Using activity based costing and simulation to reduce cost at a Police communications centre

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>PIJPSM-03-2016-0044.R4</td>
</tr>
<tr>
<td>Manuscript Type</td>
<td>Research Paper</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Workload, United Kingdom, Stress, Performance measure, Effectiveness</td>
</tr>
</tbody>
</table>
Using activity based costing and simulation to reduce cost at a Police communications centre

Abstract

Purpose
This study is based at a Police Force’s communications centre which undertakes a vital role in receiving and processing emergency and non-emergency telephone calls from the public and other agencies. The purpose of the study is to evaluate a method for addressing the conflict between the need to reduce cost and the requirement to meet national standards in terms of a timely response to customer calls.

Design/methodology/approach
In a 2-stage methodology a Activity Based Costing (ABC) approach is used as a framework to show how costs are generated by the three ‘drivers’ of cost which relate to the design efficiency of the process, the demand on the process and the cost of resources used to undertake the process. The study then provides an analysis from a resource driver perspective using discrete-event simulation to model workforce staffing scenarios.

Findings
Cost savings within the police communications centre are identified from an analysis of the three drivers of cost. Further analysis from a resource driver viewpoint using a simulation study of an alternative workforce schedule predicted an overall reduction in staffing cost of 9.4%.

Originality/value
The study outlines an innovative method that identifies where cost can be reduced using activity-based costing and then provides an assessment of strategies that aim to reduce cost whilst maintaining service levels in a police operation using simulation.

Keywords: police; activity based costing; simulation; workforce scheduling

Paper Type Research paper
Introduction

In the UK, as in other countries such as the USA (Srinivasan et al., 2013), Police Forces have been subjected to a number of significant cuts in funding while still being obliged to meet the service standards that are widely imposed on public sector organisations by government. This study is based in a Police communications centre which undertakes a vital role in receiving and processing emergency and non-emergency telephone calls from the public and other agencies. The UK National Call Handling Standards provide a minimum target of 90% of emergency calls from the public answered within 10 seconds and 90% of other calls answered within 30 seconds to all police forces to ensure a high level of service delivered nationally (Association of Chief Police Officers, 2005). These targets are required to be met 24/7 and under varying levels of demand. To provide this service staff work in shifts that provide varying levels of cover to the communications centre. The need to make cuts in expenditure whilst maintaining service levels is particularly stark in this case with a requirement to provide a continued timely response to emergency telephone calls from members of the public whilst implementing a 20% reduction in the cost of operations. This study will present a two-stage methodology that uses activity-based costing as a framework to identify how cost is generated in a government service provider and then employs a simulation model to enable the assessment of operating scenarios regarding workforce scheduling in terms of performance and cost.

Literature Review

The literature review will generalize the police communications centre to call centre operations and outline previous work in analyzing call centre performance. A review of activity based costing (ABC) is then provided before an outline of previous work that uses a combined simulation and activity-based approach upon which this study builds.

The Police communications centre can be considered as a generalized call centre which constitutes a set of resources – usually personnel, computers and telecommunications equipment – which enable delivery of services via the telephone (Gans et al., 2003). Call centres can be classified into “inbound” call centres which receive calls and are generally concerned with assisting callers and “outbound” call centres which initiate calls and are generally concerned with selling activities. The police communications centre can be considered a specialised case of an “inbound” call centre which is generally regarded as more stressful for operating staff than
an “outbound” environment (Rod and Ashill, 2013). The traditional approach to modelling staffing within call centres is through the analytical results of simple queuing models (Atlason et al., 2008) such as provided by Jagerman and Melamed (2003). However due to limitations in this approach simulation has emerged as a powerful method for analysing call centres (Saltzman and Mehrotra, 2001). Reasons for using simulation in preference to analytical methods are summarised by Bouzada (2009) and include the use of statistical distributions more compatible with the input data, the ability of simulation to incorporate the evaluation of other factors of performance such as abandoned calls and the experimental approach of simulation which allows the investigation of system behaviour leading to an understanding of why call waiting times are high. Examples of the use of simulation include Ahgari and Balci (2009) which use the technique to assess the effect of cross-training staff to deal with a mix of phone and email enquiries to a contact center, Min and Yu (2004) which use simulation to validate a reengineering effort based in a call centre operation, Lam and Lau (2004) which use simulation to explore different options for restructuring call centre operations into a single call centre and Kim and Ha (2010) which use simulation to explore the planning of staffing levels in a call centre. Doomun and Jungum (2008) look at the use of simulation to model a call centre as part of a business process reengineering effort. In terms of police communication centres in particular Gunal et al. (2008) look at modelling the centralisation of these centres with the replacement of multiple geographically located call centres with a single ‘virtual call centre’.

