Feedback Control Ideas for Call Centre Staffing

M. R. Arahal, D. R. Ramirez* and E. F. Camacho

Dpto. de Ingeniería de Sistemas y Automática, Escuela Técnica Superior de Ingenieros, Universidad de Sevilla, Camino de los Descubrimientos, s/n., 41092 Sevilla, Spain

(Received 00 Month 200x; final version received 00 Month 200x)

Call centres are nowadays a widespread solution to deal with customer support and as platform for different kind of business. Call center staffing is crucial to provide adequate service levels at acceptable costs. The task is usually accomplished using heuristics with the help of a human experts or with some static off-line optimization based on operations research. Simulators based on queue theory and statistical analysis are, in some cases, also used. The aim of the paper is to show that call centre staffing can be posed as a feedback control problem with the advantage of getting a higher level of automation, and a wealth of results from control theory that can help to obtain the best possible staffing. In the paper the working procedures of call centres are described and how the staffing is usually made. A feedback controller is proposed and assessed in simulation. The results show that good call centre staffing can be obtained even with a not very sophisticated controller.

1. Introduction

Call centres are nowadays present in companies that have to deal with a large number of customers, and some businesses (e.g. airlines, hotels, and credit-card companies) rely almost exclusively on call centres to provide service and to obtain customer feedback (Andrews and Cunningham 1995).

The capacity of a call centre is determined mostly by the human resources used. Since these are expensive, the quality of the service is often balanced with capacity, so that the call centre can provide the best service with minimum costs (Pinedo et al. 1999). A widely used approach to scheduling is to use Markov chain (Deslauriers et al. 2005) and queueing models to simulate the behavior of the call centre under different circumstances. In Koole and Mandelbaum (2002), a survey of queueing theory applied to telephone call centres can be found with many relevant references.

Forecasting techniques can be useful for predicting arrival rates (Jongbloed and Koole 2001). This information can be used by a simulator of the call centre to test different staffing situations, allowing the planner to produce a staffing plan that balance quality of service and costs. Prediction techniques can also be used in a dynamic staffing scheme such as in Whitt (1999). A more elaborate procedure is to compute the call centre staffing for a long period using static optimization, more precisely, solving an integer programming problem in which costs and service are considered through a model like the well known Erlang C formula (Gans et al. 2003, Tijms 2003). In this case, the staffing is applied in an open-loop manner and unpredictable fluctuations in the number of incoming calls have a negative impact in the service levels. In fact, using such staffing forces the management staff to decide on which policy to follow: a service-oriented policy in which more agents than needed on average are hired (with higher costs), or a cost oriented policy in which less agents are hired (with lower service levels).

*Corresponding author. Email: danirr@cartuja.us.es
In this work the authors propose the application of feedback control techniques to the scheduling of the number of agents needed to provide a desired quality of service at any time. To the best of authors knowledge, the call centre staffing has not been previously considered as a feedback control problem (although controllers based in discrete events systems have been used in Seidl (2006)). In this problem, the number of incoming calls that must be answered is considered as the set point for the number of answered calls. In this work the requirements and working procedures of a typical call centre are considered to build a realistic model of a call centre. This model has been implemented as a simulator that can be used to test both forecasting and feedback control policies. The control strategies considered in the paper are proportional plus integral controllers together with feed-forward compensation. The results included in the paper show that feedback control can be used to obtain a better call centre staffing than that provided by the usual operation.

The paper is organized as follows: section 3 presents the working principles of call centres. The model of the call centre is presented in section 4, whereas the usual ways of call centre staffing are briefly discussed in 5. The results of scheduling based in feedback control are presented in section 6. Finally, the conclusions and future work plans are discussed in section 7.

