Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

BUILDING CROWD MOVEMENT MODEL USING SAMPLE-BASED MOBILITY SURVEY

Larry J. J. Lin
Shih-Fen Cheng
Hoong Chuin Lau

School of Information Systems
Singapore Management University
81 Victoria St, 188065, SINGAPORE

ABSTRACT

Crowd simulation is a well-studied topic, yet it usually focuses on visualization. In this paper, we study a special class of crowd simulation, where individual agents have diverse backgrounds, ad hoc objectives, and non-repeating visits. Such crowd simulation is particularly useful when modeling human agents movement in leisure settings such as visiting museums or theme parks. In these settings, we are interested in accurately estimating aggregate crowd-related movement statistics. As comprehensive monitoring is usually not feasible for a large crowd, we propose to conduct mobility surveys on only a small group of sampled individuals. We demonstrate via simulation that we can effectively predict agents' aggregate behaviors, even when the agent types are uncertain, and the sampling rate is as low as 1%. Our findings concur with prior studies in urban transportation, and show that sampled-based mobility survey would be a promising approach for improving the accuracy of crowd simulations.

1 INTRODUCTION

Agent-based modeling and simulation (ABMS) is becoming increasingly popular among researchers from various domains. However, a major challenge facing all ABMS researchers is how to accurately create behavioral models for human agents. Traditionally, such models can be created based on theories (without empirical inputs), surveys, controlled lab experiments, or small-scale field observations. Yet for larger-scale models and simulations, these approaches could easily suffer from sampling or study biases and more comprehensive observations are usually necessary. However, comprehensive behavioral studies are very expensive and mostly infeasible in practice, therefore researchers still have to rely heavily on recruiting a small number of committed participants.

In recent years, with the proliferation of sensor-rich smartphones, the use of smartphones in complementing traditional behavioral studies is starting to pick up among ABMS researchers and practitioners. In particular, the field of transportation is leading this trend in utilizing such technologies in conducting *mobility surveys*. The purpose of traditional mobility surveys is to understand city-wide time-dependent origin-destination flows (which is time-dependent), and also how commuters choose their transportation modes. Such information is critically necessary for building city-wide traffic simulations that could assist city planners in evaluating important policies such as infrastructure building, public transport planning, or urban development (some notable ABMS for urban transportation include MATSim (Raney et al. 2003) and DynaMIT (Ben-Akiva et al. 1998)). The estimation of these vital parameters used to depend on paper-based mobility surveys that last for several weeks; during the survey period, all survey participants will have to diligently record their transportation-related decisions on a daily basis.

Some of these fact-collecting processes can now be replaced by smartphone Apps. For example, in the case of the Future Mobility Survey (FMS) (Pereira et al. 2013), smartphone Apps for Android and iOS were developed to capture user GPS locations without major human interventions. Besides significantly

lowers the data collection cost, the data collected in the FMS is also automated and much more accurate (eliminating human errors such as forgetting to log down the daily responses can be eliminated), owing to the range of sensors (e.g., accelerometer, assisted GPS) present in smartphones. Although traditional surveys are still necessary in collecting information related to "traveler's intentions" or reasons/motivations behind decisions, the overall efforts and cost needed for collecting all these facts have been significantly reduced.

While smartphones have definitely helped in the collection of trajectory movement data for identifying *recurring daily routines*, it is less obvious whether such approach would be equally effective if we intend to measure *ad hoc* human behaviors (e.g., visitor's movement patterns within a museum or a recreational facility such as theme park). There are some preliminary studies on the use of smartphones in collecting mobility traces in a leisure setting (Cheng et al. 2013), yet unlike the transportation domain, the effectiveness of smartphone-based mobility survey in this context is difficult to validate.

In the application area of transportation, aggregate city-wide traffic flows can be easily measured using sensors such as induction loops at intersections, probe vehicles, or traffic cams (Aslam et al. 2012, Xian et al. 2013). These aggregate measurements allow us to validate the correctness of ABMS created using mobility surveys. However, in domains other than urban transportation, the lack of environmental sensors makes it extremely difficult to validate the correctness of the data collected via smartphone-based mobility survey.

The lack of environmental sensors also adds to the difficulty in designing sampling strategy. In urban transportation, it is not too difficult to evaluate the effectiveness of different sampling schemes, since aggregate flows can be easily measured (Ortuzar et al. 2011). Without these global measurements, we cannot even decide the size of samples and who to sample.

