VU Research Portal

Optimal call center forecasting and staffing

Ding, Sihan; Koole, Ger

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

RESEARCH ARTICLE

Optimal call center forecasting and staffing

Sihan Ding[1] and Ger Koole[2]

[1] CWI, Amsterdam, Netherlands. E-mail: dingsihan@hotmail.com

[2] Vrije Universiteit Amsterdam & CCmath, Amsterdam, Netherlands. E-mail: ger.koole@vu.nl.

Abstract

In this paper, we consider a two-stage call center staffing model. In the first stage, the interval staffing levels are set under arrival rate uncertainty. In the second stage, these initial staffing levels are corrected to the right value based on more precise arrival rate information. We show that this problem is of newsvendor type, where the costs are the initial staffing costs plus the second stage adaptation costs. We show that we should initially staff according to a quantile of the distributional forecast, rather than the mean. It is also shown that the errors in staffing are approximately linear in the forecasting errors. This leads to the conclusion that the weighted sum of errors should be the error measurement in call center forecasting, since minimizing, it minimizes the total staffing costs. In special cases where the costs are symmetric for over- and understaffing, this is equivalent to minimizing the weighted absolute percentage error.

1. Introduction

In practice, call center forecasting and staffing are done in multiple stages. Usually, 4 weeks in advance, the operational planning cycle is started by making a forecast. This forecast is used to determine the initial staffing levels, for every 15- or 30-min interval, often using the Erlang C or Erlang A formula. The staffing levels in their turn are input for making the agent schedules.

However, between the time, the schedule is made and its execution things change. Some factors are still unknown by the time that the forecasts are made, such as the weather and the effect of advertisement campaigns. Theory tells us that updated forecasts have a higher accuracy ([20]). Ideally, staffing levels and schedules should be adapted every time new information becomes available. In practice, updates are dealt with in an ad hoc manner. Rarely forecasts are updated, and schedules are mostly updated on the basis of changes in the availability of agents. On the day itself, adaptations to the schedule are made to make sure, to the extent possible, that service levels (SLs) are met, exploiting flexibility in agent schedules and task assignments. This process is extremely hard to model but the outcome is surprisingly often that the SL is met by the end of the day, meaning that somehow the right staffing levels were attained.

For these reasons, we study the following two-stage model in this paper. In the first stage, for every say 15-min interval, an initial forecast of the arrival rate is made on which the initial staffing level is based. These staffing levels are later, during the second stage, corrected to the right value based on the real rates. The goal is to minimize the total sum of initial staffing and intra-day adaptation costs. We show two results. The first is that, if a distributional forecast of the arrival rate is made, then the first-stage staffing problem is of newsvendor type, and the staffing level is the staffing belonging to the right quantile of the arrival rate distribution.

As an example, suppose you expect an arrival rate of 100, with a normally distributed error with a standard deviation of 20, leading to an expected absolute error of around 17%, which is not uncommon. Assume the average handling time is 4, thus an expected load of 400. In this example, we use Erlang C

with an SL of 80/20 (80% answered within 20 s) and a non-integer number of agents, calculated as a simple linear interpolation between the integer staffing level just below and above 80%. If we schedule according to the expected rate, then we schedule close to 411 agents, thus 11 more than the load. Suppose that regular staffing costs are 10, adding flexible staff in the second step costs 15, sending agents home has final costs 1. Because of the asymmetry of the costs, we have to overstaff with respect to the expected arrival rate, according to Theorem 1 to the quantile of $5/(1 + 5)$, which is close to 119, leading to a staffing level of 488.5.

Sampling from the arrival rate and averaging the results for the samples show the following. Under full information, the costs are close to 4,107. By staffing according to the expected arrival rate, total costs are 4,300, thus the intra-day adaptation costs are close to 4%. Staffing 77 more agents gives total costs 4,228, which is a reduction by close to 2%. This is a relatively small reduction, but note that it comes at no extra work, the staffing formula just needs to be replaced. And we should realize that for different parameters, the reduction can be bigger. The R code of this example and other calculations can be found in "minimal_WAPE" on [13].

The second result is that, if a point forecast is made, then the total costs are approximately linear in the error with different coefficients for under- and overstaffing. Thus, the first result tells us how to staff, the second gives an approximation of the costs under different staffing levels. This approximation leads to the results that, when staffing multiple periods such as the quarters of a day, the point forecast that minimizes a weighted form of the WAPE (the weighted absolute percentage error) also minimizes the total staffing costs. In the case of equal under- and overstaffing costs, this reduces to the regular WAPE. This tells us that the (weighted) WAPE is the preferred forecasting error measure in call centers, which allows us to compare arbitrary point forecasts. Note that the weighted WAPE is minimized by the forecast that consists of the quantiles used in the newsvendor.

