Estimating Building Occupancy: A Machine Learning System for Day, Night and Episodic Events

Marie Urban (urbanml@ornl.gov)
Oak Ridge National Laboratory https://orcid.org/0000-0001-9571-832X
Robert Stewart
Oak Ridge National Laboratory
Scott Basford
Oak Ridge National Laboratory
Zachary Palmer
Oak Ridge National Laboratory
Jason Kaufman
Oak Ridge National Laboratory

Research Article

Keywords: Building Occupancy, Bayesian Learning, Probability Estimates, Uncertainty Quantification, Systematic Process, Population Modeling

Posted Date: April 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-427953/v1

Estimating building occupancy: a machine learning system for day, night, and episodic events

Marie Urban* (0000-0001-9571-832X), Robert Stewart (0000-0002-8186-7559), Scott Basford (0000-0002-1923-884X), Zachary Palmer (0000-0002-4651-5291), and Jason Kaufman (0000-0001-5482-6914)

Oak Ridge National Laboratory, One Bethel Valley Road, Oak Ridge, TN 37831

*Corresponding author: urbanml@ornl.gov

Abstract

Building occupancy research increasingly emphasizes understanding the social and physical dynamics of how people occupy space. Opportunities in the open source domain including social media, Volunteered Geographic Information, crowdsourcing, and sensor data have proliferated, resulting in the exploration of building occupancy dynamics at varying spatiotemporal scales. At Oak Ridge National Laboratory, research into building occupancies through the development of a global learning framework that accommodates exploitation of open source authoritative sources, including governmental census and surveys, journal articles, real estate databases, and more, to report national and subnational building occupancies across the world continues through the Population Density Tables (PDT) project.

This probabilistic learning system accommodates expert knowledge, experience, and open source data to capture local, socioeconomic, and cultural information about human activity. It does so through a systematic process of data harmonization techniques in the development of observation models for over 50 building types to dynamically update baseline models and report probabilistic diurnal and episodic building occupancy estimates. This discussion will explore how PDT is implemented at scale and expanded based on the development of observation model classes and will explain how to interpret and spatially apply the reported probability occupancy estimates and uncertainty.

Keywords

Building Occupancy, Bayesian Learning, Probability Estimates, Uncertainty Quantification, Systematic Process, Population Modeling

Declarations

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan)

Funding

This research was funded in part by the National Geospatial-Intelligence Agency; Approved for Public Release, 21-142

1. Introduction

As migration into urban areas continues unabated (United Nations 2019), policy makers are forced to navigate new challenges of sustainability and risk management in the midst of increased emissions and erratic climatic events (Barbour et al. 2020; Reinhart et al. 2016; Diffenbaugh 2020). Adhering to sustainability measures associated with updating existing buildings to new green standards, understanding the movements of people and their use of buildings, and managing the risk to population associated with natural or man-made events have resulted in more research into building occupancy at various spatiotemporal scales for green building technologies, risk analysis, and population modeling (Stewart et al. 2016).

Research on energy consumption reveals that 40% of the energy consumed in the United States in 2018 went to residential and commercial buildings (U.S. Energy Information Administration 2020). This research continues by capturing the building dynamics for optimization of building energy consumption, from exploiting sensor and digital traces within buildings for indoor positioning and mapping (Hossain 2019), or quantitative methods for modeling occupant behavior (Hong et al. 2020), to modeling transportation movements within a city to estimate population movement to transform an existing database of building occupancies (Berres et al. 2019).

Exploration into better methods to report where people may be expected in the built environment has led to advances in remote sensing technologies and machine learning for extraction of global settlement layers (Palacious-Lopez 2019) to focus population modeling in the built environment. Building extractions have facilitated the development of building footprints (Yuan et al. 2018) and, coupled with census and survey data, supported the modeling of populations at the residential building level (Weber et al. 2017; Fecht et al. 2020). While the population is modeled at the building

83 level, there remains opportunity to focus research on local building dynamics to capture
84 visitors or transient populations. Social media and mobile phone data have been
85 harnessed to capture population dynamics at higher temporal scales through Points of
86 Interest (POI) and the open and close times of businesses (Sparks et al. 2020). Facebook
87 and Google popularity curves have also been used in the development of 24-hour
88 building occupancies (Lu et al. 2020).

89 The seismic community also continues to perform research into hazards and risk
90 involving exposure datasets, including building occupancies for development of
91 probabilistic risk models at various spatial scales (Silva 2018; Yepes-Estrada et al.
92 2017; Silva 2020). This is accomplished through the collaborative support of the
93 worldwide Global Earthquake Model (GEM 2020) and Pager (USGS 2020) programs.
94 The World Housing Encyclopedia (WHE 2014) is an open community website that
95 reports various building types, occupancies, and construction for the seismically active
96 areas of the world to support more earthquake-resistant building practices. This
97 information is developed through a collection of sources and harmonization techniques
98 of expertise and open source data (Jaiswal et al. 2011; WHE 2014).

99 Oak Ridge National Laboratory developed the web-based Population Density
100 Tables (PDT) global learning framework to report national and subnational building
101 occupancies supported by a Bayesian framework (Stewart et al. 2016). The occupancy
102 estimates are reported as people/1000 ft2 for over 50 facility types (Fig. 1) for night,
103 day, and episodic (larger gatherings of people; stadiums – soccer matches; churches –
104 sermons). Baseline statistical occupancy models are developed at the national or
105 subnational level using expert knowledge, experience, or open source data and are
106 updated by open source observation models (Stewart et al. 2016).

Stewart et al. (2016) discusses the Bayesian statistical modeling process in depth and how the Bayesian learning process easily handles disparate open source data, surveys, expert knowledge, and experience. The use of observation models to report ambient occupancies was proven mathematically acceptable by Morton (2013). These models update and refine baseline occupancy estimates (priors) by dynamically updating reported building occupancy estimates (posteriors).

[Insert Fig. 1 here]

In developing a building occupancy learning system, challenges were encountered when scaling this research to the world: (1) the diversity of building types and geographic divisions, which requires an array of observation models to capture local socio-cultural activity through open source data and the development of proxies for data-poor areas; (2) reporting complete transparency and data provenance into the open source observation models and the variability and uncertainty about the reported occupancies; and finally, determining how to interpret the reported occupancies and uncertainty for spatial application.

2. PDT Open Source Observation Models

The PDT system currently holds over 50K open source observation models for different geographic and spatial scales. A systematic process for the development of observation models using open source data was operationalized to allow multiple analysts to support PDT learning updates. This robust approach, which includes a review cycle prior to publication into the PDT system, establishes a level of consistency and accuracy in data input and a reduction in development error and results in transparency in occupancy reporting.

[Insert Fig. 2 here]

Fig. 2 lays out the systematic process to developing observation models. These models use authoritative open source data including academic journals, official government statistics, corporate and university webpages, tourism brochures, and surveys. Relevant observable data about the facility are extracted, and an observation model is selected. Uncertainty is reported through ranges (i.e., almost 300 employees or 1–3 floors) and surrogates developed for data variables not reported but required. The model is submitted for review, and once approved a facility observation model PDF is developed. Finally, the priors are updated to posteriors by the resulting facility PDF within the Bayesian framework.

Ambient building occupancy estimates are reported in PDT as people/1000 ft2. Unfortunately, a query of open source statistics reveals data reported in precisely these units are indeed rare. Fortunately, there is a wealth of information from which one can infer these density units, given the right observation model. Observation models are simply equations or algorithms that allow one to estimate ambient density from a limited but sufficient set of open source observables. Values like total number of visitors, number of employees, job shifts, average visit durations, and building square footages are sufficient to estimate ambient occupancy and bypass intractable methods such as direct and constant surveillance of an operating facility.

Creation of observation models is itself a necessary but insufficient condition. While Morton (2013) shows strict mathematical equivalence given strict input values, we often have uncertain observable values. Indeed, we sometimes find ourselves in a pseudo quantitative description such as “the museum had almost 1 million visitors last year” or “most visitors stay 2–3 hours”. Furthermore, sometimes building characteristics are unclear. For example, in near NADIR satellite imagery, we may only be able to discern that a building has 1–3 stories. This ambiguity from source data can

be used to model reasonable uncertainty. For example, one may choose a heavily left right Beta distribution on 0–1,000,000m for the visitors, a family of log-normal distributions with modes between 2–3 for visit duration, and a discrete, equiprobable distribution over 1–3 for building height. We can measure the sensitivity and impact of different choices of posteriors to these choices. Treating observable inputs as random variables means that the observation models are indeed random functions, and we arrive at a means for representing (albeit imperfectly) uncertainty in the measured density observation. Using the methods outlined in Stewart et al. (2016), those data then move the prior to the posterior while considering the uncertainty in both.

