11-2015

Where are the passengers? A Grid-Based Gaussian Mixture Model for taxi bookings

Meng-Fen CHIANG
Singapore Management University, mfchiang@smu.edu.sg

Tuan Anh HOANG
Singapore Management University, tahoang.2011@smu.edu.sg

Ee-Peng LIM
Singapore Management University, eplim@smu.edu.sg

DOI: https://doi.org/10.1145/2820783.2820807

Citation

CHIANG, Meng-Fen; HOANG, Tuan Anh; and LIM, Ee-Peng. Where are the passengers? A Grid-Based Gaussian Mixture Model for taxi bookings. (2015). *SIGSPATIAL '15: Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems: November 3-6, 2015, Seattle, Washington*. 32:1-10. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/3170

Where are the Passengers? A Grid-Based Gaussian Mixture Model for Taxi Bookings

Meng-Fen Chiang
Living Analytics Research Centre
Singapore Management University
mfchiang@smu.edu.sg

Tuan-Anh Hoang
Living Analytics Research Centre
Singapore Management University
tahoang.2011@smu.edu.sg

Ee-Peng Lim
Living Analytics Research Centre
Singapore Management University
eplim@smu.edu.sg

ABSTRACT

Taxi bookings are events where requests for taxis are made by passengers either over voice calls or mobile apps. As the demand for taxis changes with space and time, it is important to model both the space and temporal dimensions in dynamic booking data. Several applications can benefit from a good taxi booking model. These include the prediction of number of bookings at certain location and time of the day, and the detection of anomalous booking events. In this paper, we propose a **G**rid-based **G**aussian **M**ixture **M**odel (GGMM) with spatio-temporal dimensions that groups booking data into a number of spatio-temporal clusters by observing the bookings occurring at different time of the day in each spatial grid cell. Using a large-scale real-world dataset consisting of over millions of booking records, we show that GGMM outperforms two strong baselines: a Gaussian Mixture Model (GMM) and the state-of-the-art spatio-temporal behavior model, Periodic Mobility Model (PMM), in estimating the spatio-temporal distribution of bookings at specific grid cells during specific time intervals. GGMM can achieve up to 95.8% (96.5%) reduction in perplexity compared against GMM (PMM). Further, we apply GGMM to detect anomalous bookings and successfully relate the anomalies with some known events, demonstrating GGMM's effectiveness in this task.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Application—*Data Mining, Spatial databases and GIS*

Keywords

Taxi demand modeling, spatial-temporal dynamics, unified grid-based Gaussian mixure model

1. INTRODUCTION

Motivation. Many large cities today are facing increasing challenges in managing high commuting demand using public transportation. While public transportation is often perceived to involve trains and buses only, one should also consider taxis as yet another form of public transportation as it is a kind of resource shared among commuters. Taxi is a flexible and efficient means for commuters to get from one place to another. The demand for taxis dynamically changes due to daily patterns of people movement and possibly other events. A good understanding of how the demand changes relative to space and time is thus critical to the formulation of effective transportation policies to regulate the number of taxis and to devise a reasonable taxi fare system. Taxi operators, knowing the demand ahead of time, would also be able to direct taxis to the right place at right time to maximize taxi utilization thereby reducing the cost of empty taxis running on the roads and maximizing revenues.

In this paper, our research goal is to model spatio-temporal dynamics of taxi bookings to capture the demand for taxis. We conduct our research using three months of taxi booking data with more than few million bookings from approximately hundreds of thousands commuters in Singapore [1]. The data is collected as commuters make taxi bookings using a mobile app. While our research is motivated by taxi bookings, the models developed in this work can also be applied to modeling other spatio-temporal events (e.g., longitudinal sightings of forest fires, disease outbreak reports, etc.).

Taxi booking modeling research must consider several important factors arising from both the *static* and *dynamic* nature of taxi demand. The static factor concerns the landscape constraint on the possible locations where taxi bookings can be made. For example, very few or zero taxi bookings should be observed in forest or water areas. The dynamic factor concerns: (1) the daily movement of commuters which changes with location and time of the day, and (2) the daily movement patterns of commuters which are different between weekdays and weekends. For example, the spatial distribution of bookings in Figure 1(a) suggests that during morning rush hour [09:00, 10:00), most of the taxi bookings come from the peripheral areas of Singapore, which are residential areas. This suggests a high demand for taxis to ferry commuters to work. In contrast, the taxi bookings during evening rush hour [18:00, 19:00) mainly come from the central business district of Singapore as commuters make their way home. Thus, our first challenge is: *How can we infer the underlying spatial and temporal distributions that continuously change*

SIGSPATIAL'15 November 03 - 06, 2015, Bellevue, WA, USA.
Copyright 2015 ACM 978-1-4503-3967-4/15/11 ...$15.00
DOI: http://dx.doi.org/10.1145/2820783.2820807 .

[1] We cannot provide precise statistics and app name in order to comply with the non-disclosure agreement.

overt time from past taxi bookings?

There have been a number of studies devoted to modeling human mobilities [1][5]. Lichman et al. [5] studied the behavior of geolocation/check-in data without timestamps and proposed a mixture model to predict spatial distributions. In consequence, the temporal dynamics cannot be reflected. Cho et al. [1] proposed Periodic Mobility Model (PMM) to model human mobilities from check-in data. PMM is a spatial-driven model that generates spatial and temporal dimensions of bookings in a pipeline. As a result, the cluster structure of PMM exhibits clear distinctions in spatial dimension as well as in temporal dimension. Limited by its description power, PMM is ineffective especially when there are more than one spatial temporal clusters sharing the same spatial region or when there are clusters that have clear temporal boundaries but do not have clear spatial boundaries. Thus the second challenge is: *How can we expressively model the nature of spatial and temporal distribution from past taxi bookings?*

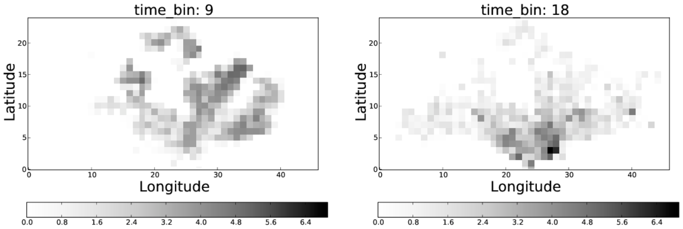
To address the challenges, we propose a continuous and unified spatio-temporal model for the purpose of modeling and predicting spatio-temporal dynamics of taxi demands. Our research combines Poisson processes that give the number of daily events occurring within a grid cell and three-dimensional Gaussian processes that generate multiple spatio-temporal centres of a day. To the best of our knowledge, the problem of unified modeling spatio-temporal dynamics of taxi demand using time and pick-up locations of bookings from passengers has not been studied so far.

Research Objective. In this work, we propose a Grid-based Gaussian Mixture Model (GGMM) to model spatio-temporal taxi bookings using a mixture of Gaussian processes that determine the locations and time points of bookings. This model summarizes the dynamic locations and times of bookings into a few clusters. GGMM also divides the city into small grid cells using a Poisson process to model the daily rate of bookings in each grid cell. This addresses the static factor due to landscape constraint as grid cells in a sparsely- (or highly-) populated areas will be assigned with low (or high) booking rates. To address the dynamic factor caused by the weekday and weekend difference in daily movement patterns, we learn two models, one for weekdays and another for weekends.

