Innovative Applications of O.R.

On the estimation of the true demand in call centers with redials and reconnects

S. Ding [a,b,∗], G. Koole [b], R. D. van der Mei [a,b]

[a] *Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands*
[b] *VU University Amsterdam, The Netherlands*

ABSTRACT

In practice, in many call centers customers often perform *redials* (i.e., reattempt after an abandonment) and *reconnects* (i.e., reattempt after an answered call). In the literature, call center models usually do not cover these features, while real data analysis and simulation results show ignoring them inevitably leads to inaccurate estimation of the total inbound volume. Therefore, in this paper we propose a performance model that includes both features. In our model, the total volume consists of three types of calls: (1) *fresh calls* (i.e., initial call attempts), (2) *redials*, and (3) *reconnects*. In practice, the total volume is used to make forecasts, while according to the simulation results, this could lead to high forecast errors, and subsequently wrong staffing decisions. However, most of the call center data sets do not have customer-identity information, which makes it difficult to identify how many calls are fresh and what fractions of the calls are redials and reconnects.

Motivated by this, we propose a model to estimate the number of fresh calls, and the redial and reconnect probabilities, using real call center data that has no customer-identity information. We show that these three variables cannot be estimated simultaneously. However, it is empirically shown that if one variable is given, the other two variables can be estimated accurately with relatively small bias. We show that our estimations of redial and reconnect probabilities and the number of fresh calls are close to the real ones, both via real data analysis and simulation.

© 2015 Elsevier B.V. and Association of European Operational Research Societies (EURO) within the International Federation of Operational Research Societies (IFORS). All rights reserved.

1. Introduction

In an inbound call center, a manager typically uses historical call data sets to forecast the future call volumes. Based on the call volume forecast, one can make staffing decisions. An inaccurate forecast inevitably leads to inaccurate staffing decisions (see Steckley, Henderson, & Mehrotra, 2010). There is extensive literature on different forecasting methods applied to call centers. Andrews and Cunningham (1995) used the Autoregressive Integrated Moving Average (ARIMA) method to forecast the inbound call volume of the L. L. Bean's call center. Taylor (2012) adjusted the traditional Holt-Winters exponential smoothing method to the Poisson count model with gamma-distributed arrival rate, and took both intraweek and intraday patterns into account in his model. Taylor (2008) compared the accuracy of a few forecasting models for a British retail bank call center. He concluded that for forecasting horizons up to two or three days ahead, seasonal ARIMA and Holt-Winters model are more accurate, while for longer lead times, simple historical average is more accurate. Shen and Huang (2008) used the Singular Value Decomposition (SVD) method to reduce the dimension of square-root-transformed call center data. Then they applied time series and regression analysis techniques to make distributional forecasts. Besides the forecasts, they also developed a method to dynamically update the forecasts when early realizations of the day are given. The doubly stochastic model built by Jongbloed and Koole (2001) addresses the issue of high variability in call arrival volume. This model was then further developed in Avramidis, Deslauriers, and L'Ecuyer (2004), where three variants of doubly stochastic model were analyzed and compared. Ibrahim and L'Ecuyer (2013) added the correlation between different call types into a model with additive seasonality, interday correlation and intraday correlation. A multiplicative way to model the intraweek and intraday pattern was used by Gans, Shen, Zhou, Korolev, McCord, and Ristock (2009).

Call center forecasting models aim to achieve the minimum error in the forecasts, where total inbound volumes are used. In this paper, we show that the true inbound volume (we refer to it as the *fresh*

---

volume from now on) is more appropriate to be used when one makes forecasts, since it is independent of the service levels, the number of agents and other factors in the call center. In contrast, the total inbound volumes are influenced by the service levels and staffing decisions of the call centers, due to the redial and reconnect customer behaviors. We define a *redial* as a reattempt of the abandoned calls, and reconnect as re-attempt of answered calls. Data analysis of a real call center reveals that a significant fraction of the inbound call volume involves redials and reconnects. The reason for customers to redial is clear, since abandoned customers did not get their questions answered in their initial attempts. There are several reasons for customers to reconnect. For example, a customer may check what is the status of his previous request. Also, solutions offered by agents may not be effective for customers, hence, they may reconnect. Koole (2013) gives more insights on redials and reconnects.

To identify the fresh volume, one would need customer-identity information in the data set, such that redials and reconnects can be filtered out. However, in most of the call center data sets, customer-identity information is either not recorded or not accessible, i.e., we do not know who is the caller of each call. In other words, we do not know whether a call is a fresh call, a redial, or a reconnect. Furthermore, the fresh volume is not stable due to the existence of seasonality and trend. On the other hand, the redial and reconnect probabilities are more stable over time, since they represent the customer behaviors. In this paper, we will show how to estimate the number of fresh calls with the assistance of the redial and reconnect probabilities.

Besides the fact that estimating redial and reconnect probabilities is crucial in estimating the fresh volume, estimating both probabilities themselves is also interesting. Much scientific effort has been spent on analyzing the performance of queueing systems with retrial behaviors (see Artalejo and Pozo, 2002; Falin, 1995 and the references therein). Some retrial models are developed for call centers, e.g., Stolletz (2008), Mandelbaum, Massey, Reiman, Stolyar, and Rider (2002), Aguir, Karaesmen, Akşin, and Chauvet (2004), Aguir, Akşin, Karaesmen, and Dallery (2008). The reconnect behavior (also referred as feedback or re-entrant in some papers) in service industry has been studied by Yom-Tov and Mandelbaum (2014). They consider a queueing model to represent hospitals where patients might return to service several times, and they apply fluid and diffusion approximations to develop some staffing principles to support healthcare staffing. In Yom-Tov and Mandelbaum (2014), customer abandonment is not included in their model. In all the existing works mentioned above, it is assumed that the retrial/reconnect probability is known, whereas it can be difficult to calculate in practice.

Hoffman and Harris (1986) are the first ones who address the issue that the total volume does not represent the true demand in call centers. Aiming to have a more accurate forecast for the call volume, they estimate the redial probability for the U.S. tax-payer service telephone center. However, Hoffman and Harris (1986) only consider the redial behavior, and they neglect the reconnect behavior. Also, the fresh call arrival rate is assumed to be a constant among certain hours of the day in their model, whereas in most call centers the arrival rate is far from constant over the day, exhibiting certain intraday pattern, see (Gans et al., 2009; Ibrahim & L'Ecuyer, 2013; Shen & Huang, 2008). In this paper, we propose a queueing model that has two extra orbits than the Erlang C model, where abandoned customers redial via one orbit, and answered customers reconnect via the other orbit. We show that these two extra orbits cannot be ignored, otherwise it will lead to inaccurate estimation of the total arrival volume, and thus in accurate staffing decisions. Having developed and validated the queueing model, we then estimate the fresh volume, the redial and reconnect probabilities. This estimation problem is formulated as an optimization problem, where the minimum objective value is attained when the actual redial and reconnect probabilities are chosen. We show that these three variables cannot be accurately estimated simultaneously. Nevertheless, if one variable is given, it is verified
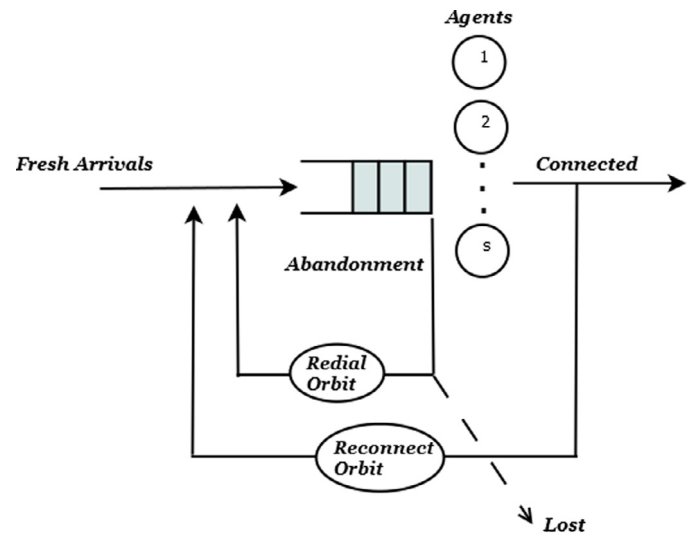

Fig. 1. Queueing diagram of a call center.

numerically that the other two variables can be estimated accurately with small relative bias. To allow intraweek seasonality, we adjust our model to a linear programming problem, which is easy to solve. We show both via simulated data and real call center data that our estimations are close to the real values. A shorter version of this paper has appeared in Ding, Koole, and van der Mei (2013).

The remainder of the paper is organized as follows. In Section 2, we describe the queueing model. We also show simulation examples of such a model to understand the influence of redials and reconnects on the total volume, as well as the necessity of distinguishing the fresh calls, redials and reconnects. In Section 3, we present our estimation models both for constant arrival rate and arrival rate with intraweek seasonal patterns. These estimation models are also validated via simulation as well as real call center data sets.

2. Model description

Consider the queueing model illustrated in Fig. 1. We assume that calls arrive according to a Poisson process. We refer to these calls as *fresh calls*. There are s agents who handle inbound calls. An arriving call will be handled by an available agent, if there is any. Otherwise, it will wait in a queue with infinite buffer size. The calls are handled in the order of arrival. After a generally distributed amount of time H, a waiting customer who did not get connected to an agent will lose his patience and abandon. We assume $\mathbb{E}H = \theta < \infty$. With probability p, an abandoned customer will enter the redial orbit, and he will redial after some generally distributed amount of time Γ_{RD}, with $\mathbb{E}\Gamma_{RD} = \delta_{RD} < \infty$. We refer to these calls as *redials*. With probability $1 - p$, this customer will not call back, and this call is considered as a 'lost' call. We assume that the service time B of a customer has a general distribution with mean $\mathbb{E}B = 1/\mu < \infty$. After the call has been finished, this customer will enter the reconnect orbit with probability q, and he will reconnect after some generally distributed time Γ_{RC}, with $\mathbb{E}\Gamma_{RC} = \delta_{RC} < \infty$. We refer to these calls as *reconnects*. We assume that p and q do not depend on customers' experiences in the system. These experiences include holding time, waiting time and the number of times that customers have already called. We use this queueing model to represent the situation of a single-skill call center.