Costing systems are often rigid and traditional in the public sector (Halachmi and Bovaird, 1997) and this study will outline the use of ABC as a framework for identifying how cost is generated in a public sector organisation. The Activity Based Costing (ABC) approach (Kaplan, 1992; Gupta and Galloway, 2003) allows the user to distinguish between resource usage and resource expenditure, the difference being unused capacity. Once identified this capacity can either be eliminated, reducing costs, or re-deployed, improving effectiveness. The ABC model incorporates three main drivers of cost of resource drivers which determine the cost of resources, activity drivers which determine the use of these resources and cost drivers which determine the effort needed to undertake the activity (Turney, 1996). Cost can be reduced by either re-configuring the resources needed for an activity (resource driver) for example by using different personnel on lower pay rates, or reducing the amount of resource required (activity driver) for example by reducing the amount of demand for the service, or by reducing the resources needed to perform an activity (cost driver) for example by incorporating information technology in the design. Following on from the activity based cost analysis the main focus for this study is from the resource driver perspective and entails an investigation into the efficient deployment of call handler staff who are required to make a timely response to customer calls.
In terms of the use of simulation in conjunction with activity based costing (ABC) Helberg et al (1994) describe an educational ABC simulation software package that demonstrates the use of simulation in providing sensitivity analysis for ABC scenarios. Kostakis et al (2011) use simulation with ABC to provide cost estimations in the restaurant industry. Özbayrak et al (2004) use a simulation-based model that uses ABC to estimate manufacturing cost behaviour. Greasley (2001) uses simulation with ABC to provide cost information for a police custody operation. However these studies simply use simulation to report costs from an ABC perspective whilst this study proposes a two stage methodology which uses the drivers of cost perspective of ABC to identify how cost is generated and then employs a simulation model to evaluate alternative operating scenarios in terms of performance and cost.

Methodology

A case–based research methodology was chosen with the aim to provide an example of practice and test the proposition that the tools of activity based costing and discrete event simulation can identify where and how cost can be reduced whilst maintaining service levels in the context of a public sector organisation. Although a single-site study has obvious limitations with respect to the generalisability of the findings, the case is not aimed at being representative, but rather exemplary, thus the researcher does not need to assume that what is observed is truly representative of all similar situations (Stuart et al., 2002). In this example a regional police force in the UK shows many of the attributes of many public sector organisations in its structure and reporting relationships, and the involvement of a number of stakeholders in determining strategic targets and operational rules. Collection of data such as process durations, task allocations and process relationships for the study was undertaken by the author and employees of the organisation in order to construct the process maps and simulation used in the analysis of the process designs. These tools thus provided a way of achieving an understanding about the current and future potential design of organisational processes without the risk of disruption to the real system itself.
The Case Study

Background

There are 43 independent police forces in England and Wales with each force responsible for answering and responding to incoming calls for service through the use of communication centres. The case study concerns the Warwickshire Police Force, based in the county of Warwickshire in the West Midlands area of England serving a local population of around 550,000 people located over an area of 763 square miles. Its workforce accounts for around 80% of costs. The Warwickshire police communications centre examined in this case study has a total annual call volume of around 285,000 calls.