Using GGMM, we can perform at least two important tasks: Firstly, we can predict the demand for taxis anywhere, anytime. Secondly, we can also use GGMM to detect anomalous events that trigger an abnormal distribution of taxi bookings. These are also the two major application tasks illustrated in this paper. We thus summarize the contributions of this paper as follows:

- We propose a continuous and unified Grid-based Gaussian Mixture Model (GGMM) to model spatio-temporal taxi bookings considering both their static and dynamic factors. This model extends the well-known Gaussian Mixture Model (GMM) to consider grid-based booking rate so as to reflect the landscape constraints.

- We perform empirical analysis of large-scale booking data using GGMM and obtain interesting insights about the taxi booking patterns in Singapore. In addition to the differences between weekday and weekend bookings, we observe that rush hour bookings in the morning and evening are quite distinctive.



(a) Morning Rush Hour (b) Evening Rush Hour

Figure 1: Spatial dynamics of bookings: spatial distributions of number of bookings in log scale during morning rush hour(a) and during evening rush hour(b).

- We conduct a rigorous evaluation of GGMM in comparison with GMM and PMM when applying them to a prediction task using the booking data. Our experiments show that GGMM yields more accurate prediction results than the other two models.

- We further apply GGMM to detect anomalous booking data and relate them to offline events in the city. The results show that GGMM is also effective in this task.

The remainder of this paper is organized as follows. A survey of related work is presented in Section 2. Section 3 covers data description and problem definition. Section 4 and Section 5 detail proposed model and empirical analysis. Section 6 reports the experimental evaluation. Section 7 presents the application of anomaly detection using our model. Section 8 concludes and discusses research directions.

2. RELATED WORKS

Urban Computing. Our research is related to urban computing [20]. This research includes human mobility modelling[1, 5, 4, 13], urban planning in transportation [15, 21, 14, 16, 10], etc.. Lichman et al. [5] proposed a mixture model to predict spatial distributions of geolocation/check-in data. While their model effectively captures human mobility by interpolating mixtures of individual and population spatial distributions, the temporal aspect is not addressed. Cho et al. [1] proposed to model human mobilities(check-in data) using separate spatial and temporal Gaussian components with social influence. Unlike [1], we propose a unified model to model and predict spatio-temporal dynamics of taxi demand.

One of major interests of urban planning in transportation is traffic route prediction[15, 16]. Yuan et al. [16] proposed a recommendation framework for both taxi drivers and passengers based on passenger mobility patterns and taxi driver behaviour from their GPS trajectories. With detailed trajectories, the parking places of taxi drivers are determined. The paper also proposed a probabilistic model to detect the states (e.g., occupied/cruising/parked) of a trajectory segment for a working taxi. [8]

Recently, there have been a number of emerging applications driven by urban planning, such as travel cost estimation [11, 12] and refuel behavior sensing[18]. Zheng et al. proposed a travel time estimation model for any path based on the trajectories of vehicles [11]. Liu et al. used taxi trajectory data to learn drivers' routing decisions [6]. The above works have however relied heavily on vehicle trajectory data, which is different from passenger bookings which are a kind of point data instead of movement data.

Table 1: Definitions of symbols

Sym.	Definition
D	number of days
X	bookings over D days
$X^{(d)}$	bookings on day d
$X_g^{(d)}$	bookings in grid g on day d
$X_{g,t}^{(d)}$	bookings in grid g, time bin t on day d
G	number of grids
K	number of booking clusters
T	number of time bins in a day
θ_k	Gaussian parameters of the k-th cluster ($=(\mu_k, \sigma_k)$)
λ_g	Poisson parameter of grid g
π_g	Multinomial distribution of clusters in grid in g
Z_g	cluster assignment for bookings in grid g
$N_{g,k}^{(d)}$	number of bookings in cluster k of grid g on day d
S	samples of bookings

Traffic Anomaly Detection. Several pioneering studies have investigated the problem of traffic anomaly detection using GPS trajectories [19]. Liu et al. [7] proposed to infer causal relationships among detected traffic outliers using their spatio-temporal properties. The authors proposed an outlier tree structure and reported empirical evidence using GPS trajectory of taxis. Ge et al. [3] proposed a parameter-free method to detect anomalous trajectories (in particular driving frauds) that deviate in terms of driving distances or driving routes. Zhang et al. [17] proposed the iBAT method to detect anomalous taxi trajectories (e.g., driving frauds or road network changes) in urban cities from taxi trajectories. Pan et al. [9] proposed a traffic anomaly detection and description method based on driver routing behaviours on road networks. The authors define an anomaly as a sub-graph of a road network with significant routing changes. Chua et al. [2] proposed a network transmission model and localisation algorithm to detect locations of anomalies. The transmission model can infer the spatio-temporal transportation data using the temporal information of passenger boarding and alighting bus stops without using the detailed trajectory information. For detecting anomalies, they proposed to rank anomalous events by the degree of their impacts on others.

Our work is different from the above in three aspects: Firstly, our work only relies on time and locations of bookings. Other works mainly focus on analyzing taxi trajectory data. Secondly, taxi bookings are generated from the commuters and are far more dynamic than the movement of the taxis. This introduces more challenges in modeling without knowing the exact commuter mobilities. Lastly, instead of finding anomalous driving behaviors or road conditions, we propose to identify the anomalous booking behaviors which again are driven by commuters. Potentially, these anomalies can reveal interesting events.

3. PROBLEM DEFINITION AND DATASET

In this section, we first introduce the notations and the formal definition of the modeling problem. We then describe a real taxi booking dataset to be used throughout this work.

3.1 Taxi Booking Modeling Problem

We define a *booking* x in some day by a 3-tuple $x = \langle x.lat, x.lng, x.t \rangle$, where $(x.lat, x.lng)$ represent the pick-up location (in latitiude and longitude coordinates respectively) and $x.t$ represents the time of day when the booking is made.

A collection of bookings X over D days is defined as:

Definition 3.1. *(Booking Collection) A* booking collection *X is a collection of booking sets $X^{(d)}$'s, i.e., $X = \langle X^{(1)}, X^{(2)}, ..., X^{(D)} \rangle$*

To model landscape constraints on bookings, we divide the entire city area into equal-sized square grid cells and assign them unique grid indices. Every booking thus falls into a grid cell. We use $X_g^{(d)}$ to denote the set of bookings on d^{th} day located in the grid cell g. If we further divide the bookings in $X_g^{(d)}$ into T time bins, the bookings of time bin t of day d in grid cell g is represented by $X_{g,t}^{(d)}$.

Using the above notations, we now define the problem of modeling taxi booking data as follows.

Problem 3.1. ***(Taxi Booking Modeling)*** *Given a collection of taxi bookings $X = \langle X^{(1)}, X^{(2)}, ..., X^{(D)} \rangle$, the problem of taxi booking modeling is to design a probabilistic model that generates the observed bookings at different locations and time points using a few spatial-temporal clusters of bookings.*

The objective of taxi booking modeling is to summarize the common booking patterns shared by commuters. The underlying assumption of this modeling research is that many commuters share common taxi booking behaviors. For example, many of them may book taxis from similar locations at similar time of the day. To tell if this assumption is reasonable, we test it on a real world taxi booking dataset described in Section 3.2.