According to the model description, the total volume is influenced by the service level in the call center, since a bad service level leads to more abandonments, which in turn leads to a larger number of redials. In this way, the total call volumes depend on the staffing decisions. To illustrate this, consider the following example, illustrated by Fig. 1. The fresh arrival rate is set to be 10 calls per minute every day,
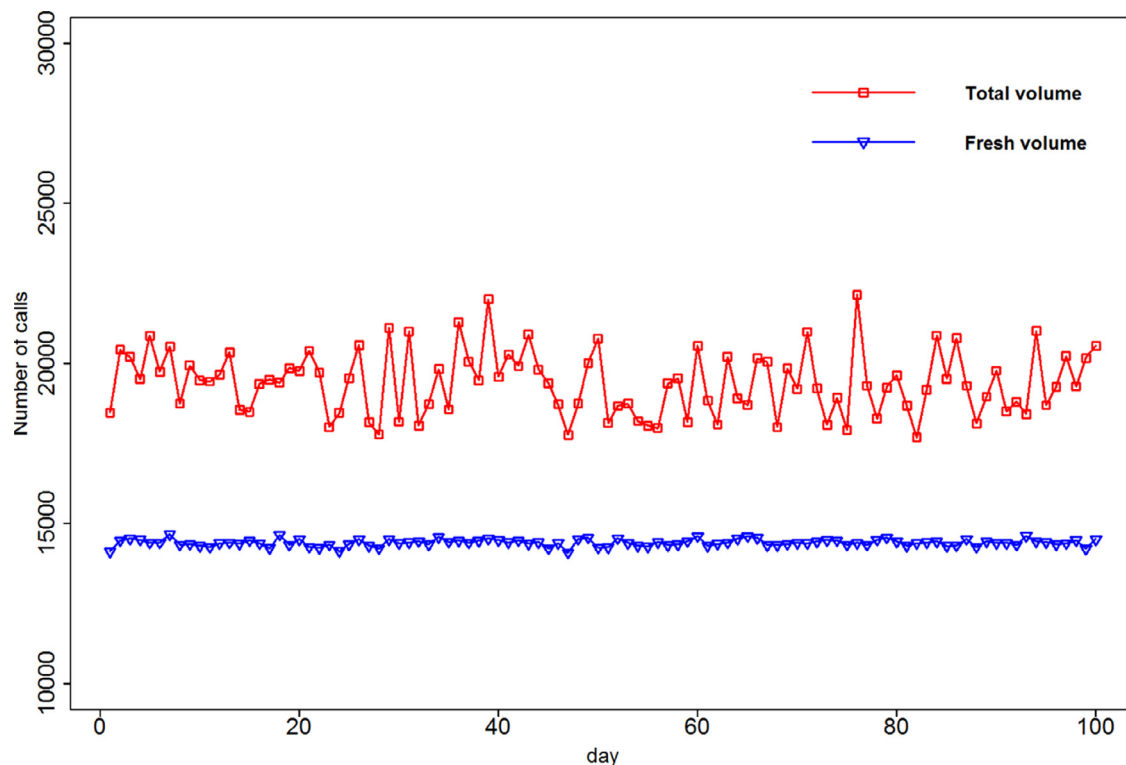
Fig. 2. Call volumes from a simulation example.

and the mean service duration is set to be 4 minutes. Since the sum of independent Poisson random variables is again Poisson distributed, F_i (i.e., the fresh volume in day i) is then Poisson distributed with rate $10 \cdot 60 \cdot 24 = 14400$. We take B, H, Γ_{RD} and Γ_{RC} to be exponentially distributed. The total call volume and fresh call volume of each day are plotted in Fig. 2 for a 100-day time interval. In this example, we set $p = 0.5$ and $q = 0.2$. The number of agents varies per day, and is drawn from a Poisson distribution with mean 43, which is slightly above the fresh arrival load per time unit, i.e., fresh arrival rate times the mean service duration. To conduct this simulation, one does not need to assume the number of agent being Poisson distributed; we make this assumption merely to model the fact that the staffing level changes each day in call centers, which is caused by several reasons, such as call centers having different shifts for different weekdays and agents' absenteeism. In the simulation, we generate a call center data set of 100 days.

Interestingly, Fig. 2 shows that not only the total volumes are much higher than the fresh volumes (as they should), but also that they exhibit much higher variability. If a manager were able to see the fresh volume, it would be easy to predict future call volumes, since they are just Poisson realizations with constant rate. However, since the manager cannot identify who the caller is in the data set, he will only see the total volume in the data set. Fig. 2 shows that the total volume depends on the staffing decisions and thus is highly volatile, which is due to the fact that both the number of redials and reconnects are influenced by the staffing decisions. In contrast, the fresh volume has less variability and is independent of the staffing decisions, which should be used to make forecasts.

In practice, managers usually use the total arrival counts to make forecasts and staffing decisions. Merely having less volatility may not be sufficient to convince them to use the fresh volumes. Naturally, following questions may arise: 'Why is it important to distinguish between the fresh calls, redials and reconnects?', 'Is knowing the total volume not enough?', 'Are there more reasons to use the fresh volume besides being less volatile?'

To address these questions, we simulate two call centers, called CC1 and CC2, whose parameters are shown in Table 1. CC1 is represents the case *without* redials and reconnect, while CC2 is its counterpart *with* redials and reconnects, constructed such that the total call volumes are the same. Both call centers have the same service level requirement, with SL ≤ 80 percent and r ≤ 10 percent, where SL is the proportion of answered customers that waited less than 20 seconds, and r is the abandonment percentage. In CC1, we let $p = q = 0$, and it receives 5669 fresh calls in a day. With $s = 40$, it achieves SL = 80.1 percent and $r = 9.6$ percent. In CC2, we let $p = 0.5$ and $q = 0.2$, and the fresh arrival rate per day to be 3700. With $s = 20$, it receives also 5669 total calls and achieves SL = 21.9 percent and $r = 4.9$ percent, which is far from achieving the service level requirement. Other parameters such as μ and θ are identical in both call centers. Assume that the manager in CC2 wants to add as little agents as possible such that CC2 reaches its service level requirement, which is very close to SL and r in CC1 in this case. Without making a distinction between the fresh calls, redials and reconnects, the manager uses the total volume to make forecasts. We assume that he simply uses the previous observation as the forecasts for the next day, which is also 5669. Consequently, to achieve the same SL and r as in CC1, the manager in CC2 derives that $s = 40$, since all parameters besides p and q are the same for CC1 and CC2. In the third row of Table 1, one can see that the SL is far beyond the 80 percent and r is far less than 10 percent in CC2 by letting $s = 40$. This means that staffing 40 agents for CC2 causes overstaffing and hence generates unnecessary staffing costs. Moreover, one can see that the realization for the next day is 4628, which is far from the original forecasts. The large forecasting error in this example is caused by not distinguishing the fresh volume from the total volume rather than using the wrong forecasting method. We could easily construct other examples to show that for other forecasting methods, such as ARIMA or exponential smoothing, large forecasting error may still exist. This means that not differentiating the fresh volume from the total volume can lead to large error in forecasts as well as in staffing decisions. In summary, this example emphasizes

Table 1
Two simulation results, $\mu = 1/10$, $\theta = 1/2$, $\delta_{RD} = 20$, $\delta_{RC} = 50$.

	p	q	F	s	Total arrival	F	SL	r
CC1	0	0	4	40	5669	5669	80.1 percent	9.6 percent
CC2	0.5	0.2	2.6	20	5669	3705	21.9 percent	49.0 percent
CC2	0.5	0.2	2.6	40	4628	3689	94.8 percent	0.7 percent

the necessities of knowing the fresh volumes, as well as using it in making operational decisions in call centers. When we say that the fresh volume represents the true demand, it is actually a subjective claim. People could also choose to claim that the total volume represents the true demand. However, this would make the demand being more complicated rather than being a simple number. For example, if one uses the total volume to represent the demand, and assume that the demand is mentioned to be 100, then the number 100 is ambiguous, i.e., is the total volume being 100 obtained by staffing 20 agents or 40 agents; is the SL low or high when we receive 100 calls; when we have different service times, would the demand still be 100? These complications or questions will not arise if we use the fresh volume to represent the demand.

3. Estimation model

Many call center data sets are similar to the simulated data set we generated in the previous section: customer identity information is not available. Thus, in such call center data sets for N days, we would only know A_i and $C_i \sim (i = 1, \ldots, N)$, which stands for the number of abandoned calls in day i and the number of connected calls in day i, respectively. We denote T_i as the total number of calls in day i, and $r_i := A_i/T_i$, which is the abandonment percentage of day i.

To estimate F_i, p and q, we start with the simple case where $F_i \sim \text{Pois}(F^*)$, i.e., each day has the same arrival rate of fresh calls, but we do not know how big F^* is. Note that, by this assumption, we ignore the intraweek arrival pattern in the call center data set. We will extend our model to address this pattern in Section 3.2. For the rest of this paper, we refer to $\hat{p}$ and $\hat{q}$ as estimated values of p and q by using our model, respectively, and p^* and q^* as the true values of p and q, respectively.

