博士論文

Passenger Forecast and Staffing at Airport Immigration

(空港の入国審査場における到着客予測とスタッフ配置に関する研究)

ドワン　フン　フィー

Abstract

In the period after the war Tokyo was the central hub of intercontinental flights in Asia. Narita Airport has traditionally served as Tokyo's airport for international traffic. Over the last few decades Narita Airport has lost its position as Asia's main transfer hub. Its Asian competitors have increased their capacity while Narita's capacity has stagnated. In addition its Asian competitors have consistently won airport awards for the best service. Immigration at Narita Airport has received complaints from passengers about long waiting times during peak hours. Narita Airport has asked us to help to reduce the waiting times for foreign passengers. In this thesis we achieve this by setting the number of staff during the day such that the waiting time is at most 10 minutes for a certain percentage of the passengers.

We have developed three models: an arrival forecasting model, a queueing model and a staffing model. Based on the flight schedule and the number of passengers on each flight we first make a distributional forecast with the arrival probabilities at each time of the day. The arrival forecast is then used as input for the staffing model. To meet a certain service level requirement the staffing model determines the staffing function, i.e. the necessary staffing levels during the day. We can then simulate the performance of the staffing function with the queueing model.

Statistical models and discrete-event simulation models are commonly used for arrival forecasting. However statistical models require a large amount of historical data and simulation models generally require many iterations. We have developed a different approach to determine the arrival probabilities by using the sum of random variables and the convolution operation. We have collected data at Narita Airport to infer the probability distributions of the parameters used in the arrival forecasting model. In addition we have developed a Monte Carlo simulation model and a deterministic approximation. All three models give reasonable results when compared to the observed arrival rates of a single flight and multiple flights.

In the staffing literature a queueing system is often assumed to be in steady-state condition in each staffing interval. From observation data we have shown that the immigration queueing system is heavily overloaded during long periods of the day. Three queueing models have been implemented that can deal with overload: the numerical integration of ODE, the deter-

ministic fluid model and the stationary backorder-carryover approach. We have counted the passenger arrivals and the number of open service counters, and recorded the queues at Narita immigration on five occasions. The estimated waiting times by all three models are in close agreement with the observed waiting times. The deterministic fluid model is our preferred model because of the short computation times while still being accurate if a 1-minute time interval is used.

Staffing at airport immigration has been studied in the literature before but uncertainty in passenger arrivals due to flight delays was not taken into account. First we have assessed the service level performance of the deterministic staffing model. We found that the daily service level performance with uncertain demand is inadequate. Second we have extended the deterministic staffing model into a probabilistic model for which we have determined the appropriate quantiles to set the staffing levels. Third we have determined the appropriate parameter values for square-root staffing at immigration. And fourth we have developed an iterative algorithm to meet a service level requirement in each staffing interval.

Contents

Chapter 1

Introduction

1.1 Background

Airport Competition

In the period after the war Tokyo was the central hub of intercontinental flights in Asia. Tokyo's role as the main transfer hub has eroded continuously after the introduction of longer range airplanes in the 1970s and the liberalization of global air transport. Other Asian metropolises expanded their airports while the capacities of Tokyo's airports stagnated. These days the Asian airport system is a multiple hub system. The neighboring airports in East and South-East Asia have several competitive advantages: expandable operating areas, lower fees, professional airport management, extensive services and offensive competition strategies [20].

Narita Airport has traditionally served as Tokyo's hub for international traffic while Haneda Airport handled domestic traffic. This situation changed in 2010 when Haneda Airport opened a fourth runway as well as a third terminal dedicated to international flights. Narita Airport now faces competition both at the national and international level. Throughout the years Narita Airport's efforts to expand have been opposed by nearby residents. The consequences of the limited growth in flight slots can be seen in Figure 1.1. Incheon International Airport (ICN) surpassed Narita Airport (NRT) in terms of the number of international passengers in 2010. The airports in Hong Kong (HKG), Singapore (SIN), Bangkok (BKK) also serve more passengers than Narita Airport [83].

Not only has Narita Airport fallen behind in terms of international passenger volume, its competitors also rank higher in terms of overall service. In a global benchmark of airport excellence [71] Narita Airport placed 16th while its Asian competitors in Singapore, Seoul, Hong Kong and Beijing placed in the top five (Table 1.1). One of the services at Narita Airport that can to be improved is immigration. According to the news site Japan Today [44] ministry officials said that "many people arriving at Narita airport—

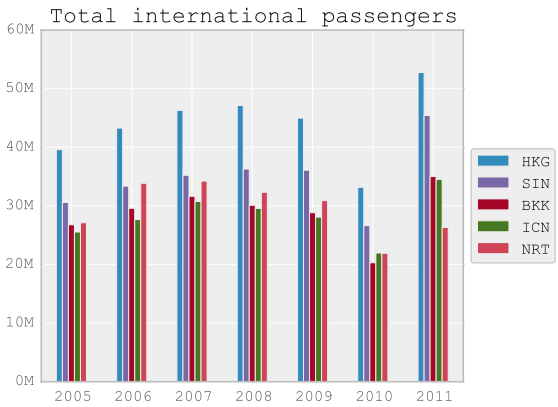


Figure 1.1: The top 5 Asian airports in terms of international passenger volume from 2005 to 2011.

Table 1.1: Ranking of the Skytrax World's Best Airports Awards.

	2012	2013	2014
1	Incheon	Singapore Changi	Singapore Changi
2	Singapore Changi	Incheon	Incheon
3	Hong Kong	Amsterdam Schiphol	Munich
4	Amsterdam Schiphol	Hong Kong	Hong Kong
5	Beijing Capital	Beijing Capital	Amsterdam Schiphol

Table 1.2: Ranking of the Skytrax Best Immigration Awards.

	2012	2013	2014
1	Incheon	Singapore Changi	Singapore Changi
2	Singapore Changi	Incheon	Incheon
3	Hong Kong	Amsterdam Schiphol	Munich
4	Amsterdam Schiphol	Hong Kong	Hong Kong
5	Beijing Capital	Beijing Capital	Amsterdam Schiphol

both foreign nationals and Japanese—have complained about the long lines at immigration, especially during peak times when several aircraft arrive one after the other." Narita Airport's competitors on the other hand have some of the best immigration services in the world (Table 1.2). Incheon Airport claims it has the world's fastest and most convenient immigration service with processing times that are more than three times faster than the international standard [41]. In order to improve its competitiveness, the decision makers at Narita Airport immigration asked us to help to reduce the waiting times.

Waiting Time

For service facilities it is not only the number of minutes in the waiting line that is important but also how the passenger experiences those waits. Maister [53] was the first to investigate the psychological aspect of waiting. He proposed a general law of service:

$$S = P - E \tag{1.1}$$

where S stands for service, P for perception and E for expectation. If a passenger perceives the received service higher than than his expected level, then he will be satisfied. Durrande-Moreau [18] reviewed 10 years of empirical research and concluded that there is a hierarchy among the factors that affect the customer experience. First the real waiting time and expectation are the main factors. Second, individual factors such as habit, motivation, mood and time pressure are decisive factors. And third, environmental factors (e.g. background music, information signs) have not been proven to be effective in altering the customers' perceptions. In this thesis the focus will be on improving the real waiting time.

The mathematical study of waiting lines identifies several elements in a queueing system that can affect the waiting time. First, a shorter service time will reduce the waiting time. At immigration the service procedure could be made faster by improving IT technology. Automated gates have been installed at Narita Airport immigration but the results of the experiment are unknown to us. The queue discipline, i.e. the order in which the

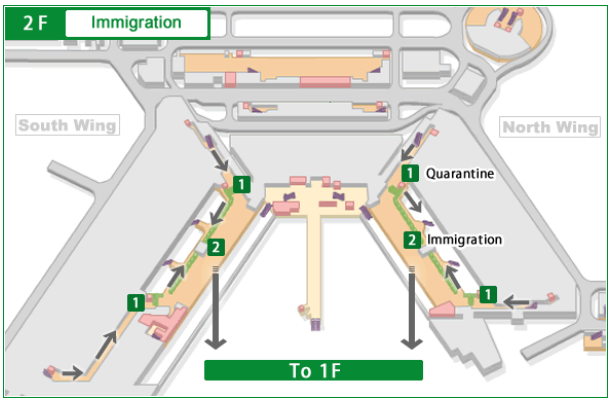


Figure 1.2: Immigration at the South wing and the North wing of Narita Terminal 1 [60].

passengers are served, also plays a role in the resulting waiting time. At Narita foreign passengers with a reentry permit have a higher priority than other foreign passengers. Another option would be to adjust the flight arrivals such that the arrival rate at immigration has lower peaks. However the flight schedule is not something that can be influenced by immigration management. An element that immigration managers can control is the number of staff at the service counters. We will explore methods to determine the required number of staff to meet a waiting time constraint.

Narita Immigration

Narita Airport has two terminals: Terminal 1 and Terminal 2. In this report we will only discuss the immigration service at Terminal 1. Terminal 1 has a satellite configuration that concentrates the gates at the end of the fingers. Terminal 1 is divided into a North wing and a South wing. The North wing is mainly occupied by the SkyTeam airline alliance while the Star Alliance (including ANA) is located in the South Wing. Each wing has its own immigration service (Figure 1.2). More flights arrive at the South Wing. Figure 1.3 shows the layout of the immigration area at the South Wing. There are two entrances. Passengers from gates 51 to 58 arrive at the left entrance while the passengers from gates 29B to 47 arrive at the right entrance. Gate 29B is a bus gate which means that the passengers are transported from the aircraft to the gate by bus.

The people arriving at immigration can be categorized into six types: airline crew, Japanese passengers, foreign passengers without a reentry permit, foreign passengers with a reentry permit and passengers that use the automated gates. In this report we will refer to the foreign passengers with a reentry permit as reentry passengers and those without a reentry permit as foreign passengers. These two groups combined will be called alien pas-

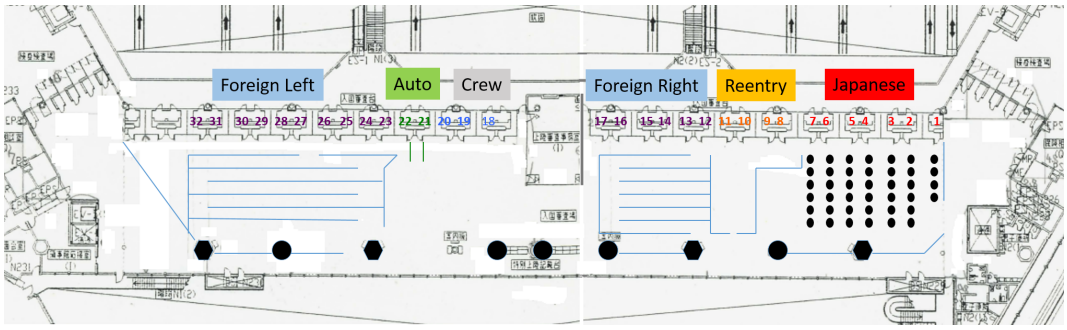


Figure 1.3: The immigration area at the South Wing of Terminal 1.

sengers. Figure 1.3 shows how the service counters are typically allocated for each type. The service counters allocation per passenger type changes dynamically during the day based on the flight schedule and the experience of the immigration managers. The figure shows just one possible scenario but the relative position of the service counters is always the same. On the right side, service counters 1 to 7 are often used for Japanese passengers. Service counters 8 to 11 are allocated to reentry passengers. Counters 12 to 17 are for all other foreign passengers. On the left side, counters 18 to 20 are always used for the airline crew. The automated gates 21 and 22 are fixed. Service counters 23 to 32 are also used for foreign passengers but these will be only opened if there is not enough capacity on the right side. In our study we will not consider the queues for the airline crew and the automated gates.

1.2 Research Objectives

To determine the required staffing levels we need to develop three models: an arrival forecasting model, a queueing model and a staffing model. Based on the flight schedule we first estimate the passenger arrivals at immigration. The number of arrivals is the input for the staffing model which provides the necessary number of staff such that the waiting time does not exceed 10 minutes for a certain percentage of the passengers. The waiting time performance can determined by a queueing model with the arrival rate and the number of staff as input.

The arrival forecasting model gives the arrival probabilities at each time of the day. In the call center literature various forecasting models have been proposed: a doubly stochastic Poisson model [79], time series methods [73], a mixed effects model [2], and a multiplicative model with Bayesian techniques [78]. These forecasting models however require large amounts of historical data that is not available in our case. In the airport literature discrete-event simulation models [64, 55] and a deterministic arrival model [52] have been

applied to airport immigration. We develop a faster method based on the sum of random variables and the convolution operation to determine the arrival probabilities.

Immigration is a collection of different queueing systems for foreign passengers, reentry passengers and Japanese passengers. In this thesis we only consider the immigration service for foreigners and we combine the foreign passengers with and without a reentry permit. Classical queueing theory assumes steady-state [32]. However at immigration the arrivals are non-stationary and the foreign queueing system is frequently overloaded. We compare three queueing models that are appropriate for overloaded systems: the numerical integration of differential equations, the deterministic fluid model and the stationary backlog-carryover approach. In the literature the performance of these models have not been compared. We have gathered data at Narita immigration on multiple days to validate the models with actual waiting times.

Staffing has been studied to a great extent in the context of call centers. Traditional staffing models assume a constant arrival rate and use the steady-state queueing theory. However assuming a constant arrival rate is unrealistic in most cases. For non-stationary arrivals we can use simple heuristics such as the square-root staffing formula or the pointwise stationary approximation [29]. In recent years the effect of uncertain demand on the staffing requirements has been studied [15] and solutions have been proposed with the newsvendor problem [17] and stochastic programming [6]. In the airport literature immigration staffing has been solved with a deterministic model [52] but delay uncertainty was not taken into account. In their discussion of a staffing model for airline services at an airport Green, Kolesar, and Whitt [29] suggest that "Future research should assess and address the uncertainty in demand." We extend the deterministic model with delay uncertainty by introducing staff probabilities. Furthermore we develop an iterative algorithm to meet a service level requirement in each staffing interval.

The objectives of this study can be summarized as follows: (1) develop a fast method to forecast the passenger arrivals under uncertain conditions, (2) determine an appropriate queueing model for the overloaded foreign queueing system, and (3) develop heuristics to quickly obtain the required staffing levels to meet a service level requirement.

The arrival forecasting model is discussed in Chapter 2. In Chapter 3 we investigate the appropriate queueing models for Narita immigration. The staffing model for uncertain demand is described in Chapter 4. The conclusions of our study and suggestions for further research are discussed in Chapter 5.

Chapter 2

Arrival Forecasting

In this chapter we will forecast the passenger arrivals at immigration. There are two types of forecasts: a point forecast that gives the expected arrival rate at each time, and a distributional forecast that gives the arrival probabilities at each time of the day. A point forecast can be obtained from the distributional forecast. Our objective is to determine the arrival probabilities based on flight arrival times and the number of passengers.

First we review the forecasting models that have been proposed for airport immigration arrivals and customer arrivals at call centers. Then we develop our arrival forecasting model for Narita immigration based on the sum of random variables. The model requires the probability distributions of the flight delay, the disembarkation delay, the disembarkation rate and the walking speed. We will describe the statistical properties of these parameters and how we gathered the data for these parameters at Narita immigration. In the last section we compare the distributional forecast and the point forecast with the observed arrival rates for a single flight and for multiple flights.

2.1 Literature Review

In this section we review arrival forecasting models that have been used for airport immigration and call centers.

2.1.1 Airport Arrivals

Airport models that forecast the passenger flow through the whole terminal have been developed since the 1970s. These models have been used for terminal design, airport planning and staff scheduling. An overview of these total airport terminal models are given by Tosic [76] and Wu and Mengersen [86]. In this section we will discuss studies that deal specifically with the passenger flow to immigration or customs.

Nikoue et al. [64] developed a passenger flow prediction model for immigration at Sydney International Airport. For their study they had access to three data sources. First, passenger tracking tools (DWELL) were placed throughout the terminals: 400 Wi-Fi access points, 130 people-counters and 50 Bluetooth sensors. Second, they had access to the Flight Information Display System (FIDS). The FIDS data contained information about the gate, the estimated and scheduled time of arrival. And third, they could use the data that was recorded at immigration by the Australian Department of Immigration and Multicultural and Indigeneous Affairs (DIMIA). The immigration dataset contained the time stamp, nationality, origin airport and flight number of each passenger. They did not have any information on the number of passengers on a flight, the changes in the estimated time of arrival, or the time of departure at the origin airport. Furthermore the Wi-Fi location data had low accuracy and a low frequency of collected signals. For example more than 80,000 devices were observed on a day but only 500 devices could be used to determine the flow to immigration. They used the DIMIA and FIDS data to generate a distribution of the number of passengers of each flight. The DWELL data was used to determine the walking time but because of the lack of useful tracking data they could not determine the walking time as a function of the congestion in the terminal. Instead they only determined the walking speed distribution for each gate.

Mason, Ryan, and Panton [56] developed a staffing model for customs personnel at Auckland International Airport in New Zealand. Their initial approach was to use historical flight loads in their simulations but the forecast was not sufficiently accurate for the model. Instead they consulted with the airlines to determine the flight loads and the scheduled flight times.

A staffing and scheduling model for immigration staff at Auckland International Airport was developed by Mason, Ryan, and Panton [55]. Prediction of the arrival flow to immigration of Auckland International Airport is more difficult that at other airports because passengers need to collect their luggage before going to immigration. This means that it is also necessary to model the baggage flow from the aircraft to the baggage claim, and to model the tendency of passengers in a group to wait until everyone has picked up their luggage. However they did not have to develop the immigration arrival model themselves because Auckland International Airport had already developed simulation models for the arrivals. The simulation model produced a deterministic arrival distribution with one-minute intervals.

Littler and Whitaker [52] also developed an algorithm for immigration staffing at a New Zealand airport terminal where the baggage claim was placed before the immigration service. Initially they developed a detailed event-based simulation model with stochastic parameters for the flight load and service processes. However they found it was difficult to use a simulation model for staffing with a service level. Instead they developed a simpler deterministic model. The stochastic variations in the arrivals was small so the

expected number of arrivals could be use for setting staffing requirements. In the forecast model the transport and service rates were deterministic and an approximation was used to estimate the proportion of passengers of a flight who finished the luggage pick-up at each time.

2.1.2 Call Center Arrivals

Ibrahim et al. [39] identified several key features of a call center arrival process. The arrival rate is non-stationary and varies with the time of the day. There can be considerable daily, weekly and yearly seasonal patterns. There can also be significant correlations between different periods of the same day and between successive days. And the variance is much larger than the mean instead of equal as for the Poisson distribution. They also distinguished between two types of models: single day models and multiple day models. Gans, Koole, and Mandelbaum [24] reviewed arrival forecast methods and argued that more accurate distributional forecasts are essential for highly utilized call centers.

Single Day Arrival Models

For single day arrival models, the day is divided in p time intervals. Let $X = \{X_1, \ldots, X_p\}$ be the vector with the number of arrivals in each interval. The random variable X_j has a Poisson distribution with mean Λ_j.

The uncertainty in the arrival rate can be captured with a doubly stochastic Poisson model [79]. Assume that the deterministic arrival rate function $\lambda(t)$ describes the variation of the arrivals over the day. Then we introduce a busyness factor B which is a random variable with mean $E[B] = 1$. Uncertainty is added to the arrival rate function as follows

$$\Lambda(t) = B\lambda(t). \tag{2.1}$$

Avramidis, Deslauriers, and L'Ecuyer [3] studied this model where the busyness factor has a gamma distribution with $Var[B] = 1/\gamma$. As a consequence vector X has a negative multinomial distribution with parameters $(\gamma, \lambda_i, \ldots, \lambda_k)$ where $\lambda_i = \int_{t_{(i-1)}}^{t_i} \lambda(t)dt$. For the negative multinomial distribution there exist equations for the maximum likelihood estimators of these parameters.

Channouf and L'Ecuyer [11] proposed a model where the dependence between time intervals is modeled via a normal copula. A copula is used to describe the correlation between two random variables. Compared to the model by Avramidis, Deslauriers, and L'Ecuyer it provided a better estimation of the correlations and variance of the arrival rates.

Multiple Day Arrival Models

In these models the arrival rates are modeled over several days or several months. Again we assume a Poisson process for the arrival rate during interval j but now we also consider the day of the week $d_i = \{1, \ldots, 7\}$ where 1 refers to Monday, 2 to Tuesday and so on [39]. $X_{i,j}$ is then the arrival count in interval j on day i and it has a Poisson distribution with mean $\Lambda_{i,j}$. Often a square-root transformation $Y_{i,j} = \sqrt{X_{i,j} + 1/4}$ is applied to stabilize the variance for Poisson data [9].

Taylor [73] compared five time series methods for forecasting intraday arrivals in half-hour intervals. The arrival rate functions that they investigated featured intraweek and intraday seasonal cycles. The methods include seasonal autoregressive integrated moving average (ARIMA), exponential smoothing method and dynamic harmonic regression. For short-term forecasts the exponential smoothing method was most accurate. For long-term forecasts (more that a week ahead) the best results were achieved with a simple method that averages past observations on the same day of the week.

Aldor-Noiman, Feigin, and Mandelbaum [2] proposed a mixed effects model where the transformed arrival counts is a linear function of fixed and random effects. Fixed effects are the day-of-the-week effect α_{d_i}, the interaction between days $\theta_{d_i,j}$, and the interval-of-the-day effect β_j. Random effects are the daily volume deviation from the fixed weekday effect D_i, and the noise or residual effects $\epsilon_{i,j}$. The model can be formulated as

$$Y_{i,j} = \alpha_{d_i} + \beta_j + \theta_{d_i,j} + D_i + \epsilon_{i,j}. \tag{2.2}$$

A multiplicative model with Bayesian techniques was proposed by Weinberg, Brown, and Stroud [78]. The model parameters are estimated with Markov chain Monte Carlo sampling methods. The advantage is that it can provide distributional forecasts but at the cost of long computational times. An additional advantage is that the experience of call center managers can be incorporated in the model.

The arrival models for call centers require large amounts of data to estimate the model parameters. Brown et al. [9] analyzed the data of 101 days of quarter-hourly arrival rates from a bank call center. Ibrahim et al. [39] collected data over 275 days with half-hour intervals. Aldor-Noiman, Feigin, and Mandelbaum [2] did a case study of an Israeli Telecom company call center with 150 agents. The data consisted of arrival counts from February 2004 to December 2004.

2.2 Immigration Arrival Model

The arrival models for immigration that were discussed in the literature review, produced point forecasts. However we are interested in a distributional forecast for Narita immigration. Also these models use simulation to

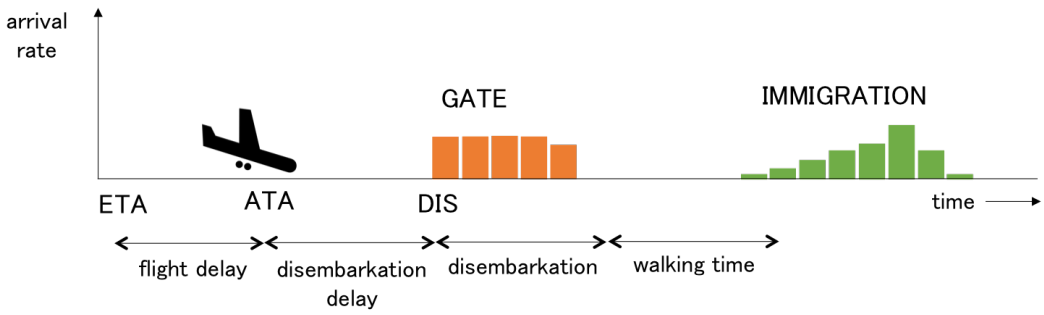


Figure 2.1: The arrival process of one flight.

estimate the arrival rate. In general simulation requires many iterations and can be relatively slow. The models for call centers require large amounts of counting data to estimate the model parameters. This amount of data is unavailable for our study. The only data of the passengers arrivals was obtained from the observations that we conducted for several afternoons. We also did not have access to the data of the immigration computer system, nor were there any passenger tracking tools available. Our objective is to develop a distributional arrival forecasting model without requiring large amounts of counting data and with relatively high computational speed.

There are two pieces of available information that we can use to forecast the arrivals at immigration: the number of passengers on a flight and the flight schedule. The immigration office receives the number of arriving passenger of each flight from the airlines on the day of arrival. If forecasting is done days or weeks in advance then the number of passengers needs to be estimated. The information about the expected time of arrival of the flights is continuously updated during the day. The expected time of arrival is only an estimation and the flight delay is an uncertain parameter. Other uncertainties in the arrival process from the gate to immigration are shown in Figure 2.1. The flight delay is the difference between the actual time of arrival (ATA) at the gate and the expected time of arrival (ETA). When the aircraft arrives at the gate, passengers need to wait inside the aircraft until the disembarkation procedure (DIS) starts. We will call this time inside the aircraft the disembarkation delay. Passengers exit the aircraft with a certain rate. The orange bars represent the number of passengers disembarking during a time interval. After disembarkation the passengers walk from the gate to immigration. The walking time or the walking speed is another uncertain parameter. The arrival rate at immigration is represented by the green bars. Ideally we would like to forecast the number of passengers arriving at each queueing system for foreign, reentry and Japanese passengers.

In this section we describe our method to estimate the probability of having k passengers arriving during an interval while including the flight

delay distribution, the disembarkation delay distribution and the walking time distribution.

Arrival Time Distribution Function

Suppose a flight has N passengers on board and the first passenger leaves the aircraft at time t_0. Let τ_i be the time that passenger i arrives at immigration relative to t_0 and let $F_i(t)$ be the distribution function of the arrival time of passenger i

$$F_i(t) = P(\text{arrival time of passenger } i \leq t) = P(\tau_i \leq t). \tag{2.3}$$

We assume that the arrival times of the passengers are statistically independent

$$P(\tau_1 \leq t_1, \tau_2 \leq t_2, \dots) = F_1(t_1)F_2(t_2)\dots \tag{2.4}$$

Let $\lambda(t)$ be the arrival function then the total number of passengers $A(t)$ of the flight that have arrived at immigration by time t is

$$A(t) = \int_0^t \lambda(x)dx. \tag{2.5}$$

Newell [63] treated the case that the arrival times are i.i.d., in other words $F_A(t) = F_1(t) = \cdots = F_N(t)$. The probability that k passengers have arrived by time t is the probability of k successes in N trials [63]

$$P(A(t) = k) = \frac{N!}{(N-k)!k!}[F_A(t)]^k[1 - F_A(t)]^{N-k} \tag{2.6}$$

where success is defined as a passenger arriving by time t and failure if he arrives after t. The expected number of cumulative arrivals is

$$E[A(t)] = \sum_{k=0}^{N} kP(A(t) = k) = NF_A(t) \tag{2.7}$$

with variance

$$Var[A(t)] = \sum_{k=0}^{N}[k - NF_A(t)]^2 P(A(t) = k) = NF_A(t)[1 - F_A(t)]. \tag{2.8}$$

Sum of Random Variables

However we are interested in the arrival probability $P(\lambda(t) = k)$ instead of the cumulative arrival probability $P(A(t) = k)$. Also the arrival distribution function $F_i(t)$ is not identical for each passenger, therefore we cannot use equation (2.6). Let $F_W(t')$ be the walking time distribution function relative to the disembarkation time δ_i of passenger i. $F_W(t')$ can be assumed to be

i.i.d for all passengers of a flight. The arrival distribution function $F_i(t)$ is not identical because the disembarkation time of each passenger is different

$$F_i(t) = F_W(t - \delta_i) \tag{2.9}$$

where $t = \delta_i + t'$.

To determine the arrival probability $P(\lambda(t) = k)$ we divide the time period into T intervals and let period j correspond to the interval $(t_{j-1}, t_j]$. Let p_{ij} be the probability that passenger i arrives at immigration during interval j. For clarity we drop the subscript j and write $p_i = p_{ij}$. The arrival probability is

$$p_i = P(t_{j-1} < \tau_i \leq t_j) = F_i(t_j) - F_i(t_{j-1}). \tag{2.10}$$

Let the discrete random variable X_i describe the outcome of passenger i arriving or not arriving at immigration during interval j. If passenger i arrives then X_i takes the value 1 and 0 if he does not arrive during the interval. The distribution function $m_i(x)$ for X_i is

$$m_i(x) = \begin{pmatrix} 0 & 1 \\ 1 - p_i & p_i \end{pmatrix}. \tag{2.11}$$

For passengers 1 and 2 the distribution functions of X_1 and X_2 are respectively $m_1(x)$ and $m_2(x)$. Let S_2 be the sum of the random variables X_1 and X_2. The distribution function $a_2(x)$ of S_2 is equal to the convolution of $m_1(x)$ and $m_2(x)$ given by [31]

$$a_2(x) = m_1 * m_2 = \sum_v m_1(v) \cdot m_2(x - v) \tag{2.12}$$

for $x = \ldots, -2, -1, 0, 1, 2, \ldots$. To determine the distribution of $S_3 = S_2 + X_3$ we convolve $a_2(x)$ with $m_3(x)$ of a third passenger. We continue this until we have convolved all N passengers of a flight. The result is the sum of N random variables $S_N = X_1 + X_2 + \cdots + X_N$ with distribution function $a_N(x)$. Let λ_j be the total number of passenger arrivals during interval j. The probability that k passengers arrive in the interval is then

$$P(\lambda_j = k) = a_N(k). \tag{2.13}$$

If we do the above procedure for all intervals $j = 1, \ldots, T$ then we get the arrival probability matrix Z

$$Z = \begin{array}{c} \begin{array}{cccccc} 1 & 2 & \ldots & j & \ldots & T \end{array} \\ \begin{pmatrix} & & & & & \\ & & \ddots & & & \\ & & & a_{Nj}(k) & & \\ & & & & \ddots & \\ & & & & & \end{pmatrix} \end{array} \begin{array}{c} 0 \\ 1 \\ \vdots \\ k \\ \vdots \\ N \end{array} \tag{2.14}$$

where $a_{Nj}(k)$ is the probability of having k arrivals during interval j. The expected number of arrivals in interval j is

$$E[\lambda_j] = \sum_{k=0}^{N} k a_{Nj}(k). \tag{2.15}$$

The arrival probability matrix Z represents the arrival probabilities for a flight without delay uncertainty.

Delay Uncertainty

Next we add the delay uncertainty to the arrivals. The delay uncertainty of a flight is described by the delay function $d(x)$

$$d(x) = \begin{pmatrix} \dots & -2 & -1 & 0 & 1 & 2 & \dots \\ \dots & p_{-2} & p_{-1} & p_0 & p_1 & p_2 & \dots \end{pmatrix} \tag{2.16}$$

where the upper row indicates the delay time in minutes and the bottom row indicates the probability for the delay time. There are two kinds of delays: the flight delay d_f and disembarkation delay d_d. We can determine the overall delay function with the convolution of d_f and d_d

$$d = d_f * d_d. \tag{2.17}$$

Because a realization of the delay is identical for all passengers of a flight we cannot add the delay uncertainty to the arrival distribution function $F_i(t)$ of the individual passengers. Instead we need to add the delay uncertainty to the arrival probability matrix Z of the flight. The arrival probability matrix Z_d with delay uncertainty can be computed by the convolution of d and each row of Z

$$Z_d(k) = Z(k) * d \tag{2.18}$$

where $Z(k)$ is the row of Z corresponding to k arrivals.

Multiple Flights

The last step is to combine the arrival probabilities of multiple flights. For the arrival probability matrix Z of a flight, the column at interval j represents the distribution function $a_{Nj}(x)$ of the random variable S_{Nj}. Let Z_d^1 and Z_d^2 be the arrival probability matrix of respectively flight 1 and 2. $\overline{Z}_d$ is the arrival probability matrix of both flights combined. For interval j we want to determine the distribution of the sum of the passengers of flight 1 and 2, i.e. $S_j = S_{Nj}^1 + S_{Nj}^2$. The distribution function $a_j(x)$ of S_j is the convolution of $a_{Nj}^1(x)$ and $a_{Nj}^2(x)$

$$a_j(x) = a_{Nj}^1 * a_{Nj}^2. \tag{2.19}$$

In other words we need to convolve the columns of Z_d^1 and Z_d^2 at the same interval j. For n flights the arrival probability matrix $\overline{Z}_d$ can be calculated with

$$\overline{Z}_d(j) = Z_d^1(j) * Z_d^2(j) * \cdots * Z_d^n(j) \tag{2.20}$$

where $Z_d(j)$ is the column at interval j.

2.3 Parameters Estimation

For the immigration arrival model we need to determine the following parameters: the flight delay distribution d_f, the disembarkation delay distribution d_d, the disembarkation time of each passenger δ_i, the walking time distribution function $F_W(t')$ and the number of passengers on a flight N. In this section we describe how the data for these parameters was collected at Narita immigration and how we derived the distributions from the data.

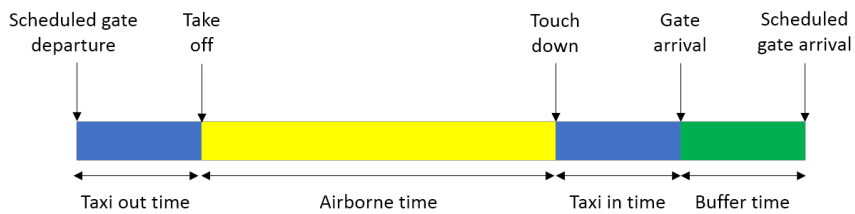
2.3.1 Flight Delay

Literature Review

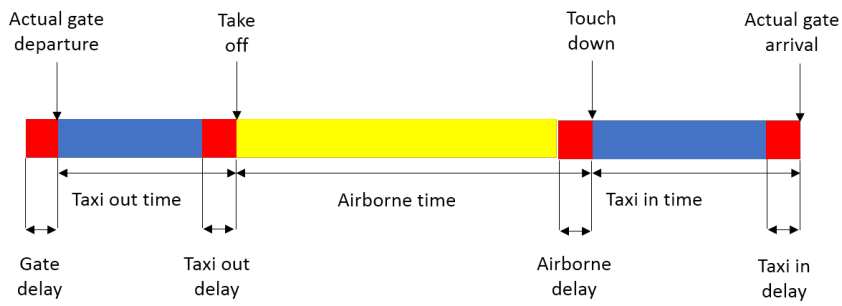
Flight delay is defined as the difference between the actual gate arrival time (ATA) and the scheduled gate arrival time (STA). The causes for delays can be derived from Figure 2.2 which shows the components of the scheduled and actual flight times. The scheduled flight time consists of the taxi-out time, the airborne time and the taxi-in time. In addition airlines add a buffer time to cope with potential delays. The actual flight time shows that there are four different delay components: the gate delay, the taxi-out delay, the airborne delay and the taxi-in delay.

