Toward Advanced Indoor Mobility Models Through Location-Centric Analysis: Spatio-Temporal Density Dynamics

Mimonah Al Qathrady, Ahmed Helmy
Computer and Information Science and Engineering Department
Gainesville, Florida, USA
{mimonah,helmy}@ufl.edu

ABSTRACT

Building's density, as its number of nodes at a specific period, is a significant parameter that affects mobile and smart applications performances and evaluations. Consequently, the buildings’ temporal density predictions and their nodes spatial distribution modeling have to follow real-world scenarios to provide a realistic evaluation. However, there is lack of real-world building-level density studies that examine these aspects thoroughly. As a result, this work is a data-driven study that investigates the temporal density predictability and spatial density distributions of more than 100 real buildings with ten different categories, over 150 days across three semesters. The study covers the buildings nodes’ temporal modeling and predictions, and their spatial distributions in the building. Seasonal predictive models are utilized to predict hour-by-hour density for a variable length of consequent periods using training data with different lengths. The models include Seasonal Naive, Holt-Winters’ seasonal additive, TBATS, and ARIMA-seasonal. The results show that the Seasonal Naive model is often selected as the best predictive model when training phase covers a shorter period. For example, Seasonal Naive predicted with the least error in 73%, 63% and 57% of cases in summer, spring and fall respectively when using only one week to predict its consecutive five weeks with mean normalized error ∼25% on average. However, when using five weeks of data to predict the sixth week, the TBATS model predicted with the least error in 60%, 54% and 43% of cases in fall, spring and summer respectively with mean absolute error ∼19% on average. When investigating the spatial density distributions, power law, log-logistic and lognormal distributions are usually selected as the first best-fit distributions for 82%, 65%, 62% of buildings in the summer, spring and fall respectively.

1 INTRODUCTION

Several recent mobility models studies have taken a data-driven approach where real data traces are collected and analyzed using user-centric or location-centric approach. Most of the studies were user-centric where they focused on the mobile user historical individual, pairwise or collective behavioral patterns [16]. Location-centric approach focuses on the location and model the statistical characteristics of its users’ visitations regardless of the identity or mobility histories of their users. It involves studying the location density distributions or the location pair-wise characteristics such as the flow of users between locations [8, 15]. Indoor nodes density, as a location-centric individual pattern, is an important parameter since it affects many aspects of wireless characteristics such as capacity and connectivity. Besides, the correlation between encounter events and density are strongly positive in most buildings [7]. Also, many indoor operations depend on the population density updated information, for example, system management of pedestrian flow inside crowd buildings [15]. Therefore, understanding the different aspects of indoor density, and reproducing them is requisite for realistic indoor modeling and smart service designing and evaluation. Some previous models preserve the density distributions, but they targeted outdoor environment [10, 14], and do not cover the temporal density modeling and prediction. This paper, however, studies the spatial distributions of users at the building levels and covers the temporal density predictions using a data-driven approach. Our data include more than 99 million mobile records from more than 100 buildings during three different semesters: spring, summer, and fall. It covers several buildings categories including academic, museums, libraries, labs, administrations offices, sports facilities, dining, theaters, housing and health-care facilities. We study hourly population prediction since it is an important mechanism when designing future advanced indoor services or evaluating them. For example, predicting the nodes number in a critical situation such as emergency evacuation helps to implement the right and efficient plan. Since the spectrum analysis of time series data shows a day as the dominant cycle [7], we use several seasonal models to predict hour by hour density in the buildings. Models include Seasonal Naive [12], Holt-Winters’ seasonal (additive) [12], TBATS [9], and ARIMA-seasonal [13]. More than 5 million hour prediction operations are performed to establish the validity of these models using training and testing data with different length. Moreover, the paper investigates the indoor nodes statistical distributions, and reports the best fit distributions using Kolmogorov-Smirnov statistic (KS-stat). While the power-law was demonstrated as the best fit for outdoor density distribution, it is only reported for 29% to 35% of cases as the first best fit distribution for nodes at the building level.
This section describes the wireless data collection, processing, filtering and its buildings' categorization.