3.1. Basic setup

By definition, we know that an inbound call can either be a fresh call, a redial or a reconnect. Hence, the following equation holds

$$T_i = F_i + RC_i + RD_i, \tag{1}$$

where RD_i and RC_i are the number of redials and reconnects in day i, respectively. $RD_i \sim B(A_i, p^*)$, $RC_i \sim B(C_i, q^*)$, where $B(k, p)$ stands for the binomial distribution with parameters k and p. If we let $F_i = F^* + \varepsilon_i$, $RD_i := A_i p^* + e_i$, and $RC_i := C_i q^* + \eta_i$, then Eq. (1) can be rewritten as

$$T_i = F^* + A_i p^* + C_i q^* + \epsilon_i + e_i + \eta_i. \tag{2}$$

Also, since a call is either answered or abandoned, we know that $T_i = A_i + C_i$. Insert this equation into Eq. (2), we obtain

$$(1 - p^*)A_i + (1 - q^*)C_i - F^* = \epsilon_i + e_i + \eta_i. \tag{3}$$

For given data points $A_1, \ldots, A_N$, and $C_1, \ldots, C_N$, we consider the following minimization problem to estimate p, q and F,

$$(\hat{p}, \hat{q}, \hat{F}) = \underset{0 \leq p, q < 1}{\text{argmin}} \sum_{i=1}^{N} |(1 - p)A_i + (1 - q)C_i - F|, \tag{4}$$

where the objective function is the sum of absolute deviance (SAD). Note that in problem (4), we use SAD as the estimation error measurement rather than using other error measurements, such as the sum of squared errors. There are two reasons for this. The first reason is that for call center models, the forecast that minimizes the absolute errors will also minimize the error in number of agents (see Ding & Koole, 2014). Another reason is that minimizing the SAD is more robust against outliers, comparing to minimizing the sum squared errors. In Section 3.5, we empirically verify this claim using two real call center data sets.

In fact, the errors measured by SAD are scaled errors, in the sense that if we choose large numbers for $\hat{p}$ and $\hat{q}$, the error would be smaller. An extreme example that indicates this scaling problem is letting $\hat{p} = 1$ and $\hat{q} = 1$, and SAD would always be 0 by choosing $\hat{F} = 0$. Therefore, we introduce the following minimization problem, which uses the Weighted Absolute Percentage Error (WAPE) instead of SAD as the objective function to remove this scaling problem,

$$(\hat{p}, \hat{q}, \hat{F}) = \underset{0 \leq p, q < 1, F}{\text{argmin}} \frac{\sum_{i=1}^{N} |(1 - p)A_i + (1 - q)C_i - F|}{\sum_{i=1}^{N} ((1 - p)A_i + (1 - q)C_i)}. \tag{5}$$

One can notice that we choose term $\sum_{i=1}^{N} ((1 - p)A_i + (1 - q)C_i)$ as the denominator of WAPE rather than the term $\sum_{i=1}^{N} F_i$. This is for computational purposes. Because p and q are always bounded between 0 and 1, we can calculate the minimum WAPE on a grid of p and q ranging from 0 to 1. In contrast, we have no information on how big F is, which makes it more difficult to find the minimum WAPE if we let $\sum_{i=1}^{N} F$ to be the denominator.

Above, we have shown a regression method for estimating (p^*, q^*, F^*). However, one can notice that we have three degrees of freedom (namely, p, q and F), while only observations for A_is and C_is are being made. This means that in a call center data set without customer identity information, we cannot estimate (p^*, q^*, F^*) simultaneously, and one parameter needs to be given before any regression method can be applied.

In a call center, there are different ways to estimate the reconnect probability. For example, the manager can ask agents to do some polling (e.g. for one whole day), we staff enough agents, so that almost all calls are handled, and we ask each agent to record each connected call's customer name or identity, then by the end of the day, we can calculate how many customers have called back. For the redial probability, this is more difficult to do, since abandoned customer's information is often not recorded. Using polling to determine the number of fresh calls is also difficult. Because the number of fresh calls is not stable over time, due to presence of trend and seasonality (see Ibrahim & L'Ecuyer, 2013; Shen & Huang, 2008).

Assuming $q = q^*$, we present an algorithm to numerically compute $(\hat{p}, \hat{F})$.

Algorithm 1.

Step 0: Let $p = 0$, WAPE $= 1$, and let the grid size to be ξ.
Step 1: Calculate $L_i = (1 - p)A_i + (1 - q^*)C_i$, for all $i = 1, \ldots, N$, and $F = \text{median}(L_1, L_2, \ldots, L_N)$, $a = \frac{\sum_i |L_i - F|}{\sum_i L_i}$.
Step 2: If $a <$ WAPE, then let WAPE $= a$, $\hat{p} = p$, $\hat{F} = F$.
Step 3: If $p \geq 1$, then stop; else, $p = p + \xi$, go to step 1.

In this estimation model, we only consider redial and reconnect in the same day of the fresh call. We will motivate this assumption when we analyze the redial and reconnect behaviors from a real call center data set.

3.2. Intraweek seasonality

In model (5), we made the assumption that each day has the same fresh call arrival rate. Often this is an unrealistic assumption in a real call center. We will show in Section 3.4 that for two real call center data sets, both the total volume and the fresh volume show strong intraweek patterns. Thus, to make our model applicable in call center data with intraweek seasonality, we make adjustments to estimation model (5). To this end, we assume that the weekly total fresh calls distributed to each day of the week in a multiplicative way, i.e.,

$$\mathbb{E}F_i = \mathbb{E}WF_{w_i} \cdot \beta_{d_i},$$

where w_i and d_i are the week number of day i and the weekday of day i, respectively, $d_i \in \{1, 2, 3, 4, 5\}$ (since we ignore the weekends), $w_i = 1, 2, \ldots n$, where n stands for the number of weeks. WF_{w_i} is a random variable that stands for the total number of fresh calls of week w_i. Thus, β_{d_i} can be interpreted as the proportion of calls on weekday d_i than in the whole week. A key assumption of this multiplicative model is that β_{d_i} does not depend on the week number. Such a multiplicative model has been applied in several call center forecasting models (see Brown, Gans, Mandelbaum, Sakov, Shen, Zeltyn, & Zhao, 2005; Gans et al., 2009; Weinberg, Brown, & Stroud, 2007). Therefore, our estimation model changes to

$$(\hat{p}, \hat{q}, \hat{\beta}_{d_i}) = \underset{0 \leq p,q,\beta_{d_i} < 1}{\operatorname{argmin}} \frac{\sum_{i=1}^{N} |(1-p)A_i + (1-q)C_i - WF'_{w_i} \cdot \beta_{d_i}|}{\sum_{i=1}^{N} ((1-p)A_i + (1-q)C_i)}$$

s.t. $\sum_{j=1}^{5} \beta_j = 1,$

$$WF'_{w_i} = WA_{w_i}(1-p) + WC_{w_i}(1-q), \quad w_i = 1, 2, \ldots n, \quad (6)$$

where WF'_{w_i} is the estimated number of fresh calls in week w_i, WA_{w_i} and WC_{w_i} are the total number of abandoned calls and total number of connected calls in week w_i, respectively. Since A_i and C_i are given observations, we can easily obtain their aggregated weekly volumes WA_{w_i} and WC_{w_i}. The intuition behind model (6) is that the daily fresh call volume is proportional to the weekly total fresh call volume. Once we have $(\hat{p}, \hat{q}, \hat{\beta}_{d_i})$, $\hat{F}_i$ can be obtained via

$$\hat{F}_i = (WA_{w_i}(1-\hat{p}) + WC_{w_i}(1-\hat{q})) \cdot \hat{\beta}_{d_i}, \quad i = 1, 2, \ldots, N.$$

Similar to the approach we took for solving (5), we assume $q = q^*$, and we solve (6) on a grid of p. Assuming $q = q^*$, and for a given value of p, problem (6) is a quantile regression problem with a linear constraint, which is equivalent to a linear programming (LP) problem. The corresponding LP problem of model (6) can be written as

$$\min_{\beta_{d_i}, Z_i^+, Z_i^-} \sum_{i=1}^{N} (Z_i^+ + Z_i^-)$$

s.t. $\sum_{i=1}^{N} \beta_{d_i} = 1,$

$$Z_i^+ - Z_i^- = \frac{(1-p)A_i + (1-q)C_i - WF'_{w_i} \cdot \beta_{d_i}}{\sum_{i=1}^{N} ((1-p)A_i + (1-q)C_i)}, \quad i = 1, \ldots, N,$$

$$WF'_{w_i} = WA_{w_i}(1-p) + WC_{w_i}(1-q), \quad i = 1, \ldots, N,$$

$$0 \leq p, q, \beta \leq 1$$

$$Z_i^+, Z_i^- \geq 0, \quad i = 1, \ldots, N. \tag{7}$$

In fact, problems (6) and (7) are equivalent (see Charnes, Cooper, & Ferguson, 1955). We now give the idea the proof, the equivalence of (6) and (7) holds if

$$Z_i^+ + Z_i^- = \frac{|(1-p)A_i + (1-q)C_i - WF'_{w_i} \cdot \beta_{d_i}|}{\sum_{i=1}^{N} ((1-p)A_i + (1-q)C_i)}, \quad i = 1, \ldots, N,$$

and it is suffice to show that at least one of the values Z_i^+ and Z_i^- is zero in the optimal solution; otherwise, assume Z_i^+ and Z_i^- are both

Table 2
Parameters of the simulation experiments.

