

Queueing at the Tax Office

Richard TWEEDIE* and Nell HALL†
Colorado State University and NSW Department of Health

September 19, 1996

Abstract

This paper discusses a consulting project where, by focussing on the basic parameters of a probabilistic model, advice was given that could result in real improvement in the service at an Australian tax office, without raising the costs of the operation. The results are not intuitive and illustrate that non-linear behaviour in models can be hard for non-mathematicians to follow or even believe.

KEYWORDS: Waiting times, server numbers, M/M/c queues, delays, loss of customers

1 The problem

Applied probability problems are often of a scale that does not lend itself to "consulting". There are of course many outstanding examples of the use of applied probability techniques in major collaborative efforts, in for example teletraffic and networking, epidemiology, spatial pattern recognition and the like, and these often result in the type of collaboration that is commended in the advice given by the IMS New Researchers Committee [2]; but they rarely look like the sort of consulting that most clients arrive with, as noted in [3]. That is, they are usually harder or deeper, and do not have the type of constraints, benefits and rewards than one might expect if coming from a non-academic environment.

This paper describes an anomaly in this pattern: an applied probability problem that really is pure consulting, with no new methodological research, but with the rewards and problems of difficult data, client interactions, approximations to reality, and time and funding constraints, and with a happier than often

ending, since valuable advice could actually be provided and implemented within the client's budget.

The problem is simple to describe and will strike a chord with all who have been put on hold in automated telephone enquiry lines everywhere.

In the mid-1980's, both of us were working in a medium-sized private sector consultancy, SIROMATH Pty Ltd, in Australia. We were consulted by an officer of the Australian Tax Office who was concerned that the behaviours of the "dial-in" enquiry lines at different offices were inexplicably different. The tax office had recently installed management software to track characteristics of their enquiry system and had looked at several measures of effectiveness, including the following:

- (a) The average length of time \hat{w} waiting on hold before questions were addressed
- (b) The percentage of customers \hat{p}_3 who waited longer than three minutes before being answered.

In particular, the client was concerned that one office seemed to have remarkably poor behaviour compared with others. Table 1 captures most of the critical material: we have (for reasons of confidentiality) labelled three of the offices we studied as Sites A, B, C, with Site C clearly being the one in distress. One can only admire the patience of those calling to clarify their tax return status and queries: note that at Site C nearly every caller waited more than three minutes and the *average* call waited over 12 minutes!

Although there were some other, well posed, questions about the possibility of networking the system, like many consultancies this one had a regrettably implicit main question: it was of the order of "what is going on here and can we do anything about it?" We will see that there are some options, at least, for using standard queueing theory to address this satisfactorily, but that such an application requires (as do so many consultancies) particular care in acquiring appropriate data before carrying out the analysis.

*Richard Tweedie, Department of Statistics, Colorado State University, Fort Collins, CO 80523 (email: tweedie@stat.colostate.edu)

†Nell Hall, NSW Department of Health, North Sydney NSW 2060, Australia

Table 1: Effectiveness of the Calling System

Measure of Effectiveness	Site A	Site B	Site C
Average Time on Hold (secs) \hat{w}	140	200	753
Percentage \hat{p}_3 waiting > 3 mins	30%	45%	92%

2 The queueing model

In this project we had three skills to offer: the first was indeed the ability to advise on the types of models that *might* fit the situation, as discussed in [2], but in contrast to [2], the second was to help the client identify real data that might enable the model to be assessed, and the third was to carry out the analysis for him since this was rather outside his capabilities.

As in all statistics applications, the modelling should come first, or we do not know what data will be relevant. In this case we turned to the simplest queueing model: a multi-server queue with c servers (the staff in the enquiry room of the tax office), the customers arriving in a Poisson process (a relatively standard assumption, and one which implies we only needed to know the average rate λ of calls per minute) and with the time to serve customers taken as i.i.d. random variables with exponential service times of mean length $1/\mu$.

This last was, as we will describe, a rather rough assumption in this case: but with it we are able to use simple known results to assess the type of effectiveness measures being proposed. (For a good if rather old-fashioned description of how this might work, see Lee [1], who still has sound advice on how to live in a real world with the distributional assumptions involved.)

These exponential assumptions are not always appropriate. One of the truly beautiful results of queueing theory is, however, the "critical threshold" result that says that the system will, *regardless of such distributional assumptions*, have stability or instability properties according as the traffic intensity

$$\rho = \lambda/c\mu.$$

is less than 1 or otherwise. In the stable situation the queue will not get too lengthy, and in the unstable situation it will grow beyond bounds. From Table 1, Sites A and B look stable and Site C is rather like an unstable situation, and so we first sought to see what the values of ρ in our system might be. For these we only need the mean interarrival and service times λ , μ and the number of servers c . The data we were given are in Table 2: as described in the next section, the

system really is close to or above critical, especially when c is smaller than it is reputed to be, and this does help explain some of the longer waiting times observed.