Tu, Ball, and Jank [77] modeled the gate delay. There are many factors influencing the gate departure time. But instead of looking at each factor separately the factors were grouped into three categories: seasonal trend, daily propagation pattern and random residues. The seasonal trend includes seasonal demand change, weather impact, and other seasonal factors. The daily propagation pattern includes factors such as crew connection problems, delay built-up from previous flights and other daily propagation factors. The random residues contain mechanical problems, luggage problems, a late passenger and other random factors. A smoothing spline model was used to estimate the seasonal trend and the daily propagation pattern. To capture the residual delay distribution a finite mixture model was applied.

Idris et al. [40] used linear regression to identify the main factors that affect the taxi-out time. The most important factor was the take off queue size. Other relevant factors were the runway configuration, the location of the airline in the terminal, and the downstream restrictions (flow management programs to regulate flights to weather-impacted destinations).



(a) Scheduled flight time



(b) Actual flight time

Figure 2.2: Components of the scheduled and actual flight time. Adapted from Skaltsas [70].

Willemain [84] analyzed the variations around daily average airborne times between origin-destination pairs and the deviations from estimated times en route. The deviations were decomposed into four components: the regional airspace as a whole (day effect), the airspace at the departure airports (origin effects), the en route airspace (residuals), and the airspace at the arrival airports (destination effects). A simple model was used that described the deviation as the addition of the four components. Numerical estimates of the first three effects were obtained by a two-way analysis of variance (ANOVA) without interaction effects. The en route effects were the residuals after fitting the other three effects. Willemain et al. [85] investigated the influence of the origin airport, the destination airport, the month of year, day of week, hour of day, aircraft type and carrier on the flight's estimated time en route. The route, month, hour of day and carrier were found to be statistically significant influences.

Mueller and Chatterji [59] analyzed the arrival and departure delay characteristics at ten major U.S. hub airports. The contribution of the delay components to the overall delay was as follows: gate delay 50%, taxi out delay 26%, airborne delay 16%, and taxi in delay 8%. The delays were modeled by creating probability density functions. The Poisson distribution was the best fit for the departure delay while the airborne and arrival delays were best modeled with a normal distribution.

Skaltsas [70] investigated how U.S. carriers adjust the buffer time. Using linear regression models it was found that the flight distance and the time of day were the most important factors that affect the buffer time. The results also showed that the buffer time fluctuates greatly during the day.

Bai [4] analyzed the delays at Orlando International Airport using statistical models. Multivariate regression, ANOVA, neural networks were used to detect patterns of airport delay. Aircraft arrival delays were analyzed with logistic regression. The following factors were found to contribute to the aircraft delay: weather (precipitation), flight distance, season, weekday, arrival time and the time spacing between two successive arriving flights. There was a very high correlation between the delays at the airport of origin and the destination airport. Also an interaction effect was found between the flight distance and the time of day.

Narita Delay Data

To investigate the flight delays at Narita airport we used the online flight schedule from the Narita Airport website [12]. The online flight schedule contains the following information: scheduled time of arrival (STA), expected time of arrival (ETA), airline, flight code, departure city, stopover city, gate, status and total travel time. The online flight schedule is updated every 10 minutes. From November 1st 2013 until June 1st 2015 the flight schedule was downloaded every 10 minutes from 4 AM until 12 PM (no data was downloaded between January 19th and March 2nd 2014). This data allows us to analyse the flight delay and how the ETA changes during the day.

Delay with Scheduled Time of Arrival

The online flight schedule does not explicitly show the actual gate arrival time (ATA). We assume that the ATA is equal to the last ETA when the status of a flight changed to "ARRIVED". The objective is to estimate the delay probability distribution of a flight. From the literature it is clear that there are many factors that influence the actual flight delay, including time of day, day of week, month and flight distance. Figure 2.3 shows the 2D-histogram of the probability of delay for each month, day of week, hour of day and flight time. These plots were made with the data of all flights but the top and bottom 1% were discarded to remove outliers. The delay can range from -60 minutes to 137 minutes. The colors represent the probabilities; blue means low probability and red means high probability. The black line represents the expected delay. In general flights arrive earlier than the STA, i.e. the delay is negative, because of the buffer that airlines add to improve their on-time flight statistics [70]. The delay per month shows that on average aircraft arrive 12 minutes before STA in April and

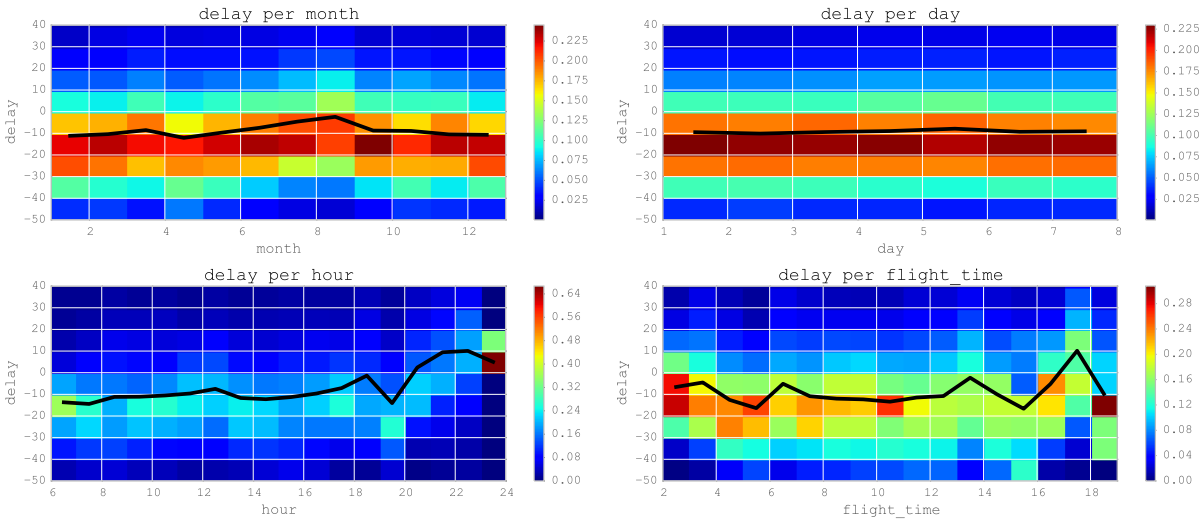


Figure 2.3: 2D-histogram of the delay probability per month, day of week, hour of day and flight time.

2 minutes before STA in August. If we look at the delay per day of week, neither the average nor the distribution of the delay differ much per day. Flights that arrive in the evening tend to have a positive expected delay while during the rest of the day the delay is expected to be negative. The expected delay for the flight time has a peculiar shape but it should be noted that the number of samples is only several hundreds for flight times of 15 hours and more, while other flight times have thousands of samples.

Figure 2.4 shows the histogram of the delays of four flights over the whole observation period. We see that the delay distribution is very different for each flight. Figure 2.5 shows the 2D-histograms for the delays per month and per day for flight 5J5054 which departs daily from Manila and arrives at Narita in the morning. For this flight there is clearly a difference in the delay distribution of each month. For the day factor we cannot detect a pattern. We can conclude that we should use the monthly delay distribution of each individual flight, and not apply the same delay distribution to all flights.

Delay with Expected Time of Arrival

Standard delay analysis deals with the difference between the ATA and the STA. When the staffing decisions are made in real-time, the most accurate prediction of the arrival time is the current ETA, not the STA. We can expect that as the time period between the ETA and the flight schedule update time t_{update} becomes shorter, the error of the ETA becomes smaller. Let $\delta = \text{ETA} - t_{\text{update}}$ be the time period from t_{update} until the ETA, and let $\epsilon = \text{ATA} - \text{ETA}$ be the error of the ETA. Figure 2.6 shows the ETA

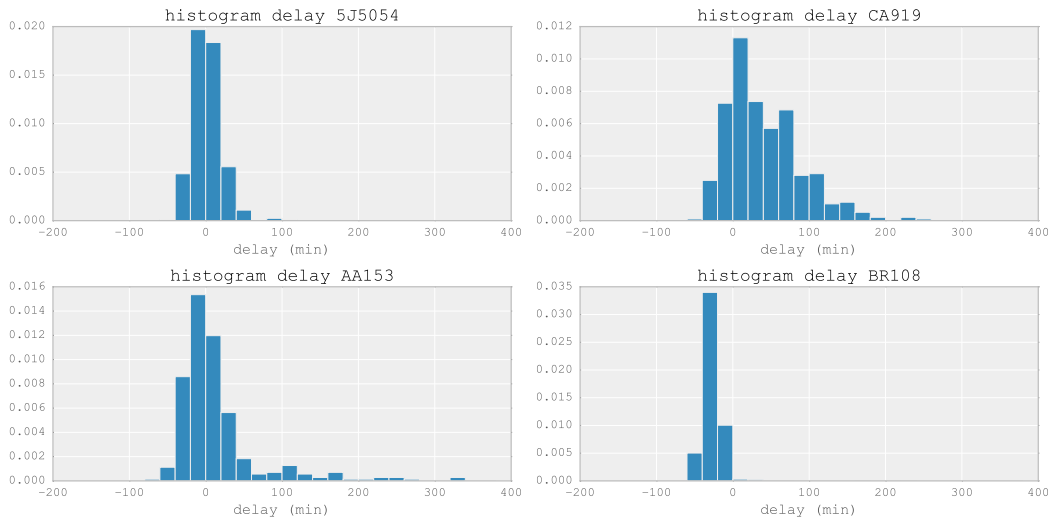


Figure 2.4: Delay histogram for four flights.

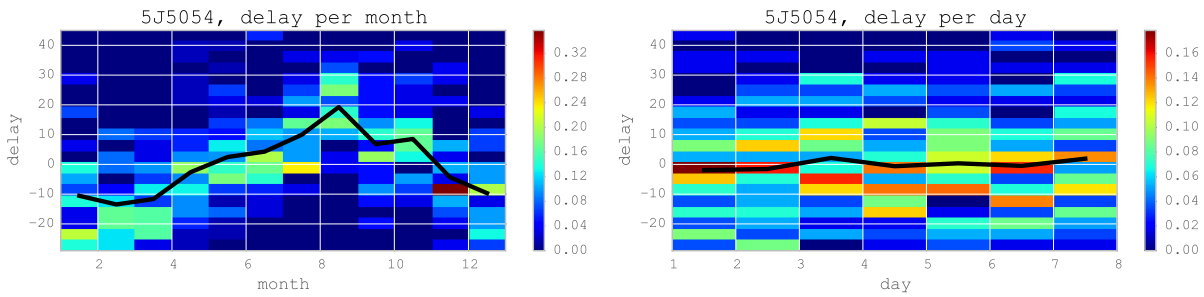


Figure 2.5: Example of the delay per month and per day of the week for one flight.

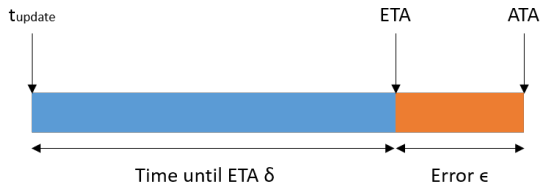


Figure 2.6: Error of the ETA relative to the flight schedule update time.

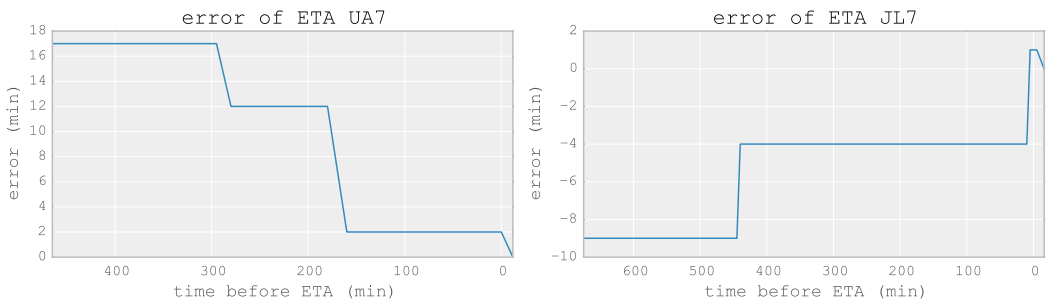


Figure 2.7: Error of the ETA for one flight.

error relative to the flight schedule update time. During the day the ETA is continuously updated every 10 minutes. Because we have downloaded the flight schedule for every 10 minute update we can analyze the changes of the ETA during the day. At the end of the day the ATA is known, from which the error distribution of the ETA can be determined. Figure 2.7 shows the error of the ETA as a function of δ for flights UA7 and JL7. Note that a negative error means that the actual arrival time of the flight turned out to be before the ETA. We see that as the time until the ETA becomes shorter, the absolute error becomes smaller.

We want to know the probability of the ETA error $P(\epsilon|\delta)$ for each flight at the flight schedule update time. From the flight schedule data that we have collected for over a year we can determine the error probability distributions for each δ. We combine the ETA error of all flights. Figure 2.8 shows the probability distributions of the ETA error for various δ. We can see that as δ becomes smaller, the variance also becomes smaller. It is also possible for δ to be negative, for example when the flight status was not properly updated after arrival. The red line represents the delay distribution with the STA. There is a large difference between the STA curve and the ETA curves for smaller δ.

2.3.2 Disembarkation

In this section we discuss how to estimate the time when each passenger leaves the aircraft. To do so the disembarkation delay and the disembarka-

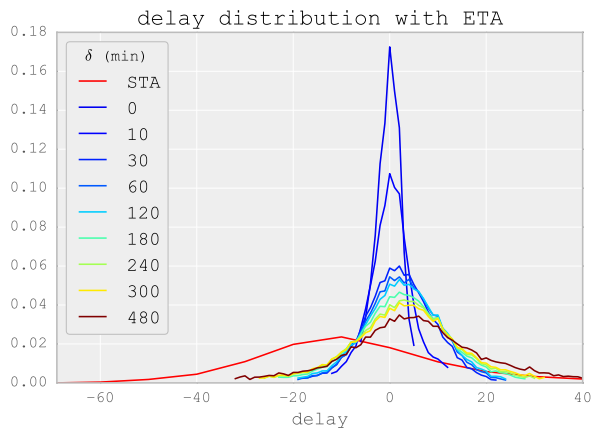


Figure 2.8: Delay probability distributions for various times before ETA.

tion rate need to be estimated. Since we could not get near the aircraft we observed the people from the moment they entered the gate instead. When we write "leaving the aircraft" it actually means passing through the gate door.

Literature Review

Barickman, Sebenius, and Sohi [5] analyzed different embarkation and disembarkation methods for the Airbus A380. They mentioned a statistic from Boeing that the exit time is 3 seconds per person per aircraft exit. Horstmeier and Haan [36] studied the handling times of turn round cycles to prepare an airport for the A380. They interviewed experts who were involved in these processes and found that the disembarkation rate fits a lognormal distribution with an average of 19 passengers per minute per bridge. We want to confirm that the same disembarkation rates apply to the passenger flow through the gate door and for various types of aircraft.

Disembarkation Delay

The disembarkation delay is the time difference between the ATA and the time that the first passenger leaves the aircraft. Some flights don't arrive at a gate. The passengers of these flights are transported by bus to a gate. The time of first disembarkation has not been observed for these flights. At Narita Airport we recorded the time of the first disembarking passenger for 41 flights. Figure 2.9 shows the normalized histogram of the observed disembarkation delay. The average disembarkation delay is 9.9 minutes. In the same figure a normal distribution is plotted with the mean and standard deviation of the observation data. As the number of samples is only 41 we want to investigate the distributions of the mean and the standard deviation

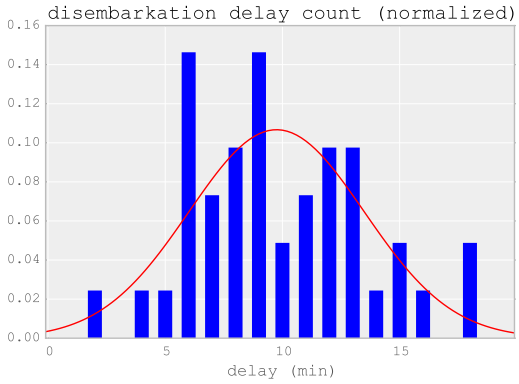


Figure 2.9: Disembarkation delay histogram.

with a Markov Chain Monte Carlo (MCMC) algorithm. The statistical model for the disembarkation delay is as follows

$$\begin{aligned} D &\sim \text{Normal}(\mu, \sigma) \\ \mu &\sim \text{Uniform}(l_\mu, u_\mu) \\ \sigma &\sim \text{Uniform}(l_\sigma, u_\sigma) \end{aligned} \tag{2.21}$$

where the disembarkation delay D is modelled as a Gaussian process, μ is the mean and σ is the standard deviation. Both μ and σ have a continuous uniform distribution bounded by a lower l and upper u limit parameter. In other words we set the prior distributions as a uniform distribution and then use the MCMC algorithm to find the posterior distribution of μ and σ by fitting the model to the data. The posterior distribution is found by generating random samples from the prior distribution. The newly generated samples are accepted or rejected based on a certain test and the current sample value. It is therefore a Markov Chain. As the number of samples increases, the distribution of the parameter converges to a stationary distribution. The posterior distribution for the mean and standard deviation are shown in Figure 2.10. The red lines are the average values of the posteriors. The average of the mean and the average of the standard deviation are respectively 9.9 and 3.7 minutes.

Disembarkation Rate

To estimate the rate of disembarkation we recorded the passengers passing through the gate door for 10 flights on video. Figure 2.11 shows the cumulative number of disembarking passengers for each flight. The disembarkation rate seems to be constant except for the tail. If only the first 90% of the passengers are considered then the average disembarkation rate is 38 passengers per minute (red line). An MCMC model was used to determine the

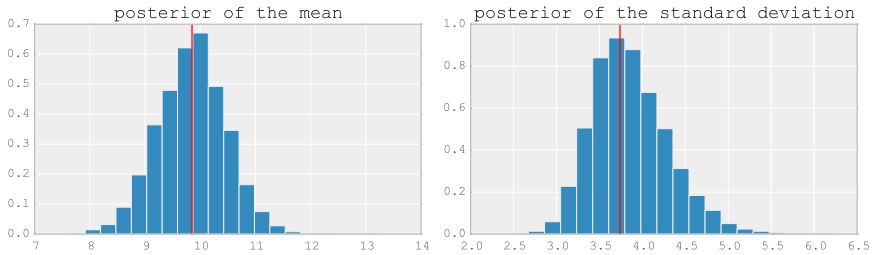


Figure 2.10: Posterior distribution of the mean and standard deviation of the disembarkation delay.

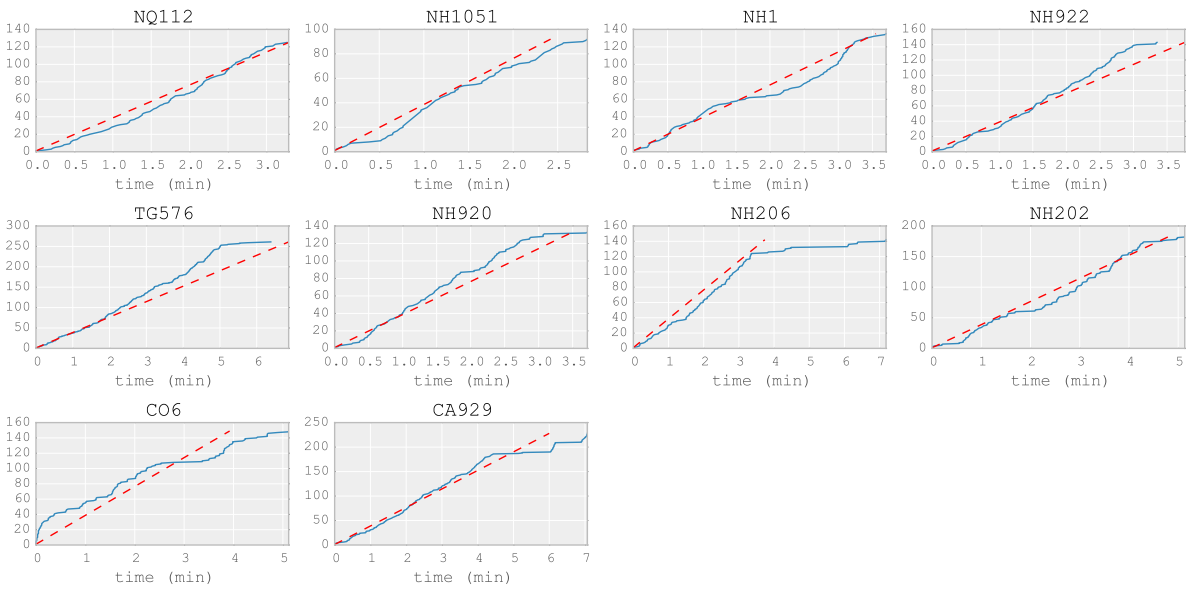


Figure 2.11: Cumulative disembarkation rate of 10 flights.

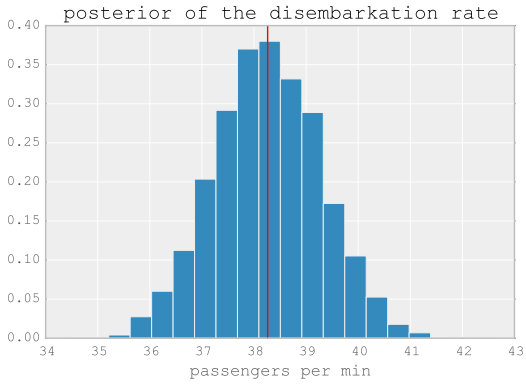


Figure 2.12: Posterior distribution of the disembarkation rate.

uncertainty of the disembarkation rate estimation. We made a statistical model for the interarrival times as follows

$$\begin{aligned} \tau &\sim \text{Exponential}(r) \\ r &\sim \text{Uniform} \end{aligned} \tag{2.22}$$

where τ is the time between two consecutive passengers and r is the rate parameter of the exponential distribution. Figure 2.12 shows the posterior distribution of the rate. The average value of the posterior distribution is 38 passengers per minute. The minimum and maximum of the distribution is 35 and 41 passengers per minute respectively. Assuming the same disembarkation rate for all flights seems reasonable. From the literature study we found an average rate of 19-20 passengers per minute per aircraft exit. If we assume two aircraft exits are used then the observed disembarkation rate corresponds to the value from the literature.

2.3.3 Walking Time

After disembarkation passengers walk from the gate to immigration. Ideally we would record the walking time of each passenger by tracking them but this was not possible. Instead we estimate the walking speed distribution and then determine the walking time from the walking speed and the walking distance.

Literature Review

Daamen [13] reviewed the free flow walking speeds in the literature. The walking speed appears to be normally distributed with a mean of 1.34 m/s and a standard deviation of 0.37 m/s.

Young [87] studied the passenger walking speed at San Francisco International Airport and Cleveland Hopkins International Airport. It was found

that the free-flow walking speed was similar to the walking speed in other transportation terminals. The average free-flow walking speed in the airport terminals was 1.34 m/s and approximately normally distributed with a standard deviation of 0.27 m/s. Also the free-flow walking speed did not vary significantly with age, gender, travel type (business or leisure), group size, number of bags carried, or direction of travel (departure or arrival). The walking speed was however affected by the presence of a moving walkway. About 20% of the passengers using moving walkways stood still or had very low walking speeds. Others on the moving walkways were obstructed by these passengers. The presence of a moving walkway reduced the average walking speed. The average walking speed from the gate to immigration is affected by the number of moving walkways, stairs, escalators and points where passengers are walking at a slower pace or stop. Passengers reduce their walking speeds when they are approaching a travel-path decision and near information signs and boards. The number of these elements differ for the path from each gate. We should therefore expect that the walking speed is different for each gate.

Nikoue et al. [64] had access to the data of WiFi and bluetooth tracking tools at Sydney International Airport. They extracted the walking paths of each gate to immigration and found that the walking times were exponentially distributed. However the data was noisy and inaccurate with low sampling rates, making it impossible to determine the walking time distribution for all gates. Instead they estimated the walking speed distributions. The walking speed distribution of each gate was modelled as a mixture of logistic distributions because the histogram of the walking speeds showed multiple modes. The walking speed distribution for all gates combined has a lognormal distribution.

Passenger Arrival Time

We model the total time from disembarkation to arrival at immigration for a single passenger. We assume that disembarkation is a deterministic process with a constant disembarkation rate r. If disembarkation starts at time t_0 then the disembarkation time d_i of passenger i is

$$d_i = t_0 + \frac{1}{r}(i-1). \tag{2.23}$$

The walking time is the distance divided by the walking speed. We assume that the walking speed V is normally distributed with probability density function

$$f_V(v) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(v-\mu)^2/2\sigma^2} \tag{2.24}$$

where μ is the mean walking speed and σ is the standard deviation. The cumulative distribution function is

$$F_V(v) = \int_{-\infty}^{v} f_V(x)dx. \tag{2.25}$$

Let t' be the time relative to the disembarkation time d_i of a passenger. Then the probability that the walking time τ' of the passenger falls within the interval $(t'_1, t'_2]$ for distance L is

$$\begin{aligned} P(t'_1 < \tau' \leq t'_2) &= P(L/t'_2 \leq V < L/t'_1) \\ &= F_V(L/t'_1) - F_V(L/t'_2). \end{aligned} \tag{2.26}$$

Let t be the time relative to the disembarkation start time t_0, then $t = d_i + t'$ for passenger i. For each passenger i we can calculate the probability p_i that the arrival time at immigration τ_i is within the interval $(t_1, t_2]$

$$\begin{aligned} p_i = P(t_1 < \tau_i \leq t_2) &= P(t_1 - d_i < \tau' \leq t_2 - d_i) \\ &= F_V(L/(t_1 - d_i)) - F_V(L/(t_2 - d_i)). \end{aligned} \tag{2.27}$$

This is the same passenger arrival probability as in equation (2.10) but it uses the walking speed distribution function. We then apply the procedure described in section 2.2 to determine the expected arrival rate $E[\lambda]$ in $(t_1, t_2]$.

Observation

To calculate the arrival probability of a passenger we need to need to determine the mean μ and standard deviation σ of the walking speed. We will infer the walking speed parameters by using the observed disembarkation start times, i.e. when the first passenger leaves the aircraft, and the observed arrivals at immigration. Usually flights arrive close to each other and we cannot distinguish which passenger arrivals at immigration belong to which flight. However there were some flights that were isolated during our observations. Figure 2.13 shows an example of the arriving passengers at the left side of the immigration area on 2012/9/9. The dots represent the ATA of the flights. The vertical lines represent the disembarkation start times for the flights with the same dot color. In this example we can identify two isolated flights: NH956 with disembarkation starting around 13:30 and NH11 with disembarkation starting around 13:55. In total we have identified ten isolated flights in our observation datasets.

Walking Speed Inference

The method of least-squares is used to minimize the error between the observed arrival rate per minute and the estimated arrival rate $E[\lambda]$ in order to find the best fitting walking speed mean μ and standard deviation σ for the

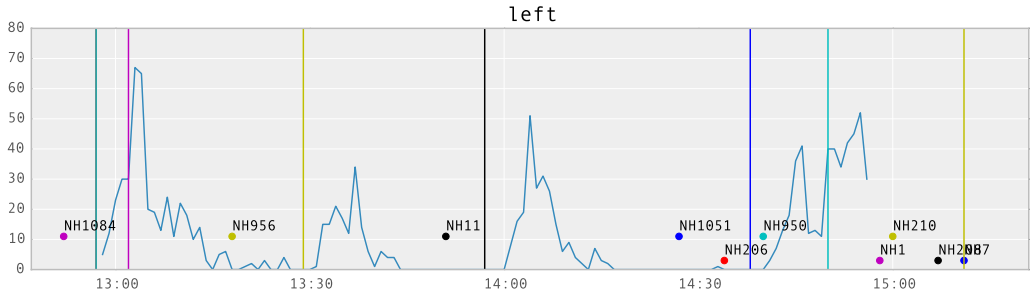


Figure 2.13: Arriving passengers at the left side of the immigration area on 2012/9/9.

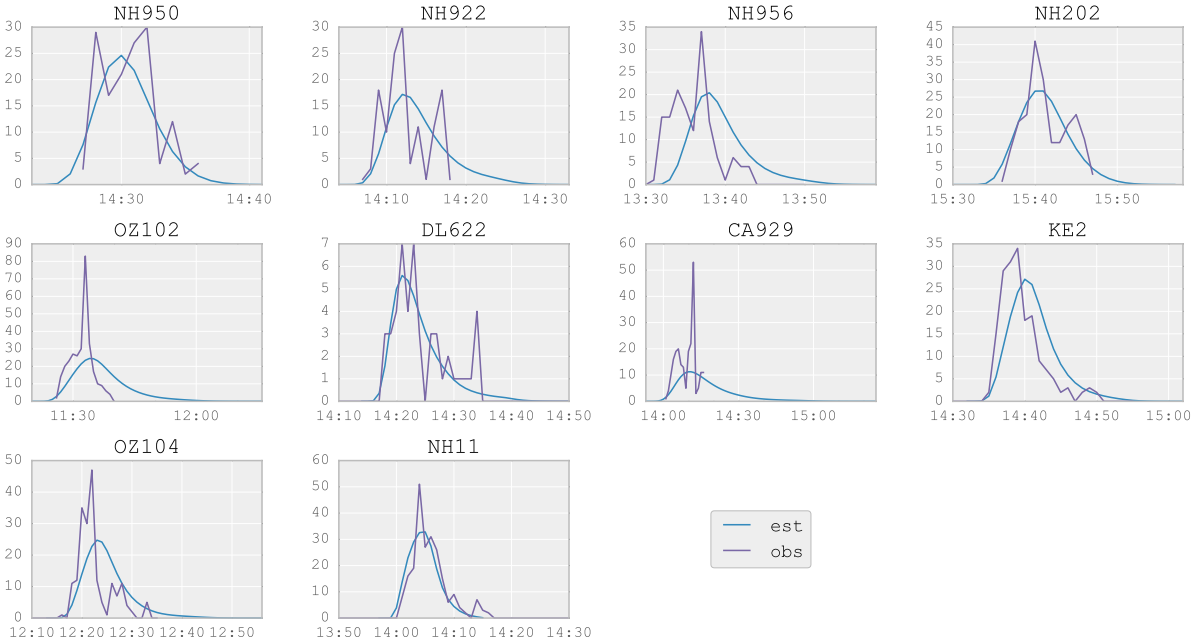


Figure 2.14: Observed and estimated arrival rates with the best fitting walking speed parameters for the 10 isolated flights.

Table 2.1: Walking speed parameters (m/s) and flight details for the 10 isolated flights.

	avg	stdev	dist	pax	gate
NH950	0.48	0.14	166	134	51
NH922	0.56	0.22	190	133	52
NH956	0.60	0.28	190	150	52
NH202	0.62	0.16	238	201	56
OZ102	0.64	0.29	343	329	35
DL622	0.68	0.32	253	45	24
CA929	0.69	0.33	625	224	46
KE2	0.71	0.34	179	193	17
OZ104	0.96	0.46	267	222	33
NH11	1.15	0.28	238	210	56

10 isolated flights. Figures 2.14 shows the observed and estimated arrival rates with the best fitting walking speed parameters for each flight.

The walking speed parameters for each flight are shown in Table 2.1 together with the number of passengers, the distance and the gate. There is a large variation in the mean walking speed and the standard deviation. The highest walking speed is three times as high as the lowest. We expected that flights from the same gate would have the same walking speed parameters. For flights NH922 and NH956 arriving at gate 52 the parameters are similar but for flights NH202 and NH11 at gate 56 the parameters are very different. One possible reason is that the level of congestion in the terminal might have been substantially different for those two flights. Another explanation is simply that there was a measurement error. The total average of the mean and the standard deviation of all the flights is respectively 0.71 m/s and 0.28 m/s.

The number of samples is very limited and we don't have parameter estimations for every gate. It is also reasonable to assume that the walking speed depends on the level of congestion in the terminal, and the number of obstacles on the path to immigration. Collecting more data is recommended for better parameter estimation and to analyse the factors that influence the walking speed.

2.3.4 Number of Passengers

On the day of arrival the airlines inform immigration management about the number of Japanese and foreigners on each flight who have to go through immigration, i.e. the transfer passengers are excluded. The passenger data is only known on the day of arrival which means that if long-term planning is desired, the number of Japanese and foreign passengers on a flight need

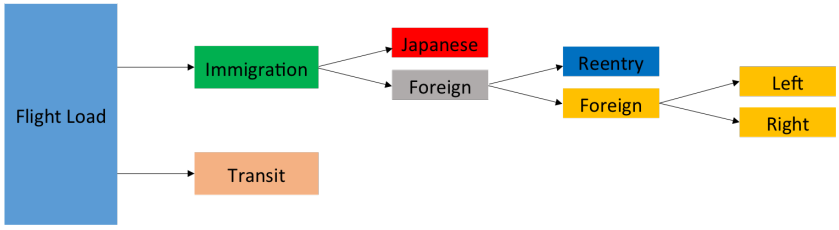


Figure 2.15: Types of passengers.

to be estimated. In the ideal case we would use historical passenger data of a flight to estimate the number of passengers, but at best we have only five historical values for a flight. However the flight schedule with the aircraft type for each flight is known far in advance. In this section we will investigate if we can estimate the number of each passenger type on a flight if only the aircraft type is known.

The total number of passengers of a flight N_{flight} can be roughly estimated with

$$N_{\text{flight}} = L \times C \tag{2.28}$$

where L is the flight load factor and C is the capacity of the aircraft type. As shown in Figure 2.15 there are different types of passengers. After arrival some passengers transit to another flight. For the passengers who go through immigration, the Japanese and foreign passengers have separate service counters. The foreigners with a reentry permit have special reentry counters. The other foreigners can choose between service counters on the left or right side of the immigration area (Figure 1.3). Overall there are four sections at immigration with each section being a separate queueing system. The number of passengers N_{section} arriving at each section can be estimated with

$$N_{\text{section}} = N_{\text{flight}} \times (1 - p_{\text{transit}}) \times p_{\text{nation}} \times p_{\text{section}} \tag{2.29}$$

where p_{transit} is the percentage of transit passengers, p_{nation} is the percentage of foreigners or Japanese passengers who will go through immigration and p_{section} is the percentage of reentry permit holders or the percentage of other foreigners who go to either the left or right service counters. Below we will discuss each parameter.

Aircraft capacity

The aircraft type that is used for a flight is known from the flight schedule months in advance. The capacity C of an aircraft depends on the seats configuration of the first/business class and the economy class. The exact seats configuration data of each flight is however unavailable. Instead aircraft data from The Travel Insider [74] will be used to estimate the average

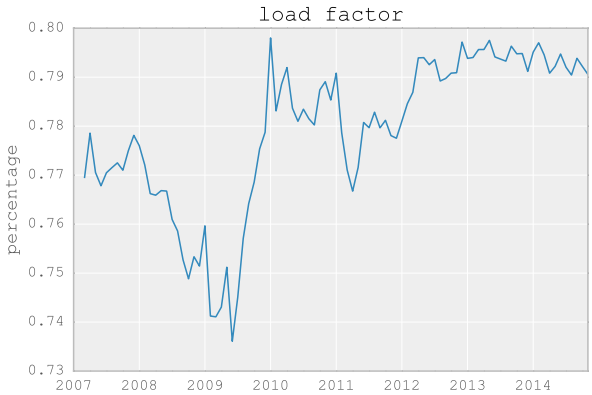


Figure 2.16: Passenger load factor for the international market (seasonally adjusted). Source: IATA [37].

number of seats. If the aircraft type is not present in the database a capacity of 155 passengers is assumed.