What happens when a facility is missing an observable? For example, we may lack information about visit duration for a particular museum. As an engineering solution, PDT turns to default distribution model observables across a range of similar facility types. These distributions can fit using similar Bayesian learning principles or frequentist fits, but they will result in larger variances. This increases overall uncertainty in the observation model but allows PDT to move forward using only partial observables.

It is worth noting that uncertainty introduced by temporal lags has not been explicitly considered. If a facility's observable data is 2 years old, how should that be weighed against a similar facility's data that is only 2 weeks old? Presently, this remains an open question handled in an ad hoc approach based on subject matter expertise about when old is too old. Alternatively, we can assess the temporal stationarity of the observable by testing the hypothesis that the older data point comes from the same distribution as newer observables. In the main, the use of observation models and surrogate distribution for missing observables allows PDT to move forward with salient inputs for advancing priors to posteriors in the Bayesian learning framework.

Development of observation models has been largely based on informing encounters with open source content. At present there are close to 50 distinct observation models that can be grouped into classes handling a typical set of open source data streams. Often these mirror facility types, which may report similarly, but not reliably. Fig. 3 lists the model classes, which mostly align with a single facility type, although two, Reported and Employee, contribute to multiple facility types. We now describe these classes and point to exemplar models found within them.

[Insert Fig. 3 here]

2.1 Reported

The Reported class is the simplest form of occupancy from known numbers of people reported within a certain space. This reporting can represent a certain time frame (hour or hours) of normal occurrence of occupancy, episodic or a large gathering of people for a function, or the occupancy capacity of the space.

$$pdt = 1000 \cdot A^{-1}[p].$$

The pdt model reports people/1000 ft2, and A is the area of the building. The population, p, is simply divided by the area to report occupancy. The area may or may not be reported, but if a specific building and location is known, the area of the building can be digitized. The number of floors of the building can be determined through imagery or photos, and any uncertainty about that number is captured as an input range. This Simple model supports all building types within PDT, whereas Episodic and Capacity models are not as widespread.

2.2 Residential

Census and housing surveys are considered the best data available for capturing residential occupancy. However, the census or survey data collected in their original

form are not available for direct use, and the reported results are masked (e.g., interval data, truncated), people and area are provided separately (e.g., average household size or average dwelling size), or are a combination of masked and averages. RevengC (Duscherer 2018) was developed to reverse engineer the various combinations of censored tables or averages through the use of statistical algorithms such as Iterative Proportional Fitting, Maximum Likelihood, Minimum Chi-Square, and Weighted Least Squares. Supplementary inputs of lower and upper bounds of household size and dwelling size area are required in addition to the table values or averages and are subsequently published within PDT to produce probable estimates of people/1000 ft2 without uncertainty.

For areas of the world where there is a lack of recent census or surveys, the Residential Land Use (Fig. 1) building functions are reported through the Housing model. This model accounts for the percentage of people working and who are of school age within a household to more accurately represent the number of people expected at home during the day.

$$pdt = 1000 \cdot A^{-1}\left[\left((a \cdot h) + g \cdot h\right) + \left((1-p)b \cdot h\right) + \left((1-s)c \cdot h\right) + \left((h \cdot d)u\right)\right].$$

To account for people at school and work at the household level, this model uses household size and percent school attendance. It is important that age groups align with primary and secondary school attendance. However, exceptions have been made where this alignment does not occur. For the model shown, the breakdown of age groups results in $a = 0-4$, $b = 5-9$, $c = 10-14$, $d = 15-64$, and $g = 65+$. Therefore, a is multiplied by h, which is the household size to determine the outcome of the number of children from 0–4 of a household at home diurnally. Again, h is multiplied by e, which is the number of people in the retired age group (65+) in a household at

home during the day. The school age children and working population are handled differently. The number of children not in school, p, which represents the percent primary school attendance, is subtracted from 1, and the result is multiplied by the household size. This is performed again for the secondary age school children to account for their percentage of the household size at home. For the working age population, d is multiplied by u, which is the unemployment for the geographic area.

Finally, refugee and internally displaced people (IDP) settlements follow the Housing model. Data collected for that particular settlement is collected from various humanitarian organizations that report on populations displaced from their homes due to conflicts or disasters.

2.3 School

Schools and universities are modeled to account for how many students, faculty, and staff will be at the facility at any time during the school hours. The types of data found vary in detail and describe demographic information on students, faculty and staff, school location, and floor and height area and its state of use. In most cases, only enrollment is reported, and efforts to verify the location are required. The model below captures the data necessary to develop the observation model, with some exceptions, such as using proxies as placeholders in the observation model until better data is secured.

$$pdt = 1000 \cdot A^{-1} \left[\frac{((e \cdot p) \cdot \bar{d}) + (f \cdot \bar{j}) + (s \cdot \bar{t})}{z} \right].$$

The student enrollment is represented by e, which is multiplied by the percent attendance, p. Many schools report enrollment, but the actual attendance is sometimes quite different. The result of $e \cdot p$ is further informed by the average number of students,

$\bar{d}$, who may not be attending school due to sickness, appointments, or other errands, for example. This process is also performed for faculty, f, and staff, s, informed by average faculty $\bar{j}$ and average staff $\bar{t}$, respectively. Last of all, z accounts for whether the school operates in shifts.

Universities are represented by the same student model and enrollment, faculty and staff, excluding shifts and percent attendance. The area of the school is either reported, or the buildings that support administrative functions, learning, and research are digitized. On-campus housing is handled separately through the Residential model for Multi Family.

$$pdt = 1000 \cdot A^{-1}\left[\left(e \cdot \bar{d}\right) + (f \cdot \bar{j}) + (s \cdot \bar{t}\,)\right].$$

The university student enrollment is also represented by e, and in this case, $\bar{d}$ represents the average attendance without accounting for the percent attendance.

2.4 Employee

The employee model is a step beyond the Simple model and accounts for employees reported for a business where activity within the building is performed by staff, e.g., Light Manufacturing.

$$pdt = 1000 \cdot A^{-1}[w \cdot \bar{w}].$$

The employees are represented by w, and the average number of employees by $\bar{w}$.

The visitor-employee model was discussed in detail in Morton (2013) and Stewart et al. (2016) and is applied here to a wider range of service activities. The annual visitors is a, average visit time is $\bar{v}$, number of days open each year is d, and $\bar{h}$ is

the average number of hours the museum is open. The number of employees is w, and $\bar{p}$ is the average number of employees at work.

$$pdt = 1000 \cdot A^{-1}\left[\frac{a \cdot \bar{v}}{d \cdot \bar{h}} + w \cdot \bar{p}\right].$$

There are variations on this model based on self-reporting by businesses such as the number of visitors reported either daily, weekly, monthly, or annually.

The Shift Worker model accounts for multiple shifts in manufacturing.

Number of shifts = 1

$$pdt = 1000 \cdot A^{-1}\,[w \cdot \bar{p}].$$

This model differs from the Employee model via the input argument for the average number of workers, $\bar{p}$. The user identifies the average number of workers for one shift, and the majority of the workers are modeled during the day and few people at night for security or cleaning crew. For two shifts, a higher average number of workers is expected during the day because of administrative and other services only offered during the daytime, but a certain number of workers are required to continue the nighttime manufacturing process.

Number of shifts = 2

$$pdt = 1000 \cdot A^{-1}\,[w \cdot \bar{q}].$$

The 2 shifts model requires a different average number of workers, represented by $\bar{q}$.

2.5 Cemetery

The Cemetery activity is modeled as open space rather than building space and is represented by activities including burials, graveside visits, memorial services, walking paths, tourists, and even IDP or refugee camps. Because there is limited information reported on cemetery visitation, the Visitation model accounts for graveside visits

through the function of the number of graves in the cemetery. This count was originally performed manually, which is a time-consuming and error prone task with potential inconsistencies in counts due to human error without acknowledging uncertainty in the results. As a result, we leveraged existing capabilities in deep learning to support automated grave counting to report uncertainty (Lunga et al. 2020).

$$pdt = 1000 \cdot A^{-1}\left[\frac{(g \cdot \bar{g}) \cdot \bar{h}}{\bar{v}}\right].$$

The area, A^{-1}, for the Visitation model is different from the previously discussed models because it is the extent of the open space, or cemetery itself. The number of graves, g, is multiplied by the average number of graves visited, $\bar{g}$, and visitation is average hours visit time, $\bar{h}$, and average hours open, $\bar{v}$.