The main processes within the communications centre will now be described and the proposed change to the shift pattern will be outlined. Figure 1 shows the main processes involved in the communication centre. 999 (emergency) calls are sent directly to call handlers with non-emergency calls sent via a switchboard. The switchboard operator will either resolve the caller query or direct the call to a call handler. Call handlers will take the relevant information from the caller and grade the call for an appropriate response. If attendance to an incident is necessary call handlers pass the call information on to a team of controllers who manage the deployment of officers on the ground. It will be a specific controller who manages the incident, controllers are responsible for a geographic area so if the incident occurs in their area they will have to manage it regardless of how busy they are. Thus controllers control the resources in a geographically area unlike call handling where the call goes to the next available call handler. The simulation study boundary is indicated in figure 1 and encompasses the incoming call and call handler process but with the controller process outside the remit of this study.
The study considers the communications centre from the perspective of each of the three drivers of cost (figure 2). The 'cost driver' relates primarily to the design efficiency of the activities within the call handler process of the communications centre. The 'resource driver' relates primarily to staffing costs for personnel involved in the arrest process. The 'activity driver' relates to the timing and frequency of the different call types made to the communications centre. Costs were computed for each call type (e.g. 999 emergency, 999 switchboard). Following on from that a more detailed investigation from a resource driver perspective was undertaken.
The Simulation Study

This project utilises the technique of discrete event simulation modelling using the ARENA (Kelton et al, 2014) simulation modelling software. The simulation study methodology will be used as in Greasley (2006).

Study Objectives

The first stage of the simulation study will use the framework provided by ABC to identify how cost is generated taking into account the variability of demand and variability in process durations that occur within the communications centre. The second stage of the simulation study will than make an assessment of the relative reduction in cost and communications centre performance between the current staff shift pattern and a proposed 9 hour, 5 shift staffing scenario. The aim is to enable managers to improve the efficiency of the deployment of staff whilst maintaining target service levels.

Data Collection and Input Modelling

The project involved the collection of data regarding the type of calls received, termed skillsets. Skillsets include 999 emergency calls, calls activated by house alarms, emergency calls from the Fire Service and calls from other Police Forces (table 1). Information was collected on each skillset regarding its priority with calls in skillsets that have the highest priority (with a lower priority ID number) jumping the queue for a response over calls in other skillsets. Information was also collected on the target response time to answer a call. The target maximum response time for 999 emergency calls from the public, which represent approximately 28% of all calls, is currently 10 seconds. For all other emergency and non-emergency calls the target response time is 30 seconds. Calls from the public are routed via the national exchange and so can be distinguished from other emergency calls for performance monitoring purposes. Emergency calls from other services such as the fire brigade will typically be a request for police to attend an incident with them so speaking with an operator from the fire contact centre.
In order to model call demand historical data was collected on the number of calls for each skillset hour by hour over a 24 hour day and on different days of the week. The demand for each skillset in terms of calls made during each hourly slot is assumed to be constant and so was modelled using an exponential distribution which may be used to model the interarrival time of “customers” to a system that occurs at a constant rate (Law and Kelton, 2000). This approach follows that normally used in input modelling of calls into a call centre (Chick, 2001). Data was also used to estimate statistical distributions for the talk time (time on the phone to a caller) for each skillset. For some skillset types such as ‘Ambulance General’, ‘Other Agency Emergency’ and ‘Coventry Airport’ the number of calls over the sample period is very low and so was modelled using a triangular distribution which is recommended for activity times when there is little or no data (Kelton et al., 2014). Other skillsets such as ‘999 Emergency’ were modelled using a log-normal distribution reflecting a general tendency for the majority of calls
to be short with a minority of much longer calls occurring. This approach was seen as an improvement on the
traditional approach of modelling call handling times using an exponential distribution based on average call
handling times derived from the call centre automated call distribution systems. In this instance sample call
handling times for each skillset were derived from skillset performance reports supplied by the force. Data was
also collected and modelled on any required communication with controllers and paperwork subsequent to a
call, termed the wrapup time.

In addition data was collected on calls that were not completed due to the caller ‘hanging up’ the telephone,
these are called abandoned calls. Abandoned calls within the target response time could be considered to be
often calls made in error while abandoned calls outside the target response time could be said to be often due
to slow response. Separate data was collected on abandoned calls both within and outside the target response
time and separately for the skillset 999 emergency. These were modelled using an exponential distribution as

Building the Model

A visual display of the ARENA simulation is configured as a dashboard and is shown in figure 3. The figure shows
performance metrics for 16 skillsets. Further skillsets could be added to meet future requirements if necessary.