Load factor

IATA [37] publishes a monthly report on the air passenger market. The load factor of the aviation industry is defined as the ratio of passenger-kilometers traveled to seat-kilometers available. The published load factor is organized by region, country or domestic/international flights. The load factor L of a single flight is the ratio of the number of passengers to the aircraft capacity. The load factor varies per flight but such data is unavailable. As an estimation we assume that all flights have the same load factor. Figure 2.16 shows the monthly load factors from 2007 to 2014. The difference between the load factor of any two months is at most 6 percent. In 2008 an increase in oil prices and a collapse in world trade caused a significant decline of the load factor [38].

Transit

To determine the number of passengers that go to immigration we need to subtract the number of transit passengers. The exact number of transit passengers on a flight is only known by the airlines but this information is not provided to the immigration office. The monthly number of transit passengers is published by Narita Airport [60]. Figure 2.17 shows the percentage of transit passengers for each month from 2003 to 2013. The difference between two months can range from -6% to +6%. The transit percentage peaks around June and December.

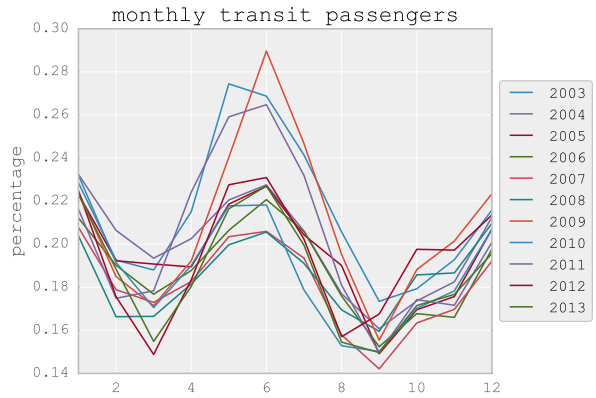


Figure 2.17: Percentage of monthly transit passengers.

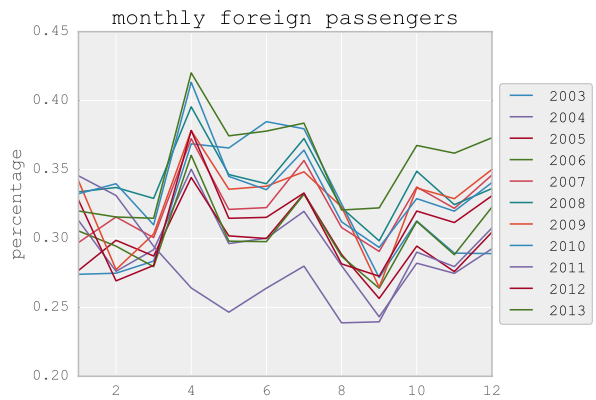


Figure 2.18: Percentage of monthly foreign passengers.

Percentage Foreigners

When the passengers arrive at immigration, foreigners and Japanese passengers join different queueing systems. The monthly passenger volume data published by Narita airport [60] also contains the total number of Japanese and foreigners separately. Figure 2.18 shows the ratio of the monthly number of foreign passengers to the total international passenger volume. The percentage of foreigners peaks in April and is lowest in September. The difference between two months can range of -7% and +11%.

Figure 2.19 shows the histogram of the true percentage of foreigners on the flights that we obtained during the observation periods. We can see that the range is very large with an average of 42%.

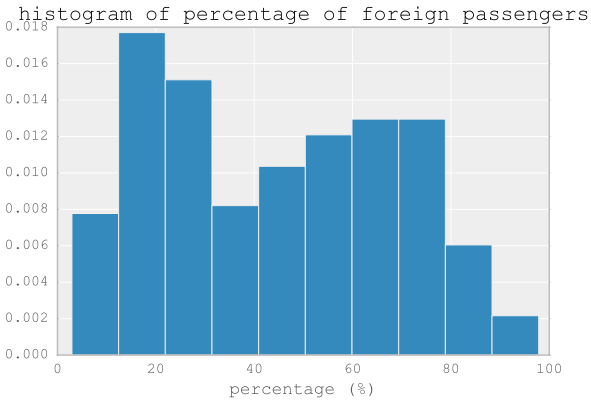


Figure 2.19: Histogram of the percentage of foreign passengers on a flight.

Table 2.2: Percentage of reentry passengers per continent.

	percentage
Continent	
Africa	0.36
Asia	0.26
Europe	0.17
North America	0.15
Oceania	0.12
South America	0.53
Stateless	0.41

Reentry

The foreign passengers need to be divided into passengers with a reentry permit and all other foreign passengers. The number of reentry passengers is difficult to estimate because there is no information about the reentry passengers on a flight or the monthly average. We did receive an internal report from Narita Airport [61] with the overall statistics of the passenger characteristics from 2011. Table 2.2 shows the percentage of the reentry passengers from each continent. The differences between continents can be quite large. The data in Table 2.3 gives the number of reentry passengers per country however we do not know the percentage of each nationality on a flight. Within a continent the reentry percentage varies significantly per country. 71 percent of the passengers with a Filipino nationality have a reentry permit while very few nationals from Hong Kong are reentry permit holders.

We observed the number of arrivals at each queueing system on several afternoons. This allows us to calculate the reentry percentage during the

Table 2.3: Percentage of Asian reentry passengers per nationality.

	percentage
Nationality	
Philippines	0.71
China	0.46
Other	0.43
India	0.40
Korea	0.22
Thailand	0.20
Indonesia	0.19
Malaysia	0.16
Taiwan	0.08
Singapore	0.05
Hong Kong	0.02

Table 2.4: Observed percentage of reentry passengers.

	flight	percentage
2011-02-06	CA929	0.34
2011-02-06	NH922	0.37
2011-02-06	total	0.17
2011-11-27	NH202	0.12
2011-11-27	NH950	0.20
2011-11-27	total	0.14
2013-04-14	total	0.14
2013-04-15	total	0.09

observation periods. Generally multiple flights arrive at the same time so we cannot distinguish the reentry passengers of each flight. However for four isolated flights we have determined the number of reentry passengers on the flight. Table 2.4 shows the reentry percentage of the isolated flights and the total reentry percentage of all flights during four observation period.

Selection of Left or Right Foreign Service

At immigration the foreign passengers without a reeentry permit can go to the service counters on the left side or the right side of immigration. The right side is always open but the left side is only used during busy periods and only if there is enough staff available. It is difficult to forecast how many foreigners go to which side because we don't know if the left side is open. And the ratio between the left and right side is also hard to predict

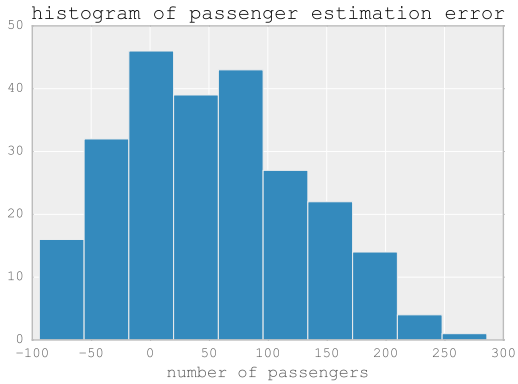


Figure 2.20: Histogram of the estimation error of the number of passengers per flight.

because the passengers are often directed by the immigration staff to go to a certain side. These problems make it impractical to forecast the arrivals of the left and right side separately. Instead we treat the left and right side as one system.

Comparison of Estimation with Observation

We compare the estimated number of passengers with the actual number of passengers for the flights that arrived during the observation periods. We consider the total number of passengers on the flight N_{total}, i.e. the Japanese plus all foreign passengers combined. The estimated number of passengers who go through immigration is

$$N_{\text{total}} = L \times C \times (1 - p_{\text{transit}}). \tag{2.30}$$

Figure 2.20 shows the histogram of the error of the estimations. On average the formula estimates 51 passengers more on a flight than in reality. If there are six flights arriving in an hour then the total error would be 300 passengers which is equivalent to one or two flights.

We can conclude that predicting the total number of passengers using an average load factor, an average transit percentage and the aircraft seat capacity is highly inaccurate. It is recommended that Narita airport builds a database with historical flight data to improve the predictions or requests early estimations from the airlines. From the analysis we can also conclude that it is difficult to estimate of the percentage of reentry passengers and to predict the side which the non-reentry foreign passengers will select. Therefore we will only forecast the arrivals of all foreign passengers including reentry permit holders as one group.

2.4 Arrival Forecast

The theoretical framework of the arrival forecasting model was given in section 2.2 and the parameters for the model were described in section 2.3. In this section we will compare the results of the arrival forecasting model with the observed arrival rates for both the point forecast and the distributional forecast. We will also compare the results with a deterministic approximation and a Monte Carlo simulation model.

2.4.1 Forecasting Models

Convolution Model

We will refer to the arrival forecasting model from section 2.2 as the convolution model. The output of the convolution model is a distributional forecast. Table 2.5 shows the parameters and the distributions that are used with the convolution model. The flight delay distribution depends on δ, the time between the ETA update time and the ETA of the flight. We then select from the empirical distributions in Figure 2.8 the distribution for the corresponding δ. The disembarkation delay is assumed to be normally distributed with a mean of 9.9 minutes and a standard deviation of 3.7 minutes. The disembarkation rate is assumed to be constant with a value of 38 persons/minute. The walking speed of a passenger is normally distributed with a mean of 0.71 m/s and a standard deviation of 0.28 m/s.

From the analysis of the walking times we learned that the parameters of the walking speed can vary significantly per flight and gate. However because of the small number of observation samples we could not determine the exact relationship. To take the uncertainty of the walking speed parameters into account we let the mean of the walking speed also be normally distributed with a mean of 0.71 and standard deviation of 0.22. The standard deviation of the walking speed is kept fixed. The uncertainty in the walking speed parameters is a flight property and not a passenger property because the walking speed distribution is assumed to be the same for all passengers of the same flight. We introduce the parameter uncertainty into the convolution model as follows. We convert the distribution of the walking speed mean to a walking time delay distribution d_w using equation (2.26). The total delay distribution d of a flight is then the convolution of all delay components

$$d = d_f * d_d * d_w \tag{2.31}$$

where d_f is the flight delay distribution and d_d is the disembarkation delay distribution.

Table 2.5: Parameters for the convolution model and Monte Carlo simulation.

parameter	distribution	value	unit
flight delay	empirical	$f(t_{\text{ETA}} - t_{\text{update}})$	min
disembarkation delay	Normal	$\mu = 9.9$, $\sigma = 3.7$	min
disembarkation rate	constant	38	p/min
walking speed	Normal	$\mu = N(0.71, 0.22)$, $\sigma = 0.28$	m/s

Monte Carlo Simulation

We develop a Monte Carlo simulation model to validate the convolution model and to use the generated arrival rate samples for performance evaluation of a queueing system. For the Monte Carlo simulation model we use the same parameters as for the convolution model (Table 2.5). The flight delay and disembarkation delay are drawn randomly from the distributions for each flight. The walking speed of each individual passenger is drawn randomly from the normal distribution. The distributional forecast using the convolution model and the Monte Carlo simulation model should produce similar results when a 1-minute interval is used.

A disadvantage of the Monte Carlo simulation model is that it is slower than the convolution method. For one flight the convolution method takes 0.5 seconds while a simulation with 1000 runs takes 2.3 seconds in total. An advantage of the Monte Carlo simulation model is that it is simple to add more uncertainties to the model. Also it is time resolution independent. The convolution model becomes less accurate for larger time intervals because the arrival probability matrix needs to be convolved with a delay distribution that is resampled to the larger time interval.

Deterministic Approximation

A point forecast gives a single value of λ_j at each interval j. One way to achieve that is to create a distributional forecast and then calculate the expected value using equation (2.15).

Another approach, that will give a different point forecast, is to assume an average flight delay and an average disembarkation delay for all flights. Furthermore we assume that the passengers who disembark within the same time interval, will all disembark at the same time. As a consequence the passengers of that interval will have the same arrival distribution. Let t be the time relative to the disembarkation start t_0 of the flight. Time is divided in intervals $j = 1, 2, \ldots$ of equal length $\Delta = t_j - t_{j-1}$. We assume that all passengers disembarking in interval n leave the aircraft simultaneously at time t_n which is the time at the end of interval n. Let $F_W(t')$ be the

Table 2.6: Parameters for the deterministic approximation.

parameter	distribution	value	unit
flight delay	constant	0	min
disembarkation delay	constant	9.9	min
disembarkation rate	constant	38	p/min
walking speed	Normal	$\mu = 0.71$, $\sigma = 0.28$	m/s

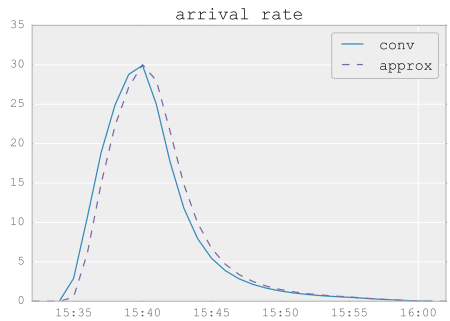


Figure 2.21: The expected value of a distributional forecast with the convolution model (conv) and the arrival rate with the deterministic approximation (approx) for a single flight without delay uncertainty.

walking time distribution for a passenger with t' being the time relative to his disembarkation time t_n. The number of passengers that disembark in interval n is represented by D_n. Then the expected number of passengers from interval n who arrive at immigration in interval j is

$$E[\lambda_j^n] = D_n[F_W(t_j - t_n) - F_W(t_{j-1} - t_n)]. \tag{2.32}$$

The total number of passengers arriving in interval j from all disembarkation intervals is

$$E[\lambda_j] = \sum_n D_n[F_W(t_j - t_n) - F_W(t_{j-1} - t_n)]. \tag{2.33}$$

The parameter values used for the deterministic approximation is shown in Table 2.6. Figure 2.21 shows an example of the arrival rate with the expected value of the convolution model without delay uncertainty (conv) and the deterministic approximation (approx) for a flight with 166 passengers. The disembarkation start time is set to the same value for both methods. The results are almost identical. The computational time for the distributional forecast and the expected value is 500 ms while the deterministic approximation only requires 6 ms.

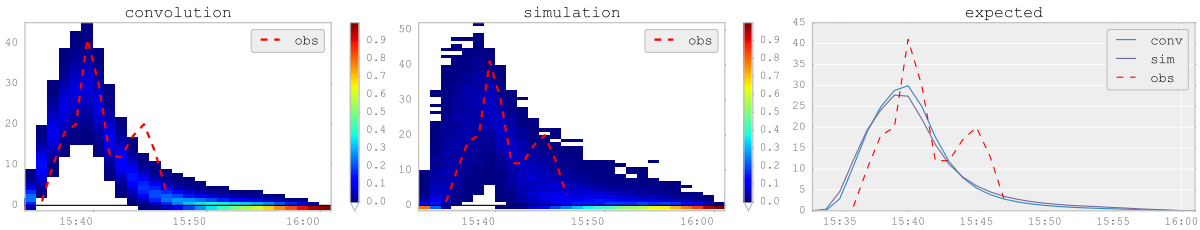


Figure 2.22: Comparison of the forecasting models with the observed arrival rates for flight NH202 with deterministic delay.

2.4.2 Single Flight

We compare the results of the convolution model, the Monte Carlo simulation model and the deterministic approximation with the observed arrival rate of a single flight. We consider the case with and without delay uncertainty. We select flight NH202 from the isolated flights. This flight had 201 passengers and it arrived at gate 56 which is 238 meters from immigration. From Table 2.1 we know that the fitted mean walking speed (0.62 m/s) for this particular flight is lower than the mean walking speed used in the forecasting models (0.71 m/s).

Arrivals With Deterministic Delay

In this case we compare the flight arrivals with a deterministic flight delay and disembarkation delay. Also we do not take the uncertainty of the walking speed mean into account. Figure 2.22 shows the results of the convolution model and the Monte Carlo simulation model together with the observed arrival rates. The resolution is 1 minute on the time axis and 1 person on the arrival rate axis. The arrival probabilities are very similar and the expected values are practically identical. The passengers of the models arrive earlier than the observed arrivals because of the higher mean walking speed used in the models.

Arrivals With Delay Uncertainty

We assume that the flight ETA time is updated 60 minutes before the ETA. The flight delay distribution for $\delta = 60$ is shown in Figure 2.23 together with the disembarkation delay distribution and the walking time delay due to the uncertain walking speed mean. The convolution of these three delay distributions gives the total delay distribution for the flight. The total delay distribution is a probability mass function with 1-minute resolution.

Figure 2.24 shows the forecasts with delay uncertainty. We sum the probabilities per 5 passengers to make the figure more clear. Without grouping per 5 passengers the probabilities are so small that there would be no differ-

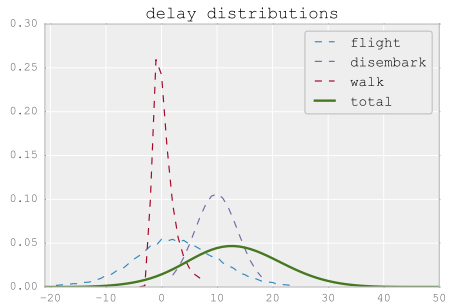


Figure 2.23: Flight delay, disembarkation delay and walking time delay distributions.

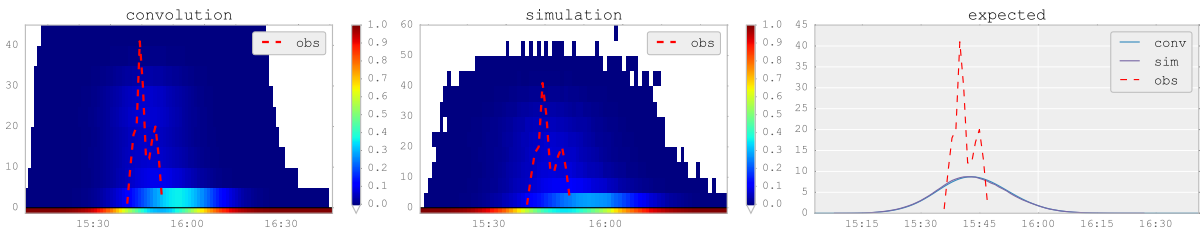


Figure 2.24: Comparison of the forecasting models with the observed arrival rates for flight NH202 with delay uncertainty.

ence in the colors of the plots. The distributional forecast of the simulation and convolution model are similar and the expected values overlap. Compared to the case with deterministic delays the arrival probability plots are more diffuse. The expected values with delay uncertainty are smaller and have a wider distribution.

2.4.3 Multiple Flights

In this section we present the arrival forecasts with multiple flights. We consider the flights in the six observation periods: five periods at the South wing and one at the North wing (2013/10/5). Figure 2.25 shows the deterministic approximation and the observed arrival rates per 5 minutes. For 2012/9/9 the deterministic approximation predicts the times of the peaks correctly. For 2011/2/6 and 2013/10/5 the deterministic approximation is reasonable. On the other days there are more mismatches of the arrival peaks. Overall the deterministic approximation gives a reasonable indication of the level of fluctuation in the arrival rates.

The distributional forecasts with the convolution model is shown in Figure 2.26 together with the expected values, the 95th percentile upper bound and the observed arrival rates. The arrival probabilities are calculated with a 1-minute accuracy and shown in the plots per 5 persons. The observed

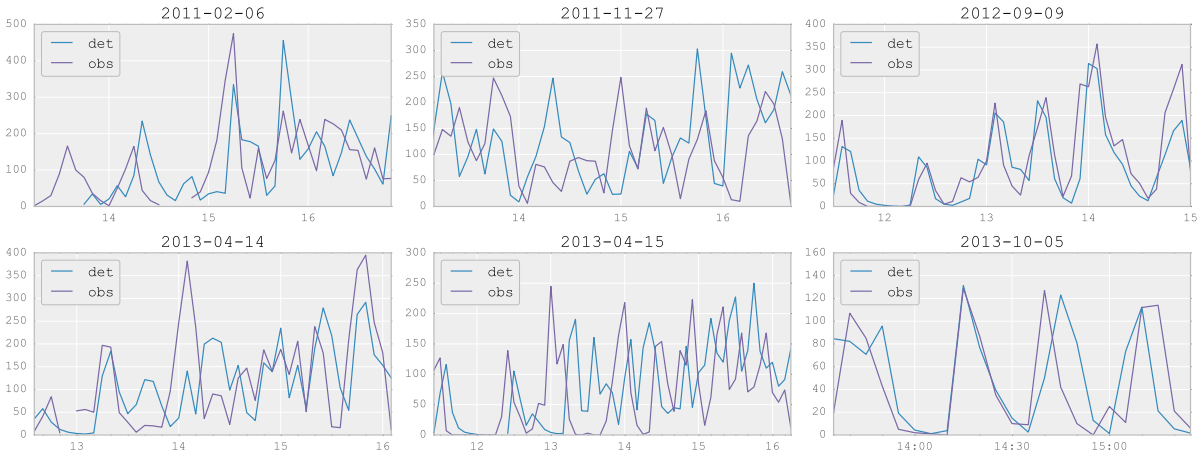


Figure 2.25: Deterministic approximation compared to the observed arrival rates with a 5-minute interval.

arrivals exceed the 95th percentile at times but the upper bound does give a reasonable indication of the maximum arrival rate per minute.

For practical use we propose to combine the convolution model with the deterministic approximation to give the decision makers an indication of the uncertainty (arrival probability), the trend (expected value), the upper limit (95th percentile) and a sample path (deterministic approximation). Figure 2.27 shows the proposed forecast output with a 1-minute interval.

2.5 Conclusion

In this chapter we have reviewed the arrival forecasting models used for call centers and airport service facilities. Because we do not have the required amount of historical arrival data we cannot apply regression models or time series models. We have developed a distributional forecasting model based on the sum of random variables and the convolution operation. The forecasting model requires the distributions of the flight delay, the disembarkation delay and the walking speed. We have discussed how we gathered data at Narita immigration to determine these distributions. In addition we have also developed a Monte Carlo simulation model and a deterministic approximation. The models give reasonable results when compared with the observed arrival rates. For use in practice the distributional forecasting model can give the decision makers at Narita immigration a good indication of the trend, the variance and the upper bound of the arrival rates while the deterministic approximation gives a sample path of the arrival rates.

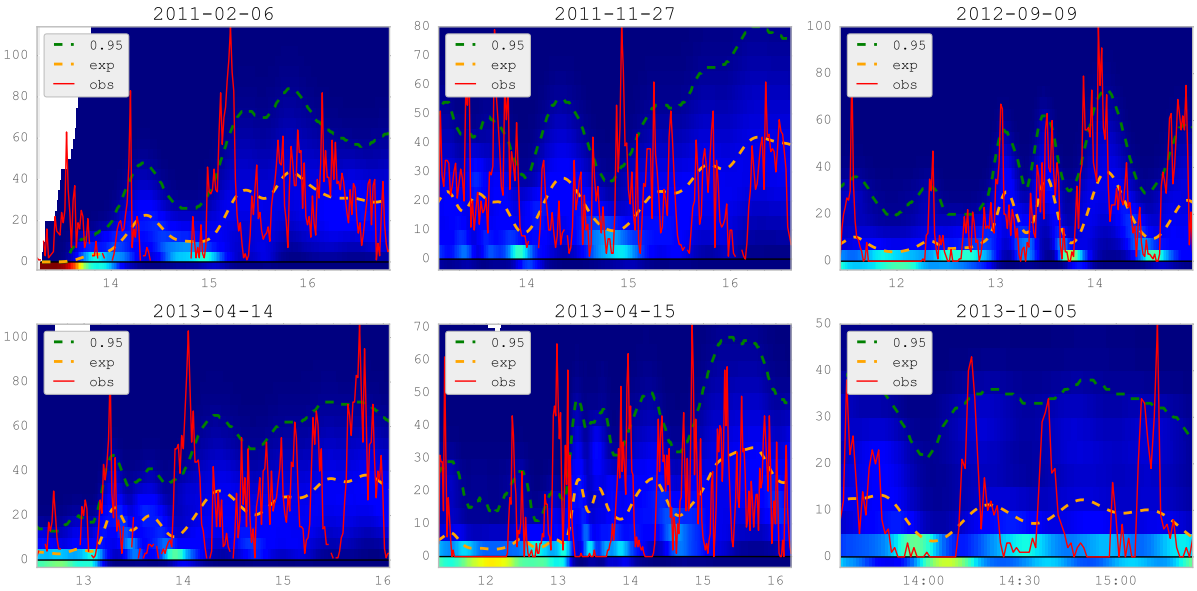


Figure 2.26: Distributional forecast compared to the observed arrival rates with a 1-minute interval.

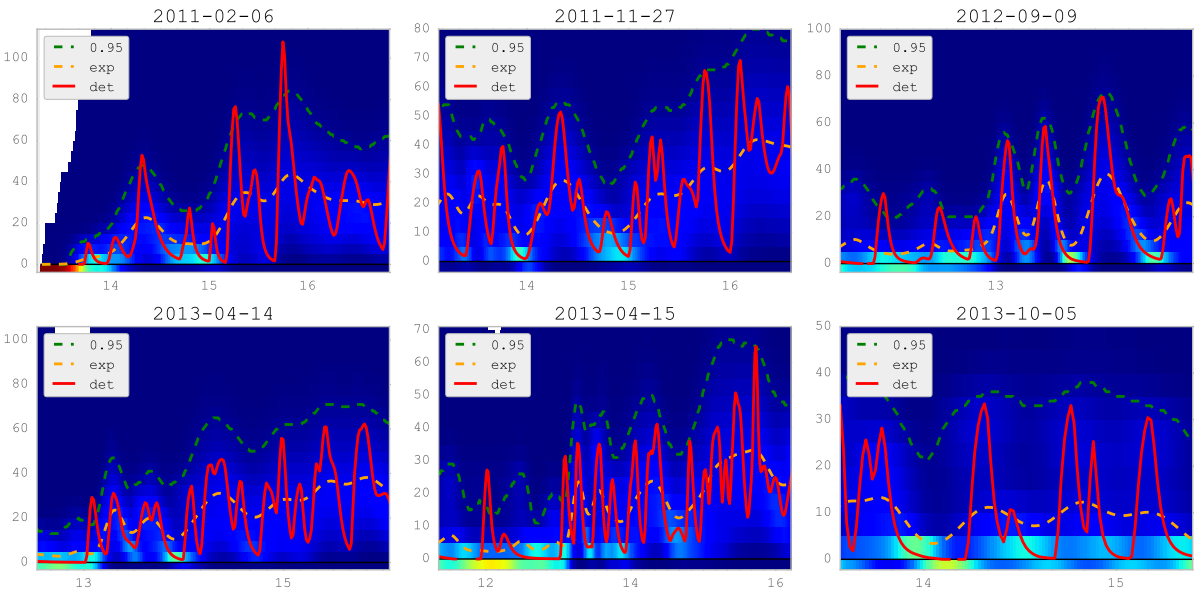


Figure 2.27: Distributional forecast combined with the deterministic approximation with a 1-minute interval.

Chapter 3

Queueing Models

A queueing system is a system where customers arrive, wait for service, receive service and leave the system. There are six characteristics to describe a queueing system [32]: the arrival pattern of customers, service patterns, queue discipline, system capacity, number of service channels and stages of service. The arrival pattern describes the probability distribution of the time between two customer arrivals. The service pattern describes the probability distribution of the service time. The queue discipline refers to the order in which customers are served, usually First-Come-First-Served is used. The system capacity is either finite or infinite. Each stage of service can have one or more parallel service channels. Narita immigration is not a single queueing system but a collection of multiple systems catering for different types of passengers: foreign, reentry and Japanese passengers. There are several approaches to model the queueing system depending on the system characteristics. The input of a queueing system is the arrival rate $\lambda(t)$, the number of servers $s(t)$ and the mean service time $1/\mu$. The output of interest is the waiting time $W(t)$.

In this chapter we first review the literature on queueing theory methods. We implement three queueing models that can deal with overload to find the most appropriate queueing model for the foreign passengers. These three queueing models have not been compared in the literature before. For the models we have gathered data at Narita Airport immigration. We present the results of our observations. The output of the models are compared with the observed waiting time values. In the last section we compare the processing times at Narita Airport to those at Incheon Airport. They claim to have the fastest immigration service in the world.

3.1 Literature Review

In this section we review queueing theory methods in order to determine performance measures such as the queue length and the waiting time of

a queueing system. We categorize the queueing theory methods in three groups: methods that involve the stationary queueing theory, methods that involve solving one or more ordinary differential equations (ODE) and deterministic methods.

3.1.1 Stationary Approximations

Stationary queueing theory deals with the long-term behavior of a queueing system. In this section we discuss the steady-state solutions for the queue length and waiting time. Then we will discuss methods to apply the stationary queueing theory to non-stationary situations.

Let the stochastic variable $X(t)$ represent the number of customers in the system at time t, i.e. the total in the queue and in service. We assume a Markovian process for the arrivals and service completions. The time until the next customer arrival is exponential with rate parameter λ. The time until the next service completion is also exponential with mean $1/\mu$, in other words μ is the service rate of a single server. The queueing system has s number of parallel servers. We assume that the queue has unlimited capacity and there are no customer abandonments. A queueing system with such characteristics is also indicated by the notation $M/M/s$. We want to describe probabilistically how $X(t)$ changes as a function of time. Let $p_n(t)$ denote the probability that there are n passengers in the system after time t has passed. These transition probabilities can be obtained by solving the Kolmolgorov differential equations [32]

$$\begin{aligned}
\frac{dp_o(t)}{dt} &= -\lambda p_0(t) + \mu p_1(t), \\
\frac{dp_n(t)}{dt} &= \lambda p_{n-1}(t) - \big(\lambda + n\mu\big)p_n(t) + (n+1)\mu p_{n+1}(t) \quad 1 \le n < s, \\
\frac{dp_n(t)}{dt} &= \lambda p_{n-1}(t) - \big(\lambda + s\mu\big)p_n(t) + s\mu p_{n+1}(t) \qquad\qquad n \ge s.
\end{aligned} \tag{3.1}$$

This set of equations can be written in matrix notation as

$$\underline{p}'(t) = \underline{p}(t)Q \tag{3.2}$$

where Q is called the intensity matrix of infinitesimal generator. Once the transition probabilities are known we can derive $L_q(t)$, the expected number of customers in the queue, with

$$L_q(t) = \sum_{n=s+1}^{\infty} (n-s)p_n(t). \tag{3.3}$$

Suppose that $\lim_{t\to\infty} p_n(t) = p_n$ then the steady-state solution can be obtained from

$$\underline{0} = \underline{p}Q. \tag{3.4}$$

For an $M/M/s$ queueing system the steady-state probabilities are given by [34]

$$p_0 = \left(\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1-\lambda/(s\mu)} \right)^{-1}, \tag{3.5}$$

$$p_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} p_0 & 0 \leq n \leq s, \\ \frac{(\lambda/\mu)^n}{s! s^{n-s}} p_0 & n > s. \end{cases} \tag{3.6}$$

The probability of delay is

$$p_d = 1 - \sum_{n=0}^{s-1} p_n. \tag{3.7}$$

The steady-state value of the number of waiting customers L_q is equal to

$$L_q = \frac{p_0 (\lambda/\mu)^s \rho}{s!(1-\rho)^2} \tag{3.8}$$

where $\rho = \lambda/(s\mu)$ is the utilization ratio or traffic intensity. A system is underloaded when $\rho < 1$, critically loaded when $\rho = 1$ and overloaded when $\rho > 1$. For the queue to be stable it is necessary that $\rho < 1$. The expected waiting time can be determined with Little's formula

$$W_q = \frac{L_q}{\lambda}. \tag{3.9}$$

Let $\mathcal{W}_q$ be the waiting time experienced by a customer. Then the probability of a customer waiting longer than τ minutes is

$$\begin{aligned} P(\mathcal{W}_q > \tau) &= (1 - P(\mathcal{W}_q = 0)) e^{-s\mu(1-\rho)\tau} \\ &= (1 - \sum_{n=0}^{s-1} p_n) e^{-s\mu(1-\rho)\tau}. \end{aligned} \tag{3.10}$$

Simple Stationary Approximation

In the simple stationary approximation (SSA) the arrival rate and number of servers are averaged over the period of interest

$$\lambda = \frac{1}{T} \int_0^T \lambda(t) dt, \tag{3.11}$$

$$s = \frac{1}{T} \int_0^T s(t) dt. \tag{3.12}$$

Green, Kolesar, and Svoronos [28] studied the accuracy of the simple stationary approximation for a sinusoidal arrival process. It was found that

if the relative amplitude is 25% or more, the relative error is at least 10% and often significantly larger. The simple stationary approach may provide reasonable estimates for small systems with one or two servers, relative amplitudes less than 10% and infrequent events. During the period the system can be temporarily overloaded but it needs to satisfy the condition

$$\rho = \frac{1}{T}\int_0^T \frac{\lambda(t)}{s(t)\mu}dt < 1. \tag{3.13}$$

As an example suppose two flights arrive in the same hour with respectively 100 and 200 passengers. The average service time is 1 minute per passenger and there are 8 servers. The average arrival rate is thus $\lambda = 300/60 = 5$ passengers per minute. The traffic intensity of the system is $\rho = 0.625$. The expected waiting time W_q is then 3.3 seconds and the average queue length L_q is 0.3 passengers. The probability of no delay $P(\mathcal{W}_q = 0)$ is 0.83.

Stationary Independent Period by Period Approach

In the example of the previous section the performance was calculated with a constant arrival rate during the one hour period. But it is unrealistic to assume that the arrivals at immigration are stationary. There will be a peak in the number of passenger arrivals every time a flight arrives and after that there will be a period with no passenger arrivals. Assuming steady-state and Poisson arrivals during the whole hour would be unrealistic. In a non-stationary environment the arrival rate fluctuates. Stationary queueing theory only deals with the steady-state behavior of a queueing system but it can be used in a non-stationary manner with the stationary independent period by period approach (SIPP) and the pointwise stationary approximation (PSA).

The PSA applies the stationary queuing theory with the instantaneous arrival rate $\lambda(t)$ and number of servers $s(t)$ at each time t. Green and Kolesar [27] studied the accuracy of the PSA and found that the estimations of the expected delay, expected queue length, probability of delay and probability of all servers busy become more reliable when the arrival rate and service rate increase. The PSA provides an upper bound for the performance of non-stationary queueing systems. The performance of the PSA is worse when the traffic intensity increases.

The SIPP is similar to the PSA but instead of using the instantaneous values of the arrival rate and service rate, the total time period is first divided into intervals of arbitrary length. Then the stationary queueing theory is applied to each interval independently with the average arrival rate and average number of servers in the interval [26].

