Forecasting Arrivals and Occupancy Levels in an Emergency Department

Ward Whitt [1], Xiaopei Zhang [2]

Abstract

This is a sequel to Whitt and Zhang (2017), in which we developed an aggregate stochastic model of an emergency department (ED) based on the publicly available data from the large 1000-bed Rambam Hospital in Haifa, Israel, from 2004-7, associated with the patient flow analysis by Armony et al. [1]. Here we focus on forecasting future daily arrival totals and predicting hourly occupancy levels, given recent history (previous arrival and departure times of all patients). For the arrival forecasting, we divide the data set into an initial training set for fitting the models and a final test set to evaluate the performance. By using 200 weeks of data instead of the previous 25, we identify (i) long-term trends in both the arrival process and the length-of-stay distributions and (ii) dependence among successive daily arrival totals, which were undetectable before. From several forecasting methods, including artificial neural network models, we find that a seasonal autoregressive integrated moving average with exogenous (holiday and temperature) regressors (SARIMAX) time-series model is most effective. We then combine our previous ED model with the arrival prediction to create a real-time predictor for the future ED occupancy levels.

Keywords: emergency departments, forecasting arrivals, predicting occupancy levels, time series analysis, neural networks.

[1] Industrial Engineering and Operations Research, Columbia University, Email: ww2040@columbia.edu

[2] Industrial Engineering and Operations Research, Columbia University, Email: xz2363@columbia.edu, Correspondence to: MailCode 4704, S. W. Mudd Building, 500 West 120th Street, New York NY 10027-6699, U.S.A.

1. Introduction

There is great interest in patient flow in emergency departments (EDs) because EDs are often plagued by congestion. Even though there have been many studies, e.g., [2, 3, 4, 5], there remain opportunities to develop new analysis methods. In this paper we contribute by studying methods to forecast future daily arrival totals and predict hourly occupancy levels, given recent history, i.e., given all previous arrival and departure times.

This paper is an extension of our recent [6], in which we developed an aggregate stochastic model to describe patient flow in an emergency department (ED) based on 25 weeks of the publicly available patient flow data from the large 1000-bed Rambam Hospital in Haifa, Israel, from 2004-7, associated with the patient flow analysis by Armony et al. [1]. The stochastic model in [6] is periodic, with a week serving as the cycle length. There is a two-time-scale model of the arrival process, in which the daily arrival totals are modeled as a Gaussian process, while the arrivals during each day are modeled as a nonhomogeneous Poisson process, given the estimated arrival rates. In other words, the arrival process is modeled as a periodic doubly stochastic nonhomogeneous Poisson process (also known as Cox process). The length of stay (LoS) variables of the successive arrivals were assumed to come from a sequence of independent random variables, where the periodic LoS distribution depends on the day of the week and hour of the day.

In this paper, we study how our proposed stochastic model can be used for prediction, which was identified as an important topic for future research in §7 of [6]. Hence we want to have a test set that is independent of the set we use to select and fit the model. For this purpose, we use a larger dataset of 200 weeks and split the data into a training set and a test set. We assess several prediction models on the same test set. By using 200 weeks of data instead of the 25 in [6], we identify (i) long-term trends in both the arrival process and the length-of-stay distributions and (ii) dependence among successive daily arrival totals, which were undetectable before. Hence, the model here becomes an extension of the previous model.

Here we forecast daily arrival totals and predict hourly occupancy levels. There is a substantial literature on forecasting, both for ED's, e.g., [7, 8, 9, 10, 11, 12], and for other service systems more generally, e.g, as in call centers [13, 14]. In this paper we examine several alternative models to forecast the daily arrival totals, including a linear regression based on calendar and weather variables, seasonal autoregressive integrated moving average with exogenous regressors (SARIMAX) model and the multilayer perceptron (MLP) model, which is an artificial neural network machine learning method. All the models can be viewed as generalizations of the two-time-scale Gaussian-NHPP model proposed in [6].

One goal of this new work is to investigate how machine learning techniques can be applied to such problems. There is great interest in advanced neural network models, because they have been successful in solving many challenging tasks in different areas. However, our exploration shows that the neural network models, with only patient arrival and departure data, do not perform as well as the highly structured time-series SARIMAX models, which include relevant extra context information, such as holidays and temperature data. With a relatively small dataset compared to the "big data" applications, as in social media applications, a good structured model evidently outperforms the extremely flexible machine learning model, which exploits the large dataset to learn the features of the system.

We also propose a real-time occupancy predictor which exploits the currently observed occupancy level and the empirical hazard functions, given the elapsed LoS for each patient in the system. That is a variant of the approach suggested in §6 of [15]; see [16, 17] and references there for related work. We conclude that exploiting the real-time information can be helpful in predicting the near future status of the system. This also illustrates how our forecasting model for the arrival process can be useful for other operational purposes.

We stress that the data we use is open to the public at the SEELab of the Technion, so that interested researchers can replicate and improve our results. The data records the arrival times and LoS of each patient that visited the ED. We refer to [1, 6] for more information about this dataset. We preprocess the data as we did in [6], only considering the patients that visited the emergency internal medicine unit (EIMU), which is the majority of all the new visits to the ED.

The paper is organized as follows: In §2 we review the model framework in [6]. In §3 we look at the larger dataset and relate it to our original model. In §4 we evaluate five potential improvement methods for predicting the daily arrival totals. Then in §5 we introduce our real-time predictor for hourly occupancy level. Finally, in §6 we draw conclusions. Additional supporting material appears in a longer online.

2. The Integrated Model for the ED Patient Flow

The model we proposed for the ED patient flow in [6] is depicted abstractly in Figure 1. The model has three parts: the process generating daily arrival totals (M_1), the arrival process within each day (M_2) and the length of stay (LoS) of each patient (M_3).

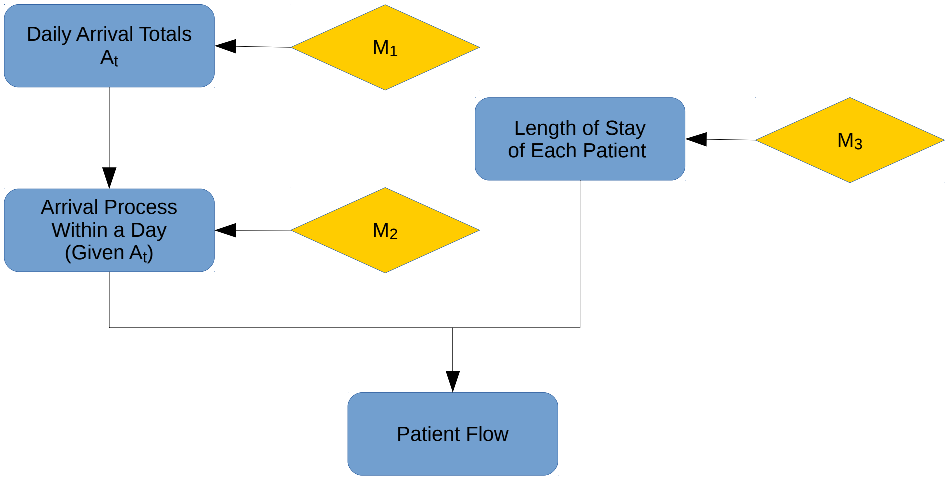


Figure 1: Illustration of the integrated ED patient flow model.

In [6], after statistical analysis, we let M_1 be a single-factor Gaussian model, which only depends on the day-of-week; i.e., the daily arrival total on day t is modeled by

$$A_t = c_0 + \sum_{i=0}^{6} d_i D_{t,i} + \epsilon_t, \tag{1}$$

where i from 0 to 6 represent Sunday to Saturday, c_0 and d_i are constants to be estimated, $D_{t,i}$ is the indicator of day-of-week (i.e. $D_{t,i} = 1$ if day t is day-of-week i, and 0 otherwise) and $\epsilon_t \sim N(0, \sigma^2)$ is the i.i.d. random term where σ is a constant. (We tested for a trend and for dependence, but based on the limited data, neither was statistically significant.) By definition, A_t could be non-integer, but we always understand A_t is an integer by rounding it to the nearest one. Also, theoretically, A_t could be negative, in which case we round up to 0. Given the estimated mean and variance, that is highly unlikely.

For M_2, we assumed that the arrival process within each day is a nonhomogeneous Poisson process (NHPP) given the daily arrival total of that day. That means, for a given total number of arrivals, the arrival epochs are i.i.d. with a probability density function (pdf) that is proportional to the arrival rate function. We also assume that the arrival rate function is piecewise constant, changing hourly. To describe it explicitly, let $\lambda_{i,j}$, $i = 0, 1, \cdots, 6$, $j = 0, 1, 2, \cdots, 23$ be the (constant) mean arrival rate for hour j on day-of-week i. For $i = 0, 1, \cdots, 6$, the pdf is

$$f_i^a(s) = \begin{cases} \dfrac{\lambda_{i,j}}{\sum_{j=0}^{23} \lambda_{i,j}}, & s \in [j, j+1), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

while $F_i^a(s) = \int_{-\infty}^{s} f_i^a(x)dx$ is the corresponding cumulative distribution function (cdf). Given A_t, for a specified day of the week i, let $a_{t,k}$, $k = 1, 2, \cdots, A_t$ be the arrival times of the patients in day t, then we assume that $a_{i,k}$ are i.i.d. with pdf f_i^a.

For M_3, we assume that the patient LoS's are mutually independent, having a distribution that only depends on the arrival time; i.e., if we let $w_{t,k}$ be the corresponding LoS of those patients that arrived at $a_{t,k}$, then $w_{t,k}$ are independent of each other and the arrival process, and $w_{t,k} \overset{\mathrm{d}}{=} f_{i,j}^s$ if day t is day-of-week i and $a_{t,k} \in [j, j+1)$, where $f_{i,j}^s$ is a given pdf of LoS for day-of-week i and hour j. As in [6]. we assume that the LoS distributions., just like the arrival processes, are time-varying and periodic, with a period of one week.

It is significant that our arrival process model captures over-dispersion, a key property observed in the arrival data of the ED; see [18] and references there. By combining M_1 and M_2, we see that the arrival process is a doubly stochastic nonhomogeneous Poisson process or Cox process. We can equivalently regard it as an NHPP where the houly arrival rates within a day are correlated Gaussian random variables. If we denote $N_i(s)$, $s \in [0, 24]$ to be the counting process of arrivals on day-of-week i, then $N_i(24) \sim N(\mu_i, \sigma^2)$, where $\mu_i = c_0 + d_i$. If we introduce the index of dispersion for counts (IDC), defined by

$$I_i(s) = \frac{\mathrm{Var}(N_i(s))}{\mathbb{E}(N_i(s))}, \tag{3}$$

as we did in [6], then according to M_2, we can easily deduce (as a special case of general Cox process) that

$$I_i(s) = 1 + F_i^a(s)(\frac{\sigma^2}{\mu_i} - 1). \tag{4}$$

Formula (4) quantifies the overdispersion. We have over-dispersion whenever, $I_i(24) = \sigma^2/\mu_i > 1$. Figure 5 of [6] shows an estimate of the IDC function, demonstrating the over-dispersion, but it is not exceptionally high. Our Gaussian time series models also approximate the generalized linear models that allow over-dispersion, such as negative binomial regression, because a negative binomial distribution can be approximated by the normal distribution when the parameter is large. Such generalized linear regression methods have been used when studying the ED patient flow [19, 20].

As observed in [6], our model can be viewed as an infinite-server queueing model with the arrival process being a Cox process and independent service times. For infinite-server queues, the arrival process can be independently thinned into two or more processes, allowing more factors to be taken into account. For example, in [6], we divided the patients into two groups according to the admission decisions, and treated them separately with different arrival rate functions and LoS distributions.

In retrospect, after looking at the larger data set, we conclude that the model components M_2 and M_3 remain quite satisfactory, but for model component M_1 (the daily arrival totals), we find ways to improve the model, especially when we consider the prediction problem.

3. Analysis of the Larger Dataset

In this section, we will do some exploratory data analysis and basic regressions for the larger dataset, i.e., the data from January 2004 to October 2007, which is in total 1400 days. We will show that there is a clear long-term trend in the daily arrival totals as well as stochastic dependence.

