Modelling Probability Distribution of Hotel Check-in and Check-out Activities

Dicong Qiu*
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
dq@cs.cmu.edu

Abstract

It has been an intriguing problem to analyze the pattern of check-in and check-out activities of hotels, which can provide insights for hotel management, guest flow triage, and so on. In this work, we use different approaches to model the probability distribution of hotel check-in and check-out activities, based on a real-world dataset that contains multiple hotel transactions. We also apply the probability distribution models to a hotel management system (HMS) coverage problem.

1 Introduction

Customer flow is a well known problem of studying the number and pattern of customers coming in and through. Different from shops or restaurants, hotels usually have customers with prolonged stay interval, which leads to the equivalent importance of studying both the check-in and check-out activities of hotels. To better understand the pattern of hotel check-in and check-out activities, we collected a real-world dataset (referred to as "the dataset" in the following discussions) with 24984 transactions from multiple hotels, from which the aforementioned activities can be extracted. With this dataset, we first analyze the check-in and check-out event frequency distribution in an empirical manner. And then we try to model the probability distribution of these events using models such as normal (Gaussian) distribution and a mixture of Gaussian distributions. The models are then applied to a hotel management system coverage rate study by using the probabilistic models to compute the normalized likelihood that a hotel is using the HMS in real-time to check in and check out guests.

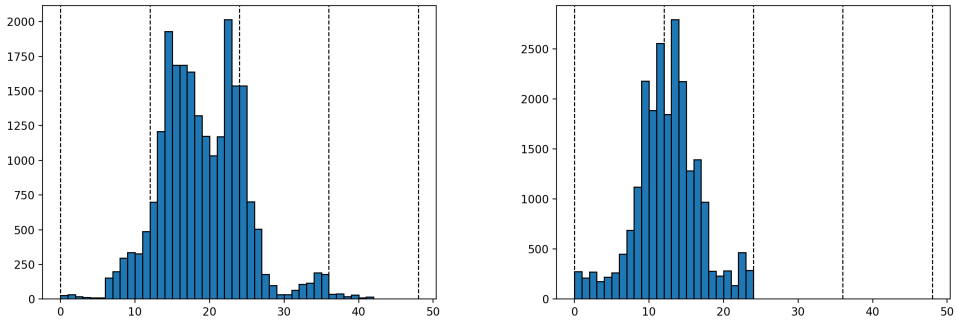
2 Empirical Analysis

The below analysis is based on three empirical assumptions from daily experience and an assumption about the dataset we used that

- **A.1** hotels usually require their guests to check out before 12:00 on their check-out dates, and guests usually check out at around 10:00 on their check-out dates;
- **A.2** hotels usually make rooms available to their guests by 14:00 to 15:00 on their check-in dates;
- **A.3** the dataset was created using transactions from trusted hotels, which means the transactions shall have minimal latency regarding their use of the HMS.

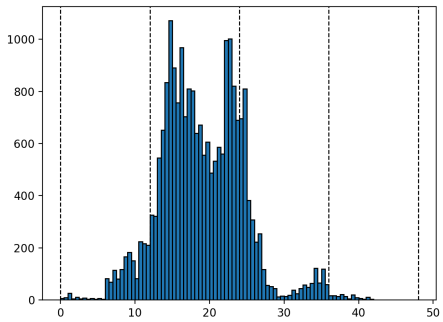
In the following analysis, these assumptions are inferred to a certain extent from the frequency distributions of valid check-in and check-out events in the dataset.

*http://cs.cmu.edu/~dq

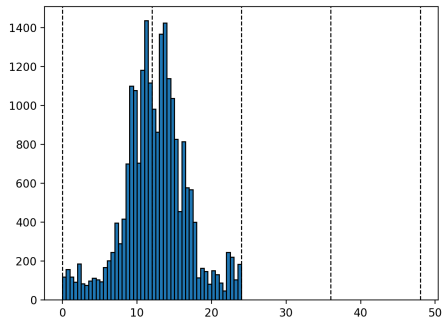


(a) Check-in event dist. (res: 1-hour)

(b) Check-out event dist. (res: 1-hour)



(c) Check-in event dist. (res: 0.5-hour)