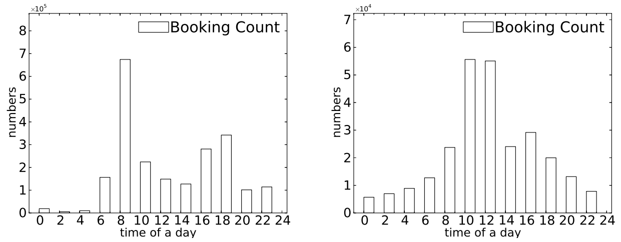
3.2 Taxi Booking Dataset

Taxi booking dataset. We used millions of taxi booking records collected in Singapore over three months from July to September 2014. The data was collected from a taxi booking mobile app[2]. Using this app, a commuter can make a taxi booking using his or her mobile phone providing her pick-up and drop-off locations. The timestamp of booking is recorded. When a booking is served by a taxi driver, the status of the booking is also updated.

Booking data preprocessing. When a commuter does not succeed in the first booking attempt, she could submit another booking again with the same pick-up and drop-off locations. Such duplicate bookings inflate the demand for taxis and were removed as follows. For any two bookings from the same commuter, x_i and x_{i+1}, sorted in chronological order, we consider x_{i+1} a duplicate booking with respect to x_i if: (1) the pick-up locations of x_{i+1} and x_i are less than 500 meters apart, and (2) the booking time stamps of the two bookings are less than five minutes apart. The above de-duplication criteria were empirically determined and have been working quite well in our experiments.

Formation of grid cells. We then divide Singapore into equal-sized $1km \times 1km$ grid cells (see Section 6.3 for the justification of grid size choice). Only grid cells with at least 100 bookings over three months are used in subsequent analysis and experiments. There are altogether 12,361 bookings removed by booking de-duplication and grid-cell pruning. At the end, the resultant dataset still contains millions of bookings after preprocessing, belonging to more than 100K unique commuters and 449 grid cells. The number of three-month bookings in the grid cells varies from 100 to 20,292, suggesting that there are spatial regions with very high booking counts.

[2] We could not reveal the app name and detailed statistics due to non-disclosure agreement.



(a) Weekdays (b) Weekends

Figure 2: Temporal dynamics of bookings and completion rate during weekdays(a) and weekends(b).

From the dataset, we observe both the spatial variation and temporal variation. We describe each as follows.

Spatial Variations. Traffic distribution changes over space. Modeling spatial variation of taxi demand is thus essential. Figure 1 (also mentioned in Section 1) illustrates the difference in the number of bookings among grid cells between the morning and evening rush hours in the three month data. The darker the fill color, the more bookings are observed in the grid cell. Among the grid cells, 1,005 and 162 are the maximum numbers of bookings observed on weekdays and weekends respectively.

Temporal Variations. Figures 2(a) and 2(b) illustrate the variations of booking count over time for weekdays and weekends. During weekdays, there is high demand for taxis at peak hours (e.g., [06:00,9:00) and [18:00,21:00)) as shown in Figure 2(a).

During the weekends, the booking counts are smaller throughout the day. Nevertheless, there are still spikes of booking requests during the [09:00,15:00) and [15:00,18:00) intervals. This suggests commuters are likely to attend social events (e.g., lunch, dinner, etc.) during those times. The served booking rate is low during the [00:00,03:00) time intervals as shown in Figure 2(b).

4. PROPOSED MODEL

In this section, we describe our proposed **G**rid-based **G**aussian **M**ixture **M**odel (**GGMM**), a generative spatial-temporal probabilistic model for dynamic taxi booking events over time and space. We then outline the parameter learning procedure.

4.1 Modeling of Spatial-Temporal Dynamics

The objective of our proposed model is to generate the spatial-temporal dynamics of taxi bookings as given in the observed booking data $X = \langle\, X^{(1)}, X^{(2)}, ..., X^{(D)} \,\rangle$.

As bookings in each $X^{(d)}$ are not uniformly distributed over space and time, we introduce K clusters to group the bookings according to their proximity in space and time. Note that each booking is a tuple $\langle x.lat,\ x.lng,\ x.t\rangle$. All values in the tuple $x.lat$, $x.lng$, and $x.t$ are continuous.

A straightforward approach is to use multivariate **Gaussian Mixture Model** (**GMM**) to cluster bookings in 2-dimensional space over D days into K clusters, each represented by a mean and covariance [5]. This simple model however suffers from a few shortcomings. Firstly, it fails to account for the landscape constraints such as forest and water areas. Secondly, it does not capture the spatial-temporal variation between grid cells. Hence, GMM may not be able to accurately model the booking data as shown in our experiment results (see Section 6).

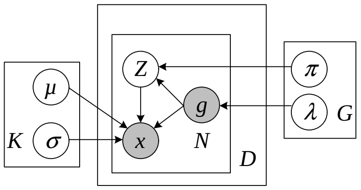


Figure 3: Plate diagram of GGMM.

```
Input:
$K$: initial cluster size;
$G$: number of grids;
$\theta$: $K$ dimensional Gaussian parameters, $(\mu_k, \sigma_k)$ for cluster $k$, $1 \leq k \leq K$ ;
$\lambda$: $G$ Poisson parameters, one for each grid cell $g$;
$\pi$: $G \times K$ multinomial distribution parameters ;
Output: $X$: a set of bookings of a single day;
for  $g = 1$ to $G$ do
    $N_g \leftarrow Poisson(\lambda_g)$;
    for  $n = 1$ to $N_g$ do
        repeat
            $Z_{g,n} \leftarrow Multinomial(\pi_g)$ ;
            $x_{g,n} \leftarrow Gaussian(\theta_{Z_{g,n}})$ ;
        until $(x_{g,n}.lat, x_{g,n}.long)$ falls into grid cell $g$
        and $x_{g,n}.t \in [0:00, 23H:59)$;
        $X \leftarrow X \cup x_{g,n}$;
    end
end
```

Algorithm 1: GGMM Generative Process

Our proposed generative model, GGMM, consists of the following three components: (1) a *Poisson process* that models the daily number of bookings in a grid cell; (2) a *Gaussian Mixture Model* (GMM) that models the spatial and temporal distribution of bookings in a single day; and (3) a grid-cell-specific *mixture weight* π that captures the probability of a booking belonging to each spatial and temporal Gaussian process of the bookings in each grid cell.

Figure 3 depicts the plate diagram of GGMM combining the three components together. The generative story of GGMM is shown in Algorithm 1. For each grid cell g, the number of bookings in g, $|X_g|$, follows a Poisson process with parameter λ_g, $1 \leq g \leq G$.

For each grid cell g, we first sample the number of bookings N_g within a day using a Poisson distribution with parameter λ_g. For each booking x_n, we sample a cluster k from the multinomial distribution over K clusters according to the mixture weight π_g. Given k and Gaussian parameters θ, we then sample the location and time point for x_n. If the location of x_n does not fall into grid cell g or its time point falls outside the time of day, we re-sample x_n until it falls into g and time of the day.