In this paper, we aim to formally establish the effectiveness of smartphone-based mobility survey under different sampling ratios in measuring ad hoc, non-repeating human mobility patterns. In particular, we will focus on measuring visitor movement behaviors in a theme park setting. As global measurements are not available, we will instead use an agent-based simulator (created using empirical insights we learned in a separate study (Cheng et al. 2013)) as the ground truth generator. Fictitious survey participants will be selected randomly from the population of visitor agents (in the simulation), and these participants will have their full mobility traces recorded. With this computational experiment, we empirically establish the effectiveness of smartphone-based mobility surveys under different sampling ratios (which are represented by percentage of total population sampled). The most important contribution of the paper is the validation of the promise in using smartphone-based mobility surveys for capturing ad hoc human activities such as visiting theme parks. Although not covered in this study, we believe the obtained human mobility model would be a good foundation for creating high-quality agent-based crowd simulation.

2 PROBLEM DEFINITION

Crowd simulation in the leisure industry is a representative case where the high-level scenario occurs repeatedly (the leisure facility opens it doors to thousands of visitors on a daily basis), yet at individual agent's level, it is very rare to have regular, repeated visits. Coupled with the fact that required environmental sensors are mostly unavailable for validation purpose, this is why despite the promise of smartphone-based mobility survey, we cannot be exactly sure whether such approach would be ideal for creating human mobility patterns in leisure industry.

In this paper, we propose to create a novel agent-based simulator, SimLeisure, based on past empirical observations, so that we can treat it as the ground truth generator that can help us in answering the above question using simulation approach. SimLeisure is designed to have many realistic features reflective of the real-world situations. In particular, the following two features are the most critical ones to our computational experiments:

1. Heterogeneous agent types: visitors in a typical leisure setting can vary greatly in preferences over attractions, energy levels (which determines how long they would stay), and group dynamics (e.g., traveling as a group of young adults or with young children). And most important of all, the composition of visitor population will change on a daily basis. As argued earlier, this is one of the major reasons why it is difficult to build behavioral models for such ad hoc, non-repeating human movements. SimLeisure is designed to ensure that a diverse collection of agent types can be easily created and generated randomly.
2. Mesoscopic modeling: a mesoscopic model is positioned to be between strategic and microscopic models. It has the benefit that strategic model in that most secondary, not-so-important details are excluded, yet it also includes the most critical microscopic feature for capturing agent's detail behaviors. In the leisure setting, details such as how visitors roam the street or visit non-essential facilities (e.g., restrooms, restaurants, or shops) are not important for our purpose. However, the most important behavior that needs to be modeled in detail is how agents queue up for major attractions. In SimLeisure, such queueing at major attractions is modeled exactly, while other non-essential agent-level details are not modeled for simplicity and tractability.

With SimLeisure, we can then perform computational experiments by sampling agents to track in the simulation, record the mobility traces of these sampled agents, and then compare the obtained mobility model against the global mobility model generated by SimLeisure. This setting also allows us to evaluate the effectiveness of different sampling ratios.

3 SIMLEISURE: AN AGENT-BASED CROWD SIMULATION FOR LEISURE INDUSTRY

Due to the lack of actual ground truth of trajectories of all visitors in most leisure settings (visiting museums, theme parks, exhibitions), we created SimLeisure to act as a ground truth generator. This is inevitable as necessary environmental sensors are mostly not available for capturing global crowd movement patterns. Therefore, it is imperative to note that the focus of this paper is not a comprehensive validation of SimLeisure, which is impossible due to the aforementioned reason (lack of environmental sensors). Instead, the focus here is on the use of SimLeisure as a ground truth generator, where the movement of agents are based on a decision model trained from data from an actual mobility survey. From the ground truth generated from SimLeisure, we then establish the effectiveness of smartphone-based mobility surveys. As highlighted earlier, SimLeisure is designed to enable the modeling of heterogeneous agent types and realistic queueing at attractions. With these two major features incorporated, we will be able to observe emergent phenomena such as congestion and queue times at attractions, which result from decisions made at agent level. This is a classic use of ABMS (Bonabeau 2002) and the reason why we adopt the ABMS paradigm. SimLeisure is developed using NetLogo 5.1, along with customized Java extensions.

3.1 Agent's Decision Model

As mentioned previously, we are interested in a special class of crowd simulation where an individual agent seeks to maximize his pleasure (expressed as a utility value) by performing ad hoc actions, such as visiting attractions. In the case of theme park visits, agent's preference can be defined as his preferred attraction features. The utility value he can receive by visiting an attraction can then be quantified by combining the base value (objectively speaking, how popular this attraction is) and the *fit* between this attraction's attributes and agent's preference.

As an agent needs to constantly make decisions regarding which attractions to visit, the set of his choices is discrete and finite, and can be best modeled using discrete choice model. At very high level, a discrete choice model is essentially a probabilistic function that takes in a vector of utility values of all alternatives, and returns a probability distribution over all alternatives. Intuitively speaking, the higher the utility value of a choice, the probability assigned to this choice should monotonically go up.

