Context-specific urban occupancy modeling using location-based services data

Journal Article

Author(s):
Happle, Gabriel; Fonseca, Jimeno A.; Schlueter, Arno

Publication date:
2020-05-15

Permanent link:
https://doi.org/10.3929/ethz-b-000404967

Originally published in:
Building and Environment 175, https://doi.org/10.1016/j.buildenv.2020.106803

Context-specific urban occupancy modeling using location-based services data

Gabriel Happle[a,b,], Jimeno A. Fonseca[a,b], Arno Schlueter[a,b]

[a]*Future Cities Laboratory, Singapore-ETH Centre, 1 Create Way, CREATE Tower, Singapore 138602*
[b]*Architecture and Building Systems, Institute of Technology in Architecture, ETH Zurich, Stefano-Franscini-Platz 1, CH-8093 Zurich, Switzerland*

Abstract

Energy-related occupant behavior is a major source of uncertainty in building and urban energy performance simulations. Standardized assumptions, published by ASHRAE and others in the form of occupancy schedules, are widely used in research and practice, especially on the district-scale. In this work, we gathered location-based services data to create context-specific, data-driven occupancy schedules. Using a web mapping service, we collected data for retail and restaurant uses in the downtown neighborhoods of 13 different U.S. cities to create data-driven schedules for each context. The schedules were compared to ASHRAE standard assumptions using the earth mover's distance approach and the schedules' energy-related features. We found that standard schedules seem to significantly overestimate weekly building occupancy, although the shapes of the schedules are generally similar. The use of standard schedules could therefore, have significant impacts on district-scale energy demand simulations, as the overestimation will be cumulative.

As compared to the differences between data-driven and standard schedules, the differences between different locations are significantly smaller. However in extreme cases, the weekly cumulative occupancy and the number of occupied hours differ by more than 30% between locations, which means that context-specific differences together with climatic differences might also impact building performance simulation results. Furthermore, we found differences in daily data between the different days of the week. In particular, the observed behavior on Fridays is significantly different from other weekdays for both considered use-types. This indicates that the conventional categorization of occupant behavior models into three day-types: weekday, Saturday, and Sunday, should be reconsidered.

Keywords: occupancy schedule, occupant behavior, urban building energy modeling, location-based services data

1. Introduction

1.1. Building energy modeling and occupant behavior

Physics-based bottom-up building energy models [1, 2] are commonly used to forecast the performance of new buildings or to assess the impacts of various retrofit measures for existing buildings. Uncorroborated assumptions

Email address: happle@arch.ethz.ch (Gabriel Happle) made regarding the behavioral aspects of energy consumption, such as the hours of occupancy and building systems usage [1], indicate possible weaknesses in such models. Specifically, *energy-related occupant behavior* is one of the main factors affecting building energy consumption and a major source of uncertainty in simulations [3, 4]. Often, occupant behavior is represented in the form of standardized schedules, also called profiles or diversity factors, which are 24-hour time series of fractions of nominal space occupancy, lighting loads, appliance loads, or building systems operation. *Standard schedules* are used when the actual behavior is unknown. This can lead to simulation result inaccuracies because these standard assumptions were intended to serve a limited number of specific purposes, which may not fit the actual purpose of the simulation.

1.2. Origins and purposes of standard occupancy schedules

Historically, the development of standard occupancy schedules was closely linked to the emergence of computers and the first building energy simulation programs. NBSLD (1974), one of the first tools to calculate annual energy demand in hourly time steps, required the input of "occupancy schedules" for weekdays and weekends as a "fraction of some maximum", which was described as a "normalized 24-hour profile of occupancy" and needed to be input together with the "maximum number of ... occupants during the 24 hour period" [5, 6].

The main driver for the creation of standard occupancy schedules was research aimed at developing energy performance standards for new construction in the late 1970s [7, 8]. For this purpose, the United States (U.S.) Department of Energy (DOE) published *Standard Building Operating Conditions (SBOC)* in 1979 [9]. The occupancy schedules for 14 building *use-types* were defined. It is critical to note that the numbers included in the schedules were determined from averaging the educated guesses of 168 building design teams, rather than from any observable data [9].

Table 1 lists the published standard schedules of occupancy for different building use-types from 1979 until the present. Table 2 lists the original purpose of publication at that time.

In 1989 the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) published standard occupancy schedules for the first time for 9 building use-types as part of the Building Energy Cost Budget (ECB) method of the *Standard 90.1* [10]. See Table 1. Many of them were based on the 1979 DOE SBOC. The same standard assumptions were published again by ASHRAE in 2004 in the user's manual to Standard 90.1-2004 [11].

In 2011 reference building energy models were developed by the DOE to standardize energy efficiency research [12]. Many of the occupancy schedules for these models were directly taken from ASHRAE documents [1]. For some building use-types, schedules were defined on the zone-level rather than on the building-level. Additionally, sub-categories for certain building uses were introduced. See Table 1.

In 2013 ASHRAE published standard occupancy schedules for 14 building use-types on their website to be used to calculate the Envelope Performance Factor for the Building Envelope Trade-Off Option procedure of their Standard 90.1-2013 [13, 14]. See Table 1. Currently, standard schedules are also available from other sources such as COMNET, which is an initiative to standardize building energy modeling [20]. On their web portal a set of 14 default schedules of building operation is available, and most schedules reference ASHRAE (2013) as the source. Although sporadically updated, many of the schedules available today are largely the same as in 1979 or 1989. According to [12], the ASHRAE schedules of 1989 were modified for the 2004 publication by a public review process. However by comparing the schedules of 2004 to 1989, it becomes evident that only the schedules for certain use-types (i.e. office, assembly, and restaurant), and only schedules for Sunday of the latter two were adjusted. By comparing different versions of schedules, some other important modifications can be noted over time. For example the health use-type was converted from day time operation to 24-hour occupancy, and restaurant use-type increased the occupancy in the AM hours. See Fig. 1b. Some schedules however have remained unchanged since 1979 such as retail use documented in Fig. 1a.

[1] For retail, restaurant, and office building use-types schedules were taken from the ASHRAE user's manual (2004) [11]. The health, hotel, school, and warehouse use-type schedules were taken from other ASHRAE documents, (i.e. Advanced Energy Design Guide (AEDG) Technical support documents) that were developed by ASHRAE in collaboration with the American Institute of Architects (AIA), the Illuminating Engineering Society of North America (IESNA), the U.S. Green Building Council (USGBC), and the DOE, based on the previously published ASHRAE standard schedules that were modified by inputs of the respective project committee members [15, 16, 17, 18, 19].

Table 1: Standard schedules of occupancy for building performance simulation of different building use-types published by different entities from 1979 until present.

Building use-type	DOE (1979)	ASHRAE (1989)	ASHRAE (2004)	DOE (2011)	ASHRAE (2013)	COMNET (2016)
Retail	Shopping Center,[Z] Store[Z]	Retail	Retail	Retail, Strip Mall, Supermarket	Schedule C[C]	Retail
Restaurant	-	Restaurant	Restaurant	Quick Service Restaurant, Full Service Restaurant	Schedule B[B]	Restaurant
Health	Clinic,[Z] Hospital[Z]	Health	Health	Hospital,[Z] Outpatient Health Care	Schedule E[E]	Health
Assembly	Community Center,[Z] Gymnasium,[Z] Theater/Auditorium[Z]	Assembly	Assembly	-	Schedule H[H]	Assembly
Office	Large Office,[Z] Small Office[Z]	Office	Office	Large Office, Medium Office, Small Office	Schedule A[A]	Office
Warehouse	Warehouse	Warehouse	Warehouse	Warehouse	Schedule L[L]	Warehouse
Hotel	Hotel/Motel,[Z] Nursing Home[Z]	Hotel/Motel	Hotel/Motel	Large Hotel,[Z] Small Hotel[Z]	Schedule F[F]	Hotel/Motel
School	Elementary School,[Z] Secondary School[Z]	School	School	Primary School,[Z] Secondary School[Z]	Schedule G[G]	School
Residential	Multifamily High-Rise Residential,[Z] Multifamily Low-Rise Residential[Z]	-	-	Midrise Apartment[Z]	Schedule D[D]	Residential
Manufacturing	-	-	Light Manufacturing	-	Schedule J[J]	Manufacturing
Gymnasium	-	-	-	-	Schedule I[I]	Gymnasium[I]
Parking	-	-	-	-	Schedule K[K]	Parking[K]
Various	-	-	-	-	-	Laboratory, Data Center

[Z] These building models contain occupancy schedules for zone-level usages
[A] for Courthouse, Office, Post Office, Town Hall
[B] for Dining: Cafeteria/Fast Food, Dining: Family, Dining: Bar Lounge/Leisure
[C] for Library, Museum, Retail
[D] for Dormitory, Multi-family
[E] for Fire Station, Health Care Clinic, Hospital, Police Station, Transportation
[F] for Hotel, Motel, Penitentiary
[G] for School, University
[H] for Convention Centre, Exercise Centre, Motion Picture Theatre, Performing Arts Theatre, Religious Building, Sports Arena
[I] for Gymnasium as part of a School
[J] for Automotive Facility, Manufacturing Facility, Workshop
[K] for Parking/Parking Garage
[L] for Warehouse

Table 2: Purposes of published standard schedules from 1979 to 2013.

Publisher	Purpose
DOE (1979) [9]	Development of Building Energy Performance Standards (BEPS) legislation
ASHRAE (1989) [10]	Building Energy Cost Budget (ECB) Method: Code compliance via simulation
ASHRAE (2004) [11]	Building Performance Rating Method: Estimation of % energy savings of advanced building designs
DOE/NREL (2011) [12]	DOE commercial building research: technology assessment, design optimization, analyze advanced controls, develop energy codes and standards, and to conduct lighting, daylighting, ventilation, and indoor air quality studies
ASHRAE (2013) [13, 14]	Envelope trade-off: Envelope code compliance via simulation

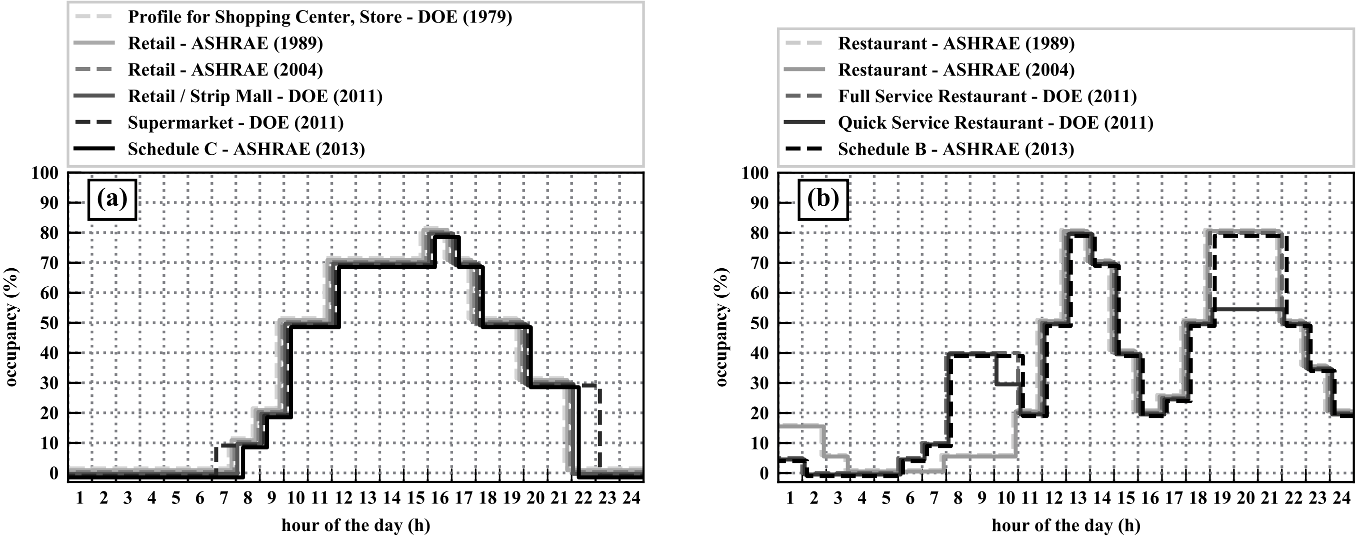


Figure 1: The historic evolution of the retail (a) and restaurant (b) standard occupancy schedule for weekdays from 1979 to present. The retail schedule (a) is quantized in steps of 10%. The restaurant schedule (b) is quantized in steps of 5%. The data is slightly shifted for visualization purposes.

On the building-scale, standard schedules are widely used for code compliance, systems design, and research purposes as intended by their publishers [12, 21, 22].

However, in our review of occupant behavior in urban building energy models [23], we found that standard schedules of occupancy are used in research beyond their original purpose. We found tools and studies that used standard schedules, among others, to:

- Generate patterns of anthropogenic heat generation,
- Generate energy demand patterns for district energy infrastructure design, and
- Generate energy demand patterns for district energy systems operation optimization.

1.3. Advanced occupant behavior models

Research efforts for improving energy-related occupant behavior modeling and simulations are consolidated in *IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings* [24] and the currently ongoing follow-up *Annex 79: Occupant-centric building design and operation* [22]. Reviews in 2015 and 2016 [25, 26] confirmed that there were not many models and tools available for advanced occupancy simulation. The tools that did exist were developed for office and residential use type buildings exclusively. As part of [25], a software module was developed to include three representative occupancy models. Namely, Chang's model [27] to simulate the occupancy state of a space, Page's model [28] to simulate the number of occupants in a space, and Wang's model [29] to generate the spatial location of each occupant and the space-level occupancy for the whole building. As part of IEA EBC Annex 66 and Annex 79 data-based models and agent-based models for building occupancy have been developed [30, 31, 32, 33, 34, 35, 36]. In [31], data-mining methods were used to derive office occupancy schedules from appliance power consumption measurements. In [32], data-mining methods were used to derive archetypal working profiles of individual occupants from measured occupancy data of 16 private offices with single or dual occupancy. In [35], machine learning techniques were used for daily occupancy patterns recognition for improving the energy efficiency of an office building. An agent-based model for office buildings has been developed by [36]. It is depending on expert user inputs, such as, e.g., the typical arrival times of each occupant, the number of planned meetings per day, and the probability distribution of meeting durations. The explicit goal of Annex 79 is the development of the next generation of "dynamic, stochastic, agent-based, and data-driven" [22] occupant models. The current common practice is, however, still the use of standard assumptions [22]. We believe that it is time to revisit these standards of commercial buildings by exploring new data sources — especially in the context of emerging sensing and data collection opportunities.

