Journal of the Operational Research Society (2008) 59, 171–181 0160-5682/08 $30.00

www.palgrave-journals.com/jors

Improving police control rooms using simulation

MM Günal, S Onggo and M Pidd*

Lancaster University Management School, Lancaster, UK

Police command and control centres are the main point of contact for the public who require help. Like other areas of UK public services, police forces are set targets for their performance. Some of these targets relate to the speed at which they respond to calls for assistance from the public. In this paper we share our experience in improving the performance of command and control centres of a UK Police Force; a project which started as a classical simulation exercise and ended up with a significant reorganization in a UK Police Force.

Journal of the Operational Research Society (2008) **59,** 171–181. doi:10.1057/palgrave.jors.2602517
Published online 7 November 2007

Keywords: simulation; applications; government service; police; call centres

Introduction

The work described here is based on a consultancy project conducted by a Lancaster University team for a large UK police service serving an urban area of about 500 square miles. The force in question, though fully cooperative in the conduct and implementation of this work, prefers to remain anonymous and so is named, hereon, as The Police Force (TPF). TPF receives approximately 2 million incoming calls every year and makes about 800 000 outgoing calls. The incoming calls show significant within-day and within-week variations. TPF initiated the work because it was failing to meet targets for responses to requests for assistance from the public. People who requested help in emergencies were waiting too long for their phone calls to be answered, the comparative performance on non-emergency calls was also poor and there was concern over the time taken to get police resources to incidents when that was necessary. Not only were these aspects of performance unsatisfactory, TPF was incurring large overtime bills that were thought to be excessive. Not unreasonably, the senior officers and staff of TPF wanted to explore different options for improvement and, having heard about computer simulation, they requested help.

Before starting to explore options for improvement, TPF officers and staff sensibly wished to start by developing their understanding of the current situation through the construction and use of an 'as-is' simulation model of its operations. This then served as the springboard for modified models that were used to explore different options for change. This exploration required a mixture of simulation modelling skills, detective work and a willingness to engage directly with police operations. This was only possible due to the close cooperation between the Lancaster team and the officers and staff of TPF. As a result of the project, TPF is planning the reorganization of the way in which it handles calls from the public and how it organizes its response by committing police resources.

Real-life work of this type is always messy and any description is likely to provide a rather sanitised and, possibly heroic, view of what happened. Various writers suggest how a simulation project should be planned and managed. Examples include Robinson (2004), Banks *et al* (2001) and Law (2007). Here we use the suggestions found in Pidd (2004, Chapter 3), for the obvious reason that Pidd is one of the authors of this paper. This account suggests that simulation projects require analysts to operate in two parallel domains (see Figure 1). The first is the technical domain, in which the team must abstract and simplify the operations of interest so as to develop and use a computer simulation model. The other is the organizational domain in which the project must be managed properly so as to gain the required insights in an appropriate timescale. The serial nature of text makes it very hard to convey the parallel nature of such activity but, where appropriate, reference will be made to both legs of Figure 1. The next section begins with the problem structuring necessary to gain some understanding of TPF operations and processes, together with an appreciation of likely options for change and improvement.

Problem structuring and conceptual modelling

Problem structuring

The term 'problem structuring' seems to be used in two distinct ways: as an end itself or as a preliminary to modelling. Writers such as Rosenhead and Mingers (2001) use 'problem structuring' as an end itself and contrast this with problem solving. The latter term carries the idea that problems can be solved and there seems little doubt that this is true, at least temporarily, in many situations. However, there exist 'wicked problems' (Rittel and Webber, 1973) that probably cannot be solved but of which increased understanding may be gained

*Correspondence: M Pidd, Department of Management Science, Lancaster University Management School, Lancaster, Lancs LA1 4YX, UK.
E-mail: m.pidd@lancaster.ac.uk

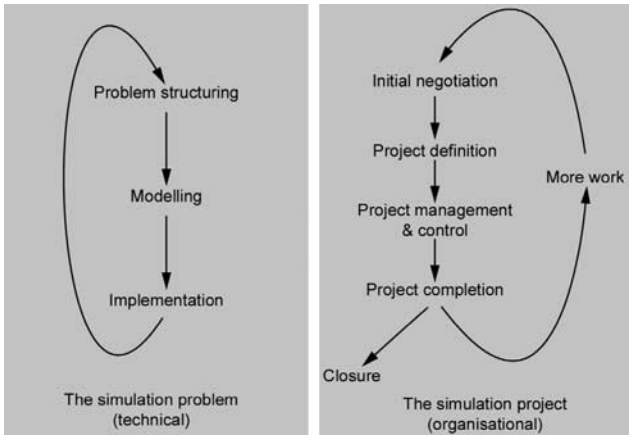

Figure 1 Simulation project domains (Reproduced from Pidd (2004) with permission).

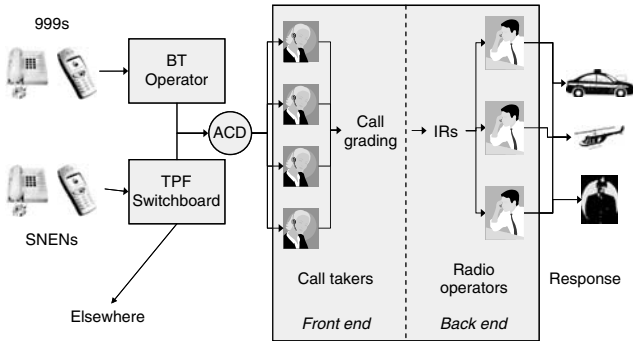

Figure 2 Current CaRC operations; call handling (front end) and radio operations (back end).

and this may provide enough clarity for appropriate action to be agreed and taken. That is, wicked problems are not solved but may be structured, so as to allow progress. As used here, however, the term 'problem structuring' refers to processes by which the analysts gain a sufficient appreciation of the presenting problem and its context so as to enable modelling and analysis to proceed. Different approaches to this form of problem structuring are discussed in Pidd and Woolley (1980).

