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.

ACM Reference Format:
Mimonah Al Qathrady, Ahmed Helmy. 2018. Toward Advanced Indoor Mobility Models Through Location-Centric Analysis: Spatio-Temporal Density Dynamics. In *21st ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems (MSWIM '18), October 28-November 2, 2018, Montreal, QC, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3242102.3242143

MSWIM '18, October 28-November 2, 2018, Montreal, QC, Canada

ACM ISBN 978-1-4503-5960-3/18/10…$15.00
https://doi.org/10.1145/3242102.3242143

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.

Table 1: Wireless Trace Information

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

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

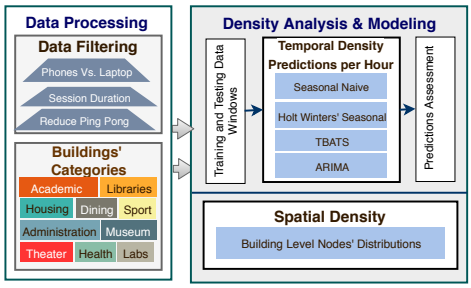


Figure 1: Data Driven Density Analysis Framework

Figure 1 shows the analysis framework where it has two main parts: data processing and spatiotemporal density analysis. The data processing involves data filtering and buildings' categorization. The analysis part consists of temporal and spatial studies. The temporal analysis studies hour by hour density modeling and prediction using variable lengths of training and testing data, while spatial analysis focuses on the nodes' distributions in the buildings.

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 a user starts and ends an association with an access point. Each access point is tagged with the buildings' code and the room number. The traces are for 150 days for three semesters: spring, summer and fall, and includes more than 99 million records. Table 1 provides more information about the wireless data.

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

Filtering phone devices. The data contain phones and laptops mobile records. Most of the people always have their smartphone at their side day and night [1], and they are more likely to use it

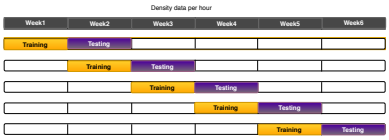


Figure 2: Training and testing data windows when training period= one week, and testing period = one week

while they are moving than laptops users. As a result, we focus on the phones' data and filter out the laptop ones. The device mac address and its website visitations are used to identify if a device is a laptop or a phone. Device types classification are explained [6].

Filtering very short session duration. Since the analysis is performed at the building level, mobile traces of people that are passing outdoor have to be discarded. As a result, records with session duration that are less than θ are filtered out. The study assigns θ to one second. Increasing θ period may risk deleting indoor records.

Filters to reduce the ping-pong effect. The ping-pong effect occurs when the wireless users are at the edge of the two access points and hop between them. The users here do not change their location, but they are being directed back and forth between the access points. Then, the users appear to be repeatedly associated with a fixed number of access points. Due to incomplete information and the ambiguity of ping-pong interpretation, there is no perfect solution to this problem [17]. To reduce the ping-pong effect in this study, we build a filter which discards the records that have association back and forth between two access points in less than λ. We run the ping-pong filters and assigns λ =10 seconds. Increasing λ value risks deleting mobile records that are not results of ping-pong effects.

2.1.3 Assigning Buildings Category. The buildings have been categorized into academic, administration, labs, dining, housing, sports facilities, museum, libraries, theaters and auditorium and health care. Online information such as campus map and building information [3–5] assisted the categorization process.

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.

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. Buildings that are not occupied during this period are filtered out. Figure 2 shows the windows when we have training data length = one week, and testing period is one week. All possible combination of different windows in terms of weeks are investigated.

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

paragraphs describe them briefly. You can see [12] for more details.
Seasonal Naive: it is the simplest model, where the predicted value is equal to the last value from the same season.

