Optimal Staffing Policy
A Service System with Stochastic Travel Times

M. Al-Foraih, P. Johnson, G. Evatt and P. Duck
School of Mathematics, University of Manchester, Oxford Road, Manchester M13 9PL, U.K.

Keywords: Workforce Staffing, Stochastic Models, Time-Varying Queues, Network Queues.

Abstract: Private sector operators of response services such as ambulance, fire or police etc. are often regulated by targets on the distribution of response times. This may result in inefficient overstaffing to ensure those targets are met. In this paper, we use a network chain of $M/M/K$ queues to model the arrival and completion of jobs on the system so that quantities such as the expected total time waiting for all jobs can be calculated. The Markov nature enables us to evoke the Hamilton Jacobi Bellman equation (HJB) principle to optimize the required number of staff whilst still meeting targets.

1 INTRODUCTION

There are strong economic arguments for introducing competition from the private sector into state emergency response services, in order to improve efficiency and reduce costs (Blackstone et al., 2007). The state requires that the response service must try to minimize the time it takes to reach an a emergency call, as in a case of an ambulance or a fire fighters it could save lives. Now, a private sector operator will always look to minimise their costs so they must be regulated effectively to maintain service to a required level. In this paper we outline a framework to solve for the optimal staffing policy of the private sector firm given a regulatory framework by weighting the benefit of not missing targets with the cost of staff given stochastic uncertainty in the demand for service. This paper has been divided into 5 sections. In section 2 we discuss the literature review and and the previous work related to this problem. Then, in section 3, we introduce and define the problem framework and the partial differential equation (PDE) to solve it. In section 4 we demonstrate the numerical result and the conclusion are found in section 5.

2 LITERATURE REVIEW

There are variety of problems associated with the optimal management of emergency response systems, and the location of new and existing distribution centers in particular has received a lot of attention (see Snyder, 2006, for an overview). A notable work is that of Singer and Donoso (2008) who combine the problem of location with queueing theory. They used key performance indicators such as the average response time that improves the operation of ambulance service providers. Some authors such as Marianov and ReVelle (1996) have attempted to model both location and resource/staffing levels with queues. In their paper, they outline how to calculate the probability that a given distribution centres will have spare capacity given current resource levels (the number of vehicles as opposed to staff in our framework), and then using a linear programming approach they can optimise the position of a limited number of vehicles at distribution centres to best cover the area. The results though are time invariant and cannot be applied to something as dynamic as staffing levels where shifts patterns change, and agency staff might be hired to cover short term spikes in demand.

The optimal resource problem of how many staff to employ has in general received much less attention in the literature than the related location or coverage. There are some exceptions such as Fry MJ (2006) who examined the problem of determining the number of employees that would minimise the cost of hiring in a fire department, subject to state staffing requirements. He applied news-vendor type models where the uncertainty depends on the decision variables. Nevertheless, this approach lacked the cost of hiring and assumed a discrete time model. The model within this paper is an extension of the stan-

Al-Foraih M., Johnson P., Evatt G. and Duck P..
Optimal Staffing Policy - A Service System with Stochastic Travel Times.
DOI: 10.5220/0004276902580263
In *Proceedings of the 2nd International Conference on Operations Research and Enterprise Systems* (ICORES-2013), pages 258-263
ISBN: 978-989-8565-40-2

dard queueing theory, utilising a Markov chain of $M/M/K$ queues. There have been countless other applications of these theories in problems as diverse as highway traffic queues Evans et al. (1964) or management of call centres (Aksin et al., 2007; Whitt, 2005). There are some examples of queueing theory within the problem of optimal staffing, such as Brooks et al. (2011). They investigated the problem of care-at-home services using a Markov decision process with queues aiming to balance the cost of staffing against the cost of rejection. In our model we assume that the response service provider wishes to balance costs of staffing against the response times, since the regulator has incentivised that with targets. Another difference between our approach and that of Brooks et al. (2011) is that the inclusion of different classes of patients resulted in the full problem becoming computationally intractable. Consequently, they can only provide heuristic bounds to guide the optimal policy. We aim to keep our model relatively simple so that the solutions can be found numerically in a reasonable time frame.