The implications for practice are as follows. Ideally, a distributional forecast should be made with staffing according to the quantile determined by the under- and overstaffing costs. However, hardly ever distributional forecasts are made. A point forecasts should preferably minimize the weighted WAPE. Then it can be used directly for staffing. If a forecast is made that minimizes the WAPE, then historical forecasting errors can be used to turn it into a distributional forecast which can again be used for staffing.

The results just described will have lower costs than the methods described in the literature, which are mostly based on expected arrival rates and error measurements such as the RMSE. For example, Shen *et al.* [19] use the MSE as error measurement to update intra-day arrivals. Ibrahim and l'Ecuyer [10] compare the MAPE, RMSE and MSE of different forecasting models. Aldor *et al.* [2] compare the MAPE and RMSE of four fixed-effects models. Brown *et al.* [4] and also Aldor *et al*. [2] are examples of papers that use the mean arrival load of the system to generate staffing levels, using queueing formulas, such as the Erlang A formula or square-root staffing. Both choices are arguable and lead to sub-optimal decisions.

For academic overviews on call center planning, see Aksin *et al.* [1] and Gans *et al.* [7]; for practitioner-level texts, see Cleveland and Mayben [5] and Koole [14]. Ibrahim *et al.* [11] is a survey on call center forecasting that also discusses arrival rate uncertainty extensively. Ideally, intra-day forecasts should be updated in a Bayesian manner, as in Shen and Huang [19], instead of the currently used ad hoc manner.

Finally, we would like to mention a number of papers which are related to the ideas of asymmetric staffing and/or intra-day management. The first is Bassamboo *et al.* [3] which considers a single-stage staffing model with arrival rate uncertainty in which the expected costs of staffing, queueing and abandonments are minimized. The optimal staffing level is a quantile of the arrival rate without safety staffing, as long as the arrival rate uncertainty is high enough. Note that this means that there is no SL guarantee, the SL depends on the realization.

Gurvich *et al.* [9] and Roubos *et al.* [18] combine single-stage staffing with change constraints, guaranteeing that the SL is met with a certain probability.

A number of papers study two-stage staffing models, especially Gans *et al.* [8] and Mehrotra *et al.* [17]. They formulate more involved mathematical programming models (e.g., modeling a whole day including the dependencies in the arrival rate distribution) containing the two decisions and solve them

numerically. Finally, Liao *et al.* [15] study a two-stage model with in and outbound calls in which the flexibility consists of the possibility to schedule agents for handling outbound calls in overtime. Solutions using stochastic and robust programming are compared.

2. Optimal staffing

We first describe our model. Arrivals in a certain interval can be seen as coming from a Poisson distribution with parameter λ. However, λ is unknown at the first moment staffing is done. In this section, we assume that we have a distributional forecast in the form of a random variable Λ. The actual arrival rate λ can be seen as a realization of Λ. On the basis of Λ, we decide on the initial staffing level s.

There is one moment at which we do intra-day management. As explained in the introduction, we assume that at this moment, we know the actual arrival rate λ. Based on this arrival rate, we adapt our staffing level to the right staffing level $S(\lambda)$, independent of the initial level $s \in \mathbb{N}$, where S is a function of λ that determines the minimal staffing level such that a certain SL requirement is met. Typically, S is determined using an Erlang formula, but depending on the type of service another function can be used as well, as long as it is non-decreasing.

When, during a certain interval, an agent is scheduled in the first step and kept then the costs for this agent are c. When an agent is scheduled but sent home or given another task to do at the second step, at the beginning of the interval, then the costs are c_o, where o stands for overstaffing. Finally, when an agent is scheduled only at the second step then its costs are $c + c_u$, with u meaning understaffing. Thus c_o and c_u are the additional costs for flexibility. They are typically 10 or 20% of c, and any call center has such a flexible layer for up- and downscaling.

Now, for initial staffing level s and realization λ, the total staffing costs $C(s, \lambda)$ are:

$$C(s, \lambda) = cS(\lambda) + c_o(s - S(\lambda))^+ + c_u(S(\lambda) - s)^+$$
$$= cs + (c_o - c)(s - S(\lambda))^+ + (c_u + c)(S(\lambda) - s)^+,$$

where $y^+ := \max\{0, y\}$.