There are a number of potential functions that we can use in converting utility values to the probability distribution over choices. In our study, we adopt *logit function* in modeling discrete choices to be made by leisure-seeking agents. The logit choice probability model is one of the most widely used discrete choice model, and chosen due to the fact that the formula for the logit choice probability model takes a closed form and is readily interpretable (Train 2009). Also, it is a good fit for our case due to its ability to model observed (e.g., queue time and visitor preferences in this case study) and latent factors (other exogenous factors not captured during mobility survey) which might affect an agent's decision. This will enable us to build an agent-based simulation model that not only acts as a ground truth generator, but also one that closely resembles *on-the-ground* movement by the theme park visitors.

Upon consultation with the theme park operator that we are collaborating with, there are four major factors that would determine agent's evaluation of the value of an attraction: there are thrill, darkness, and wetness levels of the attraction, and the distance an agent needs to travel to get to the attraction node.

Table 1: Variables in agent's logit choice model.

Variable	Description
T_n	thrill level of attraction n
D_n	darkness level of attraction n
W_n	wetness level of attraction n
$DT_{v,n}$	absolute difference between visitor $v's$ thrill preference and attraction $n's$ thrill level
$DD_{v,n}$	absolute difference between visitor $v's$ darkness preference and attraction $n's$ darkness level
$DW_{v,n}$	absolute difference between visitor $v's$ wetness preference and attraction $n's$ wetness level
$DS_{v,n}$	distance (number of zones) visitor v has to travel from current location to get to attraction node n

Following the logit choice model literature, we define the agent's decision model by first defining the following utility function, which depends on agent v's preference and attraction n's attributes:

$$U_{v,n} = \beta_1 DT_{v,n} + \beta_2 DD_{v,n} + \beta_3 DW_{v,n} + \beta_4 DS_{v,n} + \varepsilon_{v,n},$$

where $\varepsilon_{v,n}$ denotes unobserved factors that contribute to visitor's decision. Given the set of all attractions, N, the probability that agent v would choose attraction n is then defined based on the utility function.

$$P_{v,n} = \frac{e^{U_{v,n}}}{\sum_{j \in N} e^{U_{v,j}}}, \ \forall \ v,n.$$

3.2 Training of Agent's Decision Model

For the training of logit model, we will be using data collected from our mobility survey conducted (with 50 participants) on December 2012 (Cheng, Lin, Du, Varakantham, and Lau 2013). What it encompass will be the QR (Quick Response) scans by visitors, which represents the visitors entering attractions in the theme park, as well as the GPS logs of the participants, which are recorded at 1 minute intervals. Preferences of the users, such as preference towards thrill, darkness and wetness levels (on a scale of 1 to 3) are also obtained through the iOS application.

The data is then post-processed as follows. For every GPS data point, we'll have to convert the latitude and longitude (lat-lon point) captured to an attraction. This is done by setting a radius from the lat-lon point captured, and obtaining the nearest attraction to this lat-lon point. The lat-long points are next grouped into intervals of 30 minutes. For every time interval, there will be an attraction with the highest frequency. For example, if 25 minutes (out of 30) in an interval was spent at Attraction A, Attraction A will be the attraction with highest frequency for that interval.

After obtaining the attraction with highest frequency for every interval, we will then merge the GPS data with the QR scans data as illustrated in Figure 1. A visitor was found to be at attractions A, B, and

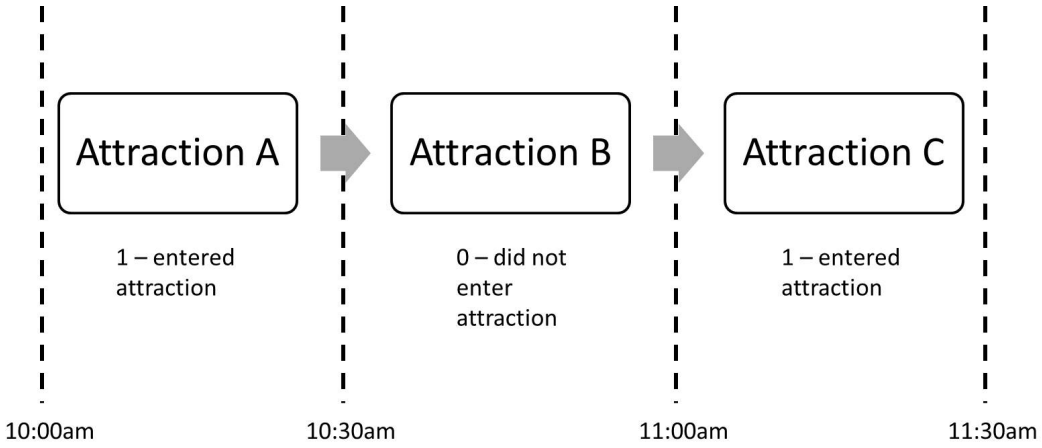

Figure 1: Post-processing of data from mobility survey.

