate dynamic population, as shown in Figure 5.

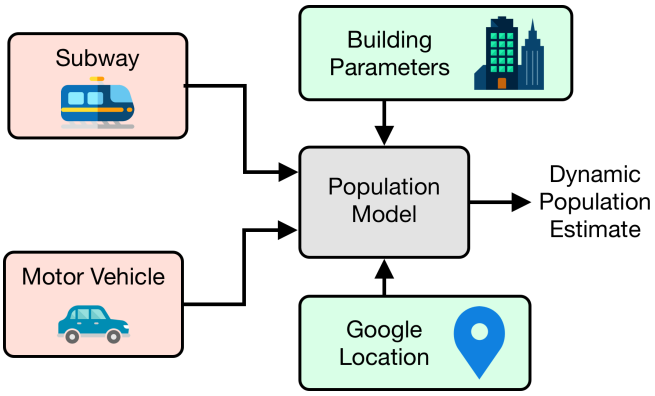


Figure 5: Illustration of the dynamic population estimation pipeline.

5.2.1 Subway. Every week, the New York City MTA publishes a new dataset detailing the commuter throughput for turnstiles in NYC subway stations at four hour intervals [2]. Each subway station has multiple turnstiles, each of which records the number of commuters entering as well as exiting the subway station.

To estimate the number of people entering or exiting from a station, we construct time series data from the past 6 weeks, which is used to train a regression model. The parameters used in the training include: historical commuter flow, hour, day of week, and weather conditions.

One consideration is that commuters enter a subway station at more uniform frequencies than they exit. The reason is that exiting commuters usually correspond to arrivals of subway trains. The regression model can estimate commuter throughput for four hour intervals; however, by using real-time data, the temporal granularity can be improved. To achieve this, *CityEnergy* utilizes the subway data feeds [1] from the NYC MTA, which provide the real-time locations of all running subway trains.

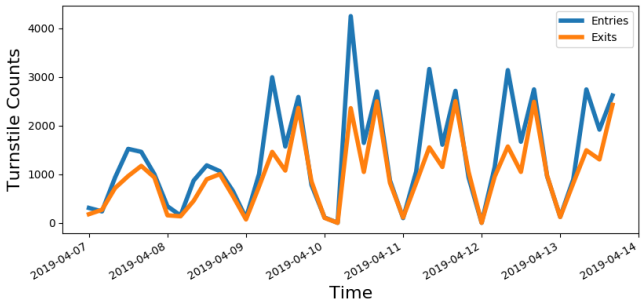


Figure 6: Illustration of the subway turnstile data and the subway throughput estimation.

When a subway reaches a stop, a number of commuters are estimated to leave the station; the estimated distribution of these commuters to the surrounding buildings is discussed in 5.3. An illustration of the whole subway estimation pipeline is shown in Figure 6. Evaluation of different regression models for hourly throughput are presented in Section 7.2.1.

5.2.2 Motor Vehicles. Besides the subway, the second most frequently used mode of transportation by commuters is by motor vehicle. For *CityEnergy*, we estimate motor vehicle commuting by analyzing real-time footage from traffic cameras, building on the work in [35]. Real-time image feeds are publicly available from the New York City Department of Transportation, which provides 752 real-time traffic cameras covering major intersections.

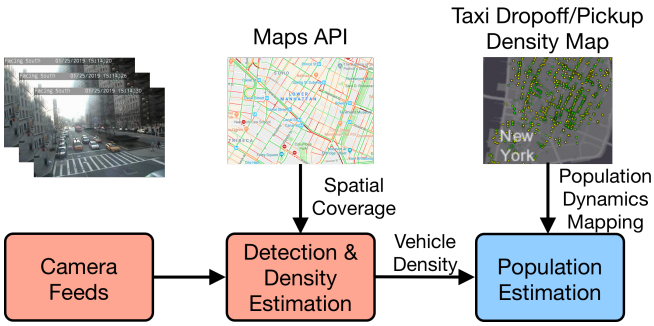


Figure 7: Pipeline for determining dynamic population at the block level from traffic cameras.