4.1.1 Gaussian Mixture Model Component

GGMM determines the spatial locations and time stamps of daily bookings denoted by X with a Gaussian Mixture Model with K clusters. The clusters are shared among bookings from all of the grid cells. Hence, each observed booking x_n is generated by one of the clusters, say k^{th} cluster with mean μ_k and covariance σ_k^2 as follows:

$$x \;\sim\; \mathcal{N}(\mu_k,\, \sigma_k^2) \tag{1}$$

As the probability of x belonging to each cluster is determined by the marginal distribution of bookings over grid cells and clusters ($\{\pi_{g,k}\}$), we have the likelihood of x_n at a specific location defined by:

$$p(x_n) = \sum_{k=1}^{K} \pi_{g_{x_n},k} \mathcal{N}(x_n|\mu_k, \sigma_k^2) \tag{2}$$

where g_{x_n} is the grid where x_n falls into. π satisfies $0 \leq \pi_{g,k} \leq 1$ and $\sum_k \pi_{g,k} = 1$.

Suppose we have a set of observations $X = \{x_1, ..., x_N\}$. To model them using a mixture of Gaussians, we assume that each observation is drawn independently as shown in Figure 3. The log of the likelihood of X is given by

$$\ln p(X|\pi, \mu, \sigma_i^2) = \sum_{n=1}^{N} \ln \left\{ p(x_n) = \textstyle\sum_{k=1}^{K} \pi_{g_{x_n},k} \mathcal{N}(x_n|\mu_k, \sigma_k^2) \right\} . \tag{3}$$

The posterior probability of x_n belonging to k can be derived using Bayes' theorem as follows:

$$p(k|x_n) = \frac{\pi_{g_{x_n},k} \mathcal{N}(x_n|\mu_k, \sigma_k^2))}{\sum_{i=1}^{K} \pi_{g_{x_n},i} \mathcal{N}(x_n|\mu_i, \sigma_i^2))}. \tag{4}$$

4.1.2 Grid-based Mixture Weight Component

For each observation x_n, the marginal distribution $\pi_{g_{x_n}}$ at grid g where x_n falls into is a multinomial distribution over K Gaussian components. Let N_g be the number of bookings at g. $N_{g,k}$ is the number of bookings at g belonging to cluster k, which can be estimated as follows:

$$N_{g,k} = \sum_{n=1}^{N_g} p(k|x_n). \tag{5}$$

Accordingly, the probability that x_n belongs to cluster k is defined as $\pi_{g_{x_n},k} = \frac{N_{g,k}}{N_g}$.

4.2 Parameter Learning

We learn the GGMM model with parameters Φ from the observed data using the well-known *expectation-maximization algorithm* (EM). EM algorithm iteratively learns the parameters Φ to maximize the likelihood of observed data X. Algorithm 2 summarizes the learning steps. We begin by determining a set of K hard clusters for X, each with a cluster mean and a covariance. These initial clusters are obtained by *k-means* algorithm. We then alternate between E-step and M-step until we find the parameter setting that satisfactorily fits observed data.

In the E-step, EM evaluates the expectation of log-likelihood with current parameter settings. Then, EM re-estimates the parameters of the GGMM model, including means, covariances, and grid-cell-specific mixture weights. We explain this in details as follows.

4.2.1 E-Step

Assume that we are given the existing parameters Φ, including (μ, σ^2, π, λ). In the E-step, EM first evaluates the membership probability of an observed booking ($x_n.lat$, $x_n.lng$, $x_n.time$) belonging to cluster k as follows:

$$\gamma(x_n, k) = \frac{\pi_{g_{x(n)},k} \mathcal{N}(x_n|\mu_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \sigma_j^2)}. \tag{6}$$

where $\sum_k \gamma(x_n, k) = 1$. γ is therefore an $N \times K$ weight matrix.

```
Input:
$X = \{X^{(1)}, ..., X^{(D)}\}$: bookings over $D$ days;
$K$: number of clusters;
$G$: number of grids;
Output:
$\theta$: Gaussian parameters for each cluster $k$;
$\lambda$: Poisson parameter at each grid cell $g$;
$\pi$: multinomial distribution over clusters at each grid $g$;
$\gamma$: posterior probability each booking $x$
$\gamma$, $\theta \leftarrow$ initial clustering by $k$-means$(X)$;
for $g$ = 1 to $G$ do
    $N_g \leftarrow \sum_{k=1}^{K} |\gamma_{g,k}|$;
end
repeat
    for $g$ = 1 to $G$ do
        for $k$ = 1 to $K$ do
            Update $\pi_{g,k}$ by Equation 9;
        end
    end
    for $k$ = 1 to $K$ do
        $\theta_k \leftarrow UpdateGaussionParam(\{\gamma\})$;
    end
    $\gamma \leftarrow ClusterAssignment(\theta, \pi, X)$;
until no changes;
```

Algorithm 2: Grid-based Gaussian Mixture Model

4.2.2 M-Step

We can now use the membership probabilities to obtain the Φ parameter values. Among them are the mean μ_k and the covariance matrix σ_k^2 for the K different Gaussian components. The mean μ_k is obtained by taking a weighted mean of all the bookings in the data set with a weighting factor for each data point x_n by the membership probability $\gamma(x_n, k)$. The mean and covariance are therefore re-estimated using current membership probability as follows:

$$\mu_k^{(new)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(x_n, k) \cdot x_n \tag{7}$$

where $N_k = \sum_{n=1}^{N} \gamma(x_n, k)$.

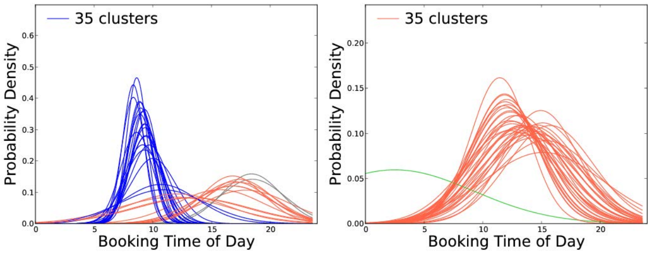
$$\sigma_k^{2(new)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(x_n, k)(x_n - \mu_k^{(new)})(x_n - \mu_k^{(new)})^T \tag{8}$$

The mixture weight for the k^{th} component is given by the average posterior probability, indicating the probability that component k generates data points falling in grid cell g.

$$\pi_{g,k}^{(new)} = \frac{N_{g,k}}{N_g} \tag{9}$$

where $N_{g,k} = \sum_{n=1}^{N_g} \gamma(x_n, k)$.

Note that time difference is derived in circular form as defined in Equation 10. For instance, the difference between 01:00 and 23:00 should be 2 hours apart; whereas the difference between 23:00 and 01:00 is -2 hours apart. Under this definition, the summation of all distances to the temporal mean would be zero.



(a) Weekdays

(b) Weekends

Figure 4: The temporal means.

$$dist(t_1, t_2) = \begin{cases} t_1 - t_2 & \text{if } 00{:}00 \le t_2 \le t_1 \text{ and } t_1 < 12{:}00 \\ t_1 - t_2 & \text{if } t_1 < t_2 < t_1 + 12 \text{ and } t_1 < 12{:}00 \\ t_1 - t_2 + 24 & \text{if } t_1 + 12 \le t_2 \text{ and } t_1 < 12{:}00 \\ t_1 - t_2 - 24 & \text{if } t_2 \le t_1 - 12 \text{ and } t_1 \ge 12{:}00 \\ t_1 - t_2 & \text{if } t_1 - 12 < t_2 \le t_1 \text{ and } t_1 \ge 12{:}00 \\ t_1 - t_2 & \text{if } t_1 < t_2 \text{ and } t_1 \ge 12{:}00 \end{cases} \tag{10}$$

5. EMPIRICAL ANALYSIS USING GGMM

We now apply GGMM to the taxi booking dataset and analyze the latent clusters and mixture weights of the learnt model. We study the model differences between weekdays and weekends.