![Image of the Police Communications Centre Simulation Model Dashboard Display]

**Figure 3. The Police Communications Centre Simulation Model Dashboard Display**

In order to explain the dashboard display figure 4 shows an enlarged view of the ‘999 emergency’ skillset.
The top of the skillset display shows the skillset number, in this case 5682, and the skillset description. The top line metric is the average wait time for customer calls in seconds. The target response time for a call of that skillset is displayed on the left hand side of the next row down (In this case the target response time is 10 seconds). This row also displays the percentage of calls that have been answered within the target time (93%). This is the key indicator of performance and should be higher than 90% for all skillsets. The next value of this row is the number of calls made to this skillset during the simulation run period (1766). The next row down displays the maximum wait time for a customer call during the simulation run (319.2 seconds) and the final row indicates the day number (day 6) and the time of day that the maximum wait time occurred (14:38). This information was requested by the Force as they were interested in being able to quickly identify lengthy calls in order to investigate the reasons behind them.

To the right hand side of the dashboard display in figure 3 there is also displayed the percentage of abandoned calls made both within the target response time and outside the target response time. The abandoned calls metrics are displayed for both 999 emergency and other calls made.

Model Validation
The model was validated by simulating the current staffing profile and comparing the results to historical data on performance collected by the Force. It became apparent at this stage that in order to provide an accurate assessment of the system it was necessary to incorporate into the model a 2 second delay that was observed between the call entering into the Call Handler system (either from the switchboard or a direct 999 call) and it connecting to the call handler and alerting them that they are about to answer a call. This was implemented as a delay block in ARENA with a fixed duration of 2 seconds.
Experimentation and Analysis

A 2 stage approach was undertaken to identify how cost is being incurred and then to evaluate operating scenarios that were designed to reduce cost. In stage 1 the simulation was used to analyse the costs of the communications centre’s operations from the three perspectives of activity driver, cost driver and resource driver. The simulation is run for a period of 4 weeks (28 days) operation and results are collected from 1000 replications of the model which was considered sufficient to achieve an appropriate confidence level. Although the communication’s centre is in continuous operation the very low number of calls likely to be in progress at the start and end of each run period is assumed to have a negligible effect on system results. This allows the communications centre to be considered as a terminating system and so no methods, such as warm up periods, are required to deal with the initial transient phase of each simulation run.

Stage 1 The ABC Analysis

In terms of activity driver analysis, figure 5 shows the average total cost for each call type (skillset) to the call centre over the simulated four week period. The graph shows the cost of the call handler staff determined by the nature of demand from members of the public seeking assistance. To decrease overall cost requires a reduction in demand and the cost information assists in showing where cost is being incurred. From an activity driver viewpoint showing costs by call type (skillset) the analysis prompted some discussion in terms of reducing the high volume of 999 emergency calls (many of which are not classified by police as actual emergencies) and discontinuing the information call line in favour of email and web based facilities.
In terms of cost driver analysis, figure 6 shows the average total cost of operation for a 4 week period, from breaking down the call handler task into its two main activities of direct talk time with a customer on the phone and a subsequent wrapup time which includes any communication with the controllers and required paperwork.

From a cost driver viewpoint the analysis prompted discussion of the relatively high proportion of wrapup time and the possibility of training to reduce this time. Laureani et al. (2009) and Piercy and Rich (2009) found that lean techniques can improve the operation of a call centre through improvements such as an increase in first-call resolution which reduces the failure created by failing to answer the query in the first place. An improvement in response resolution could be incorporated into the model and used to predict the overall impact on staffing needs.
In terms of resource driver analysis, figure 7 shows the average total call handler cost over a 4 week period. It can be seen from the graph that a large proportion of call handler cost is taken up with idle cost when a call handler is not busy either taking a call or undertaking call wrapup duties. Although there are a few administrative duties given to staff to complete during this idle time they are not so onerous that they cannot be achieved in less idle time than is currently available. Thus there is potential to improve efficiency by reducing this call handler idle time.
Figure 7. Average Total Cost by Call Handler Allocation (Resource Driver)