So the unique contribution of this paper is to introduce a time dynamic optimal control problem where the control is the number of staff available, rather than admittance or direction of calls on the system. This is enabled by offsetting the expected total time between receipt to arrival versus the cost of employing staff. The result will be an optimal policy where the optimal number of staff will be a function both of time and the current state of the system. By including one off costs at the start and the end of a shift, we can ensure that the policy will result in realistic shift patterns. The weighting of staff costs can then be varied so as to establish the minimum staffing policy to meet the required targets.

3 THE MODEL

Let us define a call to the service provider as a job J which must be completed, and the staff member that responds to that we shall denote as an engineer. The engineer could be a policeman, fireman, ambulance engineer, or a technician in the case of leakage from a distribution network depending on the service provider we are solving for. The different stages in job completion are defined by the four times

- t_b^i the time that the ith job is booked on the system.
- t_t^i the time that the ith job is assigned to an engineer.
- t_a^i the time of arrival of the engineer at the ith job.
- t_d^i the time that the engineer finishes the ith job.

Therefore, we define the ith job vector to be

$$J_i = (t_b^i, t_t^i, t_a^i, t_d^i), \tag{1}$$

and the set of jobs $\mathcal{J} = \{J_1, J_2, ..J_I\}$ represents all jobs received over a fixed period of time. Furthermore, we define $\eta = t_a - t_t$ to be the travel time for an engineer to reach their destination, and $\xi = t_d - t_a$ to be the total time which it takes for the engineer to finish the job and become available again. We divide time into K discrete periods of length Δt, where the values of K and Δt may change depending on what we wish to calculate. We can define the time at a given period to be

$$t^k = t_0 + k\Delta t \text{ with } k = 0, 1, .., K \tag{2}$$

where t_0 is the reference time that is before the first time/date in t_k period of time, t_K is after the last time/date, and $\Delta t = (t_K - t_0)/K$. Now define the number of jobs that have been received on the system within the period t_k to be B^k, and the number received but are yet to be completed at the time period t^k as Q^k. Next we define the response time as the time between the job being received on the system and the arrival time of the engineer, or $T = t_a - t_b$. So to get an estimation of the response time, we must split the jobs into three different types

- W the number of jobs waiting for the assignment of an engineer.
- R the number of jobs where an engineer is in transit to reach a target
- A the number of jobs for which the engineer has arrived but has not yet completed the job.

A job will count towards W if the call has been received before (or at) the end of the current period, but the engineer does not set off until after the end of that period, counted in transit R if the engineer has set off but not arrived, and counted as an arrived job A if the engineer arrives before (or at) the end of the current period and leaves after the start of the next period. Obviously it follows that

$$Q^k = W^k + R^k + A^k, \tag{3}$$

or Q is sum of all jobs waiting to be completed. We impose no restriction on the number of jobs awaiting completion but we must stipulate that the number of engineers S available at time t^k is greater than the number in transit or at a job:

$$R^k + A^k \leq S^k. \tag{4}$$

If we assume that the optimal way to manage the system is to serve costumers on a first-come first served basis, (although this might not always be the case if S varies with time and we know some trips might take a lot longer than others) then if $R^k + A^k < S^k$ we have spare capacity and any new job that arrives will immediately have a engineer sent out. Therefore, if $R^k + A^k < S^k$ then we have a sufficient number of engineers to S^k to cover all the jobs requested or in other words there are no job waiting for an engineer to be assigned ($W^k = 0$). Conversely, we know that if jobs are waiting to be assigned, all engineers must be out on jobs, so that $W^k \geq 0$ if $R^k + A^k = S^k$. So, even though it appears as though there are three degrees of freedom there are in fact only two, which means that predicting the total number of jobs waiting to finish and the number of engineers travelling to reach a target is enough information to tell us how many are waiting and how many have arrived. Now, we can easily calculate the cumulative total time spent by each job between it being booked onto the system and an engineer arriving, denoted by TT, which is given by the formula

$$TT = \sum_i (t_a^i - t_b^i) = \sum_k (W^k + R^k)\Delta t + O(\Delta t) \qquad (5)$$
$$= \int_{t_0}^{t_K} W(t) + R(t) dt.$$

The integral form of the equation is valid as $\Delta t \to 0$ and $K \to \infty$. This quantity gives the total time given all jobs that are taken in the period, but we are interested in the average receipt to arrival time T for a single job. Then the average receipt to arrival time for a job $T = t_a - t_b$ over the given time period is

$$E[T] = \frac{\sum_i (t_a^i - t_b^i)}{I} = \frac{1}{I} \int_{t_0}^{t_K} W(t) + R(t) dt \qquad (6)$$

where I is again the total number of jobs. The equations (5) and (6) provides a link between the distribution of T (the response times) and quantities that are easily calculated using continuous time differential equations.

