A Bayesian Machine Learning Model for Estimating Building Occupancy from Open Source Data

Robert Stewart, Marie Urban, April Morton, and Samantha Duchscherer
Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831

Keywords: population, building, occupancy, Bayesian, uncertainty, open source, elicitation

Abstract

Understanding building occupancy is critical to a wide array of applications including natural hazards loss analysis, green building technologies, and population distribution modeling. Due to the expense of directly monitoring buildings, scientists rely in addition on a wide and disparate array of ancillary and open source information including subject matter expertise, survey data, and remote sensing information. These data are fused using *data harmonization methods* which refer to a loose collection of formal and informal techniques for fusing data together to create viable content for building occupancy estimation. In this paper, we add to the current state of the art by introducing the Population Data Tables (PDT), a Bayesian model and informatics system for systematically arranging data and harmonization techniques into a consistent, transparent, knowledge learning framework that retains in the final estimation uncertainty emerging from data, expert judgment, and model parameterization. PDT probabilistically estimates ambient occupancy in units of people/1000ft^2 for over 50 building types at the national and sub-national level with the goal of providing global coverage. The challenge of global coverage led to the development of an interdisciplinary geospatial informatics system tool that provides the framework for capturing, storing, and managing open source data, handling subject matter expertise, carrying out Bayesian analytics as well as visualizing and exporting occupancy estimation results. We present the PDT project, situate the work within the larger community, and report on the progress of this multi-year project.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy 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).

1. Introduction

Estimating building occupancy is an emerging research domain with a wide array of applications including natural hazards loss analysis, green building technologies, and population distribution modeling. Building occupancy is tightly connected with human activity patterns such as home life, entertainment, religious meetings, and shopping. Patterns of life have been studied from a variety of perspectives. For example, geographers have a long history in tracking individual movement patterns using travel diaries (e.g. Axhausen et al. 2000), time use surveys, social media, and mobile phone location data (e.g. Noulas et al. 2012; Gonalez et al. 2008; and Schlich et al. 2003). Others focus on patterns of life from a building use perspective. In the energy domain, studies have focused on the relationship between occupancy and energy consumption particularly in commercial buildings. By detecting occupancy levels in real time with motion detectors, infrared devices, and WIFI connectivity, smart building technologies aim to improve energy efficiencies by adjusting climate controls, lighting, and IT infrastructure (e.g. Hong and Lin 2013; Martani et al. 2012; Melfi et al. 2011; Meyn et al. 2009). Natural hazard loss analytics estimate human fatalities over wide geographic coverages by integrating geophysical modeling with building construction practices and human occupancy patterns (Jaiswahl et al. 2011). Population distribution modeling at fine spatial scales also requires analysts to match building use with population distribution patterns (e.g. Bhaduri et al. 2007).

In this paper we focus on occupancy from a building perspective. Collecting occupancy data through direct building observation is impractical especially for large geographic extents. Because human activities are highly variable, obtaining direct and detailed occupancy data for even a small set of buildings is resource intensive. Instead, experiential accounts or simple inference methods are used to produce a single representative value or range of values for a given building. This representative value is often the *ambient occupancy*, defined as the temporally averaged number of people within a space over a given time, for example daytime or nighttime (e.g. Bhaduri et al. 2007; Sutton et al. 2003). Because of the social and geographic variation in building use dynamics, efforts at ambient occupancy have proved to be effort intensive as well. Scientists have responded to this data challenge by relying on inference methods that harmonize disparate data usually from the open source[1]. The data and corresponding inference methods form a patchwork of resources that vary widely in availability and sophistication, often relying on expert judgment for harmonization and the production of occupancy estimates. This heterogeneity reflects the kind of analytic agility required to utilize data found in open source environments that is often ancillary, incomplete, and improperly scaled both spatially and temporally for the job at hand.

Until the early 2000s, occupancy products were largely a patchwork of disjoint collections focused largely on residential dwellings, with sparse geographic coverages, low spatial detail, were unavailable to the public, or existed as part of a larger proprietary commercial system (e.g. Eguchi et al. 1997; Trendafiloski et al. 2011; Wyss et al. 2013, Jaiswal et al. 2010; and Earle et al. 2009). In 2008, the USGS released a report emphasizing the need for creating a comprehensive global building database for the earthquake community and outlined early challenges and approaches (Jaiswal and Wald 2008). This call led to a series of concerted efforts largely from the natural hazards community to develop occupancy estimates across building taxonomies either by direct observation or through practical inference methods. Important outcomes include the World Housing Encyclopedia Project (WHE 2014), the United States Geological Survey (USGS) Prompt Assessment of Global Earthquakes (PAGER) system (Earle et al. 2009) and the forth coming Global Earthquake Model (GEM) system (GEM 2014) based on the currently developing GED4GEM databases (Gamba et al. 2012). The evolution and collaborations of these efforts can be traced in the following papers: Dell'Acqua et al. (2013), Jaiswal et al. (2011), Wald et al. (2011), Jaiswal and Wald

[1] For example, the PAGER and Hazus systems (https://www.fema.gov/hazus) estimate building class populations as weighted linear combinations of demographic classes found in sources like the U.S. Census, Dun and Bradstreet data, and the CIA World Factbook data (e.g. see Table 1 in Jaiswal et al. (2010b) and Table 13.2 in the Hazus Version 2 Manual (FEMA, 2011)).

(2010), Earle et al. (2009), Jaiswal et al. (2009), and Jaiswal and Wald (2008). Much progress has been made including development of taxonomies, creation of tools for data capture, development of practical inference methods, engaging local experts, and concerted efforts to focus on regions of interest (e.g. earthquake zones).

In the final assessment, occupancy estimation is comprised of a rich collection of techniques reliant on human knowledge and experience to harmonized both direct and ancillary data to accomplish its goals. Whether information is obtained from an expert, obtained by an inferential method, directly from data, or various combinations, the resulting value(s) represent a *belief or knowledge* about the true (but unknown) occupancy. From this vantage, there is much to be to be gained from repositioning the analytics of occupancy within a Bayesian context (Bolstad 2007). Bayesian analysis permits formal linkages between expert opinion, direct observation, and inferential methods with the added value of allowing modelers to think probabilistically about occupancy. That is rather than estimating an occupancy value ($O = x$), the focus is on the *probability* of occupancy $p(O = x)$.

A standard approach is to treat occupancy as a random variable characterized by a probability distribution function, $p(O = x) \sim \delta(\theta)$, where θ is a vector of parameters (e.g. mean and variance). The probability distribution function describes the distribution of occupancy across a selected facility type (buildings of the same type such as museums). The goal is estimate proper parameter values for θ**.** Bayesian methods approach parameter estimation in a way that is well suited to occupancy modeling. Under Bayesian analytics, the parameters of the occupancy distribution are themselves modeled using a probability distribution (e.g. $\theta \sim Norm(\mu, \sigma^2)$). At first mention, this seems to increase in the problem space rather than present a solution. However, by characterizing the probability of any parameter being "correct", an opportunity exists to *formally* quantify both the state of knowledge and the level of uncertainty in a single model that reflects both subject matter experience and data. In a Bayesian context, the probability that any particular set of parameter values is the correct is given by Bayes theorem:

$$p(\theta|data) = \frac{p(data|\theta)p(\theta)}{p(data)} \quad (1)$$

Here $p(\theta)$ is a probability distribution function describing the *prior probability* for model parameters θ and $p(\theta|data)$[2] is the *posterior probability* after evidence (*data*) has been presented. If the prior represents an initial understanding, the posterior is a refinement or update of that knowledge after new data have been observed. In a Bayesian learning model, this new posterior becomes the new prior ready for updating further when newer data become available. In this way, the model learns what parameters are most likely correct over time. A brief and excellent overview of the Bayesian paradigm can be found in Bernardo (2003). The use of Bayesian analytics has expanded rapidly in the last two decades in numerous domains. Examples of Bayesian approaches in population and demographic modeling include: Bryant and Graham (2013), Raferty et al. (2012), Beresovsky et al. (2011), Mugglin et al. (2000), Elliot and Little (2000), and Mugglin and Carlin (1998). There are several important properties about Bayesian analytics that align well with the modus operandi of occupancy estimation.

