Call Center Measurements, Data Models and Data Analysis

Adapted from: **Telephone Call Centers: Tutorial, Review, and Research Prospects**

By Noah Gans (Wharton), Ger Koole (Vrije Universiteit) and Avishai Mandelbaum (Technion).
Published in Manufacturing and Service Operations Management (M&SOM), 5 (2), 2003;
downloadable from http://ie.technion.ac.il/serveng/References/CCReview.pdf

Abstract: Telephone call centers are an integral part of many businesses, and their economic role is significant and growing. They are also fascinating socio-technical systems in which the behavior of customers and employees is closely intertwined with physical performance measures. In these environments traditional operational models are of great value – and at the same time fundamentally limited – in their ability to characterize system performance.

We review the state of research on telephone call centers. We begin with a tutorial on how call centers function and proceed to survey academic research devoted to the management of their operations. We then outline important problems that have not been addressed and identify promising directions for future research.

1 Data Generation and Reporting

As it operates, a large call center generates vast amounts of data. Its IVR(s) and ACD are special-purpose computers that use data to mediate the flow of calls. (Acronyms are explained in the Appendix at the end.) Each time one of these switches takes an action, it records the call's identification number, the action taken, the elapsed time since the previous action, as well as other pieces of information. As a call winds its way through a call center, a large number of these records may be generated.

From these records, a detailed history of each call that enters the system can, in theory, be reconstructed: when it arrived; who was the caller; what actions the caller took in the IVR and how long each action took; whether and how long the caller waited in queue; whether and for how long a CSR served the call; who was the CSR. If the call center uses CTI, then additional data from the company's information systems may be included in the record: what the call was about; the types of actions taken by a CSR; related account information.

In practice, call centers have not typically stored or analyzed records of individual calls, however. This may be due, in part, to the historically high cost of maintaining adequately large databases – a large call center generates many gigabytes of call-by-call data each month – but clearly these quantities of data are no longer prohibitively expensive to store. It is also likely due to the fact that the software used to manage call centers – itself developed at a time when data storage was expensive – often uses only simple models which require limited, summary statistics. Finally, we believe that it is due to lack of understanding of how and why more detailed analyses should be carried out.

Instead, call centers most often summarize call-by-call data from the ACD (and related systems) as averages that are calculated over short time intervals, most often 30 minutes in length. These ACD data are used both for planning purposes and to measure system performance. They are carefully and continuously watched by call-center managers. Hence we shall describe them in class, and you will using then in assignments.

While the specifics of ACD reports may vary from one site to the next, the reports almost always (as far as we have seen) contain statistics on four categories:

- numbers of arrivals and abandonment
- average service times
- CSR utilization, and
- the distribution of delay in queue.

This is hardly surprising – it simply reflects the fact that call centers can be viewed, naturally and usefully, as queueing systems.

2 Data Analysis and Forecasting

The modelling and control of call centers must necessarily start with careful data analysis. For example, when used to model performance at time-period t, the simple Erlang C queueing model requires the estimation of a calling rate (λ_t) and a mean service time (μ_i^{-t}). Moreover, the performance of call centers in peak hours can be extremely sensitive to changes in these underlying parameters.

It follows that accurate estimation and forecasting of parameters are prerequisites for a consistent service level and an efficient operation. Furthermore, given the computer-mediated, data-intensive environment of modern call centers, one might imagine that highly developed estimation and forecasting methods would exist.

But in fact, though there is a vast literature on statistical inference and forecasting, surprisingly little has been devoted to stochastic processes, and much less to queueing models in general and call centers in particular. Indeed, the practice of statistics and time series analysis is still in its infancy in the world of call centers, and serious research efforts are required to bring it up to par with prevalent needs.

3 Types of Call Center Data

Call centers generate a great deal of data, which we divide into four categories: operational, marketing, human resources, and psychological.

Operational data reflect the physical process by which calls are handled. These data are typically collected by pieces of the telephone infrastructure such as IVRs and ACDs. They can be usefully organized in two, complementary fashions.