Stage 2 The simulation analysis of a proposed workforce scheduling scenario

The analysis in stage 1 showed potential improvements in operation from all 3 ABC perspectives. In this case it was decided to investigate the call handling process from a resource driver viewpoint by considering if alternative workforce schedule scenarios could achieve a closer match between supply (call handler staff) and demand (calls) and thus reduce call handler idle time leading to a decrease in direct labour costs. As with any call centre operation matching supply and demand represents a difficult task in that demand fluctuates over yearly, monthly and daily time horizons and short term demand spikes in customer calls may occur due to events such as a football match. The ability to buffer demand is very limited due to the short time span required to respond to calls. In addition there is also the need to incorporate staff holidays and staff absences in the shift pattern which represent major issues in preventing the smooth running of any shift pattern (Jezewski and Moore, 2008). The shift pattern review is now outlined in more detail.

The current 8 hour, 4 shift pattern (3 shifts used to cover each 24 hours)

Currently the communications centre staff are on an 8 hour shift working 7 days on then 2 days off covering a 24 hour period. Additional time owed to staff is made up with a floating rest day that is added either during a tour or added to the number of rest days owing. An example call handler schedule using this scenario is shown in
figure 8 which shows the ARENA Schedule block facility which is set up to allow adjustment to the call handler staff on duty at hourly intervals.

Figure 8. Example of current shift pattern entry in ARENA simulation model

The proposed 9 hour, 5 shift pattern (3 shifts used to cover each 24 hours with 3 hours of overlap)

Based upon analysis for the whole force it was decided that a shift pattern requiring a 9 hour shift working 6 days on then 4 days off would be examined. An example call handler schedule using this scenario is shown in figure 9.
Figure 9. Example of proposed shift pattern entry in ARENA simulation model

It is proposed this shift pattern scenario has the following advantages:

- 3 shifts per day working a 9 hour shift provides a theoretical 27 hours of cover over a 24 hour period. The additional 3 hours can be used to overlap staff at certain times in the day allowing greater flexibility in adapting shifts to better match demand and allowing staff to have effective briefings without impacting on performance.
- The shift pattern incorporates 5 training days per year.
- 60% of the workforce (before abstractions) are on duty most days.
- Nine hour shifts can easily be extended to 12 hour shifts at times of peak demand (public holidays, Halloween, large events happening locally).
- It’s more popular with employees offering a better balance between work and home life due to the rest period of 4 days (as opposed to 2 days in the current scenario) between tours.

Notwithstanding the advantages stated above the simulation model was used to investigate the ability of the new shift pattern to provide a closer match between staff levels and predicted call levels over time and thus improve efficiency by reducing staff idle time. The proposed shift pattern has been designed using workforce...
planning intervals of an hour in conjunction with overlapping shifts. This enables the number of employees to be scheduled to change from hour to hour with the aim of having just enough staff present to provide the desired level of service (Thompson and Goodale, 2006). The results from the simulation model are automatically exported from the ARENA simulation model to a spreadsheet for statistical analysis and the production of graphical displays. Figures 10-12 are graphs that show the performance of one shift (either ‘early’, ‘late’ or ‘night’ as defined by the current model) for emergency calls (graphs were also produced for non-emergency calls) over a 1 week (7 day) cycle. The results are based on 1000 replications and at a 95% confidence level shown by the dashed lines.

Figure 10. 999 calls response performance on the early shift
It can be seen from the graphs that the overall performance based upon this level of resource is generally meeting the required targets although there is an area of performance below target on the late shift. The reduction in call handler idle cost estimated by the simulation is presented in figure 13 and shows an average cost reduction of 9.4%.
Figure 13. Comparison of Average Total Cost by Call Handler Allocation (Resource Driver)