$$\hat{y}_{T+h|T} = y_{T+h-km}, \tag{1}$$

where m is the seasonal period and $k = \lfloor \frac{h-1}{m} \rfloor + 1$. 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.
Holt-Winters Seasonal Model:it consist of the forecast equation and three smoothing equations: one for the level, one for trend, and one for the seasonal component, with smoothing parameters α, β^* and γ. In this study, we used the additive seasonal method since the seasonal variation is roughly constant. The seasonal component is expressed in absolute terms in the scale of the observed series, and in the level equation the series is seasonally adjusted by subtracting the seasonal component [12].
TBATS Model: it is *T*rigonometric regressor with *B*ox Cox Transformations, *A*RMA errors, *T*rend and *S*easonality. It is used to model series exhibiting multiple complex seasonalities. It uses a combination of Fourier terms with an exponential smoothing state space model and a *Box-Cox* transformation [9].
ARIMA Model: it stands for Autoregressive Integrated Moving Average and describes the autocorrelations in data. A seasonal *ARIMA* model is formed by including additional seasonal terms in the *ARIMA* model and can be written as $ARIMA(p,d,q)(P,D,Q)_m$, where m= number of periods per season. The uppercase notation is used for the seasonal parts of the model, and lowercase notation for the non-seasonal parts of the model. The seasonal part of the model consists of terms that are very similar to the non-seasonal components of the model, but they involve backshift of the seasonal period. For example, d is the order of first differencing, and D is the order of seasonal differencing. Autoregressive $AR(p)$ implies current values depend on its p-previous values. Moving average $MA(q)$ means the current deviation from the mean depends on q-previous deviations, where q is the order of *MA* process. We used auto. Arima model, which return the best *ARIMA* model according to information criterion (*AIC*), but it is not necessary to be the best in term of prediction error. The order of differencing d is based on the *KPSS* test, while the order of seasonal differencing is based on *OCSB* test [13]. We have used forecast tool provided in [11] for fitting the models and predicting the desnity.
Prediction Assessments The mean absolute error have been used widely for prediction assessment. *MAE* can be computed as:

$$MAE = T^{-1} \sum_{t=1}^{T} |\hat{y_t} - y_t|, \tag{2}$$

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. Eleven distributions are used to fit the data using the maximum likelihood methods. The distributions are Power-law (Pl), Weibull (W), gamma (G), lognormal (Ln), Pareto(Pr), Normal(N), Exponential (Ex), Uniform (U), Cauchy (C), Beta (B) and Log-logistic (Ll). The *KS-stat* 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* $\leq$ 0.075 or 0.10 are reported.

3 RESULT AND DISCUSSION

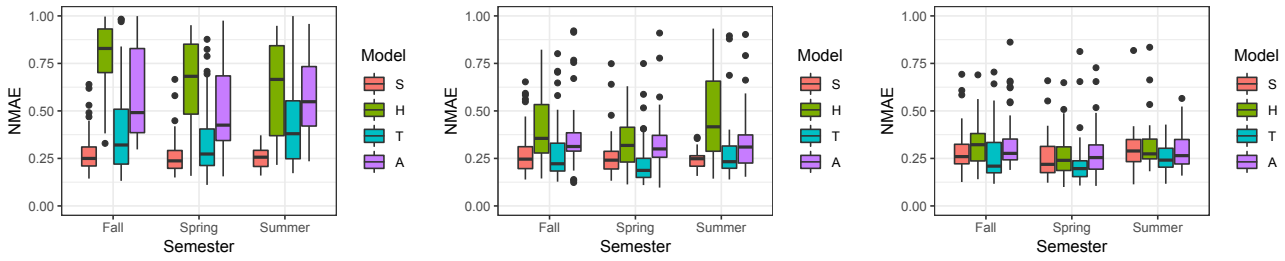
This section discusses the hour-by- hour density predictions for each semester using different lengths of training and testing data period. Also, it summarizes the nodes indoor density distributions.

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* $\leq$ 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* $\leq$0.10 in their fitted log-logistic distributions are 53% in the spring, while 29% only are fitted power-law with *KS-stat* $\leq$0.10 as it is shown in table 2. Also, power law was not reported as one of the three best-fit distributions in



(a) Training: 1, testing: 5

(b) Training: 3weeks, testing: 3weeks

(c) Training: 5weeks, testing: 1 week

Figure 3: Normalized MAE of density per hour predictions for each semester with variable length of testing and training data. S: Seasonal Naive, H: Holt-Winters, T: TBATS, A: ARIMA

Table 2: Modeling empirical density distributions of population density in 23 buildings. W: Weibull, Pl: Power Law, G: Gamma, Ll: Log Logistic, Pr: Pareto, Ln: Log Normal, N: Normal, B: Beta, E: Exponential