As understood here, problem structuring is a form of sense-making in which analysts, clients and other stakeholders explore and learn about a presenting situation so as to gain joint understanding. The approaches discussed in Rosenhead and Mingers (2001) can be used; for example, Kotiadis (2006) discusses the use of soft systems methodology in developing a conceptual model in a simulation of a health care system but they are not always necessary. Sometimes, all that is needed is a willingness to get to grips with the systems of interest so as to appreciate the appropriate simplifications that will enable modelling to proceed. During this process, the problem structuring morphs into conceptual modelling, in which a software independent representation is developed. Robinson (2006) argues that 'conceptual modelling is probably the most important aspect in the process of developing and using simulation models'. That is, an unhelpful conceptualization can lead to misunderstanding and much wasted effort in a simulation project. Though a conceptual model should be software independent, in most cases the analyst knows what software will be used and will accommodate aspects of the conceptual model to suit this. In this project, TPF wished to use Micro Saint Sharp, which focuses on entities passing through a task network. Hence the conceptual model adopts a similar view of the interactions that take place in the system.

Though modern computer simulation software certainly supports rapid model development and testing, it would be a mistake to assume that this always leads to short, sharp projects in which results and recommendations appear as if by magic. Successful work depends on the close cooperation of the analyst team and the client. In this, we were very fortunate, since the member of TPF staff who provided liaison was very astute and realized that success would depend on mutual understanding. This combined with the determination of TPF to find way to improve its performance whilst containing costs, provided a very congenial atmosphere in which the main issues could be teased out.

TPF contact and response centres

At the time of the study, telephone contact between the public and TPF was handled by four Contact and Response Centres (CaRCs). Each CaRC had two main functions: call handling, which is the responsibility of the call takers and force dispatching, for which radio operators are responsible. That is, it became clear that the operation of a CaRC could be separated into two parts (see Figure 2) and this realization determined the structure of the conceptual model.

1. *The front end*: which, in the main, acted like a conventional call centre by taking calls from the public, offering advice and noting details of the call, including its urgency.
2. *The back end*: in which radio operators took graded call information prepared by the call-takers (see below) and committed appropriate police resources. In the case of chases and continuing incidents, radio operators might be involved with a particular incident over several hours.

In the front end operation, calls from land-lines were routed on a geographic basis to the nearest CaRC, which would also dispatch police resources if these were needed. However, if the call-takers were busy, calls would be re-routed to an alternative CaRC via an Automatic Call Distributor (ACD). The call-takers dealt with two broad types of call: emergencies (known as 999s in the UK) and nonemergencies. If a member of the public dialled 999, their call was immediately answered by a British Telecom (BT) operator, whose main job was to determine which emergency service is required. If the caller needed police assistance, the call was forwarded to

Table 1 Call grading at TPF

Call grade	*Dispatch target*
Immediate response	Attendance within 10 min of call receipt
Priority response	Attendance within 1 h of call receipt
Prompt response	Attendance within 4 h of call receipt
Referred response	Referred to division for appropriate resolution
Telephone resolution of the call	Achieved by successful and appropriate first time telephone resolution

the appropriate CaRC. To discourage the use of 999 for non-emergency calls, TPF widely advertised an alternative Single Non-Emergency Number (SNEN). SNEN calls came direct to the TPF switchboard which, if appropriate, forwarded the call to an appropriate CaRC.

Since TPF wished to respond fastest to the most urgent calls for help, it employed a call-grading system that was applied to all calls, whether 999s or SNENs. The call-taker receiving the call would evaluate the situation, using a standard protocol, and assign a grade (see Table 1) to the incident reported by the caller. The first three grades generally required police resources to be deployed that day, whereas lower two grades of call were transferred to other sub-divisions in TPF for forms of resolution not requiring deployment of physical police resources. Table 1 also shows TPF's dispatch targets, and this reveals that, for instance, grade 1 calls required them to attend the incident with 10 min of the radio operator dispatching resources. In a crowded urban area, this can be a challenging target to meet. Given the need for such rapid responses, resources were distributed geographically around the TPF area.

In addition to receiving and responding to incoming calls from the public, the CaRCs also involve the call handlers and other staff making outgoing calls, which may be a suitable response to some incoming calls. Records showed that outgoing calls represented about 40% of all calls passing through the CaRCs. The effect of the outgoing calls is to reduce the time available for handling incoming calls. Figures 2 and 3 are simplified and do not show the outgoing calls. At the end of each simulated incoming call, a proportion of calls generate an outgoing call on which a call-taker is then engaged for a further period.

In the front-end operations of a CaRC, call-takers graded each call to produce an Incident Report (IR) by typing appropriate information into a pro-forma while speaking to the caller and would complete the IR immediately after the end of the call. The support software entered the IRs into a database and then ensured that a summary of the IR appeared on the screen of a radio operator working in the back-end of the CaRC. The radio operators viewed the IRs on their screens and prioritized them by their call grade, dispatching appropriate resources. On some occasions, the radio operators, assisted by their supervisors, had to further prioritize calls for two reasons. First, there might be several IRs with the same call grade awaiting resolution and, secondly, some calls might need specific resources (eg a helicopter). The increasing use of mobile phones led to many incidents being reported through multiple calls from several members of the public at or near the site of incident. Hence, the radio operators strove to avoid duplication and a supervisor might need to determine that several calls refer to the same incident.

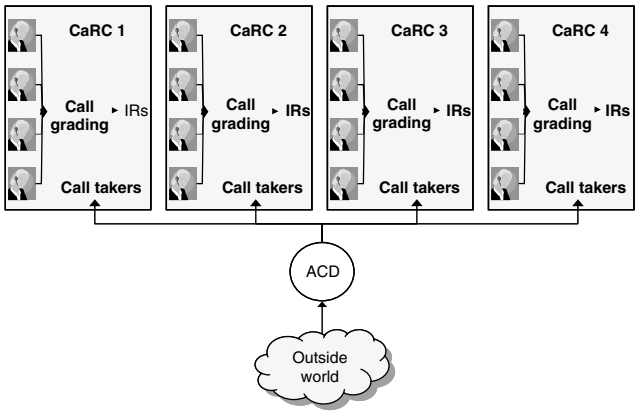

Figure 3 Proposed CaRC front end operations (showing 4 CaRCs).