Variability and Uncertainty: Rather than a single estimate for a building occupancy, a probability distribution $O \sim \delta(\theta)$ permits analysts to represent the variable nature of occupancy. Distributional values like mean, variance, ranges and quartiles provide a rich set of descriptors for describing variability in occupancy. Furthermore, a distribution allows analysts to explore the likelihood of certain occupancy scenarios (e.g. extreme densities). Occupancy uncertainty is characterized as uncertainty in the proper values for θ represented here by the probability distribution $p(\theta|data)$**.**

Subject Matter Expertise: Reliance on expert judgment is common practice in building occupancy (Jaiswal and Wald 2008). Bayesian analytics formally capture this as prior probabilities for $p(\theta)$ through the process of *expert*

[2] The symbol | means "given." The expression is read the "probability of $\boldsymbol{\omega}$ given the observed information."

elicitation. Expert elicitation is the transformation of subject matter experience or expertise into a quantifiable form. Elicitation continues to be a major interdisciplinary research area spanning numerous application domains (e.g. Wisee et al. 2007; Billari et al. 2012; Albert et al. 2012; French 2011). Enumeration of methods for developing priors is beyond the scope of this paper. Berger (2010) and Garthwaite et al. (2005) provide thorough overviews of various methods and statistical approaches. Methods for aggregating multiple opinions include the behavioral aggregation (Phillips, 1999), the Delphi approach (Garthwaite et al. 2005), logarithmic pooling (Genest et al. 1984), the classical model (Cooke, 1991) and hierarchical models (Albert et al. 2012). Prior probabilities can also be inferred from ancillary or historical data as well. The model proposed here is agnostic to the approach allowing analysts' freedom in choosing the best method for the situation.

Systematic Learning over Time: As data are encountered over time, the posteriors are automatically revised to account for the new information. Since the order in which data are introduced does not affect the outcome this approach suits computer automated systems where knowledge evolves, improves, and is further refined over time.

Data Uncertainty: The data which advance the model from the prior to posterior may also be uncertain. Data uncertainty can be accounted for in the posterior probabilities in a variety of ways (e.g. Heid et al. 2004; Tan and Xi 2003). This is particularly important for occupancy data where values are actually inferred (with uncertainty) from other data.

Data Surrogates: In areas with sparse data, it is not uncommon for analysts to use results from a similar area as surrogate values for the target area. Under Bayesian protocols, users may formalize this by using either the prior or posterior model from another area as a prior for the area of interest. This prior is then updated when and if data do become available.

We argue that adoption of Bayesian analytics permits the creation of a systematic approach that arranges and maximizes new and existing data harmonization techniques into a consistent, transparent knowledge learning framework that retains in the estimation uncertainty emerging from data, expert judgment, and inference analytics. By leveraging Bayesian analytics, the approach *formalizes* the use of subject matter expertise in occupancy estimation and explicitly links it with open source data. Uncertainty is propagated through the model allowing researchers to explicitly trace uncertainty in occupancy estimation to uncertainty in the input data through the use of recently developed sensitivity analytics. The result is the production of probability distributions which characterize the extent and variability of occupancy values within a building type for any geographic region of interest.

This research is the focus of the multi-year Population Density Tables (PDT) project at the Oak Ridge National Laboratory. PDT mines open source data with the aim of modeling occupancy for over 50 building types across the world at both the national and sub-national scale. This presents a considerable challenge addressed by the development of the PDT system, a novel spatial content web management tool for capturing, storing, and managing open source data, handling subject matter expertise, carrying out the Bayesian framework presented here as well as visualizing and exporting occupancy estimation results.

In summary, we contribute to the state of the art by developing and implementing a novel modeling framework and informatics system that:

1. Provides an explicit, systematic approach to engaging data, data harmonization techniques, and expert judgment in the production of occupancy estimation across building types,
2. Probabilistically evolves and refines estimates over time as new data and expertise emerge,
3. Formally and explicitly links expert judgment, data, and harmonization techniques,
4. Retains and characterizes uncertainty emerging from expert judgment, data, and inference,
5. Implements an enterprise informatics approach for capturing, managing, and evolving occupancy estimation over multiple contributors, geographies, building types, and harmonization methods.

The paper continues with a discussion of open source data challenges for occupancy estimation. Development of the PDT Bayesian learning model with an example follows. The paper concludes with high level discussion of the PDT informatics system.

2. Data

PDT requires occupancy values in units of people/1000ft^2. A vast amount of information is available in the open source that has important implications for occupancy estimation. On the other hand, occupancy data is rarely reported directly in the required units (people/1000ft^2). For example, museums openly report their annual number of visitors, hours of operation, and days of operation. Floor area can often be remotely measured in high resolution imagery. Hospitals often report yearly summaries that include number of beds, percent occupancy, length of patient stay, and so forth. Census data sometimes provide tables of house counts by house area and house counts by household size separately. These observable values are not always certain. For example, museum floor space may be measured with uncertainty due to poor image resolution, because the building is multi-use, or lack of certainty on the number of floors. In the building types that PDT examines, there are hundreds of different information formats that have been identified. What is required is a means to infer population in units per 1000ft^2 from data that provide a partial view of the complex dynamics each of these institutions experience under normal patterns of life. These inference methods or data harmonization methods are explicitly referred to here as *observation models*.

Observation model refers to methods by which unobservable or latent attributes are estimated from observable quantities (Buckland et al. 2004). It is a "modeled observation." In PDT over 300 hundred such observation models have been developed in response to the kinds of information available in the open source. We use the PDT category of museums to illustrate the development of observation models. For example, an observation model for museum[3] occupancy is:

$$pdt = 1000 \cdot A^{-1}\left[\frac{a \cdot \bar{v}}{d \cdot \bar{h}} + w \cdot \bar{p}\right] \quad (2)$$

Where *pdt* represents the latent ambient occupancy in people/1000ft^2, a is the annual number of visitors, $\bar{v}$ is the average visit time, d is the number of days open, $\bar{h}$ is the average number of daily hours, w is the number of employees, $\bar{p}$ is the average percent working at any given time, and A is the floor area of the museum.

Open source information about these inputs is often a mix of exact, inexact, or missing data. Ranges are sometimes given or colloquial expressions of uncertainty such as "nearly 5 million visitors" are common. Others such as square footage are rarely reported and must be measured through available satellite images, street views, and floor maps. In PDT, each model input is modeled as a probability distribution so that the museum density is itself a probability distribution. When data is missing for a particular variable, PDT retains a reservoir of probability distributions based on similar situations. For example, visit duration is rarely explicitly reported by the museum directly. However, from numerous independent sources (literature, guidebooks, etc.) we learn that for larger museums visit duration typically ranges from 1-3 hours. For each piece of information needed by the observation model, a probability distribution is adopted (e.g. $Area \sim N(\mu, \sigma^2)$) . Here the symbol $\boldsymbol{\omega}$ in Equation 3 denotes the set of observation model inputs based on open source information where bold face font indicates the variable is now a random variable described by a probability distribution function (e.g. $\boldsymbol{\omega} = \{\boldsymbol{A}, \boldsymbol{a}, \boldsymbol{\bar{v}}, \boldsymbol{d}, \boldsymbol{\bar{h}}, \boldsymbol{w}, \boldsymbol{\bar{p}}\}$).

$$p(pdt|\boldsymbol{\omega}) \sim 1000 \cdot \boldsymbol{A}^{-1}\left[\frac{\boldsymbol{a} \cdot \boldsymbol{\bar{v}}}{\boldsymbol{d} \cdot \boldsymbol{\bar{h}}} + \boldsymbol{w} \cdot \boldsymbol{\bar{p}}\right] \quad (3)$$

[3] A theoretical derivation of this model as equivalent to the actual long run average density is given by Morton (2013).