3.1. An Overview of the Arrival Data

Figure 2 shows the daily arrival totals for the entire data set. Figure 2 shows that the daily arrival totals are stable over time, but have a slight increasing trend. The blue line in the figure is the estimated regression line. From this figure, we see that there is a period around the 950th day that the arrivals are significantly lower compared to the adjacent period. This time coincides with the 2006 Lebanon war, the war between Israel and Lebanon from July 12 to August 14, 2006. We regard those points as outliers. If we ignore that Lebanon war period, almost all points fall between 75 to 200. Figure 3 shows the data without that period. Again, the blue line is the fitted regression line. We see that the war period has a significant impact on the estimated slope for the regression line. (Since we are using ordinary least square estimation, the estimators are not robust to abnormal data points.) *Throughout this paper, we exclude this war period unless specified otherwise.*

The overall mean daily arrival total is 134.7, while the variance is 571.0. The slope of the daily arrival totals is very small. The mean daily volume increases about 1 every 100 days, but there is a 13.5 difference between the estimated mean of the last (1400$^{\text{th}}$) day and the first day, which is about a 10% increment.

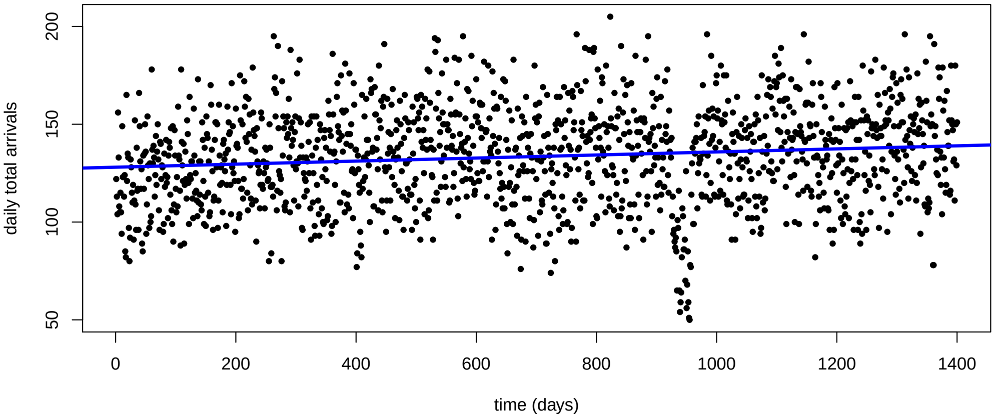


Figure 2: The daily arrival totals for the whole data set. The blue line is the regression line, $d = 128.0 + 0.00781 * t$, and the dashed red line is the average level.

In [6] we found that the system has a significant periodic structure with the period being 1 week, so we look at the daily arrival totals for each day-of-week separately and the weekly totals as well. Table 3.1 shows the sample mean, sample variance and variance-to-mean ratio of the daily arrival totals for each day-of-week. As in [6], we conclude that the variance-to-mean ratio is significantly greater than 1 on every day. Sunday, as the first day of week

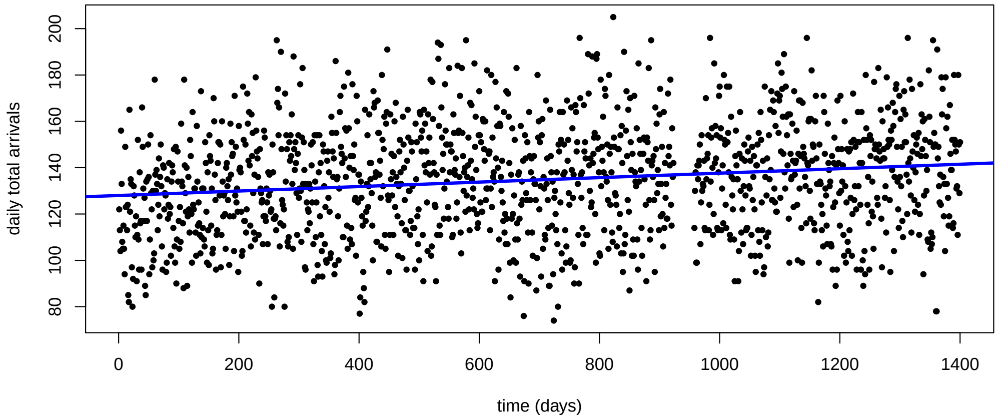


Figure 3: The daily arrival totals without the war period. The blue line is is the regression line, $d = 128.0 + 0.00964 * t$. The overall mean (shown by the red dashed line) is 134.7, while the variance is 571.0

in Israel, has the highest number of patient visits while Friday and Saturday (the weekend days) have relatively few patient visits. Figure 4 shows the weekly total arrivals, where we also fitted a regression line. It is evident that the weekly totals exhibit some time dependence structure. Figure 5 present a box plot view showing that the distributions of daily arrival totals vary over months. we expect more patients in the summer than in the winter. This suggests the daily arrival totals may be related to the temperature, as observed in [7, 10, 11, 12]. We explore this direction for improving our previous model later.

	Sun	Mon	Tues	Wed	Thurs	Fri	Sat	week
Mean	163.7	144.7	140.6	134.5	139.3	113.4	106.6	134.7
Var	276.1	292.5	294.7	253.0	302.6	181.1	163.6	571.0
V/M	1.69	2.02	2.10	1.88	2.17	1.60	1.69	4.24

Table 1: Sample mean, variance and variance-to-mean ratio of daily arrival totals for each day-of-week.

Next we look at the LoS distributions. In [6], we found that the LoS distributions are also time-varying, depending on the patient arrival time. Here we want to check if the LoS distributions changes in the long term.

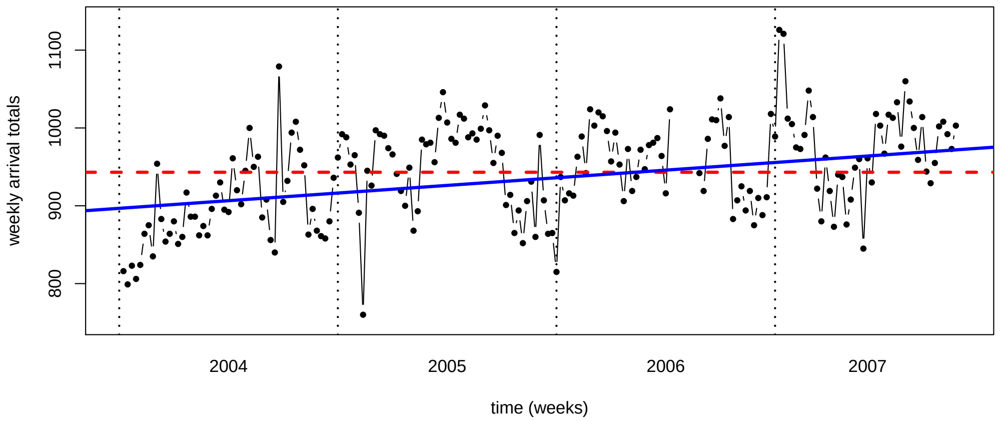


Figure 4: Number of weekly arrival totals. The blue line is what we got if we regress the weekly arrival totals on the index of day, which is $weekly\ totals = 896.72 + 0.378 * w$, where $w = 1, 2, \cdots, 199$ is the index of weeks.

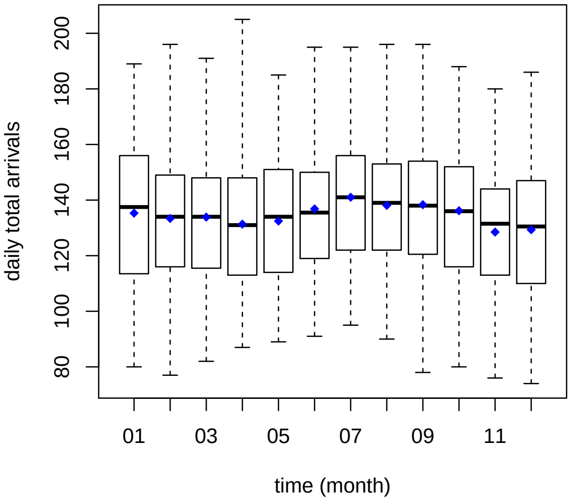


Figure 5: Number of daily arrival totals in a month view. The box together with the black bar show the quantiles of the daily arrival totals for each month, and the blue dots are the corresponding sample means. The dashed red line is the average level of daily arrival totals.

Figure 6 shows the LoS distributions in a monthly view. At a glance, the LoS distributions looks quite stable, except in 2006-08, right after the Lebanon war, where the LoS is significantly low. Figure 7 shows the regression lines for the monthly mean LoS and median LoS. It shows that both have a small but significant positive slope. This suggests that in the long term, we should not ignore the change of LoS distribution, but perhaps in a short term, we can safely assume it is stable.

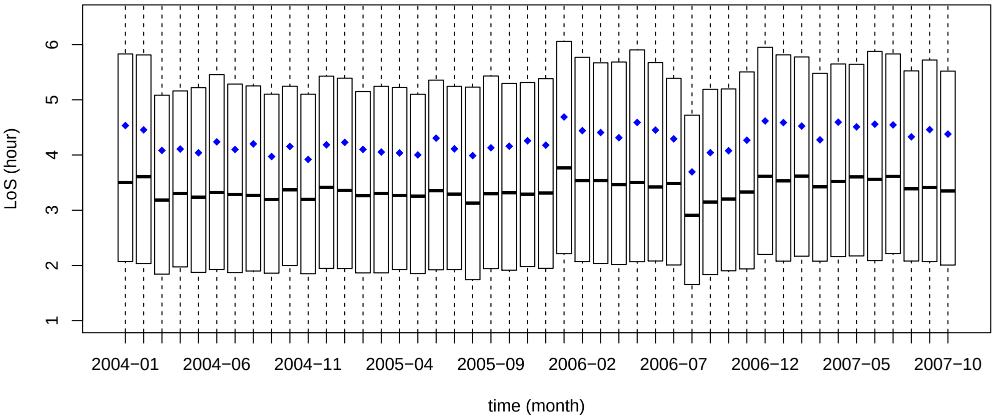
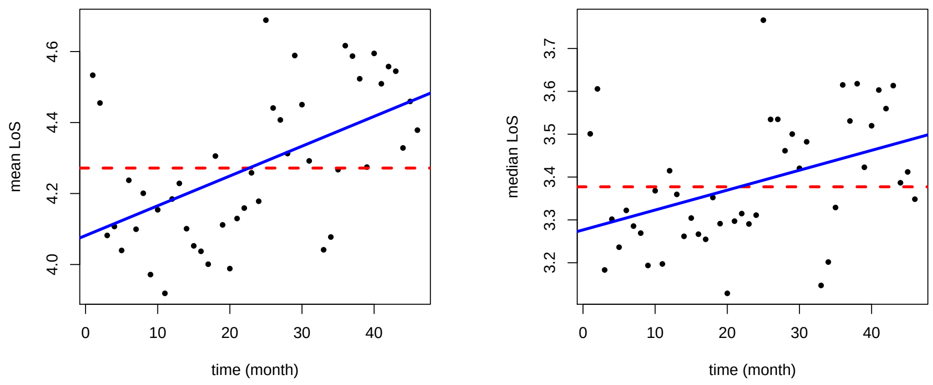


Figure 6: The quantiles of LoS distributions for each month. The box together with the black bar show the 0.25, 0.5 (median) and 0.75 quantile of the LoS distribution, and the blue dots are the sample mean of the LoS distribution.

3.2. Summary

After the analysis of the entire data set, we conclude that the model framework in [6] can still work, but with the larger dataset, we detect a trend and autocorrelation structure in the daily arrival totals. To be specific, we still consider that model components M_2 and M_3 are satisfactory, but we can be improve M_1 to better predict the daily arrival totals. An extended Gaussian model should still be appropriate. Moreover, it appears that we should be able to estimate the parameters dynamically, only using the recent data to fit the model and predict the near future. We should be careful not to apply the model for long-term forecasts, without focusing on the trend, because the daily arrival totals are increasing slowly. The same is true for the LoS distributions.



(a) Mean LoS v.s. time. Slope = 0.0084 (b) Median LoS v.s. time. Slope = 0.0046

Figure 7: Linear regression of monthly mean and median LoS on time. The slopes are very small, but statistically significant. The mean LoS increases from about 4.1 hour to 4.5 hour. while the median also grows at a lower rate. The dashed red lines are the mean and median of LoS of the entire data respectively.

But even for short term forecasting, there are alternative methods to consider. In the next section we next consider five alternative forecasting methods.

4. Forecasting the Daily Arrival Totals

In this section, we consider alternative ways to foreacst the daily arrival totals. We first focus on one-day ahead prediction, then we consider predicting more days into the future in §4.7.