Table 2 Emergency and non-emergency calls have different performance targets

Call type	*Performance indicator*	*Target (%)*
Emergency	Percentages of calls answered in 15 s	90
Non-emergency	Percentages of calls answered in 30 s	90

Table 2 shows the performance targets for the CaRCs at the time of the study. *The targets* were local, that is, they were set by TPF itself, though they were based on suggestions from external bodies. When the study began, not only was TPF not meeting its own targets, but it was likely to perform even worse against more stringent national targets that were thought to be in the offing.

Modelling of CaRC operations

The front end operations of a CaRC are, in effect, a specialized call centre and call centres have been widely studied.

Gans *et al* (2003), for example, provides a very thorough review of the operation of many types of call centre and discusses different approaches to their improvement. Since call patterns often follow non-stationary distributions and there is a need to model complex shift patterns, discrete-event simulation (DES) is a commonly used approach in planning and improving their operations (Avramidis and L'Ecuyer, 2005; Mehrotra and Fama, 2003). Within the police domain, some of the earliest such work is reported by Kolesar *et al* (1976), who worked with the 911 emergency phone system of New York City. This describes the combined use of a model based on queuing theory and a simulation, in which the latter was used to validate the queuing model. In a similar vein, Kuhn and Hoey (1987) report on work commissioned to improve the performance of Washington D.C. police call handling. Their approach included a simulation of call handling operations.

As is clear from Kolesar *et al* (1976) and from Gans *et al* (2003), analytical models of aspects of call centre operations are often used for planning shift patterns. Many of these models are based on M/M/N (Erlang C) queuing assumptions and divide the non-stationary patterns into relatively stationary short periods (eg of 30 min) that might correspond to intervals on a shift pattern. However, such models do not allow for callers who renege or balk after calling and are known not to cope well with systems that are occasionally highly congested. Taking a different tack, Chassioti and Worthington (2004) proposes the use of discrete time approximations that cope with non-stationary demand and also allow the use of realistic service time distributions.

Unlike the front-end operations, there have been very few OR&MS studies of response and resource despatch (back-end operation) and even fewer accounts of DES applications. Green and Kolesar (2004) addresses most of the issues involving police response and resource despatch and gives a short history of OR/MS work on emergency responsiveness. Larson (1973) is an early report of the use of a simulation model for urban police patrol and dispatching that was implemented in a number of large cities in the USA. Written around the same time, Ignall *et al* (1974) discusses both a simple M/M/c queuing model and a simulation approach for police patrol deployment. Published a few years later, Bohigian (1977) summarizes some of the simulation work in criminal justices, including police patrol and communication. In another early paper, Colton (1979) sets modelling work in a broader context and reports a survey on the usage of computer technology by the US police departments. Colton reports that 18% of the police departments surveyed indicated that they used analytical models to help improve their operations.

It is clear, then, that both front and back end operations of police response have been studied, though there are more reports on work attempting the improvement of front end operations. It is also clear that the front end operations can be regarded as a specialized call centre.

Our conceptual model

After discussion within our police colleagues we agreed that the objective of this study was to investigate ways to improve TPF's performance in handling emergency and non-emergency calls from the public. We also agreed to treat the front and back ends of CaRCs separately. Though each CaRC housed both front and back end operations, call handlers did not communicate directly with back end radio operators. Instead, the radio operators worked through the prioritized list of IRs that appeared on their screens. The two parts were linked by the database system into which IRs were entered and from which they were extracted. The need for detailed modelling of call handling, including non-stationary demand distributions, complex shift patterns, call balking and outgoing calls, led us to develop a simulation model of the front end operations. We had originally intended to simulate the back end operations as well; however, it soon became clear that there was little or no data available on back end operations. Hence, we focused our efforts on simulating the front end.

Our conceptual model of the front end is shown in Figure 3, in which calls arrive from the outside world. Though 999s were routed via BT operators and SNENs via the switchboard, performance was measured only at the point when the call is passed through to the ACD. Each call would be destined for a particular CaRC, but would be re-routed by the ACD if no call-taker were available in that CaRC. If no CaRC was able to accept the call, it would wait in a FIFO queue and the ACD would poll the CaRCs at frequent intervals until one was able to accept it. Each CaRC had its own manning levels and so the number of call-takers would vary across the CaRCs and through time. There was no evidence of a line shortage, so the conceptual model assumed that enough line capacity for incoming and outgoing calls would always be available.

Building the simulation model

The simulation model was implemented in Micro Saint Sharp (2005) because TPF already had a licence for its use. Though Micro Saint Sharp allows animation (eg, to show calls arriving, being answered and leaving), the nature of the conceptual model meant that such animation added very little to the simulation and hence the models were developed and run in network diagram mode. The model network is, as would be expected from the conceptual model, rather simple, and a partial network in shown in Figure 4. An oval represents a task in which resources are needed to change the state of the call. The small diamonds represents points where there is more than one route after a task is completed. The small rectangle attached to a task shows that calls may be queued before entering the task. A rectangle represents a subsystem which is drawn as another task network (the task network inside the rectangle is not shown in Figure 4). Since the front end operations in each CaRC were conceptually the same, though the number of calls and call-takers varied,