1.4. New data sources for building occupancy

Currently, new data sources of building occupancy are becoming available. Cell phone positioning systems use different signals, such as radio signals of cell towers, GPS signals, and Wi-Fi signals, to determine the accurate location of the phone [37]. Indoor positioning algorithms, using additional data of sensors embedded in smartphones, such as accelerometers and magnetometers, can determine the position of a cell phone inside a building [38].

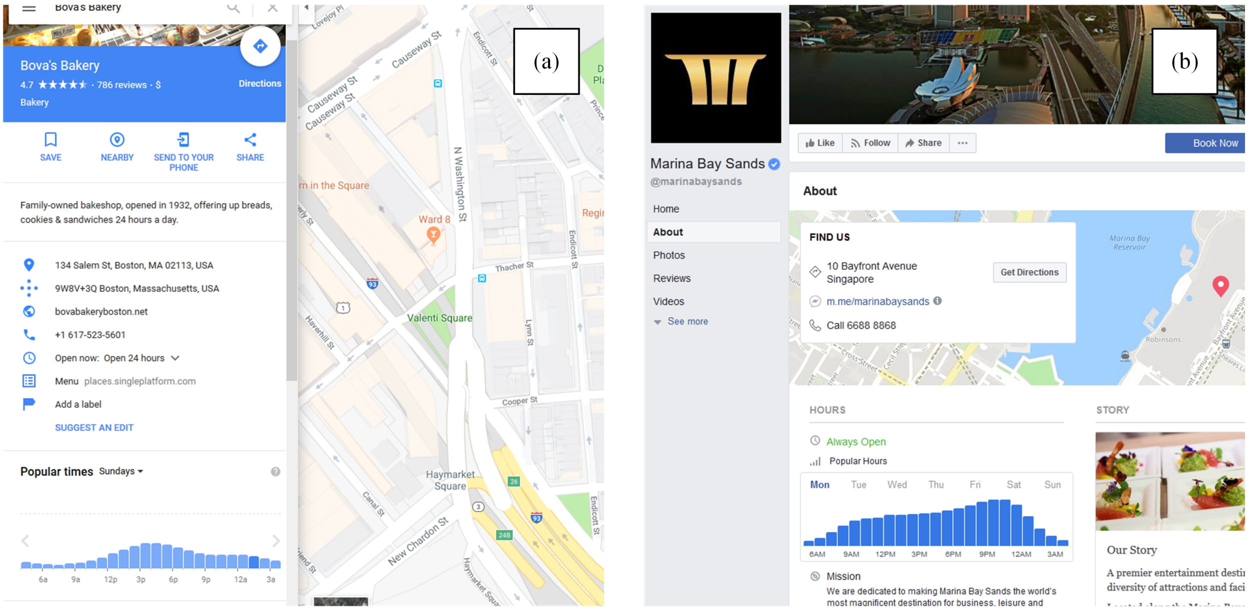
Coupled with information about the exact location and floor plans of buildings, location-based services (LBS), such as Google Maps (see Fig. 2 a) or Facebook (see Fig. 2 b) are confident to determine the relative number of people visiting a specific building or space within a building in real-time. On their respective online platforms, hourly aggregated and normalized data is published as so-called *popular times* data (Google) [39] or *popular hours* data (Facebook) [40].

The resulting data has the same structure as the standard schedules published by ASHRAE and others for building energy demand simulation. I.e., typical 24-h profiles for each day of the week.

We assume that a significant part of the population might be captured by LBS data due to the prevalence of smartphones and the popularity of the Google Maps and Facebook applications. According to a survey, 81% of U.S. adults own a smartphone. In terms of demographics, ownership varies from 53% to 96%. Smartphone ownership is common (>70%) among people from all economic, educational, and ethnic backgrounds. The only major demographic group with a lower smartphone ownership rate (53%) are people aged 65 or older [41]. Among smartphone users, Google Maps is by far the most popular mapping and navigation app with around 70% market share [42, 43], even though the market share of iOS and Android in the U.S. is around 60% to 40% in favor of iOS [44]. Also Facebook is popular in the U.S. and reaches around 60-80% of people across all demographics, except for people aged 65 or older [45].

In [46], the authors already directly extracted such data for simulation of a supermarket in the UK, due to lacking standard schedule for a 24-h operating retail building [46]. In Japan, researchers used Google popular times data to estimate quasi-real-time energy demand in a commercial district in Tokyo. They used a geospatial statistical interpolation, to infer the popularity and energy demand of commercial buildings without available data [47]. In China, researchers collaborated with a large Chinese social media company to collect occupancy data for different building use types from mobile positioning requests. They extracted typical schedules and used them in EnergyPlus building simulations and calibrations [48, 49]. Recently, [50] used mobile phone based occupancy estimates in an urban-scale energy modeling case study in Boston.

Even though currently data about non-public buildings, such as offices and residential buildings, are not published anywhere, we speculate that this data might be collected in the same way as for public buildings and possibly will be available at some point in time.



Bova's Bakery

4.7 ★★★★★ · 786 reviews · $

Bakery

Family-owned bakeshop, opened in 1932, offering up breads, cookies & sandwiches 24 hours a day.

134 Salem St, Boston, MA 02113, USA

9W8V+3Q Boston, Massachusetts, USA

bovabakeryboston.net

+1 617-523-5601

Open now: Open 24 hours

Menu places.singleplatform.com

Add a label

Marina Bay Sands

@marinabaysands

About

10 Bayfront Avenue
Singapore

m.me/marinabaysands

Call 6688 8868

Always Open

Mission

We are dedicated to making Marina Bay Sands the world's most magnificent destination for business, leisure and

Our Story

A premier entertainment desti
diversity of attractions and faci

Figure 2: Example of Google Maps information of a public place in Boston (a). The popular times feature displays normalized hourly visitorship for each day of the week. Example of a Facebook page of a public place in Singapore (a). The popular hours feature displays normalized hourly visitorship for each day of the week. Screenshots were taken on September 26, 2018.

1.5. Objectives and research questions

LBS data represents a large and global data-source with potential applications in urban energy simulations and beyond. Despite certain limitations (see section 6), we believe such data is representative for available data that can be used for the development of new methods for urban occupancy modeling. Our objective for this research is to establish a workflow that ultimately results in context-specific, representative building occupancy schedules. Such schedules then could be compared to one another using a general comparison metric, as well as using a set of energy-related features to qualitatively assess the impacts of using different schedules in energy demand simulations. From these departure points, our four research questions are:

Question 1: How can LBS data be used to establish a method to create context-specific, data-driven occupancy schedules? Question 2: How do the resulting contextual, *data-driven schedules* derived LBS compare to standard deterministic schedules, such as from ASHRAE? Question 3: How do the resulting data-driven schedules compare to one another with respect to geographical location and categorization of day types (e.g. weekends vs. weekdays)? Question 4: What are the potential implications of data-driven schedules on district energy systems' capacity and load profile?

We are using popular times LBS data published for *places* (i.e., commercial buildings or zones of commercial buildings) available on the Google Maps platform [51]. We focus on retail and restaurant use-type buildings, because these building use-types have a high density of available LBS data, and simultaneously these buildings typically have a high heating, ventilation, and air conditioning (HVAC) energy and electricity consumption. Retail buildings contribute the largest portion to overall commercial energy use in the U.S. and restaurants have the highest energy use intensity [52]. In both use-types HVAC is a major end use. The other major energy end use in such buildings, specifically the ones that handle food, is due to cooking and refrigeration [53].

Regarding the data collection, we are focusing on *downtown* areas of larger cities. They typically contain a high density of commercial buildings, and they presumably experience relatively similar characteristics of working, shopping, and leisure activities. We therefore expect a high availability of data and we hypothesize that the proposed approach will be able to detect significant differences in the data of such seemingly similar contexts. We are basing our study in the U.S. in order to compare LBS data-driven schedules to ASHRAE standard schedules for occupancy in the originally intended context.

This paper is organized as follows: The method section is split into three parts. Section 2.1 introduces the data collection and processing methods. Section 2.2 introduces the concept of earth mover's distance to compare schedules in general. And section 2.3 introduces the selected energy-related features of schedules, to be used to assess the potential impacts on building and district energy demand simulation results. Next, the selection of the case study locations is presented in section 3. In the results section, the data collection outcomes are first presented, together with the creation of average data-driven schedules based on the example of San Francisco in section 4.1. Next in section 4.2, context-specific data-driven schedules are compared to ASHRAE standard schedules, and in section 4.3 data of different locations and different days of the week is compared to one another. Section 5 contains a general discussion of the results, and section 6 contains the main limitations of the data source. Finally, the conclusions and an outlook are presented in sections 7 and 8.

2. Methods

Our method comprises three parts: 1) LBS data collection and data-driven schedule creation for any selected location, 2) using the *earth movers' distance* approach to quantify the difference between schedules, and 3) comparing schedules' *energy-related features* to assess the implications on energy demand and energy supply systems design. Parts 2) and 3) are used to compare data-driven schedules to ASHRAE standard schedules as well as to compare data-driven schedules of different locations. The three parts are introduced in detail in the subsequent sections 2.1, 2.2, and 2.3.

2.1. Methods to generate average, location-specific occupancy schedules

This section describes the LBS data collection and data processing prior to the analysis and comparison. The process is comprised of three steps: a) collection of popular times data and building use-type information at a given search location, b) categorization of places into the two studied building use-types (i.e. retail and restaurant), and c) creating a representative, data-driven schedule for each location and use-type (i.e. the mean or median of the data).

For the data collection, step a) we utilize an open-source python library [54] that connects to the Google Places application programming interface (API) [55], using its *nearby search* functionality. The nearby search query of the API requires the following parameters: 1) a geolocation, 2) a search radius, and 3) a *place-type*. The place-types that are supported by the search query are listed in [56]. See also Table A.8 in the appendix. Table 3 lists the place-types associated with the restaurant and retail building use-type according to our categorization. The API nearby search is used to obtain the names and addresses of places within the search radius. This information is then used to find the places on Google Maps and collect the popular times data. A multi-granular (i.e. larger and smaller search radius) data collection process was implemented to find as many places as possible within the defined search space[2]. The data used for this work consist of the list of place-types, and if available, the popular times 24-hour time series for each day of the week in percentage values. See Table A.9 in the appendix for an example.

A detailed classification into specific building use-types based on the existing attributes of place-types is not feasible. However, ASHRAE recommends using the same standard schedules for similar use-types [14]. For example, the retail schedule is suggested for library and museum buildings. See Table 3. The second step b), the categorization of places into ASHRAE building use-types, is based on this mapping of ASHRAE use-types to Google place-types. We used a simple categorization logic based on the count of place-types in the list belonging to the retail or restaurant use-type. In case of an equal count, the order of place-types, as returned by the Google Places API, was used as a

[2] The parameter values used to identify places were based on manual search experiments. During experiments with different search radii and supported place-types, as well as wildcard searches we realized that it is possible to find almost all places falling into the category of the restaurant use-type (see below) by searching for "restaurant", "bar", "cafe" and "night_club". For the retail use-type, the same can be achieved by searching for: "store", "shopping_mall", "museum", "library" and "art_gallery". See also Table 3. For each of the 9 place-types, a search with a 500m radius was executed in the specified area. In locations where the number of results might have exceeded the maximum number of results that the API can return (i.e. 60 places), a second search with a radius of 200m was executed in the same area.

secondary decision variable. This means that we used the first list entry corresponding to a place-type in Table 3 as the decisive one in case of a tie between restaurant and retail.

Table 3: Categorization of Google place-types to ASHRAE building use-types.

ASHRAE use-type	Google place-types		
RETAIL	art_gallery,^{x}	electronics_store,	museum,^{x}
Used for [14]:	bakery,	florist,	pet_store,
Library,	bicycle_store,	furniture_store,	pharmacy,
Museum,	book_store,	hardware_store,	shoe_store,
Retail	car_dealer,	home_goods_store,	shopping_mall,
	clothing_store,	jewelry_store,	store,^{x}
	convenience_store,	library,^{x}	supermarket^{x}
	department_store,	liquor_store,	
RESTAURANT	bar,^{x}		
Used for [14]:	cafe,^{x}		
Dining: Cafeteria/Fast Food,	restaurant,^{x}		
Dining: Family,	night_club,^{x}		
Dining: Bar Lounge/Leisure	meal_takeaway		

^{x} These are the Google place-types used in the nearby search query for the data collection.

The last step c) of the data-driven schedule creation consists of *averaging* the data to create the *representative* schedule of a building use-type for a specific location. For this step, the zero-data of days where places are closed or no popular times are displayed (e.g. "Not enough data yet for Tuesdays") were excluded.

In the context of standard schedules of occupancy, it is not clear whether this representative or average behavior is supposed to represent the mean of observations or rather the median. It seems that the originally estimated building occupancy schedules such as the SBOC mentioned above, were taking the mean of experts' estimations to create the first standard schedules of occupancy [9]. However when ASHRAE tasked researchers in the 1093-RP project [57] to compile schedules of lighting and receptacle loads in office buildings for energy and cooling load calculations, they advocated in favor of the median of measurements. In their work, the median was used to create DOE-2, BLAST, and EnergyPlus input files. The median was chosen over the mean because its value is not affected by outliers [57]. Furthermore, standard schedule values are often presented in quantized (rounded) steps of 5% or 10%. For example the ASHRAE standard schedule for retail occupancy is defined in steps of 10% and the restaurant occupancy schedule is defined in steps of 5%. To ensure a fair comparison with standard schedules (e.g. in terms of peak heights), we also explore data-driven schedules based on quantized means and medians in the analyses. We expect the quantization to have some impact on the number of occupied hours because low values of data-driven occupancy <5%, or <2.5% respectively, will be quantized to 0 and considered as non-occupied. In this work, we are performing all analyses for the quantized and non-quantized means and medians. However, we only present selected results in detail.