Operational *customer* data provide listings of every call handled by a site or network of call centers. Each record includes time-stamps for when the call arrived, when it entered service or abandoned, when it ended service, as well as other identifiers, such as who was the CSR and at which location the call was served.

Operational *agent* data provide a moment-by-moment history of the time each logged-in agent spent in various system states: available to take calls, handling a call, performing wrap-up work, and assorted unavailable states. These data allow one to deduce the numbers of agents working at any time. Often these records include identifiers of the calls being served and (with difficulty) can be matched to the operational customer data described above, for joint analysis.

Marketing or *business* data are gathered by a company's corporate information system. They may include records of the transactions that took place over the customer's entire history with the company, through call centers as well as through other channels. They may also capture information concerning the customer's current status at the business.

In theory, operational and marketing data can be seamlessly integrated via CTI software, which connects the telephone infrastructure with a company's customer databases. That is, given the existence of CTI, one might expect companies to record and analyze a full view of what happens to each call as it enters the system: marketing data concerning what happened during the service, together with operational data concerning how and when the service happened. In practice, however, the use of CTI appears, thus far, to be limited to facilitating the service process through "screen pops" which save CSRs time, not to the joint reporting of call data. Incompatibility between data storage schemes of (older) ACD and (newer) CTI systems may be the problem that prevents this integration from taking place.

Human resources data record the history and profile of agents. Typical data include information concerning employees' tenure at the company, what training they have received and when, and what types of call they are capable of handling. With one frequent exception, these data generally reside within the records of a company's human resources department. The exception is that of "skills" data, which define the types of calls that agents can handle. This information is needed, by the ACD (or those that manage it), to support skills-based routing.

Finally, *psychological* data are collected from surveys of customers, agents or managers. They record subjective perceptions of the service level and working environment.

Two additional sources of data are important to acknowledge. First, some companies record individual calls for legal needs (e.g., brokerage and insurance businesses) or training reasons. While potentially useful, we are not aware of any simple machinery that can extract these data for analysis (say, into a spreadsheet). A second source is subjective surveys in which call center managers report statistics that summarize their operations. These surveys can include both operational and marketing data, such as arrival and utilization rates, average handle times, and the average dollar value of a transaction. While they may facilitate rough benchmarking, these data should be handled with care. By their nature, they are biased and should *not* serve as a substitute, or even a proxy, for the operational and marketing data discussed above.

4 Types of Data Models

As in any statistical work, the analysis of call-center data can take a number of forms. We briefly make three sets of distinctions.

Descriptive, Explanatory and Theoretical Analysis: We first distinguish among descriptive, explanatory, and theoretical analysis. Each mode is important, and we briefly describe the three in turn.

Descriptive models organize and summarize the data being analyzed. The simplest of these are tables or histograms of parameters and performance. An example is a histogram of service duration by service type, or of customers' patience by customer type, or of waiting times for those ultimately served.

These can be contrasted with *theoretical models* that seek to test whether or not the phenomenon being observed conforms to various mathematical or statistical theories. Examples include the identification of an arrival process as a Poisson process or of service durations as being exponentially distributed.

In between descriptive and theoretical models fall *explanatory models.* These are often created in the context of regression and time series analysis. Explanatory models go beyond, say, histograms by identifying and capturing relationships in terms of explanatory variables. For example, average service times of calls may be systematically higher from 11am to 3pm and lower at other periods. At the same time, these models fall short of theoretical models in that there is no attempt to develop or test a formal, mathematical theory to explain the relationships.

Queueing models constitute theoretical models which mathematically define relationships among building blocks, for example arrivals and services, which we refer to here as *primitives.* Queueing analysis of a given model starts with assumptions concerning its primitives and culminates in properties of performance measures, such as the distribution of delay in queue or the abandonment rate. Validation of the model then amounts to a comparison of its primitives and performance measures – typically theoretical – against their analogs in a given call center – mostly empirical.

For example, as will be discussed in class, theoretical analysis of the $G/G/N$ queue gives rise to Kingman's law of congestion: in conventional heavy traffic, the waiting time of delayed customers is close to being exponentially distributed, with a calculable parameter. Empirical analysis of call centers operating in heavy traffic can then validate or refute Kingman's law. Refuting it would trigger theoretical research in order to identify alternative theoretical models, possibly in non-conventional heavy-traffic regime of Halfin-Whitt (the QED regime).