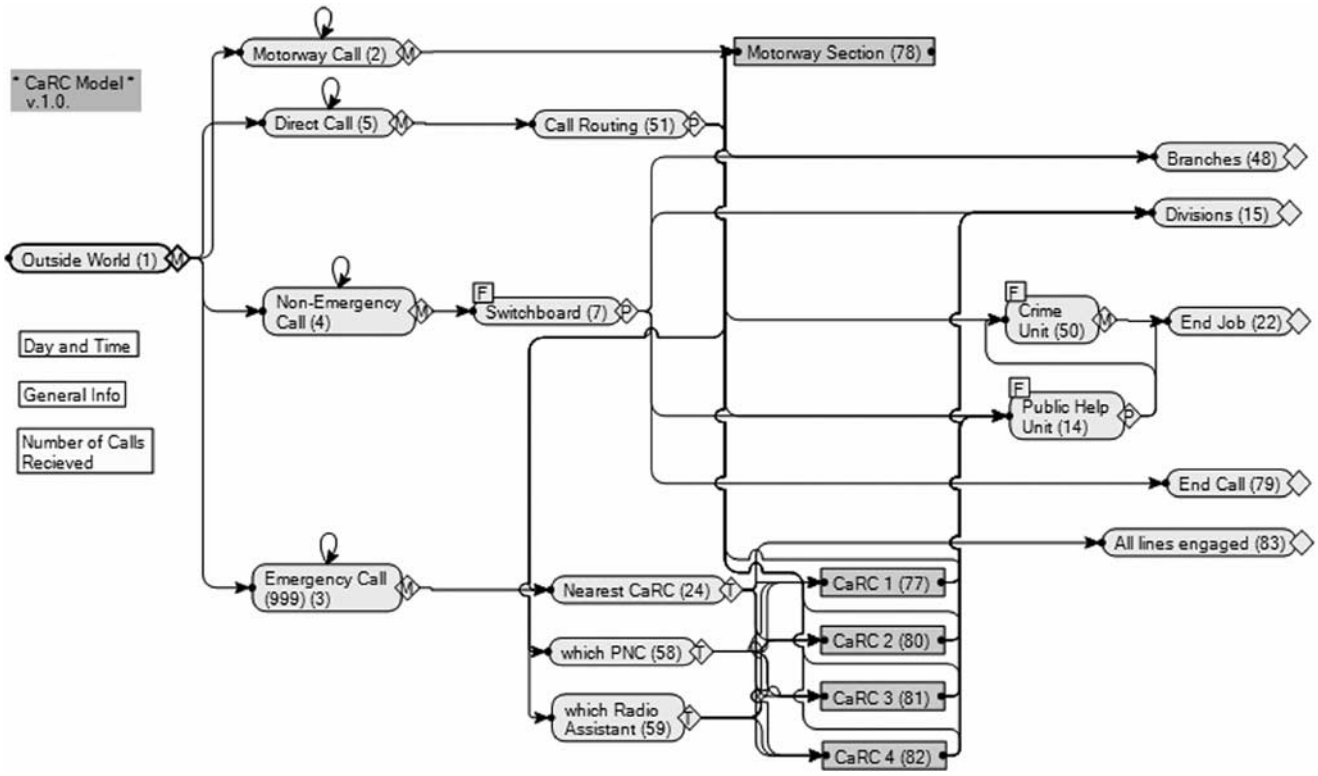

Figure 4 Simulation task network.

the actual model includes four replicates of the same CaRC logic. The call arrival distributions, call duration distributions and number of call-takers available are all provided as data via Excel files when the model runs.

The simulation model requires two types of data: the number of call-takers available at any time and details of the calls. The TPF phone information system automatically logged the start and end times of each call and so it was straightforward to obtain a sample of calls so as to analyse their statistical properties. In particular, this sample allowed us to examine the arrival patterns of the calls and their durations.

Modelling calls

Analysis of the arrival patterns quickly showed that the arrival rate of calls and, therefore, the intervals between them, varied during the day and during the week. As discussed earlier, such non-stationary behaviour is common in call centres and in police contact centres. In the case of TPF, Figure 5a shows the typical in-week variation in the number of calls received at the time of our study, and the typical within-day variation is shown in Figure 5b. Figure 5a shows the average call volumes from the data provided from the TPS systems. The highest call volumes are on Saturdays and Sundays and Figure 5b clearly shows the evening peaks. The same patterns may not occur in other police forces since TPF's area includes several large urban centres and a police force with different responsibilities is likely to have a rather different call pattern. Analysis of the data showed that call arrivals could be modelled as a non-stationary Poisson process. In our simulation model, each call is separately generated and progressed through the model. Call generation is implemented by computing Poisson rates at hourly intervals across a week and modifying these using a thinning process (Lewis and Shedler, 1979) to reflect the non-stationarity. Since Micro Saint Sharp uses C# as its simulation language, it was relatively simple to code this thinning algorithm and the call generation in a re-usable function.

The statistical patterns in the observed call data were used to build the as-is model and also served as the basis for the later investigation of questions such as: how many staff will be needed if call volumes increase each year? We could not be sure whether the patterns would remain the same and just scale up if demand increased. However, the 'what if' simulations relied on expert opinion within TPF that the patterns would remain constant, and would scale up as call volumes increased.

Since each simulated call is separately generated in the simulation, it is also necessary to simulate the length of time taken to handle that call. As with call generation, the TPF information system made this analysis straightforward, since the duration of each call can be extracted from the system. In analysing this data it was important to check for the effect of the time of day and the call grade. It might be the case