Computationally, the distribution is accomplished through Monte Carlo methods. It is important to note that uncertainty about the true occupancy is modeled by the spread or variance of the resulting observation model[4]. That is, in this case, the variance of $p(pdt|\boldsymbol{\omega})$ reflects our uncertainty about ambient occupancy and not the moment to moment variability of occupancy.

It is unavoidable that some subjective decisions will have to be made regarding the probability distributions for the model input values. For example, the phrase "nearly 500 employees" suggests a range for the number of employees with a likely upper limit of 500. A reasonable lower limit may be somewhere above 400 employees[5]. Because there is a great deal of subjectivity in choosing this value, we adopt a variance-based sensitivity analysis method called the Fourier Amplitude Sensitivity Test (FAST) (Saltelli et al. 1999) which characterizes how uncertainty in individual variables impacts overall uncertainty. The FAST method produces the *total effects index* which estimates percentage of uncertainty that would disappear if input variables were known with complete certainty. Each variable, on its own, introduces some base amount of uncertainty to the model. The contribution of this uncertainty to the overall uncertainty is measured by the *main effect index.* Because certain variables in the model may interact with each other in complicated ways, this base uncertainty may be further compounded or magnified. For example, uncertainty in variables that are multiplied (e.g. $\boldsymbol{w} \cdot \bar{\boldsymbol{p}}$ in the museum model) create a product where individual uncertainties may become multiplicative. The *interactions effect index* describes the additional uncertainty arising from this compounding effect. The total effect index is the sum of these two. These indices help analysts determine which variable(s) are most influential and whether or not valuable resources should be used in reducing their uncertainty.

In summary, the probabilistic observation model allows us to infer population density data from open source information. In combination with FAST sensitivity analytics, PDT can assess the amount and sources of uncertainty pointing to areas where outcomes are sensitive to subjective decisions or distributional choices. We continue now with a presentation of the PDT model which is designed to account for and propagate this uncertainty in the estimation of occupancy across building types.

3. Model

The PDT *building type model,* pdt_{bt},is the probability distribution function characterizing occupancy values for buildings within a specified taxonomy (e.g. museums). While any probability distribution could be used, we have found the two parameter Beta distribution function $Beta(\propto, \beta)$ to be widely applicable due to the rich variety of distributional shapes that it can assume. Bounded functions like the Beta distribution make practical sense as well given that human occupancy in finite space is itself bounded. We refer to this distribution (Equation 4) throughout the rest of the paper although any distribution could be used.

$$pdt_{bt} \sim Beta(\alpha, \beta) \tag{4}$$

An expert elicitation is used here that was developed specifically for this context and for implementation within the PDT tool. The approach, detailed in Stewart et al. (2013), transforms elicited ranges for key percentiles of the distribution model into feasible regions of model parameter space. For multiple contributors, we draw upon behavioral aggregation (Phillips 1999), the Delphi approach (see Garthwaite et al. 2005), or Cooke's classical model (Cooke 1991) for resolving varying opinions.

Posterior

We adopt Jeffreys' rule of probability kinematics, a widely used method of handling data uncertainty within Bayesian analytics (Tan and Xi 2002). Under this approach, the posterior is an average of all possible posteriors,

[4] The same approach could be applied to data harmonization techniques already found in the literature.
[5] We assume that if there were fewer than 400 the report would say "nearly 400" instead of "nearly 500".

weighted according to the likelihood they could occur. Equation 5 and 6 show the implementation of this approach for PDT occupancy estimation.

$$p(\alpha,\beta|\boldsymbol{\omega}) = \int p(\alpha,\beta|pdt)p(pdt|\boldsymbol{\omega})dpdt \quad (5)$$

$$p(\alpha,\beta|pdt) = \frac{p(pdt|\alpha,\beta)p(\alpha,\beta)}{\int p(pdt|\alpha,\beta)p(\alpha,\beta)d_{\alpha\beta}} \quad (6)$$

In Equation 5, the posterior $p(\alpha,\beta|\boldsymbol{\omega})$ is the weighted average of posteriors $p(\alpha,\beta|pdt)$ where the weights $p(pdt|\boldsymbol{\omega})$, are the probabilities of true value *pdt* supplied by the observational model. Equation 6 is the standard posterior given evidence *pdt*.

There are significant implications in accounting for data uncertainty in the posteriors. In Figure 1 we have the results of the observation model $p(pdt|\boldsymbol{\omega})$ for a hypothetical museum. The mean occupancy for the museum is 4 people/1000ft^2.

Figure 1. Observation model distribution for hypothetical museum.

To demonstrate the effect of uncertainty, we adopt a Gaussian prior (Figure 2a) and update it first considering observation model uncertainty (Figure 2b) and then without (Figure 2c). Note that when the full uncertainty (Figure 1) is considered, the posterior and prior distributions are relatively similar. This is expected as the model recognizes that the open source information is fairly uncertain and therefore moves little from its initial position. On the other hand, ignoring the uncertainty by using a single value 4, the posterior becomes excessively certain (higher peak).

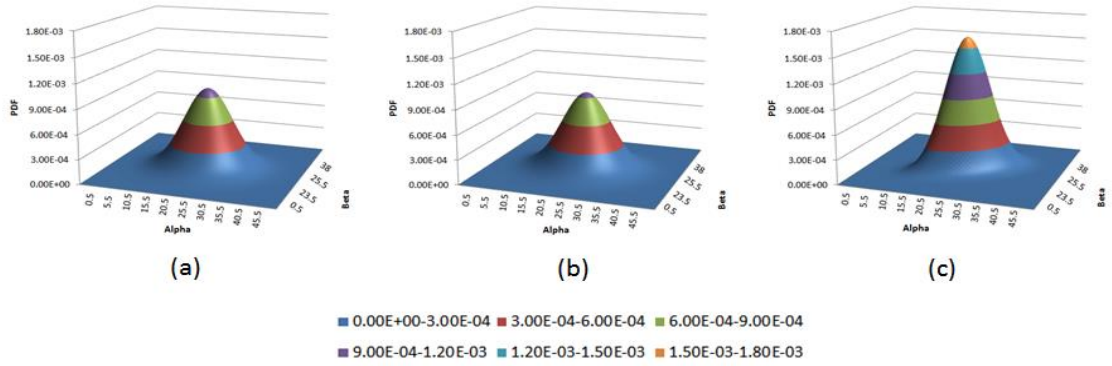


Figure 2. The effect of open source uncertainty on the posterior distribution: a) prior, b) posterior under data uncertainty, and c) posterior without data uncertainty.

Analysts can report occupancy from this process in a number of ways. The most common approach is to choose the most probable model parameters. That is, choose $(\tilde{\alpha}, \tilde{\beta})$ that maximizes $p(\alpha, \beta|\boldsymbol{\omega})$**.** From the resulting distribution, $pdt_{bt} \sim Beta(\tilde{\alpha}, \tilde{\beta})$, PDT reports a low (L), middle (M), and high (H) value respectively corresponding to the 10th, 50th, and 90th percentiles. In addition, one can sample from the posterior parameter distribution (e.g. Gibbs sampling) to report uncertainty about (L, M, H) arising from uncertainty in the model parameter values.

PDT Algorithm

The following describes the steps necessary to accomplish the PDT model in practice.

