OPRE354/COMP312

Group 3

An Analysis of Calls to a Taxi Company

Adam von Kraemer, Alex Miller, Daniel Little, Priyanka Bhula

2 June 2016

Introduction

This report explores the performance of a call centre. The call centre serves a taxi company based in Wellington city. We describe the system and data. Our analysis of the data estimates the distribution of births and deaths.

We created three models to understand the performance of the system: a theoretical M/M/C model, a fitted model and an empirical model. We address the limitations of our models and make recommendations with those in mind.

The System

The call centre uses multiple servers. Calls come to the call centre through a free call line. The dispatcher services the first call on the queue, noting the request details. Local taxis are notified of the customer’s presence. The call centre’s involvement ends at this point.

The business process has been optimised using Erlang-C. They were unwilling to disclose their full process so we were given data that represented the system as having a single service stage.

Not all requests are treated equally and the call centre doesn’t exclusively handle taxi requests. All manner of calls can come through the free call line. It is possible that some requests have priority, and pre-emption might be used to service them. The full model for this system would have multiple arrival streams and multiple queues. The service units would have multiple stages of service and the different arrivals would experience different processes.

Data Collection

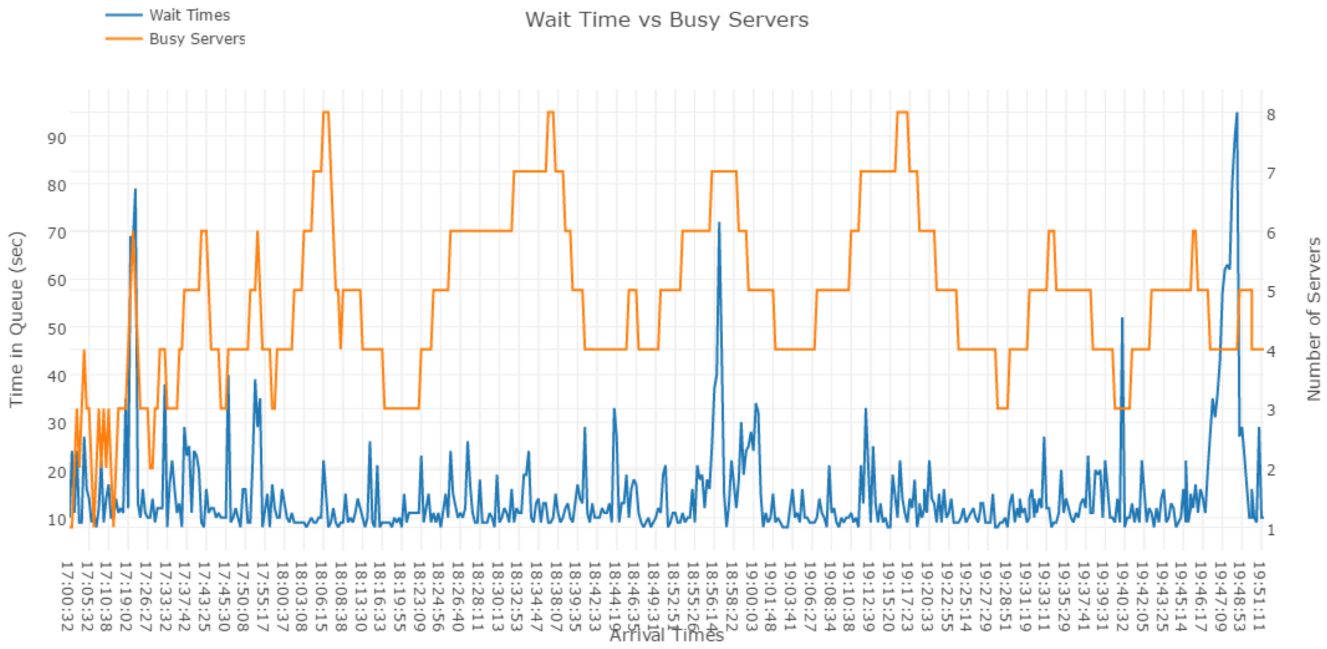
We received our data through a group contact in the organisation. This data is digitally recorded using generic call centre technology and databases.