The Episodic model supports burial activities or memorials where larger crowds are expected. The local socio-cultural activities surrounding the burial practice dictate where a burial will occur, the length of ceremony, and the number of visitors or group size at the burial ceremony. In this model, household size represents the basis for how many people may be attending a funeral for family, with group size for the extended family and additional people from the village included if local funerary practice includes community attendance.

$$pdt = 1000 \cdot A^{-1}\left[\left(j \cdot \bar{k}\right) + l\right],$$

where the household size is j, average family attendance is $\bar{k}$, and group size is l.

2.6 Prison

The Prison observation class of models captures either the baseline static prison population; prisoners and employees; or prisoners, employees, and visitors based on prison operations and data availability. Prison operations differ around the world, but

generally the baseline static inmate population represents the vast majority of the population, with small variances due to employees and visitors.

$$pdt = 1000 \cdot A^{-1}[u + (w \cdot \bar{p}) + (v \cdot \bar{V})].$$

The Prisoner model accounts for the number of prisoners, u, while the employees are once again represented by w, and $\bar{p}$ is the average number of employees. The visitors, v, represent the number reported and the average visitors, $\bar{V}$.

The Capacity model requires the prison occupancy rate in lieu of reported prisoner populations.

$$pdt = 1000 \cdot A^{-1}[(c \cdot \bar{O}) + (w \cdot \bar{p})],$$

where c is the capacity of the prison or cell, and $\bar{O}$ is the average occupancy to produce the capacity prisoner population. The employees, w, and average employees, $\bar{p}$, represent the same employee situation as in the Prison model.

2.7 Hotel

The observation models for hotels uniquely account for the area where people are expected to sleep and excludes shops, conference centers, or other similar venues. Another unique factor is the use of social media to inform whether the hotel is primarily occupied by business people or vacationers, as we expect lower occupancy in primarily business-oriented accommodations to account for fewer people under the business scenario and families under a visit for family fun scenario or a mixture of both. There are four scenarios for determining hotel area, A. (1) If room area is not available, the hotel footprint is digitized, excluding the other services within the hotel, and multiplied by the number of floors. (2) The next best option is the total number of rooms multiplied by a range of guestroom areas if the number of guestrooms for that particular

area are not reported. (3) The total number of rooms are multiplied by a range of guestroom and suite areas separately (this is a slightly better option than 2). (4) And last of all, and the best option, the number of rooms and suites and their specific area are reported.

$$pdt = 1000 \cdot A^{-1}\left[\left((t \cdot g) \cdot \bar{O}\right) \cdot \bar{p} + \left((t \cdot e) \cdot \bar{O}\right) \cdot \bar{q}\right].$$

The Occupancy model area, A, of the guestrooms and suites is found using one of the four scenarios outlined above. The first half of the equation accounts for the number of guests within the hotel, and the second half, the number of employees, which we estimate based on the hotel star rating. The total number of guestrooms (and suites) is t and guests per room is g, the average occupancy rate for the hotel is $\bar{O}$, and the average guests is $\bar{p}$. And again, the total number of guestrooms is t, and this time, the number of employees per room is e, the occupancy rate is $\bar{O}$, and the average number of employees is $\bar{q}$.

$$e = m \cdot r.$$

The number of employees per room, e, is defined by the star rating of the hotel, r, and the number of rooms, m. The star rating is based on the World Tourist Organization's report that is referenced in the city-of-hotels website about the number of employees necessary to support the hotels (https://www.city-of-hotels.com/165/hotel-staff-en.html).

2.8 Port

Ports encompass a variety of activities including freight or oil movement, cruise line terminals, and ferry services, which may all occur within one port.

366 $$pdt = 1000 \cdot A^{-1} \frac{(s+p) \cdot \bar{p}}{z}.$$

367 The area, A, is handled differently based on the function of the port. Where
368 freight or oil is involved, the entire area of the port (open areas and buildings) is
369 included. In this model, employees for ships, s, are assessed by the number of docks
370 and possible types of ships delivering goods. Both the port employees, p, and the
371 average number of employees, $\bar{p}$, at the port are included. Shifts, z, are a common
372 occurrence at ports, but they may differ based on activity.

373 $$pdt = 1000 \cdot A^{-1}(s + (p \cdot \bar{p}) + \bar{n} + \frac{\bar{v}}{\bar{h}}.$$

374 In ports that primarily handle passenger ships (ferry or cruise), the area, A, of the
375 terminal is measured for embarking and disembarking activity only. Passengers for the
376 cruise ship or ferry, n, can be accounted for daily, weekly, monthly, or annually. If
377 daily, s and p account for ship and port employees, respectively, the average number of
378 employees, $\bar{p}$, the average number of hours spent waiting for the ferry or ship, v, and
379 average hours open during the day, $\bar{h}$.

380 **3. Results**

381 A systematic approach to developing observation models optimizes the process of
382 accommodating new disparate open source data to update existing models and guides
383 analysts through the development of the models even as new updates are unveiled. This
384 approach minimizes human errors in the development of the models, establishes the
385 input data necessary for the model, and permits analysts to develop model updates to
386 accommodate new data.

387 **[Insert Fig. 4 here]**

The user is guided through the development of models with requirements for discussion about the use of input data and sources within the development process, promoting transparency. Because the user is guided through the process consistently, analysts not only easily develop observation models for a building function and geographic area, but they also optimize their time in data collection as they identify the necessary data inputs. This approach encourages continuous review of existing models as new or updated data are discovered, or as opportunities to refine the existing process to accommodate new and better methods emerge. Each part of the imputed model is available within the system for users to interrogate.

Fig. 4 consists of three panes of reported occupancies and supporting data from the PDT portal. The left pane lists the country building functions and the reported probability estimates by percentile for night and day: 10th, 50th and 90th. Shown in the middle pane is a list of observation models published in the PDT system for Philippine hospitals and their night and day occupancies. The right pane shows the uncertainty for the Philippine hospitals, from the uncertainty captured at the data input level and propagated throughout the Bayesian modeling process to reveal the full spectrum of probable building occupancy estimates for that country. Taken together, we can clearly see and interrogate the variability in building occupancies reported for Philippine hospitals and the open source observation models.

For the 50th percentile in the right pane of Fig. 4, the uncertainty range is reported below the probability distributions in the green box (day) as 3.39 to 3.81 and in the blue box (night) from 1.74 to 2.01 people/1000 ft2. Uncertainty in population distribution is given in PDF charts depicting the ensemble of probable occupancy estimates for the thickness for the Philippine hospitals. The thicker or hazier the curve, the more uncertainty about the occupancy estimates for that percentile. If another

percentile is selected, the information in the right pane updates with the new percentile uncertainty and the uncertainty ensembles are redrawn in the graphs.

In the left pane of Fig. 4, users can further explore reported occupancies by selecting from the categories “subnational” or “facility type.” At the subnational level, the occupancies are shown by facility type, illustrating differences throughout the country. When selecting facility type, there is an option to explore the occupancies for the selected type for all countries. The occupancies for a country or region can be downloaded by percentile with the accompanying uncertainty.

3.1 Model Variability

When processing large open source datasets, random sampling has been implemented to determine the requisite number of observation models necessary to capture the variability in building occupancy for that building function and geographic area.

For example, Fig. 5 displays the daytime probability distributions for the 10th, 50th, and 90th percentiles for Tokyo schools and the number of observation models published into the PDT system, beginning with the national level baseline model for Japan schools shown in the table in Fig. 5 for the 10th, 50th and 90th percentile ranges and the subsequent refinement of the baseline model as the number of models increases. The 10th percentile converges to zero, 50th percentile is close to 1, and the 50th around 15 people/1000 ft2. Most of the change occurs within the first 10 observation model publications, with a smaller progression after 10 observation models.