2. Feedback control

Control theory is an interdisciplinary branch of engineering and mathematics that has also been considered in economy and management science (Sengupta and Fanchon 1997, Sethi and Thompson 2005). Even though most of its applications are related to engineering, its principles can be applied to other kinds of problems, such as inventory control, marketing, supply chain management (Schwartz et al. 2006), etc. The purpose of a control system is to find ways to cause a system to behave in a desired manner. The most reliable way to accomplish this is by using a negative feedback control scheme. In such a scheme, the control system compares the desired behavior $R(t)$ with the actual behavior exhibited by the controlled system, $Y(t)$, and dynamically computes a control signal as a function of the difference between $R(t)$ and $Y(t)$, that is $U(t) = f(E(t))$, where $E(t) = R(t) - Y(t)$ is the tracking error. Note that the control signal $U(t)$ is a vector of that can contain whatever manipulable variables that allow to attain the desired behavior. In control theory, the desired behavior $R(t)$ is usually given in terms of a reference or set-point, the actual behavior of the system concerns the outputs $Y(t)$. There can be other input variables that are not manipulable but that affect the output, these are referred to as disturbances $P(t)$.

The main goal of control theory is the design of controllers that allow to attain the desired behavior of controlled systems while some desirable properties such as stability or control performance are guaranteed. Many different approaches to control systems design have been proposed since the initial works by J.C. Maxwell in the late 19th Century (Maxwell 1867), but most of them rely on the negative feedback control scheme described in this section (the reader is referred to classical texts on control theory such as Ogata (2001) and Ogunnaike and Ray (1994) for more details on control structures). In this work, the authors propose the use of this scheme.
to solve the problem of call centre staffing, which will be described in the following sections.

3. Call centres

Call centres are designed to cope with a variable load of incoming calls. Those calls are transferred to a human operator selected according to certain characteristics such as language, geographic source and nature of questions to be answered. The call can, in turn, be transferred among operators until it is finally attended. In this process it may happen that the call is lost because an excessive waiting time (see Zohar et al. (2002) for an interesting study of abandonment behavior), too much load on phone lines and other technical problems related with the hardware of the call centre. This work only deals with the first situation as line saturation or hardware problems do not depend on the call centre staffing (that is, the number of operators assigned to each shift). On the other hand, the number of calls lost because of a lack of free operators can be minimized with a proper staffing which is the objective of this paper.

A call centre has the following quality of service goals:

1. Low number of lost calls. The quality of service is related to the ratio of number of calls that are attended over the total number of incoming calls.
2. Low waiting time. Even if a call is attended, the customers do not like excessive waiting.
3. A correct support to customers. This is related to the training of the operators, thus it will not be considered here.

On the other hand, these objectives must be attained at the lowest possible cost, that is with the lowest number of operators. The ideal situation is that in which the number of operators (hourly offer) is balanced with the number of incoming calls (the hourly demand) in such a way that the both the number of lost calls and average waiting time is acceptable and the cost is minimum for the service given. This situation is the goal of the scheduler and can be met using feedback control techniques. This goal may not be easy to reach because:

• The future hourly demand is uncertain.
• The time required to attend each call is uncertain.
• The work capacity of each operator is also uncertain, although proper training and supervision can help to reduce this.
• The scheduling must be done a few days in advance.
• The quality of service is against the cost reduction goal.

The three first causes can be considered as stochastic effects and their impact on the call centre is variable. But the last two causes are universal and cannot be avoided. Specifically, none of the quality of service objectives cannot be perfectly assured with several days ahead unless an excessive number of operators are assigned to each shift.

4. Call centre model and simulator

The working principles of the model and simulator to test the feedback control based staffing are presented in this section. The scheme of the model is depicted in figure 2. The model is build around a queue that represents the number of calls that are waiting to be attended. Queue models are often used to test staffing strategies for call centres (Sze 1984) (see Gans et al. (2003) for a review on mathematical aspects of call centre modelling). As the incoming calls arrive they are queued until they are answered by an operator $O_i$. Operators pick calls from the queue and attend the calls. Each call takes a different time to be attended. On the other hand some calls can abandon the queue without being attended because customer impatience due to an excessive waiting time. Also those calls that arrive to a full queue are rejected due to line saturation.

The variables used in the model are discrete time variables with a sampling time of an hour.
The following variables are considered.