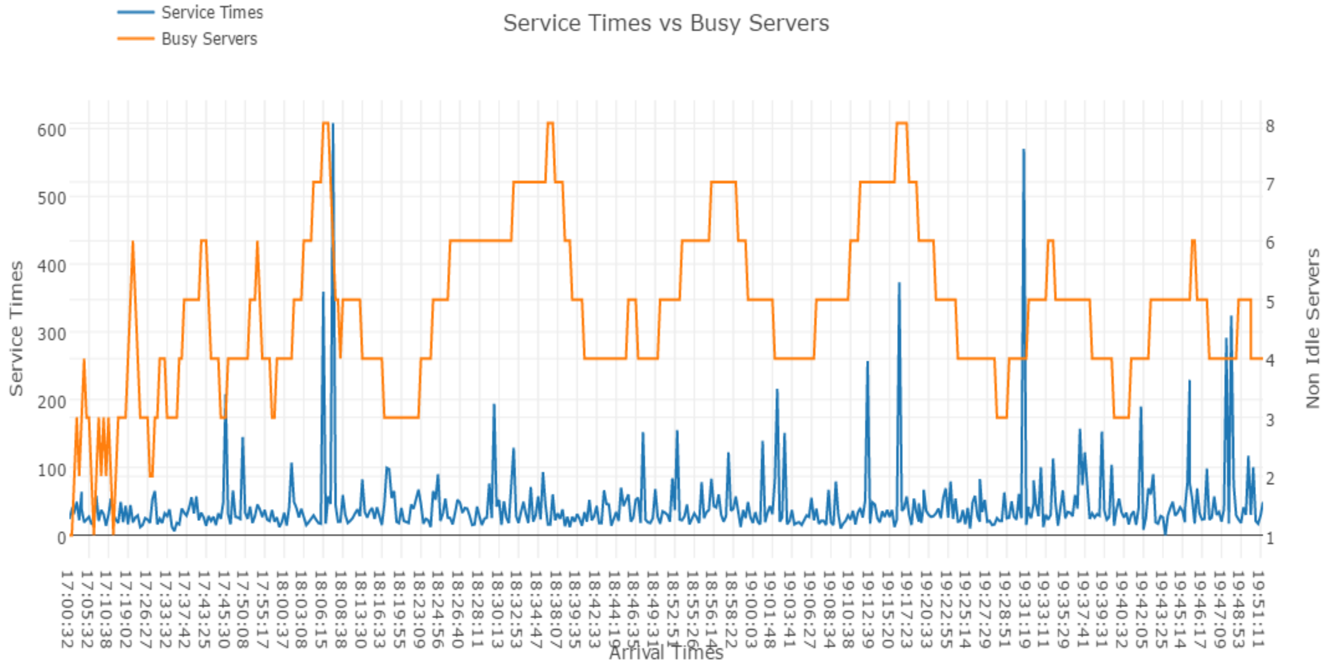
The dataset was 500 calls. Each with an arrival time (timestamp), answer time (seconds after timestamp) and completion time (seconds after timestamp).

Call Received:	*Call Answered:* *7:51:21 PM*	*Call Completed:* *7:51:50 PM*
7:51:11 PM	10 (seconds)	39(seconds)

Servers Over Time

The number of servers was not included in the data. We calculated the number of servers in use at each point in the observation period.