1. Select a building type, geographic extent, and time of day (daytime or nighttime)
2. Select a building type model $pdt_{bt} \sim \pi(\theta)$, e.g. Beta
3. Develop a prior $p(\theta)$ for parameters θ**,** e.g. (α, β)
4. For each building encountered in the open source**:**
    a. Choose a suitable observation model $p(pdt|\boldsymbol{\omega})$
    b. Develop input probability models $\boldsymbol{\omega}$ from open source/subject matter expertise
    c. Compute observation model $p(pdt|\boldsymbol{\omega})$
    d. Calculate posterior $p(\theta|\boldsymbol{\omega})$ according to equation 5 and 6
    e. Continue to next building using posterior (from d) as new prior
5. Sample the posterior distribution $p(\theta|\boldsymbol{\omega})$ to report values and uncertainty about the (L, M, H)

4. Results

To illustrate the PDT model, we continue with the museum category illustrating 1) how a prior model can be established, 2) how this model is updated to its posterior using uncertain information, and 3) how percentiles of the building type model are reported. In this discussion we explicitly follow the steps outlined in the PDT algorithm.

Step 1: The scenario will be daytime, large urban museums in the United States.

Step 2: Prior experience with museums suggests that a Beta distribution on the interval 0 to 8 people/1000ft2 will provide sufficient flexibility and latitude in modeling urban museum occupancy distribution.

Step 3: Following the elicitation model of Stewart et al. (2013), a senior analyst was elicited for ranges for the 10th, 50th, and 90th percentiles of the building type model as well as a statement about their level of acquaintance with the category. The resulting prior is given in Figure 4 (a).

Step 4a: To illustrate the model, the Art Institute of Chicago (AIC) was selected and the observation model in Equation 3 was selected.

Step 4b: The following table summarizes open source information along with the resulting input model for each variable ($\boldsymbol{\omega}$).

Table 1. Open source museum information about AIC

Variable	Information	Source	Models ($\boldsymbol{\omega}$)
Annual Visitors	1,539,716	Sharpe et al. (2013)	Constant = 1,539,716
Days Open	Daily except for Thanksgiving, Christmas,	http://www.artic.edu/[6]	Constant = 362

[6] Last accessed on 8/22/2014

	and New Year's day		
Average Hours Open	Daily 10:30 – 5:00, 10:30-8 on Thursday	http://www.artic.edu/	Constant = 6
Area (ft^2)	Nearly 1 million	http://www.artic.edu/	$Beta(10,10,980k, 1020k)$
	At 1 million	http://www.artic.edu/	
	750,000 before modern wing + 264,000 after modern wing = 1,014,000	Kolendo and Frumkin (2012) and http://www.artic.edu/	
Average Visit Time (minutes)	Recommended > 3 hours	http://www.chicagotraveler.com/guide-to-art-institute-of-chicago.htm[7]	Poisson(120 minutes) truncated at 360 minutes
	Docent tours .75-1.5hrs	http://www.artic.edu/	
	Recommended 2	http://www.coompass.com/#!search/art institute of chicago [8]	
Workers	Nearly 500 total	http://www.artic.edu/	$U[475,500]_{total}$
Average Percent Working	Percent Working By Hour For Art, Entertainment, & Recreation	American[9] Community Survey at http://www.census.gov/acs/www/	$U[0.4, 0.5]$

The first three input distributions (visitors, days open, and hours of operation) are exact. For these, the distribution model is simply a constant. Area was more problematic in the open source. Three sources were identified with mixed qualitative/quantitative characterizations of size. We adopt a beta distribution centered at one million. Average visit time was problematic as well. A limited number of visitor advice sites recommend say 1-3 hours or more for the AIC. In order to vet these recommendations other large museums were assessed with very similar suggested visit times. After considering these data, a Poisson distribution was adopted with a mean of 120 minutes (2 hours) truncated at 360 minutes (6 hours). The number of employees at AIC is reported as "nearly 500." This helpful as it establishes an upper bound and suggest the true value is close to this upper bound. The lower bound is problematic. The decision was made to model the uncertainty about the number of employees as a uniform distribution on [475-500] and assess the impact varying this lower bound on the model later. Information on the percent of employees working is obtained through American Community Survey data on the percentage of employees working by hour for a predefined number of industries. These account for both full and part time workers. Table 2 shows data for the category of Art, Entertainment, and Recreation during the AIC hours of operation.

Table 2. Percentage of Employees Present by Hour of Day

AM		PM								
10:00	11:00	12:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00
47.8	50.0	36.5	46.4	47.2	45.6	44.4	33.9	27.1	24.3	19.3

Assuming this trend holds on Thursday even though operating hours are longer, the average employee percentage is 44.3%. Under a second scenario, we assume that the museum maintains a high number of employees through the entire hours of operation. Replacing the hourly percentage for 5, 6, and 7-8pm with the employee percentages averaged over the 1:00pm to 5:00pm interval we obtain a slightly higher total mean percentage of 45.3%. This provides a narrow range of uncertainty of only 1%. All of this presumes that the AIC's worker schedule is identical (at least in averages) to the ACS. Although this is unlikely, the average percentages encountered through ACS is highly instructive in building an input distribution. Building on this ACS finding, we broaden the range to 40-50 and assume a uniform distribution in order to account for possible deviations in the AIC work schedule.

[7] Last accessed on 8/22/2014
[8] Last accessed on 8/22/2014
[9] Last accessed on 8/28/2014

Step 4c: Figure 3 shows the resulting observation model with a peak at 1.6 people/1000ft2 with values ranging primarily from 1.2 to 2.2. The spread of this distribution function is due to uncertainty in the inputs and is notably small given uncertainty in the inputs. This reinforces a trend observed by PDT analysts that suggests uncertainty in ambient occupancy estimation is robust to uncertainty in the input. This is largely attributable to the fact that occupancy here is an average.

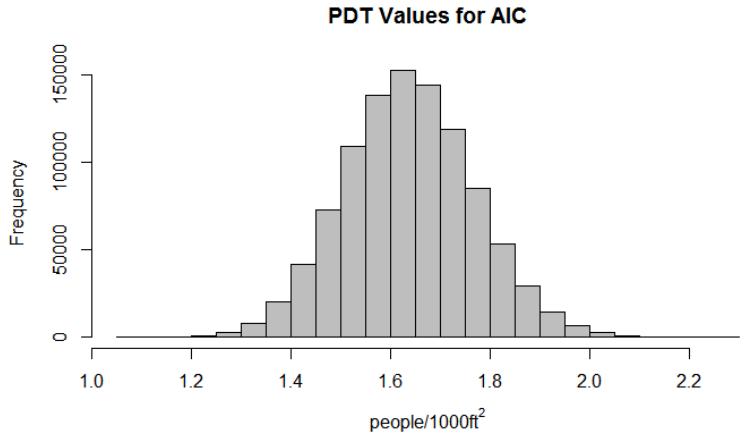


Figure 3. Uncertainty in daytime ambient occupancy for the Art Institute of Chicago

While the input distributions in Table 2 are arguably reasonable choices, they are ultimately subjective. These types of subjective decisions (both in the table and in the text) are unavoidable given the varied terrain of open source data. This is indeed precisely the point of the PDT Model, that is, to formalize these choices within a transparent empirical model where subjective choices such as these can be interrogated and assessed for their relative importance in the overall model uncertainty using the FAST. FAST quantitatively assesses the impact these choices make in the overall uncertainty suggesting a method for prioritizing deeper examination of input values models and parameters. Table 3 summarizes the results after applying the R "sensitivity" package's FAST method (Pujol 2007) to the PDT observation model with inputs provided in Table 2.