We divide the dataset into a training set and a test set. For simplicity, we use the data before the Lebanon war as the training set (923 days from Jan. 1, 2004 to July 11, 2006) and the rest as the test set (443 days from Aug. 15, 2006 to Oct. 31, 2007). We use the training set to find the optimal number of weeks we should include when we want to predict next week and to estimate the parameters. Then we use the test set to check our choice and compare with other methods. We use mean square error (MSE, by which we mean the estimate) to measure the precision of our prediction, defined by

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{5}$$

where N is the sample size, y_i are the true values and $\hat{y}_i$ are the predicted values.

4.1. A Dynamic Model

An easy way to revise our old model for the daily arrival totals is to use it in a dynamic way, i.e. if we want to predict the daily arrival totals for next week, then we fit the model only using a few weeks of history data right before it. We keep updating our model parameter according to what we observe. In particular, if we let A_t represent the daily arrival totals for day t and $\hat{A}_t$ be our estimate for it, then the estimator is

$$\hat{A}_t = \frac{1}{n}(A_{t-7} + A_{t-14} + \cdots + A_{t-n*7}). \tag{6}$$

We need to determine n in the above equation. We try different value of n from 1 to 30 weeks and pick the one with the smallest training MSE. Figure 8 and Table 2 show the training MSEs using different n. We see that it shows a typical U-shape as a function of n, reaching the minimal training MSE 248.3 at $n = 13$, so we choose this value of n and apply it on the test set. The test MSE is 269.94. (The difference provides an estimate of the overestimation of statistical precision caused by testing on the same data used to fit the model.)

For comparison, we observe that the overall mean and variance are 134.7 and 571.0. If we use the single-factor model on this new test set from [6], then the estimated residual variance is 264.6, which yields $V/M = 1.9$.

4.2. A SARIMA Time Series Model

We observe that the predictor in (6) is actually a special case of the classic autoregressive model

$$A_t = c + \alpha_1 A_{t-1} + \alpha_2 A_{t-2} + \cdots + \alpha_p A_{t-p} + \epsilon_t, \tag{7}$$

where we take $c = 0$, $p = 7n$, $\alpha_i = 1/n$ for $i = 7, 14, \cdots, 7n$ and 0 otherwise. So we next try a more general and more flexible autoregressive

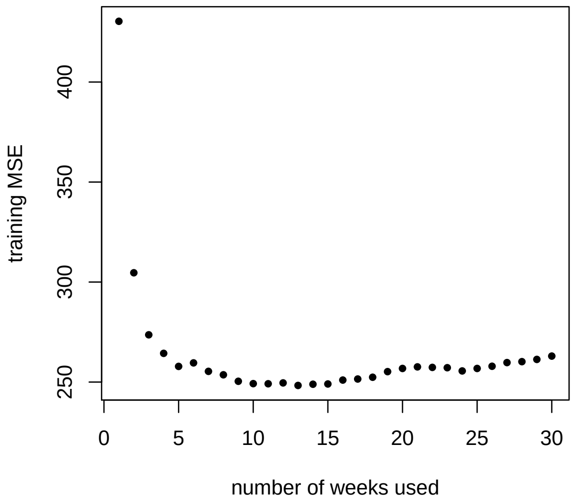


Figure 8: Training MSE as a function of the number n of weeks history to predict the daily arrival totals of the next day for the dynamic model in §4.1. The minimum training MSE is achieved when $n = 13$.

n	1	2	3	4	5	6	7
MSE	430.4	304.6	273.6	264.4	257.8	259.6	255.4
n	8	9	10	11	12	**13**	14
MSE	253.6	250.4	249.2	249.2	249.6	**248.3**	248.9
n	15	16	17	18	19	20	21
MSE	249.1	251.0	251.5	252.4	255.2	256.8	257.6
n	22	23	24	25	26	27	28
MSE	257.3	257.2	255.6	256.8	257.9	259.8	260.2

Table 2: Training MSE as a function of the number n of weeks history to predict the daily arrival totals of the next day for the dynamic model in §4.1.

integrated moving average (ARIMA) model. Because we have determined that there is periodicity, we use a seasonal ARIMA (SARIMA) model. It has 7 hyperparameters that need to be determined and can be denoted as SARIMA$(p, d, q)(P, D, Q)_m$, where p, d and q are the order of AR terms, the order of difference and the order of MA terms, respectively, while P, D and Q are the corresponding seasonal orders, and m is the period length of each season. This model is quite standard in time series analysis; see Chapter 6 of [21].

Obviously we should take $m = 7$, because we think the period is 1 week. From our analysis in §3.1, we know that the arrival rate is increasing slowly. So it is reasonable to conduct a difference for the original series, which is equivalent to assume that the time series has a stationary increasing trend. But whether to take the difference directly (i.e., setting $d = 1$, $D = 0$, to model $\{A_t - A_{t-1}\}$ by an ARMA model) or do it seasonally (i.e., setting $D = 1$, $d = 0$, to model $\{A_t - A_{t-7}\}$) needs to be determined. We will check them one by one.

Suppose that we make a difference directly and let $\{X_t \equiv A_t - A_{t-1}\}$. First, Figure 9 examines $\{X_t\}$. We conduct the Dickey-Fuller test to check whether this time series can be regarded as stationary. Because the p-value is $1.69 * 10^{-23}$, we reject the hypothesis that this time series has a unit root, and so tentatively conclude that the process is stationary. Figure 10 shows the autocorrelation function (ACF) and partial autocorrelation function (PACF) of $\{X_t\}$, which could help us determine the orders we choose in ARMA. We see that the partial correlation function vanishes after some point for both the seasonal factor and non-seasonal factor, while the autocorrelation function does not go away, so that we try to model it as an AR sequence. We try p from 1 to 6 and P from 1 to 16. According to Akaike information criterion (AIC), SARIMA$(6, 1, 0)(15, 0, 0)_7$ is the best model, whose AIC is 7663.47 and training MSE is 227.31.

Similarly, for the other alterantive, suppose we make a seasonal difference and model $\{Y_t \equiv A_t - A_{t-7}\}$. Figure 11 shows its rolling mean and standard deviation, while Figure 12 shows its ACF and PACF. Again we conduct the Dickey-Fuller test and reject the null hypothesis that the series has a unit root with a p-value $1.66 * 10^{-15}$. We see that in contrary to $\{X_t\}$, the ACF of $\{Y_t\}$ cuts off after the first period, while the PACF does not vanish. This suggests adding seasonal MA terms to the model. We try p from 1 to 6. We find that a large value of Q causes the maximum likelihood estimation to converge poorly, so that the AIC does not improve significantly when we

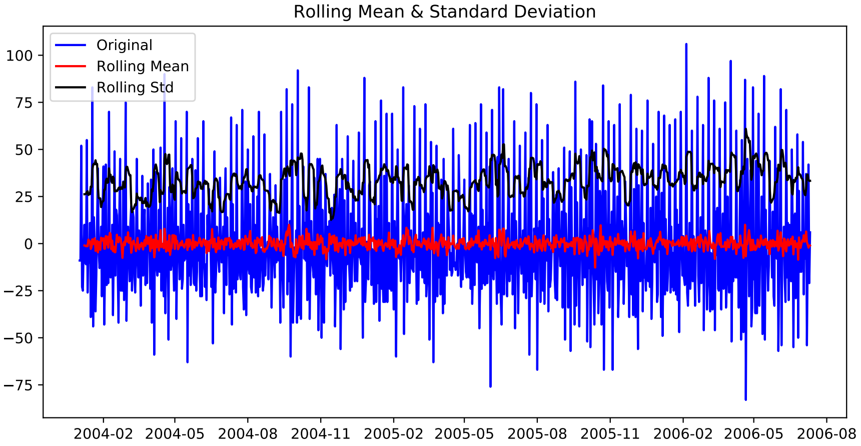


Figure 9: Daily arrival totals after taking a difference ($\{X_t \equiv A_t - A_{t-1}\}$) for the SARIMA model in §4.2. We also show the rolling sample mean and sample standard deviation using a window with width 7.

increase Q. Hence, we choose the model SARIMA$(6,0,0)(0,1,1)_7$. Its AIC is 7564.73 and training MSE is 221.51.

We conclude that the SARIMA$(6,0,0)(0,1,1)_7$ outperforms the previous SARIMA$(6,1,0)(15,0,0)_7$. Thus our fitted model is

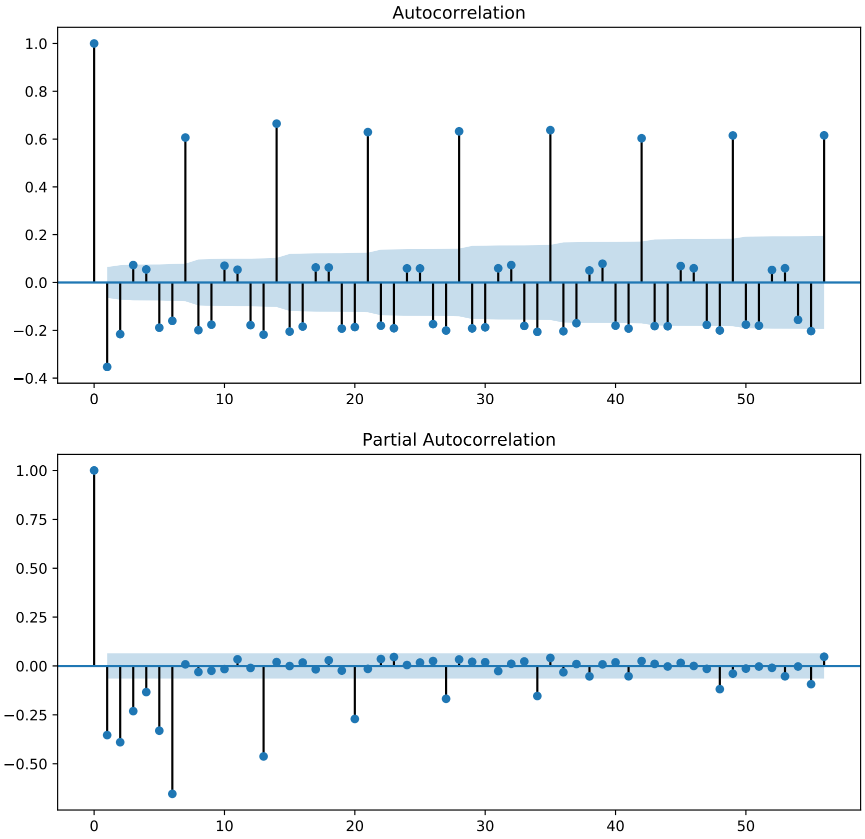
$$(A_t - A_{t-7}) - \sum_{i=1}^{6} p_i(A_{t-i} - A_{t-i-7}) = \epsilon_t + Q_1\epsilon_{t-7}, \tag{8}$$

where p_i, $i = 1,2,3,4,5,6$, and Q_1 are coefficients, $\{\epsilon_t \sim N(0,\sigma^2)\}$ are independent normal distributed noise. The maximum likelihood estimation for those parameters are shown in Table 3.

Parameter	p_1	p_2	p_3	p_4	p_5	p_6	Q_1	σ^2
MLE	0.181	0.012	0.062	0.036	0.070	0.144	-0.948	218.30

Table 3: MLE for the parameters of the SARIMA$(6,0,0)(0,1,1)_7$ model in (8).

For this model, we also checked whether the residuals are approximately independent and normally distributed. A positive conclusion is supported by the ACF and PACF in Figure 13 and the quantile-to-quantile (Q-Q) plot compared to normal distribution in Figure 14.



Autocorrelation

Partial Autocorrelation

Figure 10: The autocorrelation function and partial correlation function of $\{X_t\} = \{A_t - A_{t-1}\})$ for the SARIMA model in §4.2.