We assume that jobs arriving on the system follows a Poisson process with a time varying intensity denoted as $\lambda(t)$. Since the queues in $M/M/K$ Markovian process has a Poisson process and therefore the inter arrival times should be exponentially distributed. The memoryless nature of the exponential distributions ensures that we have a Markov chain even if the mean of the process is time varying. Therefore,we set travel times to be exponentially distributed (and therefore memoryless) with an average length of $\bar{\eta}(t)$, likewise completion times of the job are also exponentially distributed with $\bar{\xi}(t)$. Now define the quantity $V(Q, R, t)$ to be some value function depending on the stochastic processes followed by Q and R. Given these stochastic jump processes, we can evaluate what is the expected change in value of the system over a small increment of time.

$$\begin{aligned} E[dV] = {} & \frac{\partial V}{\partial t} dt + \qquad (7) \\ & dt \Big[\lambda(t)(V(Q+1, R+\psi, t+dt) - V(Q,R,t)) + \\ & \frac{R}{\bar{\eta}(t)}(V(Q, R-1, t+dt) - V(Q,R,t)) + \\ & \mu(Q,R,t)(V(Q-1, R+\phi, t+dt) - V(Q,R,t)) \Big] \end{aligned}$$

where $\mu(Q,R,t) = \frac{\min(Q,S)-R}{\bar{\xi}(t)}$ and

$$\begin{aligned} \psi = 1, \phi = 0 & \text{ if } Q < S \qquad (8) \\ \psi = 0, \phi = 0 & \text{ else if } Q = S \\ \psi = 0, \phi = 1 & \text{ else if } Q > S \end{aligned}$$

3.1 Optimising the Number of Engineers

In order to optimise the number of engineers, we need to define some cash value to $E[TT]$, and offset it against the cost of employing engineers. The value TT has the units [time jobs] and we assign this conversion from time jobs to cash or in other words, the cost of delay in jobs by the variable α. The variable α must have the units [£ time^{-1} jobs^{-1}] and can loosely be interpreted as the negative value added to the system by increasing the total receipt to arrival time by one unit. If there is some fixed cost to hire engineers that is paid continuously at the rate ε (units [£ time^{-1} jobs^{-1}]) then we must set

$$dV = -\alpha(R + \max(Q-S, 0))dt - \varepsilon S dt. \qquad (9)$$

Equation (9) is used to calculate the total cost of hiring S engineers plus the cost of TT over the time period $t \in [t_0, t_K]$. Obviously we would like to find the strategy that hires engineers in some sort of shift pattern in order to cover the jobs in the optimally. In order to do this, we must allow S to vary with time - in fact S will be our control variable. We need to consider what happens when the value of S is increased, decreased or stays the same throughout time. This means that our value function will now become an implicit function of the control variable S as well as the other variables, so that $V = V(Q, R, t; S)$.

3.1.1 Increasing the Number of Engineers

Since changing the number of engineers happens over an instant, we define $^+$ to mean just after and $^-$ to mean just before. If $S^+ > S^-$, then we have hired more engineers. This means that any spare jobs waiting to be assigned can be handed to the new engineers, which results in

$$Q^{+} = Q^{-} \qquad \text{and} \tag{10}$$
$$R^{+} = R^{-} + \max((\min(Q^{-},S^{+}) - S^{-}),0)$$

and if there is some fixed cost ε_s assigned to the engineers starting the shift (this will stop engineers being used for a short period of time) we have

$$V(Q^{-},R^{-},t^{-};S^{-}) = V(Q^{+},R^{+},t^{+};S^{+}) \tag{11}$$
$$+\varepsilon_s|S^{+} - S^{-}|$$

3.1.2 Decreasing the Number of Engineers

When an engineer finishes their shift, we can assume that if they are already on their way to a job then they must finish it, and those leaving must be either spare engineers or those already at jobs. This means that the number of jobs waiting for an engineer and the number in transit must stay constant. So if $S^{+} < S^{-}$ then $R^{-} = R^{+}$.