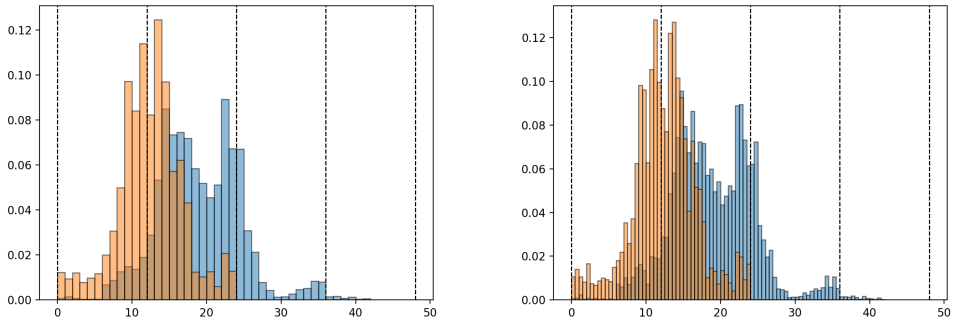
(d) Check-out event dist. (res: 0.5-hour)

Figure 1: Frequency distribution of 22,425 respective valid check-in and check-out events in 1-hour and 0.5-hour resolutions within a 48-hour interval from the beginning of the expected check-in/check-out date, where the five vertical dashed lines indicate the elapse of 0, 12, 24, 36 and 48 hours, respectively.

As indicated by figures 1a and 1c, check-in events primarily begin after 7:00 and grow to the first peak at around 3:00, which follows the assumption **A.2**. The number of check-in events started to drop and reach a local minimum at around 20:00, after which it rises to another peak at around 23:00. The interpretation for the second peak is twofold: on the one hand, it is possible that the hotel receptionists were too busy to update all the check-in transactions to the HMS; on the other hand, considering assumption **A.3**, it is highly possible that there is actually a second peak of check-in events happening near the midnight after the dinner time. The latter interpretation will be mostly considered. It is also very interesting to observe a third check-in peak the next day during noon, which suggests there are some people who arrived at their hotels a day after the expected check-in date. In other to model the check-in event probability distribution, one can consider either a double modal Gaussian distribution by ignoring the third peak, or a triple modal Gaussian distribution with the third peak. Another alternative approach is to constrain the check-in event distribution modelling within a 24-hour range, with an additional Bernoulli distribution to capture the $+1$ day delayed check-in.

In figures 1b and 1d, there are three prominent peaks of check-out events, happening around 9:00, 11:00 and 2:00. The first and second peak make sense in accord to assumption **A.1**. The temporary drop at 10:00 can be interpreted as a watershed of two groups of people: the people who were more conservative so that they checked out relatively earlier or who stayed in hotels that required earlier check-out, and those who liked to sleep until noon or who stayed in hotels that had more flexible check-out schedule. The third peak does not actually conform to the assumptions. It may be interpreted as two possible situations: (1) some hotels actually allowed their customers to stay until 3:00, which does conflicts with assumption **A.1**; (2) receptionists of some hotels were busy handling the incoming check-in transactions, which prolonged the check-out procedures, but such a situation will conflict with assumption **A.3**. So further understanding of the third peak and its later on check-out events will be beneficial. As for modelling, a more elaborated approach would be to fit

three Gaussian distributions to the three peaks, but to simplify the model, it is also feasible to assume the entire distribution as a uni-modal Gaussian distribution with its mean at around 12:00.



(a) Cross-event dist. (res: 1-hour) **(b)** Cross-event dist. (res: 0.5-hour)

Figure 2: Cross-event frequency distribution of valid check-in/check-out events in 1-hour and 0.5-hour resolutions within a 48-hour interval from the beginning of the expected check-in/check-out date, where the five vertical dashed lines indicate the elapse of 0, 12, 24, 36 and 48 hours, respectively.

To conduct a cross analysis of check-in and check-out events, the normalized frequency distributions of two events are overlapped. It can be observed that most of the check-out events antedated the check-in events, which makes sense considering the assumptions. And there is a simultaneous drop for both events around 20:00, which is usually the dinner time. It is a new discovery by superposing one distribution on top of the other.