The baseline model depicts a conservative baseline occupancy for schools in Tokyo representing the uncertainty with that model. The upper bound occupancy on the x axis is 200 people/1000 ft2, and based on learning from one school, the occupancy is refined quite substantially to an upper bound of 20 people/1000 ft2. There is more uncertainty about the 50th percentile from learning with the one observation model than

at the baseline estimate. This is expected since the baseline model and the occupancy for one school are different enough to increase (or flatten the alpha beta plane) the number of possible occupancies, resulting in more uncertainty about the variability in occupancy for schools in Tokyo. As nine more observation models are published into the system, the uncertainty about the 50th percentile is reduced. After 40 schools are published into the system, the uncertainty ensemble appears to remain nearly the same as with 10 schools, but the ensemble continues to tighten, resulting in fewer probabilities concerning the school occupancy. After 68 data points are added, the overall trend shows a tightening of the distributions toward one best possible answer.

[Insert Fig. 5 here]

The changes are quantitatively observed through the Change in Median Range graph in Fig. 5. When a new observation model is published into the system, the resulting probability distribution changes shape. However, if there is enough uncertainty in the data input in the published model, the reported median occupancy may change even if the median range does not. This is the result of having more than one best median probability estimate as the sampling method is set to randomly select from all possible median probability distributions at each update.

4. Spatial Application

To further explore use, validate, and interpret the results of the PDT building occupancies and accompanying uncertainty, we spatially apply PDT values for a small

[Insert Fig. 6 here]

research area that reports population and economic data. The research area is Shoto (Fig. 6), a small primarily residential district in the Shibuya ward of Tokyo prefecture. Adjoining districts of Shibuya are home to the two busiest railway stations in the world and one of Tokyo's primary shopping corridors. Given its proximity to some of the

463 most bustling areas of the city, property values in Shoto are high, and the neighborhood
464 has an upscale character.
465 Each city, town, or village maintains a Basic Resident Register in which
466 residents are required to register and update if they move or the household occupancy
467 changes. As of January 1, 2020, the total residential population for Shoto is 1414
468 (Tokyo Metropolitan Government 2020), and the number of households is 688, which
469 indicates a household size of 2.1. For Shoto 1 Chome, the 2015 census (2015 census
470 reference) reports 1316 residents and 678 households, while the 2014 economic census
471 reports 145 businesses and 1460 employees.
472 **[Insert Fig. 7 here]**
473 To map the built area of Shoto, we first downloaded building footprints from the
474 collaborative mapping website OpenStreetMap (www.openstreetmap.org) and opened
475 them in the geospatial processing program ArcMap. These data were then displayed
476 over the most recent available Digital Globe Worldview-3 (WV3) satellite imagery for
477 the neighborhood, which was dated April 4, 2020. The downloaded footprints were
478 adjusted as necessary to encompass newly built structures as determined by observation
479 of the WV3 imagery. Once each building in the neighborhood was mapped, we used the
480 area calculation tool within ArcMap to measure the floor area of the structures.

481 Next, we used ground imagery, primarily obtained from the Google Street View
482 service, to ascertain building functions and numbers of floors. Google Street View has
483 extensive coverage in the neighborhood, allowing nearly every building to be observed
484 from street level. Some buildings, however, were located too far from a street to be

Japan Download: 06-22-2020, M-50, H-90, L-10 (Percentile)										
	Day					Night				
Building Function	10th	50th Mid-Low	50th Mid	50th Mid-High	90th	10th	50th Mid-Low	50th Mid	50th Mid-High	90th
Residential										
Japan										
Single Family, Urban Upper/Middle Class	0.24	NA	0.73	NA	1.41	0.50	NA	1.83	NA	4.22
Multi-Family	0.60	NA	1.44	NA	2.85	1.31	NA	3.57	NA	7.04
Shoto 1 Chome										
Single Family, Urban Upper/Middle Class	0.23	NA	1.02	NA	1.54	0.58	NA	1.54	NA	3.00
Multi-Family	0.25	NA	0.58	NA	1.08	0.62	NA	1.69	NA	3.25
Institutions/Public Service										
Religion	6.58	15.3	30.18	49.88	57.28	0.47	1.81	3.51	6.36	5.05
Museum-Urban	0.67	1.59	1.95	2.3	4.01	0	0	0.01	0.02	0.04
School (D-12)	0	0.89	0.94	1.01	15.58	0	0	0	0	0.12
Fire Stations	0.71	3.17	6.98	12.26	4.21	0.43	1.39	2.88	5	2.04
Retail and Service Outlets										
Stores	3.66	12.91	19.54	26.5	45.3	0.84	2.07	3.96	6.52	10.75
Restaurant	13.92	17.07	24.8	34.52	38.31	3.43	12.41	22.3	36.08	48.05
Hotel/Motel	0.56	1.1	1.58	2.12	3.24	0.66	1.76	2.69	3.57	7.11
Commercial										
Office Building	1.81	4.37	7.84	12.46	11.96	0.08	0.47	1.03	1.84	1.24
Transportation										
Warehouse	0.11	1.48	3.93	7.21	7.12	0.45	0.96	1.72	2.79	2.99
Recreation/Entertainment										
Indoor Recreation Center	0.55	2.03	3.83	6.06	7	0.48	0.52	0.85	1.33	0.84
Theater	0.77	2.71	5.24	9.07	13.45	0.76	3.88	8.52	13.93	24.41
Night Club	0.22	0.68	0.93	1.15	1.56	1.6	4.64	9.3	16.22	19.53

Table 1 PDT building occupancies for the 10th, 50th and 90th percentile and night and day

485 viewed, while others were obscured behind large walls or trees. In those cases, floor
486 numbers were determined by close examination of satellite imagery of the properties.

487 POI and photos provided by OpenStreetMap and Google were also used to assist
488 in determining the building functions served by each structure in the district (Fig. 7).
489 The national level PDT building occupancies were downloaded (https://pdt.ornl.gov;
490 May 20, 2020) for application in the spatial mapping process for Shoto (Table 1) and
491 the uncertainty included for the 50th percentile.
492 To validate the PDT reported building occupancies, we chose to use residential
493 subnational estimates for Shoto. The PDT building occupancies for night and day for
494 the 10th, 50th and 90th percentiles were applied to the area of the single-use buildings.
495 For mixed-use buildings, PDT building occupancies were applied separately to those
496 areas and subsequently added together to report a total population for each building and
497 percentile. Table 2 lists areas for residential versus non-residential structures as well as
498 the resulting building occupancies (people/m2).

Building Occupancy for Residential and Non-Residential							
	Area (m^2)	**Day Occupancy**			**Night Occupancy**		
		10^{th}	50^{th}	90^{th}	10^{th}	50^{th}	90^{th}
Residential	148,274.0	796.5	1,902.1	3,721.2	990.7	3,076.6	6,130.8
Non-Residential	43,093.7	818.1	3,315.3	7,167.3	156.4	1,146.1	2,612.0
Total	579,211	1,644.68	5,217.5	10,888.5	1,147.1	4,222.7	8,742.8

Table 2 Area (m^2) of residential and non-residential for Shoto and the resulting residential, non-residential, and the total building occupancy for day and night

499 The resulting building occupancies for the Shoto area are mapped and shown in
500 Fig. 8 for the 10th, 50th, and 90th percentile for night and day. First, the building
501 occupancy range varies from 0 to over 232.0 people/m2, as shown in the legend. The
502 theater reports the highest occupancy for the 90th percentile at 1443.0 people/m2. This
503 range changes for each percentile and for night and day, and expectedly, an increase in

building occupancy reported from the 10th to the 90th percentile for both night and day is represented by the blue at the low end of the occupancy to the red at the high end of the occupancy. The Total Building Occupancy (TBO) also reports the increase from 10th to 90th for both night and day including the Residential (Res) total for Shoto.

[Insert Fig. 8 here]

Non-residential results are found in Table 2 along with Residential and TBO. The day TBO is higher than the night for each percentile, and this is expected given the number of employees alone is higher than the residential count in the area (see Table 1).

Further, the Residential Population for this area is reported as 1414 and found to be between the 10th and 50th percentile (Table 1), within the overall lower range of PDT residential building occupancies reported. One of the challenges in applying PDT building occupancy to the area (footprint and height) of apartments or hostels is that they may include areas that do not contribute to the dwelling. For example, an apartment building contains hallways, staircases, and storage, which are not accounted for in the census-reported dwelling size. This difference in area can cause an increase in the occupancy estimate if we are applying the PDT Multi Family estimate to the area of the apartment building (footprint and height). If the area of apartments within a specific building is known, it is best to apply building occupancies based on that information versus the size of the total building. However, that information is not readily available for all buildings.