2.2. Methods to quantify the difference between schedules

We propose to use the concept of the earth mover's distance (EMD) for part 2) of the method, to quantify the difference between any two schedules (i.e. their general similarity or dissimilarity). The EMD was developed for comparing color histograms for image retrieval in computer science[3], where it is usually formulated as a transportation problem solved with linear optimization [60]. The next two subsections, will first introduce the concept, definition, and graphical explanation of the EMD method selected for our analysis. Second, we provide a comparison to other similarity metrics and highlight the EMD's benefits when compared to accuracy measures for time series comparisons, which are usually used in the field of building energy simulations.

2.2.1. Definition and graphical explanation of the EMD

To intuitively understand the EMD, one schedule is pictured as a mass of earth spread in space, while the other is pictured as a collection of holes in that same space. The EMD measures the least amount of work needed to fill the

[3] The same concept applied in mathematics to probability distributions is known as the 1st Wasserstein distance or Mallows distance [58]. For the formal definition in mathematics, see [59].

holes with earth, where one unit of work is quantified by transporting one unit of earth by one unit of *ground distance* in that space. The definition of ground distance is chosen according to the application. This original definition of EMD is only valid for two schedules with the same mass of earth (i.e. the same integral) [60]. To overcome the limitation of having to normalize schedules, or histograms in their case, Pele and Werman proposed a variant of the EMD for non-normalized histograms [61, 62]. In this variant, extra earth can be removed or added with a distance penalty. Given two histograms $P = \{P_1, \dots, P_m\}$, $Q = \{Q_1, \dots, Q_n\}$, they define $\widehat{EMD}$:

$$\widehat{EMD}_\alpha(P,Q) = (\min_{\{f_{ij}\}} \sum_{i,j} f_{ij} d_{ij}) + |\sum_i P_i - \sum_j Q_j| \times \alpha \max_{i,j}\{d_{ij}\} \; s.t \text{ Eq. } 2 \tag{1}$$

$$\sum_j f_{ij} \leq P_i \; , \; \sum_j f_{ij} \leq Q_j \; , \; \sum_{i,j} f_{i,j} = \min(\sum_i P_i, \sum_j Q_j) \; , \; f_{ij} \geq 0 \tag{2}$$

where $\{f_{ij}\}$ denotes the flows between P and Q. Each f_{ij} represents the amount of earth transported from bin i in P to bin j in Q. d_{ij} is the ground distance between bin i and bin j in the histograms. If the masses of P and Q are not equal, $\widehat{EMD}$ adds or removes mass such that both sides become equal. The transport distance for this added or removed mass to all other locations is set to be α times the maximum ground distance.

In our application of comparing schedules, the earth or mass $\{P_i\}$,$\{Q_j\}$ are the people present at different times of the day, and the ground distance d_{ij} is the clock-time, which means we are shifting people in time to transform one occupancy schedule into another. Because of the circular nature of clock time (i.e. the time of the day), we use the modular distance as the distance between two points in time. For the definition of ground distance $d_{ij} = D_{MOD}(i, j)$, see eq. 3. For example, when comparing two 24-hour schedules, the distance between 03:00 and 23:00 is 4 hours and not 20 hours. In this example, the maximum shift in time is therefore 12 hours or $\max_{i,j}\{d_{ij}\} = 12h$. We use this maximum distance value as the distance penalty associated with creating or removing people from the schedule ($\alpha = 1$).

$$D_{MOD}(i, j) = \min(|i - j|, N - |i - j|) \tag{3}$$

For the remainder of this paper we will use *EMD* to refer to $\widehat{EMD}_\alpha(P,Q)$ as defined in eq. 1 with ground distance $d_{ij} = D_{MOD}(i, j)$ and $\alpha = 1$. An efficient algorithm to compute this EMD is available for python [63]. With this definition of EMD, the absolute maximum value the metric can assume between two 24-hour schedules is 28,800 (24 values * 100% * 12 h penalty). This maximum value corresponds to the difference between a zero schedule, where every hour of the day is 0% occupied and a full occupancy schedule, where every hour of the day is 100% occupied.

To illustrate how the value of the EMD is calculated, we can look at an example of the difference between different versions of the ASHRAE occupancy schedules for restaurants on weekdays in Fig. 3. The occupancy in the morning hours was modified between the versions of 2004 and 2013, as mentioned above in section 1.2.

The majority of occupancy or earth in grey, remains the same. The green, purple, and orange occupancy is moved from 1 AM-3 AM to the time at 6 AM-8 AM. The work for moving this occupancy is 165%h (green: 5% * 5 h + 5% * 7 h = 60, purple: 10% * 5 h + 5% * 6 h = 80%h, orange: 5% * 5 h = 25%h) and a total of 90% (yellow) has to be created with a penalty of 12h, yielding an additional 1080%h to the total of 1245%h, corresponding to 4.3% of the maximum possible EMD between two 24-h schedules.

2.2.2. Comparison of the EMD to time series forecasting accuracy measures

We chose two examples to illustrate the benefits of using the EMD to compare two schedules. See Fig. 4. We argue that the difference (distance) between the schedules A and B (top left vs. bottom left) is maximal. On the other hand, the difference between the schedules C and D (top right vs. bottom right) is small, but not zero, as all of their values are just shifted by 1 hour of clock-time. Table 4 presents a comparison of the EMD to common accuracy measures for time series forecasting applied to the examples in Fig. 4. The selected accuracy measures are the root mean square error (RMSE), the mean squared error (MSE), the mean absolute error (MAE), the mean percentage error (MAPE), the symmetric mean percentage error (sMAPE), the mean bias error (MBE), and the coefficient of variation of the root mean squared error (CV(RMSE)). For definitions of these measures see [64].

As evident in Table 4, the MAE, the MSE, the RMSE, and the sMAPE give the same value for both examples and therefore, are not indicative of the differences we would like to quantify. The MAPE, MBE, CV(RMSE) are not

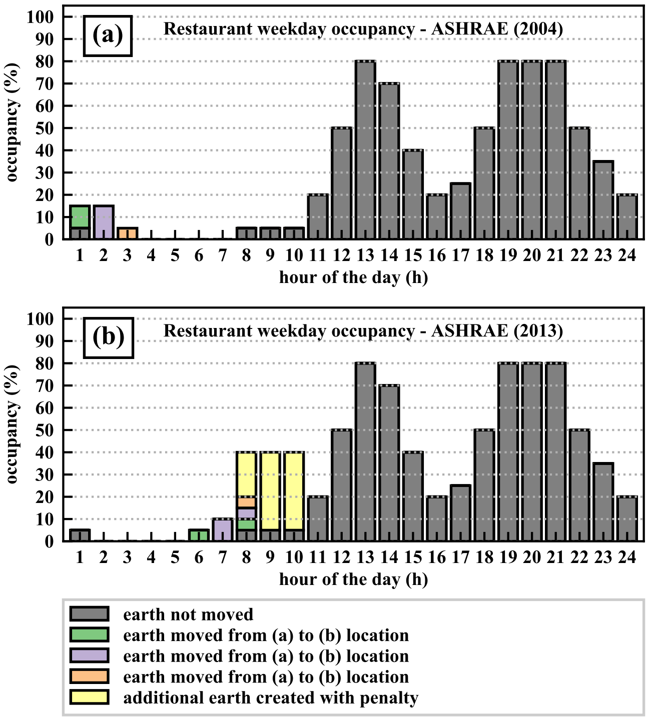


Figure 3: ASHRAE standard schedules of occupancy for restaurant on weekdays. The version of 2004 (a) is missing the additional occupancy in the morning hours that was added to the 2013 version (b). The colors of the bars illustrate the calculation of the EMD between the two schedules. The occupancy or earth moved to transform the schedules into each other is color-coded: grey is the occupancy that remains in place, yellow is the occupancy that is additionally created with a distance penalty of 12 hours, the other colors are parts of occupancy moved from 1-3 AM to 6-8 AM.

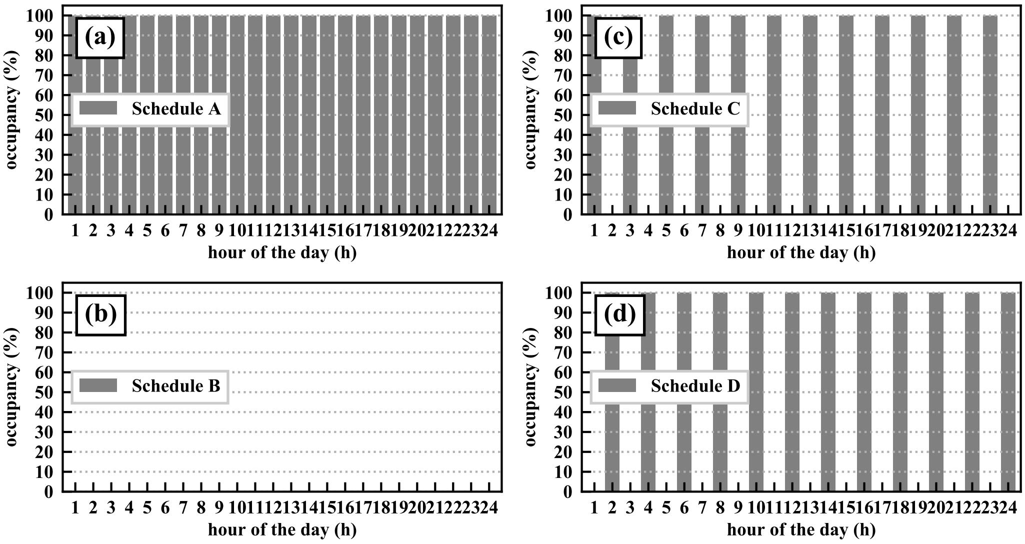


Figure 4: Examples to illustrate the usefulness of the EMD to assess the similarity or difference of two schedules. Schedule A (a) represents a full load schedule to be compared to the zero schedule B (b). Schedules C (c) and D (d) are two schedules alternating between zero and 100% that are shifted in time by one hour.

Table 4: Various accuracy measures for time series forecasting and the EMD as per definition in equations 1, 2, and 3 applied to the two illustrative examples in Fig. 4

	A vs. B	B vs. A	C vs. D	D vs. C
MAE	100	100	100	100
MSE	10000	10000	10000	10000
RMSE	100	100	100	100
MAPE	inf	1.0	inf	inf
sMAPE	2	2	2	2
MBE	inf	100	0	0
CV(RMSE)	inf	100	200	200
EMD	28800	28800	1200	1200

symmetrical and undefined in some cases. They can not be used to compare two schedules.

On the other hand, the EMD indicates what we would like to see. The maximum distance between an empty building and a fully occupied space (A vs. B) and a small but quantifiable distance between two schedules that only differ in their timing (C vs. D).

Because the EMD is symmetrical and considers changes in scale (adding or removing occupants) as well as changes in the dynamic behavior (moving occupants in time), we assume it does not directly translate to implications in energy demand. In the next section, we introduce methods to assess the relationships between differences in schedules and their potential energy implications.

2.3. Methods to assess potential energy implications of schedules

The third part of our method is to assess the potential energy implications of differences in schedules. Because schedules of occupancy are intended for energy simulations, their direct comparison might not adequately represent the potential differences or errors in simulation results. Therefore we define the following energy-related features to compare two schedules with respect to their potential impact on simulated building and district energy demand and energy supply systems design:

- The full load hours of a schedule,
- The occupied hours of a schedule,
- The maximum occupancy of a schedule,
- The maximum ramp gradients of a schedule, and
- The average daily peak times and peak values.

The number of full-load hours $FL(s)$ directly impacts the energy consumption of building systems that are occupancy-controlled. For example, ventilation systems that are controlled via the CO_2 concentration. The number of full-load hours also directly impacts the internal gains (sensible and latent) due to occupancy. Set-point controlled HVAC systems react to those internal gains. We calculate the full load hours of an occupancy schedule $s = \{s_1, \ldots, s_h, \ldots, s_n\}$ with Eq. 4.

$$FL(s) = \frac{\sum_{h=1}^{n} s_h}{100\%} \quad (h) \tag{4}$$

The number of occupied hours $OH(s)$ directly impacts the energy consumption of building systems that are binary presence controlled. For example, lights that are either on or off, or ventilation systems that are either on or off. The occupied hours of an occupancy schedule s are a summation of boolean values, represented with Iverson bracket notation, $[P] = 1$ if P is true; $[P] = 0$ otherwise, that are 1 when there is occupancy see eq. 5.

$$OH(s) = \sum_{h=1}^{n} [s_h > 0] \quad (h) \tag{5}$$

The maximum value of occupancy $MAX(s)$ is related to the expected peak of energy use for occupancy controlled building systems (e.g. ventilation) or set-point controlled building systems (e.g. latent and sensible cooling demand). The maximum value observed in the daily or weekly time series is defined as eq. (6).

$$MAX(s) = \max_{1 \leq h \leq n} s_h \quad (\%) \tag{6}$$

The maximum ramp gradients are defined here as the largest positive and negative change in hourly occupancy. Large ramps in occupancy can be indirectly related to stress in power grids or thermal networks via the sudden increase or decrease in energy demand of occupancy controlled building systems. For example, an occupancy controlled ventilation system suddenly has to provide much more fresh air to the building. We calculate the maximum ramps with Eq. (7) and (8).

$$MRU(s) = \max_{1 \leq h \leq n} s_{h+1} - s_h, \quad with \quad s_{n+1} = s_1 \quad (\%/h) \tag{7}$$

$$MRD(s) = \min_{1 \leq h \leq n} s_{h+1} - s_h, \quad with \quad s_{n+1} = s_1 \quad (\%/h) \tag{8}$$

Peak times and values are related to temporal energy demand patterns. Especially in cooling dominated climates, we assume that in most cases the peak of the sensible and latent cooling energy demand of setpoint-controlled building systems coincides with the peak of occupancy due to passive interactions of occupants with building systems. We use the peak finding algorithm of [65] to find local peaks in weekly or daily schedules. In signal processing, a local peak is commonly defined as a data sample that is larger than its two neighboring samples. Additional optional parameters in peak finding algorithms include required heights of peaks, threshold values, the minimum distances to neighboring peaks, and others [65, 66]. We use a minimum distance of 4 hours in cases where two local peaks of similar height are close to each other. We consider the center of a plateau peak as its peak time. For the comparison, we calculate the average daily peak time $\varnothing PT$ and the average daily peak value $\varnothing PV$, which we define as the means of peak times and values of all days of the week. In case schedules with two daily peaks, subscripts to indicate the first and second peak are added to the notation (e.g. $\varnothing PT_1$ and $\varnothing PT_2$).