The estimate of dynamic population from traffic cameras is set in two stages as shown in Figure 7. First, the density of motor vehicles throughout the city is estimated through computer vision methods. Second, the vehicle density, along with city information of parking spaces, is used to estimate population change at the block level. The traffic camera streams present two main challenges: low resolution

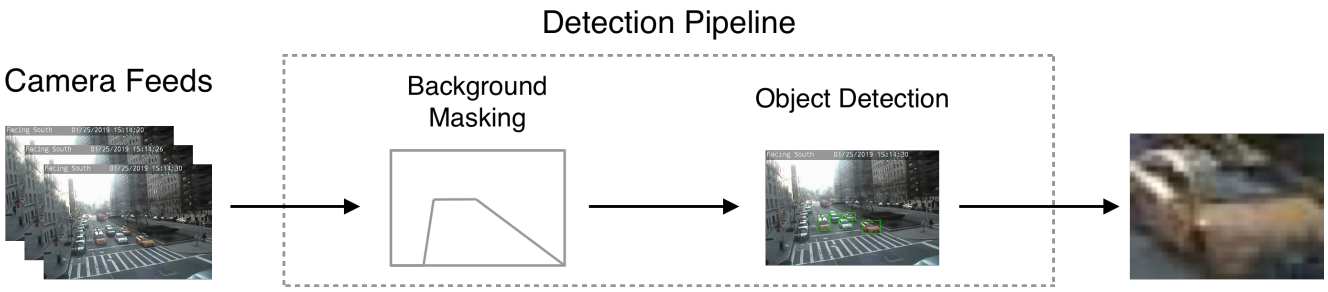


Figure 8: Computer vision pipeline for detecting vehicles. A) the background is masked from the incoming video stream image, and B) an object detection network extracts the candidate bounding boxes.

and low frame-rate. The stream resolution is 352x240 pixels, and are updated at a frequency between 0.3 – 1 Hz. Thus, to determine motor vehicle density, we implement a vehicle detection pipeline composed of two parts: background masking and object detection as shown in Figure 8. Initially, the background of the images of the video stream are masked as in [39]. This ensures that low resolution patterns in the background will not be falsely detected as vehicles.

The masked image is then fed to an object detection network. To increase accuracy, we utilize training data from two sources: the CityCam dataset [39], and a custom hand labeled dataset comprising of two thousand images and approximately five thousand vehicles. We use transfer learning to tune a pre-trained SSD-Mobilenet model [19] to better recognize vehicles at low resolution. Transfer learning is achieved by freezing all layers except for the final layer, and retraining the neural network. Once trained, the network provides an approximate vehicle density for the roads in view of the traffic camera.

Even with the large number of deployed traffic cameras, a majority of the streets remain unobserved. To provide full spatial coverage, we query the Here location framework API[1] (similar to the Google Maps API) which provides indications of traffic density on road segments not covered by traffic cameras.

Vehicle density, however, is not sufficient to predict dynamic population. For example, there are roads with high vehicle density but low dynamic population, such as highways. Unfortunately, there is no available data describing the relationship between vehicle density and population dynamics. There is, however, a dataset describing millions of taxi trips and destinations, which is sufficient to give a frequency of transfer to specific areas. The traffic data is then used to scale the population dynamics of the surrounding buildings, as shown in Figure 7.

5.2.3 Google Places API. Google's Places API[2] provides information about the popularity and the current estimated occupancy ("live" data) of many retail locations in New York City. When current estimated occupancy is available, this value is used as an estimate for the retail location; otherwise, the popularity is used.

5.3 Population Models

Given dynamic population estimates from the subway and motor vehicles, we estimate the dynamic population in buildings by using a population model. First, we define "catchment area" in this application as the spatial regions which are serviced by a transportation hub (a subway station, parking space/lot). Catchment area is defined by a Voronoi diagram, where distance is defined by street distance. The catchment areas of the subway stops in New York City are shown in Figure 9.