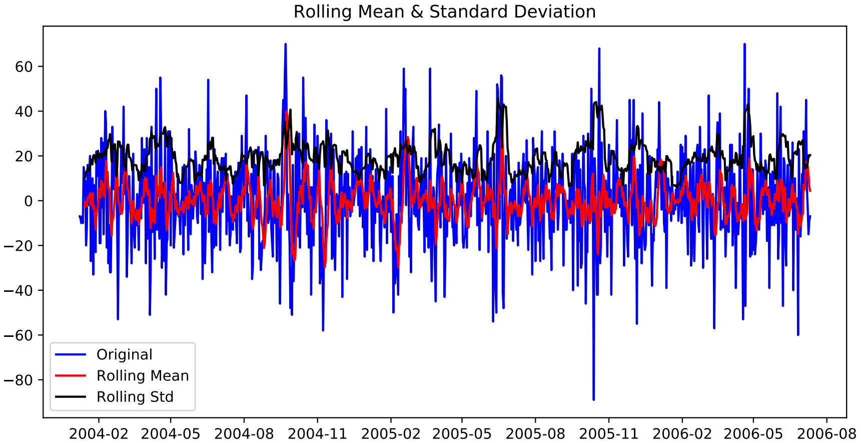


Figure 11: Daily arrival totals after taking a seasonal difference ($\{Y_t \equiv A_t - A_{t-7}\}$) for the SARIMA model in §4.2. We also show the rolling sample mean and sample standard deviation using a window with width 7.

Finally, we test SARIMA$(6, 0, 0)(0, 1, 1)_7$ on our test set and find that the MSE is 263.40. We show the predicted daily arrival totals versus the true values in Figure 15, and in Figure 16 we show the 95% confidence interval for part of test set.

4.3. A Regression Method with More Calendar and Weather Information

Many researchers have found that calendar and weather variables can be very helpful for predicting the daily arrival totals in the ED [7, 10, 11], so we want to explore this direction. Of course, our original model in [6] already is a simple version, but it considers only the day-of-week factor. We find that the day-of-week factor is most important, but there is also potential for taking advantage of other factors, including month, temperature and holiday.

We consider a regression model that includes the following independent variables: day-of-week, month, max daily temperature, min daily temperature, daily precipitation, and holiday$\pm k$. The day-of-week and month factors are straightforward. For the daily max/min temperature and the daily precipitation, since we are considering a prediction model, ideally we should use the temperatures in the weather forecast one day before, but we could not find that information, so we used the real historical data which is published by Israel Meteorological Service (available at https://ims.data.gov.il). We

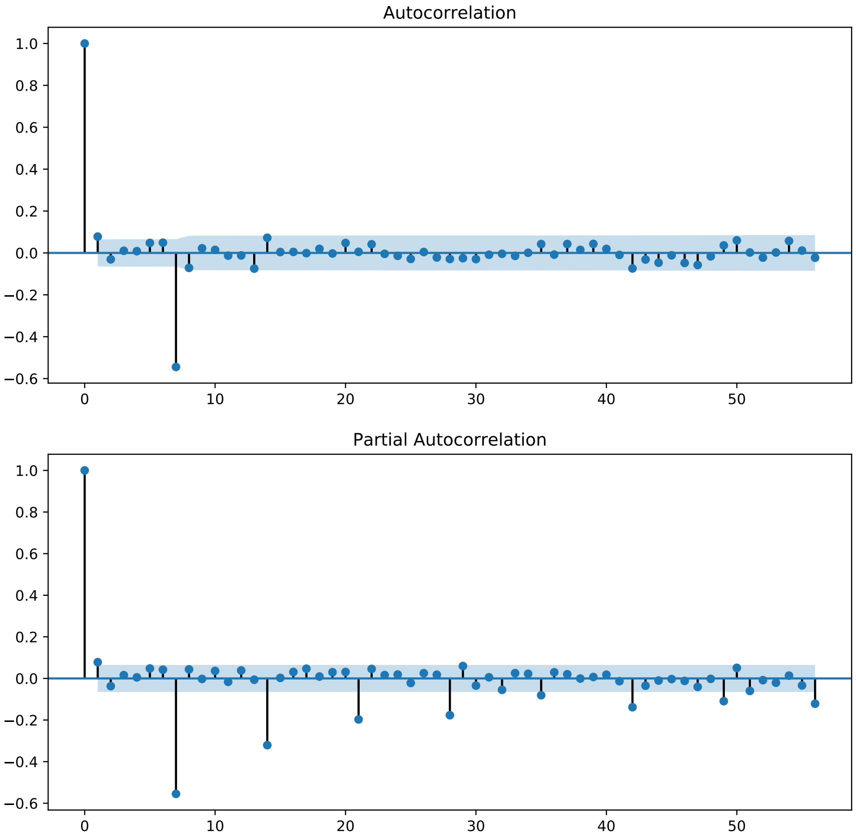


Figure 12: The autocorrelation function and partial correlation function of $\{Y_t \equiv A_t - A_{t-7}\}$ for the SARIMA model in §4.2.

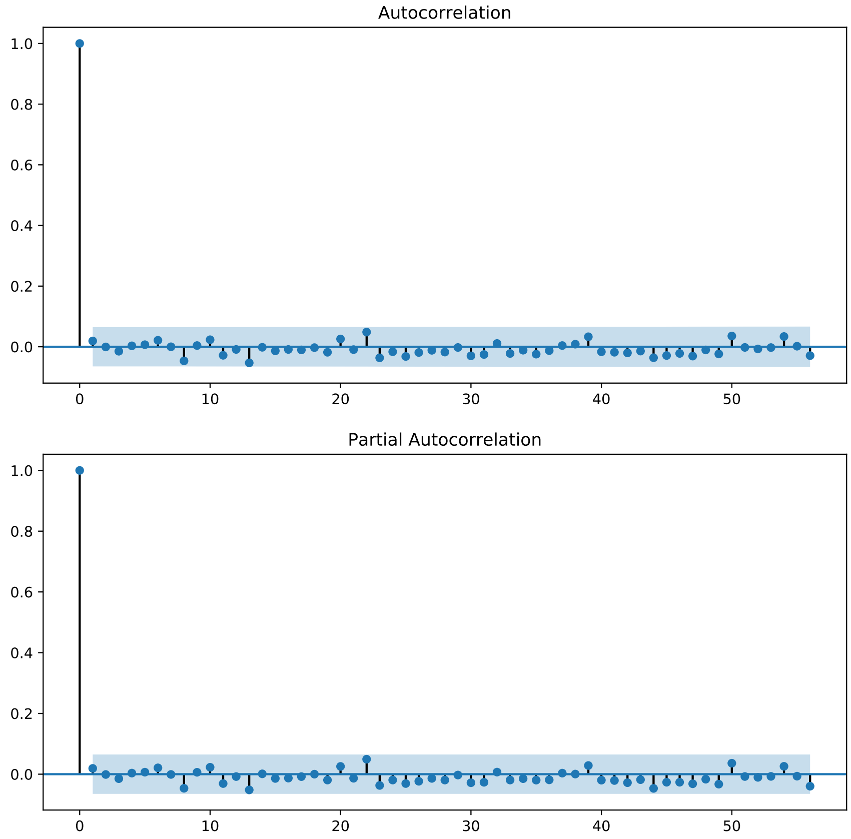


Figure 13: The autocorrelation function and partial correlation function of the residuals for model SARIMA$(6, 0, 0)(0, 1, 1)_7$ in (8).

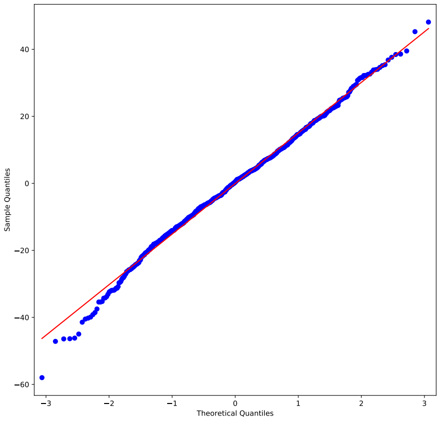


Figure 14: The Q-Q plot of the residuals for model SARIMA$(6, 0, 0)(0, 1, 1)_7$ in (8).

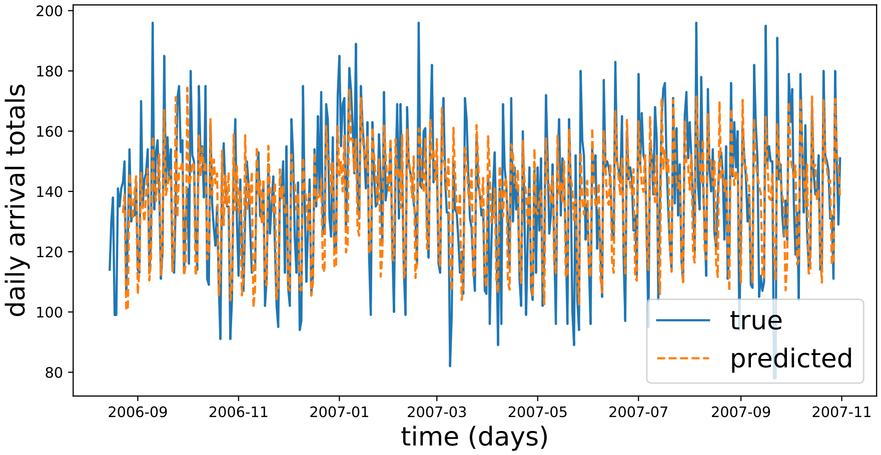


Figure 15: Test SARIMA$(6, 0, 0)(0, 1, 1)_7$ as in (8) on the test set.

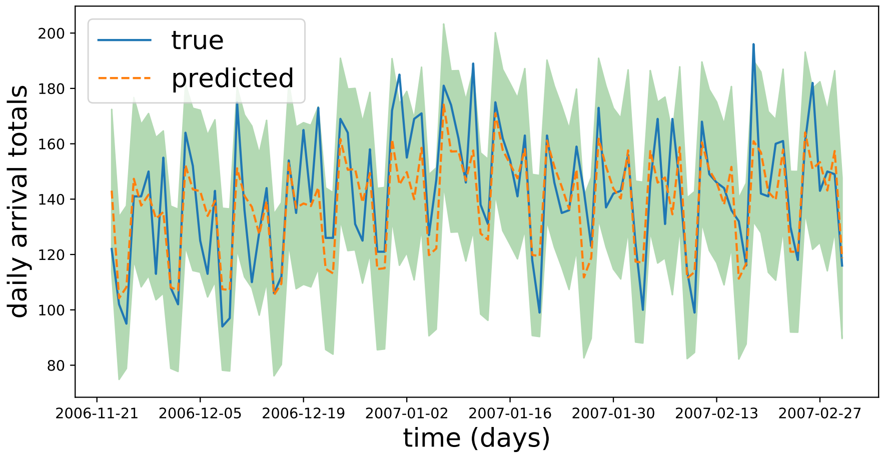


Figure 16: Confidence interval (95%) for the predicted daily arrival totals by SARIMA$(6,0,0)(0,1,1)_7$ as in (8) on the test set.

can understand this as the true relationship between weather and the daily arrival totals, and when applying, we can use the information from weather forecast as estimators. We think this approach is reasonable because, with modern technology, we can predict the weather on the next day quite accurately. Another issue is the location where the weather data is collected. The hospital itself does not have a meteorological station, so we choose the nearest one to the Rambam Hospital which is located at Haifa port. It turns out that there is missing data for one week in September 2007 and one day in January 2007. Hence, for those days we use the temperature data from the nearest meteorological station, which is located at the Technion as proxies for the missing data.

Holidays can be another potential factor affecting the daily arrival totals. In [7], they considered the holiday and near-holiday factor when predicting the daily patient volume of three EDs in the United States. They defined a day as near-holiday if it is one day before or after a public holiday. Here we think we should distinguish days before and after a holiday, so we actually use 7 indicators to tell if a day is holiday$\pm k$ day, where $k = -3, -2, -1, 0, 1, 2, 3$. We mark a day as holiday only if it is a national holiday. In summary, the full model we propose is

$$\begin{aligned} A_t &= \beta_0 + \beta_{Sun} I_{Sun}(t) + \beta_{Mon} I_{Mon}(t) + \cdots + \beta_{Sat} I_{Sat}(t) \\ &\quad + \beta_{Jan} I_{Jan}(t) + \beta_{Feb} I_{Feb}(t) + \cdots + \beta_{Dec} I_{Dec}(t) \\ &\quad + \beta_{T-max} T_{max}(t) + \beta_{T-min} T_{min}(t) + \beta_{rain} R(t) + \beta_{H-3} I_{H-3}(t) \\ &\quad + \beta_{H-2} I_{H-2}(t) + \beta_{H-1} I_{H-1}(t) + \beta_H I_H(t) + \beta_{H+1} I_{H+1}(t) \\ &\quad + \beta_{H+2} I_{H+2}(t) + \beta_{H+3} I_{H+3}(t) + \epsilon_t, \end{aligned} \tag{9}$$

where A_t again represents the total arrivals of day t, $I_{Month/Day-of-week}(t)$ is the indicator of that month or day-of-week, $T_{max}(t)$ and $T_{min}(t)$ are the highest and lowest temperature of day t, $R(t)$ is the precipitation (in cm) of day t, $I_{H\pm k}(t)$ is the indicator of near holiday effect as we explained above, β's are the corresponding coefficients and constant and finally $\epsilon_t \sim N(0, \sigma^2)$ is a normal distributed error term.