Example	p	q	F^*	$1/\mu$ min	θ min	δ_{RD} min	δ_{RC} min
1	0.5	0.2	10	4	2	5	10
2	0.5	0.2	10	4	2	15	30
3	0.5	0.2	4	10	2	20	50
4	0.7	0.3	4	9	3	15	50
5	0.7	0	4	9	3	10	NA

non-zero, then one could find a solution which has a smaller objective value by substracting $\min\{Z_i^+, Z_i^-\}$ from Z_i^+ and Z_i^-.

Assuming $q = q^*$, we show the following algorithm to numerically obtain $(\hat{p}, \hat{F})$.

Algorithm 2.

Step 0: Let $p = 0$, WAPE = 1, and let the grid size to be ξ.
Step 1: Calculate $L_i = (1-p)A_i + (1-q^*)C_i$, for all $i = 1, \ldots, N$, and $WF_{w_i} = (1-p)WA_{w_i} + (1-q^*)WC_{w_i}$, for all $w_i = 1, \ldots, n$.
Step 2: Solve LP problem (7) for given p and $q = q^*$.
Step 3: Let a be the objective value to the optimal solution, and b_{d_i} be the optimal value for decision variable β_{d_i}, $i = 1, 2, \ldots, N$.
Step 4: If $a <$ WAPE, then WAPE $= a$, $\hat{p} = p$, $\hat{F}_i = WF_{w_i} \cdot b_{d_i}$, for all $i = 1, 2, \ldots, N$.
Step 5: If $p \geq 1$, then stop; else, $p = p + \xi$, go to step 1.

In this paper, we use *linp* function in package *limSolve* in R to solve the LP problem in Step 2. When choose how large ξ is, one should bear in mind that when the grid size ξ is big, the precision will be low; when ξ decreases, the computation time will increase. In this paper, we set the grid size to be 0.01, and for such a grid size, the computation time is small even for $N = 500$, i.e., less than 2 minutes.

3.3. Validation using simulation

In this subsection, we test our estimation model (5) in the data sets generated by discrete-event simulation. The data generation procedure is the same as described in Section 2. Once the data have been generated, we use the model (5) for the estimation, then the estimated values are compared with simulation inputs.

Five different parameter settings are tested. These parameters are shown in Table 2, where F^* is the fresh arrival rate per minute. For each parameter setting, we also validate our estimation model for different number of days, namely for $N = 20$, $N = 50$ and $N = 100$. For a given sample size, the estimators are themselves random variables. To understand the bias and the variability of the estimators, we replicate such simulation-estimation procedure fifty times, and then calculate the sample mean and standard deviation of the estimated values. Note that larger numbers of replications lead to more accurate estimations for the means and quantiles, but it will be more computationally expensive.

We used Algorithm 1 to calculate the estimated values. The sample mean, standard deviation, 5 percent and 95 percent quantile of the estimated values are shown in Table 3.

In Table 3, the estimations for p^* and F^* are denoted as $\hat{p}|q^*$ and $\hat{F}|q^*$, respectively, and the notation "$|q^*$" stands for the fact that it is an estimator given $q = q^*$. Furthermore, we let SD to be the sample standard deviation of the estimators, and $Q_{\alpha, \hat{p}}$ and $Q_{\alpha, \hat{F}}$ stand for the sample α quantile ($\alpha = 0.05$ or 0.95) of the estimator $\hat{p}|q^*$ and $\hat{F}|q^*$, respectively. One can see from Table 3 that the differences between $\mathbb{E}(\hat{p}|q^*)$ and the p^* is less than 0.03, even for the relatively small sample size such as $N = 20$.

Furthermore, one could see from Table 3 that $\hat{F}$ is a biased estimator, which under-estimates F^*. Here we describe the reason of this bias and argue that it is relatively small compare F^*. The source of the bias mainly rooted from the fact that the median of F would minimize the

Table 3
Estimation results.

Example	N	$\mathbb{E}(\hat{p}|q^*)$	$SD(\hat{p}|q^*)$	$(Q_{0.05,\hat{p}}, Q_{0.95,\hat{p}})$	$\mathbb{E}(\hat{F}|q^*)$	$SD(\hat{F}|q^*)$	$(Q_{0.05,\hat{f}}, Q_{0.95,\hat{f}})$
	20	0.505	0.027	(0.464, 0.547)	9.960	0.096	(9.810, 10.110)
1	50	0.502	0.010	(0.485, 0.516)	9.969	0.039	(9.913, 10.032)
	100	0.501	0.006	(0.493, 0.510)	9.971	0.021	(9.937, 10.000)
	20	0.500	0.012	(0.480, 0.519)	9.990	0.049	(9.920, 10.058)
2	50	0.501	0.009	(0.488, 0.514)	9.987	0.032	(9.936, 10.033)
	100	0.501	0.005	(0.491, 0.508)	9.992	0.021	(9.961, 10.026)
	20	0.522	0.043	(0.457, 0.595)	3.942	0.070	(3.829, 4.052)
3	50	0.512	0.016	(0.482, 0.536)	3.958	0.028	(3.920, 4.015)
	100	0.506	0.010	(0.491, 0.523)	3.969	0.017	(3.946, 3.999)
	20	0.710	0.022	(0.670, 0.744)	3.930	0.080	(3.809, 4.069)
4	50	0.702	0.009	(0.688, 0.719)	3.958	0.036	(3.898, 4.018)
	100	0.702	0.006	(0.693, 0.710)	3.956	0.017	(3.935, 3.983)
	20	0.708	0.042	(0.642, 0.771)	3.990	0.039	(3.937, 4.066)
5	50	0.701	0.014	(0.680, 0.718)	3.996	0.016	(3.971, 4.018)
	100	0.702	0.009	(0.687, 0.715)	3.996	0.011	(3.980, 4.013)
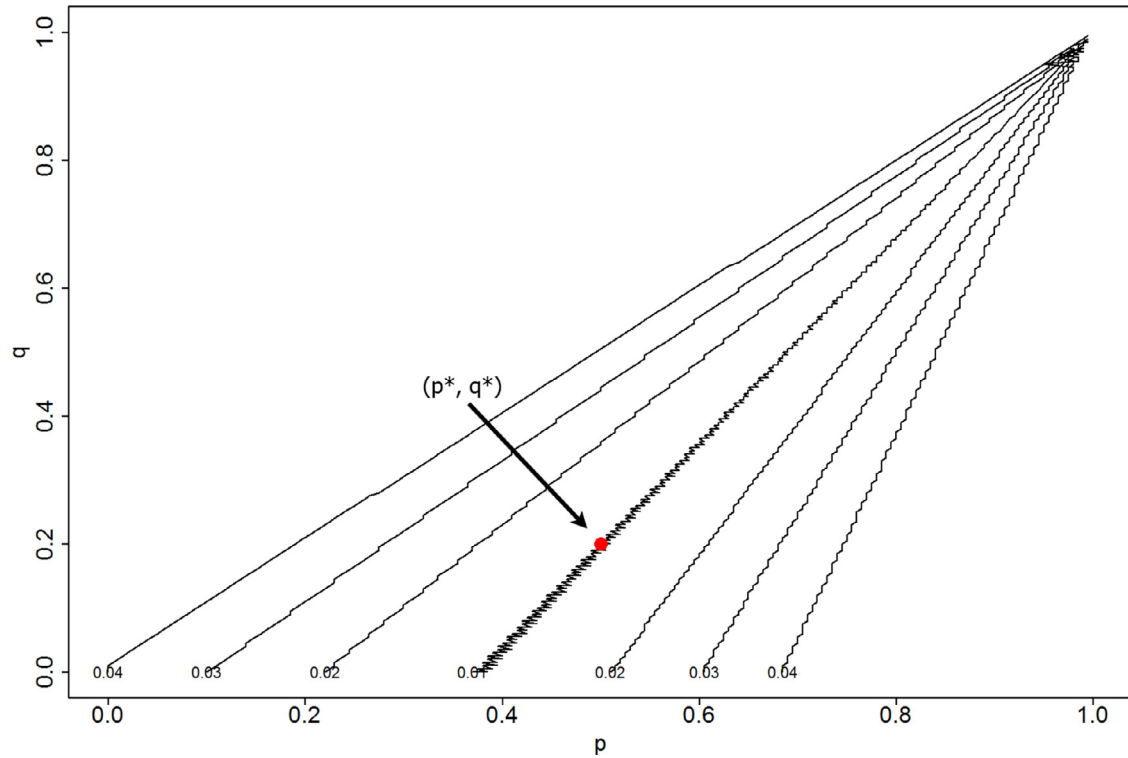
Fig. 3. Values of WAPE on the grid of p and q for simulated data example 1, with the red points standing for (p^*, q^*).

WAPE, while F^* is the mean of F. In the case that F is a Poisson random variable, the difference between its mean and median is not zero, but some small values that bounded by 1 (Chen & Rubin, 1986). However, since the estimation method uses daily aggregated volumes, the bias is relatively small comparing to F^*, as one can confirm this in the results of all examples in Table 3. To illustrate the relation between p, q and the WAPE, we plot the minimum WAPE on a grid of p and q in Fig. 3. One can see that the true parameters p^* and q^* is on the line where the minimum WAPE is attained. Other simulation examples (not shown here) gave similar graphs as in Fig. 3. This figure confirms that when $q = q^*$, the minimum WAPE leads to the accurate estimation for p. Moreover, this figure also shows how sensitive $\hat{p}$ is with respect to the choice of q. For example, if one would make a calculation error of ε for q^*, then in this example, the estimation error for p would be $\varepsilon \cdot 5/8$, since the slope of the line with minimum WAPE is 5/8 and our estimated point can only be one of the points in this line.