Discussion

This study takes a 2-stage approach to the challenge of improving efficiency during times of budget cuts whilst maintaining performance in a critical area of police operations. Traditionally functional areas such as call centres are managed with an overall budget adjusted annually to take account of increases in staff pay. However in order to reduce cost without affecting performance the first stage of the method uses the activity-based costing approach to express payroll cost as a function of time and thus enable the cause of cost incurred to be identified. Although calculating cost as a function of direct labour is standard practice, the ABC method permits identification of how cost is incurred, in this case as a result of the efficiency of the call handling processes, the timing and mix of call demand on the process and the cost of resources used to undertake the process. A reduction of cost could be tackled from all of these perspectives but in this case study analysing cost from a resource driver viewpoint led to the second stage of the analysis which presents a simulation study of an alternative staff scheduling scenario for the communications centre. The benefit of the simulation method is that it enabled the exploration of alternative staffing scenarios without disruption to the operation of what is a critical process. In terms of results it was found when comparing across the two versions of shift patterns that the simulation predicts an overall reduction in direct labour staffing cost (by reducing staff idle time) of 9.4%.
This result represents approximately half of the 20% reduction in cost required by the Force but further work on alternative workforce schedule scenarios could achieve a closer match between supply (call handler staff) and demand (calls) and thus reduce costs further. In addition further exploration of costs incurred from the alternative perspectives of the ABC approach could be undertaken. For example from an activity driver viewpoint showing costs by call type (skillset) demonstrated the potential for additional cost savings by focusing on a reduction in the high volume of 999 emergency calls by discouraging non-emergency calls to this number and discontinuing the information call line in favour of email and web based facilities. From a cost driver viewpoint the analysis showed the relatively high proportion of wrapup time which could be reduced by additional staff training.

In general terms the model was useful in that for the communications centre the cost of failure of any change is very high so it needed to be done correctly with a high level of confidence that target figures could be met before a change was made. The approach supplements other examples of the use of models to assist decision making in critical situations. For example Srinivasan et al (2013) build a simulation model to evaluate staffing levels in a police patrol workforce whilst meeting performance benchmarks and Eatock et al (2011) build a simulation model which can be used to assess the effect of changes to staff rosters to better match busy periods in a UK Accident and Emergency (A&E) department. Here the target is that 98% of patients should be either discharged or admitted to hospital within 4 hours of arrival at A&E.

In terms of limitations the study results should be viewed taking into consideration assumptions made around the input data used by the model such as call demand. Future actions such as the planned introduction of a non-emergency 101 call number aimed to reduce ‘non-emergency’ 999 calls will need to be assessed and incorporated into future data sets used by the model. In addition the performance of the call centre is dependent on a number of factors outside the scope of this study such as those indicated by Rowe et al (2011) as the profile of the call handlers, the division of labour, goals, reward systems and the use of technology to automate call distribution. For example Kekre et al (2011) explore strategies in a call centre for determining which calls should be handled by which agents. In this case the development of the staff shift pattern needed to take into account many issues such as meeting the workload demand, taking into account staff absence but also qualitative issues such as ensuring staff wellbeing in what is a particularly demanding environment. This need to consider qualitative issues, as well as quantitative performance metrics, in order to reduce stress levels in a call centre environment has been recognised by Robinson and Morley (2006). Although the concept of staff wellbeing could not be incorporated directly into the model in this study the model users could use their knowledge of the task, for example predicting periods of high stress, to adjust the staff rota accordingly. One
finding that points to the connection between shift pattern worked and wellbeing is reported in Richbell et al (1998) who provide an assessment of the application of a Canadian concept in shift working (the Ottawa shift system) in a UK policing context. This study found that 66% of respondents of the survey noticed an improvement in their own health and wellbeing since working the Ottawa shift system. In addition Tyagi and Dhar (2014) report that in the context of police operations work overload was found to be the highest contributing factor in causing stress amongst police officers.

Finally in terms of methodological limitations a single case study has been used which is appropriate for the exploratory nature of the research but additional cases which would extend the study to new populations would increase the theoretic generalisability of the findings (Meredith, 1998) and this is an area for further research.

Conclusion

Police services continue to face the challenge of meeting government imposed performance targets whilst operating under increasingly constrained budgets. This paper presents a methodology that uses activity-based costing as a framework to identify how cost is generated in a police communications centre operation. This enables a targeted approach to cost reduction to take place, in contrast to an ‘across-the-board’ approach to budget cuts. A discrete-event simulation model is then used to enable the assessment of operating scenarios to take place from a resource driver perspective, in this case in the area of workforce scheduling.

References