$$R^{+} = R^{-} \qquad \text{and} \tag{12}$$
$$Q^{+} = Q^{-} - \min(\min(Q^{-},S^{-}) - R^{-},$$
$$\max(\min(Q^{-},S^{-}) - S^{+},0))$$

and

$$V(Q^{-},R^{-},t^{-};S^{-}) = V(Q^{+},R^{+},t^{+};S^{+}) \tag{13}$$
$$+\varepsilon_f|S^{+} - S^{-}|$$

where ε_f is some fixed cost associated with finishing a shift (maybe equal to ε_f the average cost of finishing a job if they are on their way to or yet to finish one). Furthermore, we now have to consider what happens to the partial differential equation (PDE) if $R > S$, a situation which might arise if engineers finish shifts whilst they are on their way to a job. In this case we have to add an extra equation to the calculation of dV for this case when $R > S$. We are assuming that any jobs in which the engineers have arrived are lost from the system along with any jobs that arrive at their destination in the state $R > S$. This loss from the system and the cost of finishing the job is captured by fixed costs for reducing the number of engineers. Therefore we have,

$$dV = \frac{\partial V}{\partial t}dt +$$
$$\lambda(t)dt(V(Q+1,R,t+dt) - V(Q,R,t)) +$$
$$\frac{R}{\bar{\eta}(t)}dt(V(Q-1,R-1,t+dt) - V(Q,R,t)). \tag{14}$$

3.1.3 Solving the PDE

Now let us define the operator $\mathcal{L}$ such that the combination of equations (7), (8),(13) and (14) can be captured by the single equation

$$\frac{\partial V}{\partial t} - \mathcal{L}(V) = -\alpha(R + \max(Q-S,0)) - \varepsilon S \tag{15}$$

In order to solve and optimise the equations we must solve backwards in time . The value at the end of the period should be assigned so that it is suboptimal to allows jobs to remain unfinished, such as $V(Q,R,t_K;S) = PQ$ where P is some penalty for every job remaining unfinished. We proceed to solve the equations as follows. First, given that we have already solved at t^{k+1}, we must solve for all possible values of Q, R, and S in the state space using the scheme

$$V^{k}(Q,R;S) = V^{k+1}(Q,R;S) \tag{16}$$
$$+\Delta t\left[\mathcal{L}(V^{k+1}) + \alpha(R + \max(Q-S,0)) + \varepsilon S\right]$$

Once all values have been found we can apply the optimising condition (here we must minimise the value) by checking at each Q, R, what the optimal number of engineers is. This can be written as

$$(V^{*})^{k}(Q^{-},R^{-}) = \min_{E}\left[V^{k}(Q^{+},R^{+};E) + C_{S,E}\right] \tag{17}$$

where $C_{S,E}$ is the appropriate cost function from equations (10)-(13) and V^{*} is the optimal value. Once the conditions of (10)-(13) are applied in (15) we can then solve backwards in time until t_0 which will be the optimal grid.

4 NUMERICAL RESULTS

4.1 Calculating the Optimal Number of Engineers through the Day

Unless stated otherwise the results in this section assume that we have

$$\lambda(t) = 2.5 \text{ if } t \in [9,16] \tag{18}$$
$$\lambda(t) = 0 \text{ otherwise.}$$

and $\bar{\eta} = 1.5, \bar{\xi} = 0.5, t_k = 24\varepsilon = 0.4, \alpha = 3.5, \varepsilon_s, \varepsilon_f = 2.5$ and $K = 480$. In figure 1, we plot the mean path followed by both of the state variables Q and R in the problem, the corresponding choice of optimal S^{*} (given the mean path), alongside the intensity at which jobs appear on the system. We can observe that S^{*} fluctuates depending on the number of jobs Q, keeping the total number of engineers close to the current number of jobs on the system. As such in the region $t \in [0,9]$ we hire no engineers at all since there is no jobs, then increase the number of engineers up to time $t \in [12,16]$ which is the peak time, and then gradually decrease numbers until all engineers are finished their shifts at time $t = 21$. Given the parameters as stated, the number of engineers S is always close to the value of Q, keeping the cost of engineers to