We can then use the exponential distributional assumptions to enable us to consider the effectiveness parameters and predict their values, at least in the stable situation. We find, in particular, that the analytic forms are given by [1]

$$p_3 = p_0 \frac{(c\rho)^c}{c!} e^{3(\lambda - c\mu)}$$

$$w = \frac{(c\rho)^c}{c!} \frac{p_0}{c\mu(1-\rho)^2}$$

where the probability of an empty queue is

$$p_0 = \left[\sum_{r=0}^{c-1} \frac{(c\rho)^r}{r!} + \frac{(c\rho)^c}{c!} \frac{1}{(1-\rho)} \right]^{-1}$$

In this case the key question was not to predict these (or other similar quantities) with great accuracy, but rather to decide why the input parameter combinations might be leading to the particular combinations that were being observed. Note again that only λ , μ and c are relevant to these results, and so we did not need more information than this on the system.

3 Data Collection

Our initial set of data was provided from the then-new computerised telephone system. Table 2 shows that the average rate of calls at the biggest of the sites was around 4-5 per minute: this appeared relatively stable over the day, with the exception of the first hour when, not surprisingly, the rate was usually closer to the maximum of 5 per minute. The rates at the other sites seemed acceptably constant over the whole of the day.

The system also provided average call lengths. These were relatively constant across all sites. Regrettably the system did not collect the actual distribution of calls: in principle this might have been possible but the resources to reconfigure the system for better data were not available from the client. Thus we were

Table 2: Input Data

Parameter	Site A	Site B	Site C
Input Rate per Minute (λ)	4.18	2.27	1.60
Mean Call Length (secs)	180	187	200
Wrap-up Time (%)	10%	18%	40%
Mean Service Time Including Wrap-Up ($1/\mu$)	198	221	280
Minimum Number of Servers c_{\min}	10	8	6
Average Number of Servers c_a	13	12	8
Maximum Number of Servers c_{\max}	15	14	9

not able to verify if an exponential distribution was reasonable.

The most difficult information to collect was the number of servers. Internal staffing sheets showed the number of servers on an hourly basis, and these varied widely within the day. In particular the maximum number (which was possibly the number the client felt to be available) was actually 150% of the number often available. Given the role of c in ρ or in the waiting times, this is of very considerable concern.

Note that the number of servers actually assigned to each site appears in principle to be in line with the observed input rate: indeed, if anything Sites B and C seem to have pro-rata more servers than they should have in comparison to Site A, if we judge by the input rate. This helps illustrate the clients understandable concerns about the poor behaviour at Site C, since in principle the model says that this should be well under control.

Following the initial data collection, in this project we had the very real benefit of supplementing the paper data with one site visit, to Site A. There we learned rather more, as one so often does, and in particular we discovered two extra pieces of information:

- (a) On every call there was a 'wrap-up' period after the call, when the server made notes, shifted files, etc. Although these probably had at least some minimum length, we modelled them as a percentage of the service time and added them to the observed service time; and this is used in Table 2 to give the μ we used, and then in Table 3 to give the predicted values of w and p_3 ; note in particular that at Site C this wrap-up time adds much more to the actual call length than at the other sites.
- (b) There was also a period of 'idle time' for each server, some of which was time absent from the room and which might of course be reflected in the server counts, but some of which was at the

desk and in principle should be added to the service time; in Table 4 we take account of this.

These extra service times would not have been picked up without the detailed information collected on site. Some clients are insistent that statisticians visit the scene of their operations, and this can be time consuming for the statistician: others of course prefer to keep the statistician as far from the facts as possible! But whatever the attitude of the client, in order to give sensible advice within context, it is a step that one should always try to take, as this case-study illustrates.

4 Recommendations

Tables 3 and 4 show that the observed behaviour of Sites A and B is reasonably consistent with the model predictions, especially if we assume Site A is using all servers effectively, and if (in contrast) Site B is using rather close to its minimum number of servers. Site A is also very close to critical ($\rho = 1$) even when the full complement of servers is present.

If we do not incorporate the wrap-up time in Site C then in fact that site is far from critical: even with the *minimum* observed of 6 servers, they still have $\rho = 0.89$ and a predicted mean waiting time of just 180 seconds. However, the poor behaviour can be far better explained if we take into account the 40% increase to service time following the addition of the wrap-up time. Indeed, their behaviour is even more consistent with the model where we also add in some percentage of the idle time as well.