Estimation versus Prediction: We also distinguish between two closely related, but different, statistical tasks: estimation and prediction. *Estimation* concerns the use of existing (historical) data to make inferences about the parameter values of a statistical model. *Prediction* concerns the use of the estimated parameters to forecast the behavior of a sample outside of the original data set (used to make the estimate). Predictions are "noisier" than estimates, because, in addition to uncertainty concerning the estimated parameters, they contain additional sources of potential errors.

As an example, consider a simple model in which the arrival rate to a call center (each day from 9:00am–9:30am) is a linear function of the number of customers receiving a promotional mailing.

That is

$$\lambda_i = \alpha + \beta x_i + \varepsilon_i \ , \tag{1}$$

where λ_i is the arrival rate, x_i is the number of mailings, α and β are unknown constants, and the ε_i are *i.i.d.* normally-distributed noise terms with mean zero. Given n sample points (x_i, λ_i), one may use regression techniques (such as least squares) to produce parameter estimates $\hat{\alpha}$ and $\hat{\beta}$. There is uncertainty, however, regarding how closely these estimates match the true α and β. That is $\hat{\alpha}$ and $\hat{\beta}$ are random variables that are functions of the n *i.i.d.* samples, and given our estimated function

$$\hat{\lambda}_i = \hat{\alpha} + \hat{\beta} x_i \ , \tag{2}$$

the associated estimation error is distributed as

$$\lambda_i - \hat{\lambda}_i \ = \ (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) x_i \ .$$

Now suppose we are told the number of mailings that customers will receive on day $n+1$, and we are asked to predict what λ_{n+1} will be. Then we use $\hat{\lambda}_{n+1}$ to predict the $(n+1)$st arrival rate, and from (1)–(2) we see that the prediction error is distributed as

$$\lambda_{n+1} - \hat{\lambda}_{n+1} \ = \ (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) x_{n+1} + \varepsilon_{n+1} \ .$$

In particular, the ε_{n+1} term makes the prediction error larger than the estimation error that arises from the use of $\hat{\alpha}$ and $\hat{\beta}$.

5 Models for Operational Parameters

In the original article, here we review work devoted to primitives: arrivals, service times, abandonment (patience) and retrials. These, in turn, will be discussed and analyzed in class and assignments.

6 Future Work in Data Analysis and Forecasting

There has been recent progress in the analysis of call-center data. Call-by-call data from a small number of sites have been obtained and analyzed, and these limited results have proven to be fascinating. In some cases, such as the characterization of the arrival process and of the delay of arriving calls to the system, conventional assumptions and models of system performance have been upheld. In others, such as the characterization of the service-time distribution and of customer patience, the data have revealed fundamental, new views of the nature of the service process. Of course, these limited studies are only the beginning, and the effort to collect and analyze call-center data can and should be expanded in every dimension.

Perhaps the most pressing practical need is for improvements in the forecasting of arrival rates. For highly utilized call centers, more accurate, distributional forecasts are essential. While there exists some research that develops methods for estimating and predicting arrival rates, there is surely room for additional improvement to be made. However, further development of models for estimation and prediction will depend, in part, on access to richer data sets. We believe that much of the randomness of Poisson arrival *rates* may be explained by covariates that are not captured in currently available data.

Procedures for predicting waiting-times are also worth pursuing. Field-based studies that characterize the performance of different statistics and methods would also be of value. More broadly, there is need for the development of a wider range of descriptive models. While a characterization of arrival rates, abandonment from queue, and service times are essential for the management of call centers, they constitute only a part of the complete picture of what goes on. For example, there exist (self) service times and abandonment (commonly called "opt-out") behavior that arise from customer use of IVRs. Neither of these phenomena is likely to be the same as its CSR analogue. Similarly, sojourn times and abandonment from web-based services have not been examined in multi-media centers.