3. Case study locations

As mentioned above, we based our study in downtown areas of U.S. cities in order to collect as much data as possible from the same context, for which the ASHRAE standard schedules were created. In [12] the DOE defined 16 representative cities of the U.S. climate zones for their commercial reference building models. We selected the same locations to collect data for potential future studies on climate and occupancy interactions and their impact on energy demand simulation results. The 13 largest cities are Miami, Houston, Phoenix, Atlanta, Los Angeles, Las Vegas, San Francisco, Baltimore, Albuquerque, Seattle, Chicago, Denver, and Minneapolis. The three remaining cities in cold climates (e.g. Helena, Duluth, and Fairbanks) were not considered due to their small size. Although this method is generally applicable, we expect that not enough data for meaningful statistical analysis can currently be collected in smaller cities. Our results (see section 4.1) show that in Albuquerque, a city with more than half a million inhabitants, only 91 data points for retail places could be found. The populations of Helena, Duluth, and Fairbanks are all below 100,000 people [67].

We collected data from the Downtown areas of the 13 large cities. We selected areas of 4km x 4km around a central location in each city. These locations were extracted from geographical features representing the business or financial district in the downtown area of each city on Google Maps [51]. Table 5 lists the edges of search areas within each city, as well as the name of the Google Maps data anchor point in the search area.

4. Results

4.1. Generation of location-specific, data-driven occupancy schedules

In this section, the results of the LBS data collection and processing are presented. First, an overview of the available data from the 13 search locations is provided. Then, on the example of San Francisco, research question 1 is addressed: “How can LBS data be used to establish a method to create context-specific occupancy schedules?”

Table 5: Area selection for the search of LBS data. All data were collected on March 12, 2019. The cities are ordered according to climate zone. The climate zones are depicted in [12]. They range from 1 (very hot), 2 (hot), 3 (warm), 4 (mixed), 5 (cool), to 6 (cold), with subdivisions into A (moist), B (dry) and C (marine) zones. Zones 7 (very cold) and 8 (subarctic) are not considered here.

abbreviation	climate zone	Google Place name for center coordinates	southwest edge of search area (lat,lng)	northeast edge of search area (lat,lng)
MIA	1A	Downtown Miami, Miami, FL, USA	(25.753197,-80.211808)	(25.789303,-80.171932)
HOU	2A	Central Business District, Houston, TX, USA	(29.734722,-95.385841)	(29.770807,-95.344487)
PHX	2B	Downtown Phoenix, Phoenix, AZ, USA	(33.433683,-112.095971)	(33.469748,-112.052948)
ATL	3A	Downtown, Atlanta, GA, USA	(33.737678,-84.409954)	(33.773740,-84.366780)
LA	3B-CA	Financial District, Los Angeles, CA, USA	(34.032496,-118.278705)	(34.068557,-118.235382)
LV	3B-other	Downtown, Las Vegas, NV, USA	(36.149033,-115.157778)	(36.185081,-115.113321)
SF	3C	Financial District, San Francisco, CA, USA	(37.776553,-122.422646)	(37.812591,-122.377232)
BAL	4A	Downtown, Baltimore, MD, USA	(39.273990,-76.639899)	(39.310019,-76.593532)
ABQ	4B	Downtown, Albuquerque, NM, USA	(35.073642,-106.677586)	(35.109696,-106.633720)
SEA	4C	Downtown, Seattle, WA, USA	(47.587033,-122.360960)	(47.623010,-122.307764)
CHI	5A	Chicago Loop, Chicago, IL, USA	(41.860626,-87.649142)	(41.896639,-87.600954)
DEN	5B	Central Business District, Denver, CO, USA	(39.726778,-105.017710)	(39.762804,-104.971041)
MIN	6A	Downtown West, Minneapolis, MN, USA	(44.956211,-93.298675)	(44.992204,-93.247966)

4.1.1. Data availability, collection, and processing

Fig. 5a shows the number of retail and restaurant places found in each city after categorization. Fig. 5b shows the share of retail and restaurant places in each city with available popular times data. The smallest sample of 91 retail places with popular times data was collected in Albuquerque, and the largest sample of 1250 restaurant places was in San Francisco. On average, popular times data was available for 54% of restaurant and 13% of retail locations in the investigated downtown areas. In this result section and the remainder of this paper we will refer to all hourly statistics extracted from the LBS data set as *data-driven occupancy*.

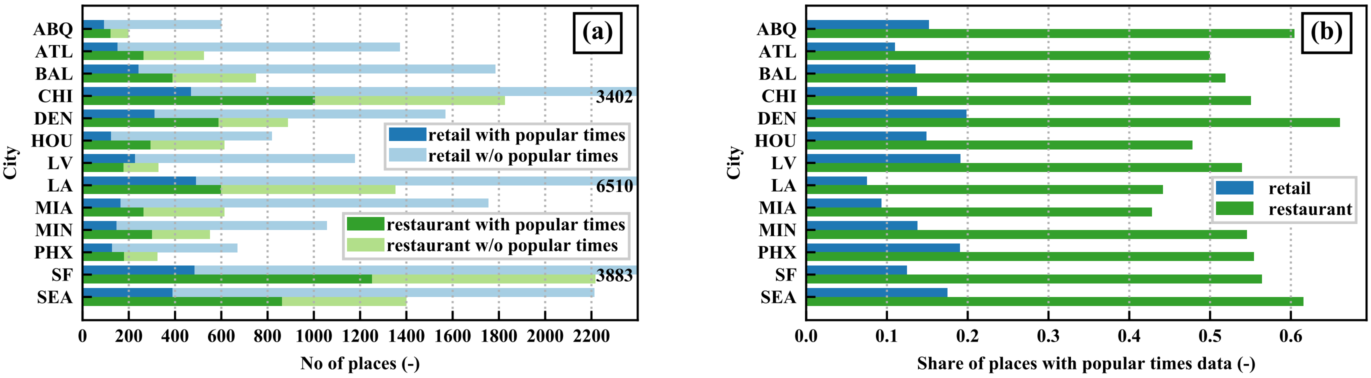


Figure 5: Number of retail and restaurant places with and without popular times data displayed in a 4km by 4km area in the vicinity of downtown neighbourhoods in different cities in the U.S. (a) and relative share of places with available popular times data relative to the total number of places (b).

4.1.2. Weekly data-driven occupancy schedules

Fig. 6 shows the data-driven occupancy schedule for retail places in San Francisco. The graph shows all four variants of the data-driven schedule (i.e. the mean and the median, as well as the the quantized mean and the quantized median). The quantization is 10%, like the ASHRAE retail standard schedules.

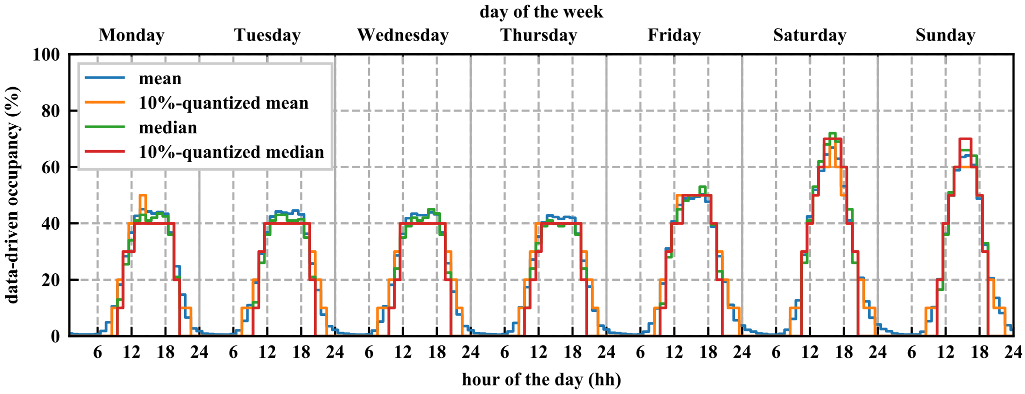


Figure 6: Data-driven retail occupancy schedule for retail locations in and around the Financial District in San Francisco. The mean, the quantized mean, the median, and the quantized median is shown. For the quantization a step of 10% is used analog to the ASHRAE Retail standard schedule.

From Fig. 6, we can observe that retail places in San Francisco follow a regular pattern with one peak per day in the afternoon. The maximum values of around 70% are observed on Saturdays and Sundays in the early afternoon. From Monday to Thursday, the pattern looks very similar. On Fridays, a higher peak is observed compared to the other weekdays.

Table 6 lists the energy-related features of the four different variants of the data-driven schedule. Choosing the mean instead of the median results in significantly more full load hours, more occupied hours, and less steep ramp gradients. Quantization only impacts the mean occupied hours significantly. This is due to the distribution of the data at night time and early morning, where the median is usually zero, but the mean is still larger than zero but smaller than 10%.

Table 6: Energy-related features of the four variants of the data-driven, weekly retail occupancy schedule for San Francisco. The data is depicted in Fig. 6.

data-driven schedule variant	*FL* (h/w)	*OH* (h/w)	*MAX* (%)	*MRU* (%/h)	*MRD* (%/h)	∅*PT* (hh)	∅*PV* (%)
mean	34.09	168	67	+16	-16	15.9	51
10%-quantized mean	33.2	102	70	+20	-20	15.4	50
median	29.11	74	72	+26	-33	16.4	52
10%-quantized median	28.8	74	70	+30	-30	15.8	50

Fig. 7 shows the variants of the data-driven restaurant occupancy schedule in San Francisco. The quantization is 5%, like the ASHRAE restaurant schedules.

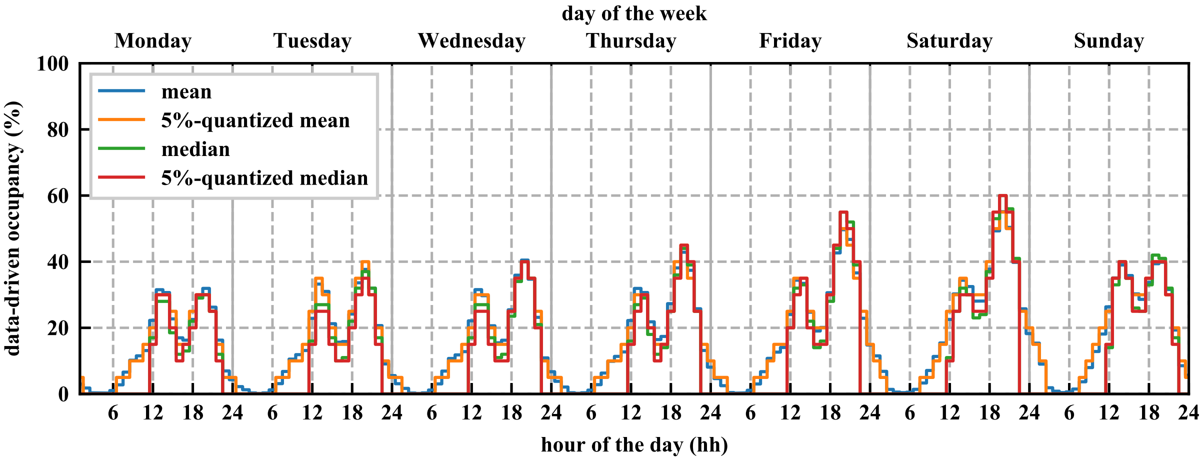


Figure 7: Data-driven restaurant occupancy schedule for the Financial District in San Francisco. The mean, the quantized mean, the median, and the quantized median is shown. For the quantization a step of 5% is used analog to the ASHRAE Restaurant standard schedule.

The data-driven restaurant occupancy schedule for San Francisco contains two peaks per day, see Fig. 7. The first peak is observed around noon, and the second peak happens in the evening. A maximum value of around 60% is observed on Saturday evening. From Monday to Saturday, the second peak of the day seems to increase in magnitude from around 30% to 60% for subsequent days, while the first peak value of the day remains between 25% and 35% depending on the day of the week and the quantization. On Sunday, both peaks are around 40%.

Table 7 lists the energy-related features of the four different variants of the data-driven schedule. The findings are similar to retail (above). The impact of quantization on mean occupied hours per week is less drastic as for the retail data because the quantization step is only 5%.

In order to provide a more comprehensive analysis across all cities, the following section 4.2 deals with the EMDs between data-driven schedules and standard schedules, as well as potential energy simulation impacts caused by these differences.

Table 7: Energy-related features of the four variants of the data-driven, weekly restaurant occupancy schedule for San Francisco. The data is depicted in Fig. 7

data-driven schedule variant	FL	OH	MAX	MRU	MRD	$\varnothing PT_1$	$\varnothing PV_1$	$\varnothing PT_2$	$\varnothing PT_2$
	(h/w)	(h/w)	(%)	(%/h)	(%/h)	(hh)	(%)	(hh)	(%)
mean	28.46	168	55	+12	-14	13.3	34	20.0	43
quantized mean (5%)	28.25	133	55	+15	-15	13.6	34	19.9	43
median	21.17	77	60	+20	-41	13.8	31	19.9	44
quantized median (5%)	20.85	77	60	+20	-40	13.9	31	19.9	44

4.2. *Comparison of location-specific schedules to standard schedules*

In this section, the EMD between data-driven and standard schedules is calculated to address research question 2 "How do the contextual, data-driven schedules derived from LBS, compare to standard deterministic schedules, such as from ASHRAE?", and in part question 4 "What are potential implications of data-driven schedules district energy systems' capacity and load profile?" First, weekly data-driven schedules of all 13 cities are compared to ASHRAE standard schedules of occupancy. Second, energy-related differences between the quantized mean data-driven schedules and standard schedules are discussed for all locations.