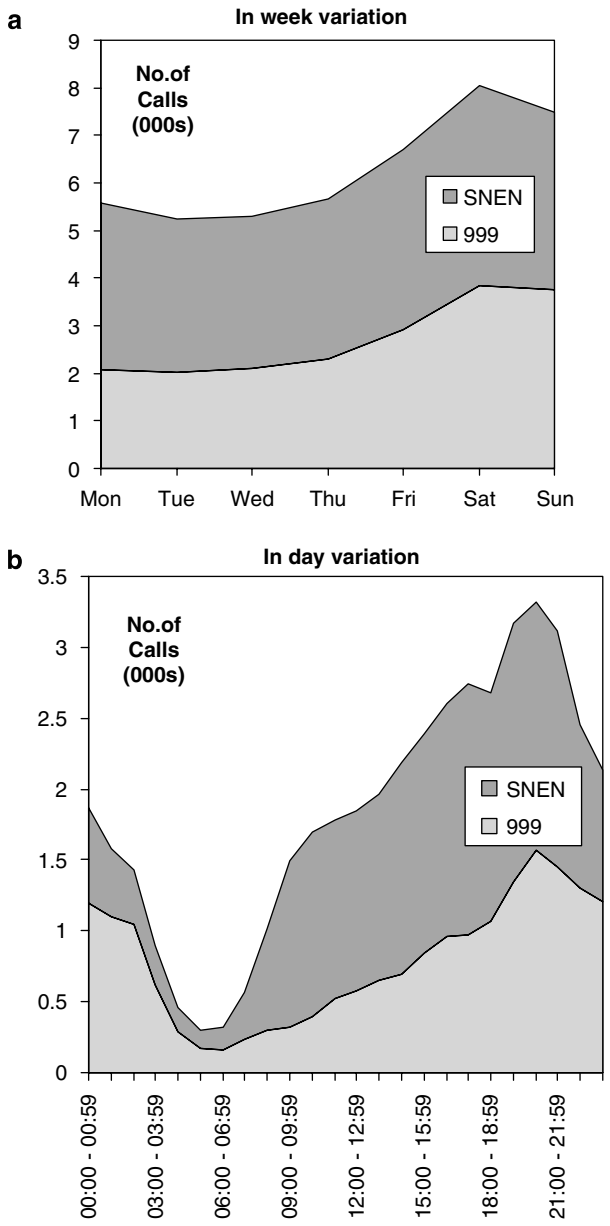

Figure 5 Variation in number of calls received. (a) In week variation and (b) In day variation.

that calls at one time of day take longer than others and it might be that the call duration varies according to the call grade. Statistical analysis showed that neither of these effects was evident and that call durations could be modelled using lognormal distributions. Analysis of the IRs enabled us to develop a histogram of different call grades, allowing each call to be probabilistically given a call grade during a simulation by taking a sample from that histogram.

Two other aspects of call handling also needed to be included in the model. The call logs stored in the information system showed that some callers reneged—that is, they ended the call before a call-taker was able to answer. There are many possible reasons for this, including congestion in the system. Based on the log data, we modelled the reneging using a staircase function, that is, the longer a person waits, the more likely it is that the call will be abandoned. Since there was no evidence of a line shortage, we ignored the issue of balking. Finally, the log data showed that about 40% of total calls were outgoing from the CaRCs and this proportion was applied in the model.

Representing staffing levels

Since one aim of the project was to understand why performance against targets was so poor, it was obviously important to represent the number of call-takers available at any time during a simulation. For the as-is model it was important that the staffing levels were an accurate representation of staff availability during the period that the as-is model was to represent. Were this not so, it would be impossible to have any confidence in the model, nor could it be used to understand why the actual performance was poor. Unlike call data, it seemed, initially at least, that staffing information could not be simply taken from a database and analysed, instead we had to rely on other records kept by TPF. The shift patterns in use were based on hourly changes, which allowed this same interval to be used for the thinning process employed for call generation in the model.

However, incorporating this hourly staffing data into the simulation model led to some very puzzling behaviour. When the model's performance was compared with the actual performance for a defined period, there were very large discrepancies between the two. In particular, the model showed the average performance to be much better than it actually had been. Digging deeper into the data, significant anomalies became apparent. For example, examination of the actual performance in the period showed that this was best when recorded staffing levels, as incorporated into the model, were at their lowest. For example, the recorded staffing levels during the last three days of the week were low, yet the percentage of calls answered within the target time was high—though there was no evidence of any significant change in call handling times. This was even more puzzling when placed against the excessive overtime that was one of the drivers for the project.

After further investigation it became apparent that there were two reasons for this. The first was that we had been given planned staffing numbers and not the actual values. Absenteeism is a problem in all call centres and answering emergency calls in a CaRC can be a very stressful job, further worsening the problem. In addition, further investigation revealed that the front and back end operations were not quite as independent as we had thought. At many times of day there are empty seats in the call-takers' area, since demand is expected to be below the maximum. However, the radio operators who act as dispatchers must always have a full complement since each is assigned to one or more physical

geographical areas. If a radio operator is absent from his/her station, for short periods or an entire shift, an appropriately trained calltaker will take over the station—depleting the actual numbers of available call-takers.

Digging deeper, we found that other parts of the information system did indeed store information about staff availability, though not in a very convenient form. When a call taker starts a shift, (s)he must press a button to notify the call distribution system of their availability. If a call taker needs to take a break or must take over a radio operator role, (s)he again presses the availability button to notify the system that their call station is unmanned. Hence, we dug into the raw information system data and were able to produce staffing figures that more or less reflected actual availability—though it could still be subject to error if a call taker took a break without logging off from the system. Using this data allowed us to demonstrate that, for the weeks in question, the model provided a good indication of the call handling performance of the CaRCs. That the CaRC supervisors were unable to provide accurate staffing data is worrying, and suggests that the CaRCs were not very well managed.

ACD logic

In the as-is model, the ACD plays a very important role in balancing the workload across the four CaRCs. It checks the availability of call takers in each CaRC and distributes the calls accordingly. The ACD would be unnecessary if there were one single CaRC. As about 15% of calls are re-routed to a non-geographic CaRC, it was important to capture ACD logic as accurately as possible. As with the staffing data, this seemed simple at first, but soon became less clear.

We were first told that the ACD operated with geographic logic. Thus, if a call arrived at a busy CaRC, it would be re-routed to the physically closest alternative CaRC. If this second CaRC was also busy, the call would be further re-routed to a third CaRC that was the next closest, and so on. If all CaRCs were busy, then the call would return to the first one where it would remain in the queue until answered. We built this logic into the simulation model and discovered that it led to unexpected results. It produced a situation in which far too many calls were passed to the first CaRC in the loop; far too many, that is, compared to the actual data, which showed a good balance of re-routed calls across all CaRCs. Something was wrong, though the CaRC supervisors were convinced that this simple ACD logic was correct.

Modelling the ACD correctly is important, since we wished to model the different performance of each of the CaRCs so that, later, we could examine different CaRC configurations and staffing levels. The ACD swings into action when the system is congested, which usually occurs at times of high call volumes, such as night-time at weekends and these are times when performance is most likely to drop. As with the staffing data, the CaRC supervisors were curiously ill-informed about the workings of the ACD.

Further investigation showed that the ACD did not operate on a geographic basis, but on a 'free the longest' basis. That is, if a call arrived at a congested CaRC, it would be re-routed to the CaRC for which the interval between the current clock time and the arrival of the previous call was the longest. This then left the question of how the ACD knew which CaRC was busy. Further investigation showed that a CaRC was regarded by the ACD as not busy if two or more call handlers were available but not engaged in current calls. On reflection, this is not sensible, since some CaRC had only two call handlers on duty during slack periods. A one-free rule would be much more sensible.