Table 3. FAST sensitivity results

Variables	Total Effect	Main Effect	Interaction Effect
Average Visit Time	0.984	0.946	0.038
Average Percent Working	0.066	0.011	0.055
Square Footage	0.058	0.003	0.055
Workers	0.056	0.001	0.055

The results reveal that the uncertainty in the average visit time accounts for approximately 98% of the uncertainty in the final PDT distribution, followed by a roughly equal contribution from average percent working, area, and workers[10]. The main effect and interaction effect indices show that average visit time largely influences the final output on its own, without interacting with the other variables, while the area, workers, and average percent working variables influence the final output variance through strong interactions. These indices suggest the greatest reduction in uncertainty will come from better data on average visit time. Table 4 provides greater detail regarding the sensitivity of the model to different lower bound choices for employee count. Little change in the overall uncertainty is attributable to differing opinions about this input variable. In practice, this deeper interrogation of expert choices and assumptions can be carried out for other variables as well.

Table 4. Worker lower bound choices has little impact on uncertainty

[10] Note that the total effects do not sum to one because the individual interactions between variables are unknown and counted multiple times in the interaction effect indices.

% Contribution	400	425	450	475
Total Effect	0.063	0.060	0.058	0.056
Main Effect	0.009	0.006	0.003	0.001
Interaction Effect	0.054	0.054	0.055	0.055

Step 4d: Figure 4 shows the prior (a) and posterior resulting from the AIC observation model (Figure 3). Both distributions are scaled to same vertical axis for easier comparison.

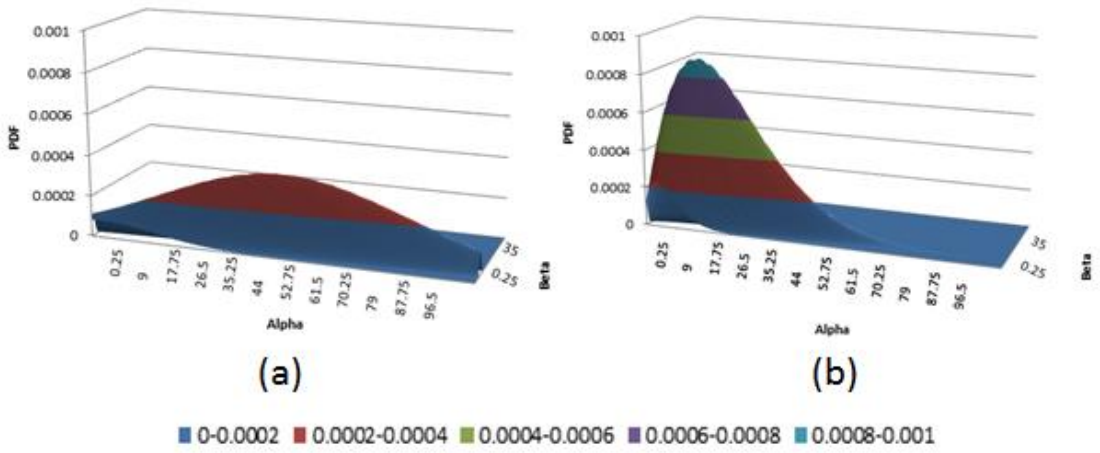


Figure 4. (a) Prior parameter distribution for Urban Museum, Daytime Occupancy density, (b) Posterior distribution after information about the Art Institute of Chicago is introduced.

Step 4e: In this example we only demonstrate the posterior for one museum. Analysts would normally continue with additional museums (e.g. Museum of Modern Art, NY). In this case, the posterior Figure 4 (b) would become the new prior and steps 4a-4d would be repeated again using information about the next museum.

Step 5: Figure 5 shows the building type model for the most probable parameter pair (3, 17.5).

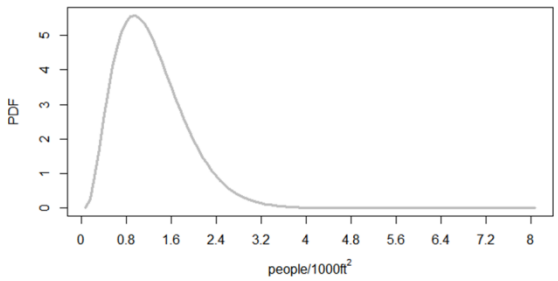


Figure 5. The most probable distribution of urban museum occupancy for the U.S. after updating with open source data on the Art Institute of Chicago.

Drawing samples from the posterior distribution (Figure 4b) produces uncertainty bounds on the Low (10^{th}), Middle (50^{th}), and High (90^{th}) percentiles of the building type model distribution in table.

Table 5. Uncertainty in PDT from uncertainty in open source data

PDT Report	Range (people/1000ft^2)
Low (10^{th})	0.2 – 1.1
Middle (50^{th})	0.6 – 1.8

High (90^{th})	0.7 – 3.4

While Figure 5 indicates the most likely museum distribution model at this point, it is apparent from the overlapping ranges in table, that a great deal of uncertainty remains about what the proper (α, β) pair should be. Of course, this is as it should be since only two pieces of information were provided (the prior and one museum). As more museums are encountered, these ranges for key percentiles will decrease. This is the strength and applicability of the Bayesian learning approach. As the model continues to learn from new data encounters, the uncertainty about model parameters decreases (see e.g. Figure 2). Systematically, this allows PDT analysts to identify and prioritize building categories and geographies for data acquisition, targeting first those combinations with high uncertainty due to poor or sparse data.

5. Validation

Ultimately, the goal of the PDT model is to transparently produce a range (i.e. L and H) that represent the diversity of occupancy rates for a particular type building type. Verification of the accuracy of L (e.g. 10^{th} percentile) and H (e.g. 90^{th} percentile) is difficult given the lack of direct observations. In a very few cases, a handful of direct human observations is available which, although highly subjective, can be compared to the L and H values. For the most part, only ancillary data processed through the observation models is available to produce facility occupancy data. Morton's mathematical proof (Morton, 2013) shows that given true input values, the observation model produces the true average occupancy that would be obtained through continuous observation. Uncertainty emerges from uncertainty in the inputs. Assuming the input distributions do contain the true ancillary values, then the model's distribution of possible average values contains the true but unknown average. A method of evaluating the model and the tool is then to apply a bootstrap method where facilities are partitioned into training and validation sets. Because uncertainty is retained throughout the algorithm, a full disclosure of uncertainty in the L and H values as well as the occupancy estimates is possible. It is possible then to view where most probable values of each facility occupancy occurs and compare this to the overall uncertainty about L and H. While not the same as comparing to available ground truth data (which is infeasible to collect), this approach does provide a reasonable approach to monitoring the progress of model development and its applicability to real world dynamics.

The museum model was trained further by 6 additional museums and compare to 3 museums selected randomly for validation. In Figure 6, the vertical red lines represent the best estimate for L and H after training from 7 museums. The red bands around each line represent uncertainty about L and H attributable to uncertainty in the training data. Blue boxes present a range for the true but unknown ambient occupancy for museums on the Y-axis. Dark blue boxes are training museums and light blue boxes are validation museums. Of the ten total museums, one (Solomon P. Guggenheim) fell outside the range. The de Young museum band falls partially outside the "best model H value" but well within the bounds of uncertainty. Because L and H represent the 10^{th} and 90^{th} percentiles of the facility model it should be expected that 20% will fall outside the vertical red lines.