The catchment area determines the buildings which are serviced by either a subway station or parking lot. However, the different buildings receive a disproportionate number of people for different times of the day; for example, the probability that a person is traveling to the office rather than home is higher in the morning, and the reverse for the afternoon. We utilize the Citywide Mobility Survey [26], which provides the destinations of a sample population in New York City for different modes of transportation, including motor vehicles and subway. From this survey, we extracted the following distributions for trip destinations as shown in Table 1.

	Office	School	Retail	Errand	Medical
Motor Vehicle	32.5%	4.9%	45.1%	14.1%	4.3%
Subway	61.6%	4.6%	21.9%	7.8%	4.0%

Table 1: Distribution of traveler destinations by mode of transportation, extracted from the NYC DOT Citywide Mobility Survey.

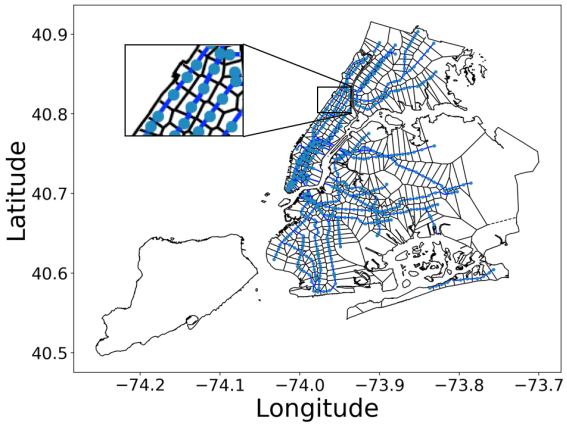


Figure 9: Illustration of catchment areas of subway stops in New York City.

[1] https://developer.here.com/
[2] https://developers.google.com/places/web-service/search

Figure 10a shows a heatmap of the real-time population in Manhattan at the block level. To our knowledge, this is the highest granularity temporal-spatial mapping of population using dynamic estimation from vehicle modalities.

6 DESIGN AND IMPLEMENTATION

Our system is composed of four connected subsystems: energy estimation, population estimation, transportation mode detection, and energy footprinting. As described in Sections 4 and 5, the energy and population estimation subsystems provide corresponding estimates for each building at fine temporal granularity. The transportation mode detection subsystem uses sensor data from the user's mobile device to determine mode of transportation, and to provide an energy estimate. Lastly, the energy footprinting module combines the results from the other three subsystems to produce an energy footprint estimate. Energy footprinting information is relayed to the user via a mobile application. The system architecture is shown in Figure 11.

CityEnergy is deployed in New York City, which is an ideal testbed for two reasons. Firstly, the city publishes a number of relevant datasets containing information about energy consumption in the built environment and population mobility for various modes of transportation. The abundance of datasets makes possible the study of energy footprinting at a city scale, while presenting new challenges in data representation, cleaning, and modeling.

Secondly, there is a wide variety of buildings and people in New York City. These variations lead to interesting differences in energy usage and mobility. For example, the age of buildings in New York City varies tremendously, which can correlate with energy usage differences due to the available building management systems (BMS) technology. On the other hand, the number of people in Manhattan varies greatly between weekdays and weekends due to people commuting from outside of the city, according to NYU Wagner [22]. These variations lead to interesting challenges in the design of a city-scale energy footprinting system.

6.1 Transit Estimation

To complete the idea of "full coverage" for energy consumption, *CityEnergy* includes a transit estimation module. This module is responsible for analyzing location information from the users, determine whether the user is in transit, and provide the corresponding personal energy footprint. If the user is determined not to be in transit, the personal energy footprint is based on apportioned energy, as described in Section 6.2; otherwise, the personal energy footprint is based on the specific mode of transportation of the user. We do not claim any novelty in this section, but it is critical to maintain high temporal granularity of the user's energy footprint.

6.1.1 Transportation Mode Detection. We focused on classifying the two main modes of transportation estimated in Section 5, as well as walking. Detection of the mode of transportation relies on the location traces collected from the user's mobile device. A few simple thresholds are used to separate the modes of transportation. If the location change between multiple samples is below a lower threshold, L1, the user is considered to be stationary. Above the lower threshold L1 and below an upper threshold L2, the user is considered to be walking; and above L2, the user is considered to be in a vehicle. Finally, due to the nature of the NYC subway system being primarily underground, the location traces create large jumps, which separates subway transportation from motor vehicle transportation. This section is not the primary focus of this work, and we do not claim any novelty.

