Summary

Workforce Management in Call Centers: Forecasting, Staffing and Empirical Studies

Many call center managers and planners strive to make better workforce planning, and to balance well the quality of service and operational costs, which is essentially balancing customers waiting times and the number of agents. In specific, call centers need to satisfy service level targets with the least number of agents. There is no simple solution to this dilemma. However, researchers and practitioners over the decades have developed experiences and models to assist this decision making process. The whole procedure generally includes four steps: forecasting future call arrival volume, make staffing decisions, make rosters, traffic management.

Predicting future call volume is difficult, as the incoming number of calls involves large uncertainty. In fact, one can never know for sure how many inbound calls in advance. Forecasters aim at forecasts that are as accurate as possible. However, one question naturally arises: “how to measure accuracy?” There are many different measurements for the forecasting errors, such as the mean squared error, the mean percentage error, etc. Choosing different error measurements can sometimes lead to different choices of forecasting models or forecasts. In Chapter 2, we investigate this problem, and we discover that under the rate uncertainty, the weighted mean absolute percentage error is the optimal error measurement. This is because the forecasts that minimize it will also asymptotically minimize the initial staffing costs plus the traffic management costs. Also, we show that under certain assumptions, the staffing decision should be made based on certain percentile of the distributional forecasts, rather than the mean.

In Chapter 3, we show the literature on forecasting models. Besides choosing an appropriate model, another important factor is the data to be used in the forecasts. We show that in call center data sets there are redials and reconnects, as customers call back after abandonments and connected calls. Both behaviors have significant influence on the call volumes. Furthermore, we show that one should use the number of fresh calls (number of unique callers) to make forecasts instead of the total number of calls, since the number of fresh calls do not depend on the staffing decisions, while the total number of calls do. It is shown that by using the total number of calls, it may lead to inaccurate estimation of the call volume.

The redial and the reconnect behaviors are significant, yet, there is no staffing model
that supports both features. In Chapter 4, we develop such a model and approximate
the performance metrics by using fluid approximation. The fluid approximation gives
a first order approximation on the number of callers in the queue and in service, in the
redial and reconnect orbits. Based on those numbers, we approximate the mean total
arrival rate to the system, and use it as an input to the Erlang A formula to derive the
waiting time distribution.

Call centers may treat differently between the redials, the reconnects and the fresh calls.
We show one example in Chapter 5, where we study the performance of a call center
model with an call-back option, and the long-waiting callers are suggested to call back
some time later if the system is congested, and when they call back, they will receive
priorities over the fresh callers. It is discussed and shown that this model is efficient in
reducing the mean waiting time and it is fair to those long waiting callers.

In most of literature in call centers, researchers usually propose models to mimic the real
situation in call centers, where assumptions and simplifications of certain processes are
made. Then analytical solutions or approximation methods are developed, and eventu-
ally the methods are evaluated numerically. In Chapter 6, we take a different approach
by validating some of the commonly made assumptions and simplifications. This is
done by comparing the reality with simulation. To be more specific, we compare the
simulated service levels of a few staffing models with the actual service levels from the
data for a multi-skill call center. We empirically show that models are in general accu-
rate despite the assumptions made; however, certain simplifications and assumptions
should be treated with care. For example, agents’ breaks are important and should not
be ignored, which usually are not taken into considerations in staffing models. Fur-
thermore, we provide some empirical results in shrinkage, agents heterogeneity, agents
learning curve, etc., which give insight to managers and planners.