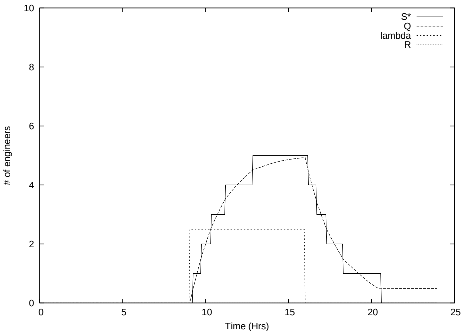

Figure 1: The optimal hiring policy S^* with $\bar{\eta} = 1.5, \bar{\xi} = 0.5, t_k = 24\epsilon = 0.4, \alpha = 3.5\epsilon_s, \epsilon_f = 2.5$ and $K = 480$.

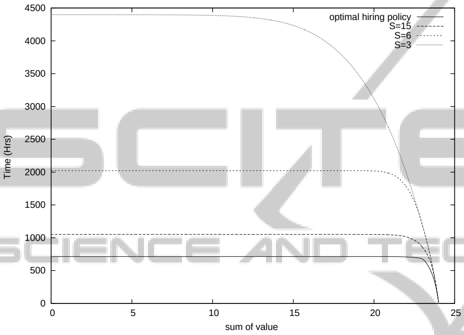

Figure 2: The optimal hiring policy of the engineers through out the day in comparison with $S = 3, 6$ and 15.

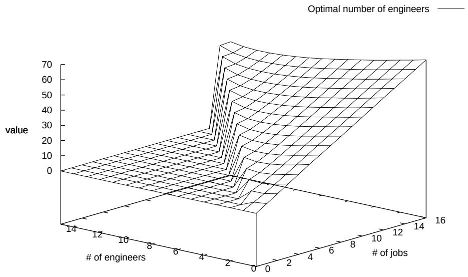

Figure 3: The optimal hiring policy of the engineers through out the day.

a minimum. This ensures staffing and prevents jobs left unattended. However, we can also observe how the cost of engineers can vary under certain parameter choices. In the same figure, we can see that R fluctuates with the number of of engineers hired. The difference between the number of engineers hired S and the travelling engineers R is controlled by $\bar{\xi}$ and $\bar{\eta}$, such that as we increase $\bar{\eta}$ and $\bar{\xi}$ then the difference between S and R decreases and *vice versa*.

To demonstrate that optimal choices are made for the number of engineers we compare

$$\min_{E}[\sum_{i=0}^{Q_{max}} \sum_{j=0}^{Q_{max}} V(Q_i, R_i, E)] \text{ for all } E \qquad (19)$$

for each time t and choice of engineer S. with $\lambda = .5$ for all t. We can then compare the effect of different hiring policies with the respective optimal solutions. In figure 2 we can see that this sum over all values for constant hiring policies $S = 3, 6$ and 15 is always higher than the optimal policy S^* for all t. Although this is no way rigorous, it does show the effect of hiring the optimal number of engineers on $V(Q, R, t; S)$.

In figure 3 we set $Q_{max} = 16$. In this figure, we have $V(Q, R, S)$ at time $t = 0$ for different values of Q and R where each represents the expected time it takes R number of travelling engineers finish Q jobs with $S = 0$. Most significantly, for $Q = Q_{max}$ and $R = 0$ or in other words, there is Q job demand and no engineers available to travel, then $V(Q, R, S) = 68.9$ as evident from the figure. It represents the expected total travelling (or waiting time for calls). Additionally, we can see that if $R > Q$ then $V(Q, R, S) = 0$ and all the jobs will be covered by R engineers and there is no waiting time. To show the effect of optimizing the number of engineers, we compare the optimal hiring policy with $S = 6$ and $S = 15$. As we can see in figure 4, using the same previously mentioned parameters, $V(Q, R, S)$ at time $t = 0$ for $S = 15$ is always higher, in comparison with the optimal hiring policy S^*, for all values of Q and R. The expected total waiting time $V(Q, R, S) = 2944$ for $S = 15$ is significantly higher than using the optimal hiring policy. Furthermore, in figure 5 we compare the optimal solution with $S = 6$ and in this case we set $Q_{max} = 20$. In this figure, we can see that the expected travelling time is for $S = 6$ is higher than the optimal hiring policy for all values of Q and R where for $Q = Q_{max}$ and $R = 0$ then $V(Q, R, S) = 4002$ for $S = 6$, and $V(Q, R, S) = 120.5$ for S^*.