C (through frequency of GPS) at time intervals 1, 2 and 3 respectively. If the visitor did visit attractions A and C at time interval 1 and 3, the visitor will obtain the logit result of 1. For time interval 2, the visitor will obtain a logit result of 0 (for attraction B), which means that the visitor was found to be there (according to frequency of GPS), but did not enter the attraction.

The above process is repeated for all visitors, and the final data is used for training (through a logistic regression) to obtain the coefficients of the logit model.

3.3 Heterogeneous Agent Classes

The actual distribution of preferences of the overall visitor population should be unknown to the planner of mobility surveys. Thus, for the generation of visitor classes, we will generate C number of visitor classes, each with their own preferences as shown below. The preferences of each visitor class will be generated randomly, so as to simulate the real environment, where the distributions of preferences of visitors in the theme park are unknown. These preferences are as stated below:

$$\text{thrill preference of visitor } v\text{: } T_v = [0,3]$$
$$\text{darkness preference of visitor } v\text{: } D_v = [0,3]$$
$$\text{wetness preference of visitor } v\text{: } W_v = [0,3]$$

On top of that, for the V number of user classes generated, there will be no two visitor classes with the same $< T,D,W >$ combination.

3.4 Generation of User Itineraries

After the visitor classes are generated, we will proceed to generate N_c number of itineraries for each visitor class c. Figure 2 depicts the process of adding attractions to a visitor class's itinerary through the logit model.

Throughout the generation process, we will maintain two list, $attraction\_list$ and $itinerary\_list$. During initialization, $attraction\_list$ will contain the full list of attractions (17 attractions), while $itinerary\_list$ will be empty. An attraction will then be chosen randomly (with equal distribution) from $attraction\_list$, and a random float ($rand\_float$: 0 to 1) will be generated. At the same time, we will generate a logit probability, $P_{v,n}$ (further elaborated in subsequent sections) of visitor v going to this attraction n. If $rand\_float$ is less than or equals to the $P_{v,n}$ calculated, we will proceed to check the time budget constraint. The time budget for the itinerary is set as 320 minutes, which is the average time that visitors (data collected from mobility survey) spends at the theme park. The time spent for an attraction is the average time a user spends at an

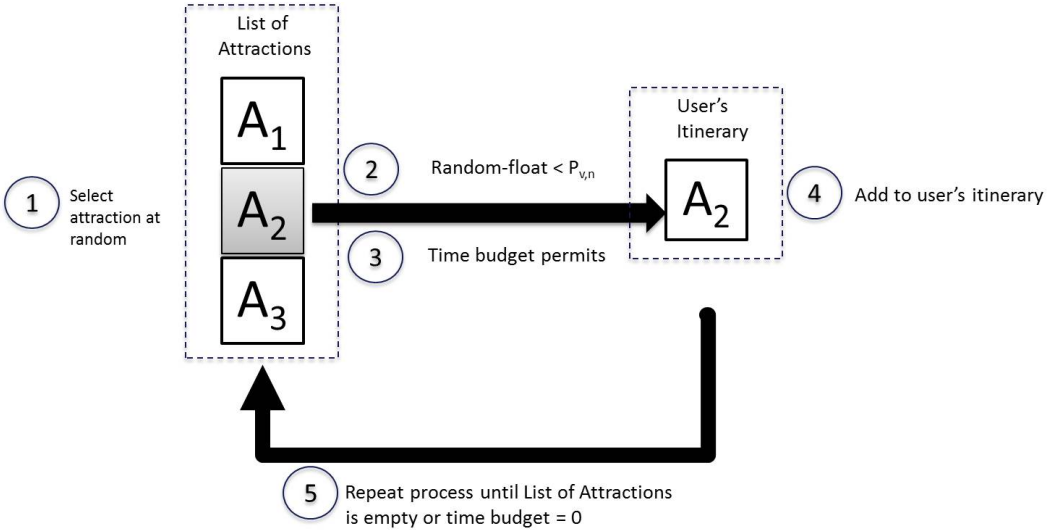

Figure 2: Process depicting the generation of itinerary of visitors.

attraction (from mobility survey). If the time budget constraint is passed, the attraction is then added to the *itinerary_list*, as well as removed from the *attraction_list*.

Step 1 to 4 in Figure 2 is repeated until all attractions are removed from *attraction_list*, or if the time budget runs out. This process for a visitor class c is then repeated C times, and the total number of itineraries generated will be $N_c \times C$.

After the itineraries are generated, they are then passed as inputs to our simulation model. In our simulation model, every agent (visitor) will select an itinerary at random (with equal probability). The simulation process is repeated 30 times, signifying 30 days of simulation. The itinerary selection process is repeated for each round or "day" of simulation. This means that the distribution of itineraries may / may not be the same for every round of simulation.