Following our own detective work on the ACD logic, we were able to correctly model call re-routing and to produce figures that were close to those observed in real life and to complete our black-box validation. It was, though, strange that the CaRC supervisors had no real grasp of the operation of the ACD. Once the model incorporated the correct staffing levels and ACD operation, its use showed that, properly managed, the call taking operation should be able to meet the target answering times shown in Table 2 without adding extra staff. The understandable priority given to radio operation was a concern, though, since it would always make it difficult to meet the answering targets whenever a radio operator was absent.

Model validation

As is often the case, model validation proceeded alongside model development. It is widely recognized that most simulation models cannot be fully validated, in the sense of knowing that they are a completely adequate representation of some system. One obvious reason for this is that such models are often used to investigate possible futures, which do not yet exist and against which model performance cannot therefore be compared. Hence, both black-box and white-box validation (Pidd, 2004) are better regarded as a process of building confidence that a simulation is appropriate for its intended use, an issue discussed in Kleindorfer *et al* (1998). Therefore, in this project, validation proceeded hand in hand with model development as call arrivals, call handling, staffing rules and call hand-off were included in the model.

Thus, the models were gradually refined by adding detail until they were considered adequate for the project aims—to understand why performance was poor and to investigate possible improvements. Figure 6 shows the model output against actual performance for a two 2-week period, based on the percentage of calls answered within the specified period.

Initial use of the front end simulation model

With an adequate model of ACD logic in place, the simulation model could mimic the actual operations of the CaRC front ends. That is, it was regarded as a valid representation of current operations. This allowed the team to investigate likely performance under different staffing assumptions.

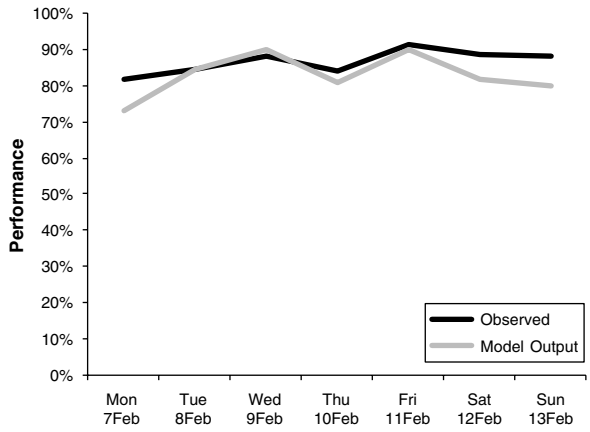

Figure 6 Model output versus actual performance over a two-week period.

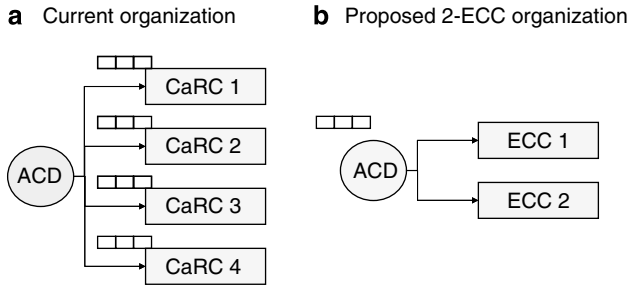

Figure 7 Possible new organizations. (a) Current organization and (b) Proposed 2-ECC organization.

These showed that the CaRCs should be able to meet performance targets without the excessive staffing levels which, in prospect, had simulated the simulation project.

Resource deployment—the back end

As discussed earlier, the simulations concentrated on the front end, in which calls are received, handled by operators, graded and passed to the back end for attention, which might include resource deployment. The CaRCs had existed for about 10 years and had replaced a set of control rooms owned by each division of TPF. Since resources had, 10 years earlier, been owned by each division, this meant that the control rooms operated as command and control centres. However, the areas covered by each division were rather small and this led to constant problems, such as those caused when police officers were absent. Hence the CaRC system had four large centres and aimed to integrate call handling and resource dispatch. Since we had a good understanding of the call handling operations and had created a front end simulation model, we decided to investigate resource deployment and so had two OR postgraduate students spend time in the divisions and on patrol with police officers.

Radio operators

It soon became clear that the CaRCs had very little real control over resource deployment. In effect, the CaRC radio operators requested resources and the subdivisions decided what to do. Within this, the police officers on the beat had considerable discretion as to which calls to respond to in which sequence. That is, much of their work was fairly loosely, unsupervised. Thankfully, 'Immediate Response' calls were always given priority, but lower grade calls might be processed in any convenient sequence—this might actually be sensible, since the officers may have local knowledge. Owing to this locally exercised discretion, it was almost impossible for the radio operators to meet the response targets shown in Table 1 since they could not control resource deployment. This, of course, raised the question of whether the attempt to integrate call handling and resource dispatch within the CaRC system made sense. If resource control was exercised by the subdivisions and divisions, perhaps the radio operators should be co-located with these resources. Since all contact between call handlers and radio operators in the CaRCs was via database records, perhaps the CaRCs should be replaced by call centres, linked electronically to local Control Centres in sub-divisions?

That is, a system very different from the current CaRCs might be preferable. This could consist of one (or two, for resilience) call centres—if more than one, any ACD should not operate geographically, but should merely aim to see calls answered as quickly as possible. If radio operators were no longer collocated, there was less opportunity for borrowing call takers as stand in operators. Admittedly, this would require better control of radio operator staffing, but this was likely to be a good thing. Hence, we proposed an alternative organization of front and back end operations to TPF and this is discussed below.

What-if modelling

Conceptual models: alternative organization

Thinking at a conceptual level, it seemed clear to us from queuing theory and intuition that operating with a single queue and multiple servers would lead to better performance than having a queue for each server. In terms of the Emergency Call Centres (ECCs: CaRCs with radio operations removed) this suggested, as mentioned above, an end to geographic call distribution. TPF agreed with this view but needed evidence from which to mount a case for the large capital expenditure that would be needed. Hence, models were needed that represented different ways of organizing the ECC operations to allow performance comparison. Broadly speaking, this required the development of two variations on the simulation.