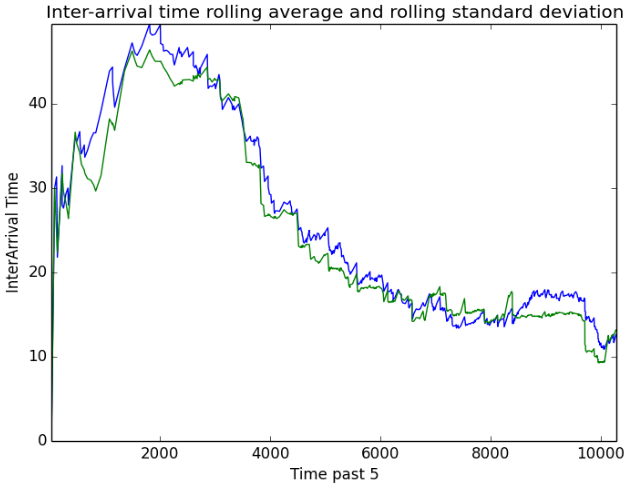




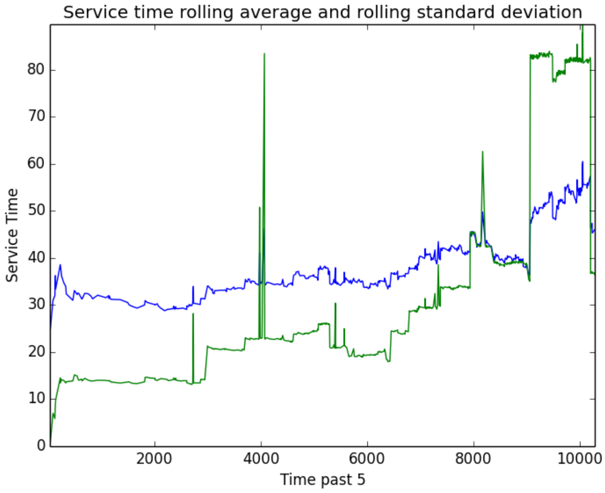
The number of servers doesn't appear to be constant during our observation period. The number appears to increase as demand peaks. Likewise, the number appears to drop during low demand.

Data

Using Python, we created two graphs to visualise the data. The first is for inter-arrival time and the second is for service time. The lines show rolling statistics, using 50 observations, with means in blue and standard deviations in green.



The arrivals graph shows an inconsistent inter-arrival rate. The inconsistency appears to be time dependent. The standard deviation of inter-arrival time mirrors the rate. This is a characteristic of the exponential distribution.



The standard deviation of the service time is wildly inconsistent. This implies outliers.

Finding the right dataset

Our tools for modeling the system aren't complex enough to deal with a time dependent arrival rate. Outliers will impede estimates of the service time distribution. Before analysing the system, we needed to find a consistent period of arrivals and remove the service time outliers.

Inter-arrival times

We regressed the inter-arrival times against the hour of day. The first hour differed significantly from the last two hours. The mean inter-arrival times for 5pm, 6pm and 7pm were 40 seconds, 19 seconds and 14 seconds respectively. We discarded calls before 6pm. This balances inconsistency and sample size.

Completion times (time in system)

An exploratory distribution based on the full dataset has a mean of 40.5 seconds and standard deviation of 53.7 seconds. Using a 99% confidence interval, outliers are defined as any completion times greater than 177 seconds (2 minutes 57 seconds).

This excludes 12 calls. Of these, 9 occur in the last half of the observation period and 7 in the last quarter.