In no cases were the fits of the data perfect for the model, of course. In particular, if we assume the value 5.4 minutes for the mean service time for Site C, then we get value of around 750 seconds for the mean waiting time (consistent with reality), but we find that we only have $p_3 = 75\%$; conversely we get close to the observed value of $p_3 = 92\%$ by assuming a mean service

Table 3: Predicted behaviour excluding idle time

Parameter	Site A	Site B	Site C
Minimum Traffic Rate $\rho_{\min} = \lambda/c_{\min}\mu$	1.38	1.04	1.24
Average Traffic Rate $\rho_a = \lambda/c_a\mu$	1.06	0.69	0.93
Maximum Traffic Rate $\rho_{\max} = \lambda/c_{\max}\mu$	0.92	0.59	0.82
Assumed c for prediction	15	9	8
Predicted (observed) mean wait W	110 (140)	262 (200)	423 (753)
Predicted (observed) p_3	22 (30)	45 (45)	57 (92)

Table 4: Predicted behaviour including idle time

Parameter	Site A	Site B	Site C
Idle Time (%)	8%	8%	14%
Mean Service Time Including Wrap-Up and Idle Time $1/\mu^*$	213	238	319
Minimum Traffic Rate $\rho_{\min} = \lambda/c_{\min}\mu^*$	1.48	1.12	1.41
Average Traffic Rate $\rho_a = \lambda/c_a\mu^*$	1.14	0.75	1.06
Maximum Traffic Rate $\rho_{\max} = \lambda/c_{\max}\mu^*$	0.99	0.64	0.94
Assumed c for prediction	15	10	9
Predicted (observed) mean wait w	1259 (140)	162 (200)	537 (753)
Predicted (observed) p_3	90 (30)	31 (45)	63 (92)

time of just 5.57 minutes, but the expected waiting time is at a (noticably theoretical!) 62 hours or so. This might perhaps be explained by a distribution of service times with a "lump" of probability near the origin, corresponding perhaps to part of the wrap-up times being of fairly fixed length, but we were not in a position to look further at this. Nonetheless the general operation seemed well described by this simple model and it was possible to give some rational advice based on it.

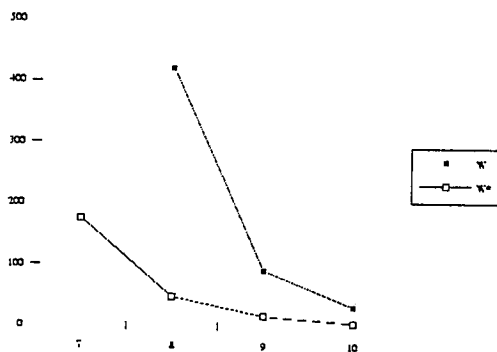


Figure 1: Waiting times for different values of c at Site C; here W is for the observed service plus wrap-up times, W^* is for the service plus short wrap-up times

In Figure 1 we illustrate the prime recommendations

we gave to the client: these were

- At Site C the average predicted waiting time (w on Figure 1) could be reduced to around 1.5 minutes (from the current 12.5 minutes) by adding just one extra server (to have an effective group of 9), and indeed w could be virtually halved merely by ensuring that all 8 current servers were constantly available;
- Even more usefully, training should be instituted at Site C to reduce the percentage of time spent on wrap-up at most to 15% of the length of the call, as was achievable at all other sites; the resulting waiting time, given as w^* on Figure 1, is less than three minutes even with only 7 servers. It is well under a minute if all 8 are working. Considerable further reductions are achieved if wrap-up is only 10%, as at Site A.

One of the effects that the client found hardest to believe was that, as just illustrated, the whole system was so sensitive to very minor improvements in the parameters when close to critical. It is not intuitive that just one extra server, or more dramatically, just saving some seconds in mean service times, could have such a powerful effect.

Various other recommendations were made, especially as the client was seriously investigating the possibility of routing calls between the sites, so the effective server pool would suddenly become around 35-50. We were able to predict that such an action

would reduce the average waiting time to well under a minute and ensure no more than 10-15% of customers would be waiting for a 3 minute period: this would give far better service than at any other single site we discussed with the client.

Did this consultancy improve the service to the taxpayers?

Sadly, I have no idea. And this is the last of the lessons in this article for the new consultant: do not always expect to make a great difference, and be grateful if you get any level of recognition. This project led to no paper (except, a decade later, this one), even though it involved much time, so there was no reward in an academic sense; it potentially helped many people at almost no cost to the client, since it clearly identified simple management changes that would give the desired result; but as so often is the case, the client did not feel the statistical consultant was relevant to implementing these, and we heard no more of it.

So why bother with such consulting? For many reasons: firstly, and not to be overlooked, we were in this instance being paid to be professionals, and, like lawyers and doctors and other professionals, we should provide our skills and not necessarily expect to be further involved, and not on center stage; secondly, statistics is designed to solve real problems, and here we did just that, and this can be its own reward; and thirdly, the project *did* achieve one of the values noted in [2], namely it was fascinating, and gave (at least to us) a real knowledge of yet another area in which statisticians can play a part that cannot be played by anyone else.

References

- [1] A.M. Lee. *Applied Queueing Theory*. MacMillan, London, 1968.
- [2] New Researchers Committee of the IMS. Meeting the needs of new statistical researchers. *Statistical Science*, 6:163-174, 1991.
- [3] R.L. Tweedie. In and out of applied probability in Australia. In J.M. Gani, editor, *The Craft of Probabilistic Modelling*, pages 291-308. Springer-Verlag, New York, 1986.