1. *Current*: The as-is model of a system with geographic call distribution, four integrated CaRCs and call queues before each CaRC.

Table 3 2-ECC option outperforms 4 CaRCs option

	999*s*		*SNENs*	
	4 *CaRCs (%)*	2-*ECC (%)*	4 *CaRCs (%)*	2-*ECC*
Average (%)	94	99	86	99.8
Best weekly (%)	98	100	93	100
Worst weekly (%)	86	97	72	99
Standard deviation (%)	3	1	5	0.3

Table 4 2-ECC option under increasing call duration

	999*s*		*SNENs*	
Call duration (%)	+10%	+20%	+10%	+20%
Average (%)	97	95	99	98
Best weekly (%)	99	99	99.8	99.7
Worst weekly (%)	91	87	97	95
Standard deviation (%)	2	3	0.6	1

2. *Proposed*: An alternative system with 2 or 4 ECCs and a single queue, with dispatch and response devolved to subdivisions. This would allow local control of resources.

Figure 7 shows the essential difference between a 4 CARC and a 2-ECC organization: clearly an organization based on 4 ECCs would show 4 ECCs rather than 2. In addition, some callers may renege if the system is too busy and waiting time becomes excessive. In the as-is model (4 CaRCs), the ACD is geographic. In the alternative 2-ECC and 4-ECC models, the ACD logic is simple. Calls enter a single queue and the ACD routes a call to the ECC with the first available call taker, which leads to a balanced workload. Under such a system, TPF would operate what is effectively a single 'virtual' call centre.

Experiments and output analysis

The main objective of the experimentation was to evaluate whether the proposed call management system (a 'virtual call centre') would perform better than the current system of four geographic CaRCs. As expected, the experiments showed that the proposed system should outperform the current system and that there was very little performance difference between 2 or 4 ECCs. Since simulation of the 2 and 4-ECC options showed very little difference in their performance, the following discussion concentrates solely on the 2-ECC option.

Following this result, we conducted sensitivity analyses to evaluate the effect of different call durations and call volumes on performance. Table 3 shows the percentage of calls handled within the target times based on current call volumes over a 12-month (52 weeks) period. This indicates the performance of the current system of 4 CaRCs against the 2-ECC option and it is clear that the proposed 2-ECC model consistently outperforms the current 4 CaRC model. Our analysis shows that the weekly performance could be increased by up to 11 and 27%, and annual performance to 5 and 14%, for emergency and non-emergency calls, respectively.

Having established that the performance of 2- and 4-ECC virtual call centres were similar and would outperform the current 4 CaRC system at current call volumes, we next investigated the performance of a 2-ECC system against tighter targets, longer call durations and increased call volumes.

TPF were keen in see how the new system would perform against new, tighter targets that they thought might be suggested by the Home Office, which includes police forces in its responsibilities. In the UK, each government department negotiates Public Service Agreements (PSAs) with the Treasury. These specify, among other things, performance targets for the department in question, in return for the budget that it is granted. Thus, the PSA for the Home Office included targets for reduced crime and increased public confidence. These national targets cascaded down to individual police forces and it seemed likely that targets for call handling would be introduce in the future. Hence, TPF wished to know what staffing levels would be needed in a 2- or 4-ECC system to meet these targets for current call volumes. Hence, we investigated the model performance with tighter targets of answering 90% of 999s within 10 s (reduced from 15 s) and SNENs within 20 s (reduced from 30 s). The experiments showed that the proposed ECC systems should be able to meet the more stringent targets with no increase in staffing.

The next experiment aimed to anticipate the effects of the National Incident Recording Standards and National Crime Recording Standards. These were intended to improve quality of information received and recorded by police forces and

were likely to increase the average call duration. As expected, the increase in call duration reduces the call handling performance (see Table 4). However, the experiments showed that the ECC-based systems should be able to meet current performance targets with increased call durations for both 999s and SNENs. There experiments show that there are likely to be occasional performance breaches at times of peak demand, but overall performance is still well above the annual targets. The same results are also observed for the non-emergency calls. Finally, any reorganization of the 4 CaRCs into 2 or 4 ECCs with dispatch devolved to sub-divisions will be expensive and will need to provide robust performance even if call volumes increase. Hence, in the final experiments, we increased call volumes by 2% pa (as specified by the Home Office) for the next 15 years. In this scenario, current call durations and current performance targets were used. As would be expected, performance declines year by year as the call volumes increase at an annual compound rate of 2%. Similar results are observed for non-emergency calls.

For the first 10 years, it seems that an ECC-based system would meet current average performance targets on an annual basis. Thereafter, staffing levels would need to be increased. However, there are some concerns behind this seemingly rosy façade. First, if performance is measured on a weekly or monthly basis, there will be increasingly frequent breaches of the performance targets—over the year, the slack periods compensate for busy periods. Secondly, it is important to consider the working conditions of the call takers. As time proceeds, the call-taker utilizations increase and, by year 10, they would be continuously busy for several hours. It seems likely that such busyness would affect performance and could lead to problems with increased absenteeism. Hence, we might reasonably expect worse performance than suggested by the simulation analysis.

We have not examined whether staffing would be adequate if caller demand and call duration both increased at the same time. It seems likely that staffing levels would be under pressure rather sooner than 10 years under such a combined scenario.

Summary and future work

Our work with TPF demonstrates the value of close cooperation between the client (TPF) and the analyst group (Lancaster). As mentioned earlier, we were very fortunate to have a member of TPF staff acting as liaison in an extremely effective way. He was trusted by the police officers and also by the Lancaster team. This enabled the work to proceed as a fascinating combination of analysis and detective work. Without the careful construction and use of simulation models and their use for analysis, hunches about current performance and alternative configurations would not have been tested. Thus, the careful analysis added much value to TPF's investigation of its CaRCs. In addition, though, some fascinating detective work was needed to establish the limited control actually exercised by radio operators over police resources, to get hold of actual staffing level for model validation, to establish the actual ACD logic and to understand the ways in which radio operations interfered with call handling.