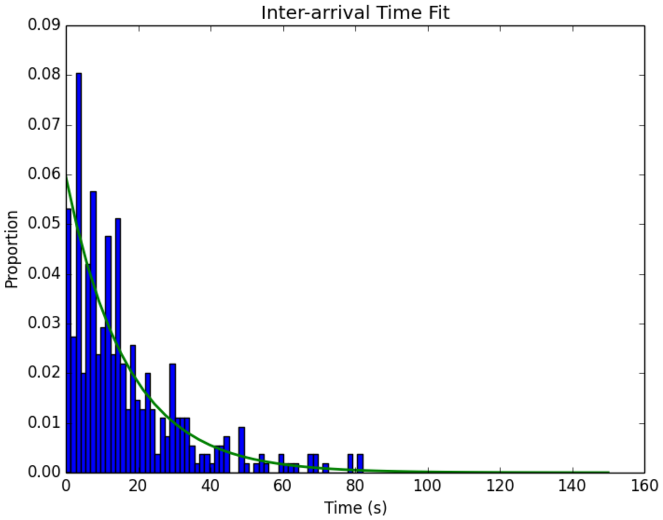
Distribution estimates

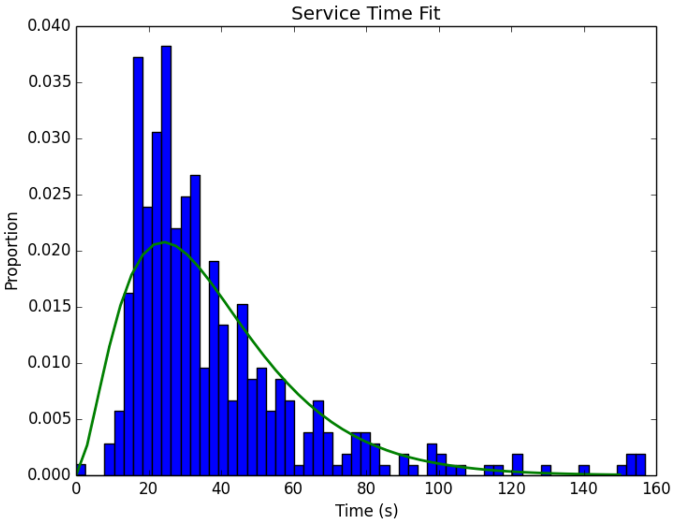
Our final dataset ignored any calls received before 6pm or with service times over 177 seconds. This subset has much lower kurtosis (the distributions propensity to produce outliers) and skewness (symmetry of the data around its mean) for both distributions.

We tested the data against the beta, Poisson/exponential and gamma distributions. We fitted an exponential distribution for inter-arrival times. We fitted a gamma distribution for service times.

Fitted Distributions

Exponential Inter-arrival		Gamma Service	
Lambda	3.574	shape	2.867
		rate	0.221
Chi squared value	49.279104	Chi squared value	73.008328
P value	0.000167	P value	0





The chi squared goodness of fit test conclusively rejects both of our fitted distributions. Nonetheless, they are the best fits. The time dependence is still an issue for inter-arrival times. The system we have analysed here is a simplification of the real system. The real system has separate arrival streams. The different arrival streams are served differently. Thus, it is no surprise that we couldn’t accurately fit distributions to the inter-arrival times and service times.

Performance Models

Theoretical Model

We estimated exponential arrival time and service time parameters

E[Inter-arrival time] = 16.41 seconds
E[Service time] = 37.98 seconds

ƛ = 3.66 / minute = 0.061 / second
µ = 1.58 / minute = 0.026 / second

Given those values, there must be at least three servers to reach steady state. Thus, our theoretical model is of an M/M/3 system.

Simulating this model produced the following results:

Parameter	Point estimate	95% confidence interval
W seconds:	68.81	(65.46 , 72.17)
Wq seconds:	30.94	(27.87 , 34.02)
Ws seconds:	37.87	(37.47 , 38.27)
L:	4.19	(3.98 , 4.40)
Lq:	1.89	(1.70 , 2.09)
Ls:	2.30	(2.27 , 2.33)

Other estimators

Estimator	Estimate
π_0	0.067
π_1	0.154
π_2	0.178
π_3	0.137
Utilisation %	0.965
P (people queueing)	0.464

The 95% confidence intervals for queueing time (Wq) and average number of customers queueing (Lq) are quite large. This variability is probably due to a small sample and the inter-arrival rate's variability within our sample.

Fitted Model

The fitted model takes parameters from the best matches of the curves fitted to the inter-arrival and service time distributions.

The results of this model agreed with the theoretical model's prediction- a system with less than 3 servers does not reach a steady state. Increasing the number of servers has no impact on service time (Ws).

