Can social media play a role in developing building occupancy curves for small area estimation?

Robert Stewart*, Jesse Piburn, Eric Weber, Marie Urban, April Morton, Gautam Thakur, Budhendra Bhaduri

Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831

Abstract

As small area population estimation continues to advance toward increasingly finer spatial scales, population modelers are ultimately confronted with understanding the occupancy of actual buildings and facility spaces. The need for characterizing building occupancy intersects a number of other related disciplines including green building technologies, natural hazards risk analytics, and public safety occupancy standards. While much has been gained in using survey diaries, sensor technologies, and downscaling data fusion techniques, there is a growing interest in the use of social media data to further illuminate building use patterns. This paper initiates a research agenda in this domain by proposing an explicit social media model that assists in delineating and articulating the opportunities, challenges, and limitations of using social media in the specific case of building use dynamics.

Keywords: social media, building, occupancy, population, agenda.

1. Introduction

Small area population estimates the distribution of individuals at high spatial and temporal resolution by disaggregating coarse population data with local ancillary data (e.g. land use, imagery, etc.). For example LandScan USA estimates populations for day and night at a 3 arc-sec (~90m^2) spatial resolution by considering local ancillary information about facilities, transportation, and businesses (Bhaduri et al. 2007). Continuing this form of disaggregation at finer spatiotemporal scales requires focus on *facility* dynamics. The Population Density Tables (PDT) project has responded in part by estimating ranges for average day and night population density for 50+ facility types (Stewart et al., 2015). To move to finer temporal resolution, modelers need occupancy signatures that characterize population density over smaller time intervals within specific timeframes. Some facilities (e.g. theaters) may be able to monitor occupancy (e.g. ticket sales) in near real time. Research in using building sensor technologies as means for detecting occupancy continues to grow (e.g. Melfi et al. 2011) as well. Unfortunately, paying for these data over many facilities is expensive and only applies to a small subset of facility types.

How viable is social media data in developing occupancy signatures? Specifically, given aggregate count data for a specified timeframe (e.g. daily visitors), can social media further disaggregate this into smaller time steps (e.g. hourly visitors). We motivate the problem by proposing an occupancy model that explicitly situates social media at the

center of estimation. We conclude with a research agenda based on the needs of applying this model in practice.

2. Unit Occupancy

To begin, we establish a timeframe and temporal resolution (time step). Examples include timeframes of monthly, daily, or hours of operation. The time step is smaller and is the desired temporal resolution. Because a selected timeframe may vary by institution (e.g. hours of operation), we normalize the timeframe and time step onto the unit domain [0, 1] without loss of generality. Because different institutions vary by popularity, we normalize the amplitude onto [0, 1] as well. This produces a *unit occupancy model* which can be used to disaggregate timeframe data to the finer scale given by the selected temporal time step. Figure 1 shows an example.

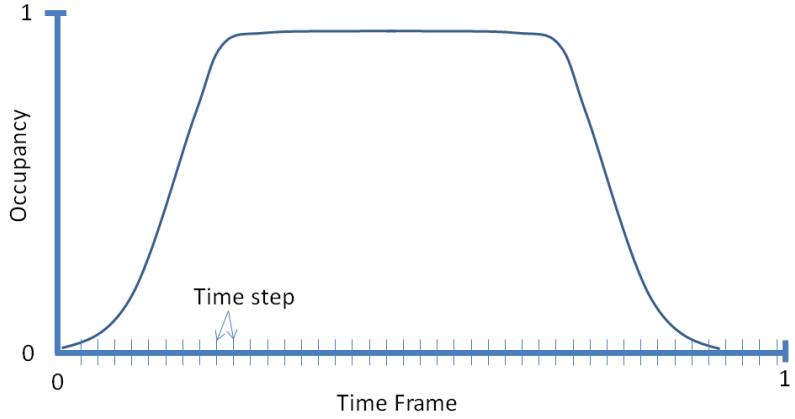


Figure 1. A unit occupancy signature can be scaled to specific facility and time frame.

The model for occupancy at time t, O_t, is structured in terms of normalized visitor[1] arrival times a_i and visit duration v_i of the ith visitor (Equation 1).

$$O_t \sim K \sum_{i=1}^{V} f(a_i, v_i, t) \tag{1}$$

$$\text{Where } f(a_i, v_i, t) = \begin{cases} 1 & if\ t \in [a_i, a_i + v_i] \\ 0 & other\ wise \end{cases}$$

Here K normalizes maximum occupancy to 1 and V is the number of arrivals occurring within the timeframe. The idea is simply that the proportion of people present at any given moment is the sum of people arriving before and leaving after the moment. The focus is therefore on developing arrival and visit duration models. Scaling the model to any particular setting then amounts to scaling the number of arrivals during the timeframe (*V*).