Inputs

- $d(k)$: The hourly demand, that is, the number of incoming calls at hour $k$. Note that here there is no distinction between incoming calls that arrive for the first time and those which are retrials (the relation between the sizing of the call centre and retrials have been recently studied in Aguir et al. (2008)).
- $u(k)$: The number of operators assigned at hour $k$.
- $C$: Size of the waiting queue. It has been considered constant in this paper.

Outputs

- $y(k)$: Number of calls attended at hour $k$.
- $u_i(k)$: Average number of idle operators at hour $k$.
- $u_a(k)$: Average number of operators actually answering calls at hour $k$.
- $w(k)$: Average waiting time for calls attended at hour $k$.
- $a(k)$: Number of abandonments at hour $k$.
- $s(k)$: Number of saturations at hour $k$.

It is clear that $d(k) = y(k) + a(k) + s(k)$. Also the model considers some parameters to represent the properties of any given call centre. These parameters are:

- The probabilistic function that models the incoming calls. From a control point of view, this is not a parameter but an input variable. Queue literature uses an stochastic approach and consider that the number of incoming calls and their duration are parameters of the call centre. In the simulator implemented the number of incoming calls can be computed from this probabilistic function or be extracted from a given sequence $d(k)$.
- The probabilistic function that models the duration of the calls.
- The probabilistic function that models the probability of losing a call for excessive waiting.
- The size of the queue. As stated before, this is considered a fixed parameter.
- The probability $P_p$ that an idle operator stays the current minute without attending a waiting call.
- The speed at which operators start working at the beginning of a given shift. This is modeled as a parameter $A_f$ which is used as the time constant of a filter $u'(k) = A_f u'(k) + (1 - A_f) u(k)$, where $u(k)$ are the operators assigned at hour $k$ and $u'(k)$ the operators that are already available (either attending calls or idle).

The previous parameters can be tuned to produce a simulator that produces simulations statistically similar to the observed behavior of any given call centre. Normal probability density
functions have been used in the reported simulations for

- The waiting time that a call stays in the queue before it is lost, defined by the average $\mu_e$ and the typical deviation $\sigma_e$.
- The time that a given call takes to be attended, defined by the average $\mu_d$ and the typical deviation $\sigma_d$.

The model can be enhanced to include relationships between variables. For instance, the dependance of $P_p$ on the time that the operator is working without a break or the relation between $P_p$ and the total number of waiting calls. Finally, the simulator uses an integration step of 1 minute, although the results given by its outputs are the cumulative values of each variable over 1 hour. The reason is that statistics data in call centres are given at each hour. Also, some input variable must be integer values, thus the simulator rounds non-integer values in that variables.

4.1. Open loop tests

The dynamic nature of call centre behaviour is not clearly seen in practice due to the complex trajectories of the variables following daily and weekly patterns as well as stochastic variations. The classical step response of process control is not encountered in this realm. However the simulator can provide some results for such tests, providing insight to develop the proposed controller.

Some simple examples that show the behaviour of the call centre model are given in the following. Consider the figure 3 in which there are some steps in the number of scheduled operators and the hourly demand. The test begins with a demand that cannot be satisfied with the current number of working operators. Thus, there is a significant number of abandonments every hour, and a high number of incoming calls rejected due to line saturation. In this situation, the number of idle operators is near zero and the waiting time for each call is moderate, about 5 minutes. Then, at hour 170, the number of working operators is incremented so that the demand can be satisfied, the number of lost and rejected calls is about zero and the waiting time is reduced to about 2 minutes. In fact, the number of operators working from hour 170 is excessive as it can be seen in the number of idles. At hour 330 the demand increases but it can be handled with the current number of working operators. In this situation the call centre is operating with a better productivity-costs ratio as there are few idle operators, no rejected calls, an average waiting time of about 3 minutes and the number of abandonments at every hour is quite low. From hour 500 to 670 the number of working operators is reduced and the situation is similar to the initial period, but with a lower difference between the demand and the number of attended calls, and also lower numbers of rejected and abandonments and lower waiting time (about 4 minutes). Then at hour 670 the demand is lowered and the situation is similar to the period from hours 170 to 330. In the last period, from hour 800 to 1000, the number of working operators is reduced so that the demand is barely satisfied, with no idle operators and a low number of abandonments and no rejected calls due to line saturation.