Mode of Transport	Estimated Power Consumption
Subway (Weekday)	36.7 Wh
Subway (Weekend)	72.2 Wh
Bus	587 Wh/km
Vehicle	874 Wh/km
Walking	0.0 W

Table 2: In-sample mean absolute error and mean root squared error of different machine learning regressors on historical turnstile data.

6.1.2 Transit Energy Footprint. If an occupant is in transit, *CityEnergy* provides an energy footprint depending on the mode of transportation. Because of the lack of relevant real-time information such as whether a motor vehicle is electric or gasoline, or how many people are sharing a subway car/bus/motor vehicle, we can only offer high level methods for energy consumption estimates. In future works, it may be possible to obtain better transit energy estimates by incorporating various sensors or user feedback.

An energy footprint estimate for a user on the NYC subway can be computed as the average energy of the entire NYC subway system, divided by the average ridership (weekday or weekend) [1].

$$E^{p}_{subway} = \frac{E_{subway}}{R_{weekday/weekend}}$$

An energy estimate per kilometer for a user on a NYC bus is computed as the energy consumption per kilometer over the average number of riders per trip (Epkm is energy per kilometer) [1, 8].

$$Epkm^{p}_{bus} = \frac{Epkm_{bus}}{\frac{ridership}{trips}}$$

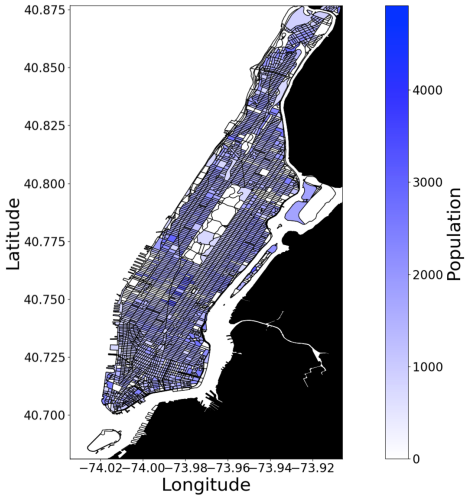
Lastly, to compute the energy estimate per kilometer for a user in a motor vehicle, we divide the approximate energy potential of gasoline by the US average distance per gallon of gasoline ($\overline{FE}$ is average fuel efficiency) [3].

$$Epkm^{p}_{car} = \frac{Epkm_{car}}{\overline{FE}}$$

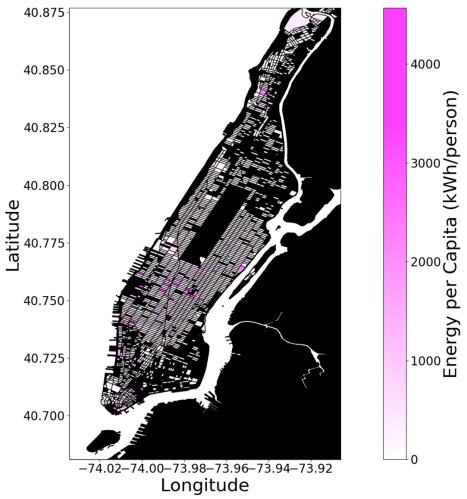
A summary of energy consumption estimates used in *CityEnergy* are provided in Table 2.

6.2 Apportionment

Once an estimate for a building's energy consumption and population are calculated, a user's personal energy footprint can be estimated by applying an apportionment policy. From [9], different apportionment policies can be applied depending on the situation. The simplest apportionment policy is uniform apportionment, such that the total energy consumption is distributed uniformly over the number of people. The energy consumption per capita for a typical morning in Manhattan is shown in Figure 10b.



(a) Heatmap of population in Manhattan by block.



(b) Energy consumption per capita, with a uniform apportionment policy.