Note that in the simulation, we take B, H, Γ_{RC} and Γ_{RD} to be exponentially distributed. However, since A_i and C_i are realizations which can be obtained from the data, how B, H, Γ_{RC} and Γ_{RD} are distributed

becomes irrelevant for estimation Model (5). We now explain the reason. Assume for simplicity, there is only one fresh call during the day, and this call is connected. However, this customer would like to reconnect today, and this reconnect is answered again. Whether this customer calls back at 2 pm or 5 pm will not change the fact that there are two answered calls, one of them is a reconnect, and $q = 1/2$. Thus, as long as customers call back within the same day as their corresponding fresh calls, how large δ_{RC} is does not have any influence on estimation results. Consequently, we can extend our estimation model to call centers where these variables have general distributions.

3.4. Validation using real data

In this section, we analyze real call center data to understand the redial and reconnect behaviors as well as to validate our estimation model (7). The call center data is obtained from a call center company called Vanad Laboratories in the Netherlands. The data set consists of call arrival records to the municipality of Rotterdam in the
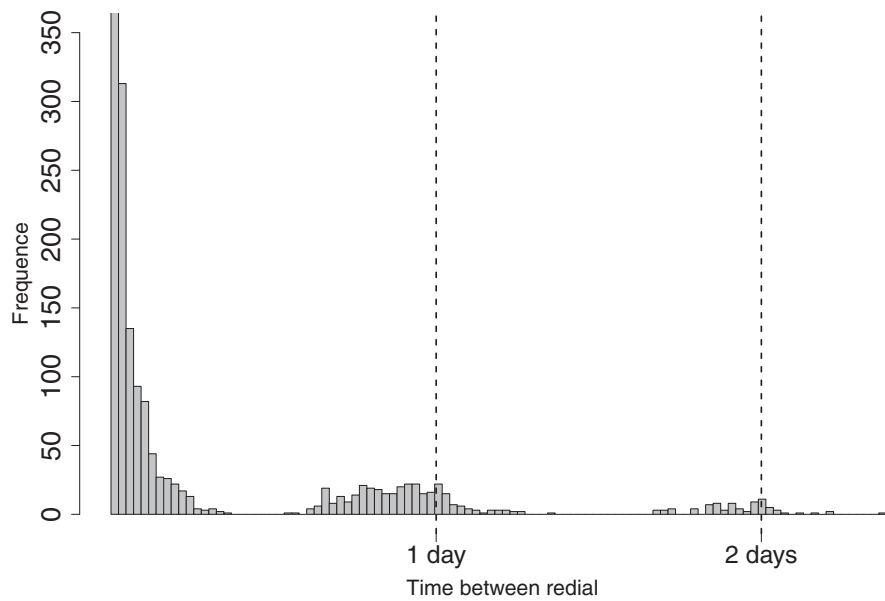
Fig. 4. The histograms of realizations of Γ_{RD} for Router A.
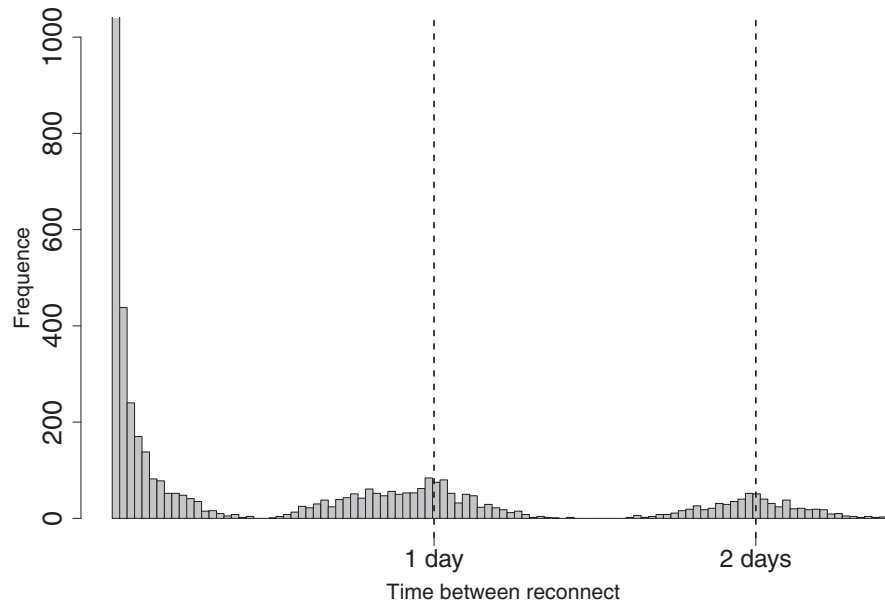

Fig. 5. The histograms of realizations of Γ_{RC} for Router B.

Netherlands. The calls are recorded from 1st April 2012 to 29th September 2012. There are in total 498508 call records during these periods. On Sundays, the call center is closed. On Saturdays, the arrival volume is quite low, i.e., 5508 total call records for 26 Saturdays. Therefore, we may ignore the weekends call data, and focus only on the weekdays. We also remove the weeks which consist of one or few days of holidays. This leaves us with 22 weeks of data. Each call record consists of seven attributes, i.e., call arrival date, arrival time, caller's phone number, router name, agent number, time that the call is answered and the time that the call is hanged up. We assume that each caller is identified by its phone number. Approximately 20 percent of the caller's phone numbers are unidentified, since some callers set their phone number to be invisible by the call receivers. There are eleven different types of routers that can be selected by a caller. The selection of router is done by customers via Interactive Voice Response (IVR) unit. After the customer has made the selection, his call will be distributed by an Automatic Call Distributor (ACDs). Each

router represents one or multiple types of questions that a customer may have. Among those eleven routers, there are four major routers, which consist of approximately 71 percent of all calls. Among those four routers, we will focus our study on two specific routers which are referred to as routers A and B. The reasons that we choose these two routers are the following; (*i*) other routers may represent multiple types of questions, and customers who have different types of questions may have different redial and reconnect behaviors; (*ii*) some routers have been merged or changed their names during the data collection periods.

For this data set, we have the caller-identity information, which allows us to follow each customer and see whether he called back or not. In Figs. 4 and 5, we plot the histograms of realizations of Γ_{RD} and Γ_{RC} for router A. For router B, we obtain similar figures. We can see that both for redial and reconnect, most of the customers call back in the same day as their fresh calls, and they call back shortly after abandonments or connected calls. A small fraction of the customers
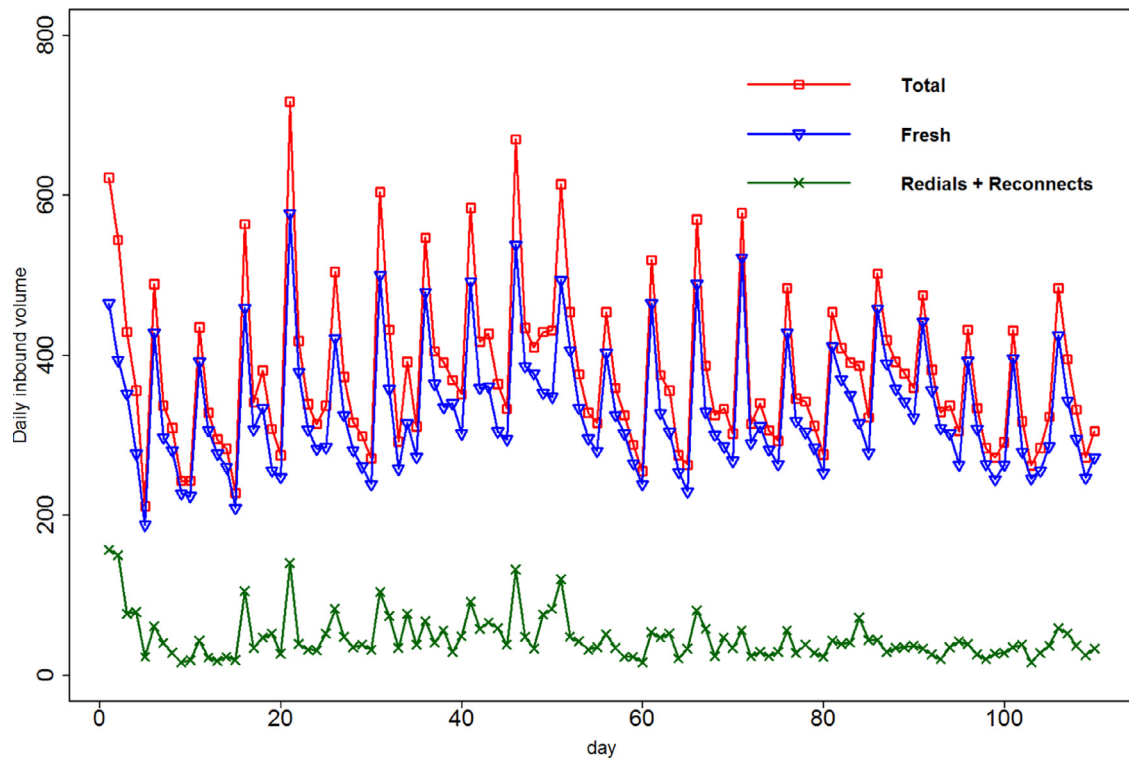
Fig. 6. The plots of the total number of calls, fresh number of calls, redials plus reconnects for Router A.

Table 4
Descriptive statistics.

Router	Total volume	Fresh volume	Redials	Reconnects
A	41,624	36,515	2142	2967
B	28,526	23,782	1117	3627

Table 5
Real redial and reconnect probabilities of each weekday for Router A.

	Mon	Tue	Wed	Thu	Fri
p^*	0.52	0.52	0.46	0.49	0.43
q^*	0.08	0.08	0.07	0.09	0.08

Table 6
Real redial and reconnect probabilities of each weekday for Router B.