3 Modelling

Formally, the transaction data are split into two datasets, the check-in events dataset $\mathcal{D}_{\text{in}} = \{x_i\}_{i=1}^{N_{\text{in}}}$ and the check-out events dataset $\mathcal{D}_{\text{out}} = \{y_j\}_{j=1}^{N_{\text{out}}}$, such that $8 \leq x \leq 30, \forall x \in \mathcal{D}_{\text{in}}$ and $0 \leq y \leq 24, \forall y \in \mathcal{D}_{\text{out}}$. Here, x_i's are samples of check-in events, y_j's are samples of check-out events, $N_{\text{in}} = 21130$ is the number of all check-in events considered and $N_{\text{out}} = 20885$ is the number of all check-out events considered.

3.1 Modelling Check-in Distribution

We consider the check-in event distribution F as a bi-modal Gaussian distribution (mixture of two Gaussian distributions).

$$F\left(x \middle| \{\mu_m\}_{m=1}^{M_F}, \{\sigma_m^2\}_{m=1}^{M_F}, M_F = 2\right) = \frac{1}{M_F} \sum_{m=1}^{M_F} \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{(x-\mu_m)^2}{2\sigma_m^2}\right)$$

The Expectation-Maximization (EM) algorithm [1] is used here to find the set of optimal parameters for distribution F, by starting from $\left[\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right]^\mathsf{T} = [15, 22, 1, 1]^\mathsf{T}$ and re-assigning samples from $\mathcal{D}_{\text{in}}$ to the Gaussian with higher probability (E-step) and maximizing the likelihood of each distribution (M-step). After multiple iterations with termination condition $(|\Delta\mu_1| + |\Delta\mu_2| + |\Delta\sigma_1| + |\Delta\sigma_2|) \leq \epsilon = 1.0 \times 10^{-6}$, a set of parameters were computed.

$$\begin{aligned}
\mu_1 &= 15.223748 \\
\mu_2 &= 23.006930 \\
\sigma_1 &= 6.544225 \\
\sigma_2 &= 4.265957
\end{aligned}$$

The model fitting result is visualized in figure 3 below.

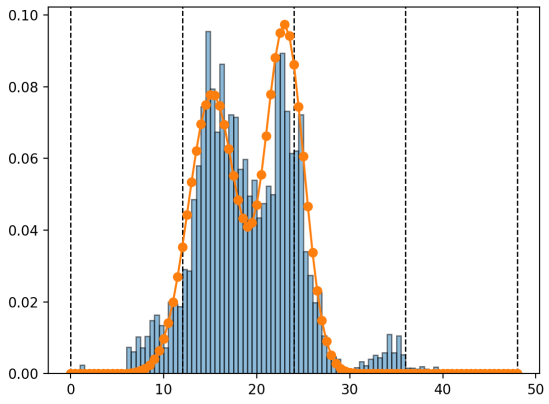


Figure 3: Model fitting for check-in event distribution.

3.2 Modelling Check-out Event Distribution

We consider the check-out event distribution G as a uni-modal Gaussian distribution.

$$G\left(y|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Its mean and variance can be directly estimated from the samples in $\mathcal{D}_{\text{out}}$.

$$\mu = \frac{1}{N_{\text{out}}}\sum_{j=1}^{N_{\text{out}}} y_j = 12.436871$$

$$\sigma^2 = \frac{1}{N_{\text{out}}}\sum_{j=1}^{N_{\text{out}}} (y_j-\mu)^2 = 18.619825$$

And the result is visualized in figure 4.

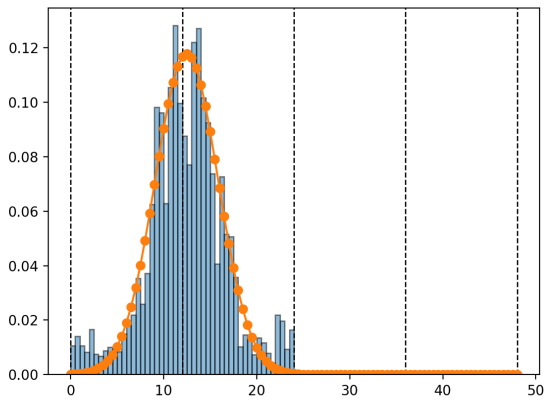


Figure 4: Model fitting for check-out event distribution.