Occupancies in Non-Residential areas of Shoto are more of a challenge to validate, even with the available economic information. The reported 1460 employees are found within the resulting night and day range of occupancies, and given the type of businesses in the area, such as hotels, restaurants, and theaters, there is an expectation that some percentage of employees will work overnight.

529 The uncertainty range of a probability distribution or percentile is available for
530 exploration both to map out options or better understand reported occupancies. The
531 PDT uncertainty indicates either data availability is scarce, there is uncertainty about the
532 data at the input level, or both. In Table 1, the night and day uncertainty about the 50th
533 percentile is found in the columns labeled 50th Mid Low and 50th Mid High, while the
534 50th Mid is the best midpoint reported for that range. For Residential, there is no
535 uncertainty about the data reported as the census or housing surveys statistically capture
536 the range of housing sizes for a geographic area, which is the best possible answer
537 available. However, the Non-Residential categories do report an uncertainty range. For
538 example, an urban museum for daytime reports 1.95 people/m2 for the midpoint and
539 ranges from a low of 1.59 people/m2 and a high of 2.3 people/m2.

540 **5. Conclusions**

541 This paper sets out to inform the reader about scaling PDT to report building
542 occupancies at the national and subnational level throughout the world through
543 expansion of model classes of open source observation models. The resulting
544 occupancy estimates include transparent reporting pertaining to model development and
545 uncertainty. Refinements to the school baseline estimates in Tokyo were explored,
546 paying particular attention to the roles of variability and uncertainty in formulating the
547 estimates. Finally, spatial mapping of the building occupancies was employed for
548 validation and interpretation of the results.

549 The PDT learning system reports night, day, and episodic probability occupancy
550 estimates for every country in the world for over 50 building types. There are several
551 classes of observation models within the PDT system to accommodate the range of
552 disparate data or proxies, accounting for the lack of data, for countries around the world.
553 The probability estimates report uncertainty, which is captured at the data input level

554 and propagated throughout the learning process. Complete transparency accompanies
555 the resulting estimates and includes formulas used within the models, the open source
556 data inputs, and any other necessary information about the model.
557 Investigating the observation models for 68 Tokyo schools reveals how the
558 Bayesian process refines the established baseline occupancy estimates. The greatest
559 refinement of the baseline estimate occurs within the first 10 models, with smaller, but
560 continued, refinement thereafter. More investigation is needed into how variability of
561 occupancies for open source observation models affects the learning of PDT baseline
562 estimates and learning differences over the various building functions for the “best
563 possible answer”. If the variability of the models is quantified, that could affect the
564 number of observation models required to adequately inform the baseline occupancies.
565 A validation of spatial application of PDT occupancy estimates was performed
566 in the Shoto 1 Chome area in Tokyo, Japan. The results revealed that the reported Shoto
567 residential counts were within the range of PDT residential occupancy estimates.
568 Validating the other building types is more of a challenge because exact counts of
569 people visiting an area is not readily available. However, the PDT results for non-
570 residential building types does capture the reported number of employees within the
571 10th to 90th population range. Those are promising results in developing population
572 estimates from the PDT occupancy estimates for capturing transient and residential
573 populations within an area.
574 While building footprints with height or 3D buildings are not yet available
575 worldwide, they continue to become more available. Application of PDT building
576 occupancies is a viable option to determine populations other than through census or
577 surveys. The PDT building occupancies account for transient populations through

facility types such as airports, bus and train stations, seaports, hotels, and urban and rural museums.

Additionally, introducing uncertainty for the date of open source models would weight the model impact on the probable occupancy estimates and, over time, phase out the oldest models unless updates to the model were made. In addition, there is opportunity to conduct sensitivity analyses into the reported uncertainty to identify the sources of error to direct data input updates. Finally, data discovery and subsequent model updates to accommodate new sources of data will continue, resulting in an expansion of the model classes.

6. References

Axhausen K, Zimmermann A, Schönfelder S, Rindsfüser G, Haupt T (2000) Observing the rhythms of daily life: a six-week travel diary. Transportation 29(2):95–124. https://doi.org/10.1023/A:1014247822322

Barbour E, Davila CC, Gupta S, Reinhart C, Kaur J, González MC (2019) Planning for sustainable cities by estimating building occupancy with mobile phones. Nature Communications 10(1):3736. https://doi.org/10.1038/s41467-019-11685-w

Berres A, Im P, Kurte K, Allen-Dumas M, Thakur G, Sanyal J (2019) A Mobility-driven Approach to Modeling Building Energy. IEEE International Conference on Big Data. https://doi.org/10.1109/BigData47090.2019.9006308

Diffenbaugh N (2020) Verification of extreme event attribution: Using out-of-sample observations to assess changes in probabilities of unprecedented events. American Association for the Advancement of Science 6(12):2375–2548. https://doi.org/10.1126/sciadv.aay2368

Dong J, Xiao Y, Ou Z, Cui Y, Yla-Jaaski A (2016) Indoor Tracking Using Crowdsourced Maps. ACM/IEEE International Conference on Information

607 Processing in Sensor Networks. https://doi.org/10.1109/IPSN.2016.7460679
608
609 Duchscherer S, Stewart R, Urban M (2018) Revengc: an R package to reverse engineer
610 summarized data. The R Journal 10(2):2073–4859. https://doi.org/10.32614/RJ-
611 2018-044
612
613 Fecht D, Cockings S, Hodgson S, Piel FB, Martin D, Waller LA (2020) Advances in
614 mapping population and demographic characteristics at small-area
615 levels. International Journal of Epidemiology 49(1):15–
616 25. https://doi.org/10.1093/ije/dyz179
617
618 GEM (2020) Global Earthquake Model. https://www.globalquakemodel.org/. Accessed
619 1 April 2021
620
621 González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human
622 mobility patterns. Nature 453(7196):779–
623 782. https://doi.org/10.1038/nature06958
624
625 Hong T, Langevin J, Luo N, Sun K (2020) Developing quantitative insights on building
626 occupant behavior: Supporting modeling tools and datasets. In: Lopes M,
627 Antunes CH, Janda KB (eds) Energy and Behaviour: Towards a Low Carbon
628 Future, 1st edn. Academic Press, London, pp 283-319.
629 https://doi.org/10.1016/B978-0-12-818567-4.00012-0
630
631 Hossain Mahtab AKM (2019) Crowdsourced Indoor Mapping. In: Conesa J, Pérez-
632 Navarro A, Torres-Sospedra J, Montoliu R (eds) Geographical and
633 Fingerprinting Data to Create Systems for Indoor Positioning and
634 Indoor/Outdoor Navigation, 1st edn. Academic Press, London, pp 97-114.
635 https://doi.org/10.1016/B978-0-12-813189-3.00005-8
636
637 Jaiswal KS, Wald DJ, Earle PS, Porter A, Hearne M (2011) Earthquake casualty models
638 within the USGS prompt assessment of global earthquakes for response
639 (PAGER) system. In: Spence R, So E, Scawthorn C (eds) Human Casualties in
640 Earthquakes: Progress in Modelling and Mitigation, 1st edn. Springer

Netherlands, Dordrecht, pp 83–94. https://doi.org/10.1007/978-90-481-9455-1_6

Kahneman D, Krueger AB, Schkade DA, Schwarz N, Stone AA (2004) A survey method for characterizing daily life experience: the day reconstruction method. Science 306(5702):1776–1780. https://doi.org/10.1126/science.1103572

Lu X, Feng F, Pang Z, Yang T, O'Neill Z (2020) Extracting typical occupancy schedules from social media (TOSSM) and its integration with building energy modeling. Building Simulation 14:25-41. https://doi.org/10.1007/s12273-020-0637-y

Lunga D, Dhamdhere R, Walters S, Bragg L, Makkar N, Urban M (2020) Learning to count grave sites for cemetery observation models with satellite imagery. IEEE Geoscience and Remote Sensing Letters 99:1-5. https://doi.org/10.1109/LGRS.2020.3022328

Morton A (2013) A process model for capturing museum population dynamics mathematics. Dissertation, California State Polytechnic University