3.5 Representing Mobility Patterns

The major output we want to obtain from SimLeisure is the aggregated mobility pattern. In our context, we define mobility pattern as *time-dependent transition between nodes*, and this information can be summarized as a T by N by N matrix, where each matrix element (t,i,j) denotes the aggregate transition probability $P_{i,j}^t$, which is defined as the probability of visitors going from node i to node j at time t. $P_{i,j}^t$ can be calculated directly by:

$$P_{i,j}^t = \frac{F_{i,j}^t}{\sum_{j \in N} F_{i,j}^t}, \forall\, i,t,$$

where $F_{i,j}^t$ refers to the number of visitor agents going from node i to node j at time t. Since $P_{i,j}^t$ refers to transition probability, $\sum_{j \in N} P_{i,j}^t = 1$.

For the sampled population to be track via mobility survey (we set the sampling ratios to be 5% and 1% in our computational experiments to be presented later), we will generate two other time-dependent transition matrices $\{\widehat{P_{i,j}^t}\}$ and $\{\check{P}_{i,j}^t\}$ for the 5% and 1% sample respectively (in the same manner as shown above). This will be used as a basis of comparison between the sampled participants and full population. The comparison metrics will be elaborated in the subsequent sections.

4 DIFFUSION DYNAMICS MODEL

In order to evaluate the effectiveness of the sampling-based mobility survey, we will be comparing the results to the ones obtained by the diffusion dynamics model, a model that can be used to derive flow patterns using observed congestion at attraction nodes (Du et al. 2014). The diffusion dynamics model is essentially an optimization model that learns the probability of a visitor going from node i to node j, using only congestion information at nodes (i.e., waiting times at attractions), instead of directly deriving from the actual *on-the-ground* movements (data is not available) of all visitors at the theme park.

In the diffusion dynamics model proposed by (Du et al. 2014), the movement of visitors are modeled as a multinomial distribution based diffusion model, which is a form of dependent cascade model. Formally, the likelihood is defined as:

$$\mathscr{L}(p;x,n) = \prod_{d\in D}\prod_{i\in A}\prod_{t\in T}\frac{(\sum_j x_{d,t,i,j})!}{\prod_{j\in A} x_{d,t,i,j}!}\prod_{j\in A} p_{t,i,j}^{x_{d,t,i,j}}, \tag{1}$$

where the total outflow of visitors from attraction node i at time t is $\sum_j x_{d,t,i,j}$, which corresponds to the total number of trials in the multinomial distribution.

Table 2: Variables in diffusion dynamics model.

Variable	Description
D	observed cascades
A	set of all attractions in the theme park
T	set of time slices
S_i	service rate at attraction i
$n_{d,t,i}$	number of visitors waiting to be serviced at node i, time t in cascade d
$x_{d,t,i,j}$	corresponds to the number of people moving from node i to j
$p_{t,i,j}$	probability of a visitor moving from node i to node j at time t

Through the modeling of visitors movement, the probability of a visitor moving from node i to j at time t ($P^t_{i,j}$)) is then estimated (through a maximum likelihood estimation) from aggregate (queue-time) observations. According to the authors, the diffusion dynamics model proposed is able to provide a prediction accuracy of about 80% for popular attractions. We are interested in comparing our *sampling-based mobility surveys* against the diffusion dynamics model, and see if our approach can generate better mobility pattern estimations.

One major reason why we compare against the diffusion dynamics model is its practicality: the input (queue time observations) for the model are readily available, either via real-world observations or the execution of SimLeisure. The output from the diffusion dynamics model is identical to our estimation approach, which is essentially a time-dependent transition matrix ($P^t_{i,j}$). Therefore, in the simulation model, other than obtaining the sequence and time of visit of visitors, we will additionally generate the queue or number of visitors at attraction node n at time t (denoted by $Q_{n,t}$). This $Q_{n,t}$ is then fed as input to the diffusion dynamics model to generate $\widetilde{P^t_{i,j}}$, which is used as comparison against $\widehat{P^t_{i,j}}$ and $\check{P^t_{i,j}}$ (generated by the sampling process).

5 PERFORMANCE EVALUATION AND RESULTS

The three time-dependent transition matrices generated by 1% and 5% sampled mobility survey, and the diffusion dynamics are compared using two performance indices. The first criterion is a straightforward *sum of errors*, and the second criterion is the *weighted sum of errors*, considering the importance of different links. These two comparison criteria are described in detail below.