4.2.1. *EMD between data-driven schedules and standard schedules*

Fig. 8a shows the EMD between standard schedules of occupancy (ASHRAE schedules 2013 [14]) and the four variants of data-driven retail schedules of all 13 case study locations. The EMD for the week is calculated as the sum of EMDs for each day of the week. Fig. 8b shows the same information for the restaurant building use-type.

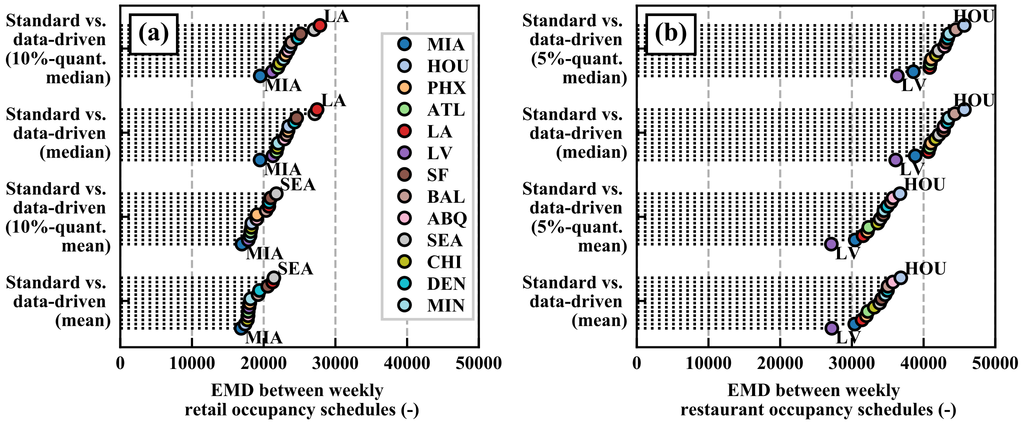


Figure 8: Earth mover's distance between the ASHRAE weekly standard schedule of retail (a) and restaurant (b) occupancy and the data-driven schedules of 13 cities in the U.S.

On average for both use-types, the mean data is closer to the standard schedules in terms of EMD. For retail, the range of values lies between around 17,000 to 22,000 (mean) and between 20,000 to 28,000 (median), corresponding to 8.4% to 13.8% of the maximum possible EMD between two 7x24-hour schedules. See Fig. 8a. For all four data-driven schedule variants, the Miami data has the smallest difference to the standard occupancy schedule. When taking the data mean, Seattle has the largest EMD. When taking the median, Los Angeles has the largest EMD, closely followed by Seattle. The impacts of quantization to 10% steps are marginal on the EMD calculation. Although the relative ranking of locations is impacted, absolute EMD values only change slightly, and the smallest, as well as the largest differences, remain the same for the mean and the median of the data.

For restaurants, the range of values lies between 27,000 to 37,000 (mean) and 36,000 to 46,000 (median), corresponding to 13.5% and 22.7% of the maximum possible EMD between two 7x24-hour schedules. See Fig. 8b. For all four data-driven schedule variants, the Las Vegas data has the smallest difference to the standard occupancy schedule, and Houston data has the largest difference to the standard. The impacts of quantization to 5% steps are marginal on the EMD calculation. The relative ranking of locations is impacted only once, between quantized and non-quantized median data.

As an example of data-driven retail schedules, Fig. 9 shows the quantized mean popular times data for Miami (smallest EMD to the standard), Seattle (largest EMD to the standard), and the ASHRAE standard schedule of retail occupancy. Fig. 10 shows the same example for restaurant data from Las Vegas (smallest EMD to standard), Houston (largest EMD to standard), and the ASHRAE standard schedule of restaurant occupancy.

As can be understood from Fig. 9 and 10, the largest contribution to the EMD stems from creating additional occupancy for the data-driven schedules. For the retail data-driven schedule, occupancy has to be added on weekdays and Saturday and removed on Sunday. For restaurants, occupancy has to be created for all days of the week. Notably, the locations with the smallest differences to the standard schedules (i.e. Miami for retail and Las Vegas for restaurant) have larger cumulative occupancy when compared to other locations.

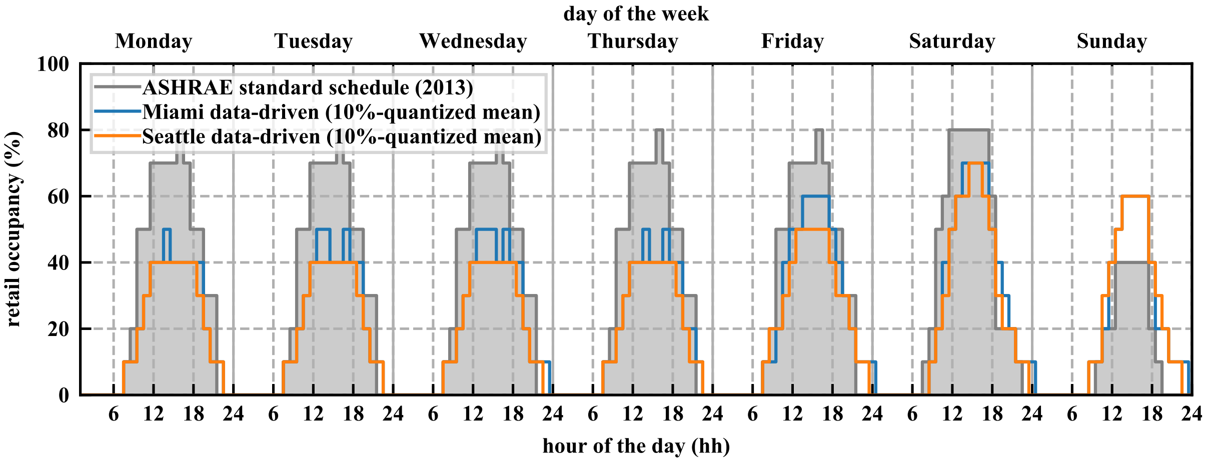


Figure 9: ASHRAE (2013) weekly standard schedule of retail occupancy (grey) and data-driven retail schedules (10%-quantized mean) for Miami (blue) and Seattle (orange).

4.2.2. Energy-related differences between data-driven schedules and standard schedules

In this section, the comparisons between energy-related features of standard and data-driven occupancy schedules are presented to address the potential implications of data-driven schedules on energy demand and energy supply systems' capacity and load profile. We compare ASHRAE standard schedules to quantized mean data because the mean is closer in terms of EMD, and non-quantized mean data has an unrealistically high number of occupied hours per week.

Fig. 11 shows the different energy-related features of context-specific data-driven retail schedules and the ASHRAE standard schedule for retail occupancy. Clear trends are visible: The standard schedules seem to overestimate full-load hours, but underestimate occupied hours. In the standard, maximum ramp-up gradients are slightly overestimated (+30%/h vs. +20%/h), while maximum ramp-down gradients are significantly overestimated. The average time of the daily peak seems to be observed a bit earlier than assumed, and the values are smaller than in the standard. The weekly maximum observed in the mean data is a bit lower than estimated.

Fig. 12 shows the comparison of all energy-related features of data-driven restaurant schedules (5%-quantized mean) to the standard schedule across all 13 case-study locations. The trends are mainly the same as for the retail building use-type, except for the occupied hours.

In general, the number of occupied hours per week seems well estimated by the standard restaurant schedule. The outliers are Albuquerque, with lower occupied hours and Las Vegas, with higher occupied hours. In contrast, the cumulative occupancy measured in full load hours is much lower in the data compared to the standard schedule.

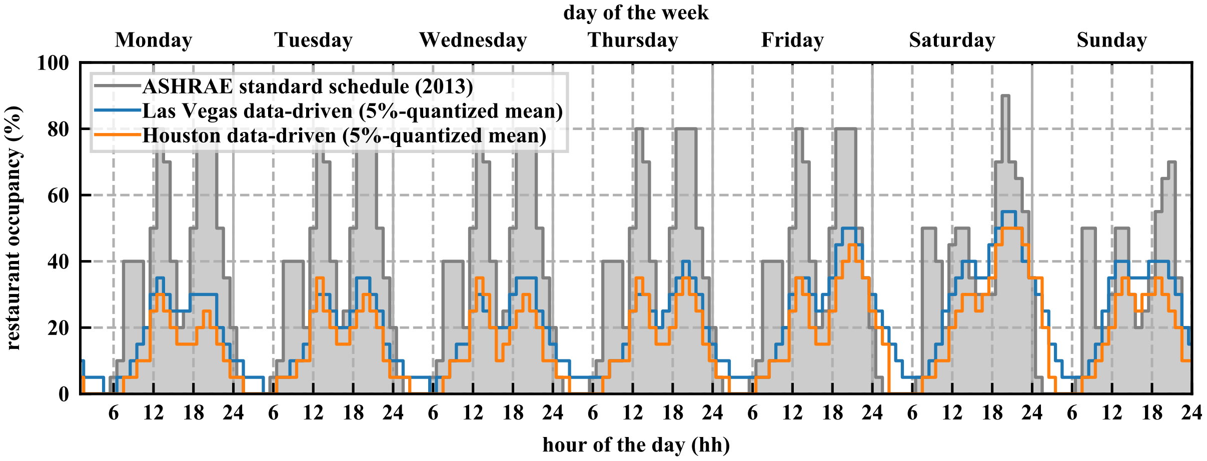


Figure 10: ASHRAE (2013) weekly standard schedule of restaurant occupancy (grey) and data-driven restaurant schedules (5%-quantized mean) for Las Vegas (blue) and Houston (orange).

It is roughly half of the full load hours assumed by ASHRAE for restaurants. The maximum value of occupancy, according to the standard schedule, is 90%. In the data, we observe maxima of 50 to 60%.

The data-driven restaurant schedules have much smoother ramps. The maximum ramp-up is +15%/h, compared to +45%/h in the standard schedule. Likewise, the maximum ramp down is -15%/h compared to -35%/h. The timing of the two peaks around noon and evening is similar in the standard schedule and the data-driven schedules. Generally, the noon peak happens a bit later compared to the standard, while the evening peak happens a bit earlier in the day. There are some exceptions, however the differences in peak times are smaller than ±1 h.

4.3. Pairwise comparison of location-specific schedules

In this last results section, we are comparing data-driven schedules from different locations to address research question 3: "How do the resulting data-driven schedules compare to one another with respect to location and categorization of day types?" and the follow-up research question 4: "What are the potential implications of data-driven schedules on district energy systems capacity and load profile?". First, we present a pairwise comparison in terms of EMD and energy-related features between weekly data-driven schedules for different locations. Second, we present a pairwise comparison of daily data-driven schedules in terms of EMD between the different days of the week for all locations.

4.3.1. Differences between data-driven schedules for different locations

Fig. 13a shows the pairwise EMD between weekly data-driven (mean) retail schedules in the 13 case-study cities in the U.S. The values of EMD range between around 1,400 to 8,200, corresponding to 0.7% to 4.1% of the maximum possible EMD between two 7x24-hour schedules. The differences between context-specific, data-driven retail schedules are around 3 to 12 times smaller, as compared to the differences between standard and data-driven schedules, with values of around 17,000 to 22,000 (mean). See Fig. 8a above.

In Fig. 13a we can observe that for the retail building use-type, Miami seems to experience a distinct behavior from other cities. Miami has the largest differences to 9 out of 12 other cities. The exceptions are Las Vegas (largest difference to Phoenix), Atlanta (largest difference to Seattle), and Chicago (largest difference to Phoenix). The overall largest difference is observed between Miami and Seattle. The smallest difference is observed between the mean behavior of Phoenix and Houston, followed by Baltimore and Denver.

Fig. 13b shows the pairwise EMD between weekly data-driven (mean) restaurant schedules for the 13 case-study cities in the U.S. The values of EMD range from around 1,200 to 11,000, corresponding to 0.6% to 5.4% of the

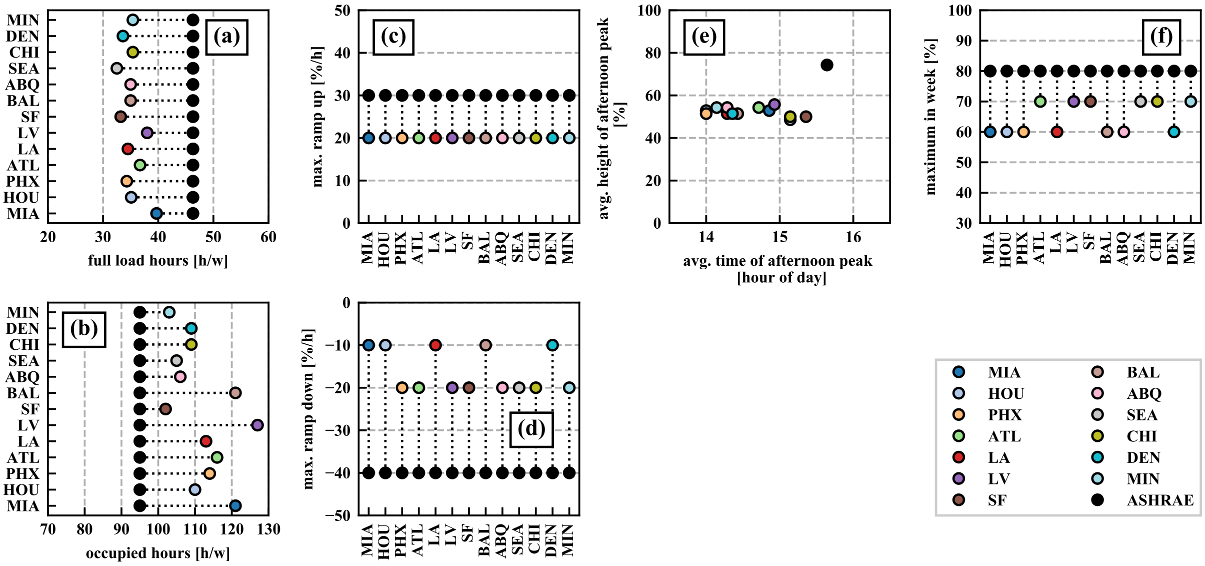


Figure 11: Lollipop and scatter plots of energy-related features of the standard (ASHRAE 2013) retail occupancy schedule for one week (black) and data-driven schedules (quantized mean) of the 13 case-study cities for one week (colored). The individual graphs show the comparison of full load hours (a), occupied hours (b), ramp gradients (c) and (d), the average time and value of the daily peak (e), and the weekly maximum value (f). Chicago has two peaks close to each other on Tuesday, Wednesday, and Thursday around 14h and 17h. The average peak time of those days was calculated as the average time between the first and the second peak.

maximum possible difference between two 7x24-hour schedules. The differences between context-specific, data-driven restaurant schedules are around 3 to 22 times smaller, as compared to the differences between data-driven and standard restaurant schedules, with values of around 27,000 to 37,000 (mean). See Fig. 8b above.