When the service times are medium to long, the system performance can change significantly because of time lags in congestion [29]. A solution is

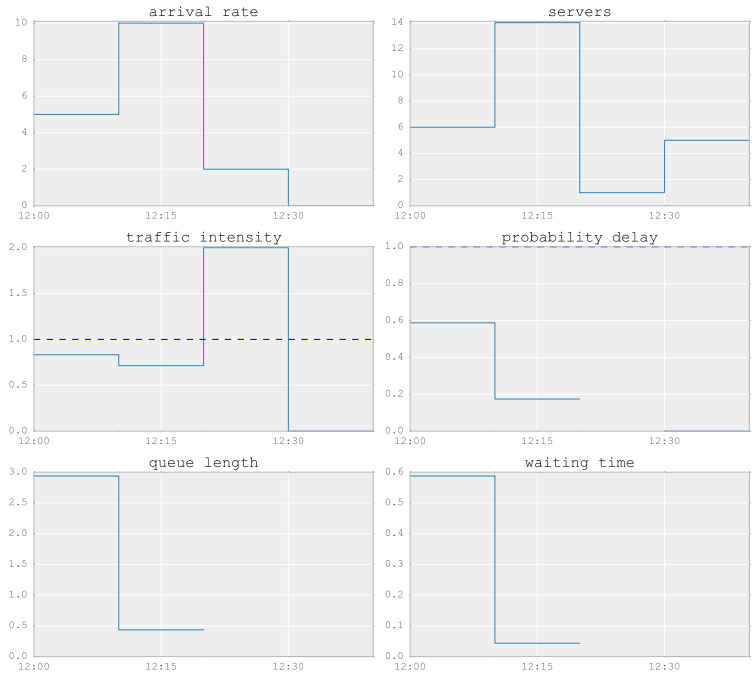


Figure 3.1: Example of the SIPP.

to shift the arrival rate by the mean service time first and then apply the stationary queueing models. The methods with this adjustment are called the lagged PSA and lagged SIPP.

A major drawback of these approaches is that the system cannot be overloaded at any time (PSA) or on average in an interval (SIPP). The utilization ratio needs to be less than 1. In call centers with a high level of service standard this is generally no problem but other service systems such as airport facilities tend to be frequently overloaded.

As an example suppose the passengers of a flight arrive at immigration from 12:00. The number of passenger arrivals per 10-minute interval are known to be 50, 100, 20 and 0 passengers. The number of servers per interval are respectively 6, 14, 2 and 4. The results of the SIPP are shown in Figure 3.1. Each time interval is analyzed independently. In the third interval the traffic intensity is larger than one and therefore no estimation for the waiting time or queue length is available.

Effective Arrival Rate Approximation

If a customer arrives at time t and experiences waiting time W_q then he will receive service during the window $[t + W_q, t + W_q + 1/\mu]$. Customers who arrive in a time interval but whose service window does not completely fall within the same interval, will impact the demand for service in later time

intervals. Thompson [75] developed an algorithm to adjust the demand for service in each interval. In the first step the simple stationary approximation (SSA) is used to determine the expected waiting W_q for the total time period, i.e. the waiting time is the same for all intervals. Then in the second step the effective arrival rate in each interval is calculated. The effective arrival rate during an interval is equal to the number of customers that receive service in the interval. For interval j the arrival rate λ_j is increased by the number of customers who arrived in previous intervals but are served in interval j. And the arrival rate λ_j is decreased by the number of customers who arrived in interval j but are served in later intervals. Once the effective arrival rates are determined a stationary queueing model is applied to each interval.

The algorithm was developed for an interval-based arrival rate function. For a continuous arrival rate function, the algorithm reduces to the calculation of the moving average [43]. The effective arrival rate at time t equals the average of the original arrival rate during the window $[t - W_q - 1/\mu, t - W_q]$

$$\lambda_{\text{eff}}(t) = \int_{t-W_q-1/\mu}^{t-W_q} \mu\lambda(r)dr. \tag{3.14}$$

Thompson [75] compared the performance of the effective arrival rate approximation (EAR) with the SIPP approximation when setting staff requirements. The EAR resulted in 8% fewer staff hours and a 3% higher service level.

Stationary Backlog-Carryover Approach

In the PSA, SIPP and EAR approximation the arrival rate at time interval t needs to be less than the service rate to ensure that there is no overloading of the system. Also each interval is assumed to be independent. In reality queueing systems can be overloaded for an extended period of time and the queues that are built up during one interval will spill over into the next. Stolletz [72] proposed the stationary backlog-carryover approach (SBC) to solve this issue. In this approach a backlog is created when the service capacity is insufficient. The backlog is carried over to the next interval as additional demand.

The SBC works as follows. The time period of interest is divided into time intervals. An appropriate interval length is the mean service time. The number of arrivals b that could not be served in an interval is determined using the Erlang B loss system, also known as an $M/M/s/s$ system. In a loss system there are no queues and customers are blocked from entering the system when all servers are occupied. Let λ_t be the arrival rate for the current interval and let b_{t-1} be the number of blocked passengers in the previous interval, i.e. the backlog rate. The artificial arrival rate $\tilde{\lambda}_t$ in the current interval is then

$$\tilde{\lambda}_t = \lambda_t + b_{t-1} \tag{3.15}$$

The probability of blocking $P_t(B)$ in the current interval is calculated with the Erlang B blocking formula using the artificial arrival rate

$$P_t(B) = \frac{(\tilde{\lambda}_t/\mu)^{s_t}}{s_t! \sum_{k=0}^{s_t} \frac{(\tilde{\lambda}_t/\mu)^k}{k!}} \tag{3.16}$$

where s_t is the number of servers in the current interval. The number of blocked passengers b_t for the current interval is

$$b_t = \tilde{\lambda}_t P_t(B). \tag{3.17}$$

Finally we determine the modified arrival rate λ_t^M as

$$\lambda_t^M = \tilde{\lambda}_t - b_t = \lambda_t + b_{t-1} - b_t. \tag{3.18}$$

The modified utilization ratio ρ_t^M in the current interval is

$$\rho_t^M = \frac{\lambda_t^M}{s_t \mu}. \tag{3.19}$$

To estimate the queue length and waiting time in interval t we apply the $M/M/s$ model with λ_t^M as the arrival rate. Even if the original traffic intensity is larger than one, the modified traffic intensity is always smaller than one. Therefore the SBC can also be applied in overload situations.

Figure 3.2 shows the results of the SBC for the same queueing system as in the example of the SIPP. We can see the buildup of the queue during both underloaded and overloaded intervals.

3.1.2 ODE Methods

In this section we will discuss methods to obtain the the transition probabilities $p_n(t)$ by solving the set of differential equations $\underline{p}'(t) = \underline{p}(t)Q$ (3.2) or by solving a simpler set of ODE. Once the transition probabilities $p_n(t)$ have been determined we can obtain the expected number of customers in the queue $L_q(t)$ at time t with

$$L_q(t) = \sum_{n=s(t)+1}^{m} [n - s(t)]p_n(t). \tag{3.20}$$

The expected waiting time from time t of joining the queue is the number of customers in the queue at time t multiplied by the service time $1/\mu$

$$W_q(t) = \frac{L_q(t)}{\mu}. \tag{3.21}$$

This formula is only valid if the number of servers does not change during the waiting time. When the waiting time is so long that the service rate

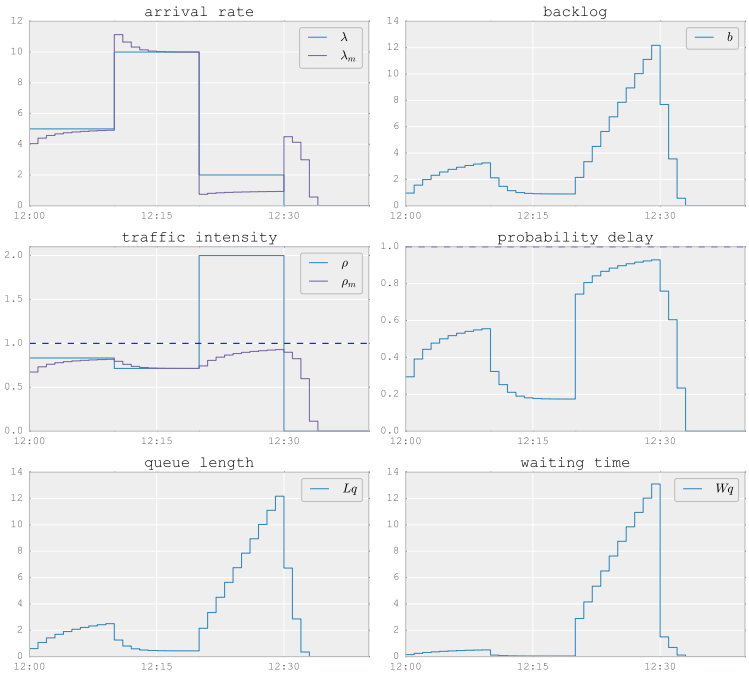


Figure 3.2: Example of the stationary backlog-carryover approach.

changes, the formula for the waiting time needs to be modified. The waiting time of a customer who arrives at the system at time t, depends not only on the state transition probabilities and the number of servers at time t but also on the number of servers after time t. In that case we also need to decide how to deal with customers who are receiving service when the number of servers decreases. The first way is to stop the service and let the customer rejoin the queue. This is called a pre-emptive discipline. The second way is to finish the current service task. This is called an exhaustive discipline.

Green and Soares [30] considered an $M(t)/M/s(t)$ queuing system with an exhaustive discipline and presented algorithms to calculate the waiting time tail probability $P(\mathcal{W}_q(t) > \tau)$ where $\mathcal{W}_q(t)$ is the waiting time experienced by an individual arriving at time t. They derived an exact expression for the waiting time probability for the special case where the number of servers can change at most one time in the interval $(t, t+\tau]$. For the general case where the numbers of servers can change infinitely an algorithm was presented to calculate the tail probability and the upper and lower bound. They also gave a recursion formula to calculate the expected waiting time $E[\mathcal{W}_q]$. For queueing systems with an pre-emptive discipline Ingolfsson et al. presented exact expressions for the waiting time tail probability when the number of servers changes once [43] and when multiple changes occur during the waiting time [42]. Let $v(t)$ be the total number of servers that begin

or stop service in the interval $(t, t+\tau]$. The number of service completions during the interval is a Poisson process with mean

$$m(t, t+\tau) = \int_t^{t+\tau} \mu s(u) du. \tag{3.22}$$

In the case of multiple changes in the number of servers, the expression for the waiting time tail probability is given by [42]

$$P(\mathcal{W}_q(t) > \tau) = \sum_{q=v(t)+1}^{\infty} \big[p_{s(t)+q}(t) \sum_{i=0}^{q-v(t)-1} e^{-m(t,t+\tau)} \frac{m(t, t+\tau)^i}{i!}\big]. \tag{3.23}$$

The expected waiting time $W_q(t)$ is then

$$W_q(t) = E[\mathcal{W}_q(t)] = \int_0^{\infty} \big(1 - P(\mathcal{W}_q(t) \leq \tau)\big) d\tau = \int_0^{\infty} P(\mathcal{W}_q(t) > \tau) d\tau. \tag{3.24}$$

Numerical Integration Approach

The set of differential equations of (3.1) describes the time-varying behavior of the system but except for simple cases this set of an infinite number of equations does not have an analytical solution. A transient solution can be found by numerical integration if the number of equations is finite. This can be achieved by setting a capacity limit m to the number of customers in the system. The value of m is chosen such that the probability that the capacity limit is exceeded is small. In the transient case both the arrival rate and number of servers can vary with time. The equations of (3.1) then become [49]

$$\begin{aligned}
\frac{dp_o(t)}{dt} &= -\lambda(t)p_0(t) + \mu p_1(t), \\
\frac{dp_j(t)}{dt} &= \lambda(t)p_{j-1}(t) - \big(\lambda(t) + j\mu\big)p_j(t) + (j+1)\mu p_{j+1}(t) \quad 1 \leq j < s(t), \\
\frac{dp_j(t)}{dt} &= \lambda(t)p_{j-1}(t) - \big(\lambda(t) + s(t)\mu\big)p_j(t) + s(t)\mu p_{j+1}(t) \qquad j \geq s(t), \\
\frac{dp_m(t)}{dt} &= \lambda(t)p_{m-1}(t) - s(t)\mu p_m(t).
\end{aligned} \tag{3.25}$$

This is a system of ordinary differential equations that can be solved with numerical integration methods such as the Runga-Kutta method [32].

Koopman [50] applied the numerical integration approach to analyse the takeoff and landing queues of aircraft at J. F. Kennedy and LaGuardia airports. He compared the expected queue lengths and the probability of an empty queue for three different models: a transient queueing model with Poisson service, a transient queueing model with fixed service times and a

deterministic flow model. In the transient queueing models the differential equations were solved numerically. The results were insensitive whether the service time was assumed to be fixed or a Poisson one. The deterministic flow model did not capture any stochastic delays. The numerical integration approach on the other hand could describe the queues in both overloaded and non-overloaded situations.

Kolesar et al. [49] developed an iterative method to generate and evaluate schedules of police patrol cars. First steady-state queueing theory is used to generate hourly staffing requirements. Then an integer linear program is solved that satisfies the the constraints such as the length of tours of duty and mealtime breaks while minimizing the number of police cars. The resulting schedule is then evaluated with a time-dependent queueing model that uses the numerical integration approach.

Bookbinder and Martell [7] optimized the allocation of firefighting helicopters among various bases in a region. The occurrence of a fire is assumed to be a non-stationary Poisson process. The solution of the differential equations gives the probability of the number of fires during the day. This result is used as the input for a dynamic programming algorithm that optimizes the helicopter allocations. The available helicopters are allocated such that the maximum expected queue length of all bases is minimized. The queue lengths in each sector are weighed according to the relative fire damage of the sector.

Randomization Technique

The randomization technique [32] is another computational method to solve the set of differential equations in (3.1). The randomization technique approximates the continuous-time Markov chain (CTMC) by a discrete-time Markov chain (DTMC). It is also known as Jensen's method or as uniformization.

As discussed before the goal is to obtain the transition probability vector $\underline{p}(t)$ by solving $\underline{p}'(t) = \underline{p}(t)Q$ where Q contains the elements q_{ij}. The general solution of the ODE is $\underline{p}(t) = \underline{p}(0)e^{Qt}$. The randomization technique decomposes the CTMC into a DTMC $X = \{X_n, n = 0, 1, ...\}$ and a Poisson process by applying a uniformization parameter $\gamma \leq \max(q_{ij})$ [16]. The DTMC can be constructed with a probability transition matrix given by

$$P = I + \frac{1}{\gamma}Q. \tag{3.26}$$

The solution for the transition probability vector of the DTMC is

$$\underline{p}(t) = \underline{p}(0) \sum_{n=0}^{\infty} e^{-\gamma t} \frac{(\gamma t)^n}{n!} P^n. \tag{3.27}$$

Let

$$\underline{p}_n = \underline{p}_{n-1}P, \quad n = 1, 2, ... \tag{3.28}$$

with $\underline{p}_0 = \underline{p}(0)$. Equation (3.27) can then be written recursively as

$$\underline{p}(t) = \sum_{n=0}^{\infty} e^{-\gamma t} \frac{(\gamma t)^n}{n!} \underline{p}_n. \tag{3.29}$$

We see that $\underline{p}_n$ describes the probability distribution of the states of X after n transitions or jumps. The CTMC can thus be represented by a DTMC where n jumps occur according to a Poisson process with rate parameter γ.

Ingolfsson et al. [43] compared the randomization technique with the numerical integration approach and the randomization technique was just as accurate but required only half of the computation time.

Closure Approximation

Closure techniques are used to reduce an infinite system of equations to a finite system by making a closure assumption. A closure assumption describes a relationship between the variables of the system. Let $m(t)$ be the mean number of customers in an $M/M/s$ queueing system then its derivative is

$$m' = \lambda - \mu s + \mu \sum_{n=0}^{s-1} (s - n) p_n. \tag{3.30}$$

The derivative of the variance $v(t)$ of the number of customers in the system is given by

$$v' = \lambda + \mu s - \mu \sum_{n=0}^{s-1} (2m + 1 - 2n)(s - n) p_n. \tag{3.31}$$

The closure assumption made by Rothkopf and Oren [67] expresses p_n in terms of the mean and variance by using the negative binomial distribution. We can approximate p_n as

$$p_n = \binom{r + n - 1}{n} k^r (1 - k)^n, \quad n = 0, 1, 2, ... \tag{3.32}$$

where $k = m/v$ and $r = m^2/(v - m)$. The infinite set of differential equations (3.1) is reduced to two differential equations that describe the transient behavior of the mean and the variance of the number of customers. The system can be solved by integrating the differential equations (3.30) and (3.31) with p_n given by (3.32) using standard numerical integration methods.

Ingolfsson et al. [43] found that the closure method is slower and less accurate than the randomization technique and the numerical integration method but it might be faster for systems larger than the ones they tested.

Infinite-Server Approximation

Equation (3.1) is intractable but by making the assumption of an infinite number of servers it is possible to obtain a solution for the time-dependent behaviour of an $M(t)/M/\infty$ system. The number of busy servers in an infinite-server model (ISA) has a Poisson distribution with mean $m_\infty(t)$. For an $M(t)/M/\infty$ system the mean number of busy servers $m_\infty(t)$ is described by an ordinary differential equation [19]:

$$m'_\infty(t) = \lambda(t) - \mu m_\infty(t). \tag{3.33}$$

In an infinite-server model the number of customers in the system is equal to the number of busy servers [43]. Therefore the number of customers in the system also follows the Poisson distribution. If we solve (3.33) then the transition probabilities $p_n(t)$ can then be determined by using a Poisson probability mass function

$$p_n(t) = \frac{m_\infty(t)^n e^{-m_\infty(t)}}{n!}. \tag{3.34}$$

In their comparison of time-dependent queueing methods Ingolfsson et al. [43] found that when calculating the probability of no delay, the relative error of the ISA was 32% compared to the exact solution. However the ISA performed better than the lagged PSA. On the other hand the lagged PSA was on average 30% faster but in absolute terms the speed difference was small.

Modified-Offered-Load Approximation

In the infinite-server approximation an exact solution is obtained for the system performance at the cost of having an infinite number of servers. Possibly better results can be achieved with the modified-offered-load (MOL) approximation [57]. It consists of two steps. In the first step the infinite-server $M(t)/M/\infty$ model is used to determine the expected number of busy servers $m_\infty(t)$. Then in the second step we apply the PSA with a stationary $M/M/s$ system but with an adjusted arrival rate. Equation (3.6) shows that for an $M/M/s$ system the steady-state probabilities are a function of the offered load λ/μ, i.e. $p_n = f(\lambda/\mu)$. Because for an $M/M/s$ system the expected number of busy servers is equal to λ/μ, the MOL approximation is

$$p_n(t) = f(\lambda(t)/\mu) \approx f(m_\infty(t)). \tag{3.35}$$

This means that the stationary model can used at each time t with a modified arrival rate

$$\lambda_{\text{MOL}}(t) = \mu m_\infty(t). \tag{3.36}$$

Because we apply the PSA the system can never be overloaded. Massey and Whitt [58] compared the MOL approximation with the numerical integration

method for systems with slowly varying arrival rates. They found that the MOL approach is accurate when the probability of delay is small.

3.1.3 Deterministic Fluid Approximations

Overload delays occur when the arrival rate λ is larger than the service rate μs. Stochastic delays occur when $\lambda < \mu s$ because of stochastic variations in the interarrival times and the service times. When a system is overloaded for a substantial amount of time, the fluid approximation works well [29]. In the fluid approximation the flow of customers in a system is modeled as the flow of a fluid through a container. The arrival rate and service rate are then deterministic. The fluid approximation can only estimate queues when a system is overloaded. In underload situations the fluid approximation will never predict any queues. In this section we will describe two approaches for the fluid approximation.

Cumulative Flows

Suppose we divide the time period in intervals $t = 1, 2, \ldots$ and the average arrival rate $\lambda(t)$ and number of servers $s(t)$ in each interval is known. With the deterministic fluid approximation it is assumed that both the arrivals and departures from a queue are evenly spaced in an interval. Queues build up when $\lambda(t) > s(t)\mu$ and decrease when $\lambda(t) < s(t)\mu$ or stay zero if there was no initial queue. The dynamic behaviour of the queue length $L_q(t)$ can be described as follows

$$L_q(t) = \max\{0, L_q(t-1) + \lambda(t) - \mu s(t)\}. \tag{3.37}$$

An alternative way to describe the queue can be obtained by using the cumulative flow diagrams [63, 62]. Let $A(t)$ represent the cumulative number of arrivals and $D(t)$ represent the cumulative number of departures from the queue. The cumulative arrivals equals

$$A(t) = \sum_{t=0}^{t} \lambda(t). \tag{3.38}$$

The cumulative number of queue departures depends on the service rate $\mu s(t)$ but it can never exceed the number of cumulative arrivals at any time

$$D(t) = \min\{A(t), D(t-1) + \mu s(t)\}. \tag{3.39}$$

The cumulative flow diagrams of $A(t)$ and $D(t)$ are plotted in Figure 3.3. The queue length $L_q(t)$ is equal to the vertical distance between the two cumulative diagrams

$$L_q(t) = A(t) - D(t). \tag{3.40}$$

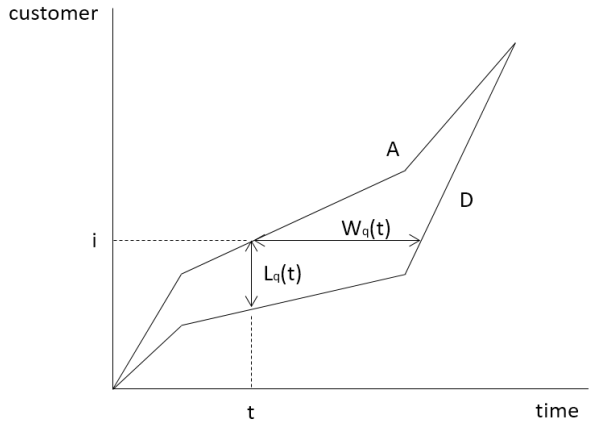


Figure 3.3: Cumulative diagrams.

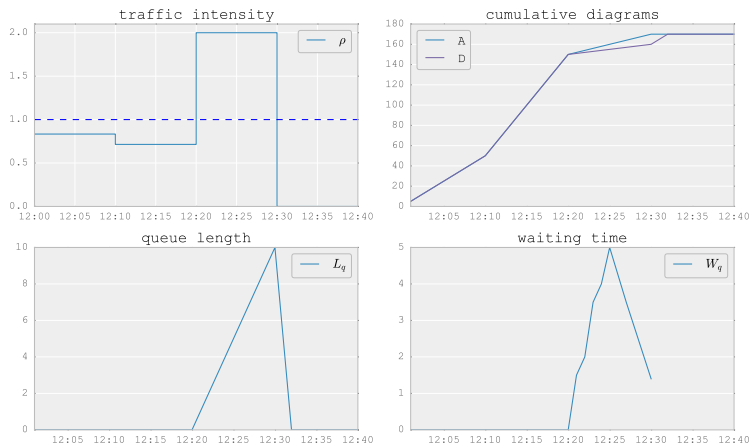


Figure 3.4: Example of the deterministic fluid approximation.

If we assume a FCFS queueing discipline, the waiting time W_q^i experienced by passenger i who arrives at the queue at t is then equal to the horizontal distance between the two cumulative diagrams

$$W_q^i(t) = D^{-1}(i) - A^{-1}(i). \tag{3.41}$$

Newell [63] described various engineering applications of the deterministic fluid model, e.g. queues at traffic lights and at airport baggage claims. Neufville and Odoni [62] used this approach to analyse the flows and queues at airports. For the example queueing system from previous sections the queue length and waiting time are plotted in Figure 3.4. We see that the deterministic method predicts no queues during periods with a traffic intensity smaller than one.

Asymptotic Limit

An alternative approach is to consider the asymptotic limit by scaling the arrival rate and number of servers. The limit is deterministic and it can interpreted as a fluid model. Let $Q(t)$ be the queue length process of an $M(t)/M(t)/s(t)$ queue. The sample paths for $Q(t)$ are the solutions to the functional equation [54]

$$Q(t) = Q(0) + A_1\left(\int_0^t \lambda_r dr\right) + A_2\left(\int_0^t \mu_r \cdot \min\{Q(r), s_r\} dr\right) \tag{3.42}$$

where $A_1(\cdot)$ and $A_2(\cdot)$ are independent Poisson processes. Let $k > 0$ be a multiplier then $Q^k(t)$ is the queue length process of an $M(t)/M(t)/s(t)$ queue with arrival rate $k\lambda(t)$ and number of servers $ks(t)$. Mandelbaum, Massey, and Reiman [54] created a family of associated queue processes $Q^1(t), \ldots, Q^k(t)$ and then determined the asymptotic behaviour of this family as $k \to \infty$. They found that the fluid limit $Q^{(0)}(t) = \lim_{k\to\infty} \frac{1}{k} Q^k(t)$ is the solution to the ordinary differential equation

$$\frac{d}{dt} Q^{(0)}(t) = \lambda(t) - \min\{Q^{(0)}(t), s(t)\}\mu \tag{3.43}$$

Jiménez and Koole [46] studied a call center model with 32 servers. They compared the fluid limit with a simulation model. The queueing system was overloaded in the first half of the time and underloaded in the second half. The fluid approximation compared favourably when $\rho = 1.378$ but performed worse with $\rho = 1.125$. It was also found that the stationary approximation gave reasonable results in the underloaded period. They concluded that the combination of the fluid approximation and the stationary approximation is considerably better than the fluid approximation alone.

3.2 Observation

A queueing model requires as input: the arrival rate $\lambda(t)$, the service time $1/\mu$ and the number of servers $s(t)$. To develop a model we need observation data. The output of the model is the waiting time $W(t)$. To validate the model we also have to measure the true waiting times at Narita immigration. In this section we describe how we measured the data and the results of the observations.

At immigration there are multiple queueing systems: queues and service counters for foreigners, reentry passengers and Japanese passengers (see Figure 1.3). For the foreign passengers there are two clusters of service counters on the left and right side of immigration. Each cluster is also a separate queueing system. In general each passenger type has its own queueing system with a separate input source. However the foreign passengers on the

Table 3.1: Service time for each passenger type in seconds.

	mean	std
Japan	13	7
foreign	63	25
reentry	59	34

right side can use the service counters of the reentry permit holders if these are free. Therefore these two systems have to be analyzed together. In Figure 1.3 we can see that the type of waiting line is different for the Japanese, foreign and reentry passengers. Foreign passengers line up in a snake queue on both sides. When the service counters for foreigners are open on both sides then the foreign passengers can freely choose which side to go to but immigration staff will usually try to direct the passengers to the side which they think is less crowded. Japanese passengers line up in parallel queues. Reentry passengers also line up in parallel queues.

3.2.1 Observation Results

We visited the immigration area of the South Wing of Terminal 1 on five occasions: 2011/2/6, 2011/11/27, 2012/9/9, 2013/4/14 and 2013/4/15. The North Wing immigration area was observed on 2013/10/5. The observation time period was usually from around 13:00 to 16:00. During these times we recorded videos of the arrivals and the queues. The number of arrivals and open service counters were counted manually on the spot.

In this section the results of the observations on 2011/11/27, 2013/4/14 and 2013/4/15 are presented. The measurements of the other days are incomplete. The data from 2011/2/6 was used to determine the service times.

Service Times

The service time is defined as the period from the moment a passenger leaves the queue until the time he leaves the service counter. From the video recordings 132 random samples were taken for Japanese passengers, 105 random samples for foreign passengers and 50 random samples for reentry passengers. The distributions of the service times are shown in Figure 3.5. The mean and standard deviation in seconds are shown in Table 3.1. The average service time for foreigners is similar to the average service time for reentry passengers. The average service time for Japanese passengers is four times shorter.

Brown et al. [9] showed that the service time distribution for call centers tends to be approximately lognormal. We plotted the fitted lognormal

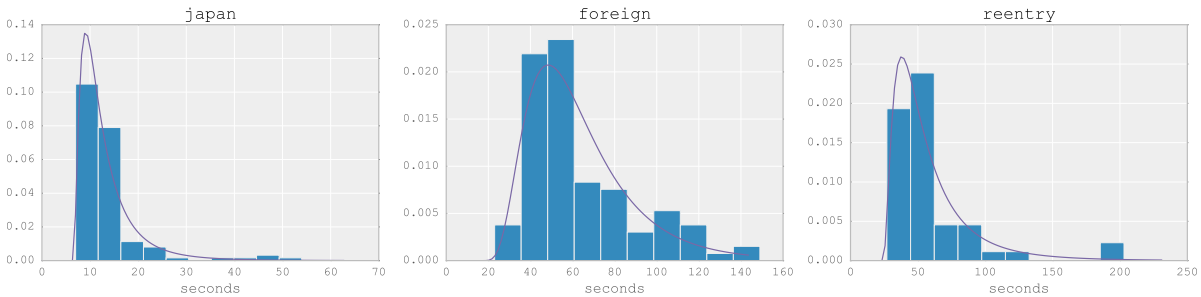


Figure 3.5: Service time distribution per passenger type.

distribution in Figure 3.5. The lognormal distribution follows the observed service time distribution remarkably well for all three passenger types.

Arrivals

The passenger arrivals at the immigration entrances and the arrivals at the tail of the waiting lines of each queueing system were counted either in real-time on the spot or afterwards from the video recordings. There were several difficulties with counting the passengers. During busy periods the number of Japanese arriving at parallel queues was too large to count accurately. The foreign passengers could be counted more easily because there is only one entry for a snake queue. However during peak times the queue could extend far beyond the entry point. The passengers arriving at the tail of the queue were not visible for the observer in real-time or from the video recordings.

The arrival time recording precision is seconds but for most calculations we use the total arrivals per minute and for visualization of the arrivals larger time intervals are more practical. The number of passenger arrivals per 10 minutes are shown in Figure 3.6. For the foreign passengers there are service counters on the left and right side of the immigration area. On 2011/11/27 there were few foreign arrivals on the left side. On 2013/3/14 the foreigner arrivals were reasonably balanced between the left and right side. Overall we can observe large fluctuations in the number of arrivals. The ratio of foreign and Japanese passengers was also very different for each day. For the three observation periods the average percentage of aliens (foreign plus reentry passengers) was respectively 0.40, 0.66 and 0.48.

Service Counters

We counted the number of open service counters during the observation periods. Because it was difficult to count from the video recordings we had one observer on the spot who registered the number of service counters that were in use for each queueing system for every couple of minutes.

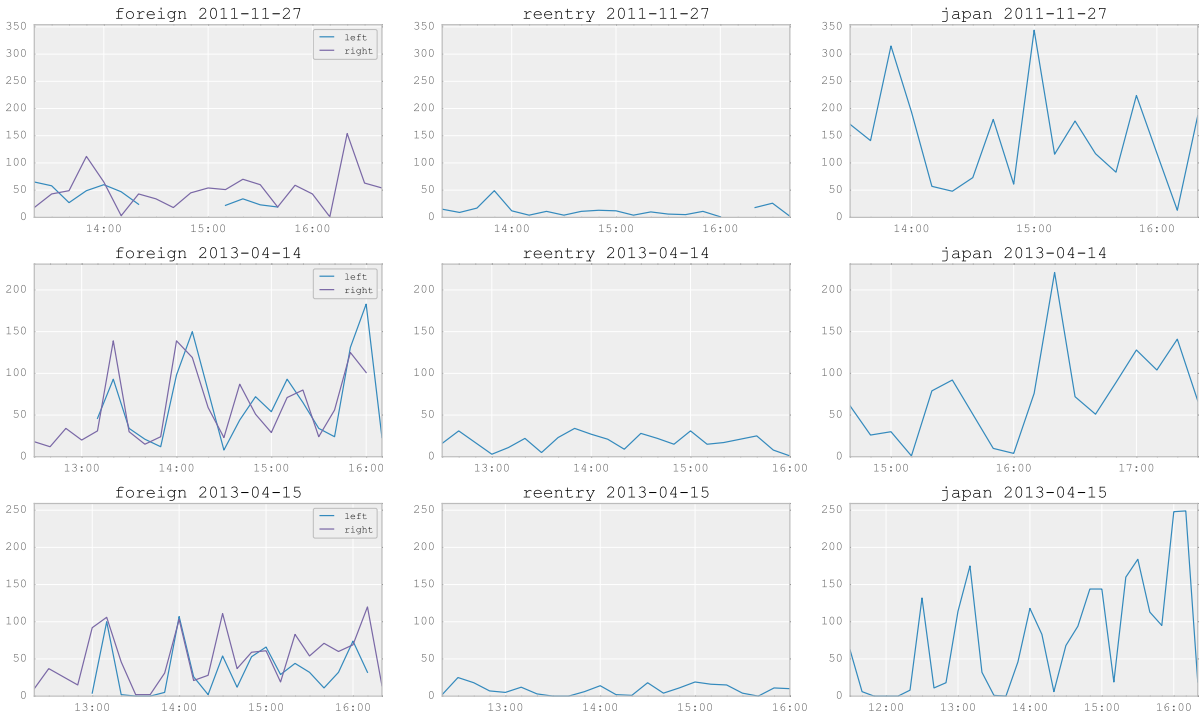


Figure 3.6: Observed number of arrivals per 10 minutes.

The number of open service counters are plotted in Figure 3.7. We see that the service counters are adjusted constantly. The service counters for foreign passengers on the right side were always open while the left side was only used when the capacity on the right side was insufficient. On average there were 9 service counters for foreign passengers, 4 on the left side and 5 on the right side. 4 service counters were available for reentry passengers. Japanese passengers could line up in front of 5 service counters on average.

Queue Lengths

The queues were recorded on video to measure the waiting times and the queue lengths. The queue lengths were determined for every 5-10 minutes. Counting the number of waiting passengers could not be done 100% accurately because of the video resolution and because some passengers were outside of the view or blocked by other passengers. Figure 3.8 shows a screenshot of a video recording that was used to determine the number of Japanese passengers in front of the service counters. In this screenshot is it also difficult to distinguish where the Japanese queues end and the reentry queues start.