5.1 Gaussian Clusters

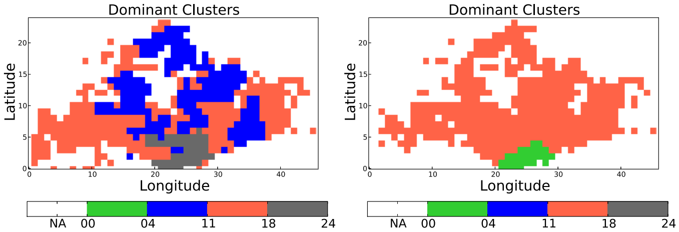
Figures 4 and 5 illustrate the spatial means/standard deviations and temporal means/standard deviations learnt by GGMM for K=35 clusters (the choice of K is elaborated on in Section 6.2). We first study the temporal aspect of the clusters followed by the spatial aspect.

Temporal Means. Figure 4(a) shows the temporal means of weekday taxi bookings. We observe three dominant time periods. The first set of clusters (blue) is centred around morning rush hours (e.g., 09:00) on weekdays. It is followed by the set of clusters (red) centred during mid-days (e.g., noon and 16:00). The remaining clusters (grey) emerge during evening rush hours (e.g., 07:00). The existence of temporal means reflects recurring commuting behaviors on weekdays.

Figure 4(b) reveals the distribution of temporal means from taxi bookings on weekends. The dominant temporal means on weekends are the blue cluster occurring after midnight (e.g., 01:00) followed by the remaining clusters (red) during mid-day (e.g., noon).

Spatial Means. Figures 5 reveals spatial means after applying the GGMM model to booking locations. To observe the correlations between spatial and temporal means, we divide the time of day into four time windows: early morning [00:00,04:00), morning [04:00,11:00), mid-day [11:00,18:00) and evening [18:00,00:00). Then, we illustrate a collective view of spatial means into one plot from the clusters whose temporal means fall into a time window. For example, the set of spatial means on weekdays can be separated into three temporal subgroups, where each temporal subgroup corresponds to one of the time windows.

Figure 5(a) depicts the spatial means on weekday mornings, which are mainly located in the peripheral suburban areas. This coincides with typical morning commuting behaviours, where massive amount of taxi bookings originate from residential areas. The second dominant group of bookings comes from central Singapore during the mid-days ($t \in$[11:00,18:00)) on weekdays as shown in Figure 5(b). Figure 5(c) reveals that the third dominant taxi booking group mainly originates from the central business districts on weekday evenings.



(a) Weekdays

(b) Weekends

Figure 6: Distribution of dominant clusters.

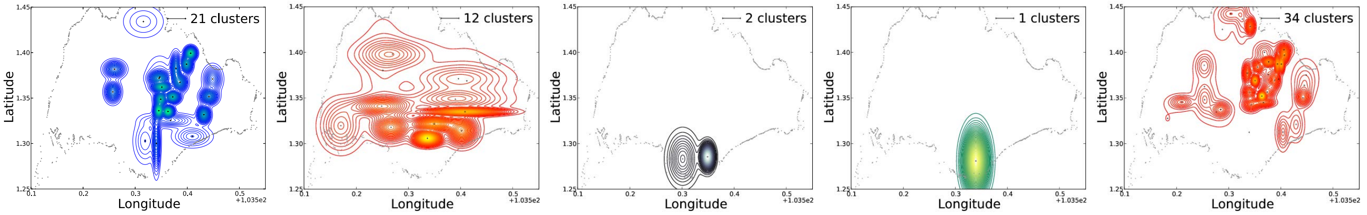
The spatial means on weekends are skewed into two time windows, as shown in Figures 5(d) and 5(e). Figure 5(d) illustrates the spatial means of booking locations mainly in the central business districts after midnight (i.e., 00:21). This suggests that the first dominant group of taxi bookings is contributed by midnight activities from downtown areas. Figure 5(e) shows that the spatial means of remaining clusters are in the northeast area of Singapore during mid-days on weekends. The emergence of mid-day taxi bookings reflects the mid-day activities originating from suburban areas.

5.2 Grid-based Mixture Weights

To observe the correlation between grid-based mixture weights and spatial-temporal clusters, we divide spatial-temporal clusters into four time windows: early morning [00:00,04:00), morning [04:00,11:00), mid-day [11:00,18:00) and evening [18:00,00:00). For each grid cell g, we compute the aggregate of mixture weights for each time window. Specifically, we define the aggregate of mixture weights during a given time window [t_i,t_j) as follows:

$$\pi_{g,k}(t_i, t_j) = \sum_{\mu_{k.time} \in [t_i, t_j)} \pi_{g_x,k}. \tag{11}$$

For each grid cell g, we select the time window [t_i,t_j) that gives the maximum value of aggregate mixture weights among the four time windows as the dominant time period of g. We refer to the set of clusters whose temporal means fall into the dominant time period as *dominant clusters*. We assign a unique color to indicate each dominant cluster. The distribution of dominant clusters across grid cells are illustrated in Figures 6(a) and 6(b) for weekdays and weekends respectively. During weekdays (Figure 6(a)), the blue grid cells show their dominate clusters in the morning [04:00,11:00), whereas the grey areas show their dominate clusters in the evening [18:00,00:00). This coincides with our observations of taxi bookings driven by morning commuters from suburban areas and taxi bookings driven by evening commuters from central business districts. During weekends (Figure 6(b)), the green areas show their dominant clusters in early morning [00:00,04:00) and the remaining areas show their dominant clusters during mid-day [11:00,18:00). This supports the observations that taxi bookings are mainly contributed by midnight events and mid-day events on weekends as shown in Figures 5(d) and 5(e). Note that the white areas are the grid cells with less than 100 bookings over three months. These grid cells typically relate to the water reservoir area in the north and the oceans outside of Singapore territories.



(a) Weekdays $t \in$[04,11) (b) Weekdays $t \in$[11,18) (c) Weekdays $t \in$[18,24) (d) Weekends $t \in$[00,04) (e) Weekends $t \in$[11,18)

Figure 5: Spatial means of booking locations. Figures 5(a) and 5(c) show that taxi bookings are largely contributed by morning commuters from suburban areas and evening commuters from the central business districts. Figure 5(b) coincides with our observations that the main pick-up locations of taxi bookings during mid-days are within the central Singapore. Figure 5(d) shows the taxi bookings contributed by midnight activities from downtown areas on weekends. Figure 5(e) shows the taxi bookings contributed by mid-day activities from suburban areas on weekends.

6. EXPERIMENTAL EVALUATION

We conducted two evaluation tasks on our model using the taxi booking dataset. In the first task, we seek to determine the behavior of both GGMM and baselines for different parameter settings. Secondly, we evaluated the GGMM and baselines in a prediction task.

6.1 Baselines

We evaluate our model by examining its performance in spatio-temporal behavior modeling and prediction. In this study, we consider two strong baseline methods, including: the Gaussian Mixture Model (GMM) and the extended Periodic Mobility Model (PMM).

GMM: a classic three-dimensional Gaussian Mixture Model which continuously and jointly models spatio-temporal variations for behavior modeling and prediction. GMM learns data in unified spatial and temporal dimensions simultaneously. GMM does not model grid cells and the grid-based mixture weights. In our context, it assumes that an observed booking x is generated by a mixture of K Gaussians each with a mean (covering both spatial and temporal dimensions), covariance, and a K-dimensional mixture weight vector. That is, GMM is a special case of GGMM when $G = 1$.