Palacios-Lopez D, Bachofer F, Esch T, Heldens W, Hirner A, Marconcini, M, Sorichetta A, Zeidler J, Kuenzer C, Dech S, Tatem AJ, Reinartz P (2019) New perspectives for mapping global population distribution using world settlement footprint products. Sustainability 11(21):6056. https://doi.org/10.3390/su11216056

Prelipcean AC, Susilo YO, Gidófalvi G (2018) Collecting travel diaries: current state of the art, best practices, and future research directions. Transportation Research Procedia 32:155-166. https://doi.org/10.1016/j.trpro.2018.10.029

Reinhart CF, Cerezo Davila C (2016) Urban building energy modeling – a review of a nascent field. Building and Environment 97:196-202. https://doi.org/10.1016/j.buildenv.2015.12.001

674 Silva V, Crowley H, Jaiswal K, Acevedo AB, Journey M, Pittore M (2018) Developing
675 a Global Earthquake Risk Model. 16th European Conference on Earthquake
676 Engineering
677
678 Silva V, Amo-Oduro D, Calderon A, Costa C, Dabbeek J, Despotaki V, Martins L,
679 Pagani M, Rao A, Simionato M, Vigano D, Yepes-Estrada C, Acevedo A,
680 Crowley H, Norspool N, Jaiswal K, Journeay MP (2020) Development of a
681 global seismic risk model. Earthquake Spectra 36(1):372-
682 394. https://doi.org/10.1177/8755293019899953
683
684 Sparks K, Thakur G, Pasarkar A, Urban M (2020) A global analysis of cities' geosocial
685 temporal signatures for points of interest hours of operation. International
686 Journal of Geographical Information Science 34(4):759-776.
687 https://doi.org/10.1080/13658816.2019.1615069
688
689 Stewart R, Urban M, Duchscherer S, Kaufman J, Morton A, Thakur G, Piburn J, Moehl
690 J (2016) A Bayesian machine learning model for estimating building occupancy
691 from open source data. Natural Hazards 81(3):1929-
692 1956. https://doi.org/10.1007/s11069-016-2164-9
693
694 Sun L, Axhausen KW (2016) Understanding urban mobility patterns with a
695 probabilistic tensor factorization framework. Transportation Research Part B:
696 Methodological 91:511-524. https://doi.org/10.1016/j.trb.2016.06.011
697
698 Tokyo Metropolitan Government (2020) 住民基本台帳による東京都の世帯と人口
699 .https://www.toukei.metro.tokyo.lg.jp/juukiy/2020/jy20q10501.htm. Accessed 1
700 April 2021
701
702 United Nations (2019) World Urbanization Prospects: The 2018 Revision
703 (ST/ESA/SET.A/420). United Nations, New York
704
705 USGS (2020) PAGER. https://earthquake.usgs.gov/data/pager/. Accessed 1 April 2021
706

U.S. Energy Information Administration (2020) How much energy is consumed in U.S. buildings. https://www.eia.gov/tools/faqs/faq.php?id=86&t=1. Accessed 1 April 2021

Weber EM, Seaman VY, Stewart RN, Bird TJ, Tatem AJ, McKee JJ., Bhaduri BL, Moehl JJ, Reith AE (2018) Census-independent population mapping in northern Nigeria. Remote Sensing of Environment 204:786-798. https://doi.org/10.1016/j.rse.2017.09.024

WHE (2014) World Housing Encyclopedia. http://db.world-housing.net/. Accessed 1 April 2021

Yepes-Estrada C, Silva V, Valcarcel J, Acevedo AB, Tarque N, Hube M, Coronel GD, Santa-Maria H (2017) Modeling the residential building inventory in South America for seismic risk assessment. Earthquake Spectra 33(1):299-322. https://doi.org/10.1193/10191EQS155DP

Yuan J, Roy Chowdhury PK, McKee J, Yang HL, Weaver J, Bhaduri, B (2018) Exploiting deep learning and volunteered geographic information for mapping buildings in Kano, Nigeria. Sci Data 5:180217. https://doi.org/10.1038/sdata.2018.217

Figures

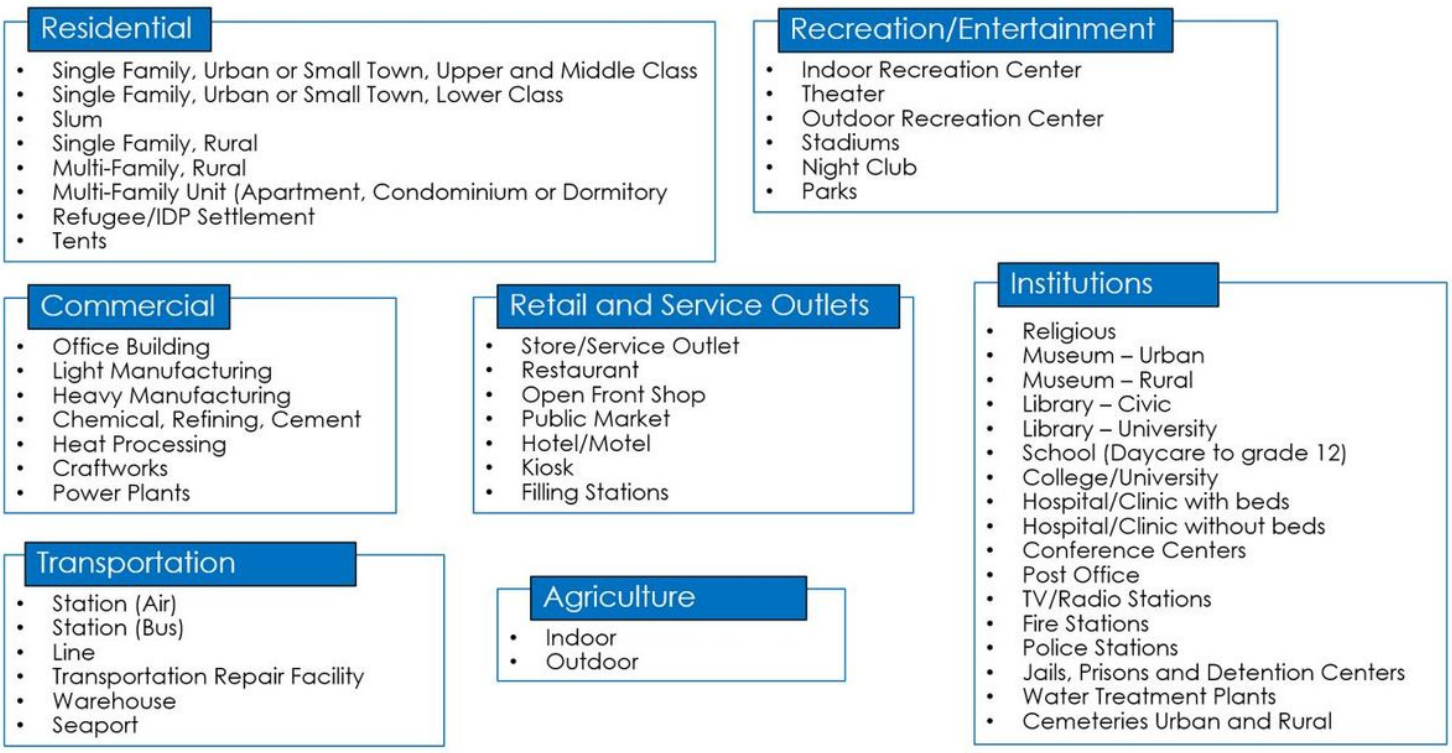


Figure 1

PDT facility types captured under seven land use categories

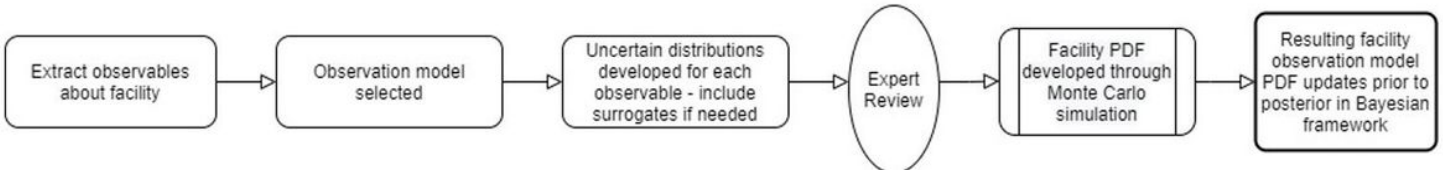


Figure 2

PDT observation model workflow from open source to data measurement under uncertainty. PDF = Probability Distribution Functions