From Fig. 13b, we can observe that for the restaurant building use-type, Las Vegas seems to experience a distinctly different behavior compared to most other cities. Las Vegas has the largest EMD to 9 out of the 12 other cities. The exceptions are Los Angeles (largest difference to Houston), Atlanta (largest difference to Albuquerque), and Miami (largest difference to Albuquerque). The smallest difference is observed between Seattle and San Francisco, followed by Baltimore and Denver, and Baltimore and Minneapolis.

Energy-related differences between data-driven schedules can be extracted from Fig. 11 and 12 above. See section 4.2 for additional details. They are generally small, but there are some exceptions. For restaurants, these exceptions are the difference in the number of occupied hours between Las Vegas and Albuquerque, the spread in full load hours among the cities, and to some extent, the timing of the afternoon peak in Miami when compared to other cities. The values of full load hours in the data-driven schedules (5%-quantized mean) range from 26.5 to 35.1 h/week, corresponding to a factor of 1.32 between the smallest and the largest value. Meaning that for example, using Las Vegas data to simulate Houston, would lead to an overestimation of full load hours by 32%. The values of occupied hours range from 122 to 163 h/week, corresponding to a factor of 1.34. Meaning that for example, when using Las Vegas data to simulate the energy demand in Albuquerque, the number of occupied hours would be overestimated by 34%.

Notably, this means that rather than the peak loads or the shape of demand patterns, the annual energy demand prediction is likely to be most affected by the differences between geographic locations.

For the retail building use-type, energy-related differences can be extracted from Fig. 11 in section 4.2.2 above. The data-driven schedules of different locations are very similar with regards to mean peak values, peak times, and ramp gradients. However, full load hours range from 32.5 to 39.7 h/week, corresponding to a factor of 1.22 from the smallest to the largest value. The values of occupied hours range from 102 to 127 h/week (i.e. factor 1.25). Like for the restaurant use-type above, these differences could translate to significant differences in annual energy demand in

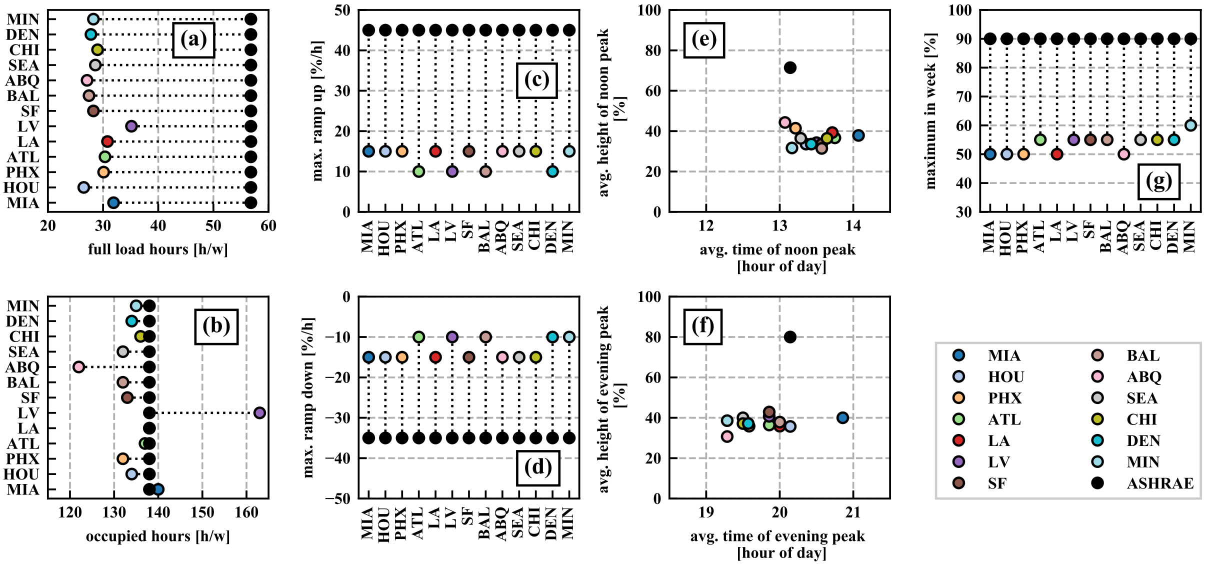


Figure 12: Lollipop and scatter plots of energy-related features of the restaurant standard (ASHRAE 2013) occupancy schedule for one week (black) and data-driven schedules (quantized mean) of the 13 case-study cities for one week (colored). The individual graphs show the comparison of full load hours (a), occupied hours (b), ramp gradients (c) and (d), the average time and value of the daily peaks (e) and (f), and the weekly maximum value (g). PHX and LA quantized mean data does not have an evening peak on Sunday. ATL and MIN quantized mean data does not have a noon peak on Saturday. For those four cities, the average shown is the mean of the 6 other days of the week.

some cases.

4.3.2. Differences between data-driven schedules of different days of the week

Fig.14ab shows the pairwise EMD between data-driven (mean) *daily* retail schedules of the 13 case-study cities (color-coded) in the U.S. The weekdays from Monday to Thursday are compared against each other in Fig. 14a. In Fig. 14b, the data-driven Friday schedules are compared against every other day of the week. The daily data-driven retail schedules have smaller differences and a smaller spread among different cities from Monday to Thursday. The Friday data has large differences to all other days of the week and a larger spread among different cities.

Fig. 14cd shows the same comparison for restaurants. The weekdays from Monday to Thursday are compared against each other in Fig. 14c. In Fig. 14d, data-driven Friday schedules are compared to every other day of the week. In Fig. 14c, we observe small differences between data-driven schedules of adjacent weekdays from Monday to Thursday (e.g. Monday vs. Tuesday, Tuesday vs. Wednesday, and Wednesday vs. Thursday). Larger differences are observed between non-adjacent weekdays from Monday to Thursday. In Fig. 14d, large differences between Friday data and other days of the week are evident. For some locations, Friday is closer to Saturday and Sunday than to other weekdays.

Fig. 15 shows two examples for the restaurant use-type of the best (a) and worst (b) agreement between data-driven (mean) Friday schedules and schedules of other weekdays. Left is the data of San Francisco for Friday and Thursday, which is the smallest difference between Friday and another weekday, observed in Fig. 14d. On the right is the data of Chicago for Friday and Monday, which is the largest difference observed in the same figure. In general, the data-driven daily full load hours and the peak values are larger on Friday. This means that using average weekday data to model Fridays, would likely lead to a misprediction of daily energy demands sensitive to cumulative occupancy (e.g. via internal loads), as well as a misprediction of energy demands sensitive to the evening peak of occupancy.

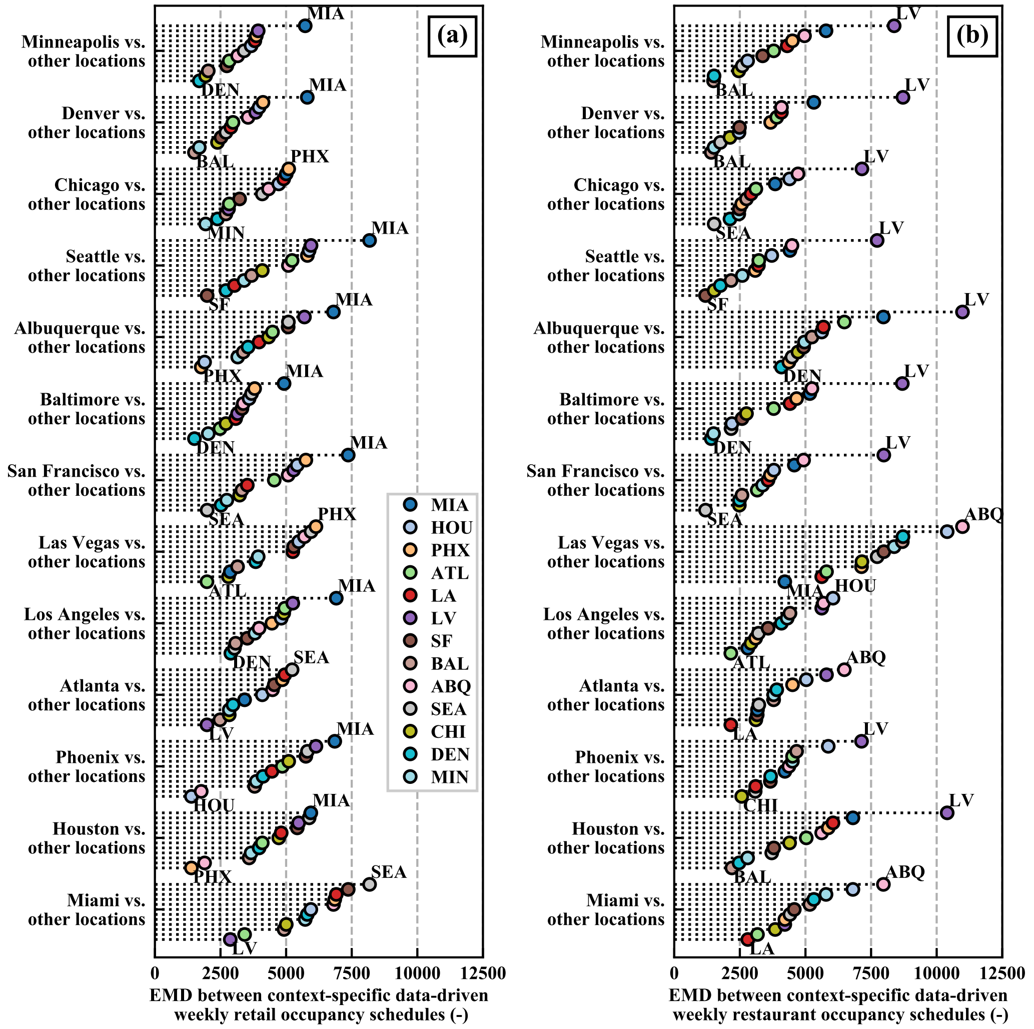


Figure 13: Pairwise earth mover's distances between weekly data-driven (mean) retail occupancy schedules (a) and restaurant occupancy schedules (b) of the 13 case-study cities in the U.S.

5. Discussion

In this work we are using LBS to collect data from places that can be categorized as retail or restaurant building use-type. The data availability is highly variable across the selected cities in areas of equal size. This is due to parameters such as urban density, use mix of neighborhoods, and of course, the geographical shape of the city (i.e. not every downtown fits into a square area). Despite this, it seems possible to collect enough data from large cities to run meaningful statistical analyses and compare locations to one another. We assume that the reason for a place not displaying popular times data is that there is insufficient visit data available. For a discussion on the potential biases introduced by considering only places with available data, we refer to the limitations in section 6. In this initial exploratory analysis, we are interested in the representative data-driven schedule of occupancy for each geographic location and use-type for direct comparison with the status-quo of occupancy modeling, which especially on the district- and urban-scale use ASHRAE standard schedules [23, 50].

The choice of mean vs. median and quantization to 10% vs. 5% vs. no quantization for the creation of data-driven schedules has an impact on the comparison results. The examples in section 4.1 reveal that the largest influences on

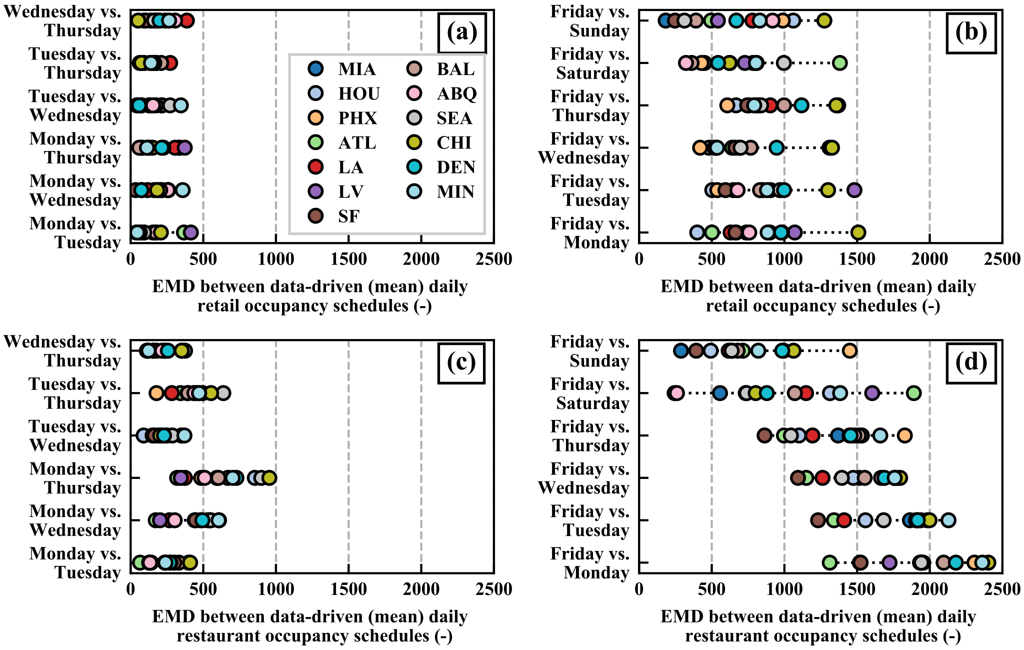


Figure 14: Pairwise EMDs between daily data-driven (mean) retail schedules (top row). Comparison of weekdays from Monday to Thursday (a) and Friday to all other days of the week (b). Pairwise EMDs between daily data-driven (mean) restaurant schedules (bottom row). Comparison of weekdays from Monday to Thursday (c) and Friday to all other days (d). Different colors represent different cities.

energy-related features concern the number of full load hours, occupied hours, and the ramp gradients. The results of these feature comparisons should therefore be investigated further for additional verification. Regarding the use of EMD for the comparison between standard schedules and data-driven schedules, we can observe that the creation of additional occupancy dominates the value of EMD. The reason being that the data-driven cumulative mean or median occupancy is much lower than the ASHRAE standard assumptions. These general findings agree with the results in [50] who used mobile phone data to model building occupancy.