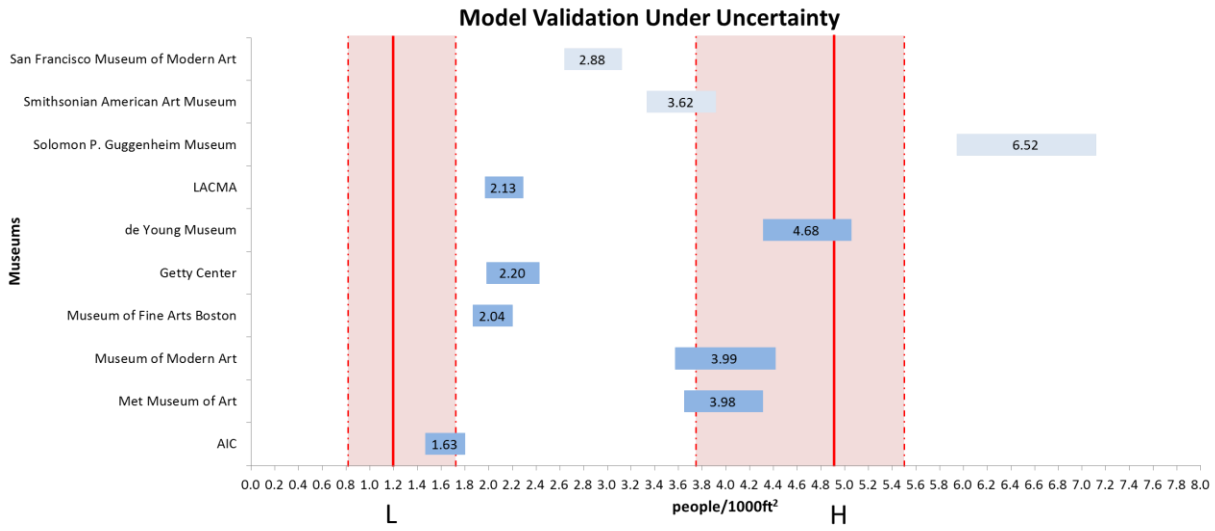


Figure 6. Validation of model results under uncertainty.

This validation technique is used to suggest when training may have achieved sufficient levels and to assess when the distribution model (e.g. Beta) may be in appropriate.

In summary, the results demonstrate how

- the model formally engages uncertain open source data using Bayesian analytics to produce low, middle, and high estimates for occupancy with corresponding measures of uncertainty.
- the model permits analysts to interrogate (through the FAST model) the sensitivity of the outcome to their informed but often subjective choices for input distributions.
- model uncertainty metrics guide the use of valuable resources, and
- to monitor and validate the process while considering both model and data uncertainty

6. Tool

The goal of the PDT project is to scale occupancy density modeling to cover over 50 building types at the national and subnational level across the world. This is a tremendous challenge requiring a comprehensive informatics system capable of handling expert elicitation, capturing, storing, and managing open source data, dealing with hundreds of observational models, their associated input distributions, and carrying out the Bayesian learning model that advances occupancy knowledge over a long time horizon. Information must be managed and carried out at national and even subnational extents thereby multiplying the content management and modeling demands to account for these geographic sub-regions. Researchers have responded to this by creating the PDT system, a novel spatial content web management service powered by a system of databases, computational services, visualization analytics, and web interfaces that has evolved and matured over the past 5 years to meet these considerable demands. A complete discussion of the PDT system is outside the scope and length of this current paper and will be carried out in a forthcoming paper focused on the tool itself. Only the most important elements are provided here to provide clarity on how the system operates.

The PDT system is a browser interface that presents analyst with a rich analytical environment for capturing subject matter expertise, managing data, modeling, and visualizing the results through maps, graphs, and tables. Using the **expert elicitation module** (with assistance by statisticians), users can provide answer to the elicitation questions (Stewart et al. 2013) and obtain feedback on how well their responses reflect their knowledge by showing the most

likely Beta distribution. Visual aids include tools for visualizing population densities in built environments (implemented outside web tool).

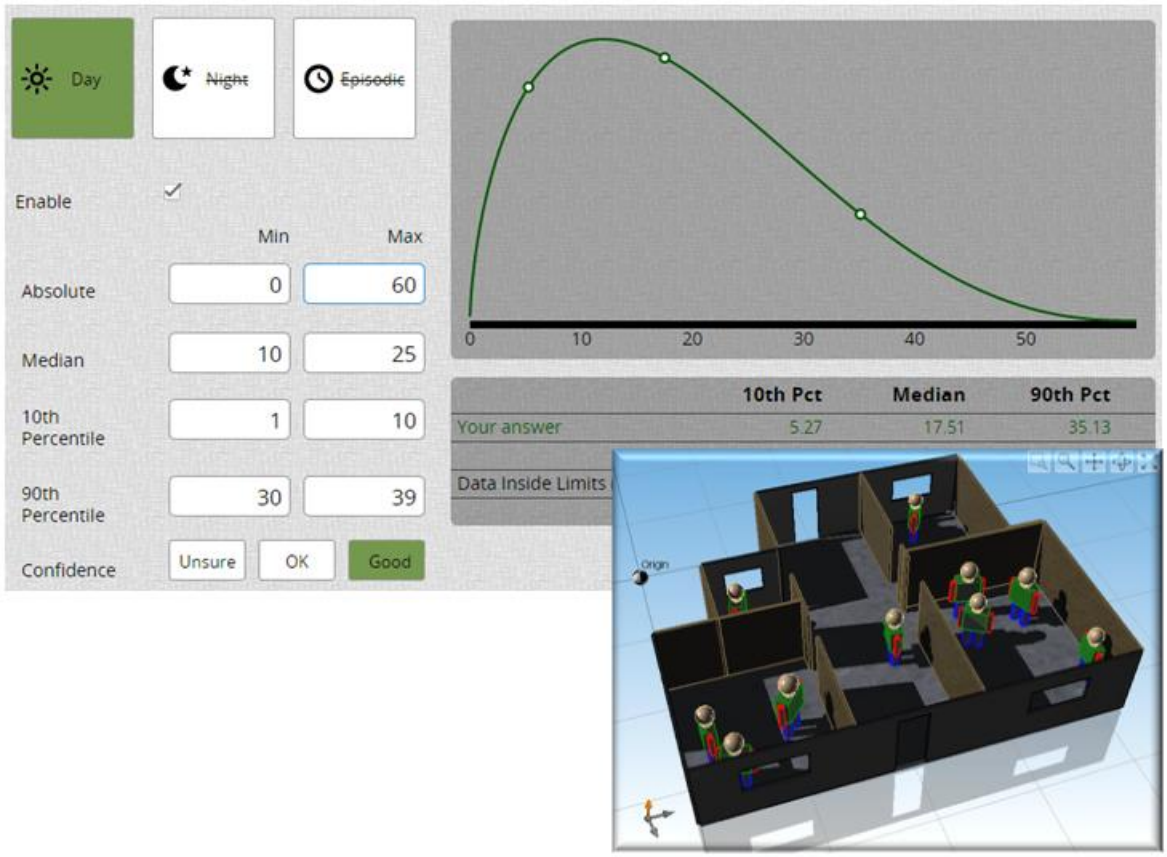
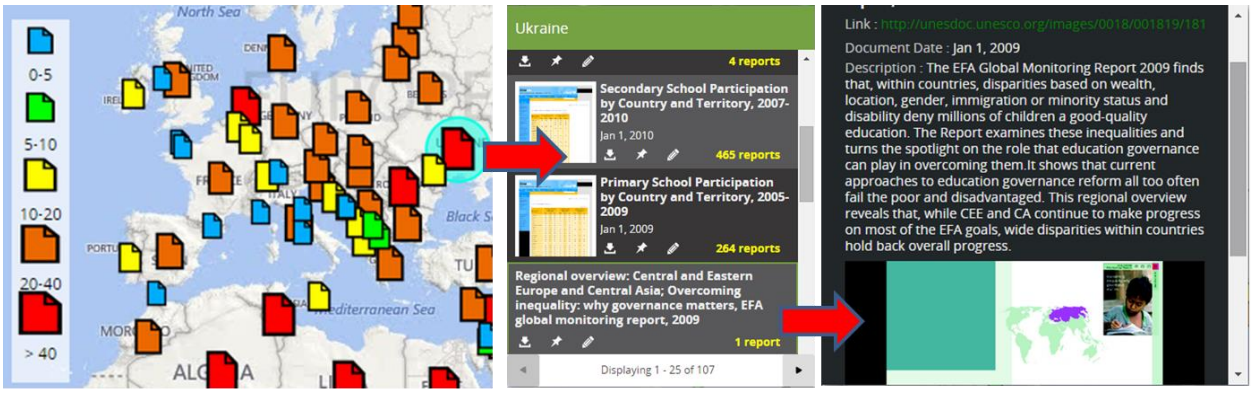


Figure 7. Expert elicitation in PDT system

Using the **Sources module**, analysts can quickly see the geographic distribution of reports across countries (Figure 8a) and drill down into any country for a listing of sources (Figure 8b) including the report itself (Figure 8c). Figure 7 shows the sources for *Urban/Upper Middle Class Residential*.