The total expected costs are

$$\mathbb{E}C(s, \Lambda) = c\mathbb{E}S(\Lambda) + c_o\mathbb{E}(s - S(\Lambda))^+ + c_u\mathbb{E}(S(\Lambda) - s)^+. \tag{1}$$

The cost-optimal staffing $s^* := \arg\min_s \mathbb{E}C(s, \Lambda)$ can then be found by

$$s^* = \arg\min_s\{c_o\mathbb{E}(s - S(\Lambda))^+ + c_u\mathbb{E}(S(\Lambda) - s)^+\}. \tag{2}$$

Eq. (2) has the form of the newsvendor problem, with the random demand replaced by $S(\Lambda)$. Therefore, if we denote with H the cdf of the random variable $S(\Lambda)$, then, by solving Eq. (2), we obtain

$$s^* = H^{-1}\left(\frac{c_u}{c_o + c_u}\right),$$

where H^{-1} is the *quantile function* of H. For any cdf F, its quantile function is defined by $F^{-1}(y) := \inf\{x \in R : F(x) \geq y\}, 0 \leq y \leq 1$.

We assume that S is a non-decreasing function. This is a natural assumption: when there are more arrivals, then we need more agents to obtain the required SL. We can show that S is non-decreasing for a number of often-used models and performance measures, as is shown in the appendix.

Theorem 1. *If $S(\cdot)$ is non-decreasing, then*

$$s^* = S\left(F_\Lambda^{-1}\left(\frac{c_u}{c_o + c_u}\right)\right),$$

with F_Λ the distribution function of Λ.

Proof. It suffices to show that $H^{-1}(p) = S(F_\Lambda^{-1}(p))$ for any $0 \le p \le 1$. To this end, let $\lambda_p := F_\Lambda^{-1}(p)$. Due to the properties of the quantile function, we have

$$F_\Lambda(\lambda_p) \ge p.$$

Furthermore, we have

$$\mathbb{P}(S(\Lambda) \le S(\lambda_p)) \ge \mathbb{P}(\Lambda \le \lambda_p) \ge p,$$

which leads to

$$H(S(\lambda_p)) \ge p,$$

from which it follows that $S(\lambda_p) \in \{x \in R : H(x) \ge p\}$. Due to the definition of $H^{-1}(p)$, we have

$$H^{-1}(p) \le S(F_\Lambda^{-1}(p)).$$

Assume $H^{-1}(p) < S(F_\Lambda^{-1}(p))$. Then, we can always find some $\epsilon > 0$, such that $S(F_\Lambda^{-1}(p)) = H^{-1}(p) + \epsilon$. Moreover, we define $B := \{\lambda \in R : S(\lambda) = H^{-1}(p)\}$, and $\lambda' := \sup B$. Clearly, $B \ne \emptyset$. Therefore, under such assumptions, $\lambda' < F_\Lambda^{-1}(p)$ would be true, due to the fact that $S(F_\Lambda^{-1}(p)) = H^{-1}(p) + \epsilon > S(\lambda')$ and S being a non-decreasing function.

$H(H^{-1}(p)) \ge p$ must hold, because H is a cdf. Now we show that $H(H^{-1}(p)) \ge p$ contradicts $\lambda' < F_\Lambda^{-1}(p)$. If $H(H^{-1}(p)) \ge p$, then due to the definition of λ', we must have

$$p \le \mathbb{P}(S(\Lambda) \le H^{-1}(p)) = \mathbb{P}(\Lambda \le \lambda'). \tag{3}$$

Inequality (3) leads to $F_\Lambda(\lambda') \ge p$, which contradicts the fact that $\lambda' < F_\Lambda^{-1}(p)$. □

Theorem 1 proves that staffing according to the $c_u/(c_o + c_u)$ quantile of the arrival rate distribution minimizes the expected staffing costs. In the special case of $c_o = c_u$, staffing according to the median of Λ is optimal. This means that staffing according to the mean, which is often done, is not optimal, not even in the symmetric cost case, unless the mean is equal to the median.

In practice, it is often simpler to scale up than to scale down. Scaling up is often done by hiring flexible workers, who are often available on a short notice, especially when they work at home. Scaling down is sometimes not even possible, because of the unflexible contracts of the agents. In that case, $c_o = c$ and at the first stage staffing should be done very conservatively. Furthermore, many call centers have different layers of flexibility. First, flexibility is sought in the task assignment. If this is not sufficient then the number of agents is changed. This leads to increasing costs of up- and downscaling, giving a piece-wise linear costs function C in s. In general, there is no closed-form solution for s^*. A solution can be found numerically by calculating $\mathbb{E}C(s, \Lambda)$ for various values of s. Vice versa, call centers try to avoid high intra-day management costs by contracting enough flexible agents.