The cost of continuous observation (e.g. video), makes estimation of arrival and visit times over repeated time frames estimation intractable. In lieu of this, we focus on more readily available social media as an indicator of facility population dynamics. We begin

[1] The term visitor means more broadly the "occupant" and includes visitors, employees and so forth.

by defining the term *social media author* (SMA) as any individual producing social media content that indicates presence at an institution under study. We further our inquiry by writing the occupancy model for the SMA sub-population.

3. Social Media Unit Occupancy

Given a social media filters applied to the social media stream(s) a set of SMA authors and posts indicating presence at a facility is identified. Let $SMA \equiv \{sma_1, \ldots, sma_Q\}$ represent the set of social media authors and $e_i \equiv \{e_i^1, \ldots, e_i^N\}$ represent the set of relevant posts for the ith SMA. If $e_i^{range} = e_i^N - e_i^1$ indicates the minimum visit duration, we model the full visit duration of the ith SMA as the conditional distribution $v \sim \varphi(\cdot \,|e_i^{range})$ where the visit duration is modeled as a random variable parameterize by external, ancillary visitor data (e.g. Stewart et al. 2015, Morton, 2013). The arrival time of an individual SMA is modeled as

$$a_i \sim \lambda(a_i^{min}, e_i^1 | \delta) \tag{2}$$

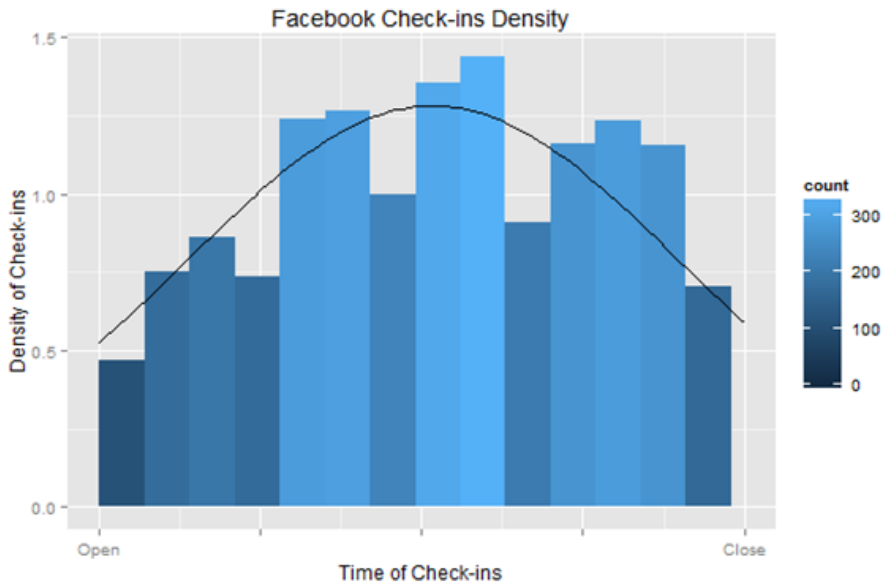
Where $a_i^{min} = e_i^1 - (v_i - e_i^{range})$ where a_i^{min} and e_i^1 respectively the earliest and latest arrival time possible for the ith SMA and δ parameterizes the individual arrival time λ (e.g. uniform). Given the set of *sample* arrivals a_i from one or more institutions, we fit a continuous *population* arrival model (Eq. 3).