Of course, we do not regard the model above as the best one, because so far we have included all possible factors. The regression result shown in Table 4 indeed indicates that some factors such as precipitation may not be important. To select our final model, we use a two-way stepwise model selection procedure based on AIC; i.e., we start from the full model as above and in each step, we exclude or include one factor at a time, based on the AIC; see section 9.4 of [22]. After this procedure, the remaining factors are day-of-week, month, holiday+0, holiday+1, holiday+2, holiday−1, holiday−3, max daily temperature and min daily temperature. In the final model, we further exclude the holiday+2 and holiday−3, because they are the least two important factors among those above and the AIC will increase only 0.5 from 4962.8 to 4963.3 if we exclude them. Also we think including holiday−1, holiday−3 but not holiday−2 is not very reasonable. So in the end we select the model in (10).

$$\begin{aligned} A_t &= \beta_0 + \beta_{Sun} I_{Sun}(t) + \beta_{Mon} I_{Mon}(t) + \cdots + \beta_{Sat} I_{Sat}(t) \\ &\quad + \beta_{Jan} I_{Jan}(t) + \beta_{Feb} I_{Feb}(t) + \cdots + \beta_{Dec} I_{Dec}(t) \\ &\quad + \beta_{T-max} T_{max}(t) + \beta_{T-min} T_{min}(t) \\ &\quad + \beta_{H-1} I_{H-1}(t) + \beta_H I_H(t) + \beta_{H+1} I_{H+1}(t) + \epsilon_t. \end{aligned} \tag{10}$$

Table 5 shows the estimation of coefficients for the final model using the training set. We can see that the day-of-week factor and the holiday+0 are

	Estimate	Standard Error	p-value
Intercept(β_0)	92.56	4.95	<0.001
Sunday(β_{Sun})	51.46	1.80	<0.001
Monday(β_{Mon})	32.59	1.80	<0.001
Tuesday(β_{Tue})	28.09	1.80	<0.001
Wednesday(β_{Wed})	20.91	1.80	<0.001
Thursday(β_{Thu})	24.75	1.79	<0.001
Friday(β_{Fri})	0		
Saturday(β_{Sat})	-7.31	1.79	<0.001
January(β_{Jan})	0.20	2.68	0.942
February(β_{Feb})	-0.93	2.67	0.729
March(β_{Mar})	2.70	2.44	0.268
April(β_{Apr})	0		
May(β_{May})	-5.29	2.32	0.023
June(β_{Jun})	-5.17	2.77	0.062
July(β_{Jul})	-5.83	3.25	0.073
August(β_{Aug})	-7.75	3.43	0.024
September(β_{Sep})	-5.75	3.14	0.067
October(β_{Oct})	-5.43	2.65	0.041
November(β_{Nov})	-7.61	2.62	0.004
December(β_{Dec})	-3.83	2.67	0.151
Holiday+0(β_H)	-21.17	2.86	<0.001
Holiday+1(β_{H+1})	9.33	3.09	0.003
Holiday+2(β_{H+2})	4.83	3.08	0.117
Holiday+3(β_{H+3})	1.50	3.07	0.626
Holiday−1(β_{H-1})	-17.25	3.10	<0.001
Holiday−2(β_{H-2})	-1.56	3.11	0.616
Holiday−3(β_{H-3})	-4.69	3.10	0.130
Max Temp.(β_{T-max})	0.45	0.22	0.042
Min Temp.(β_{T-min})	0.70	0.30	0.018
Precip.(β_{rain})	0.02	0.11	0.876
σ^2	211.12		

Table 4: Estimated coefficients for the full linear regression model with calendar and weather variables in *(9)* in §4.3. (Friday and April are chosen as the base line for the categorical variables day-of-week and month based on alphabetical order.)

the most important ones. There are fewer patients the day before holiday and more patients the day after holiday. We also see that both the max and min daily temperature have a positive coefficients.

We make prediction using the test set, and the MSE is 234.33.

4.4. The SARIMAX Model

Though the model exploiting holiday and weather data in §4.3 has pretty good results, it does not capture any internal dependence. When we look at the residuals of final regression model in §4.3, as is shown in Figure 17 and 18, we can see that the residuals have an increasing trend and positive autocorrelation structure. We conducted the Mann-Kendall trend test (see [23, 24]) on the residuals and it strongly rejects the null hypothesis that the series does not have a trend.

On the other hand, the SARIMA model we use in §4.2 does capture internal dependence, but does not include any external information such as the holiday and temperature factors. We now consider a SARIMAX model, which combines the two approaches. In particular, it is an extended version of the SARIMA model, includes both (seasonal and non-seasonal) AR and MA terms like the SARIMA model and other independent variables.

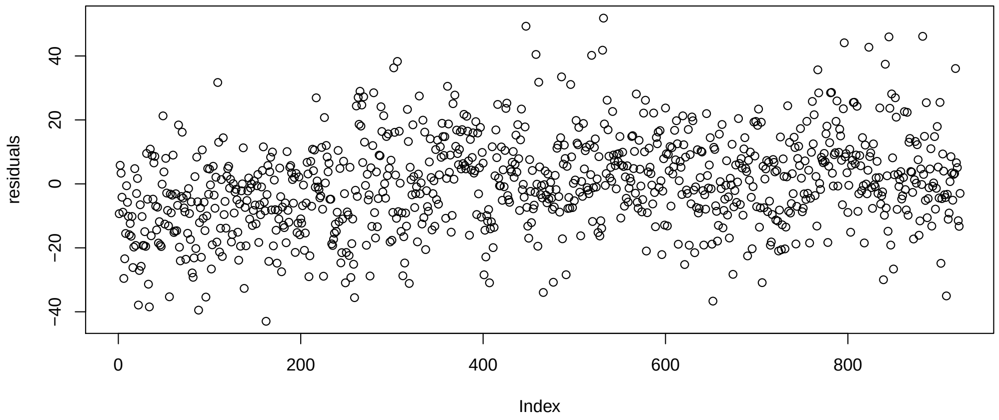


Figure 17: The residuals of the regression model (10) in §4.3.

To express the SARIMAX model clearly in a concise way, we introduce the backshift (or lag) operator B, with $By_t = y_{t-1}$, which is commonly used

	Estimate	Standard Error	p-value
Intercept(β_0)	92.83	4.76	<0.001
Sunday(β_{Sun})	51.23	1.79	<0.001
Monday(β_{Mon})	32.34	1.80	<0.001
Tuesday(β_{Tue})	27.96	1.80	<0.001
Wednesday(β_{Wed})	20.89	1.80	<0.001
Thursday(β_{Thu})	24.83	1.80	<0.001
Friday(β_{Fri})	0		
Saturday(β_{Sat})	-7.43	1.79	<0.001
January(β_{Jan})	0.29	2.53	0.908
February(β_{Feb})	-0.86	2.53	0.734
March(β_{Mar})	2.77	2.29	0.226
April(β_{Apr})	0		
May(β_{May})	-5.07	2.28	0.026
June(β_{Jun})	-4.94	2.69	0.067
July(β_{Jul})	-5.64	3.16	0.074
August(β_{Aug})	-7.54	3.34	0.024
September(β_{Sep})	-5.66	3.11	0.069
October(β_{Oct})	-5.13	2.64	0.053
November(β_{Nov})	-7.48	2.49	0.003
December(β_{Dec})	-3.72	2.59	0.152
Holiday+0(β_H)	-21.20	2.81	<0.001
Holiday+1(β_{H+1})	9.08	3.05	0.003
Holiday−1(β_{H-1})	-17.15	3.05	<0.001
Max Temp.(β_{T-max})	0.44	0.22	0.044
Min Temp.(β_{T-min})	0.70	0.30	0.019
σ^2	211.12		

Table 5: Estimated coefficients for the selected linear regression model with calendar and weather variables from *(10)* in §4.3. (Friday and April are chosen as the base line for the categorical variables day-of-week and month based on alphabetical order.)

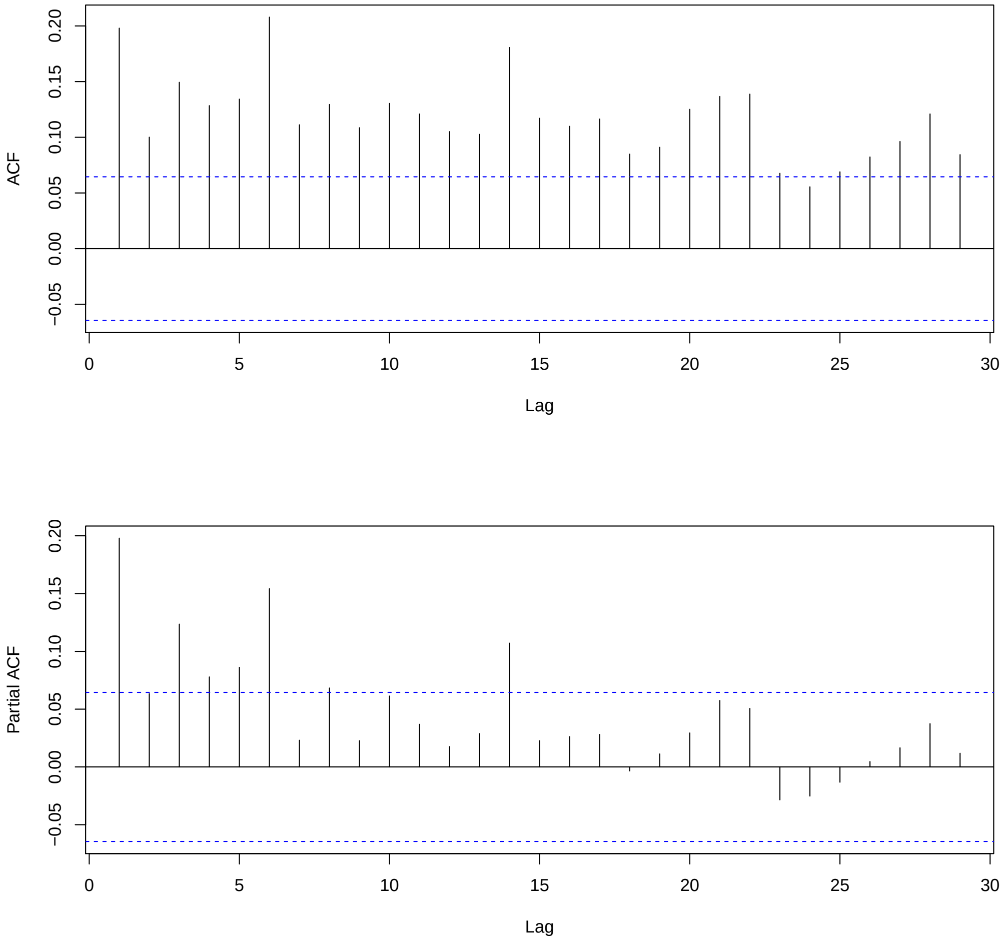


Figure 18: ACF and PACF of the residuals of the regression model (10) in §4.3.

in time series analysis. We also define the associated operators

$$\begin{aligned}
\phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \cdots \phi_p B^p, \\
\Phi(B) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}, \\
\theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q, \\
\Theta(B) &= 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots \Theta_Q B^{Qs}, \\
\Delta &= 1 - B, \\
\Delta_s &= 1 - B^s, \qquad (11)
\end{aligned}$$

where $\phi(\cdot)/\Phi(\cdot)$ is the non-seasonal/seasonal AR polynomial, $\theta(\cdot)/\Theta(\cdot)$ is the non-seasonal/seasonal MA polynomial and Δ/Δ_s is the non-seasonal/seasonal difference operator respectively. A SARIMA$(p, d, q)(P, D, Q)_s$ model can be formally represented by

$$\phi(B)\Phi(B)\Delta^d\Delta_s^D y_t = \theta(B)\Theta(B)\epsilon_t,$$

where $\epsilon_t \sim N(0, \sigma^2)$ is a Gaussian white noise. If we allow external variables to explain the mean of the transferred time series, then we get the SARIMAX model

$$\phi(B)\Phi(B)\Delta^d\Delta_s^D y_t = x_t^T\beta + \theta(B)\Theta(B)\epsilon_t, \qquad (12)$$

where x_t is the external variables and β is the corresponding coefficients. We see that this can be viewed as a generalization of both the SARIMA model and the ordinary linear regression model. See §6.6 of [21] for more about this model.