Next, we compare numerically staffing according to Theorem 1 and the usual staffing based on the expected forecast. We use the Erlang A model with an 80% answered within 20 s SL requirement based on the virtual waiting time (see Appendix for the definition). The average handling time is 4 min, the average patience is 5 min, $c = 1$. The other parameters and results can be found in Table 1. Define $s_a = S(\mathbb{E}\Lambda)$ and $s_n = S(F_\Lambda^{-1}(c_u/(c_o + c_u)))$. The expected costs are obtained by simulating Λ and calculating the Erlang A values. We obtain non-integer values for the numbers of agents by linear interpolation between neighboring integer values which have SL just above and below the required one. The first case is a rather symmetric one, both in terms of demand as in costs. We see that the optimal staffing is slightly higher than the usual staffing method, because overstaffing is slightly less expensive than understaffing. The next two situations are more asymmetric: in the first, it is not possible to scale down ($c_o = c$), in the second, it is not possible to scale up ($c_o = c$). We see the consequences in terms of

Table 1. *Staffing based on newsvendor model vs. $\mathbb{E}\Lambda$ in the Erlang A model*

Λ	$\mathbb{E}\Lambda, \sigma(\Lambda)$	c_u, c_o	s_a	s_n	$\mathbb{E}C(s_a, \Lambda)$	$\mathbb{E}C(s_n, \Lambda)$
Normal	20, 2	0.2, 0.1	82.2	85.5	82.8	82.7
Lognormal	20, 4	0.1, 1	82.2	62.8	88.2	83.8
Lognormal	20, 4	1, 0.1	82.2	103.8	87.6	84.5

the staffing levels and costs. Evidently, for more extreme input values, the results will be more striking, but these input values were chosen because they are realistic. Note that the values for the lognormal distribution are those of Λ, not those of the normal distribution from which the lognormal is constructed.

3. Staffing costs

In Section 2, we saw how to compute the optimal staffing level using the random arrival rate Λ. In this section, we quantify the costs of the error we inevitably make. We do this for λ sufficiently big, and an error that increasing proportionally with λ. Theorem 2 states that the costs are approximately linear in the error, with the coefficient depending on the sign of the error. This allows us to compare forecasts, also if they are made for multiple time periods. We characterize the optimal forecast in Theorem 3 in terms of the errors. It is the one minimizing the well-known WAPE in the case of symmetric costs.

We are interested in $C(S(\hat{\lambda}), \lambda)$, in which $\hat{\lambda}$ is the forecast on which the initial staffing level is based. This function is hard to characterize, therefore we will look at it for λ and $\hat{\lambda}$ large. To obtain a limiting argument, we have to define how $\hat{\lambda}$ behaves as $\lambda \to \infty$.

It is important to realize that call center actuals (the commonly used word for realizations) are best modeled by multiplicative models. They consist of the base level (the trend) multiplied by factors for the different seasonal components (intra-year, intra-week and intra-day) and special events (such as marketing campaigns, public holidays and bill runs) that influence volume. Thus, for example, a 5% error in the day-of-the-week factor results in a 5% error in the daily volume, but also a 5% error at the interval level. If volumes grow, for example, because of an increasing trend or an end-of-year peak, we do not expect forecasting errors to disappear; instead, we expect errors to grow in a multiplicative way. The only term that scales sublinearly is the error coming from the Poisson distribution. However, this term, in the order of $\sqrt{\lambda}$, is negligible for larger λ. This leads to the following model for $\hat{\lambda}$: $\hat{\lambda} = h\lambda$, with $h > 0$ and likely to be close to 1, as it indicates the accuracy of the forecast. The factor h is not explicitly calculated: it is the product of the percentage errors of all the components of the forecast $\hat{\lambda}$. It can be calculated once λ is observed.

For more details on multiplicative models, see Chapter 3 of [14]. Further evidence for the multiplicativity comes from the small dataset of Figure 1. We see actuals from two different weeks, one with high and one with low volume. We see that the intra-week pattern scales with the overall volume. A simple linear model with day and week as explanatory variables shows the same thing: an additive model has a WAPE of 5.5%, a multiplicative model a WAPE of less than 1% (for a definition of the WAPE, see below). The additive fit is shown as a dashed line in Figure 1, the multiplicative fit is dotted and coincides almost exactly with the actuals.