	Mon	Tue	Wed	Thu	Fri
p^*	0.42	0.40	0.38	0.38	0.39
q^*	0.15	0.13	0.14	0.15	0.12

Table 7
Estimation results for both routers.

Router	p^*	q^*	$\hat{p}\|q^*$	WAPE$_F$
A	0.49	0.08	0.54	2.6 percent
B	0.40	0.14	0.49	3.0 percent

redial or reconnect one or two days later after the fresh call. Therefore, in our model, it is sufficient to assume that the redials and the reconnects arrive in the same day as the fresh call, i.e., customer who calls again one or more days later will be regarded as another fresh call. Some descriptive statistics are shown in Table 4. The total number, the fresh number and the redials and reconnects are plotted in Figs. 6 and 7. In these two graphs, the unidentified calls are removed.

In Figs. 6 and 7, one could still see high variability in the number of fresh calls in contrast to Fig. 2, where very little variability is observed in the number of fresh calls. This is because besides the redials and reconnects, another source of variability in the total volumes in Figs. 6 and 7 is the intraweek seasonality. In other words, one cannot use the redials and reconnects to explain all the variabilities in real call center data, since intraweek seasonality is also a major cause of variability. This observation would confirm the necessity to include seasonality in the estimation model.

After removing the unidentified calls, RD_i and RC_i become observations in this data set with customer identity information, and we use following formulas to calculate the actual redial and reconnect probabilities,

$$p^* = \frac{\sum_{i=1}^{N} RD_i}{\sum_{i=1}^{N} A_i}, \tag{8}$$

$$q^* = \frac{\sum_{i=1}^{N} RC_i}{\sum_{i=1}^{N} C_i}. \tag{9}$$

We also calculate the probabilities p^* and q^* of each weekday and show them in Tables 5 and 6. We see that the redial and reconnect behaviors are quite significant, i.e., the reconnect probability can reach 15 percent. This further confirms the necessity of including both orbits in the queueing model. Furthermore, we see that both probabilities are different, i.e., redial probability is usually larger than the reconnect probability. Intuitively, this makes sense, since an abandoned customer has higher urge to call back than an answered customer. For different routers, redial probability has more fluctuations than reconnect probability. However, within the same router for every weekday both probabilities are stable, only except for the redial probability for router A on Friday. Therefore, it is sufficient to have two parameters for all weekdays together for redial and reconnect probabilities of each router.
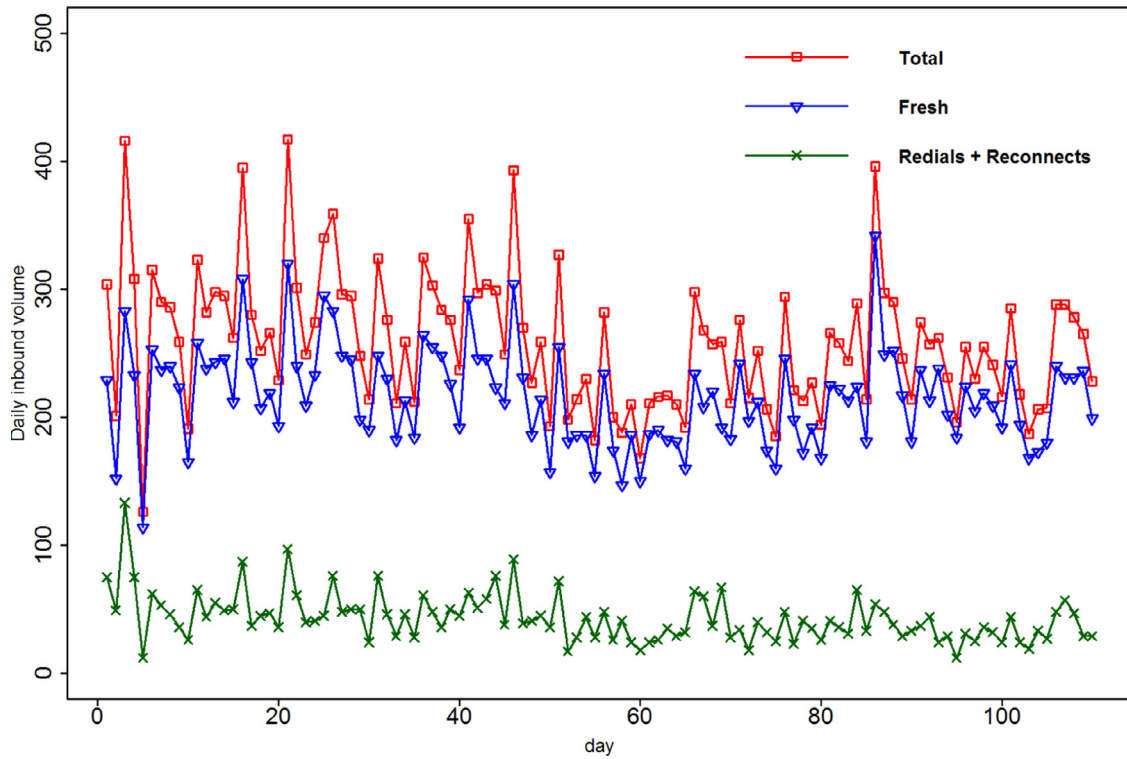
Fig. 7. The plots of the total number of calls, fresh number of calls, redials plus reconnects for Router B.
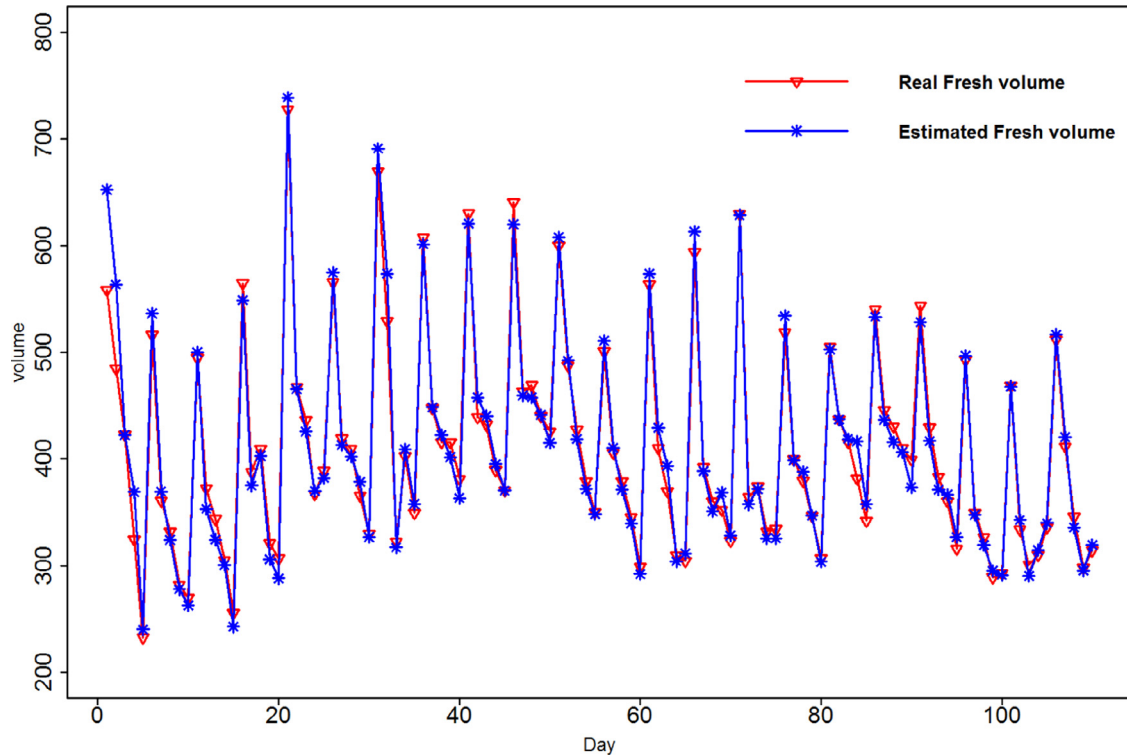

Fig. 8. Real number of fresh calls vs. estimated number of fresh calls for Router A.

We apply Algorithm 2 to the Vanad Laboratories data set (with grid size 0.01). The estimation results are shown in Table 7, where WAPE_F is used to measure the percentage difference between $\hat{F}_i$ and F_i^*, and it is defined as

$$\text{WAPE}_F := \frac{\sum_{i=1}^{N} |\hat{F}_i - F_i^*|}{\sum_{i=1}^{N} |F_i^*|},$$