$$p(a) = \theta(\pi) \tag{3}$$

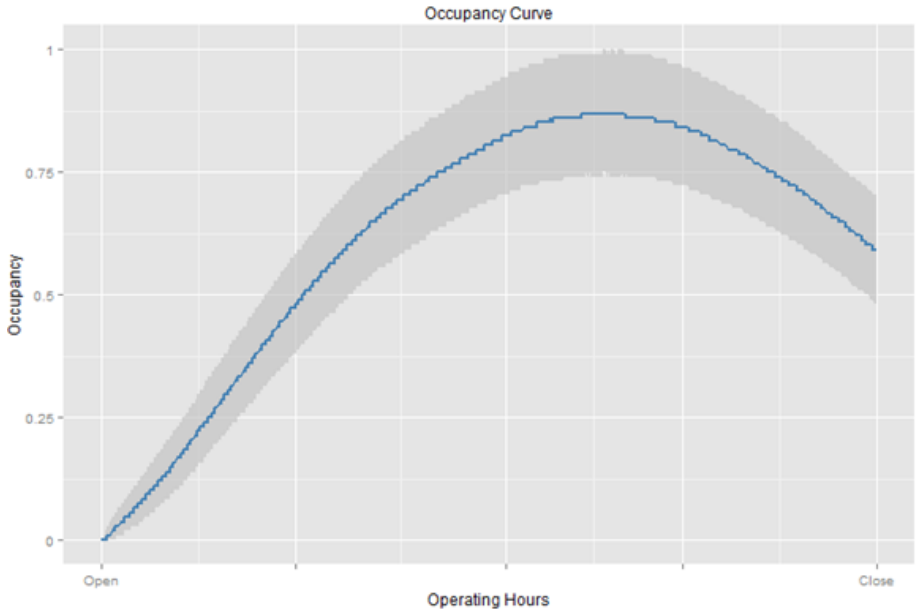
Occupancy curve estimation is carried out by drawing realizations from the population arrival model and then conditional realizations from the duration model giving pairs $(a_k, v_k | a_k)$. Calculation of O_t for SMAs is according to Eq. 1 is straightforward. Inferring occupancy to population assumes that SMA arrivals and visit durations are not fundamentally different than those of the population. If this critical assumption is valid, then SMA occupancy dynamics is an estimate of population dynamics.

Results

We apply the model to High Museum Facebook check-ins collected on 15 minute intervals from 9/6/2013-1/8/2014 using the Facebook Graph API. The timeframe is daily hours of operation and the time step is 1 minute intervals. We adopt a uniform visit duration model U[1,3] for large museums (Stewart et al., 2015) with individual check-in times coincident with arrival times. Figure 2a shows the fitted population arrival model and Figure 2b the resulting SMA occupancy model with 95% confidence bands.



(a)



(b)

Figure 2. a) Arrival times data and model and b) Unit occupancy (blue line) with uncertainty (gray bands)

The results suggest that SMAs arrive uniformly throughout the day creating a gradual increase in the facility population which peaks about 2/3 through the day. With over 400,000 annual visitors visiting across 311 days of operation (www.high.org), the museum averages 1280 visitors per day. We simulate 1280 visitors per day using the unit occupancy to disaggregate people counts throughout the day. The shape will be the same,

but the amplitude now reflects total visitor counts. For example, peak occupancy occurs near 2pm and average between from 380-424 people

4. A Model Based Research Agenda

The model structure is valuable in that it supports articulation and discussion of specific research needs in responsibly applying the model across multiple facilities.

- Which facilities will generate a viable amount of social media data and which do not? For example, it is unlikely that libraries generate as much SMA content as museums.
- Are SMA arrivals and visit times substantially different than the general population? Is there anything different about SMAs that would cause them to arrive and remain in very different patterns than the rest of the population?
- The previous two points suggest great value in furthering knowledge about the specific relationship between demographics and facility popularity. Could separate surveys of SMAs and non-SMAs support this inquiry?
- Can we develop adequate filters to identify spatially unreferenced SMAs and how much error might be incurred? Georeferenced data is at best only a small percentage of the social media volume and may create problems for narrowing filters.
- What are cost effective means of validating this model results? Validating the model will require some collaboration/direct observation for at least some time steps.

The first two questions are really about data sufficiency and are not unique to this problem. The third question points to the novelty and benefit of this approach. Here the question is moved from how well SMA counts represent population counts to how well SMA arrivals represent population arrivals. The SMA subpopulation for one facility may be 1% and another 30%, but if SMA arrival and visit duration curves are separately representative, then unit occupancy can theoretically disaggregate timeframe population totals into smaller time steps. Validation could be addressed for some facilities through collection of open source time step data (e.g. museum monthly totals for validating annual-to-month disaggregation) or through limited real time observations. More research and development is required in each of these areas.

We present here a novel model for leveraging and illuminating the challenges of social media data in estimating occupancy dynamics. The model presented here formally encapsulates social media contributions and focuses data quality concerns explicitly on when SMAs arrive and departure and narrowly focuses the goals in follow-on research efforts.

5. Acknowledgement

This manuscript has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, 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.

6. References

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

Melfi R, Rosenblum B, Nordman B, and 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

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

Stewart, R.N., M Urban, A Morton, and S Duchscherer (2015/submitted), *A Bayesian Machine Learning Model for Estimating Building Occupancy from Open Source Data*, Natural Hazards.