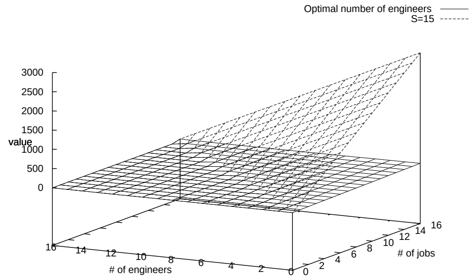

Figure 4: The hiring policy of the engineers through out the day with $S = 15$.

4.2 The Effects of the Cost Parameters ϵ and α

In figure 6, we set $\lambda(t)$ as in equation (19) and $\bar{\xi} = 0.5, \bar{\eta} = 1.5\alpha = 3.5, K = 480, \epsilon_s = 2.5, \epsilon_f = 2.5$ with the maximum of 10 engineers. As we decrease ϵ then the optimal number of engineers required to complete

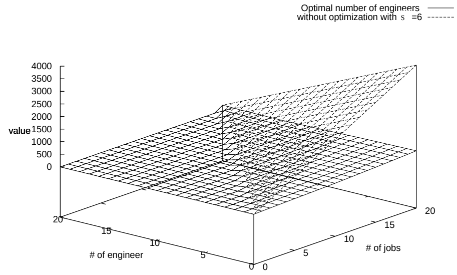

Figure 5: The optimal hiring policy of the engineers through out the day with $S = 6$.

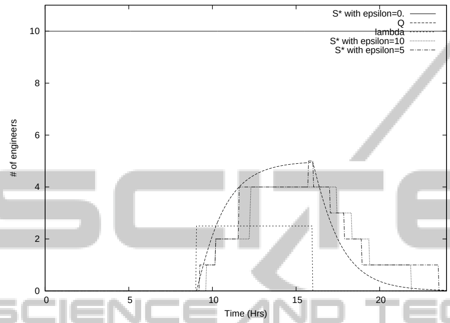

Figure 6: The effect of the cost employing ε the optimal hiring policy of engineers through out the day.

the job increases. This is because ε is the cost of hiring and because it becomes cheaper to hire engineers therefore we can invest in more of them. We notice that for $\varepsilon = 0$ it is optimal to hire $S = 10$ for all t since it is effectively free to hire them. For $\varepsilon = 5$ and $\varepsilon = 10$ the number of engineers hired are actually less than the number of jobs Q since it is too expensive to hire engineers.

5 CONCLUSIONS

In this paper, we have developed a model to optimize the number of engineers hired by the response service operator. We modelled the arrival and completion of jobs of a response as $M/M/K$ queues using a time dynamic stochastic processes and allowed the operator to the control the number of staff. We defined the problem as a jump process PDE and categorized the jump into three different types depending on the current state of the system. The structure of the model is developed from queueing theory but we put a real options style slant on the model in which the operator must invest fixed costs to start and end the shift patterns of their staff. The main results show that the optimal staffing policy can effectively balance the costs of employing staff with the total waiting time, allowing an operator to find the balance of staffing levels required to meet the regulatory targets.

REFERENCES

Aksin, Z., Armony, M., and Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688.

Blackstone, E. A., Buck, A. J., and Hakim, S. (2007). The economics of emergency response. *Policy Sciences*, 40(4):pp. 313–334.

Brooks, J., Edwards, D., Sorrell, T., Srinivasan, S., and Diehl, R. (2011). Simulating calls for service for an urban police department. In *Simulation Conference (WSC), Proceedings of the 2011 Winter*, pages 1770 –1777.

Evans, D. H., Herman, R., and Weiss, G. H. (1964). The highway merging and queuing problem. *Operations Research*, 12(6):pp. 832–857.

Fry MJ, R. U. (2006). Firefighter staffing including temporary absences and wastage. *journal of operation research*, 54(2):pp. 351–365.

Marianov, V. and ReVelle, C. (1996). The queueing maximal availability location problem: A model for the siting of emergency vehicles. *European Journal of Operational Research*, 93(1):110 – 120.

Singer, M. and Donoso, P. (2008). Assessing an ambulance service with queuing theory. *Computers & operations research*, 35(8):2549–2560.

Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Science*, 51(2):pp. 221–235.