The queue lengths on 2011/11/27 are shown in Figure 3.9. We see that the foreign queue at the right side is at the worst point twice as long as on

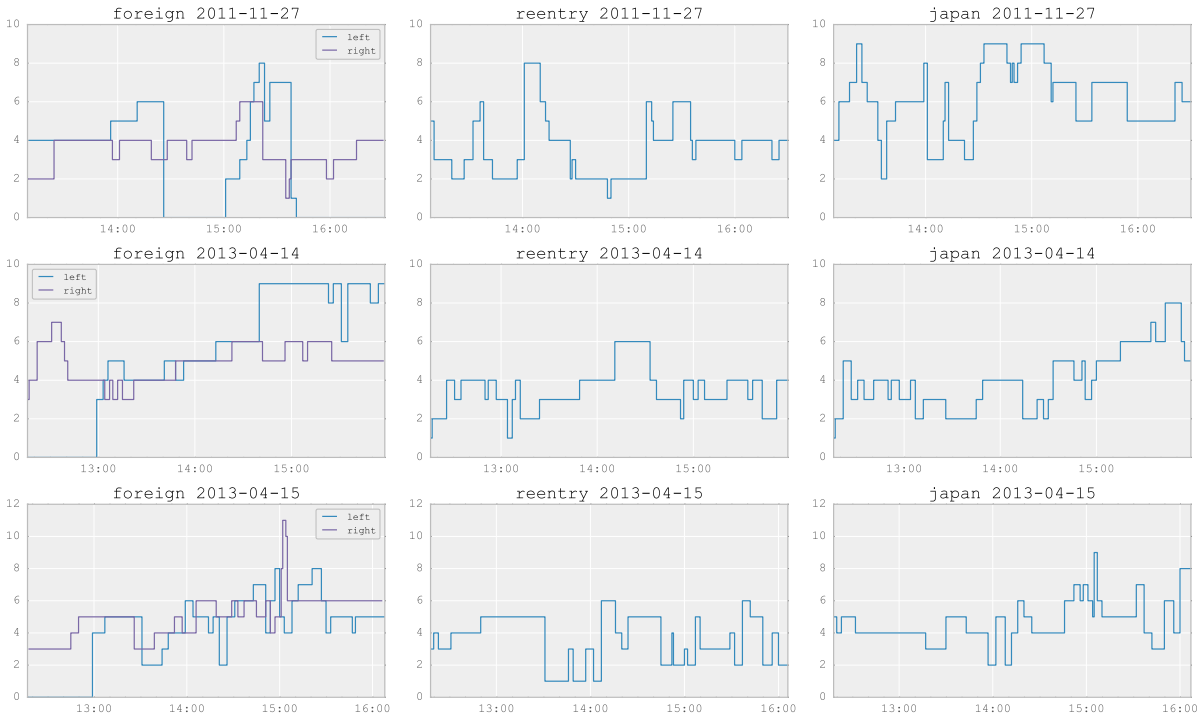


Figure 3.7: Observed number of open service counters.

Figure 3.8: Screenshot of the Japanese queue.

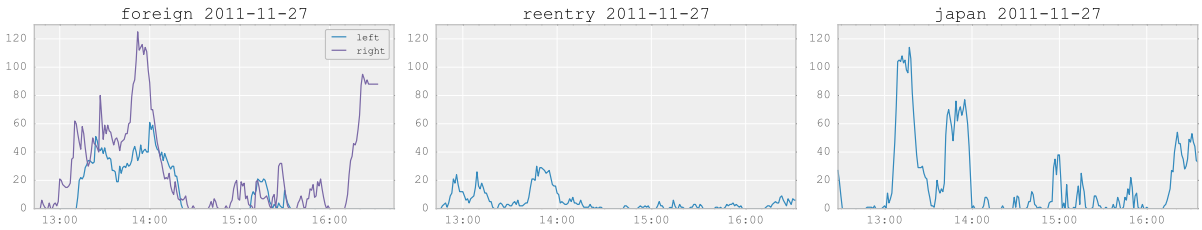


Figure 3.9: Observed queue lengths.

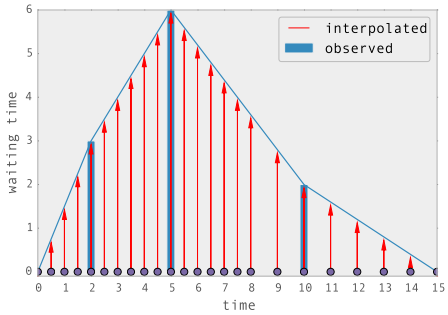


Figure 3.10: Interpolation of waiting times.

the left side.

Waiting Times

The waiting times were measured for random passengers from the video recordings. We need to interpolate these observed waiting times to estimate the waiting time of the other passengers in the queue. Figure 3.10 shows how the waiting times are interpolated. The blue bars represent the observed waiting times for the passengers arriving at $t = 2, 5, 10$. We make the assumption that the waiting time changes linearly between each observed waiting time. The observed arrival times of all passengers are indicated by the purple dots. After linear interpolation we can estimate the waiting time of each arrived passenger as shown by the red arrows.

The waiting times on three observation days are shown in Figure 3.11. These waiting times represent the time spent in the queue for a passenger arriving at time t. The waiting times for foreign passengers can differ significantly between the left and right side. At the same time of the day there can be a difference of more than 10 minutes. The maximum waiting time experienced by any foreign passenger is 25 minutes. The passengers with a reentry permit have a much shorter waiting time than the other foreigners. The waiting times for Japanese passengers are very short. The average waiting times during the observation period are shown in Table 3.2.

Because the interpolated waiting times per passenger are known, it is

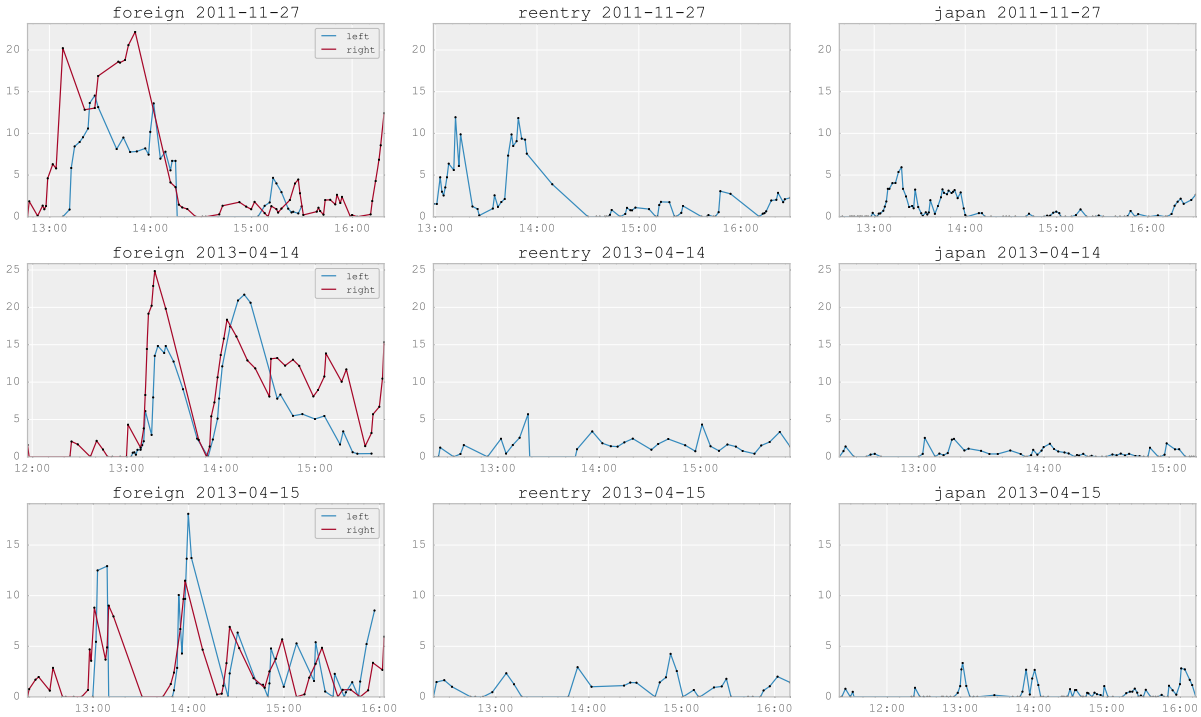


Figure 3.11: Observed waiting times.

Table 3.2: Observed average waiting times in minutes.

	foreign left	right	reentry right	Japan right	pax all
2011-11-27	6.8	7.2	2.8	0.7	2.9
2013-04-14	8.4	10.8	1.4	0.3	5.3
2013-04-15	4.6	3.4	0.9	0.7	2.0

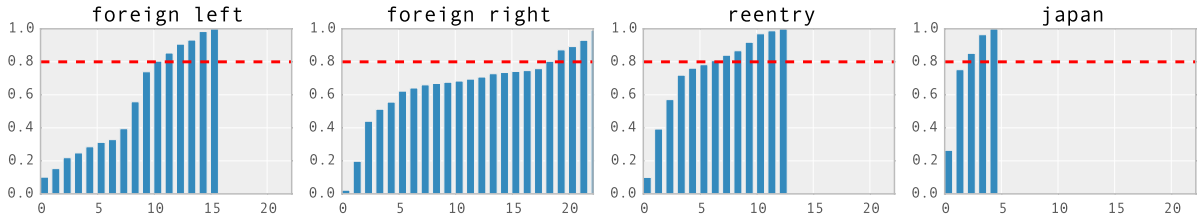


Figure 3.12: Cumulative histograms of the observed waiting times on 2011/11/27.

possible to draw a distribution for the waiting time. Figure 3.12 shows the cumulative histograms for the waiting times on 2011/11/27. The red lines indicate the 80% level. For the foreigners on the left side 80% has a waiting time less than 10 minutes. On the right side the waiting time is 18 minutes at this level. For the reentry and Japanese passengers the waiting times are respectively at most 6 and 2 minutes for 80% of the people.

3.3 Immigration Queueing Models

Ingolfsson et al. [43] did a comparison of seven queueing theory methods and compared the computational speed and accuracy for calculating the waiting time probability. The seven methods were: numerical integration approach (NUM), randomization method (RND), closure approximation (CLS), infinite server approximation (ISA), modified offered load approximation (MOL), effective arrival rate approximation (EAR) and lagged stationary approximation (LST). For their experimental design they used a sinusoidal arrival rate function and a discretized sinusoidal function for the number of servers. The waiting time probabilities were calculated at 5-minute intervals for 640 different cases. They assumed that the numerical integration approach is the exact solution. For a target waiting time of 0 minutes, the ranking in order of decreasing accuracy was RND, MOL/EAR, ISA, LST, CLS. For a target waiting time larger than zero the ranking became RND, CLS, ISA, MOL/EAR, LST.

In this section we compare the numerical integration approach, the deterministic fluid approximation and the SBC approach. The last two methods were not studied by Ingolfsson et al. We also don't use artificial functions for the number of arrivals and servers. And instead of assuming that the numerical integration approach is exact, we compare the results with the observed queue lengths and waiting times of three observation periods.

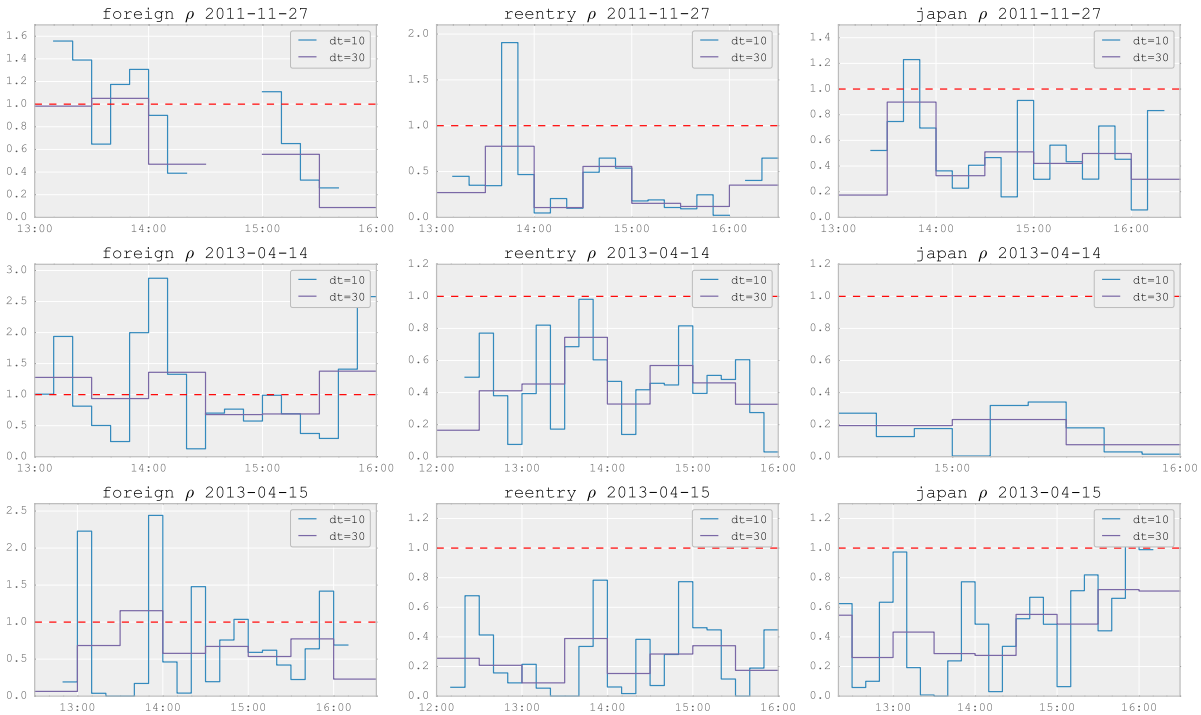


Figure 3.13: Observed traffic intensities.

3.3.1 Traffic Intensity

From the literature review we know that models based on the stationary queueing theory can only be applied if the traffic intensity ρ is less than one. The exception is the SBC approximation which can be used for all values of ρ. The numerical integration methods and the deterministic fluid approximations can also be used in overload situations. But fluid approximations cannot capture stochastic delays. The value of the traffic intensity depends on the length of the time interval. In general the traffic intensity is lower for longer time intervals. We plot the traffic intensity for a interval length of 10 minutes and 30 minutes (Figure 3.13). Note that we only calculate the traffic intensity for the left foreign queueing system because for the right side we don't know the exact service rate because the foreign passengers can also use the reentry counters if these are free. For the left foreign counters the traffic intensity is always high and often exceeds one for both interval lengths. For the reentry queueing system the 30-minute interval length does not give any overload situations but for the 10-minute interval length the traffic intensity exceeds one in one period. A similar conclusion can be drawn for the Japanese queueing system.

In Table 3.3 the percentage of intervals that are overloaded are shown for interval lengths of 1, 10 and 30 minutes. The foreign queueing system is

Table 3.3: Percentage of overloaded intervals for time intervals of 1, 10 and 30 minutes.

		foreign left	reentry	japan
2011-11-27	1	0.45	0.15	0.17
	10	0.36	0.05	0.06
	30	0.20	0.00	0.00
2013-04-14	1	0.32	0.16	0.02
	10	0.39	0.00	0.00
	30	0.50	0.00	0.00
2013-04-15	1	0.21	0.08	0.14
	10	0.25	0.00	0.04
	30	0.12	0.00	0.00

overloaded during a substantial amount of time for all three interval lengths. The reentry queueing system is frequently overloaded if a time interval of 1 minute is used and sporadically overloaded if a 10-minute interval length is used. Overload occurs for the Japanese queueing system with interval lengths of 1 and 10 minutes.

3.3.2 Service Time

In queueing models the service time distribution is often assumed to be exponential. Exponential service times allow us to analytically solve the queueing model. The exponential distribution has one parameter: the mean of the service times. According to Green, Kolesar, and Whitt [29] the most important parameter for any service time distribution is the mean and the second most important parameter is the squared coefficient of variation (SCV). The SCV is defined as the square of the variance divided by the mean. For an exponential distribution the SCV is equal to 1. If the SCV is smaller than 1 the exponential distribution will be too conservative. In general the exponential-distribution assumption for the service times is a reasonable approximation if the SCV is smaller than 2. The SCV of the Japanese service times is 0.31, 0.165 for foreign service times and 0.34 for reentry service times. However the exponential-distribution assumption can still be adequate for small SCV. Kolesar [48] analyzed the congestion at ATM's. The measured service times were not exponentially distributed and the SCV was only 0.25. Simulations showed that there was no practical difference between a queueing model with the empirical service time distribution and an exponential service time distribution.

We found that using the average service time did not give optimal results for the queueing models of the foreign and reentry passengers. We compared the error between the observed waiting times and the estimated waiting

times when using service times smaller and larger than the average. A service time of 90% of the average gives the best results. For the Japanese passengers the average of the observed service times gives good results.

3.3.3 Queueing Models

Based on the percentages of overload time we can conclude that for the foreign queueing systems we can only use queueing theory methods that can deal with overload. For the reentry and Japanese queueing systems we could use the steady-state queueing theory with the longer interval lengths but with an interval length of 1 minute there will also be overload. Therefore we will analyse the immigration queuing systems with three models that can deal with overload: the stationary backlog-carryover approach (SBC), the numerical integration approach (NUM) and the deterministic fluid approximation with cumulative flows (DET).

We choose 1 minute as the interval length for higher accuracy. Also this is an appropriate interval length for analysing the foreign and reentry queues with SBC. A shorter interval length however increases the computation time. The computation time for a 3 hour period with 1 minute intervals is 15 ms for DET, 87 ms for SBC and 1 minute for NUM. The numerical integration approach is much slower than the other two models.

The input for each model is the observed arrival rate, the observed number of open counters and the service time. For the foreign and reentry queueing systems we use a service time of 90% of the average while for the Japanese queueing system the service time is the average observed value.

All queueing systems are modeled as independent systems except for the right foreign system. The foreign passengers on the right side can also use the reentry service counters if these are free. We adjust the service rate for the right foreign system as follows. First the queues for the reentry system is calculated. Then at each time interval we check if the reentry service counters are open but there is no queue. When there is no queue we add the available service counters to the right foreign system. However we cannot add the full service capacity. By minimizing the error between the observed and estimated waiting times it was found that we can add only 50%-80% of the available service capacity. We can interpret this factor as the efficiency of using shared service counters. With the additional service counters we then calculate the queues for the right foreign system. Figure 3.14 shows the difference between the estimated queue lengths with and without the additional service capacity for an efficiency factor of 80%. We will use an average efficiency factor of 60% for the models in the remainder of this chapter.

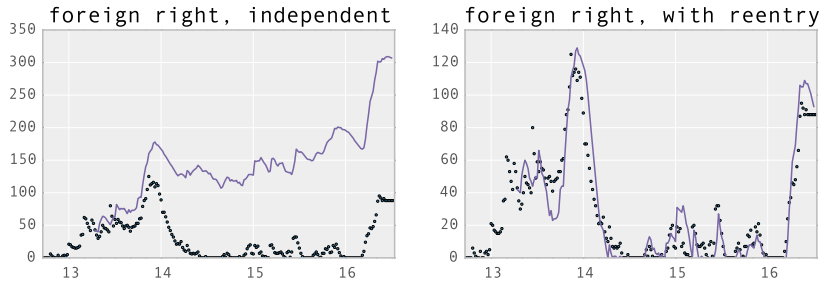


Figure 3.14: Observed and estimated queue lengths for the right foreign queueing system without (left) and with (right) the available service capacity from the reentry queueing system.

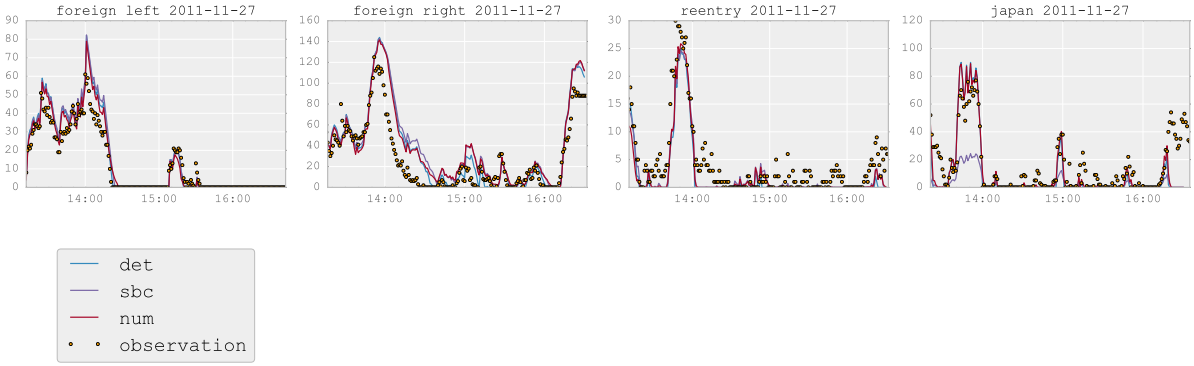


Figure 3.15: Estimated queue lengths compared with observed queue lengths on 2011/11/27.

3.3.4 Model Results

Queue Lengths

Figure 3.15 shows the estimated queue lengths for the three queueing models with the observation data from 2011/11/27. All three models give almost identical results and follow the trend of the observed queues quite well except for SBC with the Japanese system. It is not unexpected that SBC performs poorly for the Japanese system because the interval length is longer than optimal for this model. The optimal interval length would be equal to the average Japanese service time, i.e. 13 seconds instead of 1 minute.

Waiting Times

Figure 3.16 shows the observed and estimated waiting times for three observation days. In general SBC gives higher maximum waiting times than

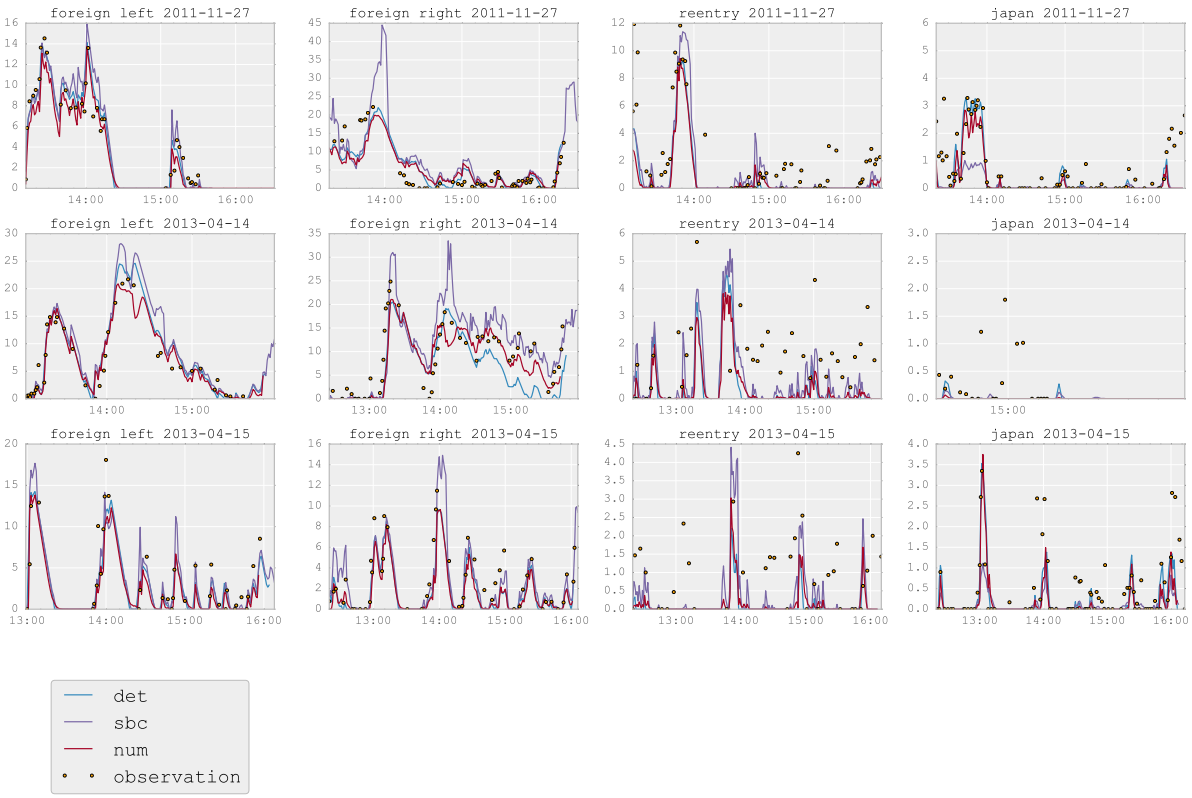


Figure 3.16: Estimated waiting times compared with observed waiting times on 2011/11/27, 2013/4/14 and 2013/4/15.

DET and NUM. For the left foreign system all three models fit the observation data well. For the right foreign system SBC tends to produce longer waiting times than observed while DET gives lower waiting times than observed on 2013/4/14. This is also influenced by the estimation of the free reentry service counters. All three models give good results for the reentry system. For the Japanese queues on 2011/11/27 and 2013/4/15 DET and NUM give good results. SBC however produces much lower peaks than the observed values because of the inappropriate interval length as explained in the previous section. Note that for 2011/11/27 the queue lengths estimated by all three methods were very similar but the estimated waiting times can differ more significantly.

3.3.5 Combining Foreign and Reentry Service

In this section we investigate two scenarios where we combine the left foreign, right foreign and reentry service counters. The current setup where

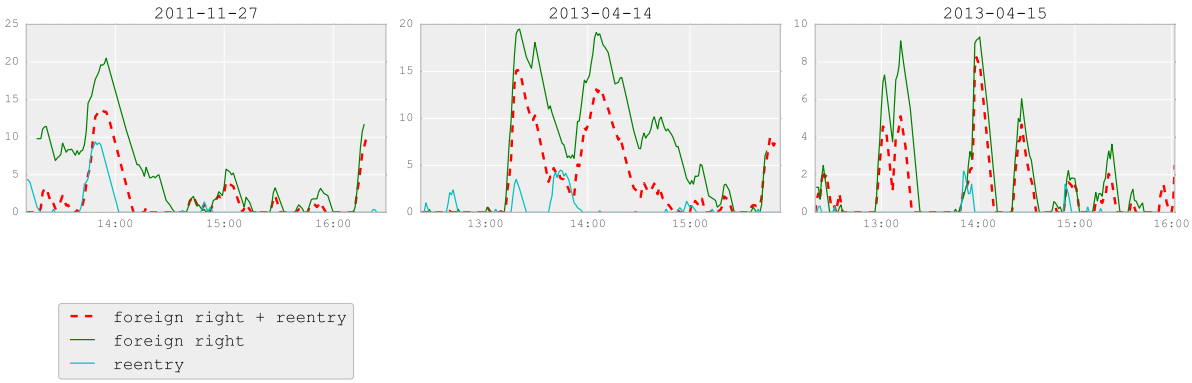


Figure 3.17: Estimated waiting times for the combination of right foreign and reentry service counters.

foreigners can use reentry counters when these are free, has an efficiency of 0.5 to 0.8. Without a priority queueing system for reentry passengers the full capacity can be utilized. The effect of each scenario is calculated with the deterministic fluid approximation. Because the average service time for the reentry passengers is almost the same as for the other foreign passengers we can simply add the arrival rates and service counters.

In the first scenario the foreign service counters are still split between left and right side but the reentry service counters are merged with the right foreign counters. The resulting waiting times are shown in Figure 3.17. The maximum waiting time goes down from 21 minutes to 14 minutes on 2011/11/27, down from 20 to 15 minutes on 2013/4/14, and down from 10 to 8 minutes on 2013/4/15. Of course this comes at the cost of the reentry passengers whose original maximum waiting times were only 10, 5 and 3 minutes respectively.

In the second scenario we merge both left and right foreign service counters and the reentry service counters into one service system for all alien passengers. The resulting waiting times are shown in Figure 3.18. Here we see that this scenario is also very beneficial for the foreign passengers who were originally on the left side. In this scenario the maximum waiting times drop from 21 to 10 minutes on 2011/11/2, from 25 to 15 minutes on 2013/4/14, and from 14 to 8 minutes on 2013/4/15. We can conclude that removing the priority of the reentry passengers has a large impact on the maximum waiting time for the foreign passengers.

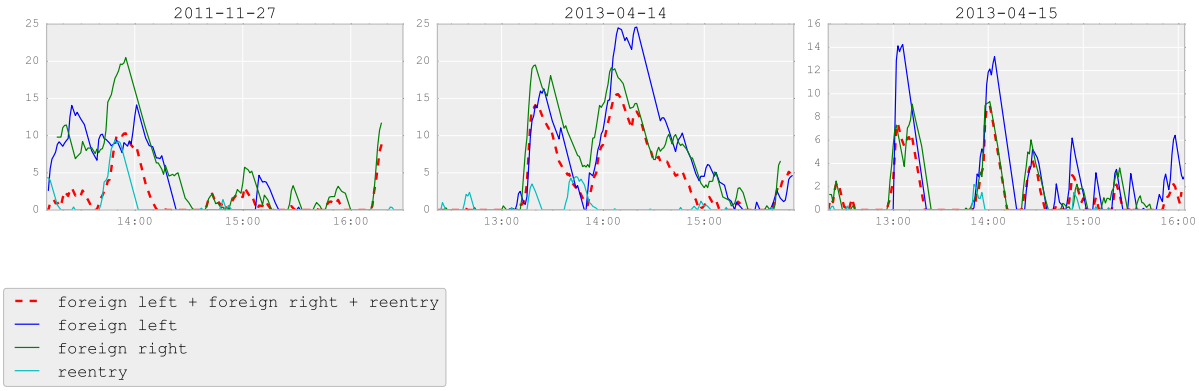


Figure 3.18: Estimated waiting times for the combination of left foreign, right foreign and reentry service counters.

3.4 Processing Times

Managers at Narita Airport immigration have stated that they wish to reduce the waiting times to the same levels as Incheon Airport. In this section we will compare the processing times of Narita Airport immigration and Incheon Airport immigration.

Incheon Airport Immigration

According to Incheon Airport's brochure [41] their immigration service is "the world's fastest, most convenient." More specifically it is stated that "The average immigration processing time for arriving passengers at Incheon Airport is only 12 minutes compared to 45 minutes internationally, while departure processing takes only 19 minutes as opposed to 60 minutes worldwide." The international average processing time refers to the "ICAO Service Target Average."

If we look at the the ICAO document [66] where the service target average is specified, we find that the international processing times are not actual worldwide averages but only recommendations for airports. For the departing passengers the recommendation is "a total time period of 60 minutes for the completion of departure formalities for all passengers [...] from the time of the passenger's presenting himself at the first processing point at the airport [...] to the scheduled time of his flight departure." For the arriving passengers the recommendation is "clearance within 45 minutes of disembarkation from the aircraft of all passengers." In other words the processing time recommendations do not refer to the time at immigration only but the total time it takes for a passenger to go through all service facilities.

Table 3.4: Average processing times at Incheon Airport.

	2005	2010	2012
departure	29	18	19
arrival	20	13	12

According to Incheon Airport they have developed the world's fastest immigration service by [41] "harnessing the latest advances in IT and biotechnology." More information about the systems deployed at Incheon Airport to reduce the processing times is given in an ICAO working paper [65]. Incheon Airport utilized two IT systems: the Congestion Relief System and the U-Immigration System. The Congestion Relief System was introduced in 2004 with the purpose of easing congestion in the passenger terminals. This was achieved by forecasting the passenger flows two hours in advance and by deploying more personnel to bottlenecks in the traffic flow. Forecasting was done by the statistics-based Passenger Forecasting System that predicted the total number of passengers per hour. The system provided the information for flexible personnel deployment with small teams that were placed at congestion points. This resulted in more personnel at peak times. The U-immigration System is an automatic immigration service and it was introduced in 2008. It was used by 14% of the departing passengers in 2009 and it reduced processing time up to 60%. Overall the introduction of the IT systems resulted in 30% personnel savings and a 40% reduction in processing time.

The Incheon immigration processing times were measured twice a year: once in the low season and once in the high season. The average measured times in minutes are shown in Table 3.4. We see that after the introduction of the IT systems the immigration processing times reduced greatly for both arrival and departure. It is however unknown how the processing times were measured. It was not stated whether the measurements were done on peak days and during which hours of the day. We also don't know how the passengers were selected and if it included both Koreans and foreigners. Most importantly it is not clear how they defined the immigration processing time, whether it refers strictly to only the immigration service or whether it also encompasses all other arrival and departure services before and after immigration as in the ICAO definition.

Narita Airport Processing Times

We define the processing time as the total time from arriving at the airport to leaving immigration, i.e. the sum of the walking time, the waiting time

Table 3.5: Average processing times at Narita Airport.

	foreign	reentry	Japan	all
2011-11-27	18	14	11	13
2013-04-14	21	12	10	16
2013-04-15	15	12	10	12

and the service time at immigration,

$$T_{process} = T_{walk} + T_{wait} + T_{service}. \tag{3.44}$$

Using the observation data we can estimate the waiting time and the service time of each passenger. We don't have data of the walking time of individual passengers. Therefore we estimate the average walking time from each gate by using the walking speed distribution as described in the arrival forecasting chapter. The average walking time on a day varies between 9 and 10.5 minutes. The observed average waiting times were shown in Table 3.2. The average waiting times of all passengers during the three observation periods were 2.9, 5.3 and 2.0 minutes. The average service times for Japanese, foreign and reentry passengers are respectively 0.22, 1.1 and 0.98 minutes (Table 3.1). If these data are combined then the average processing times have the values as shown in Table 3.5.

The average processing time of all passengers combined ranges from 12 to 16 minutes. The average processing time for arrivals at Incheon Airport is 12 minutes. One could conclude that that Incheon Airport immigration is more efficient. However we can see that the results can be easily manipulated by selecting different days. Also the largest portion of the processing time consists of 9 to 10 minutes walking. We cannot draw any conclusions about the performance of the immigration service only because we don't know the average walking times and waiting times at Incheon airport.

3.5 Conclusion

In this chapter we have reviewed multiple queueing approaches. An important distinction is that some approaches can deal with overload and other approaches cannot. A queueing system is overloaded if the traffic intensity is larger than one. From observation data the traffic intensities were determined for the queueing systems at Narita immigration and we found that the foreign queueing system is frequently overloaded. Overload also occurs for the reentry and Japanese queueing systems when a 1-minute time interval is used. We have implemented three queueing models that can cope with overload: the numerical integration of ODE, the stationary backlog-carryover approach and the deterministic fluid model. These three models have not

been compared in the literature before. The estimated waiting times from the three models give good results compared to the observed waiting times. Even though the deterministic model cannot capture stochastic delays in underload situations while the other two methods can, the results with a 1-minute interval are just as good for the foreign queues and even for the reentry and Japanese queues. The deterministic fluid model requires the shortest computation time which makes it our preferred method to calculate the waiting time performance of a queueing system. Finally we have compared the total processing timse from arriving at the gate to leaving immigration with the processing times published by Incheon Airport. Because it is unknown how the processing time was exactly defined and measured by Incheon Airport, we cannot conclude if Narita immigration performs better or worse.

Chapter 4

Staffing

Staff planning consists of four steps [15]. First the customer demand is forecasted. Based on the demand the staffing requirements, i.e. then number of staff for each interval of the day, are determined. The third step is shift scheduling where the number of employees are determined while taking shift requirements into account. The last step is rostering to determine the employee assignments to each shift. In real-time the staffing levels can be adjusted by calling additional employees who are doing other jobs. We will deal only with setting the staffing requirements.