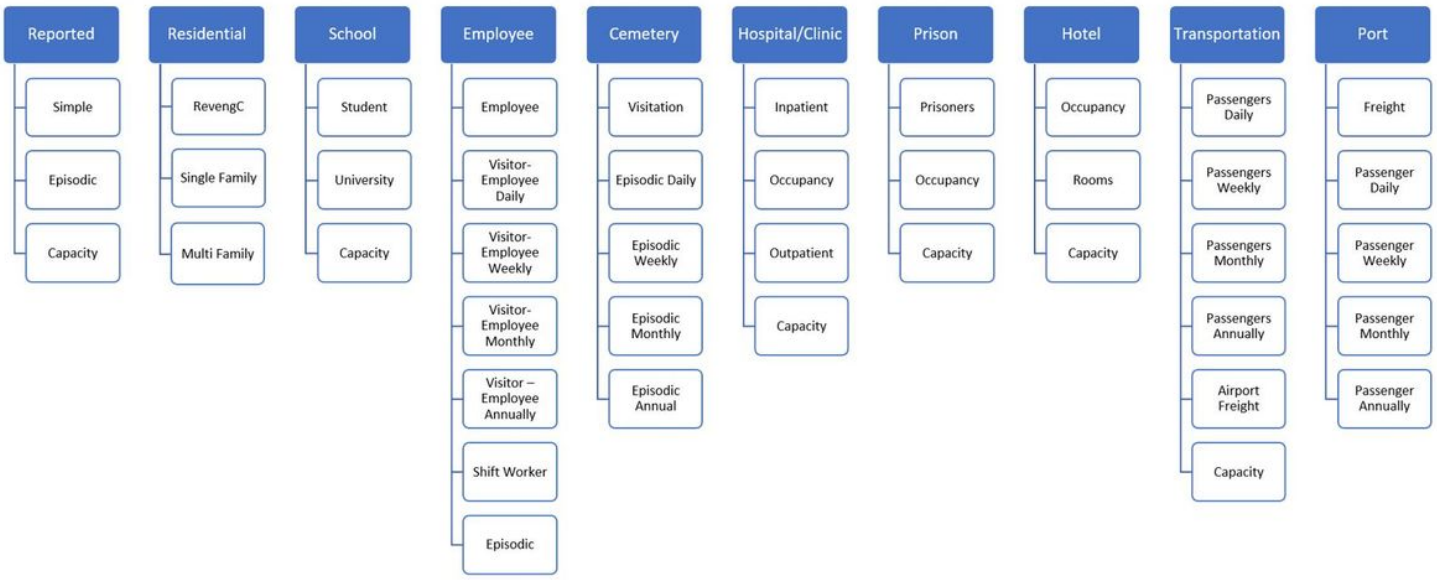


Figure 3

Open source observation model classes in blue at the head of the list and models below. A few models report alternative approaches to accommodate available data

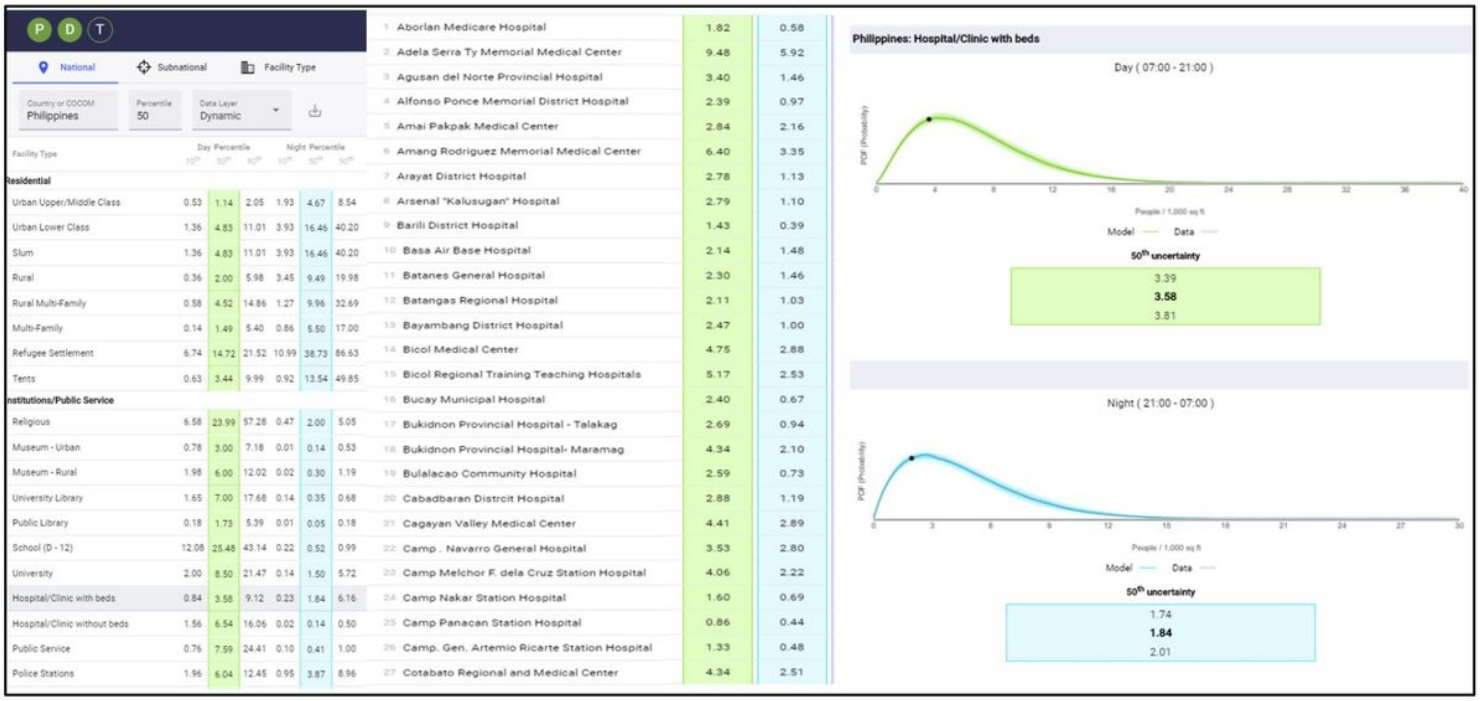
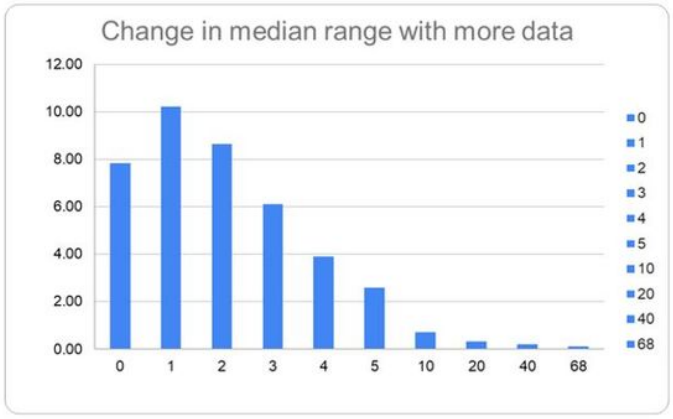


Figure 4

National ambient building occupancies reported in the left panel for 10th, 50th, and 90th probabilities for day and night. Observation models reporting ambient hospital occupancies for day and night in the middle, and the probability distribution for all observation models after publication into the PDT learning system on the right

	10th percentile range				50th percentile range				90th percentile range			
# Reports	Low	Middle	High	Range	Low	Middle	High	Range	Low	Middle	High	Range
0	1.62	2.03	4.37	2.75	14.28	15.84	22.12	7.84	47.44	51.83	60.16	12.72
1	0.23	2.03	2.33	2.10	6.51	15.84	16.71	10.20	34.72	51.83	53.47	18.75
2	0.57	2.03	2.70	2.13	9.04	15.84	17.68	8.64	39.60	51.83	54.35	14.75
3	0.00	0.05	0.28	0.28	1.48	3.73	7.59	6.11	22.88	29.54	38.92	16.04
4	0.00	0.05	0.18	0.18	1.29	3.73	5.20	3.91	21.02	29.54	34.55	13.53
5	0.00	0.00	0.05	0.05	1.07	1.31	3.66	2.59	17.90	21.39	29.39	11.49
10	0.00	0.00	0.00	0.00	1.04	1.31	1.74	0.70	17.14	21.39	24.93	7.79
20	0.00	0.00	0.00	0.00	1.01	1.12	1.34	0.33	16.89	18.43	21.78	4.89
40	0.00	0.00	0.00	0.00	0.94	0.96	1.14	0.20	15.58	15.77	18.73	3.15
68	0.00	0.00	0.00	0.00	0.89	0.94	1.01	0.12	14.79	15.58	16.65	1.86