Parallel, descriptive studies are also needed to validate or refute the robustness of initial findings. For example, lognormal service times have been reported in two call centers, both of which are part of retail financial services companies. Perhaps the service-time distribution of catalogue retailers or help-desk operations have different characteristics. Similarly, one would like to test some finding that the waiting-time messages customers hear while tele-queueing promote, rather than discourage, abandonment.

It would also be interesting to put work on abandonment (Palm, Roberts, Kort, Mandelbaum with Sakov and Zeltyn) in perspective. These studies provide empirical and exploratory models for (im)patience on the phone in Sweden in the 40's, France in the late 70's, the U.S. in the early 80's, and Israel in the late 90's. A systematic comparison of patience across countries, for current phone services, should be a worthy, interesting undertaking.

There is the opportunity to further develop and extend the scope of explanatory models. Indeed, given the high levels of system utilization in the QED regime, a small percentage error in the forecast of the offered load can lead to significant, unanticipated changes in system performance. In particular, the state of the art in forecasting call volumes is still rudimentary. Similarly, the fact that service times are lognormally distributed enables the use of standard parametric techniques to understand the effect of covariates on the (normally distributed) natural log of service times.

In well-run QED call centers, only a small fraction of the customers abandon (around 1-3%), hence about 97% of the (millions of) observations are censored. Based on such figures, one can hardly expect any reasonable estimate of the whole patience distribution, non-parametrically at least. Fortunately, however, theoretical analysis suggests that only the behavior of impatience near the origin is of relevance, and this is observable and analyzable.

Indeed, call-center data are challenging the state-of-the-art of statistics, and new statistical techniques seem to be needed to support their analysis. Two examples are the accurate non-parametric estimation of hazard rates, with corresponding confidence intervals, and the survival analysis of tens of thousands, or even millions, of observations, possibly correlated and highly censored.

Last but certainly not least, a broader goal should be, in fact, the analysis of *integrated* operational, marketing, human resources, and psychological data. That is, the analysis of these integrated data is essential if one is to understand and quantify the role of operational service quality as a driver for business success.

7 A Vision: Central Repository for Call-Center Data and Expertise

A prerequisite for understanding the financial effects of operational decisions is the ability to analyze an integrated data set that includes both operational (ACD) and marketing / business (customer information systems) data. With this information, one can attempt to tease out the longer-term, financial effects of operational policies.

Our experience has been that both types of data are very difficult to access, however. One reason for this is technical. Only recently have the manufacturers of telephone equipment given customers something of an "off the shelf" ability to capture, store, and retrieve detailed, call-by-call data. Similarly, the integration of these operational data with the business data captured in customer information systems is only now becoming widely available. Another reason stems from confidentiality concerns; companies are rightly wary of releasing customer information. Once managers recognize the great untapped value of these data, we believe they will employ mechanisms for preserving confidentiality in order to reap the benefit.

Ultimately, we envision a data-repository that is continuously fed by many call centers of varying types. The collected data would be continuously and automatically analyzed, from both operations and marketing perspectives. Then the data would be both archived and fed back to the originating call centers, who would use it (through visualization tools) to support ongoing operations, as well as tactical and strategic goals.

Little imagination is required for appreciating the value of such a data-base. As a start, its developer could become a benchmark that sets industry standards, as far as customer-service quality and call-center efficiency are concerned. As already mentioned, such a data-base would enable the identification of success-drivers of call-center business transaction.

A Glossary of Call-Center Acronyms

Acronym	Description	Definition
ACD	automatic call distributor	p. ??
ANI	automatic number identification	p. ??
ASA	average speed of answer	p. ??
CRM	customer relationship management	p. ??
CSR	customer service representative	p. ??
CTI	computer-telephony integration	p. ??
DNIS	dialed number identification service	p. ??
IVR	interactive voice response unit (also called VRU)	p. ??
PABX	private automatic branch exchange (also called PBX)	p. ??
PBX	private automatic branch exchange (also called PABX)	p. ??
PSTN	public switched telephone network	p. ??
QED	Quality and Efficiency Driven (operational regime)	p. ??
TSF	telephone service factor (also called the 'service level')	p. ??
VRU	interactive voice response unit (also called IVR)	p. ??
WFM	workforce management	p. ??