PMM [1]: a state-of-the-art spatio-temporal behavior model which incorporates the temporal influence and spatial interests for behavior modeling and prediction. PMM is a spatial-driven model which generates spatial and temporal dimensions in a pipeline. First, the cluster structure is derived based on spatial distribution. Second, the mixture weight of each data point is derived based on the temporal distribution from spatial clusters (please refer to [1] for details). PMM is designed mainly to estimate the probability distribution over locations of a user at time t (i.e., $p(x_n.lat, x_n.lng|x_n.t)$) without modeling time. To estimate the quality of PMM, we therefore extend PMM to model time by generalizing PMM to model $p(x_n)$. Given a set of unobserved samples $X = \{x_1, x_2, ..., x_N\}$, the log-likelihood is defined accordingly as follows:

$$\ln p(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln \sum_{i=1}^{K} p(x_n.lat, x_n.lng|k)p(k|x_n.t) \cdot \sum_{i=1}^{K} p(x_n.t|k)p(k) \tag{12}$$

where PMM assumes that the location $x_n.lat, x_n.lng$ is generated by K two-dimensional Gaussian $p(x_n.lat, x_n.lng|k) \sim \mathcal{N}(x_n.lat, x_n.lng|\mu_k^s, \sigma_k^{s2})$ and the time $x_n.t$ is generated by K one-dimensional Gaussian $p(x_n.t) \sim \mathcal{N}(x_n.t|\mu_k^t, \sigma_k^{t2})$ respectively.

6.2 Evaluation using Real Dataset

The dataset was divided into training and testing sets. For each grid cell, we randomly grouped the bookings into five folds. Four folds were used as training data, and the rest data were used for testing. Since the taxi demand on weekdays and weekends are fairly different, we further separate the training and test data into weekdays and weekends accordingly.

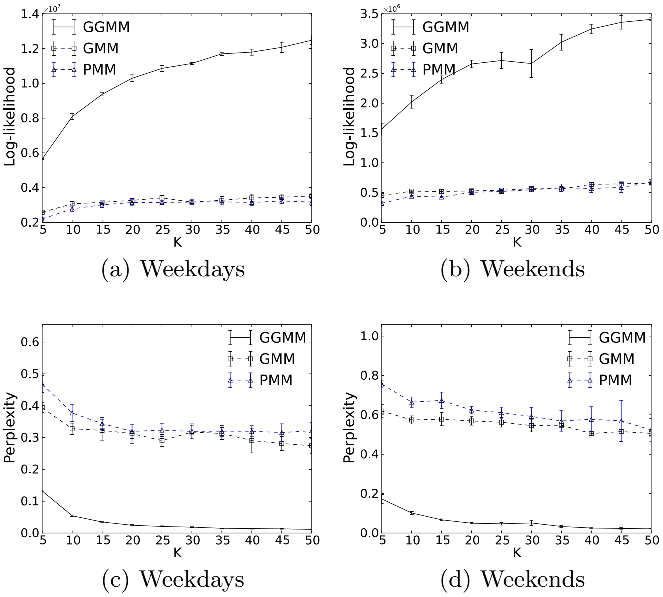
Likelihood Evaluation. To answer the research question: *how well does GGMM model the dynamic booking data?*, we conducted experiments to evaluate three models (i.e., GGMM, GMM, and PMM) from the weekday and weekend training datasets.

We fix the grid size or width (denoted by gs) to be 1km. We varied the number of clusters K from 5 to 50 and plotted the log-likelihoods of GGMM, GMM, and PMM learnt from weekday and weekend training data, as shown in Figures 7(a) and 7(b) respectively. All models show increasing log-likelihoods as K increases. GMM begins to converge when K=10 while GGMM converges when K=35. Since $K = 35$ yields good log-likelihood for GGMM, we used that for the remaining experiments.

GGMM outperforms GMM and PMM for all K for both weekday and weekend data. For weekday bookings, GGMM achieves up to 5.4 (5.1) times improvement in log-likelihood compared with GMM (PMM) with $K = 35$. For weekend bookings, GGMM's log-likelihood is 3.6 (3.7) times that of GMM (PMM). Both GGMM and GMM enjoy higher log-likelihood for weekday booking data. This may be due to either larger training data (i.e., 2,774,358 bookings for weekdays vs 876,556 bookings for weekends) or more regular booking behaviors on weekdays.

Unified spatio-temporal models (i.e., GMM and GGMM) outperform PMM for both weekday and weekend data due to their more expressive ability than PMM. PMM is a spatial-driven model that generates spatial and temporal dimensions of bookings in a pipeline. As a result, the cluster structure of PMM exhibits clear distinctions in spatial dimension as well as in temporal dimension. The clusters derived from unified spatio-temporal models on the other hand may highly overlap at one dimension while differing greatly in the other dimension. As shown in Figure 5(a), 21 clusters of bookings derived from GGMM emerge during [04,11) while they are distinctive from one another in spatial dimension.

Perplexity Evaluation. To evaluate the prediction accuracy of each model, we applied the model to estimate



(a) Weekdays

(b) Weekends

(c) Weekdays

(d) Weekends

Figure 7: Effect of k on log-likelihood(a)(b) and perplexity(c)(d).

the perplexity on testing data using 5-fold cross validation. Perplexity measures how well a set of unobserved samples $X = \{x_1, x_2, ..., x_N\}$ is predicted by the model as follows:

$$Perplexity(X) = exp[-\frac{1}{N}\sum_{n=1}^{N}\log_2 p(x_n)]. \quad (13)$$

A good model will assign high probabilities $p(x_n)$ to the test data thus resulting in low perplexity.

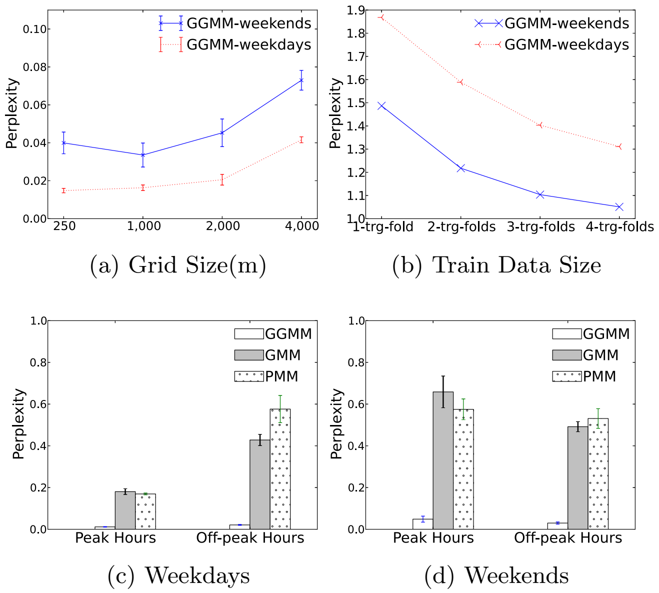
To observe the effects of K, we plotted the perplexity with varying K for weekdays and weekends in Figures 7(c) and 7(d) respectively. All models see their perplexities decrease with larger K for both weekday and weekend data. All models perform better on weekday data than on weekend data, suggesting that they have better predictive power on weekdays. Again, GGMM outperforms both PMM and GMM models across all K's, which suggests that GGMM can model the observed data with mixture of gaussian spatio-temporal distributions in different grid cells. For example, GGMM reduces 95.7% (95.8%) perplexity against GMM in the weekday(weekend) model for K=50.