(a) (b) (c)

Figure 8. PDT source management: a) geographic distribution of sources, b) listing of sources for a country (Ukraine), and c) a source with required information.

Using the **results module** analysts can see the facility model based on the posterior's most probable parameter pair along with the Low, Middle, and High (10^{th}, 50^{th}, and 90^{th} percentiles) for each building type in the geographic region of interest. Figure 8 shows the spatial distribution and tabular PDT model estimates for Ukrainian subnational daytime urban upper/middle class residential. Users can download national and regional occupancy results in Excel, XML or PDF format.

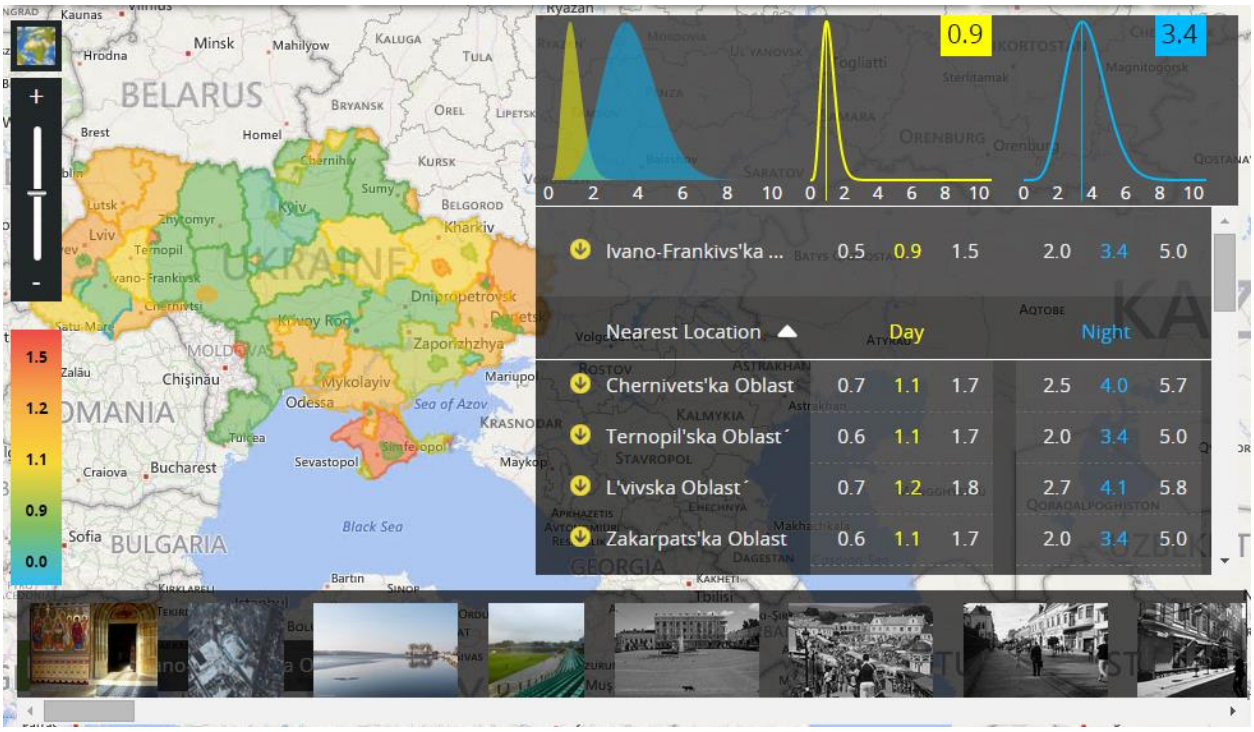


Figure 8. PDT model results for Ukraine sub-national urban upper/middle class residential dwellings.

Although the PDT system currently holds[11] 6200+ documents, 2900+ web pages, 5900+ images, and 162 datasets, for many country/category combinations no data or priors are yet available. Though much work remains to be done in providing complete geographic coverage, a couple of factors may help expedite the progress. First , both the method and the tools have matured sufficiently to engage external partners in academia, agencies operating in the countries of interest, and other data sources (e.g. GED4GEM). We anticipate this will considerably expand the growth and maturity of the data in the years to come. Secondly as a PDT categorical model matures within a particular area (e.g. Europe) posterior distributions could inform new priors in previously unmodeled countries with similar socioeconomic and cultural conditions. In this way, a knowledge ecology could emerges that continues to expand geographically gaining defensible and transparent ground in estimating occupancy globally. This is currently an active area of research and the results are not available as of yet.

7. Summary

We have developed a probabilistic system for the estimation of building occupancy called the Population Data Tables. PDT contributes to the advancement of building occupancy estimation in several ways. First, by adopting

[11] Statistics are current as of April 2014.

Bayesian analytics, a systematic approach is developed that arranges data harmonization techniques into a knowledge learning framework that evolves over time. Secondly, rather than producing a single answer through deterministic mechanisms, PDT stochastically models occupancy as a probability distribution function. This permits analysts to assess variability and extent, to assess likelihood of specific occupancy values, and to produce practical ranges for occupancy (e.g. 10^{th} -90^{th} percentile). Additionally, expert judgment has been a mainstay of occupancy estimation for many years. Under this approach, subject matter expertise and its associated uncertainties are now formally and explicitly integrated with data in the production of these probability distribution functions. Using Bayesian analytics, the PDT model facilitates the evolution of occupancy estimation first initiated by subject matter expertise (prior) and then fine-tuned over time by data (posterior). Model uncertainty is represented by parameter uncertainty and is given by the posterior distribution. Occupancy data, inferred from the open source via observation models, is also uncertain and the PDT model accounts for the impacts to posterior distribution estimation. Furthermore, data uncertainty is decomposed and attributed to particular input variables which allow modelers to assess the relative importance and impact that open source uncertainty has on estimation. Finally, the challenge of estimating occupancy globally under the PDT model led to the development of an interdisciplinary geospatial informatics system that provides the framework for evolving occupancy estimation systematically over time. Contributions to or requests for access to the PDT system are welcomed by the authors. Finally we anticipate this approach is flexible and may be useful to other scientists interested in developing and using similar occupancy data (e.g. PAGER and GEM).