Figure 10: Energy Footprinting of Manhattan on a typical weekday at 8 AM.

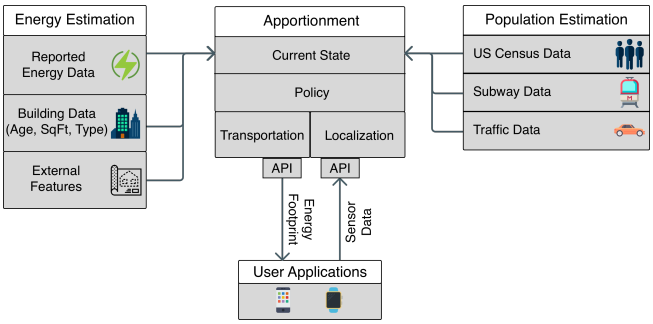


Figure 11: System architecture block diagram.

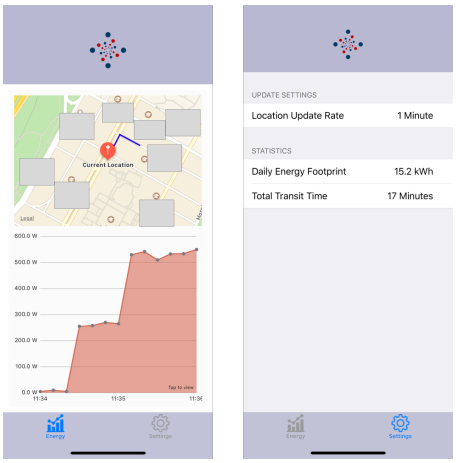


Figure 12: Screenshots of the *CityEnergy* iOS application. Left: User's recent estimated energy footprint and location trace. Right: Settings screen.

6.3 User Applications

Mobile Energy Footprinting: To provide energy footprints and other actionable feedback to everyday users, we developed a mobile application for iOS. The application is responsible for sensing location data, such as GPS and WiFi information, which is then encoded and sent to the server. Once *CityEnergy* has localized the user and determined an energy footprint estimate, corresponding data is returned and displayed to the user in real-time. Screenshots of the mobile application are shown in Figure 12.

City Planner Applications: We also plan to make public web applications displaying city-wide information about energy and population estimates at the block level, such as in Figures 4, 10a and 10b. These tools can be useful for a variety of applications such as transportation planning and energy planning.

7 EVALUATION

We conducted the evaluation of *CityEnergy* in three parts: energy estimation, population estimation, and as a complete system.

7.1 Energy Estimation

As described in Section 4, energy consumption is estimated by passing building parameters through a trained regression model, and downscaled to the hourly scale by using a model building derived from the DOE Reference Buildings. To evaluate the energy estimation pipeline, we evaluated in-sample estimation with different regression models. In addition, we gathered data from a few representative buildings over two months to assess out-of-sample estimation. Note that evaluation uses the logarithm of energy consumption, as in [17].

7.1.1 In-Sample Estimation. We tested a few different machine learning models in an effort to improve prediction accuracy. Support vector regression, random forest, and linear regression models were

Model	In-Sample MAE	In-Sample MRSE	R^2
SVR	0.74	1.11	0.28
Linear Regression	0.77	1.06	0.34
Random Forest	**0.30**	**0.48**	**0.86**

Table 3: In-sample mean absolute error and mean root squared error of different machine learning regressors on the Local Law 84 and NYC Housing Authority datasets.

trained on the Local Law 84 and NYCHA datasets using five-fold cross validation. The mean absolute error and mean root squared error for each model are shown in Table 3.

Model	SVR		Linear Regression		Random Forest	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Residential	**0.66**	**0.69**	1.48	1.49	0.87	0.93
Commercial	1.62	1.65	2.47	2.49	**0.47**	**0.48**
Retail	**0.93**	**1.08**	2.0	2.07	2.61	2.67

Table 4: Comparison of regression models for out-of-sample energy estimation of four ground truth buildings.