Based on our analysis in §4.3, we directly use the variables we chose there in (10) as external regressors. Then we conduct the same model selection procedure as we did in §4.2 based on AIC, and it turns out the model SARIMAX$(6, 1, 0)(0, 0, 2)_7$ is the best one. If we write it explicitly, it is

$$A_t = x_t^T\beta + A_{t-1} + \sum_{i=1}^{6}\phi_i(A_{t-i} - A_{t-1-i}) + \epsilon_t + \Theta_1\epsilon_{t-7} + \Theta_2\epsilon_{t-14}, \qquad (13)$$

where x_t includes all the variables (except the constant term) in (10). We see that the optimal model here is different from the final model in §4.2. In §4.2 we take a seasonal difference on the original time series while here we take a nonseasonal difference. But either implies that the original time series (daily arrival totals) has a long-term trend. The maximum likelihood estimation of

	Estimate	Standard Error		Estimate	Standard Error
Sunday(β_{Sun})	51.04	2.08	ϕ_1	-0.90	0.03
Monday(β_{Mon})	32.04	2.01	ϕ_2	-0.90	0.04
Tuesday(β_{Tue})	27.66	2.01	ϕ_3	-0.84	0.05
Wednesday(β_{Wed})	20.68	2.08	ϕ_4	-0.81	0.06
Thursday(β_{Thu})	24.77	1.72	ϕ_5	-0.78	0.07
Friday(β_{Fri})	0		ϕ_6	-0.70	0.08
Saturday(β_{Sat})	-7.48	1.72	Θ_1	-0.70	0.09
January(β_{Jan})	1.01	4.00	Θ_2	0.08	0.03
February(β_{Feb})	-1.03	3.58	σ^2	187.9	
March(β_{Mar})	2.12	2.81			
April(β_{Apr})	0				
May(β_{May})	-5.71	2.67			
June(β_{Jun})	-6.03	3.57			
July(β_{Jul})	-5.92	4.23			
August(β_{Aug})	-7.43	4.81			
September(β_{Sep})	-6.54	4.97			
October(β_{Oct})	-8.46	4.91			
November(β_{Nov})	-11.44	4.79			
December(β_{Dec})	-8.62	4.60			
Holiday+0(β_H)	-20.23	2.64			
Holiday+1(β_{H+1})	9.12	2.81			
Holiday$-$1(β_{H-1})	-16.31	2.81			
Max Temp.(β_{T-max})	0.55	0.20			
Min Temp.(β_{T-min})	0.62	0.28			

Table 6: Estimated coefficients for the SARIMAX$(6,1,0)(0,0,2)_7$ model in (13) from §4.4. (Again Friday and April are chosen as the base line for the categorical variables day-of-week and month based on alphabetical order.)

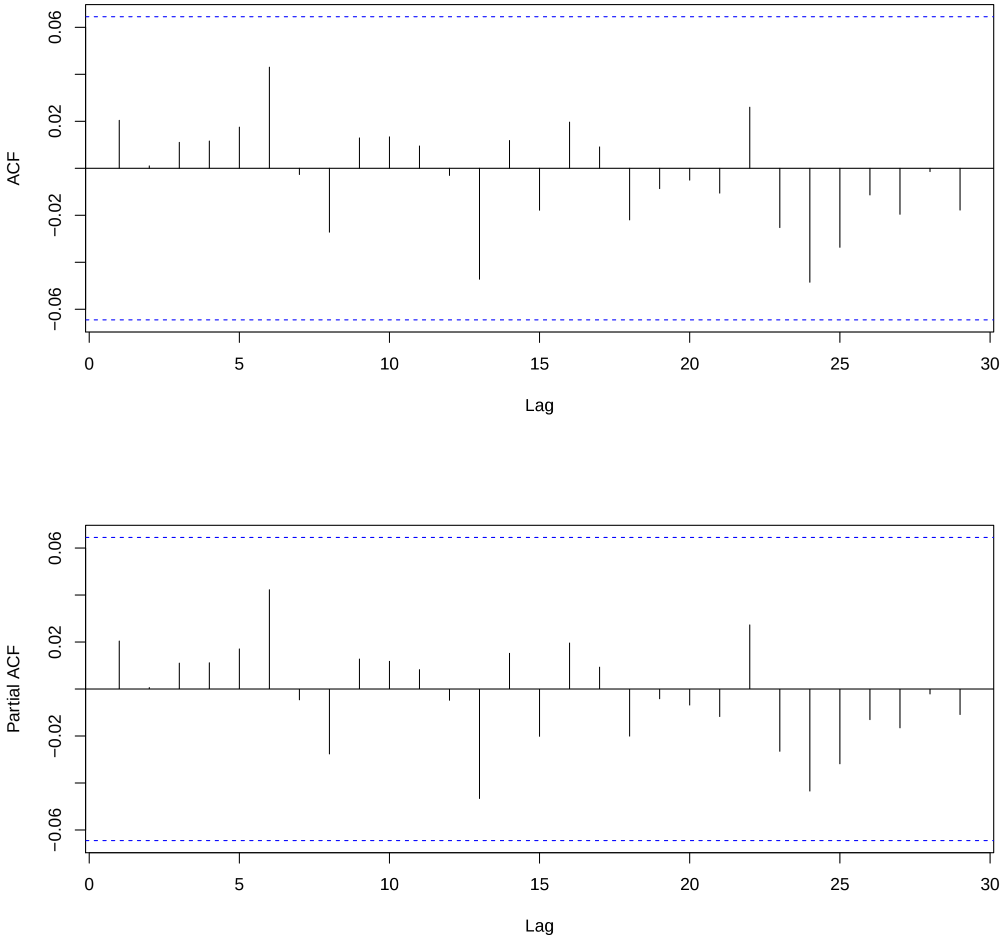


Figure 19: ACF and PACF of the residuals of the SARIMAX$(6, 1, 0)(0, 0, 2)_7$ model as in (13).

the coefficients are shown in Table 6. We also checked the ACF and PACF of the residuals in Figure 19 and find that they no longer have any significant autocorrelation structure.

We apply our fitted SARIMAX model on the test set and find that test MSE is 191.76, which is significantly better than both the SARIMA model in §4.2 and the regression model with only external variables in §4.3.

4.5. A Neural Network Method

Finally, we consider predicting the daily arrival totals with an even more flexible machine learning method, in particular, the multilayer perceptron (MLP). We refer to Chapter 11 of [25] for the basic concepts of MLP and to [26] for more on implementation issues. Here we only give a very briefly introduction to assist readers who are not familiar with machine learning get a rough idea about what we are doing.

Suppose that we have samples $(x_t, y_t) \in (\mathbb{R}^d, \mathbb{R})$, $t = 1, 2, \cdots, T$, where we want to use x_t to predict y_t, i.e. we think

$$y_t = f(x_t) + \epsilon, \tag{14}$$

for some unknown but determined function f and random error ϵ. If we take f to be a linear function, then (14) is a classical linear regression model. With machine learning, f is a relatively complicated nonlinear function, more like a black box. (Neural network models originally were inspired by trying to abstract how the human brain works.)

For the basic MLP version of the neural network model that we use here, we take f to be an iterated function, i.e., $f = g_l \circ g_{l-1} \circ \cdots \circ g_1$, where each g_i takes the form of

$$g_i(x) = (\psi(b_{i1}+x^T w_{i1}), \psi(b_{i2}+x^T w_{i2}), \cdots, \psi(b_{ih_i}+x^T w_{ih_i})), \quad i = 1, 2, \cdots, l,$$

with $b_{ij} \in \mathbb{R}$ and w_{ij}, $j = 1, 2, \cdots, h_i$ are coefficients of the proper dimension to be determined and $\psi(\cdot)$ is usually an S-shape function called “the activation function”. Common choices of ψ are the logistic function, the hyperbolic tangent and the rectified linear unit. The key point is that the non-linearity of ψ allows that f could approximate a very broad family of functions. Within this MLP model, l is the number of layers and h_i is the number of neurons in the i^{th} hidden layer. Each function ψ mimics a neuron which can be ”activated” or not depending on the input and the feature that the neuron can detect.

The training of a neural network (finding the best values for the large number of coefficients) used to be an extremely challenging problem, but significant progress has been made recently when general processing units (GPU's) or even more specialized hardware were developed to make the heavy computation feasible. Also some new optimization algorithms were invented to speed up the training process, such as stochastic gradient decent. Usually the hardest part of traditional gradient-based optimization algorithm is to compute the gradient. The main idea of stochastic gradient decent is that since the form of objective function (usually a loss function we defined, e.g., the training set MSE) is a summation of similar components, instead of computing the full gradient, we can "sample" a small portion of the terms in the objective function and compute the gradient and use that to approximate the full gradient. At the expense of some lost accuracy, the computation load is greatly reduced.

Now we will specify the inputs, the activation function and other settings. Given that A_t is the number of arrivals on day t, we aim to predict A_t given $(A_{t-1}, A_{t-2}, \cdots, A_{t-s})$ as well as other variables, just as in previous sections, where s is a parameter to be determined. Given s, we will have $923 - s$ samples (each sample is a pair of input variable and output variable) in the training set. We use all the candidate external variables introduced in §4.3 together with 28 days history (i.e. $s = 28$) as input. We assume that the number of hidden layers (d) can be 1 or 2. Since the final dimension of the input data is $57 = 28$(days of history arrivals) $+$ 3(weather variables) $+$ 7(holiday indicators) $+$ 7(week indicators) $+$ 12(month indicators), and the training sample size is about 900, we avoid more hidden layers or hidden neurons, because that could cause over-fitting.

In summary, we tried models with a single hidden layer and let number of hidden neurons (h) range from 2 to 10. We then selected the best according to cross-validation error. In each round of training we randomly set 10% of the training samples and validation set. We also add a l_2 regularization to each hidden layer in order to prevent over-fitting. We used the "Adam Optimizer" stochastic gradient decent algorithm. We let the batch-size be 1 and stop the iteration if the validation loss is not improved in 5 iterations. Then we record the validation loss (which is mean square error on the validation set).

The training results are reported in Table 7. Since the optimization algorithm and the cross-validation set are both random, we repeated the training of each model 20 times and computed the average cross-validation MSE and its standard deviation. We see that a simple MLP with 2 hidden neurons actually works best. So we use it by repeating the optimization 10 times and picking the one with the minimum cross validation MSE.

Then we test it using the test set, and the test MSE is 265.84. Figure 20 compares the predicted number of arrivals to the actual data. We see that the MLP approach for prediction performs approximately as well as the dynamic model in §4.1 and the SARIMA model in §4.2, but not as well as the other two models, even though we included temperature and hospital data with MLP, as in SARIMAX. (The MLP performs even worse if we omit the extra holiday and temperature regressors.) This may be due to the low dimension of our problem and/or the relatively small sample size. Usually such flexible machine learning method needs require large datasets for training. Compared to the number of parameters that need to be estimated, our sample size is still relatively small.

h	mean cross validation MSE (standard deviation)
2	247.1 (12.68)
3	278.5 (21.68)
4	279.0 (27.27)
5	284.1 (23.30)
6	277.5 (29.12)
7	275.0 (22.44)
8	272.4 (24.91)
9	272.4 (14.44)
10	269.2 (17.19)

Table 7: Validation loss for the single hidden layer MLP with different numbers of hidden neurons.

4.6. Summary for All 5 Methods

In this section we summarize the results for the five methods we considered in this section. Table 8 reports the training MSE and the test MSE for these five methods as well as for our original Gaussian model. Table 8 shows that the SARIMAX model outperforms the others, while the regression with calendar and weather variables is second best. Interestingly we see that the dynamic prediction model, SARIMA model and the MLP evidently do not perform better than our original Gaussian model with a single day-of-week factor. This indicates that the model we proposed in [6] actually does capture the main feature of the arrival process.