In the next theorem, we formulate our approximation of $C(S(\hat{\lambda}), \lambda)$. We use the following notation: $f(x) = o(g(x))$ if $\lim_{x \to \infty} f(x)/g(x) = 0$. Please note as that "$+G$" is a by now common extension of the Kendall notation indicating the distribution of the patience. In the next results, we study the $M|M|s + G$ queue, also known as the Erlang A model ([16]). Note that common performance measures in call centers are the waiting time quantile (often called the SL), the expected waiting time (*average speed of answer* (ASA)), and the abandonment rate. In the next theorem, we will use the fact that the Erlang A model is monotone in these performance measures, which is shown in the appendix.
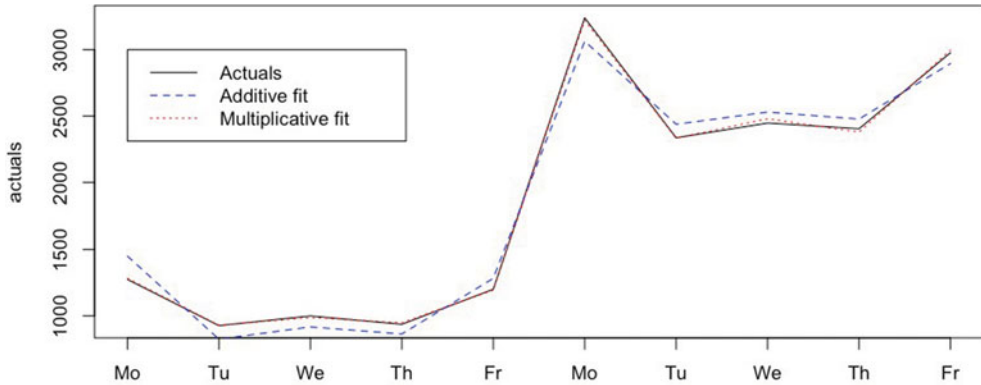
Figure 1. *Actuals and fits of multiplicative and additive models for 2 weeks.*

Theorem 2. *For the $M|M|s + G$ model, given performance constraints based on SL, ASA or abandonment ratio,*

$$C(S(\hat{\lambda}), \lambda) = cS(\lambda) + (1 - \gamma)(c_o(\hat{\lambda} - \lambda)^+ + c_u(\lambda - \hat{\lambda})^+)\beta + o(\lambda), \tag{4}$$

for some $\gamma \geq 0$ which depends on the performance constraint and β the expected service time.

Proof. Consider first the case $h \geq 1$. Then $\hat{\lambda} \geq \lambda$ and, according to Theorem A.1, $S(\hat{\lambda}) \geq S(\lambda)$. Therefore,

$$C(S(\hat{\lambda}), \lambda) = cS(\lambda) + c_o(S(\hat{\lambda}) - S(\lambda)).$$

From [16], Section 2, we know that $S(\lambda) = (1 - \gamma)\lambda\beta + o(\lambda)$ and also $S(\hat{\lambda}) = (1 - \gamma)\hat{\lambda}\beta + o(\hat{\lambda}) = (1 - \gamma)\hat{\lambda}\beta + o(\lambda)$. Thus,

$$\begin{aligned} C(S(\hat{\lambda}), \lambda) &= cS(\lambda) + c_o((1 - \gamma)\hat{\lambda}\beta - (1 - \gamma)\lambda\beta + o(\lambda)) \\ &= cS(\lambda) + c_o(1 - \gamma)(\hat{\lambda} - \lambda)\beta + o(\lambda). \end{aligned}$$

Similarly for $h < 1$. $\square$

Remark 1. *Depending on the performance objective, different operational regimes apply, with different limiting behavior. All are $o(\lambda)$, but in some cases, stronger results are obtained: for the SL objective, the limiting behavior is $O(\sqrt{\lambda})$ (with $f(x) = O(g(x))$ if $\limsup_{x \to \infty} |f(x)/g(x)| < \infty$). See Section 2 of [16] for details.*

From Theorem 2, we conclude that the total costs are approximately linear in the (weighted) error in the rate. We can interpret it as a weak version of Theorem 1. If we replace λ by Λ and take expectations then we see that, in the limit, we should staff according to $(1 - \gamma)\beta F_\Lambda^{-1}(c_u/(c_o + c_u))$. $(1 - \gamma)\beta\lambda$ is the linear approximation of $S(\lambda)$.

In our model, we assumed that we know λ in time to adapt the staffing level. In practice, we do not observe λ, we observe a realization of $N_\lambda \sim \text{Poisson}(\lambda)$. Due to the Central Limit Theorem, we have $N_\lambda - \lambda = O(\sqrt{\lambda})$ a.s. Therefore,

$$C(S(\hat{\lambda}), \lambda) = cS(\lambda) + (1 - \gamma)(c_o(\hat{\lambda} - N_\lambda)^+ + c_u(N_\lambda - \hat{\lambda})^+)\beta + o(\lambda) \text{ a.s.}$$

Thus the costs are also, in the limit, linear in the error w.r.t. the actuals, that is, the observed call volume.