Semester	1st Best Fit	2nd Best Fit	3rd Best Fit	$\leq$7.5%$[KS_stat]$	$\leq$10% $[KS\_stat]$
Fall	Pl[31%], Ll[31%], W[15%] C[10%]	Ln[26%], Ll[21%], Pl[11%] W[11%], G[11%], C[11%]	G[26%], W[21%], Ll[16%] Pl[11%], C[11%]	Ll[5%]	Pl[31%], W[%26], C[%16], Ll[%16], G[11%], Ln[11%]
Summer	Pl[35%], Ll[29%], Ln[18%] W[12%]	Ll[24%], C[18%], Pl[18%] W[12%], G[12%], Ln[12%]	W[29%], Ll[29%], G[12%] Ln[12%], E[12%]	Ll[18%], Pl[12%] Ln[12%]	Ll[%41], Pl[29%], Ln[%29], G[%29], W[%24], C[%12]
Spring	Pl[29%], Ll[18%], Ln[18%] G[12%]	Ll[41%], G[18%], Ln[18%] Pl[12%]	Ln[24%], Ll[24%], Pl[24%] W[12%], G[12%]	Ll[11%]	Ll[%53], G[%35], Ln[%35] , Pl[29%], W[11%]

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

[1] 2016. mobile market statistics. (2016). https://deviceatlas.com/blog/16-mobile-market-statistics-you-should-know-2016.
[2] 2018. Mobility Analysis Results and Tools. (2018). https://cise.ufl.edu/ Helmy; https://cise.ufl.edu/ qathrady.
[3] 2018. UF Campus classrooms. (2018). https://classrooms.at.ufl.edu/classroom-info/pictures-and-info/prettyPhoto.
[4] 2018. UF Campus Map. (2018). https://campusmap.ufl.edu/.
[5] 2018. UF Named Facilities. (2018). https://www.uff.ufl.edu/Facilities/ListingName.asp.
[6] Babak A., Leonardo T., Aaron Y., Roozbeh K., Joerg O., and Ahmed H. 2018. Flutes vs. Cellos: Analyzing Mobility-Traffic Correlations in Large WLAN Traces. In *IEEE INFOCOM*. IEEE.
[7] Mimonah A. and H.Helmy. 2018. Location-Centric Analysis For Advanced Indoor Mobility Models: Spatio-Temporal Density Dynamics-Technical Report. (2018). https://cise.ufl.edu/ qathrady/reports/BuildingsDensityAnalysis.pdf.
[8] Mimonah Al Qathrady and Ahmed Helmy. 2017. Location-Centric Flow Flux for Improved Indoor Mobility Models. In *Proceedings INFOCOM*. IEEE.
[9] Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder. 2011. Forecasting time series with complex seasonal patterns using exponential smoothing. *J. Amer. Statist. Assoc.* 106, 496 (2011), 1513–1527.
[10] Danielle Lopes Ferreira, Bruno AA Nunes, and Katia Obraczka. 2014. On the heavy tail properties of spatial node density for realistic mobility modeling. In *(SECON)*. IEEE, 504–512.
[11] Rob H., George A., Christoph B., Gabriel C., Leanne C., Mitchell O, Fotios P., Slava R., Earo W., and Farah Y. 2018. *forecast: Forecasting functions for time series and linear models.* http://pkg.robjhyndman.com/forecast R package version 8.4.
[12] Rob J H. and George A. 2014. *Forecasting: principles and practice.* OTexts.
[13] Rob J H., Yeasmin K., et al. 2007. *Automatic time series for forecasting: the forecast package for R.* Number 6/07. Monash University.
[14] Bruno N. and Katia O. 2011. On the invariance of spatial node density for realistic mobility modeling. In *MASS*. IEEE, 322–331.
[15] Lorenz S., Martin W, and Philipp M. 2014. Estimating crowd densities and pedestrian flows using wi-fi and bluetooth. In *11th Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ICST, 171–177.
[16] Gautam S Thakur and Ahmed Helmy. 2013. COBRA: A framework for the analysis of realistic mobility models. In *INFOCOM, 2013*. IEEE.
[17] Jungkeun Y., Brian D N., Mingyan L., and Minkyong K. 2006. Building realistic mobility models from coarse-grained traces. In *Mobile systems, applications and services*. ACM, 177–190.