2.1.1 Wireless data. The data are collected from more than 100 buildings on the university campus. The mobile records show when the users are at the edge of the two access points and they are being directed back and forth between the access points. The users here do not change their location, but their records are filtered out. The study assigns \( \lambda = 10 \) seconds. Increasing \( \lambda \) value risks deleting mobile records that are not results of ping-pong effects.

2.1.2 Records Filters. Filters are applied to select phone devices, reduce ping pong effect and eliminate very short sessions.

2.2.1 Temporal Density Prediction. This section describes the training and testing period windows, seasonal prediction models that are used in this study and the assessment metric. 

**Prediction testing and training windows**: we have investigated the temporal density of six consecutive weeks of data for each building in each semester. Since many buildings are not visited on the weekends, we include the weekday data only from 8: am to 8: pm, where the majority of the buildings are occupied during this period.

**Prediction models**: four seasonal predictive models are used to predict hour by hour density. The models are Seasonal Naive, Holt-Winters’ seasonal (additive), TBATS and auto.ARIMA. The following

<table>
<thead>
<tr>
<th>Semester</th>
<th>From</th>
<th>To</th>
<th>Records</th>
<th>Mac#</th>
<th>Aps#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring</td>
<td>12-Mar</td>
<td>30-Apr</td>
<td>40,975,015</td>
<td>100,871</td>
<td>1727</td>
</tr>
<tr>
<td>Summer</td>
<td>22-Jun</td>
<td>10-Aug</td>
<td>14,106,673</td>
<td>66,839</td>
<td>1938</td>
</tr>
<tr>
<td>Fall</td>
<td>6-Sep</td>
<td>25-Oct</td>
<td>43,976,833</td>
<td>103,216</td>
<td>1974</td>
</tr>
</tbody>
</table>

To the best of our knowledge, our investigation of temporal dynamic density predictions and spatial distributions at the building level is the first of its kind in term of data or level of analysis. Besides, the result of this study is expected to have a beneficial impact on the future mobility modeling, evaluation, and designing of IoT applications including the ones related to crowd management, tracking, infection tracing or wireless communications.

2 DATA-DRIVEN BUILDING LEVEL DENSITY ANALYSIS AND MODELING FRAMEWORK

2.1 DataSet

This section describes the wireless data collection, processing, filtering and its buildings’ categorization.

2.1.1 Wireless data. The data are collected from more than 100 buildings on the university campus. The mobile records show when the users are at the edge of the two access points and they are being directed back and forth between the access points. The users here do not change their location, but their records are filtered out. The study assigns \( \lambda = 10 \) seconds. Increasing \( \lambda \) value risks deleting mobile records that are not results of ping-pong effects.

2.1.2 Records Filters. Filters are applied to select phone devices, reduce ping pong effect and eliminate very short sessions.

2.2 Density Analysis and Modeling

The density analysis investigates the temporal predictions and nodes density distributions at the building level.

2.2.1 Temporal Density Prediction. This section describes the training and testing period windows, seasonal prediction models that are used in this study and the assessment metric.
which it is called the Normalized Mean Absolute Error (MAE) with different density. To make it is a useful metric to compare between different buildings.

\[
\hat{y}_{T+h} = \hat{y}_T + h\cdot m, \quad (1)
\]

where \( m \) is the seasonal period and \( k = \lceil \frac{h-1}{m} \rceil + 1 \). In this study, a power spectrum analysis to hour by hour time series of the density data reveals one day as the dominant cycle [7]. As a result, one day is considered as a season when running the model. The Kolmogorov-Smirnov statistic is used to evaluate the fitted distributions. The three best fit distributions are selected. Also, the percentage of distributions that have resulted in KS-stat ≤ 0.075 or 0.10 are reported.

3 RESULT AND DISCUSSION

This section discusses the density predictions for each semester using different lengths of training and testing data period. Also, it summarizes the nodes indoor density distributions.