In general, ASHRAE standard schedules seem to overestimate features that influence the annual and peak energy demand of buildings. On the building level, this overestimation might be beneficial because it conservatively estimates energy demands for building HVAC system sizing, especially for cooling loads. For heating systems, the opposite is the case. If occupancy-related internal gains are overestimated, the necessary heating energy demand might be underestimated. On the building-scale, such an underestimation of heating energy demand will not impact the heating system sizing, because the winter design conditions assume zero occupancy and other heat gains. See for example the design schedules in [12]. On the other hand, this overestimation of occupancy is potentially detrimental for district energy infrastructure design and sizing, since the resulting over- or underestimations of energy demands are cumulative.

Another interesting finding is that it seems that the standard assumptions generally estimate the shape of the profile correctly: A *triangle-shape* for the retail use-type with a peak in the afternoon, and a *M-shape* for the restaurant use-type with a first peak at lunchtime and second one in the evening. However, trends in the collected data did not corroborate the *breakfast peak* introduced by ASHRAE in 2011/2013, see Fig. 1 and 3.

Using data-driven schedules instead of standard schedules of occupancy for building and district energy demand simulations would result in two main consequences. First, the large differences in full load hours would impact the annual energy demand of applications, whose energy demand are sensitive to occupancy, such as demand-controlled ventilation, heating, and cooling. Second, the substantial differences in maximum occupancy values would impact the peak power requirement prediction of those energy demands and lead to different sizing of supply and distribution systems. The relatively small differences in occupied hours mean that occupant-presence controlled building systems,

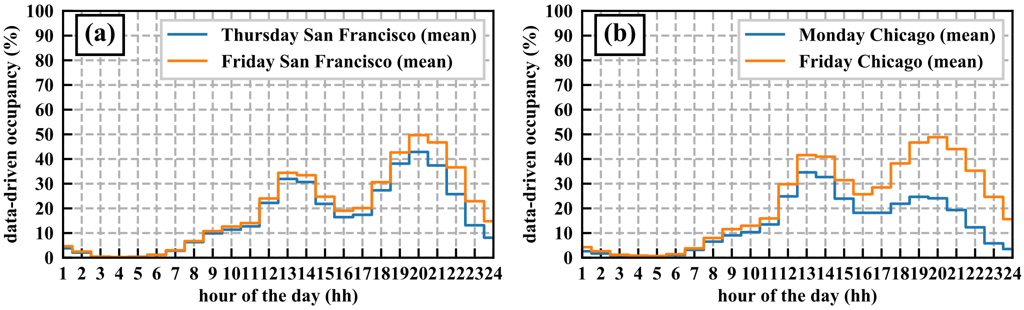


Figure 15: Examples of the smallest (a) and largest (b) differences between daily data-driven (mean) restaurant occupancy schedules for Friday and for other weekdays according to the EMD value. See Fig. 14d. Left (a) is San Francisco Thursday and Friday data, right (b) is Chicago Monday and Friday data.

such as lighting systems or ventilation systems with a constant operation, will result in similar annual energy demand when simulated with data-driven schedules and standard schedules. The shape of energy demand patterns of buildings and districts would look relatively similar when simulated with data-driven schedules and standard schedules, because the timing of the peaks is generally well estimated in the standard schedules. However due of the smaller ramps, they would look smoother. This smoother patterns will influence the steepness of the load duration curve of occupancy-sensitive energy demands.

Overall, EMD and other differences are much smaller when data-driven schedules from different locations are compared as opposed to a comparison with standard schedules. That can be interpreted in a positive way because it might be possible the represent multiple locations with one updated data-driven standard schedule.

However when comparing weekly data from different locations, we can still find significant differences despite the general agreement. Notably, those differences are found in full load hours and occupied hours, rather than peak values and peak times. Meaning, that the annual energy demand of buildings is likely to be influenced by context-specific differences, while peak demands would remain relatively similar, given similar climatic conditions.

Finally when comparing data-driven schedules from different days of the week, it seems that Friday behavior is significantly different from other weekdays for both considered use-types. The common categorization of standard schedules into three day-types: Weekday, Saturday, and Sunday is therefore questionable for these building use-types and should be further investigated.

6. Limitations

The process of conducting the novel research presented in this paper has raised a number of additional considerations for future research mainly related to the non-transparency of the data source. We found no publicly available information on the data collection, data processing, and data publishing processes. This suggests the possibility of skewed or biased results and should be considered if raw data, or similar data from other sources, becomes available in the future. With regards to data collection, potential data biases include: A bias towards places with larger capacity due to the higher probability of data collection, and a bias towards places visited by a particular population demographic due to a higher probability of cell phone ownership or higher penetration of location data collecting applications.

The shares of users opting in or out of collecting their location history [68] is not known. However, we are not aware of potential reasons that might significantly impact the shares among different demographics. Overall it can be assumed that location data can be collected from a large part of the population across all demographics, with a potential exception of older people due to lower smartphone ownership (see section 1.4). LBS data might, therefore, underestimate occupancy in places that cater to an older demographic or underestimate occupancy at times when a significant share of older people is visiting retail or restaurant buildings.

Verifying the data collected by location-based services is very challenging, as it would require the setup of large scale data collection at many locations over a long time, for example by exploiting data collected by building management system (BMS), access controls systems, or video-based people counting systems. In addition to technical feasibility, this would raise massive privacy concerns.

With regards to data processing and publication the current identified challenges are: The potential seasonal effects that might be or might not be reflected in the published data, and the estimation of the of 100% popularity or occupancy value, which is based solely on historic trends of recorded visits and might be inconsistent with standard practices of occupant density and capacity estimations. In ASHRAE and other standard schedules the 100% occupancy value corresponds to the number of people expected at *design conditions*. According to the considered standard schedules in this work, these design conditions are not expected to occur on a regular weekly basis. The maximum occupancy in the retail and restaurant schedule is 80% and 90% respectively. See Fig. 9 and Fig. 10. In the popular times data the 100% corresponds to actually observed typical peak occupancy conditions. Google states that "the typical peak popularity for the business for the week" is based on data which is collected "over the last several weeks" [39]. Therefore, it is likely that the 100% in popular times corresponds to the situation where the standard schedules assume 80%–90% occupancy.

Additionally, with regards to our presented methodology, we are aware that simply extracting the mean and median of the data aggregated into use-type and day-type categories, does not necessarily account for the potential diversity of patterns within these categories. To further refine the method, additional categories could be formed by using classification and clustering techniques on the LBS data directly. Such a method could be based on the EMD approach proposed in this paper.

Finally as discussed in detail in early sections of the paper, the results of the quantized mean of data are presented in comparison to the standard schedules. However conceptually, the median could represent a better choice as it can be understood as the actual behavior of the place in the center of the sample. The issue with selecting the mean is not only that outliers disproportionately influence the values, but also that it is unlikely in terms of occupancy to observe hours with zero mean occupancy in a large sample. This can be somewhat avoided by the quantization to 5% or 10% steps as in the comparison to the standard schedules.

7. Summary and Conclusions

Standard schedules of occupancy are a by-product of the development of simulation-based building performance evaluation. Some schedules still being used today are unchanged since as back as 1979 and were not created from verifiable, observed data. Since then, schedules of occupancy have partly been updated, but have never been based on systematic data collection. Notably despite newer modeling approaches, these unchanged schedules are still pervasively used in practice.

In this work, we demonstrate that we can utilize LBS data to create contextual building occupancy schedules. We compare the data-driven schedules to standard schedules from ASHRAE, introducing the earth mover's distance as a suitable metric for comparison. The proposed methodology allows identifying differences that might be related to contextual influences, such as surroundings, cultural, or climatic influences. Furthermore, we introduced energy-related features of occupancy schedules to estimate their impacts on energy demand and supply systems design without extensive simulations. We collected data for commercial building uses for 13 selected cities in different U.S. climate zones and compared the retail and restaurant building use-type data-driven schedules to the respective ASHRAE standard schedules and to one another.

Our main observations are:

- Differences between data-driven schedules and ASHRAE standard schedules are large, especially when compared to differences among context-specific data-driven schedules.
- Differences between data-driven schedules and standards are mainly due to a general overprediction of occupancy in the standards. Other significant differences were found in full load hours (i.e. cumulative occupancy), peak values, and ramp gradients.
- The shape of standard schedules was generally well estimated, meaning that occupied hours and peak times were generally in good agreement between standard and data-driven schedules.

- Despite the smaller differences between context-specific data-driven schedules, comparisons between geographical locations on the level of energy-related features revealed significant differences, especially in occupied hours per week, and full load hours per week. Peak values, peak times, and ramp gradients were generally similar.
- For both use-types and almost all locations, Friday data was significantly different from data of other weekdays. The current approach of categorization into three day-types (weekday, Saturday, and Sunday) for occupancy models should be reconsidered, especially for the use-types analyzed in this work (i.e. retail and restaurant).

8. Outlook

In addition to examining the greater diversity of profiles and the extraction of typical patterns of occupancy based on observation data, the next logical step of this research is to consider the potential implications of our findings in the realm of energy simulations.

The focus of further work will be the application of data-based occupancy schedules for commercial buildings in a real case study. In order to research the impacts of using data-based schedules compared to the benchmark of standard occupancy schedules on the district scale, extensive simulations with an urban energy modeling tool are planned.

9. Acknowledgements

This work was developed at the Future Cities Laboratory at the Singapore-ETH Centre, which was established collaboratively between ETH Zurich and Singapore's National Research Foundation (FI 370074016) under its Campus for Research Excellence and Technological Enterprise programme.

Appendix A. Method Details

Table A.8 provides the list of all place types supported in the Google Places API search queries [56]. Table A.9 provides an example of the data structure of the raw data collected with the populartimes script [54].

Table A.8: List of Google place-types supported by the API nearby search functionality [56].

accounting	airport	amusement_park	aquarium
art_gallery	atm	bakery	bank
bar	beauty_salon	bicycle_store	book_store
bowling_alley	bus_station	cafe	campground
car_dealer	car_rental	car_repair	car_wash
casino	cemetery	church	city_hall
clothing_store	convenience_store	courthouse	dentist
department_store	doctor	drugstore	electrician
electronics_store	embassy	fire_station	florist
funeral_home	furniture_store	gas_station	grocery_or_supermarket
gym	hair_care	hardware_store	hindu_temple
home_goods_store	hospital	insurance_agency	jewelry_store
laundry	lawyer	library	light_rail_station
liquor_store	local_government_office	locksmith	lodging
meal_delivery	meal_takeaway	mosque	movie_rental
movie_theater	moving_company	museum	night_club
painter	park	parking	pet_store
pharmacy	physiotherapist	plumber	police
post_office	primary_school	real_estate_agency	restaurant
roofing_contractor	rv_park	school	secondary_school
shoe_store	shopping_mall	spa	stadium
storage	store	subway_station	supermarket
synagogue	taxi_stand	tourist_attraction	train_station
transit_station	travel_agency	university	veterinary_care
zoo			

Table A.9: Example of a raw data set (not a real place) obtained with populartimes [54]. The place in this example is categorized as restaurant based on the place-type information according to the method introduced in this paper.

Attribute	Example Data	Origin
place id	”ChIJTxXE6Nd_j4qgulV36kcAvD9”	Google Places API
name	”Joshuas”	Google Places API
address	”855 Brannan Street, San Francisco”	Google Places API
place types	[”night_club”, ”bar”, ”restaurant”, ”point_of_interest”,”food”,”establishment”]	Google Places API
coordinates	{lat: 37.771900; lng: -122.403268}	Google Places API

popular times	[{name: ”Monday”, data: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7, 10, 12, 10,0,10, 22, 38, 39, 24, 0, 0, 0]};
	{name: ”Tuesday”, data: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 17, 22, 21, 16, 0, 17, 35, 58, 65, 48, 0, 0, 0]};
	{name: ”Wednesday”, data: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 25, 30, 6, 0, 14, 40, 80, 100, 75, 0, 0, 0]};
	{name: ”Thursday”, data: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 18, 42, 43, 20, 0, 15, 40, 72, 87, 70, 37, 0, 0]};
	{name: ”Friday”, data: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 21, 30, 30, 22, 0, 18, 38, 65, 77, 62, 33, 0, 0]};
	{name: ”Saturday”, data: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 4, 7, 10, 0, 17, 32, 51, 58, 47, 30, 0, 0]};
	{name: ”Sunday”, data: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 4, 4, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0]}]
		obtained using populartimes [54]

References

[1] M. Kavgic, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, M. Djurovic-Petrovic, A review of bottom-up building stock models for energy consumption in the residential sector, Building and Environment 45 (7) (2010) 1683–1697. doi:10.1016/j.buildenv.2010.01.021.