One can see from Table 7 that our estimation of redial probability for router A is approximately 0.05 higher than the real redial probability, while for router B, our estimation is about 0.09 higher. For a call center with $r = 20$ percent, i.e., 20 percent of all calls are abandoned, 0.09 error in redial probability would lead to less than 2 percent error in estimating the number of fresh calls. For these two routers, which have much less abandonments than 20 percent, 0.09 would lead to
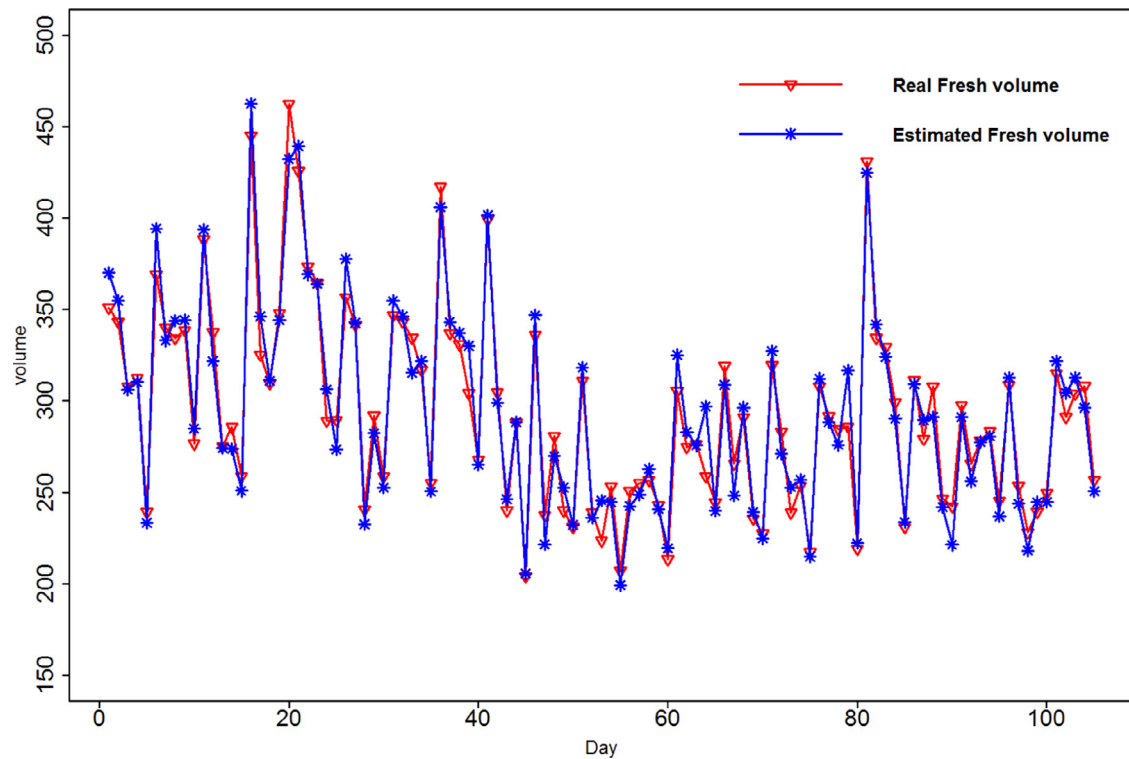
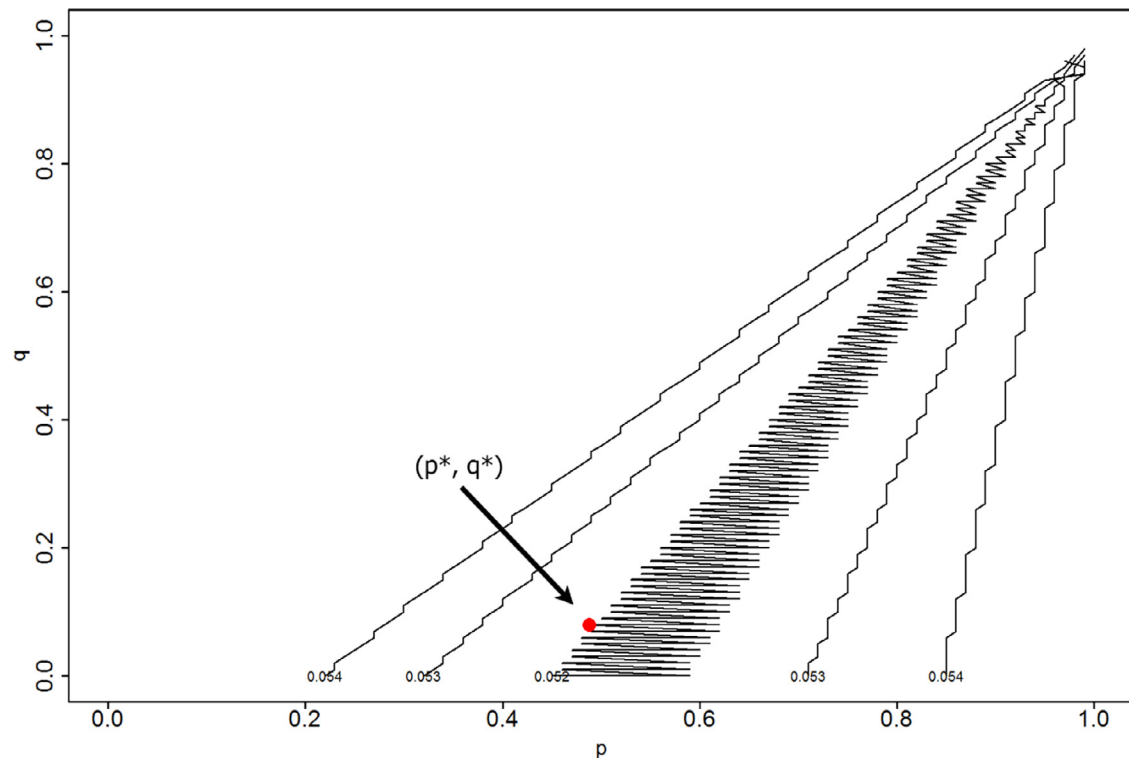**Fig. 9.** Real number of fresh calls vs. estimated number of fresh calls for Router B.

Fig. 10. Values of WAPE on the grid of p and q for Router A.

even less errors. Therefore, maximum of 0.09 error in our estimation of redial probability is acceptable. Furthermore, the $WAPE_F$ for both routers are quite small, which are both less than 3 percent. The real fresh calls and the estimated fresh calls are plotted in Figs. 8 and 9.

In both figures, our estimations $\hat{F}_i$ are quite close to the real fresh calls F_i^*.

To understand the relation between p, q and the WAPE, we plotted the minimum WAPE on a grid of p and q in Figs. 10 and 11. One can
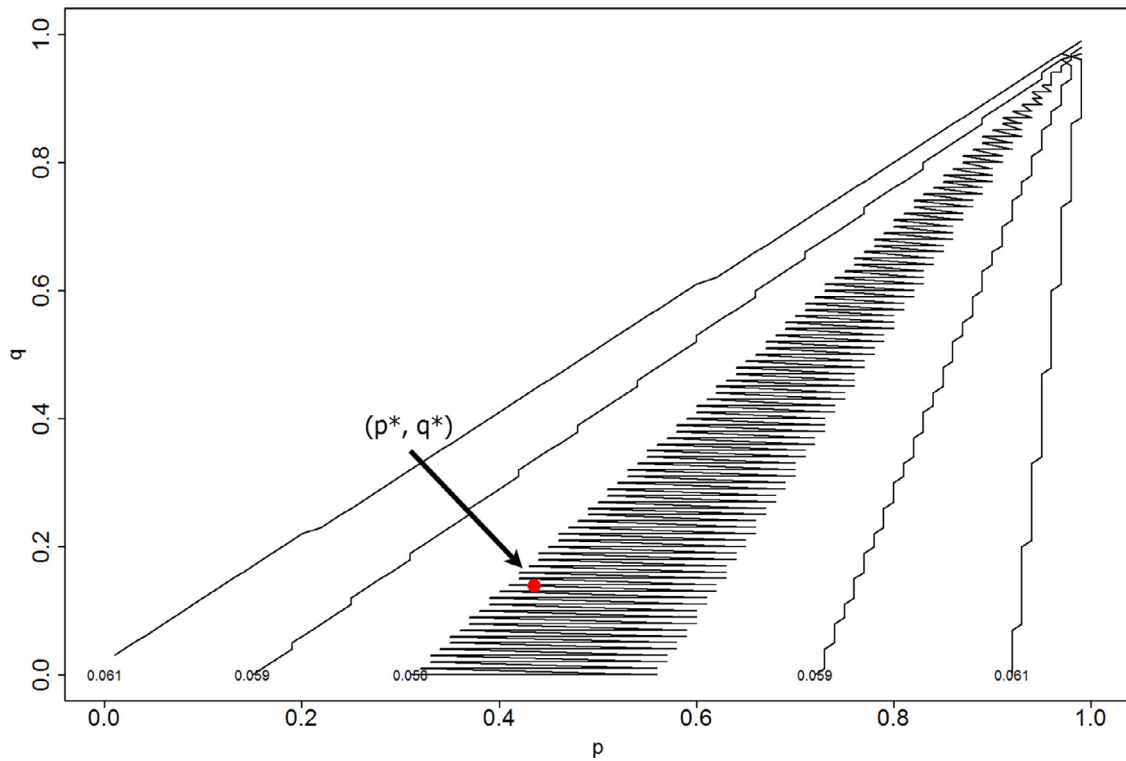
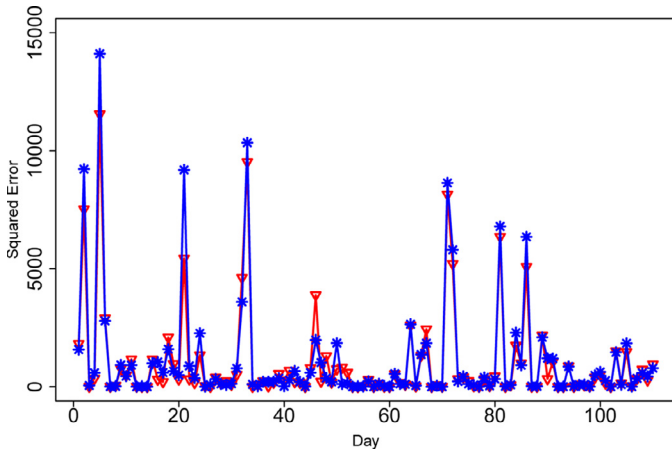Fig. 11. Values of WAPE on the grid of p and q for Router B.