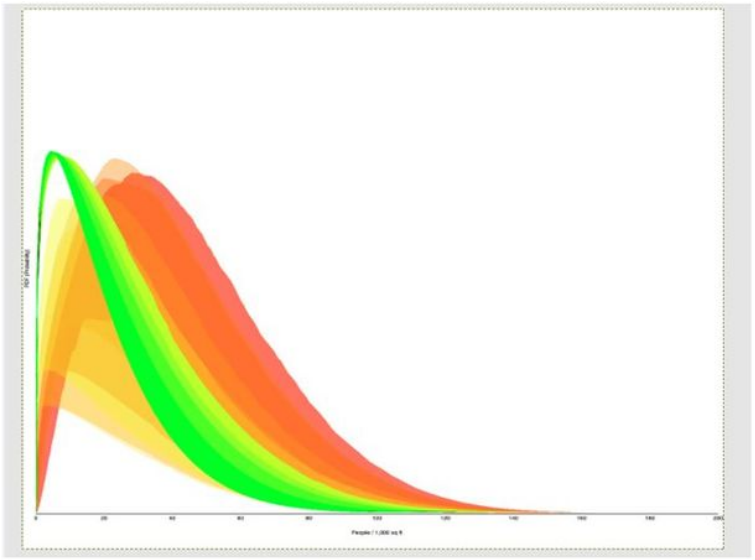


Figure 5

Model learning begins with the baseline model and number of models that refine the baseline estimate and the resulting range of occupancies for Low, Mid, and High (10th, 50th, and 90th percentile, respectively) and the change in the median range shown in the table and the bottom left chart. In the bottom right corner, visualization of probability distributions and ensembles for Tokyo schools for the 50th percentile beginning with the baseline model in red and the refinement of the probability distributions to the green ensemble (#68)

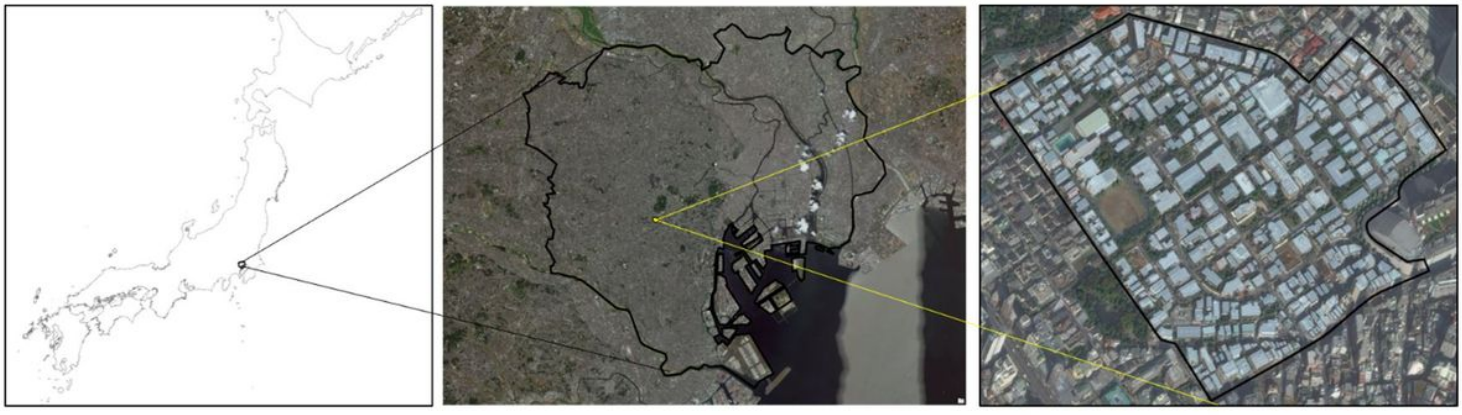


Figure 6

The Shoto (1 Chome) study area shown within Japan (far left) and Tokyo prefecture (middle) and Shoto (right) with building footprints mapped to increase visibility. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Figure 7

Building types for Shoto determined through multiple open sources. Mixed use buildings areas were determined separately for application of the PDT building occupancies

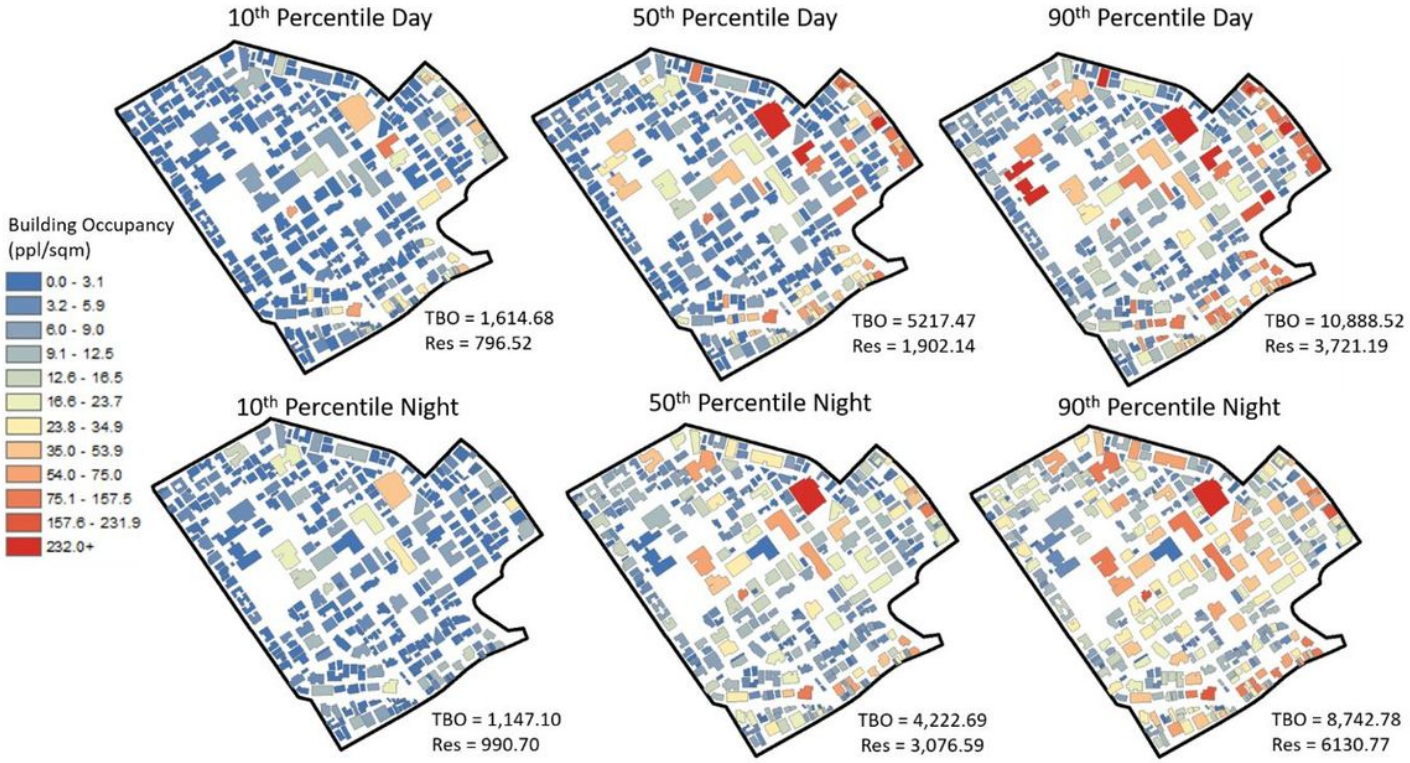


Figure 8

PDT building occupancies (ppl/1000 sqft) probability estimates for the 10th, 50th (the low and high about the 50th; the uncertainty range) and 90th percentile, and night and day for the Shoto area