8. References

Albert I, Donnet S, Guihenneuc-Jouyaux C, Low-Choy S, Mengersen K, Rousseau, J (2012) Combining Expert Opinions in Prior Elicitation. Bayesian Analysis 7(3): 503-532

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

Beresovsky V, Burt C, Parsons V, Schenker N, Mutter R (2011) Application of Hierarchical Bayesian Models with Poststratification for Small Area Estimation from Complex Survey Data. American Statistical Association Joint Statistical Meetings Miami, FL

Berger J (2010) Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York

Bernardo J (2003) Bayesian Statistics. In: Viertl R (ed) Encyclopedia of Life Support Systems, Probability and Statistics UNESCO Oxford, UK

Bhaduri B, Bright E, Coleman P, Urban M (2007) LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal 69(1): 103-117

Billari F, Graziani R, Melilli E (2012) Stochastic population forecasts based on conditional expert opinions. Journal of the Royal Statistical Society Series A (Statistics in Society) 175(2): 491-511

Bolstad W (2007) Introduction to Bayesian Statistics. John Wiley & Sons, Hoboken, NJ

Bryant J, Graham P (2013) Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources. Bayesian Analysis 8(3) 591-622 doi:10.1214/13-BA820

Buckland S, Newman K, Thomas L, Koesters N (2004) State-space models for the dynamics of wild animal populations. Ecological Modelling 171(1-2): 157-175

Cooke R (1991) Opinion and Subjective Probability in Science. Oxford University Press, New York

492 Dell'Acqua F, Gamba P, Jaiswal K (2013) Spatial aspects of building and population exposure data and their
493 implications for global earthquake exposure modeling. Natural Hazards 68(3): 1291-1309

494 Earle P, Wald D, Jaiswal K, Allen T, Hearne M, Marano K, Hotovec A, Fee J (2009) Prompt Assessment of Global
495 Earthquakes for Response (PAGER): A system for rapidly determining the impact of earthquakes worldwide. United
496 States Geological Survey

497 Elliott M, Little R (2000) A Bayesian Approach to Combining Information from a Census, a Coverage Measurement
498 Survey, and Demographic Analysis. Journal of the American Statistical Association 95(450): 351-362

499 Eguchi R, Goltz J, Seligson H, Flores P, Blais N, Heaton T, Bortugno E (1997) Real-time loss estimation as an
500 emergency response decision support system: the Early Post-Earthquake Damage Assessment Tool (EPEDAT).
501 Earthquake Spectra, 13:4, 815-833

502 FEMA (2011) Hazus 2.0 Manual. https://www.fema.gov/media-library/assets/documents/21879, Accessed 6/2014

503 French S (2011) Aggregating expert judgment. Revista de la Real Academia de Ciencias Exactas, Fisicas y
504 Naturales Serie A Matematicas 105(1): 181-206

505 Gamba P, Cavalca D, Jaiswal K, Huyck C, Crowley H (2012) The GED4GEM Project: development of a global
506 exposure database for the Global Earthquake Model initiative. 15th World Conference on Earthquake Engineering,
507 Lisbon, Portugal

508 Garthwaite P, Kadane J, O'Hagan A (2005) Statistical Methods for Eliciting Probability Distributions. Journal of the
509 American Statistical Association 100(470) 680-700

510 GEM, (2014) Global Earthquake Model. http://www.globalquakemodel.org/ Accessed 4/2014

511 Genest C, Weerahandi S, Zidek J (1984) Aggregating opinions through logarithmic pooling. Theory and Decision
512 17(1): 61-70

513 Heid I, Kuchenhoff H, Miles J, Kreienbrock L, Wichmann H (2004) Two dimensions of measurement error:
514 Classical and Berkson error in residential radon exposure assessment. Journal of Exposure Analysis and
515 Environmnetal Epidemiology 14(5)

516 Hong T, Lin H-W (2013) Occupant Behavior: Impact on Energy Use of Private Offices. Berkeley National
517 Laboratory and the Green Energy and Environment Laboratories, Industrial Technology Research Institute, Taiway,
518 ROC

519 Jaiswal K, Wald D (2008) Creating a Global Building Inventory for Earthquake Loss Assessment and Risk
520 Management. United States Geological Survey

521 Jaiswal K, Wald D, Earle P, Porter K, and Hearne M (2009) Earthquake Casualty Models within the USGS Prompt
522 Assessment of Global Earthquakes for Response (PAGER) System. Second International Workshop on Disaster
523 Casualties, University of Cambridge, UK

524 Jaiswal K, Wald D, Porter K (2010) A Global Building Inventory for Earthquake Loss Estimation and Risk
525 Management. Earthquake Spectra 26(3): 731

526 Jaiswal K, Wald D (2010b) Development of a Semi-empirical Loss Model within the USGS Prompt Assessment of
527 Global Earthquakes for Response (PAGER) System. United States Geological Survey

Jaiswal K, Wald D, Earle P, Porter K, and Herne M (2011) Earthquake casualty models within the USGS Prompt Assessment of Global Earthquakes for Response (PAGER) system. Human Casualties in Earthquakes, Springer: 83-94

Kolendo A, Frumkin P (2012) Case Study: The Art Institute of Chicago and the Decision to Start Building. The Harris School of Public Policy at the University of Chicago

Martani C, Lee D, Robinson P, Britter R, Ratti C (2012) ENERNET: Studying the dynamic relationship between building occupancy and energy consumption. Energy and Buildings 47(0): 584-591

Melfi R, Rosenblum B, Nordman B, Christensen K (2011) Measuring building occupancy using existing network infrastructure. Proceedings of the 2011 International Green Computing Conference and Workshops, IEEE Computer Society: 1-8

Meyn S, Surana A, Lin Y, Oggianu S, Narayanan S, Frewen T (2009) A sensor-utility-network method for estimation of occupancy distribution in buildings. 48th IEEE Conference on Decision and Control

Morton A (2013) A Process Model for Capturing Museum Population Dynamics Mathematics, California State Polytechnic University.

Mugglin A, Carlin B (1998) Hierarchical Modeling in Geographic Information Systems: Population Interpolation over Incompatible Zones. Journal of Agricultural, Biological, and Environmental Statistics 3(2): 111-130

Mugglin A, Carlin B, Gelfand A (2000) Fully Model-Based Approaches for Spatially Misaligned Data. Journal of the American Statistical Association 95(451): 877

Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A Tale of Many Cities: Universal Patterns in Human Urban Mobility. PLoS ONE 7(5): e37027 doi:10.1371/journal.pone.0037027

Phillips L (1999) Group Elicitation of Probability Distributions: Are Many Heads Better than One? In: Shanteau J, Mellors B, Schum D (eds) Decision Science and Technology: Reflections on the Contributions of Ward Edwards. Kluwer Academic Publishers, Norwell, MA

Pujol G (2007) Sensitivity Package, R package version 1.1

Saltelli A, Tarantola S, Chan K (1999) A quantitative model-independent method for global sensitivity analysis of model output Technometrics. 41(1), 39-56

Schlich R, Axhausen K (2003) Habitual travel behaviour: Evidence from a six-week travel diary. Transportation 30(1): 13-36

Sharpe E, Skeggs T, McNaught S, Saraceno V, Stapley-Brown V (2013) Visitor Figures 2013. The Art Newspaper, Allemandi Publishing, New York.

Stewart R, White D, Urban M, Morton A, Webster C, Stoyanov M, Bright E, Bhaduri B (2013) Uncertainty quantification techniques for population density estimates derived from sparse open source data. Proceedings of the SPIE: Geospatial InfoFusion III (refereed) 8747: 874705-874705

Sutton P, Elvidge C, Obremski T (2003) Building and Evaluating Models to Estimate Ambient Population Density. Photogrammetric Engineering & Remote Sensing 69(5) 545-553

Tan Z, Xi W (2003) Bayesian analysis with consideration of data uncertainty in a specific scenario. Reliability Engineering & System Safety 79(1): 17-31

565 Trendafiloski G, Wyss M, Rosset P (2011) Loss Estimation Module in the Second Generation Software QLARM
566 Human Casualties in Earthquakes Springer: 95-106

567 Wald D, Jaiswal K, So E, Gracia D, Marano K, Lin K, Hearne M, Greene M, D'Ayala D, Crowley H, Gamba P,
568 Porter K (2011) The Role of PAGER in Improving Global Hazard, Building, and Loss Inventories. Seismological
569 Society of America Annual Meeting, Memphis (TE), Seismological Research Letters

570 WHE (2014), World Housing Encyclopedia Project, http://www.world-housing.net/, last accessed 11/19/2014

571 Wyss M, Tollis S, Rosset P, Pacchiani F (2012) Approximate Model for Worldwide Building Stock in Three Size
572 Categories. Report for World Agency of Planetary Monitoring and Earthquake Risk Reduction