3.1 Temporal Density Predictions

This section discusses the density predictions from several dimensions: The prediction models, the predicted period length (testing data size), the training model length. Figure 3 shows NMAE for three states of different training and testing data length. The semesters show similar patterns when using the same length of training and testing data period. Another notable result is a simple predictive model such as Seasonal Naive provides a consistent result with less error even when predicting a more extended period such as the case with 3a. As a result, it could be used to model the temporal density in many buildings due to its simplicity and accuracy as it has been shown in this study. More detail about the models that are predicted with the least error for each training and testing data length are presented in [7]. The results show that the Seasonal Naive model is usually selected as the best predictive model when training the model with a shorter period to predict the following periods. For example, Seasonal Naive predicted with the least error in 73%, 63% and 57% of cases in summer, spring and fall respectively when using only one week to predict its consecutive five weeks with average NMAE ~25% on average. Other models are improved when the training period length is increased. For example, when using five weeks of training data to predict the sixth week, TBATS model predicted with the least error in 60%, 54% and 43% of cases in fall, spring and summer respectively with NMAE ~19% on average.

3.2 Density Distributions

The Kolmogorov-Smirnov statistic KS-stat is used to evaluate the fitted distributions. The power law, log-logistic and lognormal distributions are usually selected as the first best-fit nodes distributions for 82%, 65%, 62% of buildings in the Summer, Spring and Fall respectively. Also, Log-normal and power-law are the only two distributions that are reported with KS-stat ≤ 7.5%. The previous investigation at the campus level or park level concludes that the power law distributions are the best fit for the population density [14]. However, this is not always the case with indoor nodes distributions. For instance, the power law is only selected as the first best fit for 29% of the buildings in spring to 35% of buildings in the summer, other distributions such as log-logistic have been selected as the best fit distribution for several buildings. Also, the percentage of buildings that have KS-stat ≤0.10 in their fitted log-logistic distributions are 53% in the spring, while 29% only are fitted power-law with KS-stat ≤0.10 as it is shown in table 2. Also, power law was not reported as one of the three best-fit distributions in

\[
MAE = \frac{1}{T} \sum_{t=1}^{T} |\hat{y}_t - y_t|, \quad (2)
\]

where \( y_t \) denote the \( t \)-th hour density and \( \hat{y}_t \) denote its prediction values, where \( \hat{y}_t \geq 0 \). However, MAE is scale dependent. Therefore, it is not a useful metric to compare between different buildings with different density. To make MAE scale independent, the MAE is then divided by the mean density per hour in the testing period, which it is called the Normalized Mean Absolute Error (NMAE).

2.2.2 The density distributions. The number of users that have associated with an access point on a day at the building is examined.

<table>
<thead>
<tr>
<th>Semester</th>
<th>1st Best Fit</th>
<th>2nd Best Fit</th>
<th>3rd Best Fit</th>
<th>≤7.5% [KS_stat]</th>
<th>≤10% [KS_stat]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring</td>
<td>Pl[29%], LI[18%], Ln[18%]</td>
<td>W[12%], G[12%], Ln[12%]</td>
<td>Ln[12%], E[12%]</td>
<td>Ln[12%]</td>
<td>G[29%], W[24%], C[12%]</td>
</tr>
</tbody>
</table>

some cases such as a Gym building that has 48 Access points and has been visited by more than 2.6K of users in the summer period, where log-logistic, log-normal, and gamma are selected as the three best fit for its indoor nodes distributions on a summer day.

4 CONCLUSION AND FUTURE WORK

Since density, as a location individual pattern, is a very critical parameter for indoor modeling and smart applications evaluation, the study investigated the temporal density predictions and nodes’ spatial distributions at the buildings level. We have analyzed a real trace data at the building level, and we have covered buildings from various categories. Four seasonal models are fitted and evaluated with all combination of training and testing data length. This study is built with the purpose of developing an advanced indoor mobility model that represents the most important observations, and enable accurate simulation of smart indoor services by recreating real scenarios of population density. The mobility model will combine the statistical data from real traces with other contextual information such as buildings layout, constraints and vertical movement between floors. Besides, an in-depth investigation that involves buildings categories and distinguishes their behavioral patterns are planned to be executed and used for realistic category-based mobility models generation. The paper findings are beneficial for a wide range of IoT applications that require information about density or rely on the crowd. In the future, mobility analysis results and tools are expected to be published on [2].

ACKNOWLEDGMENTS

This work was partially funded by Najran University, Saudi Arabia, and NSF 1320694.

REFERENCES