Let us now consider T measurements. Let, for $t = 1, \ldots, T$, λ_t be the realizations and $\hat{\lambda}_t = h_t \lambda_t$ the forecast on which the initial staffing was based. Now we define a number of forecasting error measurements:

$$\text{MAPE} = \frac{1}{T} \sum_{t=1}^{T} \frac{|\hat{\lambda}_t - \lambda_t|}{\lambda_t}$$

The MAPE, *mean average percentage error*, is an intuitive performance measure that is easy to interpret by practitioners: "over the last week the forecast was on average 5% off." However, the MAPE is prone to small absolute errors for small volumes and gives an error when the actual is 0 in an interval. Therefore, it is better to weigh the APEs with the relative volume, leading to a very simple formula:

$$\text{WAPE} = \sum_{t=1}^{T} \frac{\lambda_t}{\sum_{s=1}^{T} \lambda_s} \frac{|\hat{\lambda}_t - \lambda_t|}{\lambda_t} = \frac{\sum_{t=1}^{T} |\hat{\lambda}_t - \lambda_t|}{\sum_{t=1}^{T} \lambda_t}$$

In what follows it will appear to be useful to define a weighted version of the WAPE, the wWAPE, for some $w \in [0, 1]$:

$$\text{wWAPE} = \frac{2 \sum_{t=1}^{T} [w(\hat{\lambda}_t - \lambda_t)^+ + (1 - w)(\lambda_t - \hat{\lambda}_t)^+]}{\sum_{t=1}^{T} \lambda_t}$$

Theorem 3. *For λ_t large, the forecast with the lowest wWAPE, with $w = c_o/(c_o + c_u)$, minimizes the total staffing costs.*

Proof. The total costs $C_T(S(\hat{\lambda}), \lambda)$, with $\hat{\lambda}$ and λ T-dimensional vectors, are approximated by

$$C_T(S(\hat{\lambda}), \lambda) \approx \sum_{t=1}^{T} cS(\lambda_t) + (1 - \gamma)\beta \sum_{t=1}^{T} (c_o(\hat{\lambda}_t - \lambda_t)^+ + c_u(\lambda_t - \hat{\lambda}_t)^+).$$

This value is minimized by the forecast $\hat{\lambda}$ that minimizes $\sum_{t=1}^{T}(c_o(\hat{\lambda}_t - \lambda_t)^+ + c_u(\lambda_t - \hat{\lambda}_t)^+)$, the weighted sum of errors, which is equivalent to minimizing the wWAPE. □

Note that in the symmetric case $c_o = c_u$, this reduces to minimizing the WAPE. It is interesting to note that the Poisson variability gives lower bounds to the achievable absolute percentage error (APE) per interval, given by $\mathbb{E}|N_\lambda - \lambda)|$ with $N_\lambda \approx \text{Poisson}(\lambda)$. This fact is often ignored by call center managers: they sometimes set forecast error targets that are impossible to achieve, simply because the variability is the Poisson distribution is higher than the target. The APE of the Poisson distribution is given in Crow [6], a simple approximation based on the normal distribution is $\sqrt{2/(\lambda\pi)}$. For example, for $\lambda = 100$, the APE is 8%.

Theorem 3 can also be used with the actuals instead of the rates, as in the single-interval case.

So far, we looked at limiting behavior as $\lambda \to \infty$, but we claim that the results can also be used for smaller values of λ, and thus that the (asymmetric) WAPE should be used under all circumstances. To support this, we executed the following experiment. We sampled from $\Lambda = (\Lambda_1, \ldots, \Lambda_{20})$ with $\Lambda_t \sim N(1, 0.1)t$. For $\hat{\lambda}_t = t$, which is optimal for the symmetric case, and parameters $c = 0$, $c_o = c_u = 1$, $\beta = 4$, and an 80/20 target SL, we computed the staffing levels using Erlang C, and from that the total costs. Making a scatter plot of WAPE and MAPE against the total costs leads to the left plot of Figure 2. The relation between WAPE and the costs is perfectly linear, supporting the claim that Theorem 3 can be used for all values of λ.