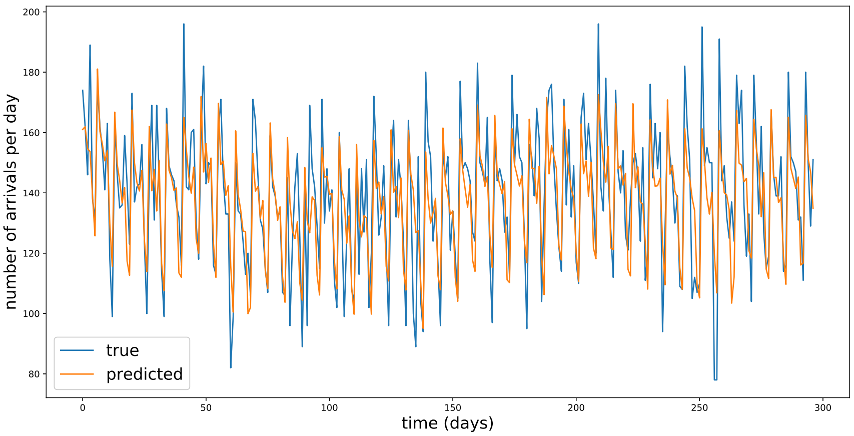


Figure 20: A comparison of the daily arrival totals in the test set to the predicted numbers using the MLP model in §4.5.

However, upon further study, we also discovered that the good result for our original model is not robust. When we looked at the larger dataset before we noticed and disregarded the abnormal Lebanon war period, we split our dataset from the beginning of 2007 (i.e. we used 01/01/2004 to 12/31/2006 as training set and 01/01/2007 to 10/31/2007 as test set). Then our original model performs badly with a test MSE larger than 300, while the dynamic prediction model and the SARIMAX model keep their test error level at 262.94 and 186.50 respectively. These are shown in the final column of Table 8.

In conclusion, for forecasting the daily arrival totals one day ahead, we find that the SARIMAX model is best, with about 25% improvement in test MSE. Since the mean absolute percentage error (MAPE) is often reported, e.g., as in [7]), we also computed the test MAPE for the SARIMAX, which is 8.4%.

4.7. Forecasting More Than One Day Ahead

A natural question is whether we can accurately forecast the daily arrival totals for 2 days, 3 days or even weeks ahead using our method. First, we remark that according to our models, the prediction results will stay the same for the original model and, as long as we are predicting less than one

Method	Training MSE	Test MSE	Test MSE*
Original	248.9	264.6	300.3
Dynamic	248.3	269.9	262.9
SARIMA	221.5	263.4	-
Regression with calendar and weather var.	206.0	234.3	-
SARIMAX	181.6	191.8	186.5
MLP	205.7	265.8	-

Table 8: Summary of the training and test results for the five methods to predict the daily arrival totals. Test MSE* is the test error when we train and test the method with a different splitting point.

week ahead, the dynamic model, because we simply use the average daily arrival totals on the same day of week in history as our prediction. However, we expect the MSE to increase if we use the SARIMAX model to predict more days ahead and, indeed, Table 9 shows that is the case. Table 9 shows the test MSE if we use our SARIMAX model to predict 1 to 7 days ahead. The MSE for 1 day ahead is a little different from Table 8 because we throw the last week away in order to predict 7 days ahead. We see that the MSE indeed increases as expected, but is not dramatically bad.

However, we need to point out again that there is a tricky flaw in the result. Because we only have historical actual weather record (not weather forecast, which we should use when make predictions), the input of weather data is better than would have in practice. This should not be a big issue if we consider 1 or 2 days ahead, but over longer time intervals the weather prediction is likely to degrade significantly. Hence, the actual prediction capability of SARIMAX for many days ahead should not be as good as reported here.

# of days ahead	1	2	3	4	5	6	7
Test MSE	193.2	197.1	198.1	202.7	204.2	208.0	211.0

Table 9: Test MSE for predicting several days ahead using the SARIMAX model (13) in §4.4.

5. Real-Time Predictor for the Occupancy Level

In this section we propose a real-time procedure for predicting the hourly occupancy level, exploiting the predictions for daily arrival totals in §4. We start by predicting the occupancy level for 1 hour ahead. Afterward, we show that the method can be extended to more hours ahead, but losing predictive power as the time interval increases. For this occupancy prediction, we exploit the elapsed service times of the patients initially in the system. Thus, an important referrence point is the average LoS, which is about $4 - 4.5$ hours. Clearly the value of information about current patients will disipatee as the time interval increases to 4 hours and beyond.

5.1. Predicting Occupancy Level 1 Hour Ahead

We use the same discrete time framework as we used in [6, 27, 28] since we assume both the arrival rate and the LoS distribution are fixed within an hour. Let $Y_{k,j}$ denote the number of patients that arrived at k^{th} discrete time period (in our case one time period represents an hour) and had length-of-stay (LoS) larger or equal to j hours, i.e. they left at or after time period $k + j$. We assume all the arrivals occurred at the beginning of each time period and departures occurred at the end of the DTP and we count the number of patients (occupancy level) in the middle of each DTP. (See [27] and [28] for discussion on such counting assumption.) Under this rule, we have the occupancy level at time period k to be

$$Q_k = \sum_{j=0}^{\infty} Y_{k-j,j}. \tag{15}$$

Our goal is to approximate Q_k, assuming we are given the all the arrival and departure epochs for each patients occured before time period k.

We built a stochastic model for the patients flow in [6] and as reviewed in §2, which well described the arrival process and the time-varying distributions of LoS. Given the daily arrival totals, hourly arrival rate curve and the LoS distributions for each hour, we can easily calculate the occupancy level of the system. But how can we predict the occupancy level for next hour given all the history up to now? We propose a real-time predictor as follow.

To predict Q_k, first we observed that we can use finite summation to approximate the infinite sum, because in reality we can always view the LoS distributions as bounded. Actually only 3.60% of all the patients had

LoS greater or equal to 13 hours. Hence we divided Q_k into two parts and estimated them separately, i.e. we wrote

$$Q_k = \sum_{j=0}^{12} Y_{k-j,j} + \sum_{j=13}^{\infty} Y_{k-j,j} \equiv \sum_{j=0}^{12} Y_{k-j,j} + R_{k,13}. \tag{16}$$

Note that $Y_{k,0}$ is the total number of arrivals at time period k. Since we assume the arrival process can be modeled as a NHPP within a day given the daily arrival totals (model M_2), given that we have already got the predicted daily arrival totals in Section 4, we only need to estimate the arrival rate function within a day. We use the empirical hourly arrival rate curve by combining the arrivals on the same day-of-week of the latest 10 weeks as our estimated arrival rate for a day because as we shown in §3, the LoS distributions also change slowly.

To conveniently express our estimator, we re-index our time period from k to a three-element tuple (w, d, h), where $w \in \{0, 1, 2, \cdots\}$ represents the week index, $d \in \{0, 1, 2, 3, 4, 5, 6\}$ is the day-of-week index and $h \in \{0, 1, 2, \cdots, 23\}$ is the hour index. The indices k have a one-to-one mapping to the tuples (w, d, h), therefore we will use them exchangeably in the following part of the paper. We also define that $(w, d, h) \pm x$ means we add/minus x time periods (hours) to time period (w, d, h). Denote $\hat{A}_{(w,d)}$ to be the predicted daily arrival totals of week w, day-of-week d, and $A_{(w,d)} \equiv \sum_{h=0}^{23} Y_{(w,d,h),0}$ to be the true daily arrival totals, then we construct the estimator for $Y_{(w,d,h),0}$ as

$$\hat{Y}_{(w,d,h),0} \equiv \hat{A}_{(w,d)} * \frac{\hat{\lambda}_{(w,d,h)}}{\sum_{h=0}^{23} \hat{\lambda}_{(w,d,h)}} = \hat{A}_{(w,d)} * \frac{\sum_{i=1}^{10} Y_{(w-i,d,h),0}}{\sum_{i=1}^{10} A_{(w-i,d)}}, \tag{17}$$

where $\hat{\lambda}_{(w,d,h)} \equiv \frac{1}{10} \sum_{i=1}^{10} Y_{(w-i,d,h),0}$ is the estimated hourly arrival rate for time period (w, d, h).

To estimate $Y_{k-j,j}$ for $j = 1, 2, \cdots, 12$, note that we can observe $Y_{k-j,j-1}$ for $j = 1, 2, \cdots, 12$ at time period $k - 1$, i.e. the number of patients that arrived at time period $k - j$ and still stayed in the system at time period $k-1$. Let $\mathcal{F}_k$ be the information filtration, i.e. $\mathcal{F}_k$ denotes all the observable arrival and LoS information up to time k. Since in M_3 we model the LoSs as i.i.d. random variables given the arrival time, so the conditional expectation of $Y_{k-j,j}$ equals to $Y_{k-j,j-1}$ times the corresponding survival probability of each patient:

$$\mathbb{E}(Y_{k-j,j}|\mathcal{F}_{k-1}) = Y_{k-j,j-1} * P_{k-j}(W \geq j | W \geq j-1) \equiv Y_{k-j,j-1} * p_{k-j,j-1}, \tag{18}$$

where W is a random variable following the LoS distribution of customers that arrived at time period $k-j$, and $p_{k-j,j-1}$ is the probability of a patient that arrived at time period $k-j$ and did not leave the system up to time period $(k-j)+(j-1)=k-1$ will still be there at time period k. Again, we use the latest 10 weeks history data to estimate that probability. We estimated $p_{(w,d,h),j}$ by

$$\hat{p}_{(w,d,h),j} \equiv \frac{\sum_{i=1}^{10} Y_{(w-i,d,h),j+1}}{\sum_{i=1}^{10} Y_{(w-i,d,h),j}}, \quad j=0,1,\cdots,11. \tag{19}$$

Then we have the estimator for $Y_{(w,d,h)-j,j}$ as

$$\hat{Y}_{(w,d,h)-j,j} = Y_{(w,d,h)-j,j-1} * \hat{p}_{(w,d,h)-j,j-1}, \quad j=1,2,\cdots,12. \tag{20}$$

Finally, we need to estimate $R_{k,13}$, the number of patients at time period k that had already been in the system for greater or equal to 13 hours. Instead of estimating it directly, we actually estimate $r_{k,13} \equiv R_{k,13}/Q_k = 1-(\sum_{j=0}^{12} Y_{k-j,j})/Q_k$ by the history data of the latest 10 weeks. Under the other indexing way, this can be writen as

$$\hat{r}_{(w,d,h),13} \equiv 1 - \frac{\sum_{i=1}^{10}\sum_{j=0}^{12} Y_{(w-i,d,h)-j,j}}{\sum_{i=1}^{10} Q_{(w-i,d,h)}}. \tag{21}$$

The reason we did like this is that we only used $Y_{i,j}$ and Q_i for $i \le k$ and $j=0,1,\cdots,\min\{12,k-i\}$, which means we only need to keep a finite dimensional record of Y matrix, other than a infinite dimensional one, which is impractical. Of course the number we chose here (12) is somewhat arbitrary, we could also keep record of $Y_{k,j}$ for j up to 11 or 13, we want to point out that if we use a small one, we will waste some information we have, on the other hand, if we used a large one, since patients with very long LoS are rare, the estimation for $p_{k,j}$ will be inaccurate for large j's.

Combining the estimation for each part in (17), (19), (20) and (21), we get the estimator for $Q_{(w,d,h)}$ as

$$\hat{Q}_{(w,d,h)} \equiv \frac{\sum_{j=0}^{12} \hat{Y}_{(w,d,h)-j,j}}{1-\hat{r}_{(w,d,h),13}}. \tag{22}$$

We use the prediction results of daily arrival totals by SARIMAX$(6,1,0)(0,0,2)_7$ as we introduced in §4.4 for $\hat{A}_{(w,d)}$ in (17). Since we need 10 weeks data to estimate the arrival rates and the LoS distributions before we start predicting the occupancy level, we make the hourly occupancy prediction from 12 p.m. Oct.25, 2006, which is 10 weeks after the start of test set, to 11 p.m. Oct.31, 2007, i.e. 8928 hours or equivalently 372 days.

The MSE of 1-hour-ahead real-time occupancy prediction is 14.65 (MAPE 10.59). As a comparison, we compare it to two other simple prediction methoeds applied to the same test period. The first one is we directly use the current hour occypancy level as the prediction for the next hour's occupancy level (i.e. $\hat{Q}_{(w,d,h)} = Q_{(w,d,h-1)}$), and the MSE of this method is 23.04. The second way is we use the average of 10 weeks' observed occupancy levels at the same hour in a week before the one we want to forecast as predictor (i.e. $\hat{Q}_{(w,d,h)} = 1/10 \sum_{i=1}^{10} Q_{(w-i,d,h)}$), which is like what we do for the daily arrival totals in §4.1. In this way the MSE is 51.91. So we make a 30% improvement to the first naive prediction method and even better to the second. Figure 21 plots part of the prediction compared to the actual values. Figure 21 shows that the prediction curve is quite close to the true curve.