5. Commonly used Staffing Strategies

The scheduler guided by experience and, in some cases, with the help of queue models and statistical assumptions establishes the call centre staffing plan with a few days in advance. The main factors that are considered are listed below.

(1) Future hourly demand forecasting. This forecasting can be obtained using formal methods or just heuristics based on experience.

(2) Past staffing. The scheduler can use historian data to see the number of attended calls by a given number of operators.
(3) Degree of satisfaction of the quality objectives. The scheduler can take into account to what extent the objectives (lost calls, waiting and number of idle operators) have been achieved with a certain staffing.

(4) Operational restrictions. The scheduler has not complete freedom to assign operators at each shift because of labor agreements and other legal issues.

The staffing obtained taking into account these factors can be interpreted as the result of a non-formal control strategy. Two cases can be established depending on how the scheduler processes the available information.

(1) The scheduler does not take into account factor 3. In this case the staffing is equivalent as an open-loop control with a feed-forward compensation. The control signal only depends on the forecast of the hourly demand. This control signal can be computed explicitly from a set of rules or be implicit in the way that the scheduler takes the staffing decisions. In either case, two strategies can be considered:

a) The feed-forward compensation do not change over time. This is not very reasonable as it implies that situations of over/under staffing are not corrected. This situation may happen in practice as hourly and weekly patterns as well as stochastic behavior of the load make it difficult to spot the problem. This is illustrated in figure 4. The test starts with a proper staffing for the usual hourly demand. Few calls are lost by excessive waiting or line saturation. Then at hour 175 the profile of the hourly demand increases and it can be seen that the staffing is now insufficient and the number of lost calls and the waiting time rise. Later, the hourly demand decreases to a point in which the staffing is able to meet the demand, but with an excessive number of assigned operators which yields in an increase of the number of idle operators.

b) The feed-forward compensation is adjusted over time to improve the staffing. In this case the strategy is equivalent to an adaptive inverse controller (Widrow and Plett 1996). This approach has also the effect of adding dynamics to the system due

---

1 A feedforward compensator is an open-loop compensator that it is used to reduce the effects of disturbances that can be measured by acting before the deviations show up. It is very sensitive to model errors and thus it should be used together with a feedback control structure.
to the outer adaptation loop. Thus, the situation in this case is similar to that in which feedback control is used to plan the call centre staffing. We will not explore this control strategy in this paper.

(2) The scheduler takes into account factor 3 either explicitly or implicitly. In this case the staffing is similar to that obtained with a feedback controller plus a feed-forward effect. The scheduler, however, does not make use of control theory or a particular control strategy when the staffing is made. This leads to the consideration of call centre staffing as a feedback control problem, with all the usual requirements of stability, robustness, etc... but also it provides an excellent way to achieve all the quality objectives of the call centre.

The dynamic assignment techniques (see Whitt (1999)) are commonly used in many call centres to transfer operators between different sections depending of the needs. With this technique, operators that are inactive in some sections become active operators of some others thus making the offer on a par with the demand. However, this technique cannot be applied to call centres to just one section or service. Also, this technique is very demanding in terms of knowledge and training of the operators.

6. Call centre staffing based on feedback control

6.1. Control scheme

The problem of computing the staffing of a call centre can be explicitly posed as a feedback control problem. In this case, the set point to be tracked is the hourly demand, the controlled variable is the number of calls attended and the manipulated variable is the number of operators assigned to each particular hour. Since the staffing must be done with some days in advance, the effect of the control signal appears on the output with a delay $L$ which in this paper is assumed to be a week, that is $L = 168$ hours. This delay is constant and no uncertainty is considered in its value.

On the other hand, the hourly demand, which acts as the set-point, can also be seen as an external measurable disturbance. Thus, although its value is not known in advance, a forecast can be used in a feed-forward compensator. This forecast of the future demand can be obtained
from any method including neural networks, time series and polynomial fitting.