5.1 Sum of Errors

The sum of errors measures the difference of an approach (versus the actual population) over all possible $< i, j, t >$ edges. Formally, it is defined as:

$$\widehat{SE} = \sum_{i,j,t} |P^t_{i,j} - \widehat{P^t_{i,j}}|$$
$$\check{SE} = \sum_{i,j,t} |P^t_{i,j} - \check{P^t_{i,j}}|$$
$$\widetilde{SE} = \sum_{i,j,t} |P^t_{i,j} - \widetilde{P^t_{i,j}}|$$

Table 3: Matrices and metrics (non-weighted) for comparison.

Variable	Description
$P^t_{i,j}$	time-dependent transition matrix for entire population
$\widehat{P^t_{i,j}}$	time-dependent transition matrix for sampled population (sampling at 5% of population)
$\check{P^t_{i,j}}$	time-dependent transition matrix for sampled population (sampling at 1% of population)
$\widetilde{P^t_{i,j}}$	time-dependent transition matrix from diffusion dynamics model
$\widehat{SE}$	sum of errors for sampling-based approach (sampling at 5% of population)
$\check{SE}$	sum of errors for sampling-based approach (sampling at 1% of population)
$\widetilde{SE}$	sum of errors for diffusion dynamics approach

Table 5.1 shows the results of comparison between the sampling approach (at 5% and 1%) and diffusion dynamics model. From the results, we can observe that the 5% sampling gives us the best estimation in all instances, as observed by the lowest sum of errors. We also observe that even at 1 % sampling rate the accuracy of the sampling-based approach (as seen in the sum of errors) outperforms that of the diffusion dynamics approach. Another observation is that when experimenting with increasing interval length (30 minutes to 2 hours) for the discrete time interval t, the sum of errors decreases across all three categories. This is not surprising, as the estimation accuracy for a method is expected to decrease with finer granularity in the time interval (lesser data points).

Table 4: Performance comparison using sum of errors.

Time Interval		$\lvert P^t_{i,j} - \widehat{P^t_{i,j}}\rvert$	$\lvert P^t_{i,j} - \check{P^t_{i,j}}\rvert$	$\lvert P^t_{i,j} - \widetilde{P^t_{i,j}}\rvert$
0.5 hr	Sum of Errors	49.77	101.08	362.82
	Standard Deviation	0.04	0.06	0.11
1 hr	Sum of Errors	16.02	35.25	176.85
	Standard Deviation	0.02	0.04	0.1
2 hr	Sum of Errors	5.6	12.04	85.54
	Standard Deviation	0.01	0.02	0.11

5.2 Weighted Sum of Errors

The second evaluation technique is the weighted sum of errors. It is essentially a weighted variant of sum of errors, where the weights represent the importance of time-dependent transition links. With full population transitions, the weight in our evaluation is simply the flow $F^t_{i,j}$. The formal definition of this

performance index is defined as:

$$\widehat{WSE} = \sum_{i,j,t} |P^t_{i,j} - \widehat{P^t_{i,j}}| \times F^t_{i,j}$$
$$\check{WSE} = \sum_{i,j,t} |P^t_{i,j} - \check{P^t_{i,j}}| \times F^t_{i,j}$$
$$\widetilde{WSE} = \sum_{i,j,t} |P^t_{i,j} - \widetilde{P^t_{i,j}}| \times F^t_{i,j}$$

Table 5: Metrics (weighted) for comparison.

Variable	Description
$\widehat{WSE}$	weighted sum of errors for sampling-based approach (sampling at 5% of population)
$\check{WSE}$	weighted sum of errors for sampling-based approach (sampling at 1% of population)
$\widetilde{WSE}$	weighted sum of errors for diffusion dynamics approach

Table 5.2 shows the results of the weighted sum of errors. Through Table 5.2, the performance of the sampling-based approach becomes even clearer, as observed by the larger differences between the techniques. This means that the sampling-based approach makes good estimates for links with high volume of flows (number of visitors), resulting in a much lower weighted sum of errors.

Table 6: Performance comparison using weighted sum of errors.

Time Interval		$\lvert P^t_{i,j} - \widehat{P^t_{i,j}} \rvert$	$\lvert P^t_{i,j} - \check{P^t_{i,j}} \rvert$	$\lvert P^t_{i,j} - \widetilde{P^t_{i,j}} \rvert$
0.5 hr	Weighted Sum of Errors	18,260.35	38,106.30	170,084.54
	Standard Deviation	0.04	0.06	0.11
1 hr	Weighted Sum of Errors	12,889.96	26,096.17	169,722.64
	Standard Deviation	0.02	0.04	0.1
2 hr	Weighted Sum of Errors	9,068.67	18,476.5	173,172.33
	Standard Deviation	0.01	0.02	0.11

5.3 Heat Map Representation

On top of the numerical differences, we used a visual approach for comparing the performance of the various techniques. Figure 3 is a heat map representation of the transition matrix, aggregated over time (i.e., $P_{i,j} = \sum_t P^t_{i,j}$). The probabilities of flows are colored using grayscale, where lighter color represents higher probabilities. From the plot, we can tell how accurate the sampling-based approach is when estimating the mobility patterns of the full population, as compared to the diffusion dynamics approach. This complements the numerical results (Tables 5.1 and 5.2) and we can clearly see that the heat map is indeed consistent with our intuition that higher flows should lead to a better prediction using sampled mobility survey.