7.1.2 Out-of-Sample Estimation. To evaluate the generalizability of the trained energy regression models to the city-scale, we gathered data from one residential building, one commercial building, and one retail building over the course of two months. We generated energy estimation traces for these buildings using *CityEnergy*, and calculated the MAE and MRSE against the ground truth in Table 4.

As described in [17], linear regression can produce low MAE for both in-sample and out-of-sample energy estimation; however, in our experiments, support vector regression yields the lowest MAE and MRSE. Although the accuracy is low (especially for commercial buildings), the model can still provide reasonable energy estimates at the city-scale. More complex models can be used to further improve the accuracy of energy estimation in *CityEnergy*.

7.2 Population Estimation

To evaluate the population estimation in *CityEnergy*, we individually evaluate the subway estimates. Instead of evaluating vehicle to population dynamics, we evaluate the total population estimates on 5 ground truth buildings throughout New York City to assess the correctness of the population models.

7.2.1 Subway. A regression model is trained for each subway station on the previous six weeks of turnstile data. We tested three different regression models to determine the best in-sample MAE and MRSE, as shown in Table 5. From our experiments, random forest provides the highest accuracy. Evaluation uses the logarithm of the outflow and inflow populations.

7.2.2 Population Models. Finally, we evaluated the full dynamic population pipeline. For 4 different buildings (two residential, one commercial and one retail), we manually counted the number of occupants in a building as ground truth. We combined the dynamic population estimates from the subway and motor vehicles as described in Section 5.3, and scaled the total to account for the remaining 33% from other transportation modalities.

Model	In-Sample MAE	In-Sample MRSE	R^2
SVR	0.27	0.44	0.80
Linear Regression	0.74	0.93	0.13
Random Forest	**0.24**	**0.43**	**0.81**

Table 5: In-sample mean absolute error and mean root squared error of different machine learning regressors on historical turnstile data.

Using our population model, we were able to achieve an MAE of 0.7 for the 4 test buildings. We found that the error varied greatly depending on the type of building, time of day, and other factors. Although the population models used were sufficient for *CityEnergy*, we believe that covering additional modes of transportation and increasing the existing models can produce a model with much higher accuracy.

7.3 Energy Footprinting Estimation

The most important characteristic for *CityEnergy* is coverage; however, the energy footprint estimate should also be accurate enough to approximate a ground truth energy footprint.

To demonstrate the potential accuracy of *CityEnergy*, we collected a real energy footprint as ground truth. The real energy footprint consisted of three building locations: one residential location, one retail location, and one commercial building. The ground truth energy footprint is collected as follows:

- In the residential location, personal energy footprint is determined using a set of plugmeters and light sensors to measure the energy consumption.
- In the retail location, a camera was deployed to monitor the building's energy meter; the population was recorded manually.
- In the commercial building location, a combination of plugmeters, light sensors, and building management system data was utilized to determine the personal energy footprint.

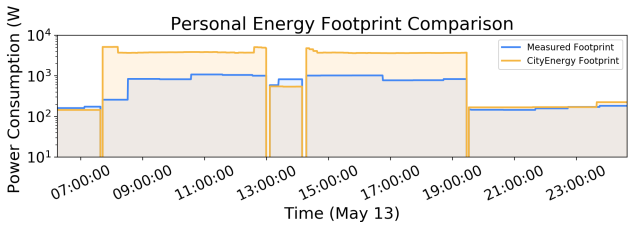


Figure 13: Semi-log plot of the measured real-time energy footprint of a user, and the energy footprint estimate from *CityEnergy*.

A comparison of the ground truth energy footprint and the *CityEnergy* energy footprint is shown in Figure 13. *CityEnergy* is able to achieve a MAE of 1.7 kWh. For different buildings and different energy footprints, *CityEnergy* may achieve higher or lower accuracy. However, better energy estimation and population estimation models can help reduce error in future works.

8 CONCLUSION

In this work, we present *CityEnergy*, a personal energy footprinting system which uses energy and population models to provide coverage throughout an urban environment. *CityEnergy* can be easily extended to other urban cities by interchanging available energy and population data specific to the urban city. *CityEnergy* is, to our knowledge, the first system to address the problem of personal energy footprinting at the city-scale without relying on building specific energy and population monitoring deployments.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation under Grant Numbers CNS-1704899 and CNS-1815274. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed implied, of Columbia University, NSF, or the U.S. Government or any of its agencies.