The controller can be designed using any control strategy that fits into the control structure shown in figure 5. The objective of this paper is to show how call centre staffing can be posed as control problem, hence it has been chosen a simple proportional plus integral controller together with a feed-forward static action. Block C in figure 5 corresponds thus to the following feedback control law:

\[ u(k) = u_p(k) + u_r(k) \] (1)

where:

- The term \( u_p(k) \) is computed from a feed-forward compensator according to:
  \[ u_p(k) = \hat{d}(k \mid k - L) \cdot \hat{\mu}_d/60 \] (2)

  where \( \hat{d}(k \mid k - L) \) is an estimation of the hourly demand at hour \( k \) made with \( L \) hours in advance and \( \hat{\mu}_d \) is an estimation of the average time needed to answer an incoming call. This average is variable and depends of which particular day or hour is considered.

- The term \( u_r(k) \) is a computed from a feedback proportional plus integral controller according to:
  \[ u_r(k) = u_r(k - L) + k_i(\hat{\mu}_d/60e(k - L) - x(k - L)) \] (3)

  where \( e(h) = d(h) - y(h) \) is the number of calls that were not attended the previous week at hour \( k \). At the beginning of the simulation, \( u_r(k) \) is chosen to be zero, that is \( u_r(1) = 0 \), because it cannot be computed using (3). The parameter \( k_i \) can be made equal to zero, which means that only feed-forward is considered.

### 6.2. Simulation results

In the following, some results of feedback control based staffing are presented. Note that in the figures presented in this section, the number of assigned operators at hour \( K \), that is \( u(k) \), were computed by the controller at hour \( k - L \), where \( L = 168 \) (the staffing is computed one week in advance). Also, the size of the queue has been considered constant and equal to \( C = 20 \). Note that this parameter affects to the number of rejected calls due to line saturation.

In the first case, a PI without feed-forward compensation and \( k_i = 0.4 \) is shown in figure 6. There is a step in hourly demand at hour 340 and the controller, reacts to attend that increase in hourly demand, but too slowly due to a poor tuning.

Figure 7 shows how a better tuned PI, with \( k_i = 0.7 \), reacts faster to the step in the hourly demand, so that the demand is meet properly. However, there is a delay between the step in demand and the time at which the controller starts to increase the number of assigned operators. This is because no feed-forward compensation was taken into account.
The effect of feed-forward compensation can be seen in figure 8 in which a more realistic profile of the hourly demand is also used. In this case a perfect forecast of the demand is used in the feed-forward, to see what can be obtained from the control strategy. Note how the effect of the delay is compensated by the feed-forward. Also the number of lost calls for excessive waiting or line saturation at every day are quite constant, regardless the variation in the demand profile observed in the figure. These results can be compared with that obtained in figure 4. A comparison of the essential data is shown in Table 1. The open-loop staffing of figure 4 gets slightly better results in service levels the first part of the experiment, in which the average number of lost calls, either by abandonment or saturation, and the average waiting time are lower than that of the closed loop. But note that the number of idle operators is slightly greater. In the second part of the experiment, the profile of the demand changes from the expected conditions. In this part, the
Table 1. Comparison between Open-loop and Closed-loop results

<table>
<thead>
<tr>
<th>time</th>
<th>( \text{avg}(l(k)) )</th>
<th>( \text{avg}(x(k)) )</th>
<th>( \text{avg}(w(k)) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 ( \leq k \leq \frac{T}{4} )</td>
<td>5.86</td>
<td>3.38</td>
<td>3.08</td>
</tr>
<tr>
<td>O.L.</td>
<td>( \frac{T}{4} \leq k \leq \frac{3T}{4} )</td>
<td>38.33</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>( \frac{3T}{4} \leq k \leq T )</td>
<td>0.33</td>
<td>14.5</td>
</tr>
<tr>
<td>0 ( \leq k \leq \frac{T}{4} )</td>
<td>7.13</td>
<td>2.52</td>
<td>3.54</td>
</tr>
<tr>
<td>C.L.</td>
<td>( \frac{T}{4} \leq k \leq \frac{3T}{4} )</td>
<td>6.75</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td>( \frac{3T}{4} \leq k \leq T )</td>
<td>7.32</td>
<td>2.41</td>
</tr>
</tbody>
</table>

