Validation of call center workforce scheduling models

Ger KOOLE
Vrije Universiteit Amsterdam, the Netherlands, ger.koole@vu.nl

Siqiao LI
CCmath bv, Amstelveen, the Netherlands, siqiao@ccmath.com

Sihan DING
Apple Inc, China, dingsihan@hotmail.com

As part of an effort to validate call center performance models we address some technical issues to be able to perform this validation, especially the role of noise and the way to eliminate forecasting errors. We also give the main results of the validation study.

Key words: validation, call centers, simulation

1. Introduction

Many models to predict call center performance exist in the literature (see, e.g., Gans et al. (2003)), but little is known about their accuracy. This study tries to fill this gap by comparing different simulation models to the realized performance of one particular call center. The models differ in a number of ways, especially in the way arrivals, patience, agent heterogeneity, and breaks are modeled. We would like to identify the model that minimizes the mean absolute error (MAE) between the predicted and realized service level.

We focus on daily performance measures. There are a number of problems with comparing daily realized performance and the outcome of simulation models. The source of these problems is that every day is different: we do not have i.i.d. replications of the same day as is usually the case (see, e.g., Kleijnen (1995)). The issues we tackle are:
- eliminating the **system noise**. Even an exact model would not have error 0, because of random fluctuations during the day. Note that this noise is substantial, even at the daily level, as shown in Roubos et al. (2012). Therefore we would like to make a distinction between that part of the error that is due to noise and that part that is due to the imperfectness of the model, the *model error*;
- eliminating the **forecasting error**. No forecast is exact. To focus on the model error and eliminate the forecasting error we would like to use the actual rate of the inhomogeneous Poisson arrival process. However, this rate is unknown, therefore we use the actual instead. Unfortunately this creates an error by itself, because service level and actuals are negatively correlated, the actuals give more information than we are allowed to use. Therefore we have to find a way to eliminate this *cheating error*.

The contribution of this study is twofold: we develop theory on how to validate service models (Section 2) and we obtain insights in which features are crucial to model in call centers (Section 3).

2. Validating inhomogeneous models

Let us formulate our model mathematically. The r.v. Λ represents the parameters that change from day to day, which are the rate of the non-homogeneous Poisson process and the agents that are scheduled and their shifts. The performance, typically the service level obtained during a day, is denoted by X. Because X depends on Λ we write $X(\Lambda)$. Note that X is a r.v., even for fixed $\Lambda = \lambda$: its value depends on the realization of the Poisson process, the handling times, times at which agents take breaks, etc. We can also simulate various models. The service level estimation given by the simulation is written as $S(\Lambda)$. However, the arrival rate is not observed. We could replace that part of Λ by a forecast, but they are usually quite bad. Instead, we use use the actual instead of the rates. Λ in which the rates are replaced by the random realizations of the arrivals is written as $A(\Lambda)$, the corresponding simulation $S(A(\Lambda))$.

With $\mathbb{E}_\bullet$ we indicate the expectation with respect to the corresponding r.v. For example, $\mathbb{E}_S S(A(\Lambda))$ is the expected simulated performance of a random day for a random realization of the rates; $\mathbb{E}_\Lambda X(\Lambda)$ is the random performance "averaged" over the days. Note that we can interchange expectations, e.g., $\mathbb{E}_X \mathbb{E}_\Lambda X(\Lambda) = \mathbb{E}_\Lambda \mathbb{E}_X X(\Lambda)$.

We are interested in estimating $\mathbb{E}_\Lambda |\mathbb{E}_X X(\Lambda) - \mathbb{E}_S S(\Lambda)|$, which corresponds to the mean absolute error (MAE) of the service level. However, we measure $X(\Lambda)$ and $\mathbb{E}_S S(A(\Lambda))$, the latter by averaging over a sufficiently high number of simulations. We will show how to get an estimate of the MAE based on the actuals and simulations.

We can show that

$$\mathbb{E}_S S(A(\Lambda)) - X(\Lambda) \approx \mathbb{E}_S S(A^2(\Lambda)) - S(A(\Lambda)) + \mathbb{E}_S S(\Lambda) - \mathbb{E}_X X(\Lambda), \tag{1}$$

where A^2 means taking a sample twice. The part $\mathbb{E}_S S(A^2(\Lambda)) - S(A(\Lambda))$ approximates the model noise and the cheating error, and can be obtained only using simulations. The remainder is what we measure.

Simulations show that $\mathbb{E}_S S(A^2(\Lambda)) - S(A(\Lambda))$ is approximately normally distributed with a mean equal to 0.4% and a standard deviation of 3.7%. Its MAE is $\mathbb{E}_{\Lambda,A} \big| \mathbb{E}_S S(A^2(\Lambda)) - S(A(\Lambda)) \big| = 3\%$.