We also considered the asymmetric case, with $c_o = 1$ and $c_u = 5$, keeping all other parameters equal. We found again, as can be seen in the right plot of Figure 2, a linear relation, both for the optimal staffing rule as for the one that staffs according to the median rate. As can be expected, the optimal staffing rule has lower costs. Again, the R code can be downloaded from [13].
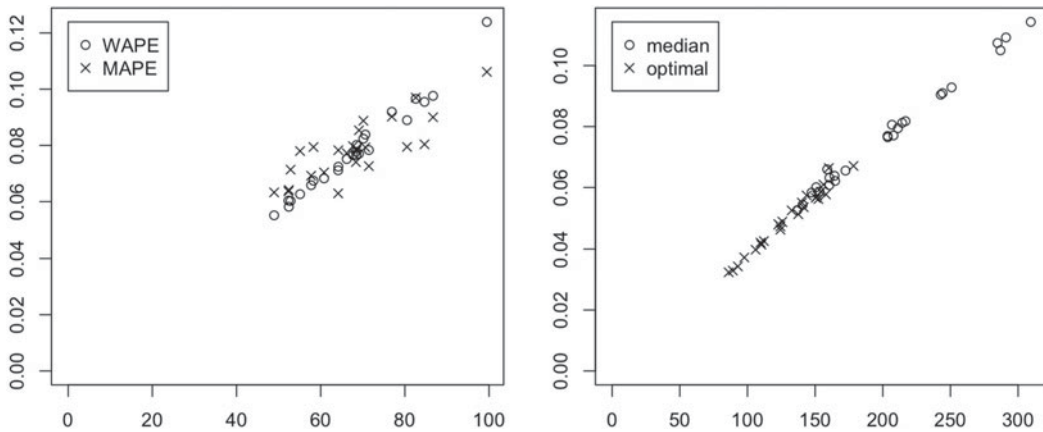
Figure 2. *Left: Scatter plot of symmetric case, total costs (x-axis) against MAPE and WAPE; Right: Scatter plot of asymmetric case, total costs against median and optimal initial staffing.*

4. Conclusion

In the context of a two-stage call center staffing problem, we have shown that the regular staffing method is not theoretically justified and that considerable savings can be obtained by staffing according to the newsvendor method described in this paper. Furthermore, we derived that, for the usual performance models, the staffing costs are approximately linear in the absolute arrival rate error. This led to the conclusion that the weighted sum of absolute errors is the asymptotical optimal forecasting error measurement, since minimizing, it leads to minimizing costs. This is equivalent to minimizing WAPE in case under- and overstaffing are equally expensive.

References

[1] Akşin, O.Z., Armony, M., & Mehrotra, V. (2007). The modern call-center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* 16: 665–688.

[2] Aldor-Noiman, S., Feigin, P., & Mandelbaum, A. (2009). Workload forecasting for a call center: Methodology and a case study. *The Annals of Applied Statistics* 3(4): 1403–1447.

[3] Bassamboo, A., Randhawa, R., & Zeevi, A. (2010). Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* 56(10): 1668–1686.

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

[5] Cleveland, B. & Mayben, J. (1997). *Call center management on fast forward*. Annapolis, Maryland: Call Center Press.

[6] Crow, E.L. (1958). The mean deviation of the Poisson distribution. *Biometrika* 45: 556.

[7] Gans, N., Koole, G.M., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5: 79–141.

[8] Gans, N., Shen, H., Zhou, Y.-P., Korolev, N., McCord, A., & Ristock, H. (2015). Parametric forecasting and stochastic programming models for call-center workforce scheduling. *Manufacturing & Service Operations Management* 17(4): 571–588.

[9] Gurvich, I., Luedtke, J., & Tezcan, T. (2010). Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science* 56(7): 1093–1115.

[10] 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.

[11] Ibrahim, R., Ye, H., L'Ecuyer, P., & Shen, H. (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. *The International Journal of Forecasting* 32: 865–874.

[12] Jouini, O., Koole, G.M., & Roubos, A. (2013). Performance indicators for call centers with impatience. *IIE Transactions* 45(3): 341–354.

[13] Koole, G.M. (2020). Github account. github.com/gerkoole.

[14] Koole, G.M. (2013). *Call center optimization*. Amsterdam: MG Books.

[15] Liao, S., Koole, G., van Delft, C., & Jouini, O. (2012). Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR Spectrum* 34(2): 691–721.

[16] Mandelbaum, A. & Zeltyn, S. (2009). Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research* 57(5): 1189–1205.

[17] Mehrotra, V., Ozluk, O., & Saltzman, R. (2010). Intelligent procedures for intra-day updating of call center agent schedules. *Production and Operations Management* 19: 353–367.

[18] Roubos, A., Koole, G., & Stolletz, R. (2012). Service-level variability of inbound call centers. *Manufacturing & Service Operations Management* 14: 402–413.