Fig. 12. Squared errors of the WAPE estimator (star symbol in blue) vs. those of the OLS estimator (triangle symbol in red) Router A. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
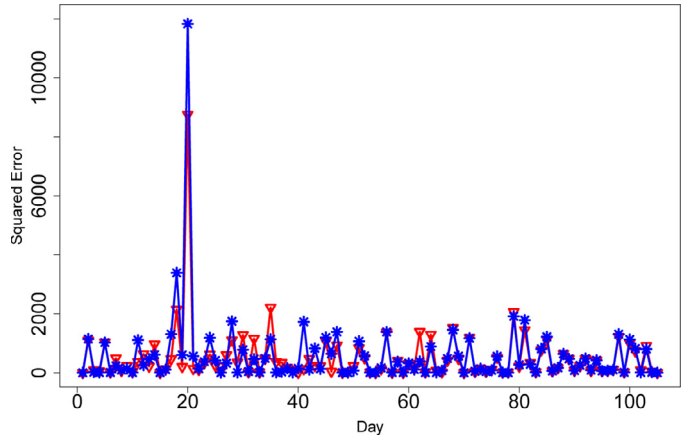

Fig. 13. Squared errors of the WAPE estimator (star symbol in blue) vs. those of the OLS estimator (triangle symbol in red) Router B. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

see that the true parameters p^* and q^* is close to the region where the minimum WAPE is attained. This also suggests that once we know q^*, the minimum WAPE will lead us to a close estimation for p.

3.5. Minimizing WAPE vs. minimizing squared errors

In this subsection, we compare the estimator that minimizes WAPE with the ordinary least squared (OLS) estimator. The WAPE or the absolute errors are more robust against outliers comparing to the squared errors (Narula, Saldiva, Andre, Elian, Ferreira, & Capelozzi, 1999). We now empirically validate this claim with our data.

Given q^*, the OLS estimator can be obtained by

$$(\hat{p}^{(OLS)}, \hat{\beta}_{d_i}^{(OLS)}) = \underset{0 \le p, \beta_{d_i} < 1}{\operatorname{argmin}} \sum_{i=1}^{N} \left((1-p)A_i + (1-q^*)C_i - WF'_{w_i} \cdot \beta_{d_i}\right)^2$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \beta_{d_i} = 1,$$

$$WF'_{w_i} = WA_{w_i}(1-p) + WC_{w_i}(1-q^*), \quad w_i = 1, 2, \ldots n. \quad (10)$$

Minimization problem (10) is a standard regression problem with linear constraints.

The fitted squared errors for routers A and B are plotted in Figs. 12 and 13. Based on these two figures, we can clearly visualize some outliers. These outliers are not holidays or being caused by any special events, thus, one could not identify them in advance. To validate the sensitivity of the OLS estimator with respect to outliers, we removed the whole week data for weeks that contain one or more days of outliers. The WAPE estimator and the OLS estimator based on data *with* and *without* outliers are shown in Tables 8 and 9. The

Table 8
Comparing different estimators for Router A.

	With outliers	Without outliers
$\hat{p}$	0.54	0.57
$\hat{p}^{(OLS)}$	0.70	0.44

Table 9
Comparing different estimators for Router B.

	With outliers	Without outliers
$\hat{p}$	0.49	0.52
$\hat{p}^{(OLS)}$	0.81	0.58

results show that with the outliers, the OLS estimator leads to larger estimation errors for both routers comparing to those for the WAPE estimator; while without the outliers, OLS estimator results in much more accurate estimation. For instance, we see in Table 9 that even a single outlier can lead to a very different estimation result for the OLS estimator. On the other hand, it can be seen that the WAPE estimator is much less sensitive to those outliers, in the sense that both with and without outliers, the WAPE estimator leads to accurate estimations. In call centers, the call volume is influenced by a lot of effects, some of which can be easily identified by date, such as holiday effects. However, not all outliers are easily identifiable in call center data. For example, in day 47 of Fig. 12, it is difficult to judge whether this day is an outlier or not. Therefore, for the advantage of being more robust against outliers, we prefer the WAPE estimator to the OLS estimator.

4. Conclusion and further research

In this paper, we propose a queueing model of call centers with redials and reconnects. We use simulation results as well as real data analysis results to show that both features are significant and should not be ignored. We claim that it is more convenient to let the fresh volume to represent the demand in call center in the sense that it does not depend on future operational decisions and other parameters such as customer patience and service durations. Simulations show that if one does not distinguish between the total volume and the fresh volume, and uses the total volume to make operational decisions, it could lead to unnecessary costs. Thus, knowing the fresh volume is important for call centers. However, direct calculation of the number of fresh calls is difficult in some call centers, since customer identity information is not available in their data. In our model, we try to estimate the redial probability, reconnect probability and the fresh calls simultaneously in call center data without customer identity information by solving a minimization problem. However, we show that these three parameters cannot be estimated simultaneously. It is empirically shown that in order to have an accurate estimation, one variable needs to be given. We propose a polling method in call centers to calculate the reconnect probability. Once the reconnect probability is given, we show via simulation examples that the other two variables can be estimated. We also validate our model via two real call center data sets. Our estimation of the redial probabilities for both data sets are close to the actual redial probabilities, with errors of less than 0.09. Furthermore, our estimation of the number of fresh calls are very close to the real number of fresh calls, with WAPE$_F$ less than 3 percent.

In addition to help call center managers to estimate the fresh volumes of the call centers, this paper also addresses the reconnect customer behavior in call centers. In the data set of Vanad Laboratories, we find out that the number of reconnects is significant. Neglecting it will lead to inaccurate prediction of the call volumes, which will

cause inaccurate staffing. Inspired by these findings in this paper, we propose the following topics for further research:

(a) For a call center manager, it would be interesting to know what are the consequences of neglecting reconnects in terms of costs or service levels;
(b) In order to make the right staffing decisions, it would be useful to evaluate the service levels of a call center with consideration of the reconnect behaviors;
(c) The redial and reconnect behaviors will introduce intraday correlation to the call center data. For example, if a manager saw a busy morning, he would expect a busy afternoon, since some "morning customers" will redial or reconnect in the afternoon. This raises an interesting question: how should the manager update the agents' schedules dynamically when morning realizations are available?

References

Aguir, S., Akşin, O., Karaesmen, F., & Dallery, Y. (2008). On the interaction between retrials and sizing of call centers. *European Journal of Operational Research, 191*(2), 398–408.

Aguir, S., Karaesmen, F., Akşin, O., & Chauvet, F. (2004). The impact of retrials on call center performance. *OR Spectrum, 26*(3), 353–376.

Andrews, B., & Cunningham, S. (1995). L.L. Bean improves call-center forecasting. *Interfaces, 25*(6), 1–13.

Artalejo, J., & Pozo, M. (2002). Numerical calculation of the stationary distribution of the main multiserver retrial queue. *Annals of Operations Research, 116*(1), 41–56.

Avramidis, A., Deslauriers, A., & L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science, 50*(7), 896–908.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., et al. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association, 100*(469), 36–50.

Charnes, A., Cooper, W., & Ferguson, R. (1955). Optimal estimation of executive compensation by linear programming. *Management science, 1*(2), 138–151.

Chen, J., & Rubin, H. (1986). Bounds for the difference between median and mean of gamma and poisson distributions. *Statistics & probability letters, 4*(6), 281–283.

Ding, S., & Koole, G. Optimal call center forecasting and staffing under arrival rate uncertainty. Available via http://sihanding.com/wp-content/uploads/2014/10/2014_DingKoole_optimal_staffing.pdf [accessed January 26, 2015].

Ding, S., Koole, G., & van der Mei, R. (2013). A method for estimation of redial and reconnect probabilities in call centers. *Proceedings of the 2013 winter simulation conference: Simulation: Making decisions in a complex world* (pp. 181–192). IEEE Press.

Falin, G. (1995). Estimation of retrial rate in a retrial queue. *Queueing systems, 19*(3), 231–246.

Gans, N., Shen, H., Zhou, Y., Korolev, K., McCord, A., & Ristock, H. (2009). Parametric stochastic programming models for call-center workforce scheduling. *Technical report* Available via http://faculty.washington.edu/yongpin/Gans-Shen-Zhou-Scheduling.pdf [accessed March 27, 2013].

Hoffman, K., & Harris, C. (1986). Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research, 27*(2), 207–214.

Ibrahim, R., & L'Ecuyer, P. (2013). Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management, 15*(1), 72–85.

Jongbloed, G., & Koole, G. (2001). Managing uncertainty in call centres using poisson mixtures. *Applied Stochastic Models in Business and Industry, 17*(4), 307–318.

Koole, G. (2013). Call center optimization. Amsterdam: MG books.

Mandelbaum, A., Massey, W., Reiman, M., Stolyar, A., & Rider, B. (2002). Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems, 21*(2), 149–171.

Narula, S., Saldiva, P., Andre, C., Elian, S., Ferreira, A., & Capelozzi, V. (1999). The minimum sum of absolute errors regression: a robust alternative to the least squares regression. *Statistics in medicine, 18*(11), 1401–1417.

Shen, H., & Huang, J. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management, 10*(3), 391–410.

Steckley, S., Henderson, S., & Mehrotra, V. (2010). Forecast errors in service systems. *Probability in the Engineering and Informational Sciences, 23*(2), 305.

Stolletz, R. (2008). Approximation of the non-stationary $m(t)/m(t)/c(t)$-queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research, 190*(2), 478–493.

Taylor, J. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science, 54*(2), 253–265.

Taylor, J. (2012). Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science, 58*(3), 534–549.

Weinberg, J., Brown, L., & Stroud, J. (2007). Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association, 102*(480), 1185–1198.

Yom-Tov, G., & Mandelbaum, A. (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management, 16*(2), 283–299.