Our goal is to compute the MAE of $\mathbb{E}_S S(\Lambda) - \mathbb{E}_X X(\Lambda)$, $\mathbb{E}_\Lambda \big| \mathbb{E}_S S(\Lambda) - \mathbb{E}_X X(\Lambda) \big|$. We cannot simply subtract 3% from $\mathbb{E}_\Lambda \big| \mathbb{E}_S S(A(\Lambda)) - X(\Lambda) \big|$. To get a better understanding we start with computing the first two moments of the model error. Our simulations show $\mathbb{E}_{\Lambda,A} \big(\mathbb{E}_S S(A(\Lambda)) - X(\Lambda) \big) = 3.9\%$ and $\sigma_{\Lambda,A} \big(\mathbb{E}_S S(A(\Lambda)) - X(\Lambda) \big) = 5\%$.

The first moment of the model error follows directly by taking expectations in Equation (1):

$$\mu := \mathbb{E}_\Lambda \big(\mathbb{E}_S S(\Lambda) - \mathbb{E}_X X(\Lambda) \big) = 3.5\%.$$

This value is of interest in itself: it tells us to which extend the model is *biased*, to which extent there is a systematic error. But even if μ is small, the errors can be big but fluctuating, sometimes positive, sometimes negative. That is why we defined the performance measure as $\mathbb{E}_\Lambda \big| \mathbb{E}_S S(\Lambda) - \mathbb{E}_X X(\Lambda) \big|$.

Next we compute the standard deviation of the model error. As we expect $\mathbb{E}_S S(A^2(\Lambda)) - S(A(\Lambda))$ to be independent of the model error, we have

$$\sigma^2 := \sigma^2_{\Lambda,A} \big(\mathbb{E}_S S(\Lambda) - \mathbb{E}_X X(\Lambda) \big) \approx$$

$$\sigma^2_{\Lambda,A} \big(\mathbb{E}_S S(A(\Lambda)) - X(\Lambda) \big) - \sigma^2_{\Lambda,A} \big(\mathbb{E}_S S(A^2(\Lambda)) - S(A(\Lambda)) \big).$$

In our case $\sigma = \sqrt{(0.05^2 - 0.037^2)} = 3.4\%$. We conclude that the first moments of the model error and will also be quite similar to the measured error.

We can go a step further if we assume the measurements to be normally distributed. For the simulated noise/cheating factor this is the case, for the measured error this is a rough approximation. For $\mathbb{E}_S S(\Lambda) - \mathbb{E}_X X(\Lambda) \sim \mathrm{N}(0.035, 0.034)$, using a straightforward calculation, we find

$$\mathbb{E}_\Lambda \big| \mathbb{E}_S S(\Lambda) - \mathbb{E}_X X(\Lambda) \big| \approx \frac{2\sigma}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(\frac{\mu}{\sigma})^2} + \mu \Big(2\Phi\Big(\frac{\mu}{\sigma}\Big) - 1 \Big) = 4.0\%,$$

with Φ the standard normal distribution function. Note that the MAE without the correction is $\mathbb{E}_{\Lambda,A} \big| \mathbb{E}_S S(A(\Lambda)) - X(\Lambda) \big| = 4.5\%$. We conclude that the added precision given by the correction is small compared to the measured error, around 10%.

The results of this section are derived for the "HT & Patience Model", one of the models we analyzed. Other models give similar results. Note that even the best model will give a non-zero measured error. The lower bound to the measured error is approximated by $\mathbb{E}_S S(A^2(\Lambda)) - S(A(\Lambda))$. Its MAE is 3%.

3. Validation results

We did an extensive data analysis of a call center of which we had data about the calls and the agents: who did which calls, but also when agents where available to take calls, when they took breaks, etc. Based on this analysis we compared several models each having a different set of features. The most important ones are:

- **Handling times**: can be taken empirical or exponential;
- **Average handling times**: the averages can be taken all the same or weighted based on the mix of agents available that day;
- **Patience**: empirical or exponential;
- **Breaks**: are yes or no taken into account, proportional to the length of the breaks.

We find that all models overestimate the SL. The best one in terms of the MAE is the model with handling times exponential, AHTs adapted to the agent mix, empirical patience, and breaks taken into account. The MAE is $\approx 3\%$, for a SL that is usually around 80%. The worst models are the ones that have a yearly overall AHT and do not take breaks into account. Although it looks evident, these two features are often not taken into account.

Acknowledgments

We thank VANAD Laboratories for supplying the data and Rob van der Mei and Raik Stolletz for useful discussions.

References

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

Kleijnen, J.P.C. 1995. Verification and validation of simulation models. *European Journal of Operational Research* **82** 145–162.

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