6.3 Empirical Findings

In the next experiment, we study the effect of parameters on the performance performance.

Effects of Grid Size (gs). We vary the size of grid cell (gs) among 250m, 1km, 2km and 4km. Intuitively, smaller grid size should yield more accurate performance for GGMM as it captures more detailed spatial structures of bookings. In Figure 8(a), the perplexity of GGMM indeed decreases when gs becomes smaller. For instance, with $gs = 250m$, GGMM achieves an improvement of over 1.8(2.8) orders of magnitude compared to the perplexity with coarser-grained grid size (gs =4,000m) in weekdays(weekends) model.

Effects of Training Data. To observe the effect of the size of training data, we used data from i folds of weekday (weekend) data for training and the remaining one fold for testing, where $1 \leq i \leq 4$. For example, when i=3, data from fold-1 to fold-3 are used for training. The perplexities using different training sets are reported in Figure 8(b). As expected, the prediction accuracy increases (as shown in decreasing perplexity) when the training data size increases.



(a) Grid Size(m)

(b) Train Data Size

(c) Weekdays

(d) Weekends

Figure 8: Effect of (a) grid size; (b) training data size; (c)(d) Perplexity of different time periods.

Effects of Time Periods. As the number of bookings and its distribution change over time, we investigat how each model behaves during peak hours and off-peak hours. The time t is considered peak hours if $t \in$[07:00,09:00) or $t \in$[17:00,19:00), otherwise t is considered an off-peak hour. Figures 8(c) and 8(d) report the perplexity at each time period. During weekdays, all models show a better prediction power during peak hours. This is because peak hours are dominant time periods during weekdays, and as a result, the model learnt from peak hours completely depicts the dominant spatio-temporal distributions of bookings on weekdays. On the other hand, because the dominant time periods during weekends are off-peak hours, the prediction accuracy during off-peak hours is better than that during peak hours. GGMM outperforms remaining models regardless of time periods in both weekdays and weekends. We calculate the standard deviations in prediction accuracy across all testing folds (shown as error bars in Figures 8(c) and 8(d)) for each model. We obtain that the average standard deviation of GGMM (0.001 and 0.009 in weekday and weekend data) is significantly smaller than that of PMM (0.034 and 0.048 in weekday and weekend data), demonstrating a more performance of GGMM with less deviation compared to others. In conclusion, GGMM can better predict booking distribution during peak hours with fine-grained grid cells.

7. ANOMALY DETECTION

7.1 Problem Definition and Method

Once we learn a model to describe the variations of taxi demands over space and time, we can determine *when and where anomalous taxi demands occur* using the model. We define the *anomaly detection* task as follows:

Task 7.1. *(Anomaly Detection) Given a taxi booking dataset, we want to find top-ranked groups of anomalous taxi bookings.*

Table 2: Top-5 anomalies in weekdays

Date	Time	Location	Num.	Per.	Z-Score	Event
07-14	[6:00,7:00)	1.25,103.816	62.0	182.9	84.4	WCF
07-14	[6:00,7:00)	1.277,103.852	117.0	98.4	45.2	WCF
07-14	[5:00,6:00)	1.259,103.816	77.0	89.3	41.0	WCF
07-14	[6:00,7:00)	1.259,103.816	111.0	61.6	28.6	WCF
09-22	[6:00,7:00)	1.286,103.852	67.0	58.0	26.7	F1

As defined, the anomalies to be detected are not individual bookings that are anomalous. Such isolated anomalies are difficult to validate and have little utility values in the transportation system. Instead, we aims to find anomalies that involve groups of anomalous bookings that happen at same location regions during same time intervals. Both taxi operators and transportation regulation units find such anomalies interesting as they may suggest revenue opportunities or events that affect transportation service (e.g., train breakdown). Even so, it is non-trivial to verify the anomalies as there are no ground truths. In this part of the work, we shall focus on GGMM model only.

There are several possible ways to detect anomalous groups of bookings using GGMM model. To verify the efficacy of GGMM in anomaly detection, we employ a simple approach as follow. We bin all taxi bookings by both space and time. For simplicity, each bin consists of bookings in a grid cell over one hour of the day. Hence, we have $G \times 24$ bins per day. A bin with many anomalous bookings as determined by GGMM is thus detected and ranked. We define the anomaly score of a bin using perplexity as defined in Equation 13. The higher the perplexity is, the more anomalous is the bin of bookings. We also flag the bins with perplexity at least three standard deviations away from mean as anomalous. In other words, we compute the z-score of the bins and return those with z-score ≥ 3.

7.2 Qualitative Evaluation

We qualitatively evaluate the application of GGMM for anomaly detection by verifying the anomalies against online content as follows. We trained a GGMM model using weekday booking data and another using weekend data with number of clusters $K = 35$. We focus only on bins with at least 50 bookings as small events are less interesting. As a results, we have 13,565 bins for the weekdays and 3067 bins for the weekends to be considered for anomaly detection. For each detected anomalous bin, we perform a search for events that happened at the stated location area during the stated time interval mentioned by online news and announcements. This approach of verification is obviously non-ideal but is the only option that can be employed. Despite this, we can find reasonably clear events that are possible associated with the anomalies as shown below.

7.2.1 Weekday Anomalies

Table 2 reports the top five anomalous bins using GGMM (K=35, trained using weekday data) according to their z-scores. The mean and standard deviation of perplexities are 0.09 and 2.17 respectively. We also found these anomalous bins likely to be associated with the following events.

World Cup Final screening (WCF). The top-1, top-3 and top-4 anomalies are very likely related to the night screening of World Cup Final screening at Sentosa, a resort island[3] which occurred from 3AM to 5AM on July 14. After screening, many users might have make taxi bookings the early morning to return home. Figures 9(a) and 9(b) illustrate the spatial and temporal deviations of the WCF event (pink box) from GGMM weekday model. Figure 9(a) shows the spatial means of two clusters of GGMM (29^{th} and 32^{th} clusters) that are temporally closest to the WCF location. It is quite clear that the WCF location is quite far from these two spatial means. Similarly, Figure 9(b) shows the probability density functions for 29^{th} and 32^{th} clusters in blue lines, and the histogram of bookings on 14 July 2014 and the anomalous bookings in grey during [06:00, 07:00) to highlight the temporal deviation of WCF bookings.

Table 3: Top-5 anomalies on weekends

Date	Time	Location	Num.	Per.	Z-Score	Event
08-16	[23:00,0:00)	1.376,103.996	96.0	1.19×10^8	55.37	NDC
08-17	[17:00,18:00)	1.322,103.654	55.0	873.6	-0.02	RE
07-12	[10:00,11:00)	1.412,103.69	52.0	197.5	-0.02	ME
07-12	[11:00,12:00)	1.412,103.69	64.0	141.5	-0.02	ME
08-16	[16:00,17:00)	1.412,103.69	141.0	102.7	-0.02	ME

7.2.2 Weekend Anomalies

Table 3 reports the top-5 anomalous bins using GGMM (K=35, trained with weekend data). The mean and standard deviation of perplexities are 3.9×10^4 and 2.2×10^6 respectively. We also found these anomalous bins likely to be associated with the following events.