The number of staff impacts the cost and the service level. We want to balance service quality and staffing cost. This can be viewed as an optimization problem with the objective to find the number of staff such that the sum of the staffing cost and the cost of waiting is minimized. In practice the number of staff not is determined through optimization [8]. In the staffing literature it is common [15] to view staffing as a constraint satisfaction problem with the objective to find the least number of staff while meeting a target service level requirement. The reasons for using the constraint satisfaction approach are the difficulty of quantifying the waiting cost and the custom of using service level agreements in call centers. Common performance metrics for call centers are the service level and the expected waiting time. The service level $P(W < \tau) \geq \alpha$ specifies the percentage of customers α that wait at most a target waiting time τ. In hospital emergency departments it is common to use the probability that the waiting time exceeds a given maximum waiting time, and the length of stay [15]. The input for the staffing model are the arrival forecast and the service level requirement, and the output is the staffing function $s(t)$.

First we review the staffing literature for different types of arrival rate input. In the airport literature there is a lack of staffing models that include uncertain demand [29, 52]. Our objective is to develop a staffing model that can deal with uncertainty in the delay. We assess the performance of the deterministic staffing model when the delay is uncertain. Then we extend

the deterministic model by converting the deterministic staffing function into a staff probability matrix. We propose a quantile-based solution to set the staffing requirements. We will also assess the performance of the square-root staffing formula for uncertain demand and we determine the best multiplication factor to the meet daily service level requirement. Finally we develop an iterative algorithm to meet the service level at each time interval.

4.1 Literature Review

The literature on staffing in call centers is large [24, 1]. Green, Kolesar, and Whitt [29] reviewed staffing methods for time-varying arrivals and discussed mainly simple heuristics. Defraeye and Nieuwenhuyse [15] provided an overview of staffing and scheduling approaches with non-stationary demand and included other application fields such as hospital emergency departments. In this section we will discuss staffing methods categorized by the type of demand: a constant arrival rate, non-stationary arrival rates and uncertain arrival rates.

4.1.1 Staffing with Constant Arrival Rate

In this section we will discuss several methods for staffing with a constant arrival rate. First we treat constraint satisfaction approaches and then optimization approaches.

Constraint Satisfaction Approaches

Deterministic Approximation

If we apply a naive deterministic approximation, with a constant time between arrivals of $1/\lambda$ and a constant service time of $1/\mu$, then the optimal staffing level equals the offered load $r = \lambda/\mu$ [21, 25, 29]

$$s = r = \frac{\lambda}{\mu}. \tag{4.1}$$

This simple approach is effective for large systems with customer abandonments and if the performance requirement is not very high, e.g. a delay probability of about 0.5 [29]. Without abandonment the staffing level must always be above the offered load [21].

Whitt [80, 81] derived a fluid approximation for the steady-state behaviour of the $M/GI/s + GI$ model. An important feature of the model is customer abandonments because abandonments stabilize the queue. If the system is overloaded ($\lambda > \mu s$) without abandonments then the system would not reach steady-state. The fluid approximation for the abandonment

rate B is the arrival rate minus the service rate

$$B = [\lambda - \mu s]^+ \tag{4.2}$$

where $y^+ = \max\{0, y\}$. The abandonment probability is

$$P(Ab) = \frac{B}{\lambda} = \frac{[\lambda - \mu s]^+}{\lambda}. \tag{4.3}$$

Customers who do not abandon, wait a length of time w. Let $F(w)$ be the time-to-abandon cdf then

$$F(w) = P(Ab). \tag{4.4}$$

If the service level requirement is moderate, e.g. $P(w < \tau) \geq 0.8$, and the abandonment requirement is $P(Ab) \leq \beta$ then the optimal staffing level is [29]

$$s^* = \frac{\lambda(1 - x^*)}{\mu}, \qquad \text{where } x^* = \min\{F(\tau), \beta\}. \tag{4.5}$$

Infinite-Server Approximation

If the number of busy servers rarely reaches the maximum capacity or if there are customer abandonments then it might be appropriate to approximate the system by a system with an infinite number of servers [25]. Infinite-server approximations are discussed in [63, 25, 82]. In an infinite-server system the customers do not interact and are thus independent. The arriving customers are served immediately which means that the number of customers in the system equals the number of busy servers. In an infinite-server system with Poisson arrivals, the number of customers in the system N is Poisson with parameter equal to the offered load r

$$P(N = k) = \frac{r^k e^{-r}}{k!}. \tag{4.6}$$

The probability of delay for a system with a finite number of servers is then approximated with

$$P(W > 0) \approx P(N \geq s) = 1 - \sum_{k=0}^{s-1} \frac{r^k e^{-r}}{k!}. \tag{4.7}$$

If the target delay probability α is specified then we can find the optimal number of servers s such that

$$P(N \geq s) \leq \alpha < P(N \geq s - 1). \tag{4.8}$$

Normal Approximation

For large values of r the Poisson distribution of N can be approximated by the normal distribution with mean μ and variance σ^2 [22]. If we standardize then $(N-\mu)/\sigma$ is normally distributed with mean 0 and standard deviation 1. Because the original distribution is Poisson, the normal approximation has mean and variance equal to the offered load, i.e. $\mu = r$ and $\sigma = \sqrt{r}$. The probability of delay can then be written as [29]

$$\begin{aligned} P(W > 0) &\approx P(N \geq s) = 1 - P(N < s) \\ &= 1 - P\left(\frac{N-r}{\sqrt{r}} < \frac{s-r}{\sqrt{r}}\right) \\ &\approx 1 - \Phi(\beta) \end{aligned} \tag{4.9}$$

where Φ is the cdf of the normal distribution with mean 0 and variance 1, and β is a quality of service parameter. Given a target delay probability α, the optimal number of servers can then be determined with a square-root formula [25, 45]

$$s = \mu + z_\alpha \sigma = r + \beta\sqrt{r} \tag{4.10}$$

where z_α is the z-score for $1-\alpha$. Thus β is specified by

$$\alpha = 1 - \Phi(\beta). \tag{4.11}$$

A simple thumb of rule [29] is to set $\beta = 2$ which is equivalent to a probability of delay equal to 0.02. We can interpret the square-root staffing formula as the offered load plus a safety staffing part. The safety staffing part is necessary to deal with stochastic variability.

Heavy-Traffic Limit

Halfin and Whitt [33] found another square-root-staffing formula by considering simultaneously the heavy-traffic limit $\lambda \to \infty$ and the infinite-server limit $s \to \infty$ while keeping the service time $1/\mu$ fixed for an $M/M/s$ queue [29]

$$\frac{s-r}{\sqrt{r}} \to \beta \tag{4.12}$$

where $\beta \in (0, \infty)$. In this limit the probability of delay becomes

$$P(W > 0) \approx HW(\beta) = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1} \tag{4.13}$$

where Φ is the cdf and ϕ is the pdf of the standard normal distribution. $HW(\beta)$ is called the Halfin-Whitt delay function. The asymptotic result in

(4.12) suggests a square-root formula $s \approx r + \beta\sqrt{r}$ where we determine β for given α with the Halfin-Whitt delay function

$$\alpha = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}. \tag{4.14}$$

For small values of α the normal approximation and the heavy-traffic limit approach give similar results but the asymptotic approach is uniformly more accurate [8].

Steady-State Queueing Model

A common staffing method is the smallest staffing level (SSL) approach [26, 29, 15]. The SSL uses a steady-state queueing model to determine the performance for different number of servers and then selects the smallest number of servers for which the performance requirement is met. Suppose the performance required is defined by the service level, i.e. the probability that α percent of the customers are served within τ minutes. The optimal number of staff is then determined iteratively [15]

$$s = \text{argmin}\{k \in \mathbb{N} : P(W_k > \tau) < \alpha\} \tag{4.15}$$

where W_k is the waiting time with k number of servers. For an $M/M/s$ system there exists an analytical expression for the service level (see equation (3.10) in the queueing models chapter).

Optimization Approaches

Deterministic Approximation

Grassmann [25] discussed three models to maximize the total profit: a deterministic model, an infinite-server model and a steady-state model. The total profit per time unit is defined as

$$T(s) = mB + gE - wQ - cs \tag{4.16}$$

where m is the revenue from sales per time unit, B is the number of busy servers, g is the revenue per customer, E is the number of customers leaving, Q is the number of customers waiting, w is the waiting cost per customer, and c is the maintenance cost per time unit.

For the deterministic model the queue length at time t is $t(\lambda - s\mu)$ if s less than the offered load r and the queue length is 0 otherwise. The total profit is then

$$\begin{aligned} T(s) &= ms + g\mu s - wt(\lambda - s\mu)/2 - cs && s < r, \\ T(s) &= mr + g\mu r - cs && s \geq r. \end{aligned} \tag{4.17}$$

However if $s < r$ then the queue length will increase forever. Therefore we can consider only $s \geq r$. Total profit maximization or cost minimization is achieved with staffing to the offered load $s = r = \lambda/\mu$.

Infinite-Server Approximation

Grassmann [25] developed a simple stochastic model where the customers being served are assumed to be independent. The infinite-server model is used for the approximation of the number of customers in the system N. The queue length is $N - s$ if $s \leq N$ and 0 if $s \geq N$. The total profit is then

$$\begin{aligned} T(s) &= ms + g\mu s - w(N - s) - cs \qquad && s \leq N, \\ T(s) &= mN + g\mu N - cs && s \geq N. \end{aligned} \tag{4.18}$$

The expected profit $E[T(s)]$ is maximized if $E[T(s)] > E[T(s-1)]$ and $E[T(s)] > E[T(s+1)]$. Let $D(s) = T(s+1) - T(s)$ be the marginal profit then equivalently the expected total profit is maximized if $E[D(s)] \leq 0 \leq E[D(s-1)]$. The marginal profit can be written as

$$\begin{aligned} D(s) &= m + g\mu + w - c \qquad && s < N, \\ D(s) &= -c && s \geq N. \end{aligned} \tag{4.19}$$

The expected marginal profit is

$$E[D(s)] = -c + (g\mu + w - c)P(N > s). \tag{4.20}$$

Then it follows that

$$P(N > s) \leq c/(m + g\mu + w) \leq P(N > s - 1). \tag{4.21}$$

If we define $\alpha = c/(m + g\mu + w)$ then we can find the optimal number of servers iteratively until $P(N > s) < \alpha$. If the arrivals are assumed to be Poisson then N is also Poisson. Using the normal approximation as discussed above we can determine the optimal staffing level with the square-root formula $s = r + z_\alpha\sqrt{r}$.

Steady-State Queueing Model

The last model discussed by Grassmann [25] is the equilibrium model. Because of the steady-state assumption, the revenue per time unit and the revenue per customer are independent of the number of servers. Let $L_q(s)$ be the number of waiting customers. Using a similar marginal approach as for the infinite-server model, the number of servers is optimal if

$$L_q(s) - L_q(s-1) \leq c/w \leq L_q(s-1) - L_q(s). \tag{4.22}$$

The values of $L_q(s)$ are determined with the analytical expressions from the stationary queueing theory.

Heavy-Traffic Limit

Borst, Mandelbaum, and Reiman [8] used asymptotic optimization to determine the optimal number of staff at large call centers for different trade-offs between efficiency and service quality. The overall cost per time unit $C(s, \lambda)$ consists of the staffing cost $F(s)$ and the waiting cost $D(W)$ for a customer waiting W time units

$$\begin{aligned} C(s, \lambda) &= F(s) + \lambda E[D(W)] \\ &= F(s) + \lambda \pi(s, \lambda) G(s, \lambda) \end{aligned} \tag{4.23}$$

where $\pi(s, \lambda) = P(W > 0)$ and $G(s, \lambda) = E[D(W)|W > 0]$. For the $M/M/s$ model they let $\lambda \to \infty$ and searched for the asymptotically optimal s that minimizes $C(s, \lambda)$.

Three regimes were considered: an efficiency-driven regime where staffing cost dominates waiting cost ($F \gg G$), a quality-driven regime where waiting cost dominates staffing cost($G \gg F$), and a quality-and-efficiency-driven regime (QED) where staffing cost and waiting cost are balanced ($F \approx G$). They derived a square-root staffing rule

$$s = r + \beta(\cdot)\sqrt{r} \tag{4.24}$$

where the expression for $\beta(\cdot)$ depends on the cost functions and the regime. If we assume linear staffing cost c per agent per time unit and linear waiting cost w per customer per time unit, then in the QED the factor β can be determined with a unimodal function

$$\beta\left(\frac{w}{c}\right) = \operatorname{argmin}\left\{ y > 0 : cy + \frac{wHW(y)}{y} \right\} \tag{4.25}$$

where $HW(\cdot)$ is the Halfin-Whitt delay function. Numerical experiments showed that the approximation for the QED is also accurate for the other regimes.

4.1.2 Staffing with Time-varying Arrival Rates

In this section we discuss various approaches to determine the staffing levels when the arrival rate varies during the day.

Stationary Independent Period-by-Period Approximation

A common approach is to divide the day in n staffing intervals. It is then assumed that in each interval i the arrival rate λ_i is constant and that the intervals are independent. For a given quality-of-service requirement, the number of staff s_i is determined with a steady-state queueing model [29]. This approach is called the stationary independent period-by-period approximation (SIPP).

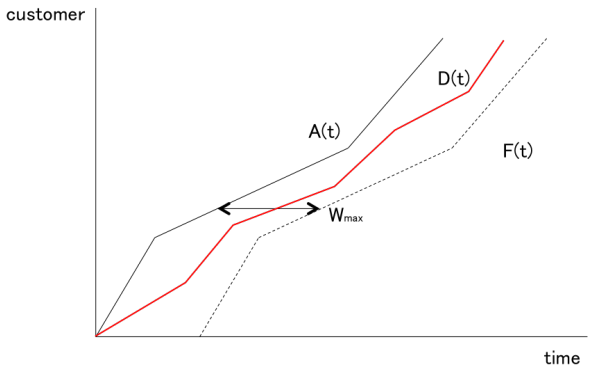


Figure 4.1: Staffing with cumulative diagrams.

Related is the pointwise stationary approximation (PSA). A steady-state queueing model is used to determine the number of staff $s(t)$ at each time t. Then the staffing level s_i in a staffing interval is the average or the maximum of all $s(t)$ in the interval.

Alternatively we can use any of the staffing approaches for a constant arrival rate, such as staffing to the offered load and the square-root staffing rule, to set the staffing levels in each interval independently.

Deterministic Fluid Model

In the deterministic fluid model the buildup of the queue can be analyzed from the cumulative arrivals $A(t)$ and the cumulative departures from the queue $D(t)$ [63]. We know that $D(t) \leq A(t)$ because at any time there cannot be more total customers departing from the queues than total arrivals. The waiting time W_j of passenger j is the horizontal distance between the two cumulative diagrams (Figure 4.1). We can set the performance requirement that the waiting time of any passenger is less than W_{max}, i.e. $W_j \leq W_{max}$. The time before which each passenger must be served is indicated by $F(t) = A(t - W_{max})$, i.e. it is the cumulative arrivals shifted by the maximum allowed waiting time. The performance requirement is met if

$$F(t) \leq D(t) \leq A(t). \tag{4.26}$$

We discretize time with a unit interval, $t = 1, 2, \ldots$. For given $A(t)$ the cumulative departures are determined with

$$D(t) = \min\{D(t-1) + \mu s(t), A(t)\}. \tag{4.27}$$

If the staffing objective is to minimize the total number of servers then this objective is achieved if $s(t)$ is set at each time interval such that $D(t-1) + \mu s(t) \leq A(t)$. Any resulting cumulative departures diagram has the same

total number of servers. If the left-hand side is larger than $A(t)$ there will be unused servers during that time interval. If there are no other requirements then it makes sense to set $s(t)$ such that $D(t) = A(t)$ because the waiting time becomes zero for all passengers. In other words the number of servers equals the offered load $s(t) = \lambda(t)/\mu = [A(t) - A(t-1)]/\mu$.

Littler and Whitaker [52] used the cumulative diagrams to determine the immigration staffing levels at an airport in New Zealand. They defined an objective that minimizes the maximum slope of the line segments of $D(t)$

$$\min \quad \max_t \{s(t)\}. \tag{4.28}$$

Simulation-Based Heuristics

Feldman et al. [21] developed a simulation-based iterative staffing algorithm (ISA) for time-varying queues. They considered an $M(t)/M/s(t)+M$ model, i.e. including abandonments, and an $M(t)/M/s(t)$ model (in the electronic companion of [21]). The performance requirement is the probability of delay

$$P(W(t) > 0) \leq \alpha \tag{4.29}$$

where α is the target probability. This requirement is equivalent to

$$P(N(t) \geq s(t)) \leq \alpha. \tag{4.30}$$

The time horizon is divided in intervals. The ISA works iteratively. Let $s_n(t)$ be the number of staff in interval t for iteration n and let $N_n(t)$ be the number of customers in the system if there are $s_n(t)$ servers. The algorithm starts with setting the staffing function $s_0(t)$ to a large value so that the probability of delay $P(N(t) \geq s(t))$ is very small. Then in each iteration i the system is simulated (with discrete-event simulation) multiple (5,000) times with $s_i(t)$ to determine the probability distribution of $N_i(t)$. In the next iteration $i+1$ the staffing level $s_{i+1}(t)$ is set to the lowest number for which the performance requirement is met in iteration i

$$s_{i+1}(t) = \arg\min\{k \in \mathbb{N} : P(N_i(t) \geq k) \leq \alpha\}. \tag{4.31}$$

The algorithm stops when the change in the staffing levels is very small. The ISA works well when compared with simulation results.

Defraeye and Nieuwenhuyse [14] adapted the ISA so that it can be used for small number of arrivals, for staffing intervals and for a performance requirement based on the excessive waiting time probability $P(W(t) > \tau) \leq \alpha$ instead of the delay probability $P(W(t) > 0) \leq \alpha$.

Dynamic Programming

Dynamic programming (DP) gives a procedure for determining the optimal sequence of interrelated decisions [34]. A DP problem has N stages and

each stage n has a number of states s. At each stage a policy decision x_n needs to be made. An important property is that the decision in a stage is independent of the decisions in the previous stages. We want to find the optimal policy that describes the optimal decision x_n^* for each possible state s such that the total cost of the overall policy is minimized. The solution procedure starts by prescribing the optimal policy decision x_N^* at the last stage for each possible state. Then given the optimal policy at stage $n+1$ the optimal policy for stage n is determined with a recursive cost relationship. Let the cost $f_n(s, x_n)$ be the sum of the immediate cost c_{sx_n} at stage n plus the minimum future cost $f_{n+1}^*(x_n)$ from stages $n+1$ onward. Let x_n^* be the value of x_n that gives $f_n^*(s)$, the minimum of $f_n(s, x_n)$. The recursive relationship is then [34]

$$f_n^*(s) = \min_{x_n}\{c_{sx_n} + f_{n+1}^*(x_n)\}. \tag{4.32}$$

With the DP procedure we successively find $f_N^*(s), f_{N-1}^*(s), \ldots, f_1^*(s)$, for each of the possible states s and the corresponding optimal policy decisions $\{x_N^*, x_{N-1}^*, \ldots, x_1^*\}$.

Fu, Marcus, and Wang [23] considered a transient queueing model with non-stationary arrivals. The objective is to set staffing levels in each time period of the day in order to minimize the expected cost. The time horizon is divided into N periods with length τ. Let the random variable $X_t^{(n)}(i, s)$ be the number of customers in the system at time t into period n, given i customers at the beginning of the period and s servers throughout the period. Let k be the cost per server per period. Then the cost $c_n(i, s)$ in period n is the sum of the system occupancy cost and the staff cost

$$c_n(i, s) = \frac{1}{\tau}\int_0^\tau E[X_t^{(n)}(i, s)]dt + ks. \tag{4.33}$$

Let X_n be the number of passengers at the beginning of period n and let σ_n be the number of servers assigned to period n if there are i initial customers. The optimization problem is to find the staffing policy matrix σ that minimizes the expected total cost

$$\min_\sigma \sum_{n=1}^{N} E[c_n(X_n, \sigma_n(X_n))]. \tag{4.34}$$

The staffing problem can be solved with dynamic programming. The recursive equation of the dynamic program is

$$f_n^*(i) = \min_s\{c_n(i, s) + E[f_{n+1}^* X^{(n)}(i, s)]\} \tag{4.35}$$

where the number of passengers $X^{(n)}$ in the system can be determined by numerical integration of the ODE of the transient $M(t)/M/s(t)$ queue. The

optimal policy $\sigma_n^*(i)$ is determined by backward induction

$$\sigma_n^*(i) = \arg\min_s \{c_n(i,s) + E[f_{n+1}^* X^{(n)}(i,s)]\}. \tag{4.36}$$

Roubos, Bhulai, and Koole [68] viewed the staffing problem as a Markov decision process (MDP) problem and solved it with dynamic programming. The goal of a MDP problem is to optimize the performance of a Markov chain [34]. Let x_t be the states of the Markov chain. In each state we make a decision about which action a_t out of several actions to take. The action affects the transition probabilities p_t from state x_t to state x_{t+1} and the short-term cost and long-term cost. For each state we want to determine the optimal action while taking both the short-term and long-term cost into account.

They considered a service system with N workplaces for s_t permanent staff and f_t flexible staff. The objective is to minimize the cost to meet a service level requirement SL by varying the number of flexible staff f_t over the day. Customer arrivals occur according to a non-homogeneous Poisson process with parameter λ_t. The day is divided into m intervals with length θ. The cost for permanent and flexible staff are respectively c_1 and c_2. There is a penalty cost P if the service level requirement is not met at the end of the day. The problem can then be formulated as follows

$$\min \sum_{i=1}^{m} (c_1 s_i \theta + c_2 f_i \theta) + P \mathbb{I}_{\mathrm{SL}<\alpha}. \tag{4.37}$$

This optimization problem is converted to a MDP problem. The state x_t represents the weighted service level realized up to t, i.e. $x_t = \sum_{i=1}^{t-1} \lambda_i \mathrm{SL}_i / \sum_{i=1}^{t-1} \lambda_i$. After observing x_t we decide on action a_t from action space $A_t = \{0, \ldots, N - s_t\}$. Taking action a_t means that a_t flexible staff are scheduled at t. The transition probabilities p_t are obtained through simulation. Given x_t and λ_t, the service level x_{t+1} is determined by assuming steady-state and simulating for each combination of x_t and a_t. This results in $p_t(x_t, a_t, x_{t+1})$, the probability of moving from state x_t to x_{t+1} after choosing action a_t. The direct cost is $c_t(x_t, a_t) = \sum_{i=1}^{t-1} (c_1 s_i \theta + c_2 f_i \theta) + P \mathbb{I}_{i=m} \mathbb{I}_{\mathrm{SL}<\alpha}$ where the permanent staff s_t can be determined with the SIPP approach. The optimal policy is then obtained by dynamic programming.

4.1.3 Staffing with Uncertain Arrival Rates

In this section we discuss staffing models for when the demand is uncertain.

Mean Arrival Rate

In practice and in the staffing literature distributional forecasts are made from historical data. Then it is common to use the mean of the distributional

forecast as the basis for staffing [17]. With the mean arrival rates any of the staffing methods for time-varying arrival rates can be applied.

Arrival Rate Randomization

To deal with forecasting errors in the deterministic model, Grassmann [25] randomized the arrivals by assuming that λ is normally distributed. Then the offered load r has a normal distribution with mean $E[r]$ and standard deviation $Std[r]$. For the deterministic model the total profit $T(s)$ was described in equation (4.17). The goal is to maximize the expectation of the total profit $E[T(s)]$ for a finite time span τ. Its derivative is

$$E'[T(s)] = E[T'(s)] = -c + P(r > s)(m + g\mu + w\tau\mu/2). \tag{4.38}$$

This gives the optimality condition

$$P(r > s) = c/(m + g\mu + w\tau\mu/2) = \alpha. \tag{4.39}$$

We can interpret α as the chance that there are not enough servers. The required number of servers is then

$$s = E[r] + z_\alpha Std[r] \tag{4.40}$$

with z_α being the z-score for $1 - \alpha$.

For the infinite-server model of Grassmann [25], randomization of the arrivals does not change the optimality condition $P(N > s) \leq \alpha \leq P(N > s-1)$, see equation (4.21). Only the distribution of the number of customers in the system N is affected. For Poisson arrivals with known r, the number of customers in the system is Poisson with $E[N|r] = r$ and $Var[N|r] = r$. If we randomize r then the distribution of N is affected as follows

$$\begin{aligned} E[N] &= E[E[N|r]] = E[r], \\ Var[N] &= Var[E[N|r]] + E[Var[N|r]] = Var[r] + E[r]. \end{aligned} \tag{4.41}$$

If N is normally distributed after randomization, the optimal number of staff is

$$s = E[N] + z_\alpha Std[N] = E[r] + z_\alpha \sqrt{Var[r] + E[r]}. \tag{4.42}$$

Arrival Rate Bounds

Jongbloed and Koole [47] used the SIPP approach with an $M/M/s$ queue and the service level requirement $P(W > \tau) < \alpha$. In each interval the arrivals are described by a Poisson variable X with parameter λ being drawn from distribution H. The distribution of X is then a Poisson mixture

$$P_H(X = x) = \int_0^\infty \frac{\lambda^x}{x!} e^{-\lambda} dH(\lambda). \tag{4.43}$$

The mixing distribution H can be determined from historical data with a parametric approach or with a maximum likelihood estimator. Then the lower and upper bounds of λ are the 5% and 95% quantiles of the mixing distribution.

Because $P(W > \tau)$ is increasing in λ for an $M/M/s$ queue, the upper and lower bounds can be used in two ways. In the first case the staffing levels cannot be adjusted once the staffing requirements have been set. One needs to consider a worst-case scenario. Staffing with the upper bound then gives the guarantee that there is a 95% chance that the service level requirement is met.

In the second case it is assumed that it is possible to call additional staff with flexible contracts. The lower bound gives the number of fixed staff and the upper bound indicates how many flexible staff are needed.

Newsvendor Problem

The newsvendor problem deals with inventory management for uncertain demand. Suppose there is a single time period and the newsvendor needs to order quantity x to sell during the period but the demand is uncertain. The demand is described by a random variable D with pdf f, cdf F and mean $E[D]$. The newsvendor buys quantity x for price c per unit. If at the end of the period $x \geq D$ then the remaining units $(x - D)$ have a holding cost h per unit. On the other hand if $x < D$ then there is a penalty b per unit for lost sales $(D - x)$. The total cost $C(x, D)$ at the end of the period is then [69]

$$C(x, D) = cx + b[D - x]^+ + h[x - D]^+. \tag{4.44}$$

The objective of the newsvendor is to minimize the total expected cost

$$\begin{aligned} E[C(x, D)] &= cx + bE[D - x]^+ + hE[x - D]^+ \\ &= bE[D] + (c - b)x + (b + h)\int_0^x F(z)dz. \end{aligned} \tag{4.45}$$

By equating the derivative to zero, the optimal order quantity x^* can be determined

$$x^* = F^{-1}\left(\frac{b - c}{b + h}\right). \tag{4.46}$$

Ding and Koole [17] formulated the staffing problem as a newsvendor problem with the objective to minimize the total staffing cost while meeting a service level constraint. The total cost consists of the initial staffing cost and the traffic management cost. The staffing level is determined with a steady-state model to meet a service level requirement. Initial staffing is usually done weeks or months in advance. The initial staffing level s is based on the distributional arrival forecast described by the random variable Λ with cdf F_Λ. During the day a more accurate arrival rate forecast λ is available.

The number of staff is then adjusted (this is called traffic management) to the correct staffing level $S(\lambda)$. Let c be the cost per agent. The overstaffing cost is $(c_o - c)$ per agent and the understaffing cost is $(c_u + c)$ per agent. The total cost $C(s, \lambda)$ is the sum of the initial staffing cost and the traffic management cost

$$\begin{aligned} C(s,\lambda) &= cs + (c_o - c)[s - S(\lambda)]^+ + (c_u + c)[S(\lambda) - s]^+ \\ &= cS(\lambda) + c_o[s - S(\lambda)]^+ + c_u[S(\lambda) - s]^+. \end{aligned} \tag{4.47}$$

The goal is to minimize the expected total cost

$$E[C(s,\Lambda)] = cE[S(\Lambda)] + c_o E[s - S(\Lambda)]^+ + c_u E[S(\Lambda) - s]^+. \tag{4.48}$$

This is a newsvendor problem with demand $S(\Lambda)$ and its cdf H. The solution for the newsvendor problem is thus

$$s^* = H^{-1}\left(\frac{c_u}{c_o + c_u}\right). \tag{4.49}$$

Ding and Koole [17] showed that $H^{-1}(p) = S(F_\Lambda^{-1}(p))$ for $0 \le p \le 1$ and proved the monotonocity of S for the $M/M/s$ and $M/M/s+G$ models with a service level constraint. The optimal staffing level becomes

$$s^* = S\left(F_\Lambda^{-1}\left(\frac{c_u}{c_o + c_u}\right)\right). \tag{4.50}$$

In other words the optimal staffing level should be determined according to the $c_u/(c_o + c_u)$ quantile of the arrival distribution forecast.

Bassamboo, Randhawa, and Zeevi [6] developed a staffing model that also has the form of a newsvendor problem. They assumed an $M/M/s + M$ queueing model with a doubly stochastic Poisson arrival process. The arrival rate Λ is a random variable with mean λ and cdf F_Λ. Customers are impatient with the abandonment times being exponentially distributed with mean $1/\gamma$. The total cost $C(s)$ is defined as the sum of the staffing cost and the customer cost. The customer cost consists of the waiting time cost w per customer per time unit and the cost of abandonment a per customer. The cost of staffing is c per agent per time unit. The expected queue length in steady state is $Q = E[N - s]^+$ and the abandonment rate is γQ. The objective is to find s that minimizes the expected total cost in steady state

$$E[C(s)] = (w + a\gamma)E[N - s]^+ + cs. \tag{4.51}$$

The objective function is approximated by ignoring stochastic variability in the customer arrivals and service times, and only considering uncertainty in the arrival rate. The system is approximated by a fluid model with arrival rate Λ. The approximate abandonment rate is the arrival rate minus

the total service rate, i.e. $E[\Lambda - \mu s]^+$. The queue length can then be approximated as

$$Q = E[N - s]^+ \approx \frac{1}{\gamma} E[\Lambda - \mu s]^+. \tag{4.52}$$

The approximate objective function becomes

$$E[C(s)] = (w/\gamma + a)E[\Lambda - \mu s]^+ + cs. \tag{4.53}$$

The solution for this newsvendor problem is

$$F_\Lambda(\mu s^*) = 1 - \frac{c}{w/\gamma + a}. \tag{4.54}$$

Let $\overline{F}_\Lambda = F_\Lambda - 1$ be the tail distribution function of F_Λ then the optimal number of servers is

$$s^* = \frac{1}{\mu} \overline{F}_\Lambda^{-1} \left(\frac{c}{w/\gamma + a} \right) \tag{4.55}$$

They derived the following rule of thumb. If the coefficient of variation of the random arrival rate $CV = Std[\Lambda]/E[\Lambda] = \sigma/\lambda$ is larger than $1/\sqrt{r}$ then uncertainty dominates variability and the newsvendor approximation is accurate. If $CV < 1/\sqrt{r}$ then variability dominates and the square-root staffing rule $s = r + \beta\sqrt{r}$ can be used to improve the solution of the newsvendor approach.

Stochastic Programming

In the previous section a closed-form solution was derived for the newsvendor problem. For more difficult cases the stochastic program can be formulated as a linear programming problem. Suppose the random demand D has a discrete distribution with outcomes $d_1, \ldots, d_K$ and probabilities $p_1, \ldots, p_K$. We can write the expected total cost as a weighted sum [69]

$$E[C(x, D)] = \sum_{k=1}^{K} p_k C(x, d_k). \tag{4.56}$$

The linear programming problem is then

$$\begin{aligned} \min \ & \sum_{k=1}^{K} p_k C(x, d_k) \\ \text{s.t.} \ & C(x, d_k) \geq (c - b)x + bd_k, \quad k = 1, \ldots, K \\ & C(x, d_k) \geq (c + h)x - hd_k, \quad k = 1, \ldots, K \\ & x \geq 0. \end{aligned} \tag{4.57}$$

Liao et al. [51] studied a call center model with doubly stochastic Poisson arrivals. The staff have two types of jobs: handle incoming calls first and

alternatively do back-office jobs that can be delayed. The day is divided into n intervals and in each interval i the mean arrival rate Λ_i is random

$$\Lambda_i = \Theta f_i, \text{ for } i = 1, \ldots, n \tag{4.58}$$

where Θ is a random variable describing the busyness of the day, and f_i describes the shape of the arrival rate intensity in each interval. The busyness variable Θ has a discrete distribution with outcomes θ_l and probabilities p_{θ_l} for $l = 1, \ldots, L$.

The back-office workload is represented by the random variable A, the number of staff needed to do the back-office workload during a single period. The discrete probability distribution of A has outcomes a_k and probabilities p_{a_k} for $k = 1, \ldots, K$. The number of staff v_i required to handle the calls in interval i are determined with an SIPP approach and an $M/M/s$ model. For a given number of staff v, there is an expression for the steady-state waiting time probability $P(W_i \leq \tau|\theta)(v) = F_{\theta i}(v)$. Given the service level requirement $P(W_i \leq \tau) \geq \mathrm{SL}_i$ and the arrival rate sample θf_i, the required number of staff can be determined with

$$v_i(\theta f_i) = F_{\theta i}^{-1}(\mathrm{SL}_i). \tag{4.59}$$

Let s be the number of staff during the day, let v_i be the number of staff required to handle the call arrivals θf_i in interval i, and let a be the number of staff required to handle the back-office workload in an interval. Note that s is assumed to be constant over all intervals. The staff salary cost is c and the overtime cost (for doing the delayed backoffice jobs) is r per agent per period, with $c < r < u$. The daily total cost is the sum of the salary cost during normal time, the under-staffing cost and the overtime cost

$$C(s, \theta, a) = ncs + u \sum_{i=1}^{n} [s - v_i(\theta f_i)]^- + r \left[a - \sum_{i=1}^{n} [s - v_i(\theta f_i)]^+ \right]^+. \tag{4.60}$$

The objective is to minimize the expected daily total cost

$$E[C(s, \theta, a)] = \sum_{l=1}^{L} \sum_{k=1}^{K} p_{\theta_l} p_{a_k} C(s, \theta_l, a_k). \tag{4.61}$$

They proved that the expected daily total cost $E[C(s, \theta, a)]$ is convex in s. Two approaches were used to solve the stochastic problem: a linear program and a robust program. The model was extended by allowing overflows between successive periods. The overflow equals the backlog b_{i-1} from a previous period. The backlog is determined with the Erlang B model as in the stationary backlog-carryover approach (SBC). Overflow can be viewed as under-staffing by $\frac{b_{i-1}(s)}{\mu}$ agents in the current period. Thus we can add an additional under-staffing penalty cost $u\frac{b_{i-1}(s)}{\mu}$ to $C(s, \theta_l, a_k)$.