The simulations show that, properly managed, even the current CaRCs should be able to meet current performance call handling targets with current staffing, but this will only be true if staff are properly managed. Scrapping geographic call distribution should lead to performance improvements that should be robust against increased call volumes or increased call durations or tighter targets. TPF were not interested in experiments that combined these three effects, though it would probably have been wise to conduct them. Our analysis shows that the proposed reorganization structure, having a 2-ECC system, should perform better than the current system. The sensitivity analysis also shows that it can meet tighter targets to anticipate the national targets which will be introduced in the near future.

We did not attempt to simulate the back-end operations, though it seems that a significant benefit might accrue from doing so. Members of the public expect more than a good call handling operation from TPF but also a good response and despatch operation. Once this responsibility is devolved to sub-divisions, then it will be important to establish staffing levels that will support a good service to the public. As well as modelling individual sub-divisions, it may be important to consider their interaction and cooperation since overall staffing levels may be higher if this is not done. All experienced operational researchers know that focusing only on one side of a system may switch the pressure elsewhere and it is important to ensure that meeting call handling targets does not lead to a failure to meet response targets.

Acknowledgements—Some of this work is described in an earlier paper: Gunal MM and Pidd M (2006). Detective work in a police force: meeting standards for call handling. *Proceedings of the 3rd OR Society Simulation Workshop*, 28–29 March 2006, Ashorne Hill, UK.

References

Avramidis AN and L'Ecuyer P (2005). Modeling and simulation of call centers. In: Kuhl ME, Steiger NM, Armstrong FB and Joines JA (eds). *Proceedings of the 2005 Winter Simulation Conference*, Orlando, FL, USA. IEEE: Piscataway, NJ, pp 144–152.

Banks J, Carson JS, Nelson BL and Nicol DM (2001). *Discrete-Event System Simulation*, 3rd edn. Prentice-Hall: Upper Saddle River, NJ.

Bohigian HE (1977). Simulation modeling of the criminal justice system and process. In: Sargent RG, Schmidt JW and Highland HJ (eds). *Proceedings of the 9th Conference on Winter Simulation* — Volume 1, Gaithersburg, MD, USA, 5–7 December 1977, *Winter Simulation Conference*. Winter Simulation Conference Pubs., pp 246–256.

Chassioti E and Worthington DJ (2004). A new model for call centre queue management. *J Opl Res Soc* **55**(12): 1352–1357.

Colton KW (1979). The impact and use of computer technology by the police. *Commun ACM* **22**(1): 10–20.

Gans N, Koole G and Mandelbaum A (2003). Telephone call centers: Tutorial, review, and research prospects. *INFORMS Manuf Service Opns Mngt* **5**(2): 79–141.

Green L and Kolesar P (2004). Improving emergency responsiveness with management science. *Mngt Sci* **50**(8): 1001–1014.

Ignall EJ, Kolesar P and Walker WE (1974). The use of simulation in the development and empirical validation of analytic models for emergency services. In: Morris MF, Steinberg H and Highland HJ (eds). *Proceedings of the 7th Conference on Winter Simulation* — Volume 2, Washington, DC, 14–16 January 1974, *WSC '74*. ACM Press: New York, NY, pp 529–537.

Kleindorfer GB, O'Neill L and Ganeshan R (1998). Validation in simulation: Various positions in the philosophy of science. *Mngt Sci* **44**(8): 1067–1099.

Kolesar P, Pedrinan A and Stein P (1976). Models for assignment of 911 emergency telephone operators. In: Highland HJ, Schriber TJ and Sargent RG (eds). *Proceedings of the 76 Bicentennial Conference on Winter Simulation*. Gaithersburg, MD, 6–8 December 1976, Winter Simulation Conference Pubs., pp 193–197.

Kotiadis K (2006). Extracting a conceptual model for a complex integrated system in health care. In: Robinson S, Taylor S, Brailsford S and Garnett J (eds). *Proceedings of the 2006 OR Society Simulation Workshop*, Leamington Spa, UK. The Operational Research Society: Birmingham, UK.

Kuhn P and Hoey TP (1987). Improving police 911 operations in Washington, D.C. *Natl Prod Rev* **6**(2): 125–133.

Larson RC (1973). On-line simulation of urban police patrol and dispatching. In: Sussman J and Hoggatt AC (eds). *Proceedings of the 6th Conference on Winter Simulation*, San Francisco, CA, 17–19 January 1973. *WSC '73*. ACM Press: New York, NY, pp 371–385.

Law AM (2007). *Simulation Modeling and Analysis*, 4th ed. McGraw Hill: New York.

Lewis PAW and Shedler GS (1979). Simulation of non-homogeneous Poisson process by thinning. *Naval Res Logist Quart* **26**: 403–413.

Mehrotra V and Fama J (2003). Call center simulation modeling: Methods, challenges and opportunities. In: Ferrin D, Morrice DJ, Sanchez PJ and Chick S (eds). *Proceedings of the 2003 Winter Simulation Conference*, 7–10 December 2003. The Fairmont New Orleans, New Orleans, LA, pp 135–143.

Micro Saint Sharp (2005). http://www.maad.com/index.pl/micro_saint (accessed 25 October, 2005).

Pidd M (2004). *Computer Simulation in Management Science*, 5th edn. Wiley: Chichester.

Pidd M and Woolley RN (1980). Just modeling through: A rough guide to modeling. *Interfaces* **10**(1): 51–54.

Rittel HWJ and Webber MM (1973). Dilemmas in a general theory of planning. *Pol Sci* **4**: 155–169.

Robinson S (2004). *Simulation: The Practice of Model Development and Use*. Wiley: Chichester.

Robinson S (2006). Issues in conceptual modelling for simulation: Setting a research agenda. In: Robinson S, Taylor S, Brailsford S and Garnett J (eds). *Proceedings of the 2006 OR Society Simulation Workshop*, Leamington Spa, UK. The Operational Research Society, Birmingham, UK.

Rosenhead JV and Mingers J (2001) (eds). *Rational Analysis for a Problematic World Revisited*. Wiley: Chichester.

Received April 2007;
accepted August 2007 after one revision