4 HMS Coverage Rate Study

The problem of HMS coverage rate study rises because different hotels may have their legacy systems for management. As a new HMS goes into the market, it will be beneficial to understand its coverage rate and market occupation speed for optimizing business strategy. A metric for evaluating the coverage rate of an HMS is to check if the hotels are using this system in real-time, by which it can be inferred that the hotels actually use the HMS for their daily management without using any other system in parallel. But figuring out the real-time utilization rate of an HMS is nontrivial, because neither is it feasible to send agents to monitor a large number of hotels all the time, nor is it possible to reveal whether the hotels are utilizing the HMS in a real-time manner without ground-truth observations.

To overcome the above challenges, we proposed to use the fitted models from section 3 to compute baseline check-in and check-out scores, which will be used to evaluate to what extent a hotel is utilizing the HMS in a real-time manner. We define the baseline score B as the normalized likelihood that the N transaction samples $\mathcal{D} = \{x_i\}_{i=1}^{N}$ from the trusted hotels are sampled from a probability distribution model M. M will be used here to denote a probability distribution model in order to make the derivation more general, which in our case can be either F or G.

$$B\left(M_\theta, \mathcal{D}\right) = \sqrt[N]{\prod_{i=1}^{N} M_\theta\left(x_i\right)}$$

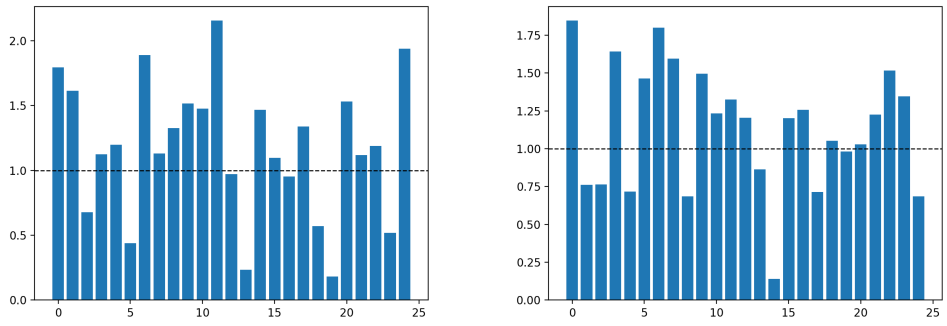
where θ represents the fitted parameters of the probability distribution model M. The logarithm form of the above equation will more practically feasible for calculation.

$$\log B\left(M_\theta, \mathcal{D}\right) = \frac{1}{N}\sum_{i=1}^{N} \log M_\theta\left(x_i\right)$$

A score that measures the extent to which a hotel is using the HMS in real-time will be a relative scale compared to the baseline score. Consider the N_k transaction samples $\mathcal{D}_k = \left\{y_i^{(k)}\right\}_{i=1}^{N_k}$ are collected from a target hotel k. The real-time utilization evaluation score R will be calculated based on a given probability distribution model M_θ, the corresponding baseline metric B and a set of trusted hotel transactions $\mathcal{D}$.

$$R\left(\mathcal{D}_k | M_\theta, B, \mathcal{D}\right) = \exp\left(\log B\left(M_\theta, \mathcal{D}_k\right) - \log B\left(M_\theta, \mathcal{D}\right)\right)$$

Experimental result using the above metric on our dataset is shown below in figure 5.



(a) Check-in scores.

(b) Check-out scores.

Figure 5: Normalized scores of hotel check-in/check-out event distributions for different hotels. The names of the hotels have been removed and each hotel is associated to an index number. Hotels with a normalized score over 1.0 utilize the HMS in a more real-time manner compared to the average situation.

5 Conclusion

We first empirically analyze the hotel check-in and check-out activity pattern based on a real-world dataset we collected, and the analysis is to a certain extent consistent with three assumptions from daily experience. In section 3, we use both Gaussian mixture model and uni-modal Gaussian distribution model to capture the overall check-in and check-out event frequency distribution pattern. And finally, the models are then applied to address the HMS coverage rate evaluation problem.

The source code of this work is accessible online[2].

References

[1] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[2] `https://github.com/davidqiu1993/hotel_in_out_analysis`