4.1.4 Airport Staffing

Green, Kolesar, and Whitt [29] discussed how queueing theory is used by airlines to schedule airport personnel at for example, check-in and ticket counters. Scheduling is often done with 15-minute staffing intervals. The arrival rate is based on the flight departures and contact ratios, which is the percentage of arrivals for each type of passengers that require service at different service counters (e.g. economy and first class). There are no passenger abandonments. The expected service times are a few minutes but can vary by the time of day. The service standards are fairly high and can vary throughout the day. Service standards could be defined as 85% of the passengers served in less than 5 minutes. The SIPP approach is commonly used to determine the required number of staff during the day.

United Airlines [35] developed a shift scheduling system for its employees at reservation offices and customer service agents at airport. The system produced $6 million labor cost savings per year. The staffing requirements were determined by a forecasting model and a queueing model. For the reservation offices the forecast was based on historical call volumes and an autoregressive moving average technique was used. Poisson arrivals and exponential service times were assumed for the queueing model which proved to be valid. The queueing model provided the number staff for 30-minute intervals such that a certain percentage of passengers was served within a desired waiting time. For check-in personnel they used passenger loads and arrival trends to forecast the work load. A similar queueing model was used for staffing airport personnel. Integer and linear programming techniques were then used to determine monthly shift schedules.

Brusco et al. [10] improved the airport personnel scheduling model for United Airlines. The staffing requirements were determined by using the flight schedule and forecasts of passengers, luggage and cargo loads. The passenger forecasts for check-in counters were based on passenger arrival curves at the airport, and the percentage of passengers that require different types of service at the counters. The arrival curves, contact ratios and service times are time-dependent and have different values depending on the time of the day, day of the week and type of queue. A queueing model was used to calculate the required number of employees per 15-minute interval. Based on the staffing function and other personnel requirements a tour-scheduling model generated schedules for full-time and part-time shifts.

Littler and Whitaker [52] developed a staffing algorithm for immigration personnel at a New Zealand airport. Their performance requirement was that 100% of all passengers are processed from landing through immigration within a specified time. Because the performance requirement refers to a whole flight the stochastic variations in the queueing system can be ignored and a deterministic model is justified. Cumulative diagrams were used to determine the staffing requirements. Their staffing problem was

more complicated because the baggage claim is placed before immigration at the New Zealand airport. A simple model was introduced to estimate the percentage of passengers who picked up their luggage at each time.

Mason, Ryan, and Panton [56, 55] developed a staffing heuristic for customs personnel at Auckland International Airport in New Zealand. The performance requirement was that 85% of the passengers of a flight needs to be finish customs within 45 minutes and the other passengers have to be processed within 15 minutes later. The staffing levels per 15 minute interval were calculated once a week. An algorithm was implemented to test if a reduction in staff hours in one of the intervals if feasible. It tries to find an optimum by reducing the number of periods with the most number of staff $s_{\max}$ and continues with $s_{\max} - 1$ and so on. Testing for feasibility is done with a simulation program. The algorithm stops when a local optimum has been found.

4.2 Staffing at Narita Immigration

The decision makers at immigration of Narita Terminal 1 base their staffing decisions on the expected flight arrival times, the number of (non-transit) foreign and Japanese passengers on each flight, and four TV screens that display the inbound and outbound immigration areas at the South and North wings of Terminal 1. When staff are not needed, they stay at a location 10 minutes from the immigration area. During peak times it is possible to request additional staff from Terminal 2. Arrival forecasting and staffing are mainly experienced-based which can lead to suboptimal staffing levels and excessive waiting times. The decision-makers at Narita Airport Immigration want to limit the maximum waiting time of a foreign passenger to 10 minutes. In this chapter we will deal only with the number of staff for foreign service counters. It is assumed that there is only one queue for all foreign passengers and reentry permit holders join the same queue as the other foreigners.

4.2.1 Staffing Model

Important characteristics of the foreign queueing system are long periods with overload and uncertainty of the arrival rate. Call centers are commonly assumed to be underloaded and many studies use the PSA or SIPP approach where the steady state queueing theory is used in independent time intervals [47]. In the studies that include overload it is also assumed that there are customer abandonments because these stabilize the queue [80, 81]. At immigration however there are no abandonments because all passengers must go through immigration. The deterministic staffing model by Littler and Whitaker [52] is appropriate for overload but they did not consider uncertainty in the flight arrivals. In the reviewed papers that deal with uncertain demand, the methods either deal with only one time period

[25, 17], or the staffing level is assumed to be constant over the whole day [51], or a steady-state queueing model is used [47].

There are two approaches to staffing: a constraint satisfaction approach and a cost optimization approach. In practice the constraint approach is commonly used because it is difficult to quantify the waiting time cost and because the performance requirement is often described by the service level. Another important consideration is the practical use of the staffing model. At Narita immigration the decision makers want a support tool to help them make decisions. The output of the model is just one of the factors that they take into account. Other factors are the real-time queues and the availability of the staff. A constraint satisfaction approach is more intuitive for the decision makers because the main goal is to keep the maximum waiting time under 10 minutes. Therefore we define the staffing problem at Narita immigration as a constraint satisfaction approach. The computational speed is also a factor in practice which means that simple heuristics are preferred.

The objective is to minimize the staff cost while meeting a service level constraint. The service level is defined in [43] as $\mathrm{SL}(t) = P(W(t) \leq W_{\max})$, and the complementary service level as $\overline{\mathrm{SL}}(t) = P(W(t) > W_{\max})$. The latter is also called the waiting time tail probability [30], the excess wait probability [14] or the total service factor [8]. For our discussion we define the service level as the excess wait probability

$$\mathrm{SL}(t) = P(W(t) > W_{\max}). \tag{4.62}$$

A lower service level is thus better. We will refer to $\mathrm{SL}(t)$ as the interval service level. As in [68] the performance requirement is to achieve the desired service level over the whole day where the service level is weighed by the arrival rate in each time interval

$$\mathrm{SL} = \frac{\sum \lambda(t) P(W(t) > W_{\max})}{\sum \lambda(t)} \leq \alpha. \tag{4.63}$$

We will refer to SL as the daily service level. We assume that the target service level at Narita immigration is that at most 1% of the foreign passengers have a waiting time longer than 10 minutes, i.e. $W_{\max} = 10$ and $\alpha = 0.01$.

Grassmann [25] suggested the following procedure to find an appropriate staffing model. Start with the simplest model and validate the model. If the results are unsatisfactory then add more features to the model and validate the new model. This is done recursively until an appropriate model has been found.

4.2.2 Deterministic Staffing

In this section we apply the deterministic fluid model from section 4.1.2. First we discuss the deterministic staffing model for a single flight and determine the waiting time performance for the case of uncertain walking speeds.

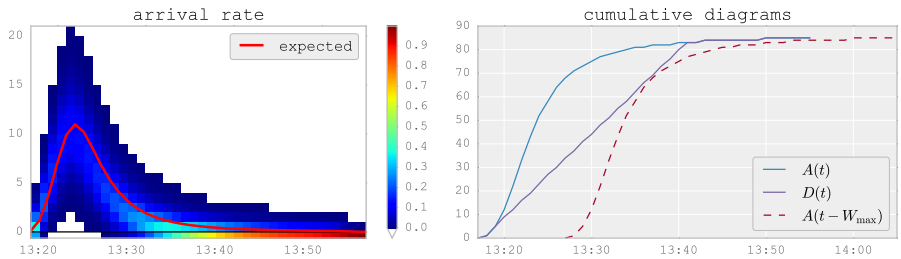


Figure 4.2: Arrival rate probabilities with the expected arrival rate of a single flight (left) and the cumulative diagrams (right).

Then we extend it to multiple flights and test the performance with simulations that include uncertain flight delays and uncertain disembarkation delays.

Single Flight

We consider the case of passenger arrivals of a single flight without flight delay uncertainty and without disembarkation delay uncertainty but with variability in the individual walking speed. Figure 4.2 (left) shows the arrival probabilities for a flight with 85 foreign passengers. The expected flight delay is 2 minutes and the expected disembarkation delay is 10 minutes. The maximum allowed waiting time $W_{\max}$ for any passenger of the flight is 10 minutes. We use a deterministic fluid model with the arrival rate $\lambda(t)$ equal to the expected value of the arrival probabilities at each time. Let $A(t) = \sum \lambda(t)$ be the cumulative arrivals and let $D(t)$ be the cumulative departures from the queue then the waiting time requirement is met if

$$A(t - W_{\max}) \leq D(t) \leq A(t). \tag{4.64}$$

The target cumulative curve $A(t - W_{\max})$ is shown in Figure 4.2 (right). There are many curves $D(t)$ that meet this requirement. We want to minimize the total number of staff. When $D(t) < A(t)$ all staff $s(t)$ are fully used. When $D(t) = A(t)$ we can avoid overstaffing by setting $s(t) = \lambda(t)/\mu$. However this still allows many curves. As in [52] we set the objective to minimize the maximum of the number of staff during the period, i.e. minimize the maximum slope of curve $D(t)$. We achieve this by drawing curve $D(t)$ such that it touches curve $A(t - W_{\max})$ as shown in Figure 4.2 (right). We developed the following iterative heuristic to find $s(t)$ and the corresponding $D(t)$. It is based on the idea of a constant number of staff during the peak time and afterwards until the queue has vanished, see Figure 4.3 (left). First set the number of staff $s(t)$ to one over the entire arrival period. Then the waiting time is calculated with a deterministic queueing model. If the waiting time requirement is not met, increase the number of staff by

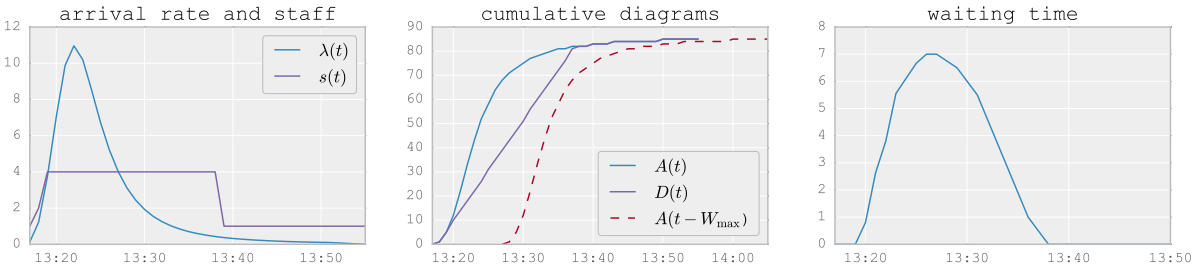


Figure 4.3: Staffing a single flight with the deterministic fluid model.

one over the entire time period and repeat the process until the maximum waiting time is less than the allowed maximum waiting time

$$s^*(t) = \arg\min\{s : W_s(t) \leq W_{\max}\} \tag{4.65}$$

We now have a constant staffing function but after the arrival peak has passed and the waiting time has become zero, the same staffing level is no longer needed. Therefore we determine when the waiting time has reached zero as in Figure 4.3 (right), and set the number of staff to one for all times after that to serve the late arrivals. Furthermore we can also reduce the number of staff at the beginning of the period when the waiting time is also zero. The number of staff is an integer therefore the maximum slope of the curve cannot be exactly minimized but the waiting time requirement is still met as shown in Figure 4.3 (middle).

We simulate the waiting time performance with the resulting staff function $s(t)$ when there is walking speed variability. The waiting time probability $P(W(t) = k)$ at each time interval t is determined by simulating the system 200 times with a random walking speed for each passenger. The waiting time probabilities are shown in Figure 4.4. The plot also includes the 99% quantile line which indicates the waiting times $W_p(t)$ for which $P(W(t) \leq W_p(t)) = 0.99$. We can conclude that the deterministic model performs well for a single flight if the delays are deterministic and the walking speed is random.

Multiple Flights

Now we consider multiple flights on the same day. It is assumed that the flight schedule data is updated at 13:00 and all flights before this time have been processed at immigration. The goal is to find the total staffing function for all flights from 13:00. As in the deterministic model for a single flight, the expected flight delay and the average disembarkation delay are used.

For each individual flight we determine the expected arrival rate $\lambda_{\mathfrak{fl}}(t)$ and the staffing function $s_{\mathfrak{fl}}(t)$. The total arrival rate $\lambda(t) = \sum \lambda_{\mathfrak{fl}}(t)$ and the total staffing function $s(t) = \sum s_{\mathfrak{fl}}(t)$ are the sum of respectively the

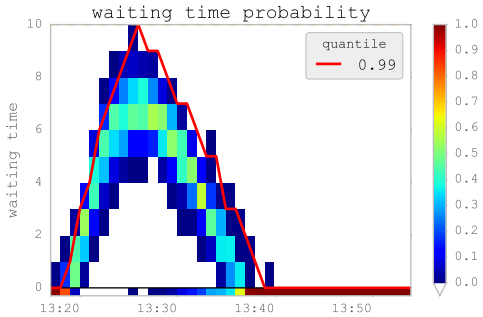


Figure 4.4: Waiting time probabilities for the case of staffing a single flight.

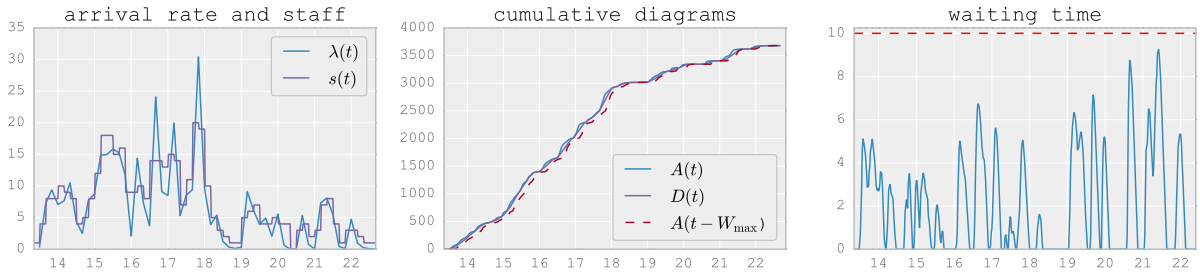


Figure 4.5: Output of a deterministic queueing system with multiple flights.

expected arrival rate and the staffing function of all individual flights. However we allow the staffing function to be non-integer with a step size of 0.1 because summing multiple flights will lead to overstaffing if the staffing functions of the individual flights are rounded up to the nearest integer. Also a staffing interval of 10 minutes is used, i.e. the number of staff is averaged over 10 minute intervals and the staffing level is constant and equal to the average of each interval. An example output of a deterministic queueing system for 44 flights is shown in Figure 4.5. The maximum waiting time experienced by any passenger is less than 10 minutes.

Performance Simulation

We have collected the flight schedule updates for 514 days from 2013 to 2015 with the actual ETA of the flights during the day. In general the flight schedules do not change much on a weekly basis. However because the actual ETA are used the resulting daily schedules are different for each day. The number of passengers of each flight is unknown therefore the passenger load is estimated with the average of the flight data that we collected on the observation days or by using the aircraft type as described in the arrival forecasting chapter. It is assumed that the passenger load is constant over the year. The daily service level SL (after 13:00) is calculated with 200

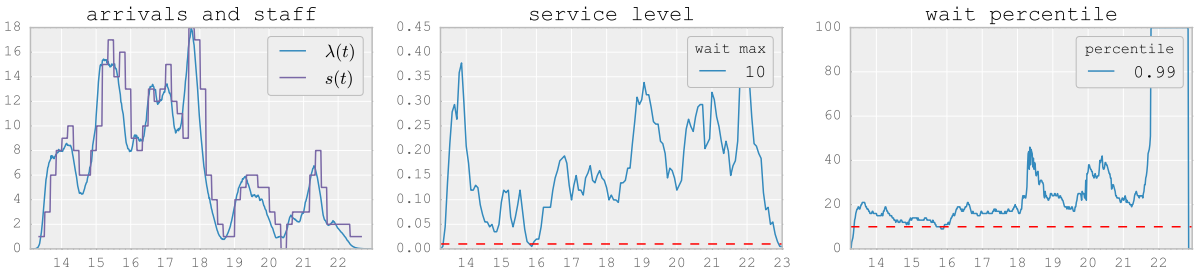


Figure 4.6: Example of the waiting time performance with multiple flights.

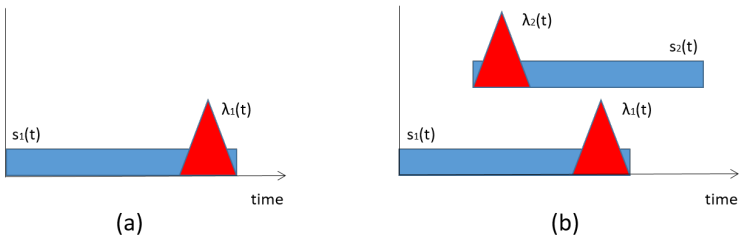


Figure 4.7: A flight arriving at the end of the staffing period without other flights nearby (a) and with another flight arriving around the same time (b).

simulations of the queueing system. Figure 4.6 shows an example of the simulation results for one day. The left plot shows the expected arrival rates and the number of staff. We can see that the staffing function lags the arrival rates. The middle plot shows the interval service level $\mathrm{SL}(t)$ for a target maximum waiting time of 10 minutes. The 99% quantile of the waiting time probabilities is shown in the right plot.

The service level plot shows that the worst service levels occur when the number of arrivals are the lowest. At the end of the day the 99%-quantile waiting time increases indefinitely which means that there are still passengers in the queue without any staff available. This can be explained with Figure 4.7. In plot (a) there is a single isolated flight with $s_1(t)$ being the the staffing function for the flight according to the deterministic model and $\lambda_1(t)$ being the actual arrival rate. If the actual flight arrival time is at the end of the staffing period then there are not enough staff available after arrival and the waiting time increases indefinitely. In plot (b) the staffing periods of two flights overlap. If flight 1 arrives at the end of its staffing period it can still use the staff that were allocated for flight 2 if that flight arrived at the beginning of its staffing period. Flight 2 can also utilize the unused staff of flight 1. The more overlapping flight, the better the service level becomes on average.

The daily service level of the example is 0.17. In other words the target service level requirement of 0.01 is not met for this day. Table 4.1 shows

Table 4.1: Daily service level statistics for various number of samples.

samples	2	5	10	20	50	100
service level						
mean	0.123	0.134	0.129	0.131	0.131	0.130
std	0.020	0.021	0.019	0.020	0.018	0.017
min	0.100	0.100	0.100	0.099	0.098	0.095
max	0.138	0.156	0.156	0.159	0.167	0.167

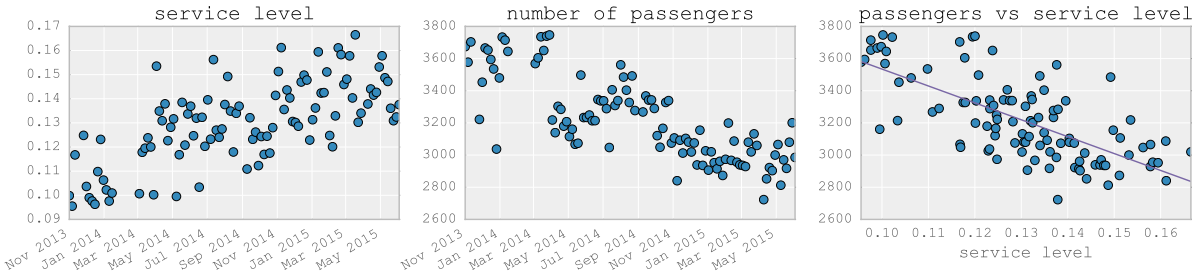


Figure 4.8: The service level (left), the total passenger volume (middle), and the service level vs. total passenger volume (right) for 100 days.

the daily service level averaged over various number of days. The samples are taken with equal distance between each sample. The mean daily service level is 0.13 with a range from 0.095 to 0.167. The number of samples do not affect the mean and the standard deviation of the daily service level significantly.

In Figure 4.8 (left) the daily service level is plotted for 100 days. The total number of passengers for the flights in the flight schedule is shown in Figure 4.8 (middle). The correlation between the daily service level and the number of passengers is -0.70. The correlation is clearly visible if we plot the daily service level and the number of passengers in one figure as in Figure 4.8 (right). Staffing for days with more passengers results on average in lower daily service levels. The number of passengers correlates strongly with the number of flights. As explained above the more overlapping flights, the better the service levels that can be achieved.

Maximum Waiting Time

In the previous sections the staffing function was determined for $W_{\max} = 10$ minutes. Figure 4.9 shows the staffing functions for a single flight with a target maximum waiting time of 5, 10 and 15 minutes. As the maximum waiting time increases, the staffing level goes down and the length of the staffing period increases. The total number of staff over the whole period is

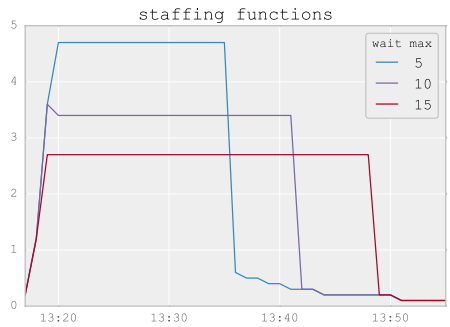


Figure 4.9: Staffing functions of a single flight for different $W_{\max}$.

Table 4.2: Daily service level statistics of 100 days for different $W_{\max}$.

wait max	5	10	15
service level			
mean	0.393	0.130	0.049
std	0.024	0.017	0.011
min	0.347	0.095	0.026
max	0.466	0.167	0.075

the same because all allocated staff are fully utilized.

For $W_{\max} = 5, 10, 15$ Table 4.2 shows the 100-day statistics for the daily service level. If $W_{\max}$ increases, the mean of the daily service levels decreases. From the results we can conclude that the deterministic model does not provide the required daily service level of 0.01 for any $W_{\max}$ in the table. The deterministic method also does not provide a way to adjust the staffing levels to meet the service level requirement.

4.2.3 Staff Probabilities

In the deterministic model the delay was deterministic. In this section we discuss how to extend the deterministic model with flight and disembarkation delay uncertainty.

Staff Probability Matrix

For the arrival forecast we added delay uncertainty by the convolution of the flight arrival probability matrix and the delay distribution. We will perform a similar procedure for the staffing function. Figure 4.10 shows the steps of the procedure. First the staff function $s(t)$ (left plot) is converted to a staff probability matrix S with probabilities equal to one (middle plot). Let d be the total delay distribution of a flight. We add the delay uncertainty by the

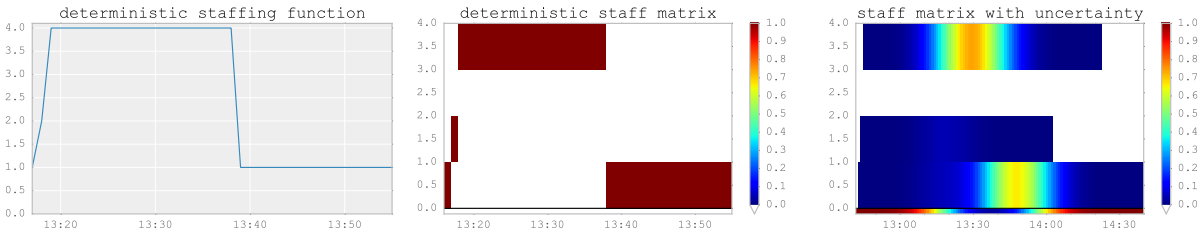


Figure 4.10: Procedure to add delay uncertainty to a deterministic staffing function of a single flight.

convolution of the staff probability matrix and the delay distribution of the flight

$$S_d(r) = S(r) * d \tag{4.66}$$

where $S(r)$ is the rth row of S, i.e. the probabilities of r number of staff over time. The resulting staff probability matrix is shown in the right plot of Figure 4.10.

For the arrival forecast the arrival probability matrix of each flight was calculated and then the arrival probabilities of all flights were summed with the convolution operation. We will do a similar procedure for the staff probability matrix. For each flight in a flight schedule we first determine the staff probability matrix. Then we convolve the staff probability distributions of each flight at time interval t, i.e. the columns of the staff probability matrices at t, to obtain the total staff probability distribution. Let $S_d^i(t)$ be the probability distribution at time interval t of the staff probability matrix S_d^i of flight i then the total distribution $\overline{S}(t)$ is given by

$$\overline{S}(t) = S_d^1(t) * S_d^2(t) * \cdots * S_d^k(t) \tag{4.67}$$

where k is the number of flights. Figure 4.11 shows an example of the arrival probabilities (left) and the staff probabilities (middle) for the flights after 13:00. In the figure the expected values and the 5% and 95% quantiles are also plotted. The range of the quantiles is smaller for the staff probabilities than for the arrival probabilities. In the right plot the expected values are compared. When the expected values increase both curves are similar but when the expected values decrease the staff curve lags the arrival curve.

Staffing with Quantiles

We propose a quantile-based staffing heuristic where the staff level is set to a certain quantile of the staff probability matrix. Let q be the quantile and let $X(t)$ be the number of staff for time interval t with probability distribution $\overline{S}(t)$ then the staffing function is

$$s(t) = \arg\min\{k \in \mathbb{N} : P(X(t) \leq k) \geq q\}. \tag{4.68}$$

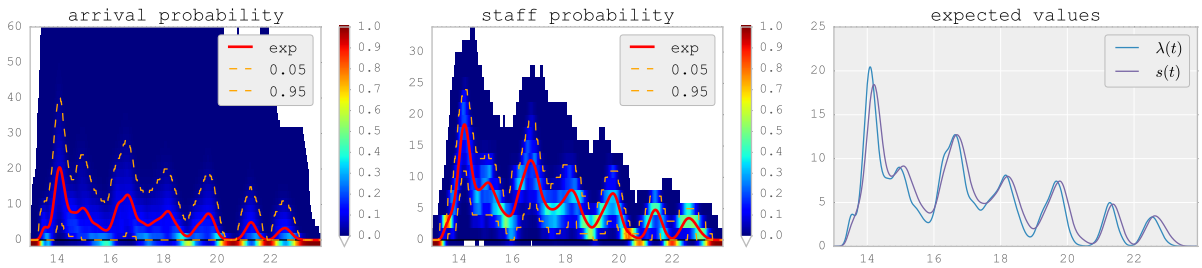


Figure 4.11: Example of arrival probabilities, staff probabilities and the expected values.

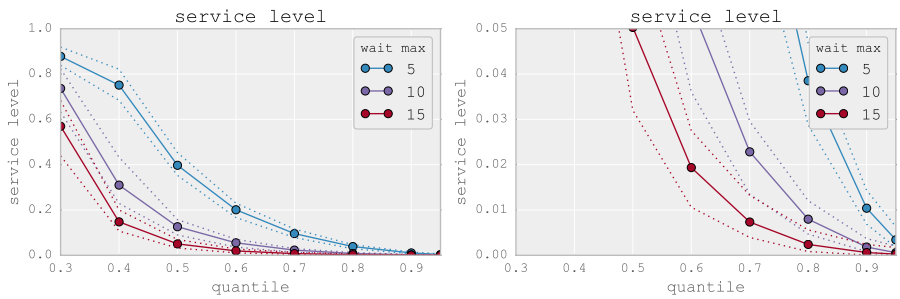


Figure 4.12: The mean daily service levels for different quantiles.

For $q = 0.3, 0.4, \ldots, 0.9$ and 0.95 the daily service levels were calculated for 100 days. The mean of the daily service levels are shown in Figure 4.12 (left). The dotted lines represent the 5% and 95% bounds. The right plot shows the same curves but zoomed in for service levels up to 0.05. For a target service level with $W_{\max} = 10$ and $\alpha = 0.01$ the appropriate quantile is 80%.

Figure 4.13 shows an example of the service level on one day when staffing with an 80% quantile. The daily service level is 0.007. The worst interval service level is 0.07 at 21:00. The waiting time percentile plot shows that at the end of the day there is a chance that the queues are not empty.

Staff Cost

The staffing levels are determined per 10-minute staffing intervals. Let c be the cost per agent per staffing interval. The total staff cost C is then

$$C = c \sum_i s(i) \tag{4.69}$$

where $s(i)$ is the number of staff at staffing interval i. We let $c = 1$ then the total staff cost is simply the sum of the staff per 10 minutes over the day. For 100 days the daily service level and the total number of staff were

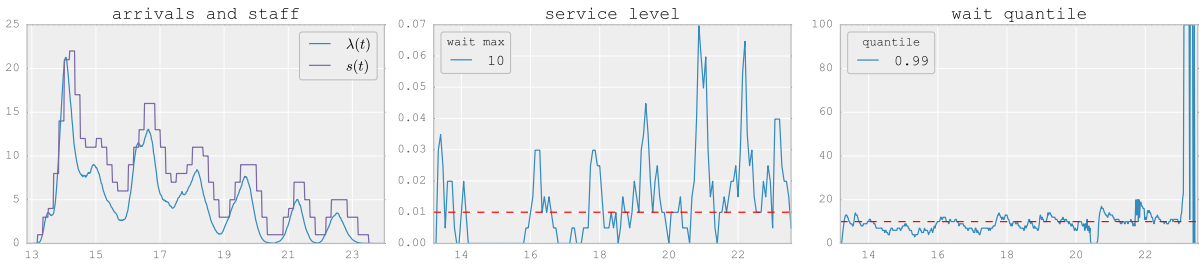


Figure 4.13: Example of the waiting time performance on one day when staffing with the 80% quantile of the staff probabilities.

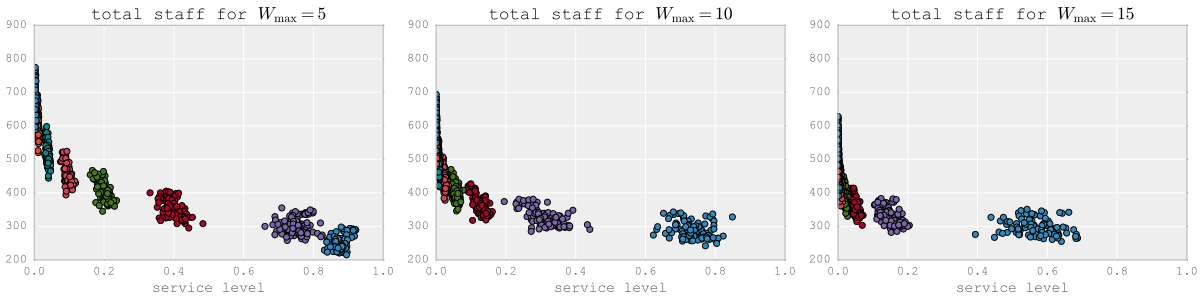


Figure 4.14: The total number of staff and the daily service levels of 100 days. The colors represent different quantiles $q = 0.95, 0.9, 0.8, \ldots, 0.3$.

calculated for different values of q. The result is shown in Figure 4.14. Each color represents a different quantile. From left to right the quantiles start at 0.95 and decrease to 0.3.

If the mean daily service level and the mean total number of staff is calculated over the 100 days then the total staff-service level curves for $W_{\max}$ equal to 5, 10 and 15 minutes are shown in Figure 4.15. This figure allows us to determine the average additional staff cost if we want to improve the service level. For a service level of 0.1 the mean number of staff is 379. To improve the service level to 0.05 requires an additional 29 staff. To achieve a service level of 0.02 requires an extra 42 staff. And if we want to improve the service level from 0.02 to 0.01 we need to add another 31 staff.

Update Time

Up to now we have assumed that the flight schedule is updated at 13:00 and only considered the flights from that time. However the flight schedule is updated continuously during the day every 10 minutes. To investigate the effect of the update time we calculated the 100-day mean of the daily service level for various update times: 6:00, 9:00, 12:00, 15:00, 18:00 and 21:00. The daily service level is thus the weighted service level from the update time

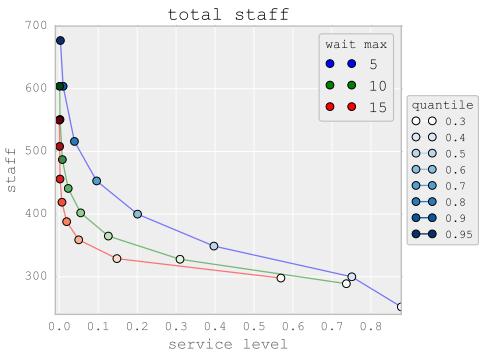


Figure 4.15: The mean total staff and the mean daily service level per quantile for $W_{\max} = 5, 10, 15$.

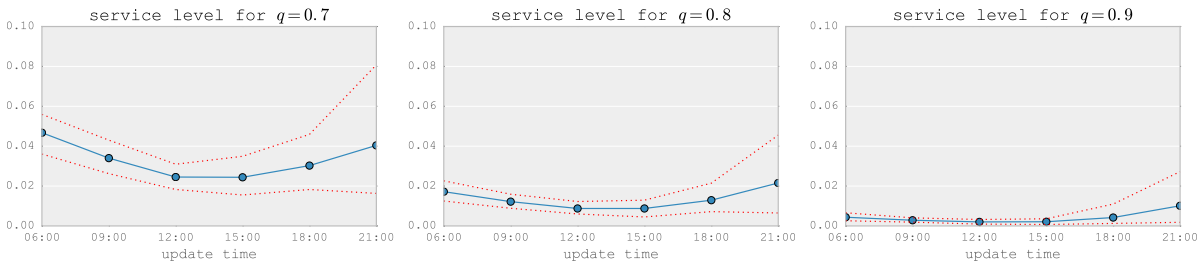


Figure 4.16: The mean daily service level at different update times and different quantiles.

to the end of the day. Figure 4.16 shows how the mean daily service level changes for $q = 0.7, 0.8, 0.9$. The daily service levels decrease from 6:00 to 12:00 and increase from 15:00 to 21:00. In other words if we want to have a constant daily service level when staffing at different update times, we have to staff with higher quantiles for update times in the morning and in the evening. The red dotted lines represent the 5% and 95% bounds of the daily service levels. The plots show that the variance of the daily service level increases significantly after 18:00. This means that we need to add extra staff during these periods.

4.2.4 Square-Root Staffing

The square-root staffing (SRS) formula, $s = r + \beta\sqrt{r}$, consists of the offered load $r = \lambda/\mu$ plus safety staffing to deal with stochastic uncertainty. The difficulty lies in finding the appropriate value of β. In this section we determine the service levels for different values for β, ranging from 0 to 1.5 with a step size of 0.1. The arrival rate is equal to the expected value of the arrival probabilities. SRS is applied with a staffing interval of 10 minutes. An example of SRS with $\beta = 1$ is shown in Figure 4.17. We see that the

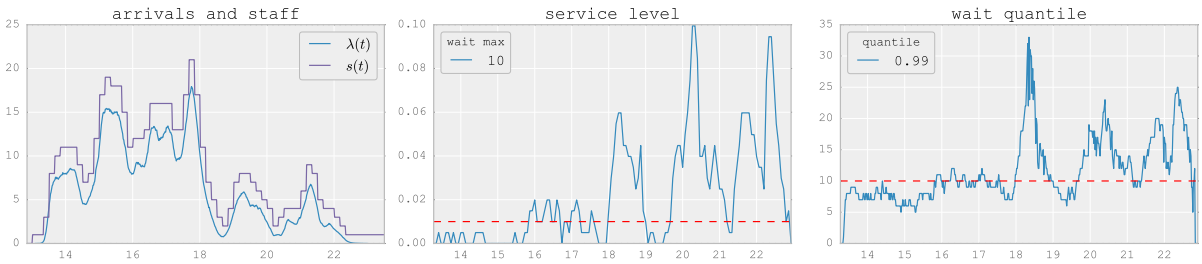


Figure 4.17: An example of the performance of SRS with $\beta = 1$.