REFERENCES

[1] Metropolitan Transportation Authority. 2018. Metropolitan Transportation Authority. (2018).
[2] Metropolitan Transportation Authority. 2018. Turnstile Data. (2018).
[3] Michael R Bloomberg. 2007. Inventory of New York City greenhouse gas emissions. *New York City MayorâĂŹs Office of Operations, Office of Long-term Planning and Sustainability* (2007).
[4] U.S. Census Bureau. 2017. U.S. Census Bureau, 2017 American Community Survey 1-Year Estimates. (2017).
[5] Drury B Crawley, Linda K Lawrie, Frederick C Winkelmann, Walter F Buhl, Y Joe Huang, Curtis O Pedersen, Richard K Strand, Richard J Liesen, Daniel E Fisher, Michael J Witte, et al. 2001. EnergyPlus: creating a new-generation building energy simulation program. *Energy and buildings* 33, 4 (2001), 319–331.
[6] Sarah Darby et al. 2006. The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays* 486, 2006 (2006), 26.
[7] Office of Energy Efficiency Department of Energy and Renewable Energy. 2018. Commercial Reference Buildings. (2018).
[8] Igors Graurs, Aigars Laizans, Peteris Rajeckis, and Aivars Rubenis. 2015. Public bus energy consumption investigation for transition to electric power and semi-dynamic charging. *Eng. Rural. Dev* 14 (2015), 366–371.
[9] Simon Hay and Andrew Rice. 2009. The Case for Apportionment. In *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*. ACM, 13–18.
[10] Samuli Hemminki, Petteri Nurmi, and Sasu Tarkoma. 2013. Accelerometer-based transportation mode detection on smartphones. In *Proceedings of the 11th ACM conference on embedded networked sensor systems*. ACM, 13.
[11] B Howard, L Parshall, J Thompson, S Hammer, J Dickinson, and V Modi. 2012. Spatial distribution of urban building energy consumption by end use. *Energy and Buildings* 45 (2012), 141–151.
[12] Samuel Humeau, Tri Kurniawan Wijaya, Matteo Vasirani, and Karl Aberer. 2013. Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households. In *2013 Sustainable Internet and ICT for Sustainability (SustainIT)*. IEEE, 1–6.
[13] Arash Jahangiri and Hesham A Rakha. 2015. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE transactions on intelligent transportation systems* 16, 5 (2015), 2406–2417.
[14] Rishee K Jain, Kevin M Smith, Patricia J Culligan, and John E Taylor. 2014. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy* 123 (2014), 168–178.
[15] Radiša Ž Jovanović, Aleksandra A Sretenović, and Branislav D Živković. 2015. Ensemble of various neural networks for prediction of heating energy consumption. *Energy and Buildings* 94 (2015), 189–199.
[16] Constantine Kontokosta. 2012. Local Law 84 Energy Benchmarking Data: Report to the New York City Mayor's Office of Long-Term Planning and Sustainability. (2012).
[17] Constantine E Kontokosta and Christopher Tull. 2017. A data-driven predictive model of city-scale energy use in buildings. *Applied energy* 197 (2017), 303–317.
[18] Seungwoo Lee, Daye Ahn, Sukjun Lee, Rhan Ha, and Hojung Cha. 2014. Personalized energy auditor: Estimating personal electricity usage. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*. IEEE, 44–49.
[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.
[20] Yunjia Ma, Wei Xu, Xiujuan Zhao, and Ying Li. 2017. Modeling the hourly distribution of population at a high spatiotemporal resolution using subway smart card data: A case study in the central area of Beijing. *ISPRS International Journal of Geo-Information* 6, 5 (2017), 128.
[21] Alessio Mastrucci, Olivier Baume, Francesca Stazi, and Ulrich Leopold. 2014. Estimating energy savings for the residential building stock of an entire city: A GIS-based statistical downscaling approach applied to Rotterdam. *Energy and Buildings* 75 (2014), 358–367.