6 SENSITIVITY ANALYSIS

For the sensitivity analysis, we wanted to investigate the effect that sampling rate has on the performance metrics (sum of errors and sum of weighted errors). Referring to 4(a) and 4(b), we can observe that with increases in sampling rate, it will lead to better results (i.e. lower sum/weighted sum of errors). This is not surprising, as with more of a population being sampled, it will produce movement pattern estimates that are closer to that of the (full) population. What is worth noting is that increases in sampling rate

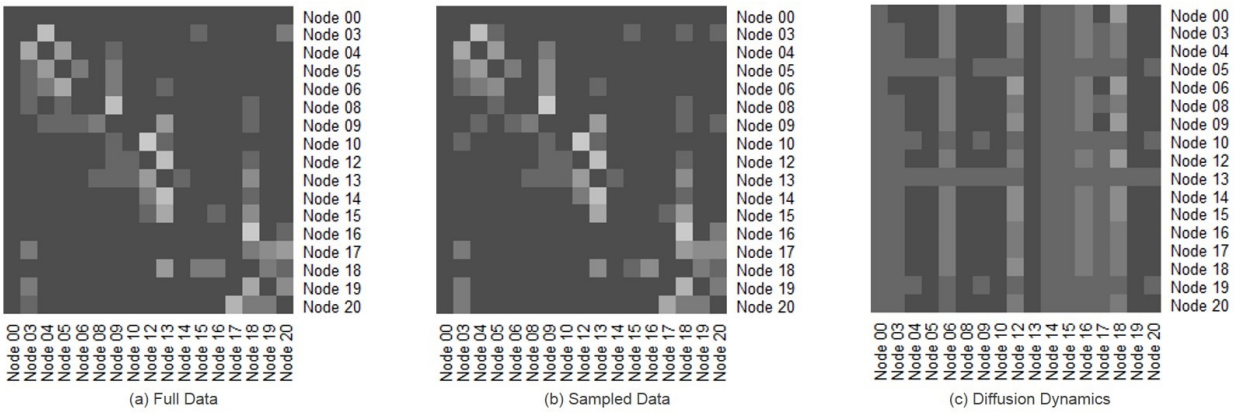

Figure 3: Transition matrices (between attraction nodes) as heat maps: (a) full population, (b) sampled at 5%, (c) diffusion dynamics.

from 1 to 10 percent will lead to dramatic increases in performance (sharp decrease in sum (weighted) of errors), after which increases in sampling rate (10 percent onwards) does not lead to as much an increase in performance. This confirms the effectiveness of smartphone-based mobility survey, where the number of participants usually accounts for less than 10 percent of the actual population.

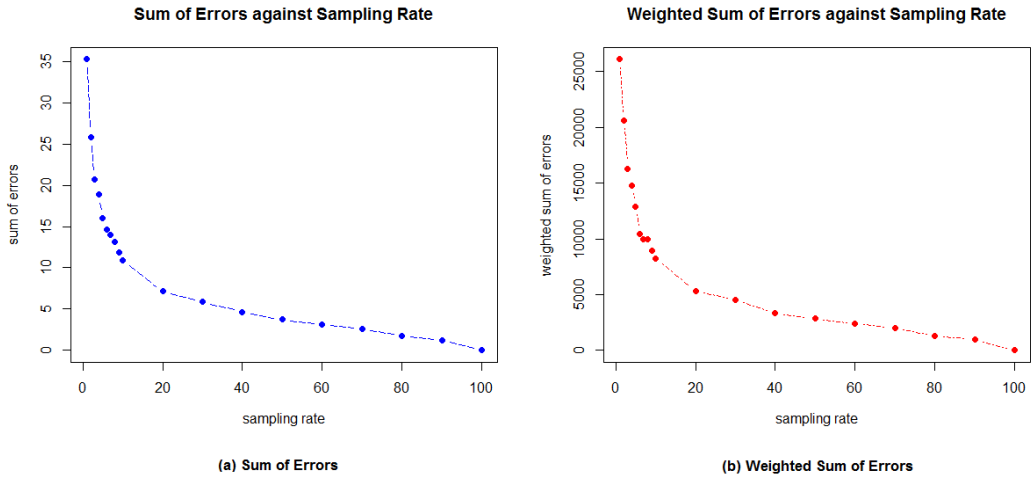

Figure 4: Sensitivity results for sampling rate (dependent variable: (a) sum of errors (b) weighted sum of errors)