Servers	3	4
L:	3.507 (3.463, 3.551)	2.490 (2.478, 2.501)
Lq:	1.243 (1.203, 1.283)	0.2264 (0.2205, 0.2323)
W seconds:	58.86 (58.172, 58.549)	41.83 (41.688, 41.972)
Wq seconds:	20.856 (20.206, 21.507)	3.803 (3.709, 3.898)
ƛ:	0.0595 (0.0593, 0.0597)	0.0595 (0.0593, 0.0597)

B%:	0.929 (0.928, 0.930)	0.904 (0.903, 0.905)

Empirical Model

The empirical model uses the observed densities of inter-arrival and service times.

Simulating the system for different numbers of servers confirmed the results of the theoretical model. Less than three servers will cause the system to fail.

Increasing the number of servers decreases the time in queue but service times are consistently around 38.6 seconds. This result is very close to the two previous models.

Servers	3	4
L:	3.503 (3.463, 3.543)	2.477 (2.464, 2.491)
Lq:	1.218 (1.185, 1.252)	0.199 (0.193, 0.205)
W seconds:	59.25 (58.667, 59.834)	41.973 (41.817, 42.128)
Wq seconds:	20.602 (20.059, 21.144)	3.372 (3.275, 3.47)
$\tilde{\lambda}$:	0.059 (0.059, 0.059)	0.059 (0.059, 0.059)
B%:	0.936 (0.935, 0.937)	0.913 (0.912, 0.914)

<u>Comparison of performance measures</u>

(3 servers)

Measure\Model	**Theoretical**	**Fitted**	**Empirical**
L:	4.19	3.507	3.503
Lq:	1.89	1.243	1.218
W seconds:	68.81	58.86	59.25
Wq seconds:	30.94	20.856	20.602
$\tilde{\lambda}$:	0.061	0.060	0.059
B%:	0.965	0.929	0.936

Our three models have very similar performance measures. The theoretical model is not as good a fit. The exponential distribution underestimates the skewness of the service time distribution, which is closer to Gamma.

We could not use Pollaczek-Khinchine analysis. It does not apply to systems with multiple servers.

Limitations

Our recommendations are limited by several factors we cannot control. The company's unwillingness to disclose the full details of the system is a severe limitation. Thus, estimating changes in the system's performance is an exercise in conjecture.

Our tools for analysing systems can't model a system with this time dependent parameters. The data required that we find a subset to model the system. Our model was simplified further, as we had to discard potentially relevant information.

The average time spent in the system is about a minute. Because this is objectively short,we don't expect much reneging. However, there is likely to be a natural loss of customers. This could be caused by customers changing their minds or by competition from alternatives.

The call centre would experience any balking as reneging. Both would appear as incomplete requests. Because customers have to enter the queue before they are aware of its state, there is no formal balking in this system.

We have no information about rejections in the system. The average queue length is close to one for the fitted and empirical models. Because of the small queue, we assume the queue effectively has infinite capacity. Therefore, rejections are unlikely.

Recommendations

If reneging is present, it still might not be noticeable. Our models indicate average time in the system will be about a minute, and that queuing time is a less significant proportion of this than service time. They also indicate that an additional server had no impact on service time, only reducing queue time. Even if reneging is noticeable, the benefits of additional servers are unlikely to outweigh the costs.

Thus, optimising the system depends on the ratio between the cost of an extra server and the cost of delay. These costs are unknown. We assume the cost of an extra server is greater than the reward for reduction in delay costs.

Since the arrival rate is time dependent, the company should ideally have flexible resources. This will keep wait times low during peaks. Outsourcing the call centre is a good way to achieve this. The taxi company can avoid the costs of planning and the costs associated with over allocation of resources and insufficient capacity.

An outsourcing company has economic incentives to achieve highly accurate traffic models and manage resources flexibly. Taxi requests are generic. It is unlikely the customer would notice the difference between in-house and outsourced services.

Conclusion

The performance of the call centre was analysed using a subset of the received data. System characteristics were estimated and performance measures in different conditions were simulated. Unfortunately, due to the large number of unknown factors involved in the operation of the system, few definite conclusions can be drawn. Though, as discussed above, it is unlikely that additional servers would make a worthwhile difference to the performance of the system. Little else of value can be determined.