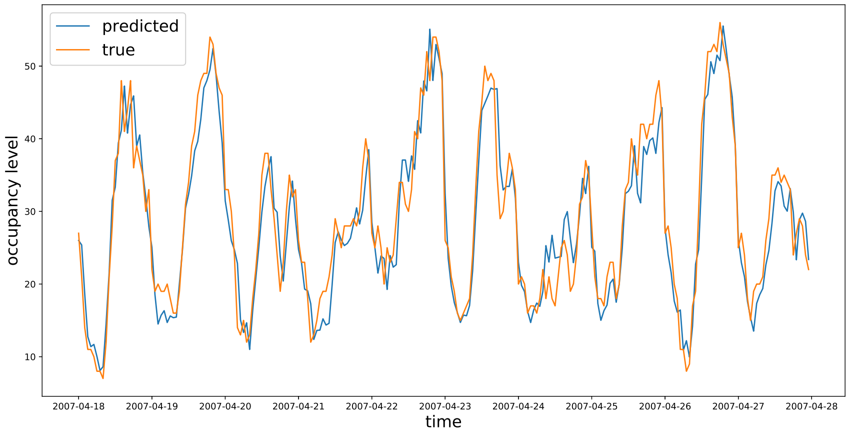


Figure 21: Predicted occupancy level one hour ahead compared to the true occupancy level.

5.2. Predicting Occupancy Level for More Than 1 Hour Ahead

The construction of real-time occupancy predictor for 1 hour ahead in the last section can easily be generalized to estimate the occupancy level for

2, 3 or even more hours in the future. But the farther we predict into the future, we encounter two problems: (i) the relevance of the elapsed times of current patients will decrease and (ii) the number of estimators we must use will increase. For both reasons, we expect that the error will increase. We take predicting 2 hours ahead as an example to show how to construct the predictor; others can be done in the same way.

Recall that

$$Q_{(w,d,h)} = \sum_{j=0}^{12} Y_{(w,d,h)-j,j} + \sum_{j=13}^{\infty} Y_{(w,d,h)-j,j} \equiv \sum_{j=0}^{12} Y_{(w,d,h)-j,j} + R_{(w,d,h),13}$$

as in (16). Since we are predicting 2 hours ahead, we assume that we are now at time $(w,d,h)-2$. The estimator for $Y_{(w,d,h)-0,0}$ is the same as in (17). However, we can no longer apply (20) to estimate $Y_{(w,d,h)-j,j}$ because $Y_{(w,d,h)-j,j-1}$ is not available to us as we assume we are now at $(w,d,h)-2$. Analogous to (18), for $j \geq 2$, we have

$$\begin{aligned} \mathbb{E}(Y_{(w,d,h)-j,j}|\mathcal{F}_{(w,d,h)-2}) &= Y_{(w,d,h)-j,j-2} * P_{(w,d,h)-j}(W \geq j | W \geq j-2) \\ &\equiv Y_{(w,d,h)-j,j-2} * p_{(w,d,h)-j,j-2,2}, \end{aligned} \tag{23}$$

where we use $p_{(w,d,h),j,l}$ to denote the probability that a patient arrived at time period (w,d,h) will stay for at least another l hours given he/she has been in the system for j hours. Similar to (19), we estimate $p_{(w,d,h),j,l}$ by

$$\hat{p}_{(w,d,h),j,l} \equiv \frac{\sum_{i=1}^{10} Y_{(w-i,d,h),j+l}}{\sum_{i=1}^{10} Y_{(w-i,d,h),j}}. \tag{24}$$

So for $Y_{(w,d,h)-j,j}$, $j = 2, 3, \cdots, 12$, we use

$$\hat{Y}_{(w,d,h)-j,j} = Y_{(w,d,h)-j,j-2} * \hat{p}_{k-j,j-2,2}, \tag{25}$$

as the estimator. For $Y_{(w,d,h)-1,1}$, since it is in the future, we use the predicted total arrivals times the corresponding survival probability, i.e. we use (19) for $j = 1$ and replace $Y_{(w,d,h)-1,0}$ on the right hand side by $\hat{Y}_{(w,d,h)-1,0}$. Finally, we use the same equation (22) to get the estimator for $Q_{(w,h,d)}$.

When we making predictions for more than 1 hour ahead, we need to use 2 days ahead predictions for daily arrival totals. This can be seen from equation (17). When we predict the hourly total arrivals, we need the forecast of daily arrival totals for that day. However, we can make 1 day ahead daily volume prediction only after 00:00 on that day. For example, assume that we are now at 23:00 on day k and we want to predict the occupancy level for 00:00-01:00 on day $k+1$. According to our real-time prediction procedure, we need to know both $\hat{A}_k$ and $\hat{A}_{k+1}$. Note that here $\hat{A}_k$ is 1 day ahead daily volume prediction while $\hat{A}_{k+1}$ must be 2 days ahead daily volume prediction because we have not observed the arrivals in 23:00-24:00 on day k so that we don't know A_k yet and the last observed daily total volume is A_{k-2}.

Table 10 shows the MSE for predicting the occupancy level more than 1 hour ahead for up to 6 hours. For comparison, Table 11 shows the corresponding MSE by using rolling history occupancy average using n weeks history.

# of hours ahead	1	2	3	4	5	6
MSE	14.65	25.21	35.05	66.33	113.59	160.16

Table 10: MSE for predicting the occupancy several hours ahead.

n	1	2	3	4	5
MSE	82.24	64.29	57.43	55.74	54.92
n	6	7	8	9	10
MSE	53.92	53.10	52.78	52.52	51.91

Table 11: MSE for predicting the occupancy by the rolling averages of n weeks.

Table 10 shows that beyond 4 hours ahead, the error is larger than the MSE 51.91 obtained by using rolling history occupancy average using $n = 10$ weeks history. For another comparison, the MSE is only 58.83 if we predict the occupancy level by the observed occupancy level two hour ago. Hence, we conclude that our real-time occupancy predictor outperforms the rolling average predictor for forecasting from 1 to 3 hours in the future, but not longer. That is roughly what we expect, given that the mean LoS for all patients is about 4 hours. Beyond that time interval, the current state information will not provide much information. At the same time, we have to making too much estimations in the real-time estimation procedure, which will make it perform worse than the rolling average estimator, which only removes noise but does not use the recent system state.

6. Conclusions

In this paper we investigated forecasting methods for the daily total arrivals and their application to predict the hourly occupancy level based on the framework we proposed [6] and its refinement. For that purpose, we exploited much more data, which enabled us to detect both (i) a long-term trend in both the arrival process and the LoS distributions and (ii) dependence in the daily arrival totals. For daily arrival totals, in §4 we studied five prediction methods, including rolling averages, highly structured time series models and a neural network model. We found that the SARIMAX time series model exploiting both exogenous variables (temperature and holiday effects) and internal dependence has the best predicting power.

In §5 we also proposed real-time predictors for hourly occupancy levels, which take account of the current system state. We found that our new method is superior to the rolling-average prediction for forecasting occupancy level in the near future (≤ 3 hours). We think that these occupancy predictors have great potential to help improve operational decisions.

There are many opportunities for future research. One direction is to study the advantage of systematically including additional information. For our current study, we only used the arrival and departure epochs of the patients. The dataset of the ED itself does not contain much more beyond that. Only gender, admission decision, age, the hospitalization duration (if any) and total number of departments visited by the patient are available. For predicting future arrivals, it is possible that some of the arrivals to the ED are actually known in advance, because they are actually scheduled.

For predicting future occupancy levels, there will likely be much more relevant information available in the future. It may be possible to know the patient medical problem, and the patient status (severity). Information about the the internal hospital wards also could be relevant. Even without extra information, further progress may be possible in the context of our study. Because the data we use is readily available, others may be able to develop improvements.

Acknowledgement. We thank Avishai Mandelbaum, Galit Yom-Tov and their colleagues at the Technion IE&M Laboratory for Service Enterprise Engineering (SEELab) for providing access to the Israeli Rambam hospital data. We thank the United States National Science Foundation for research support through CMMI 1634133.

References

[1] M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, G. Yom-Tov, On patient flow in hospitals: A data-based queueing-science perspective, Stochastic Systems 5 (1) (2015) 146–194.

[2] A. M. De Bruin, A. Van Rossum, M. Visser, G. Koole, Modeling the emergency cardiac in-patient flow: an application of queuing theory, Health Care Management Science 10 (2) (2007) 125–137.

[3] M. A. Ahmed, T. M. Alkhamis, Simulation optimization for an emergency department healthcare unit in Kuwait, European Journal of Operational Research 198 (3) (2009) 936–942.

[4] A. Kolker, Process modeling of emergency department patient flow: Effect of patient length of stay on ED diversion, Journal of Medical Systems 32 (5) (2008) 389–401.

[5] D. J. Medeiros, E. Swenson, C. DeFlitch, Improving patient flow in a hospital emergency department, in: Proceedings of the 40th Conference on Winter Simulation, Winter Simulation Conference, 2008, pp. 1526–1531.

[6] W. Whitt, X. Zhang, A data-driven model of an emergency department, Operations Research for Health Care 12 (1) (2017) 1–15.

[7] S. S. Jones, A. Thomas, R. S. Evans, S. J. Welch, P. J. Haug, G. L. Snow, Forecasting daily patient volumes in the emergency department, Academic Emergency Medicine 15 (2) (2008) 159–170.

[8] D. Tandberg, C. Qualls, Time series forecasts of emergency department patient volume, length of stay, and acuity, Annals of Emergency Medicine 23 (2) (1994) 299–306.

[9] S. A. Jones, M. P. Joy, J. Pearson, Forecasting demand of emergency care, Health Care Management Science 5 (4) (2002) 297–305.

[10] D. R. Holleman, R. L. Bowling, C. Gathy, Predicting daily visits to a waik-in clinic and emergency department using calendar and weather data, Journal of General Internal Medicine 11 (4) (1996) 237–239.

[11] A. K. Diehl, M. D. Morris, S. A. Mannis, Use of calendar and weather data to predict walk-in attendance., Southern Medical Journal 74 (6) (1981) 709–712.

[12] L. M. Zibners, B. K. Bonsu, J. R. Hayes, D. M. Cohen, Local weather effects on emergency department visits: a time series and regression analysis, Pediatric Emergency Care 22 (2) (2006) 104–106.

[13] R. Ibrahim, P. L'Ecuyer, Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models, Manufacturing & Service Operations Management 15 (1) (2013) 72–85.

[14] R. Ibrahim, H. Ye, P. L'Ecuyer, H. Shen, Modeling and forecasting call center arrivals: A literature survey and a case study, International Journal of Forecasting 32 (3) (2016) 865–874.

[15] W. Whitt, Predicting Queueing Delays, Management Science 45 (6) (1999) 870–888.

[16] R. Ibrahim, W. Whitt, Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals, Production and Operations Management 20 (5) (2011) 654–667.

[17] R. Ibrahim, W. Whitt, Wait-time predictors for customer service systems with time-varying demand and capacity, Operations research 59 (5) (2011) 1106–1118.

[18] S. Kim, W. Whitt, Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes?, Manufacturing and Service Oper. Management 16 (3) (2014) 464–480.

[19] J. Schwartz, D. Slater, T. V. Larson, W. E. Pierson, J. O. Koenig, Particulate air pollution and hospital emergency room visits for asthma in Seattle, Amer. Rev. Respiratory Dis. 147 (1993) 826–831.

[20] M. L. McCarthy, S. L. Zeger, R. Ding, D. Aronsky, N. R. Hoot, G. D. Kelen, The challenge of predicting demand for emergency department services, Academic Emergency Medicine 15 (4) (2008) 337–346.

[21] P. J. Brockwell, R. A. Davis, Introduction to Time Series and Forecasting, Springer, 2016.

[22] M. H. Kutner, C. Nachtsheim, J. Neter, Applied Linear Regression Models, 4th Edition, McGraw-Hill/Irwin, 2004.

[23] H. B. Mann, Nonparametric tests against trend, Econometrica: Journal of the Econometric Society (1945) 245–259.

[24] M. G. Kendall, Rank Correlation Methods, 5th Edition, Oxford University Press, 1990.

[25] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd Edition, Springer, New York, 2009.

[26] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.

[27] W. Whitt, X. Zhang, Periodic Little's law, to appear in Operations Research; available at Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html (2017).

[28] W. Whitt, X. Zhang, A central-limit-theorem version of the periodic Little's law, available at Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html (2018).