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

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

[21] Zeltyn, S. & Mandelbaum, A. (2005). Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. *Queueing Systems* 51: 361–402.

Appendix. Monotonicity of the staffing function

For some yet unspecified call center model, we write the (expected) performance as $P(s, \lambda)$. We assume there is some maximal allowable performance level τ. Then S can be written as $S(\lambda) = \inf\{s | P(s, \lambda) \leq \tau\}$. Examples are the expected waiting time (ASA) and its tail probability (SL) in the Erlang C model. In the case of ASA, we take P equal to the expected waiting time; in the case of SL, we take $P = \mathbb{P}(W > t)$ with W the stationary waiting time in the queue and t the so-called waiting time limit or *acceptable waiting time*. In the case of abandonments, we have to decide how abandonments are integrated into the performance measures. The abandonment % or ratio now becomes important, but we also have to decide how the abandonments are accounted for in the measures that are functions of the waiting time. Two regular choices are the time in queue W and the *virtual* or *offered waiting time* V (the waiting time of a *test customer* with ∞ patience), but other choices are possible (see [12]). The extension of the Erlang C model to general patience distributions is written as $M|M|s+G$ model. An overview of results for this model can be found in Section 9 of Zeltyn and Mandelbaum [21]. The special case $M|M|s+M$ is known as the Erlang A model.

Theorem 4. *The staffing function S is non-decreasing for the $M|M|s+G$ model and any τ and P given by $\mathbb{P}(W > t)$, $\mathbb{P}(V > t)$, $\mathbb{E}W$, $\mathbb{E}V$, or the abandonment rate.*

Proof. If P is non-increasing in s and non-decreasing in λ then S is non-decreasing. That P is non-increasing in s can be found for all performance measures in Section 2.1 of the online appendix of [21]. To show that P is non-decreasing in λ for the different performance measures, it suffices to show it for $\mathbb{P}(V > t)$: all other results follow directly from that.

We introduce the following notation, slightly adapted from [21]:

$$H(x) = \int_0^x (1 - G(u))\, du,$$

$$J_\lambda(t) = \int_t^\infty e^{\lambda H(x) - s\mu x}\, dx,$$

$$J_\lambda = J_\lambda(0),$$

$$\varepsilon_\lambda = \int_0^\infty e^{-t} \left(1 + \frac{t\mu}{\lambda}\right)^{s-1} dt,$$

where μ is the service rate and G is the cdf of patience time. Then, according to [21],

$$\mathbb{P}_\lambda(V > t) = \frac{\lambda J_\lambda(t)}{\varepsilon + \lambda J_\lambda}.$$

We now show that for fixed s, if $\lambda_1 > \lambda_2 > 0$, then $\mathbb{P}_{\lambda_1}(V > t) \geq \mathbb{P}_{\lambda_2}(V > t)$, that is,

$$\frac{J_{\lambda_1}(t)}{\varepsilon_{\lambda_1}/\lambda_1 + J_{\lambda_1}} - \frac{J_{\lambda_2}(t)}{\varepsilon_{\lambda_2}/\lambda_2 + J_{\lambda_2}} \geq 0,$$

which is equivalent to showing that

$$\frac{J_{\lambda_1}(t)\varepsilon_{\lambda_2}/\lambda_2 - J_{\lambda_2}(t)\varepsilon_{\lambda_1}/\lambda_1 + J_{\lambda_1}(t)J_{\lambda_2} - J_{\lambda_2}(t)J_{\lambda_1}}{(\varepsilon_{\lambda_1}/\lambda_1 + J_{\lambda_1})(\varepsilon_{\lambda_2}/\lambda_2 + J_{\lambda_2})} \geq 0.$$

Because $J_\lambda(t)$ is increasing and ε_λ is decreasing in λ it is readily seen that $J_{\lambda_1}(t)\varepsilon_{\lambda_2}/\lambda_2 - J_{\lambda_2}(t)\varepsilon_{\lambda_1}/\lambda_1 > 0$. Because all its terms are ≥ 0 also $(\varepsilon_{\lambda_1}/\lambda_1 + J_{\lambda_1})(\varepsilon_{\lambda_2}/\lambda_2 + J_{\lambda_2}) > 0$. Thus, we only need to show

$$J_{\lambda_1}(t)J_{\lambda_2} - J_{\lambda_2}(t)J_{\lambda_1} \geq 0.$$

Its proof is equivalent to that of Equation (2.4) in the online appendix of [21].   □

Note that the $M|M|s$ model is a special case of the $M|M|s + G$ model (with ∞ patience). Thus Theorem A.1 holds also for the often-used Erlang C model.