[2] L. G. Swan, V. I. Ugursal, Modeling of end-use energy consumption in the residential sector: A review of modeling techniques, Renewable and Sustainable Energy Reviews 13 (8) (2009) 1819–1835. doi:10.1016/j.rser.2008.09.033.

[3] T. Hong, S. C. Taylor-Lange, S. D’Oca, D. Yan, S. P. Corgnati, Advances in research and applications of energy-related occupant behavior in buildings, Energy and Buildings 116 (2016) 694–702. doi:10.1016/j.enbuild.2015.11.052.

[4] D. Yan, W. O’Brien, T. Hong, X. Feng, H. Burak Gunay, F. Tahmasebi, A. Mahdavi, Occupant behavior modeling for building performance simulation: Current state and future challenges, Energy and Buildings 107 (2015) 264–278. doi:10.1016/j.enbuild.2015.08.032.

[5] T. Kusuda, NBSLD, computer program for heating and cooling loads in buildings, Tech. Rep. NBS IR 74-574, National Bureau of Standards, Gaithersburg, MD (1974). doi:10.6028/NBS.IR.74-574.

[6] H. Lau, J. M. Ayres, Building Energy Analysis Programs, in: Proceedings of the 11th Conference on Winter Simulation - Volume 1, WSC ’79, IEEE Press, Piscataway, NJ, USA, 1979, pp. 283–289.

[7] Bruce D Hunn, Patterns of Energy Use in Buildings, in: Fundamentals of Building Energy Dynamics, MIT Press, 1996, pp. 39–111.

[8] AIA Research Corporation., United States. Dept. of Energy. Office of Conservation and Solar Applications., United States. Dept. of Housing and Urban Development. Office of Policy Development and Research., Phase One/ Base Data for the Development of Energy Performance Standards for New Buildings :Final Report, Dept. of Energy, Washington, 1978.

[9] U.S. Department of Energy, Standard Building Operating Conditions - Technical Support Document for Notice of Proposed Rulemaking on Energy Performance Standards for New Buildings, DOE/CS-0118, DOE, Washington, D.C. (Nov. 1979).

[10] ASHRAE, ANSI/ASHRAE/IESNA 90.1-1989: Energy Efficient Design of New Buildings Except Low-Rise Residential Buildings (1989).

[11] ASHRAE, User’s Manual for ANSI/ASHRAE/IESNA Standard 90.1-2004, ASHRAE, 2004.

[12] M. Deru, K. Field, D. Studer, K. Benne, B. Griffith, P. Torcellini, B. Liu, M. Halverson, D. Winiarski, M. Rosenberg, M. Yazdanian, J. Huang, D. Crawley, U.S. Department of Energy Commercial Reference Building Models of the National Building Stock, Tech. Rep. NREL/TP-5500-46861, 1009264 (Feb. 2011). doi:10.2172/1009264.

[13] ASHRAE, ANSI/ASHRAE/IES Standard 90.1-2013: Energy Standard for Buildings Except Low-Rise Residential Buildings (2013).

[14] ASHRAE Project Committee 90.1, Schedules and internal loads for Appendix C, http://sspc901.ashraepcs.org/documents/Addendum_an_Sched_and_Load.pdf (2019).

[15] S. Pless, P. Torcellini, N. Long, Technical Support Document: Development of the Advanced Energy Design Guide for K-12 Schools–30% Energy Savings, Tech. Rep. NREL/TP-550-42114, 918448 (Sep. 2007). doi:10.2172/918448.

[16] B. Liu, R. E. Jarnagin, W. Jiang, K. Gowri, Technical Support Document: The Development of the Advanced Energy Design Guide for Small Warehouse and Self-Storage Buildings, Tech. Rep. PNNL-17056, 921429 (Dec. 2007). doi:10.2172/921429.

[17] W. Jiang, R. E. Jarnagin, K. Gowri, M. McBride, B. Liu, Technical Support Document: The Development of the Advanced Energy Design Guide for Highway Lodging Buildings, Tech. Rep. PNNL-17875, 939043 (Sep. 2008). doi:10.2172/939043.

[18] E. Bonnema, I. Doebber, S. Pless, P. Torcellini, Technical Support Document: Development of the Advanced Energy Design Guide for Small Hospitals and Healthcare Facilities–30% Guide, Tech. Rep. NREL/TP-550-46314, 977289 (Mar. 2010). doi:10.2172/977289.

[19] S. Somasundaram, P. R. Armstrong, D. B. Belzer, S. C. Gaines, D. L. Hadley, S. Katipamula, D. L. Smith, D. W. Winiarski, Screening Analysis for EPACT-Covered Commercial HVAC and Water-Heating Equipment 225.

[20] COMNET, Factsheet COMNET Overview (2012).

[21] W. O'Brien, I. Gaetani, S. Gilani, S. Carlucci, P.-J. Hoes, J. Hensen, International survey on current occupant modelling approaches in building performance simulation, Journal of Building Performance Simulation 10 (5-6) (2017) 653–671. doi:10.1080/19401493.2016.1243731.

[22] W. O'Brien, S. Gilani, M. Ouf, Advancing Occupant Modeling for Building Design & Code Compliance: Part 1: Introduction, ASHRAE Journal (Feb. 2019).

[23] G. Happle, J. A. Fonseca, A. Schlueter, A review on occupant behavior in urban building energy models, Energy and Buildings 174 (2018) 276–292. doi:10.1016/j.enbuild.2018.06.030.

[24] D. Yan, T. Hong, B. Dong, A. Mahdavi, S. D'Oca, I. Gaetani, X. Feng, IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings, Energy and Buildings 156 (2017) 258–270. doi:10.1016/j.enbuild.2017.09.084.

[25] X. Feng, D. Yan, T. Hong, Simulation of occupancy in buildings, Energy and Buildings 87 (2015) 348–359. doi:10.1016/j.enbuild.2014.11.067.

[26] I. Gaetani, P.-J. Hoes, J. L. M. Hensen, Occupant behavior in building energy simulation: Towards a fit-for-purpose modeling strategy, Energy and Buildings 121 (2016) 188–204. doi:10.1016/j.enbuild.2016.03.038.

[27] W.-K. Chang, T. Hong, Statistical analysis and modeling of occupancy patterns in open-plan offices using measured lighting-switch data, Building Simulation 6 (1) (2013) 23–32. doi:10.1007/s12273-013-0106-y.

[28] J. Page, D. Robinson, N. Morel, J. L. Scartezzini, A generalised stochastic model for the simulation of occupant presence, Energy and Buildings 40 (2) (2008) 83–98. doi:10.1016/j.enbuild.2007.01.018.

[29] C. Wang, D. Yan, Y. Jiang, A novel approach for building occupancy simulation, Building Simulation 4 (2) (2011) 149–167. doi:10.1007/s12273-011-0044-5.

[30] K. Sun, D. Yan, 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. doi:10.1016/j.buildenv.2014.04.030.

[31] J. Zhao, B. Lasternas, K. P. Lam, R. Yun, V. Loftness, Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining, Energy and Buildings 82 (2014) 341–355. doi:10.1016/j.enbuild.2014.07.033.

[32] S. D'Oca, T. Hong, Occupancy schedules learning process through a data mining framework, Energy and Buildings 88 (2015) 395–408. doi:10.1016/j.enbuild.2014.11.065.

[33] K.-U. Ahn, D.-W. Kim, C.-S. Park, P. de Wilde, Predictability of occupant presence and performance gap in building energy simulation, Applied Energy 208 (2017) 1639–1652. doi:10.1016/j.apenergy.2017.04.083.

[34] X. Luo, K. P. Lam, Y. Chen, T. Hong, Performance evaluation of an agent-based occupancy simulation model, Building and Environment 115 (2017) 42–53. doi:10.1016/j.buildenv.2017.01.015.

[35] Y. Peng, A. Rysanek, Z. Nagy, A. Schlueter, Using machine learning techniques for occupancy-prediction-based cooling control in office buildings, Applied Energy 211 (2018) 1343–1358. doi:10.1016/j.apenergy.2017.12.002.

[36] Y. Chen, T. Hong, X. Luo, An agent-based stochastic Occupancy Simulator, Building Simulation 11 (1) (2018) 37–49. doi:10.1007/s12273-017-0379-7.

[37] D. Huber, Background Positioning for Mobile Devices - Android vs. iPhone, in: Joint Conference of IEEE Computer & Communication Societies, 2011, p. 7.

[38] E. Martin, O. Vinyals, G. Friedland, R. Bajcsy, Precise indoor localization using smart phones, in: Proceedings of the International Conference on Multimedia - MM '10, ACM Press, Firenze, Italy, 2010, p. 787. doi:10.1145/1873951.1874078.

[39] Google, Popular times, wait times, and visit duration, https://support.google.com/business/answer/6263531?hl=en (2020).

[40] Merchant Centric, 3 Easy Ways to Increase Revenue with "Popular Hours" on Facebook, https://www.merchantcentric.com/blog/3-easy-ways-increase-revenue-popular-hours-facebook/ (2017).

[41] Pew Research Center, Mobile Technology and Home Broadband 2019, Tech. rep. (Jun. 2019).

[42] C. Hwong, Chart of the Week: What are the most popular mapping apps?, https://vertoanalytics.com/chart-of-the-week-what-are-the-most-popular-mapping-apps/ (May 2018).

[43] R. Panko, The Popularity of Google Maps: Trends in Navigation Apps in 2018, https://themanifest.com/app-development/popularity-google-maps-trends-navigation-apps-2018 (Jul. 2018).

[44] Jkielty, Android v iOS market share 2019, https://deviceatlas.com/blog/android-v-ios-market-share#us (Sep. 2019).

[45] A. Perrin, M. Anderson, Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018, https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since (Apr. 2019).

[46] J. Parker, A. Hardy, D. Glew, C. Gorse, A methodology for creating building energy model occupancy schedules using personal location metadata, Energy and Buildings 150 (2017) 211–223. doi:10.1016/j.enbuild.2017.06.014.

[47] T. Yoshida, Y. Yamagata, D. Murakami, Energy demand estimation using quasi-real-time people activity data, Energy Procedia 158 (2019) 4172–4177. doi:10.1016/j.egypro.2019.01.813.

[48] G. Jiefan, X. Peng, P. Zhihong, C. Yongbao, J. Ying, C. Zhe, Extracting typical occupancy data of different buildings from mobile positioning data, Energy and Buildings 180 (2018) 135–145. doi:10.1016/j.enbuild.2018.09.002.

[49] Z. Pang, P. Xu, Z. O'Neill, J. Gu, S. Qiu, X. Lu, X. Li, Application of mobile positioning occupancy data for building energy simulation: An engineering case study, Building and Environment 141 (2018) 1–15. doi:10.1016/j.buildenv.2018.05.030.

[50] E. Barbour, C. C. Davila, S. Gupta, C. Reinhart, J. Kaur, M. C. Gonzlez, Planning for sustainable cities by estimating building occupancy with mobile phones, Nature Communications 10 (1) (2019) 1–10. doi:10.1038/s41467-019-11685-w.

[51] Google, Google Maps, https://www.google.com/maps (2019).

[52] L. Prez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy and Buildings 40 (3) (2008) 394–398. doi:10.1016/j.enbuild.2007.03.007.

[53] U.S. Energy Information Administration (EIA), Commercial Buildings Energy Consumption Survey (CBECS), Table E2. Major fuel consumption intensities (Btu) by end use, 2012, Tech. rep. (2016).

[54] m-wrzr, riedmaph, Populartimes, https://github.com/m-wrzr/populartimes (Mar. 2019).

[55] Google, Google Maps Platform — Places API, https://developers.google.com/places/web-service/intro (2020).

[56] Google Maps Platform, Place Types — Places API, https://developers.google.com/places/supported_types (2019).

[57] B. Abushakra, A. Sreshthaputra, J. S. Haberl, D. E. Claridge, Compilation of Diversity Factors and Schedules for Energy and Cooling Load Calculations, ASHRAE Research Project 1093-RP, Final Report (2001).

[58] E. Levina, P. Bickel, The Earth Mover's distance is the Mallows distance: Some insights from statistics, in: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vol. 2, 2001, pp. 251–256 vol.2. doi:10.1109/ICCV.2001.937632.

[59] A. Ramdas, N. Garcia, M. Cuturi, On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests, arXiv:1509.02237 [math, stat] (Sep. 2015). arXiv:1509.02237.

[60] Y. Rubner, C. Tomasi, L. J. Guibas, The Earth Mover's Distance as a Metric for Image Retrieval, International Journal of Computer Vision 40 (2) (2000) 99–121. doi:10.1023/A:1026543900054.

[61] O. Pele, M. Werman, A Linear Time Histogram Metric for Improved SIFT Matching, in: D. Forsyth, P. Torr, A. Zisserman (Eds.), Computer Vision - ECCV 2008, Vol. 5304, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 495–508. doi:10.1007/978-3-540-88690-7_37.

[62] O. Pele, M. Werman, Fast and robust earth mover's distances, in: 2009 IEEE 12th International Conference on Computer Vision, 2009.

[63] W. Mayner, Pyemd: A Python wrapper for Ofir Pele and Michael Werman's implementation of the Earth Mover's Distance., http://github.com/wmayner/pyemd (2019).

[64] C. Chen, J. Twycross, J. M. Garibaldi, A new accuracy measure based on bounded relative error for time series forecasting, PLOS ONE 12 (3) (2017) e0174202. doi:10.1371/journal.pone.0174202.

[65] The SciPy community, Scipy.signal.find_peaks - SciPy v1.3.1 Reference Guide, https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html (2019).

[66] The MathWorks, Inc., Find local maxima - MATLAB findpeaks, https://www.mathworks.com/help/signal/ref/findpeaks.html (2019).

[67] GeoNames, GeoNames geographical database, http://www.geonames.org/ (2020).

[68] Google, Manage your Location History, https://support.google.com/accounts/answer/3118687?hl=en (2020).