National Day Celebration (NDC). The top-1 anomaly is located at Singapore Changi Airport during [23:00, 00:00) right after a National Day Celebration event held in Crown Plaza Hotel at the airport around [18:30, 22:30)[4]. The anomalous taxi bookings were likely due to the crowd leaving that event. Figures 9(c) and 9(d) depict the spatial and temporal deviations, respectively, of the NDC-related bin (pink box) from the GGMM weekend model. Spatially and temporally, the bin of bookings quite far from the nearest clusters of GGMM making it appear to be very anomalous.

Racing event(RE). The top-2 anomaly occurred around 5pm on Aug 17 at a factory area in west Singapore where the population is sparse. Interestingly, we observed a racing event held on the same day[5] by Raffles Marina Club House, which is the only point of interest in the factory area. The anomalous taxi demand may be caused by this racing event.

Military event(ME). The top-3 to top-5 anomalies occurred at Sungei Gedong Camp, one of Singapore's army bases. The camp is located in northwest Singapore in a forest area where civilians are forbidden to visit. The anomalous taxi demands may be related to personnel booking taxis to leave the army base after finishing military training classes on Saturday morning (July 12) and Saturday afternoon (August 16). Unfortunately, we could not correlate this with any events online on July 12's morning or August 16's afternoon.

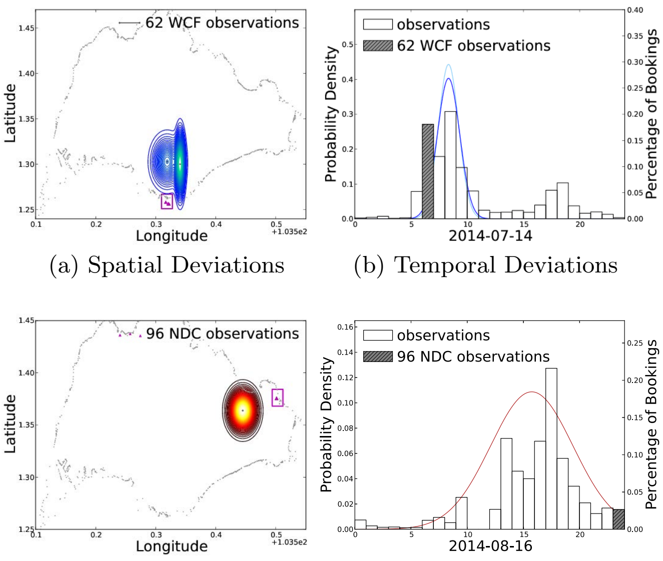
8. CONCLUSIONS

In this paper, we have proposed and evaluated a novel model, GGMM, for predicting the dynamics of taxi demands using Poisson processes and spatio-temporal Gaussian processes at the grid cell level. GGMM is especially good at coping with multiple spatio-temporal clusters sharing similar time periods or spatial regions. This is not the case for the state-of-the-art Periodic Mobility Model (PMM). The empirical studies on GGMM reveal the spatio-temporal movement

[3] https://store.sentosa.com.sg/main/events/universal-studios-singapore-ni/54#!/

[4] http://www.whatshappening.sg/events/index.php?eID=79165

[5] http://www.westerncircuit.com/schedule.php



(a) Spatial Deviations (b) Temporal Deviations

(c) Spatial Deviations (d) Temporal Deviations

Figure 9: (a) Spatial and (b) Temporal deviations of WCF discovered by GGMM during [06:00, 07:00) on July 14. (c) Spatial and (d) Temporal deviations of NDC discovered by GGMM during [23:00, 00:00) on August 16.

patterns of daily commuters on both weekdays and weekends separately. Our experiments on a taxi booking dataset show that the modeling quality of GGMM measured by log-likelihood is 5.4 times better than the standard Gaussian Mixture Model (GMM) and 5.2 times better than PMM. GGMM also have been shown to yield around 95.8% (96.5%) reduction in perplexity compared against GMM (PMM). Lastly, we demonstrated the effectiveness of our proposed model through anomaly detection and successfully verified the anomalies with real-world events.

Two directions for future work are of particular interest. The first is to update taxi demands in real time. Another direction is to explore taxi supplies from trajectory data. We would like to extend our model to analyze taxi trajectories and derive the spatio-temporal dynamics of taxi supply. Combining the two directions, we will yield a more complete set of insights on urban mobility patterns.

Acknowledgment

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).

9. REFERENCES

[1] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *SIGKDD*, 2011.

[2] F. C. Chua, E.-P. Lim, and B. A. Huberman. Detecting flow anomalies in distributed systems. In *ICDM*, 2014.

[3] Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou. A taxi driving fraud detection system. In *ICDM*, 2011.

[4] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *VLDB Journal*, 20(5), 2011.

[5] M. Lichman and P. Smyth. Modeling human location data with mixtures of kernel densities. In *SIGKDD*, 2014.

[6] S. Liu, S. Wang, C. Liu, and R. Krishnan. Understanding taxi drivers' routing choices from spatial and social traces. *Frontiers of Computer Science*, 9(2), 2015.

[7] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *SIGKDD*, 2011.

[8] Y. Liu, C. Liu, N. J. Yuan, L. Duan, Y. Fu, and H. Xiong. Exploiting heterogeneous human mobility patterns for intelligent bus routing. In *ICDM*, 2014.

[9] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *SIGSPATIAL*, 2013.

[10] M. Qu, H.-S. Zhu, J. Liu, G. Liu, and H. Xiong. A cost-effective recommender system for taxi drivers. In *SIGKDD*, 2014.

[11] Y. Wang, Y. Zheng, and Y. Xue. Travel time estimation of a path using sparse trajectories. In *SIGKDD*, 2014.

[12] B. Yang, C. Guo, and C. S. Jensen. Travel cost inference from sparse, spatio temporally correlated time series using markov models. *Proceedings of the VLDB Endowment*, 6(9), 2013.

[13] N. Yang, X. Kong, F. Wang, and P. S. Yu. When and where: Predicting human movements based on social spatial-temporal events. In *SDM*, 2014.

[14] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and POIs. In *SIGKDD*, 2012.

[15] J. Yuan, Y. Zheng, C. Zhang, W.-L. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: Driving directions based on taxi trajectories. In *SIGSPATIAL*, 2010.

[16] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie. T-finder: A recommender system for finding passengers and vacant taxis. *TKDE*, 25(10), 2013.

[17] D. Zhang, N. Li, Z.-H. Zhou, C. Chen, L. Sun, and S. Li. ibat: Detecting anomalous taxi trajectories from GPS traces. In *UbiComp*, 2011.

[18] F. Zhang, N. J. Yuan, D. Wilkie, Y. Zheng, and X. Xie. Sensing the pulse of urban refueling behavior: A perspective from taxi mobility. *ACM TIST*, 6(3), 2015.

[19] Y. Zheng. Trajectory data mining: An overview. *ACM TIST*, 6(3), 2015.

[20] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: Concepts, methodologies, and applications. *ACM TIST*, 5(3), 2014.

[21] C. Zhong, S. M. Arisona, X. Huang, and G. Schmitt. Identifying spatial structure of urban functional centers using travel survey data: a case study of Singapore. In *SIGSPATIAL International Workshop on Computational Models of Place*, 2013.