[22] Mitchell Moss and Carson Qing. 2012. *The Dynamic Population of Manhattan*. Rudin Center for Transportation, NYU Wagner School.
[23] Alex Nutkiewicz, Zheng Yang, and Rishee K Jain. 2018. Data-driven Urban Energy Simulation (DUE-S): A framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow. *Applied energy* 225 (2018), 1176–1189.
[24] New York City Housing Authority (NYCHA). 2018. Electric Consumption and Cost (2010 - June 2018). (2018).
[25] Department of City Planning. 2018. Primary Land Use Tax Lot Output. (2018).
[26] New York City Department of Transportation. 2017. Citywide Mobility Survey. (2017).
[27] Leon Stenneth, Ouri Wolfson, Philip S Yu, and Bo Xu. 2011. Transportation mode detection using mobile phones and GIS information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 54–63.
[28] Jenna Tatum. 2013. PlaNYC: New York City MayorâĂŹs Carbon Challenge Progress Report. *New York: MayorâĂŹs Office of Long-Term Planning and Sustainability* (2013).
[29] Shailja Thakur, Manaswi Saha, Amarjeet Singh, and Yuvraj Agarwal. 2014. WattShare: Detailed Energy Apportionment in Shared Living Spaces Within Commercial Buildings. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM, 30–39.
[30] Florian Toqué, Mostepha Khouadjia, Etienne Come, Martin Trepanier, and Latifa Oukhellou. 2017. Short & long term forecasting of multimodal transport passenger flows with machine learning methods. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 560–566.
[31] Liping Wang, Robert Kubichek, and Xiaohui Zhou. 2018. Adaptive learning based data-driven models for predicting hourly building energy use. *Energy and Buildings* 159 (2018), 454–461.
[32] Peter Wei, Xiaoqi Chen, Rishikanth Chandrasekaran, Fengyi Song, and Xiaofan Jiang. 2016. Adaptive and Personalized Energy Saving Suggestions for Occupants in Smart Buildings: Poster Abstract. In *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys '16)*. ACM, New York, NY, USA, 247–248. https://doi.org/10.1145/2993422.2996412
[33] Peter Wei, Xiaoqi Chen, Jordan Vega, Stephen Xia, Rishikanth Chandrasekaran, and Xiaofan Jiang. 2018. A Scalable System for Apportionment and Tracking of Energy Footprints in Commercial Buildings. *ACM Transactions on Sensor Networks (TOSN)* 14, 3-4 (2018), 22.
[34] Peter Wei and Xiaofan Jiang. 2018. A data-driven system for city-scale personal energy footprint estimations. In *Proceedings of the 5th Conference on Systems for Built Environments*. ACM, 194–195.
[35] Peter Wei, Haocong Shi, Jiaying Yang, Jingyi Qian, Yinan Ji, and Xiaofan Jiang. 2019. City-scale vehicle tracking and traffic flow estimation using low frame-rate traffic cameras. In *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. ACM, 602–610.
[36] Peter Wei, Stephen Xia, and Xiaofan Jiang. 2018. Energy Saving Recommendations and User Location Modeling in Commercial Buildings. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 3–11.
[37] Peter Widhalm, Philippe Nitsche, and Norbert Brändie. 2012. Transport mode detection with realistic smartphone sensor data. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 573–576.
[38] Tri Kurniawan Wijaya, SFRJ Humeau, Matteo Vasirani, and Karl Aberer. 2014. Residential electricity load forecasting: evaluation of individual and aggregate forecasts. In *tech. rep.* Citeseer.
[39] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. 2017. Understanding traffic density from large-scale web camera data. *arXiv preprint arXiv:1703.05868* (2017).
[40] Hai Xiang Zhao and Frédéric Magoulès. 2010. Parallel support vector machines applied to the prediction of multiple buildings energy consumption. *Journal of Algorithms & Computational Technology* 4, 2 (2010), 231–249.
[41] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 312–321.