7 CONCLUSIONS

Mobility surveys have been shown to be highly effective in domains with recurrent mobility patterns, such as urban transportation. What we demonstrate in our study is its effectiveness in domains where visitors are diverse, with ad hoc objectives, and heterogeneous. By using SimLeisure, an agent-based simulation based on real-world observations, we are able to generate ground truth, and use it to quantify the effectiveness of using sample-based mobility surveys for generating mobility patterns. We contrast the results we obtain against a state-of-the-art statistical approach in diffusion dynamics model, and demonstrate that even with only 1% of sampling ratio, the sample-based mobility survey still outperforms diffusion dynamics model by at least one order of magnitude (using weighted sum of errors as performance index). In this paper,

we demonstrate the usefulness of the sample-based mobility survey in a situation where the flow amongst attractions are uneven (as observed in Figure 3(a)). This justifies the use of mobility surveys in a wide range of similar situations (particularly in modeling human behaviors in the leisure setting). This also lays the foundation for future study on how to more effectively build highly accurate crowd simulation models for leisure settings.

ACKNOWLEDGMENTS

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

REFERENCES

Aslam, J., S. Lim, X. Pan, and D. Rus. 2012. "City-Scale Traffic Estimation from a Roving Sensor Network". In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, 141–154.

Ben-Akiva, M., M. Bierlaire, H. Koutsopoulos, and R. Mishalani. 1998. "DynaMIT: A Simulation-Based System for Traffic Prediction". In *Proceedings of the DACCORD Short-Term Forecasting Workshop*.

Bonabeau, E. 2002. "Agent-Based Modeling: Methods and Techniques for Simulating Human Systems". *Proceedings of the National Academy of Sciences* 99 (suppl 3): 7280–7287.

Cheng, S.-F., L. Lin, J. Du, P. Varakantham, and H. C. Lau. 2013. "An Agent-Based Simulation Approach to Experience Management in Theme Parks". In *Proceedings of the 2013 Winter Simulation Conference*.

Du, J., A. Kumar, and P. Varakantham. 2014. "On Understanding Diffusion Dynamics of Patrons at a Theme Park". In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems*, 1501–1502.

Ortuzar, J. D. D., J. Armoogum, J.-L. Madre, and F. Potier. 2011. "Continuous Mobility Surveys: The State of Practice". *Transport Reviews* 31 (3): 293–312.

Pereira, F., C. Carrion, F. Zhao, C. D. Cottrill, C. Zegras, and M. Ben-Akiva. 2013. "The Future Mobility Survey: Overview and Preliminary Evaluation". In *Proceedings of the Eastern Asia Society for Transportation Studies, Vol.9, 2013*.

Raney, B., N. Cetin, A. Völlmy, M. Vrtic, K. Axhausen, and K. Nagel. 2003. "An Agent-Based Microsimulation Model of Swiss Travel: First Results". *Networks and Spatial Economics* 3 (1): 23–41.

Train, K. E. 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press.

Xian, O. Y., M. Chitre, and D. Rus. 2013. "The Probe Allocation Problem". In *2013 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems*, 50–57.

AUTHOR BIOGRAPHIES

LARRY J. J. LIN is a PhD Student at the Singapore Management University. He received his BSc in Information Systems Management from the Singapore Management University. He is interested in the application of multi-agent-based modeling and simulation in complex business domains. His email address is larry.lin.2013@phdis.smu.edu.sg.

SHIH-FEN CHENG is Associate Professor of Information Systems and Deputy Director of the Fujitsu-SMU Urban Computing and Engineering Corp Lab at the Singapore Management University. He received his Ph.D. degree in industrial and operations engineering from the University of Michigan, Ann Arbor, and B.S.E. degree in mechanical engineering from the National Taiwan University. His research focuses on the modeling and optimization of complex systems in engineering and business domains. He is particularly interested in the application areas of transportation, manufacturing, and computational markets. He is a member of INFORMS, AAAI, and IEEE, and serves as Area Editor for Electronic Commerce Research

and Applications. His email address is sfcheng@smu.edu.sg.

HOONG CHUIN LAU is Professor of Information Systems, Director of the Fujitsu-SMU Urban Computing and Engineering Corp Lab, and Deputy Director of the Living Analytics Research Centre at the Singapore Management University. His research in the interface of Artificial Intelligence and Operations Research has contributed to advances of algorithms in a variety of complex problems in logistics, transportation and travel planning and operations. For his work with the Singapore Ministry of Defense, he won the National Innovation and Quality Convention Star Award in 2006, and was nominated for the prestigious Defense Technology Prize (individual category) in 2007. He was awarded the Lee Kwan Yew Fellowship for research excellence in 2008. He currently serves on the editorial board of the IEEE Transactions on Automation Science and Engineering. His email address is hclau@smu.edu.sg.