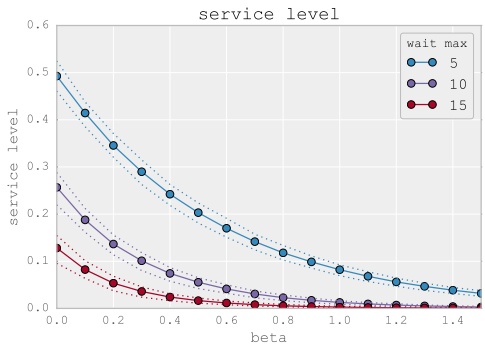


Figure 4.18: The mean daily service level as a function of β for $W_{\max} = 5, 10, 15$.

worst service levels are after 18:00 when the arrival rates are lowest. The daily service level with $W_{\max} = 10$ is 0.009.

Figure 4.18 shows the mean daily service level averaged over 100 days for $\beta = 0, 0.1, 0.2, \ldots, 1.5$. The dotted lines indicate the 5% and 95% quantile of the service level distribution. The statistics for the daily service levels of 100 days with $W_{\max} = 10$ are shown in Table 4.3. If the target daily service level is 0.01 then $\beta = 1.1$ provides good average performance.

Staff Cost

For 100 days the total staff and the daily service level are calculated for different values of β. Figure 4.19 shows the results for each separate day. Each group of dots with the same color represents a different value of β. From left to right the value of β goes from 1.5 to 0 with a step size of 0.1. If we take the mean of the service levels and the total staff for each β then the resulting total staff-service level curves are shown in Figure 4.20. Because the staffing levels vary linearly with β the curves are exactly the same as in Figure 4.18 with the x-axis and y-axis transposed. This figure allows us to determine the average additional staff cost if we want to improve the service level. For example in the case of $W_{\max} = 10$ to improve the service

Table 4.3: Daily service level statistics of 100 days for different β values with $W_{\max} = 10$ minutes.

beta	0.1	0.3	0.5	0.7	0.9	1.1	1.3	1.5
service level								
mean	0.188	0.101	0.055	0.031	0.017	0.010	0.005	0.003
std	0.016	0.012	0.007	0.004	0.003	0.002	0.001	0.001
min	0.143	0.072	0.035	0.020	0.008	0.005	0.002	0.001
max	0.226	0.125	0.073	0.040	0.024	0.014	0.008	0.005

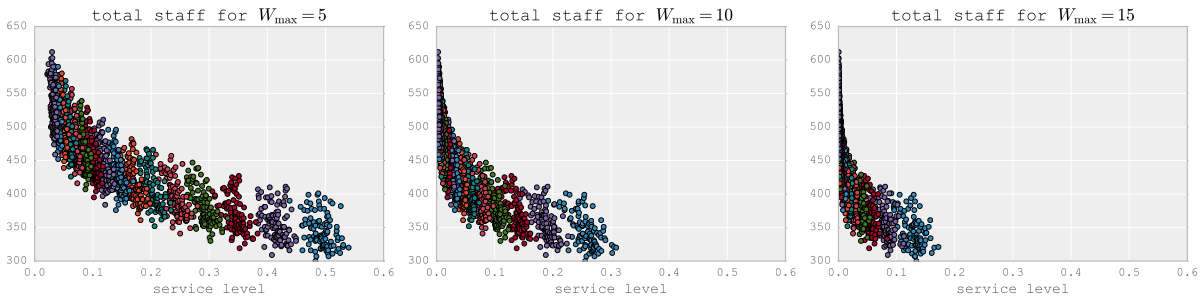


Figure 4.19: The total staff and the daily service level for different $W_{\max}$. The colors represent different values of $\beta = 1.5, 1.4, \ldots, 0$.

level from 0.1 to 0.05 requires an additional 29 staff. A similar additional cost of 30 staff is required to improve the service level from 0.02 to 0.01. In Figure 4.21 the total staff as a function of the service level for SRS and the staff probability quantile (spq) are shown. In general SRS gives better cost performance but the difference is small except for $W_{\max} = 15$.

Update Time

The mean daily service level over 100 days for different update times and values for β are shown in Figure 4.22. The mean daily service level decreases from 6:00 to 12:00 after which it stays constant. In contrast the daily service level with the staff probability quantile tends to increase for update times in the evening. Similar to staffing with the staff probability quantile the variance increases in the evening.

4.2.5 Iterative Algorithm

As can be seen in Figures 4.13 and 4.17 the service levels are worse when the arrival rates are lower. This means we have to add additional staff during these periods. Up to now the performance requirement was the daily service level SL $\leq \alpha$. Now we want to limit the service level during any staffing

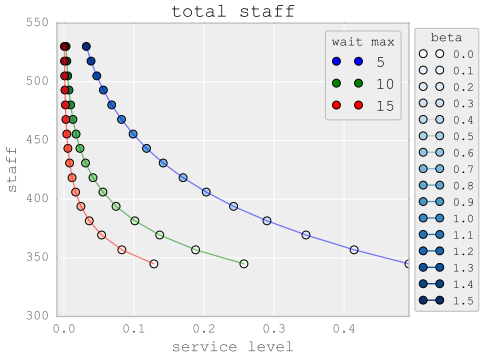


Figure 4.20: The mean total staff and the mean daily service level per β over 100 days.

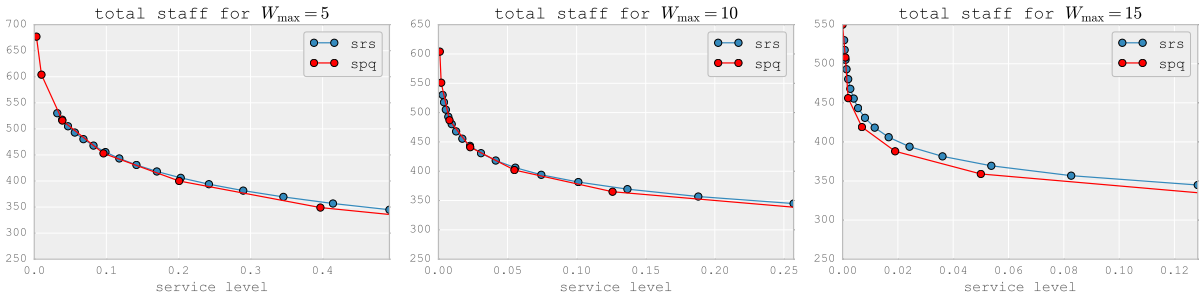


Figure 4.21: Comparison of the total staff per service level with SRS and staff probability quantile.

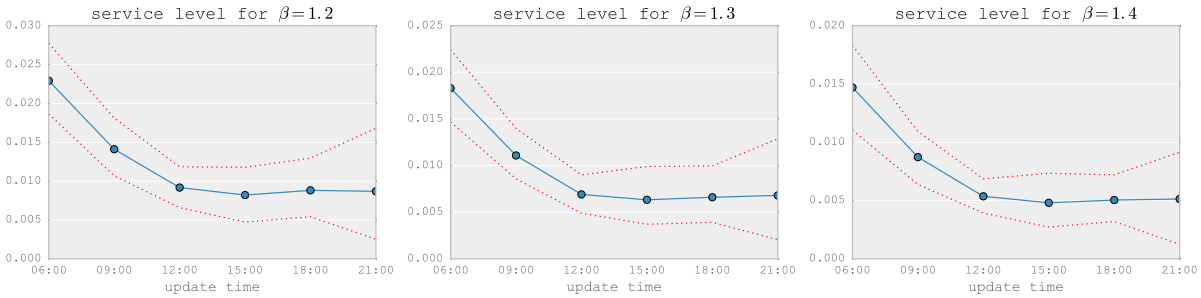


Figure 4.22: The average daily service level at different update times and different β.

Table 4.4: Service levels and total staff with the iterarive algorithm and deterministic staffing.

wait max	5	10	15
SL	0.01	0.008	0.007
max SL(t)	0.032	0.03	0.03
staff	613	480	414

interval. This means we impose a new constraint $\mathrm{SL}(t) \leq \gamma$.

We propose the following iterative algorithm to meet the new constraint. The initial staffing levels can be determined with any of the previously discussed methods. Then we calculate the service level at each time interval $\mathrm{SL}(t)$. For each staffing interval where the service level is not met, we add one extra staff member. Then the service levels $\mathrm{SL}(t)$ are calculated again. We stop if the constraint has been satisfied for all t or else we add another staff member to the intervals with excessive waiting time probabilities. An effect of adding additional staff is that the daily service level will also decrease. If the initial staffing levels were determined for α and we add additional staff then the final daily service level can be much smaller than α.

We apply the iterative algorithm for various initial staffing levels. The update time is 13:00, $W_{\max} = 10$ minutes and the performance requirement is $\mathrm{SL}(t) \leq 0.03$. First we determine the initial staffing function according to the deterministic model. For a target waiting time of 10 minutes the 100-day average of the daily service level was 0.13 (Table 4.2). We then apply the iterative algorithm to add staff to the staffing intervals where the constraint $\mathrm{SL}(t) \leq 0.03$ is violated. The resulting the multiple-day average of the daily service level, the maximum service level at any interval, and the total staff are shown in Table 4.4. We can see that changing the target maximum waiting time from 10 minutes to 5 minutes requires a large increase in the total number of staff.

Table 4.5 shows the results of the iterative algorithm with initial staffing levels according to SRS. If we start with β values from 0 to 0.6 the resulting service levels and total staff are very similar. The original staffing levels were shown in Table 4.3. A daily service level of 0.01 could be achieved with $\beta = 1.1$. If we start with $\beta = 1.1$ and then adjust the staffing levels with the iterative algorithm to meet $\mathrm{SL}(t) \leq 0.03$ the daily service level drops to 0.005. This means we are overstaffing because we can satisfy the same constraint with less total staff if the initial staffing level is set with $\beta = 0.1$.

Table 4.6 shows the results of the iterative algorithm if the initial staffing levels are determined with the staff probability quantile. For quantiles from 0.30 to 0.70 the daily service level and the total number of staff are similar.

Table 4.5: Service levels and total staff for $W_{\max} = 10$ with the iterative algorithm and square-root staffing.

beta	0.1	0.3	0.5	0.7	0.9	1.1	1.3	1.5
SL	0.01	0.01	0.009	0.009	0.007	0.005	0.004	0.002
max SL(t)	0.03	0.03	0.03	0.03	0.03	0.029	0.028	0.028
staff	477	478	478	481	488	501	517	538

Table 4.6: Service levels and total staff for $W_{\max} = 10$ with the iterative algorithm and the staff probability quantile.

quantile	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
SL	0.01	0.01	0.01	0.01	0.009	0.005	0.002	0
max SL(t)	0.03	0.03	0.03	0.03	0.03	0.029	0.025	0.017
staff	473	472	471	470	474	498	552	604

If we compare these results with the iterative algorithm with SRS then we can achieve the same service level performance but with less total staff when the initial staffing levels are determined with the staff probability quantile.

4.3 Conclusion

In this chapter we have reviewed the staffing literature for constant arrival rates, time-varying arrival rates and uncertain arrivals. Many models in the literature assume steady-state in the staffing intervals or assume abandonments if the queueing system is overloaded. Both assumptions don't seem appropriate for staffing at immigration. In the airport literature a deterministic fluid model was applied to the immigration service at an airport in New Zealand. However uncertainty in delay was not considered. We assessed the performance of the deterministic model when demand is uncertain and the requirement is that the weighed service level SL for the remainder of the day is less than 0.01. Simulation results show that the service level requirement cannot be met. We then extended the deterministic model by introducing staff probabilities. The staffing levels are determined with quantiles. In addition we have also applied the square-root staffing formula and determined the best value of β to meet the service level requirement. Although these staffing models give good performance for the daily service level, during periods with low arrival rates and at the end of the day the service levels $\mathrm{SL}(t)$ can reach unsatisfactory levels. We imposed a new constraint $\mathrm{SL}(t) < 0.03$ and we developed an iterative algorithm to adjust the staffing levels in the staffing intervals where the new constraint is violated.

Chapter 5

Conclusion

The purpose of our research is to reduce the waiting times for foreign passengers at Narita Airport immigration by optimizing the staffing levels. To do so we have developed three models: an arrival forecasting model, a queueing model and a staffing model. Based on the flight schedule and the number of passengers on each flight we first make a distributional forecast. The arrival forecast is then used as input for the staffing model. To meet a certain service level requirement the staffing model gives the staffing function, i.e. the required number of staff during the day. We can then simulate the performance of the staffing function with the queueing model.

The first step is to forecast the number of passenger arrivals at immigration. In the literature statistical models and discrete-event simulation models are commonly used. Statistical models require a large amount of historical data and simulation models generally require many iterations. We have developed a different approach to determine the arrival probabilities by using the sum of random variables. For a flight it is assumed that the walking time probability distribution is the same for each passenger. Then the arrival probabilities for the passengers of the flight without delay uncertainty are determined by the convolution of the distribution functions of every passenger at each time interval. Adding delay uncertainty to the flight requires the convolution of the arrival probabilities and the delay probability distribution. To calculate the arrival probabilities of the combined passengers of multiple flights we again apply the convolution operation. We call this forecasting model the convolution model.

The forecasting model requires knowledge of the probability distributions of the flight delay, the disembarkation delay, the disembarkation rate and the walking speed. To determine these probability distributions we gathered data at Narita Airport. Flight delay is defined as the difference between the estimated and the actual flight arrival time. In the literature on flight delays the scheduled arrival time is used as the estimated arrival time. However we estimate the flight delay with the more accurate estimated arrival time from

the real-time online flight schedule which is updated every 10 minutes. We have collected every flight schedule update for more than 1 year. We found that the delay probability distribution of a flight depends on the length of time between the estimated time of arrival and the flight schedule update time. The second uncertain parameter is the disembarkation delay. After an aircraft arrives at the gate, the passengers have to wait a certain time before disembarkation. We have recorded the time when the first passenger leaves the aircraft for 41 flights. A Markov Chain Monte Carlo algorithm was used to infer the distribution of the mean and the standard deviation of the disembarkation delay. For 10 flights the disembarking passengers at the gate were recorded on video. We found that the disembarkation rate is relatively constant for all flights. After disembarkation the passengers walk from the gate to immigration. Because the walking times could not directly measured we assumed that the walking speed is normally distributed and we inferred the walking speed mean and standard deviation from the observed arrival rates of 10 isolated flights.

In addition we have developed a Monte Carlo simulation model and a deterministic approximation. All three arrival forecasting models give reasonable results when compared to the observed arrival rates of a single flight and multiple flights. For use in practice the convolution model can give the decision makers at Narita immigration a good indication of the trend, the variance and the upper bound of the arrival rates while the deterministic approximation can give a sample path of the arrival rates. The Monte Carlo simulation model can be used to provide the arrival rate samples for the performance simulation of the staffing function.

The second step in our research is to develop a queueing model for the immigration service of foreign passengers. The purpose of the queueing model is to estimate the waiting time if the number of arrivals and staff are known. Traditional queueing theory studies the steady state of a queueing system. However at immigration the arrival rate is non-stationary and uncertain. Traditional queueing theory also assumes that the system is non-overloaded, i.e. the average arrival rate is less than the service capacity. However at immigration overload occurs frequently. In the literature there are various queueing approaches that can deal with overload. We have implemented three queueing models: the numerical integration of the Kolmogorov differential equations, the deterministic fluid approximation and the stationary backlog-carryover approach. In the literature there has not been a comparison of these three models. Also artificial arrival rates are commonly used in the literature to compare queueing models. Instead we use actual arrival rates and validated the models with actual waiting times.

To compare the three queueing models we gathered data at Narita Airport. We recorded videos of the queues to determine the waiting times and service times. We also counted the number of arrivals and the number of staff during five afternoons. The input of the queueing models is the ob-

served arrival rate, the observed number of staff and the observed service time. The output is the waiting time at each time of the day. We compared the output of the queueing models with the observed waiting times. All three models give good estimations of the waiting times. The numerical integration approach however requires long computation times. The deterministic fluid approximation is easy to implement, the fastest and accurate if a time interval of 1 minute is used. It is our preferred queueing model for the performance evaluation of the staffing function.

The final part of our study is to determine the staffing requirements for foreign passengers in order to support the decision makers at Narita immigration. In the call center and hospital staffing literature there are two approaches to determine the staffing function: the optimization approach and the constraint satisfaction approach. The optimization approach is rarely used in practice because of the difficulty of quantifying the waiting cost. For our staffing model we use the constraint satisfaction approach where the objective is to minimize the number of staff while meeting a service level requirement. The service level is defined as the excessive waiting time $\mathrm{SL}(t) = P(W(t) > W_{\max})$. We want to keep the daily service level $\mathrm{SL} = \sum \lambda(t)\mathrm{SL}(t) / \sum \lambda(t)$, i.e. the service level weighed by the arrival rate $\lambda(t)$ over the remainder of the day, under a certain percentage. For the foreign passengers the performance constraint is $\mathrm{SL} \leq 0.01$ with $W_{\max} = 10$ minutes. The performance evaluation is done with the deterministic queueing model and the arrival rate samples generated by the Monte Carlo simulation model.

In the staffing literature of call centers it is frequently assumed that the system is in steady-state during each staffing interval or that there are customer abandonments in the case of overload. At immigration there are no abandonments and the steady-state assumption is inappropriate because of severe overload. In the airport staffing literature a deterministic fluid model has been applied for immigration staffing but uncertainty in delay was not taken into account. We analyzed the waiting time performance of the deterministic fluid model with uncertain demand and we concluded that the performance is inadequate. We have extended the deterministic fluid model with delay uncertainty by converting the deterministic staffing function into staff probabilities. The staffing levels are then set by determining the appropriate quantiles of the staff probabilities. In addition we also applied the square-root staffing approach, $s = r + \beta\sqrt{r}$, to our problem and determined the appropriate values of β to meet the service level requirement. With both methods we can satisfy the daily service level SL requirement. However during periods with low arrival rates and at the end of the day the service level in an interval $\mathrm{SL}(t)$ can reach unsatisfactory levels. We propose a new service level constraint: the service level during any interval $\mathrm{SL}(t)$ should be at most 0.03. To meet this new constraint we have developed an iterative algorithm. The initial staffing levels can be set according to the determin-

istic fluid model, the staff probability quantiles or the square-root staffing formula. Then the service level in each interval SL(t) is calculated. If the constraint is violated in any staffing interval, the number of staff in that interval is increased by one. We do this iteratively until the performance constraint is satisfied in each staffing interval.

In practice we can support the decision makers at Narita immigration with our models. First we can provide the arrival forecast with the arrival probabilities, the expected arrival rate, the 95% upper bound and a sample path with the deterministic approximation. Second we can provide a staff forecast with the staff probabilities and the upper bound of the required number of staff. Using the quantiles for the staff probabilities or the square-root staffing formula we can also give a recommendation for the staffing function such that the daily service level requirement is met. If a more robust solution is desired then the iterative algorithm can be used but it will require longer computation time to generate arrival rate samples and to simulate waiting time probabilities of the queueing system.

For future research we can improve the arrival forecasting model and extend the staffing model. We have collected all of our data manually by counting and video recordings. Passenger tracking systems have recently been implemented in various airports all over the world to monitor the movement and waiting times of the passengers. Such systems should be implemented to collect more historical data and more accurate data. Additional walking time measurements are recommended to derive accurate walking speed distributions for each gate. The estimated times of arrival that we have collected every 10 minutes for more than a year can be further analyzed and a statistical model could be developed to predict the flight delay.

The staffing model can be extended in several ways. We have focused only on the foreign passengers and we have simplified the queues for the foreigners by combining the reentry and non-reentry passengers. Staffing is only one of the steps in personnel management. Further research should include the scheduling and rostering of immigration staff. We can also consider staffing of immigration at multiple terminals simultaneously because the staff can be moved between terminals during peak times. We have only looked at the arrival immigration service but on the departure side there is also an immigration service which poses different challenges because we also have to consider the check-in and security check before departure immigration.

Bibliography

[1] Z. Aksin, M. Armony, and V. Mehrotra. "The modern call-center: a multi-disciplinary perspective on operations management research". In: *Production and Operations Management* 16.6 (2007), 665–688.

[2] S. Aldor-Noiman, P.D. Feigin, and A. Mandelbaum. "Workload forecasting for a call center: methodology and a case study". In: *The Annals of Applied Statistics* 3.4 (2009), 1403–1447.

[3] A. Avramidis, A. Deslauriers, and P. L'Ecuyer. "Modeling daily arrivals to a telephone call center". In: *Management Science* 50.7 (2004), 896–908.

[4] Y. Bai. "Analysis of Aircraft Arrival Delay and Airport On-Time Performance". MSc thesis. University of Central Florida, 2006.

[5] A. Barickman, E. Sebenius, and H. Sohi. "A Practical Approach to Boarding/Deboarding an A380". In: *Control* 1724 (2007).

[6] A. Bassamboo, R. Randhawa, and A. Zeevi. "Capacity sizing under parameter uncertainty: Safety staffing principles revisited". In: *Management Science* 56.10 (2010), 1668–1686.

[7] J.H. Bookbinder and D.L. Martell. "Time-Dependent Queueing Approach to Helicopter Allocation for Forest Fire Initial-Attack". In: *INFOR* 17 (1979), pp. 58–70.

[8] S. Borst, A. Mandelbaum, and M.I. Reiman. "Dimensioning large call centers". In: *Operations Research* 52.1 (2004), 17–34.

[9] L. Brown et al. "Statistical Analysis of a Telephone Call Center: A Queueing Science Perspective". In: *Journal of the American Statistical Association* 100.1 (2005), pp. 36–50.

[10] M.J. Brusco et al. "Improving Personnel Scheduling at Airline Stations". In: *Operations Research* 43.5 (1995), pp. 741–751.

[11] N. Channouf and P. L'Ecuyer. "A Normal Copula Model for the Arrival Process in a Call Center". In: *International Transactions in Operational Research* 19 (2012), 771–787.

[12] Narita International Airport Corporation. *Flights Today*. URL: `http://www.narita-airport.jp/en/flight/today.html#all`.

[13] W. Daamen. "Modelling Passenger Flows in Public Transport Facilities". PhD thesis. Delft University of Technology, 2004.

[14] M. Defraeye and I. Van Nieuwenhuyse. "Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm". In: *Decision Support Systems* 54 (2013), 1558–1567.

[15] M. Defraeye and I. Van Nieuwenhuyse. "Staffing and scheduling under nonstationary demand for service: A literature review". In: *Omega* 58 (2016), pp. 4–25.

[16] J.D. Diener and W.H. Sanders. "Emperical Comparison of Uniformization Methods for Continuous-Time Markov Chains". In: *Computations with Markov Chains*. Ed. by W.J. Stewart. 1995, pp. 547–570.

[17] S. Ding and G. Koole. "Optimal Call Center Forecasting and Staffing under Arrival Rate Uncertainty". 2014.

[18] A. Durrande-Moreau. "Waiting for service: ten years of empirical research". In: *International Journal of Service Industry Management* 10.2 (1999), pp. 171–194.

[19] S.G. Eick, W.A. Massey, and W. Whitt. "Mt/G/∞ queues with sinusoidal arrival rates". In: *Management Science* 39 (1993), 241–252.

[20] T. Feldhoff. "Japan's Capital Tokyo And Its Airports: Problems And Prospects From Subnational And Supranational Perspectives". In: *Journal of Air Transport Management* 9 (2003).

[21] Z. Feldman et al. "Staffing of Time-Varying Queues to Achieve Time-Stable Performance". In: *Management Science* 54.2 (2008), 324–338.

[22] W. Feller. *An Introduction to Probability Theory and its Applications*. 3rd ed. Vol. I. New York: Wiley, 1968.

[23] M.C. Fu, S.I. Marcus, and I.J. Wang. "Monotone optimal policies for a transient queueing staffing problem". In: *Operations Research* 48.2 (2000), 327–31.

[24] N. Gans, G. Koole, and A. Mandelbaum. "Telephone call centers: Tutorial, review and research prospects". In: *Manufacturing and Service Operations Management* 5.2 (2003), 79–141.

[25] W.K. Grassmann. "Finding the Right Number of Servers in Real-World Queuing Systems". In: *Interfaces* 18.2 (1988), pp. 94–104.

[26] L.V. Green, P.J. Kolear, and J. Soares. "Improving the SIPP Approach for Staffing Service Systems That Have Cyclic Demands". In: *Operations Research* 49.4 (2001), pp. 549–564.

[27] L.V. Green and P.J. Kolesar. "The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals". In: *Management Science* 37.1 (1991), pp. 84–97.

[28] L.V. Green, P.J. Kolesar, and A. Svoronos. "Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems". In: *Operations Research* 39.3 (1991).

[29] L.V. Green, P.J. Kolesar, and W. Whitt. "Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System". In: *Production and Operations Management* 16.1 (2007), pp. 13–39.

[30] L.V. Green and J. Soares. "Computing Time-Dependent Waiting Time Probabilities in M(t)/M/s(t) Queuing Systems". In: *Manufacturing & Service Operations Management* 9.1 (2007), pp. 54–61.

[31] C.M. Grinstead and J.L. Snell. *Introduction to Probability.* 2nd. American Mathematical Society, 2003.

[32] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory.* 4th. New York: Wiley, 2008.

[33] S. Halfin and W. Whitt. "Heavy-traffic limits for queues with many exponential servers". In: *Operations Research* 29.3 (1981), 567–587.

[34] F.S. Hillier and G.J. Lieberman. *Introduction to Operations Research.* 7th. New York: McGraw-Hill, 2001.

[35] T.J. Holloran and J.E. Byrn. "United Airlines station manpower planning system". In: *Interfaces* 16.1 (1986), pp. 39–50.

[36] T. Horstmeier and F. de Haan. "Influence of ground handling on turn round time of new large aircraft". In: *Aircraft Engineering and Aerospace Technology* 73.3 (2001), pp. 266–271.

[37] IATA. *Air Passenger Monthly Analysis.* 2015. URL: `http://www.iata.org/publications/economics/Pages/Air-Passenger-Monthly-Analysis.aspx`.

[38] IATA. *Annual Report 2009.* 2009. URL: `https://www.iata.org/pressroom/Documents/IATAAnnualReport2009.pdf`.

[39] R. Ibrahim et al. "On the modeling and forecasting of call center arrival". In: *Proceedings of the 2012 Winter Simulation Conference.* Ed. by C. Laroque et al. 2012.

[40] H. Idris et al. "Queuing Model for Taxi-out Time Estimation". In: *Air Traffic Control Quarterly,* 10.1 (2001).

[41] Incheon Airport. *A Story Begins... 2012 Brochure.* 2012. URL: `https://www.airport.kr/iiacms/pageWork.iia?_scode=C3006020401`.

[42] A. Ingolfsson. "Modeling the M(t)/M/s(t) Queue with an Exhaustive Discipline". 2005.

[43] A. Ingolfsson et al. "A Survey and Experimental Comparison of Service-Level-Approximation Methods for Nonstationary M(t)/M/s(t) Queueing Systems with Exhaustive Discipline". In: *Journal on Computing* 19.2 (2007), pp. 201–214.

[44] Japan Today. *Japan eyes simpler immigration procedures, including automatic gate*. 2011. URL: http://www.japantoday.com/category/national/view/japan-studying-simpler-immigration-procedures-including-automatic-gate.

[45] O.B. Jennings et al. "Server staffing to meet time-varying demand". In: *Management Science* 42 (1996), 1383–1394.

[46] T. Jiménez and G. Koole. "Scaling and comparison of fluid limits of queues applied to call centres with time-varying parameters". In: *OR Spectrum* 26.3 (2004), pp. 413–422.

[47] G. Jongbloed and G. Koole. "Managing uncertainty in call centres using Poisson mixtures". In: *Applied Stochastic Models in Business and Industry* 17.4 (2001), 307–318.

[48] P. Kolesar. "Stalking the endangered CAT: A queueing analysis of congestion at automatic teller machines". In: *Interfaces* 14.16 (1984), pp. 16–26.

[49] P. Kolesar et al. "A queueing linear programming approach to scheduling police cars". In: *Operations Research* 23.6 (1975), pp. 1045–1062.

[50] B.O. Koopman. "Air-terminal queues under time-dependent conditions". In: *Operations Research* 20.6 (1972), pp. 1089–1114.

[51] S. Liao et al. "Staffing a call center with uncertain non-stationary arrival rate and flexibility". In: *OR Spectrum* 34 (2012), 691–721.

[52] R.A. Littler and D. Whitaker. "Estimating staffing requirements at an airport terminal". In: *Journal of the Operational Research Society* 48 (1997), pp. 124–131.

[53] D. H. Maister. "Psychology of Waiting Lines". In: *The Service Encounter*. Ed. by J. Czepiel. Lexington, MA: Lexington Books, 1985, pp. 113–23.

[54] A. Mandelbaum, W.A. Massey, and M.I. Reiman. "Strong approximations for Markovian service networks". In: *Queueing Systems* 30.1-2 (1998), pp. 149–201.

[55] A.J. Mason, D.M. Ryan, and D.M. Panton. "Integrated Simulation, Heuristic and Optimisation Approaches to Staff Scheduling". In: *Operations Research* 46.2 (1998), pp. 161–175.

[56] A.J. Mason, D.M. Ryan, and D.M. Panton. "Staff Planning at Auckland International Airport". In: *Proceedings of the 30th Annual Conference of the Operational Research Society of New Zealand*. 1994, pp. 112–117.

[57] W.A. Massey and W. Whitt. "An analysis of the modified offered-load approximation for the nonstationary Erlang loss model". In: *The Annals of Applied Probability* 4.4 (1994).

[58] W.A. Massey and W. Whitt. "Peak Congestion in Multi-Server Service Systems with Slowly Varying Arrival Rates". In: *Queueing Systems* 25.4 (1997), pp. 157–172.

[59] E. Mueller and G. Chatterji. *Analysis of Aircraft Arrival and Departure Delay Characteristics*. Los Angeles, California: AIAA's Aircraft Technology, Integration, and Operations Forum, 2002.

[60] Narita International Airport Corporation. *Monthly Traffic Statistics*. 2015. URL: `http://www.naa.jp/en/airport/traffic.html`.

[61] Narita International Airport Corporation. *Statistics (H23)*. 2012.

[62] R. De Neufville and A.R. Odoni. *Airport Systems: Planning, Design, and Management*. New York: McGraw-Hill, 2003.

[63] G.F. Newell. *Applications of Queueing Theory*. 2nd. London: Chapman and Hall, 1982.

[64] H. Nikoue et al. "Passenger Flow Prediction at Sydney International Airport: a data-driven queuing approach". 2015.

[65] International Civil Aviation Organization. *Advanced Technologies For Facilitation At Airports*. Working Paper. Assemby – 37th Session. ICAO, Sept. 2010.

[66] International Civil Aviation Organization. *International Standards And Recommended Practices. Annex 9 to the Convention on International Civil Aviation*. 10th. ICAO.

[67] M.H. Rothkopf and S.S. Oren. "A Closure Approximation for the Nonstationary M/M/s Queue". In: *Management Science* 25.6 (1979), pp. 522–534.

[68] A. Roubos, S. Bhulai, and G. Koole. *Flexible staffing for call centers with non-stationary arrival rates*. Working paper. VU University Amsterdam, 2011.

[69] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming*. Philadelphia: Society for Industrial and Applied Mathematics, 2009.

[70] G. Skaltsas. "Analysis of Airline Schedule Padding on U.S. Domestic Routes". MSc thesis. Massachusetts Institute of Technology.

[71] Skytrax. *World Airport Awards*. URL: `http://www.worldairportawards.com`.

[72] R. Stolletz. "Approximation of the non-stationary M(t)/M(t)/c(t)-queue using stationary queueing models: The stationary backlog-carryover approach". In: *European Journal of Operational Research* 190.2 (2008), 478–493.

[73] J.W. Taylor. "Density Forecasting of Intraday Call Center Arrivals Using Models Based on Exponential Smoothing". In: *Management Science* 58.3 (2012), pp. 534–549.

[74] The Travel Insider. *Common Airplane Types Configuration Data*. 2015. URL: `http://www.thetravelinsider.info/airplanetypes.htm`.

[75] G.M. Thompson. "Accounting for the multi-period impact of service when determining employee requirements for labor scheduling". In: *Journal of Operations Management* 11.3 (1993), pp. 269–287.

[76] V. Tosic. "A review of airport passenger terminal operations analysis and modelling". In: *Journal of the American Statistical Association* 26.1 (1992), pp. 3–26.

[77] Y. Tu, M. Ball, and W. Jank. "Estimating Flight Departure Delay Distributions - A Statistical Approach with Long-term Trend and Short-term Pattern". In: *Journal of the American Statistical Association* 103.481 (2008).

[78] J. Weinberg, L.D. Brown, and J.R. Stroud. "Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data". In: *Journal of the American Statistical Association* 102 (2007), 1185–1199.

[79] W. Whitt. "Dynamic staffing in a telephone call center aiming to immediately answer all calls". In: *Operations Research Letters* 24 (1999), 205–212.

[80] W. Whitt. "Fluid models for many-server queues with abandonments". In: *Operations Research* 54.1 (2006), 37–54.

[81] W. Whitt. "Staffing a call center with uncertain arrival rate and absenteeism". In: *Production and Operations Management* 15.1 (2006), 88–102.

[82] W. Whitt. "Understanding the efficiency of multi-server service systems". In: *Management Science* 38 (1992), 708–723.

[83] Wikipedia. *List of the world's busiest airports by international passenger traffic*. URL: `https://en.wikipedia.org/wiki/List_of_the_world%27s_busiest_airports_by_international_passenger_traffic`.

[84] T. Willemain. *Estimating Components of Variation in Flight Time*. Technical report. The National Center of Excellence for Aviation Operations Research (NEXTOR), 2001. URL: `http://www.nextor.org/pubs/WP-01-2.pdf`.

[85] T. Willemain et al. *Factors Influencing Estimated Time En Route*. Technical report. The National Center of Excellence for Aviation Operations Research (NEXTOR), 2003.

[86] P. Pao-Yen Wu and K. Mengersen. "A review of models and model usage scenarios for an airport complex system". In: *Transportation Research Part A* 47 (2013), 124–140.

[87] S.B. Young. "'Evaluation of Pedestrian Walking Speeds in Airport Terminals". In: *Transportation Research Record: Journal of the Transportation Research Board* 1674 (1999), pp. 20–26.