\( T \): length of experiment. \( l(k) = a(k) + s(k) \), i.e. total number of lost calls.

open-loop staffing gets much worse results than the proposed staffing, with almost six times more lost calls and a substantially longer waiting time. In the third part, the open-loop staff gets better service levels but at a much higher costs due to a number of idle operators six times bigger than that of the proposed staffing. It is noteworthy how the closed-loop nature of the proposed staffing strategy allows to get similar levels of service and cost under changing values of the demand.

Figure 9 shows an example of staffing with a more realistic profile of hourly demand that, in addition, cannot be perfectly forecasted. Thus in this example the demand profile is more difficult to meet than in the previous example. In this example, it has been considered an error in the demand forecast of \( \pm 2\% \) of the average hourly demand. This error is what it can be expected, for example, from a neural network based predictor (an example of prediction with neural networks applied to call centres can be found in Balaguer et al. (2008)). It can be seen that the number of lost calls is moderately low, there are some hours in which it gets higher, but the controller reacts making it lower.

Furthermore, some simulations have been performed using real data obtained from a call centre service of one of the subsidiaries of Spanish telecom Telefónica, SA. The values of the load are shown in some plots that reveal certain characteristics. The data have been scaled to respect a non disclosure agreement, but the shape of the hourly demand plot is true. Figure 10 shows one week of the data. Periodicity in the data and seasonal effects due to the weekly economic cycle are clearly visible. Also, as seen in figure 11, some days like Mondays, Saturdays, and Sundays
Figure 9. Closed-loop example with errors in the forecasted steps in hourly demand.

Figure 10. Scaled number of incoming calls at the call center.

have very distinctive profiles. On the other hand, although the load on a particular day of the week, has a distinctive profile, there are some variations from one week to another. Figure 12 shows the mean load with a solid line and above and below the mean plus and minus the variance for Mondays. It can be seen that the series is heteroscedastic and that the variance is higher for central hours of the day when the load is also higher.

Finally, figure 13 shows an example of staffing with real data obtained from a 12 weeks series of hourly demand from the same call centre. Service of one of the subsidiaries of Spanish telecom Telefónica, SA. The data have been scaled to make it comparable with the previous simulations. It has been considered an error in the demand forecast of ±4% of the average monthly hourly demand, and even with this error, the service levels are met with low idle operators. Note how the shape of the demand variations is quite repetitive and seasonal, thus it can be predicted without great errors.
Figure 11. Average load for each day of the week. Each curve contains values of the load averaged for each hour of the day using data from the same day of the week.

Figure 12. Average load for Mondays (solid line) and interval given by the mean minus and plus the variance.

7. Conclusions and future work

In this paper the authors have shown how call centre staffing can be posed as a feedback control problem. This has the advantage of getting a higher level of automation in the process of call centre staffing reducing the need of heuristics and also allows the use of a wealth of results from control theory. The working procedures of call centres and usual procedures for the staffing have been described. Also a realistic simulator of a call centre has been presented. A feedback controller with feed-forward compensation has been used to compute the staffing applied to the call centre simulator. It has been shown that good results can be obtained even with a not very sophisticated controller.

This paper only presents preliminary results obtained around a new control problem. A lot of work is to be done, and many aspects must be researched. For example, more advanced controllers should be applied, like the inverse adaptive controller. Also Model Predictive techniques can be very useful due to its predicting capability. These controllers can also take into account operating constraints much like the static optimizer often used but with the advantage of being a feedback controller. They can also take into account other issues like the fact that the number of assigned operators is an integer sequence. Other tasks like an analysis of the dynamics of the controller
Figure 13. Closed-loop example with hourly demand taken from a real call centre.

plus call centre are to be done.

References


REFERENCES


