Stochastic Data Analysis and Modeling of a Telephone Call Center

M. S. Shafae, M. H. Elwany, M. N. Fors and Y. Abouelseoud
Production Engineering Department
Faculty of Engineering
Alexandria University
Alexandria 21544, Egypt

Abstract

In this paper, stochastic data models for the arrival process of a call center are investigated. Comprehensive analysis of real-life call center data revealed several problems with modeling arrival processes. The high uncertainty in the arrival counts during different time intervals and the strong correlation between arrivals counts in consecutive time periods show that using popular models are rather insufficient. A stochastic model is thus developed, for the arrival process, which successfully handles the above-mentioned difficulties. Arguments are provided for the suitability of the proposed model to characterize the arrival counts to model real-data obtained from large call center.

Keywords
Call center, Doubly stochastic processes, Non-homogeneous Poisson model, and Parameter estimation

1. Introduction
Telephone inbound call centers could be described as a typical queuing system in which the customers are callers seeking telephone-based services presented by the call center agents, and they are queued in tele-queues waiting to be served. One of the key management decisions, in this setting, is how to manage the agents to serve the arriving calls in an optimal manner; a few percent saving in agents' salaries means several million dollars because inbound call centers are highly labor-intensive, with the cost of agents typically comprising 60–80% of the overall operating budget (Aksin, Z., Armony, M., and Mehrotra, V. 2007). This could be done by building staff schedules that satisfy certain performance measures achieving the least possible agents cost. A key input to build the agents schedules is the staffing decisions (how many agents to have in the center for each period of the scheduling horizon). These numbers of agents are, naturally, determined according to, the demand of the acquired services from the call center; that is, the arrival counts. The arrival counts, in fact, show a lucid nature where they experience significant variations with time during the same day, day-to-day variations, and different seasonality patterns over a time scale of weeks and months. The ignorance of such a fact about the calls arrival patterns, in making the staffing decisions, results in flat staff levels over the time. Pursuant to this, the call center performance would be affected negatively due to any variation in those call arrivals especially during peak load periods where the customers would encounter significant delays and might abandon without acquiring any service. Hence, developing accurate arrival data models – accounting for that highly stochastic nature of arrivals – is an essential requirement to make accurate staffing decisions without understaffing leading to unsatisfactory performance or even overstaffing leading to high agents' costs.

1.1 Modeling Approaches
The analysis and modeling of call center input data could be performed using three different approaches (Koole, G. and Mandelbaum, A. 2002). *Descriptive models* provide summaries of the empirical data obtained from the real system in the form of tables and histograms such as a histogram of total daily number of arrived calls as the call arrivals fluid-like model presented in (Mandelbaum, A., Sakov, A., and Zeltyn, S. 2002). *Explanatory models* utilize regression and time series analysis to determine and describe the desired parameters using explanatory variables; for example, a type of first-order autoregressive structure is suggested in (Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. 2005) for the random daily effects influencing the daily call volumes. *Theoretical models* provide a mathematical representation to fit the empirical data using theoretical statistical distributions and to sample from them. The arrival counts to a call center, for example, are modeled using a Poisson mixture model in (Jongbloed, G. and Koole, G. 2001).

1.2 Modeling Issues

In spite of the high capabilities of modern computerized data collection systems used in call centers, the construction of accurate arrival data models faces, in practice, a wide range of real world problems which complicate this task considerably. The level of detail, at which the operational data is collected, is limited to aggregate data parameters over fixed short time intervals varying between fifteen to sixty minutes, hence, the call-by-call data is not available. The available data, in particular, is the aggregated number of arrived calls to the center in each time interval. The main issue with such kind of data is that standard parameter estimation methods could not be used, to estimate the required statistical models, due to the unavailability of records for each individual call (Henderson 2003). Consequently, devising a suitable data modeling method that could deal with this issue is not an easy task.

In addition to the lack of call-by-call data issue, the stochastic nature of arrival counts exhibits three main properties that characterize the arrival process in telephone inbound call centers (Avramidis, A., Deslauriers, A., and L'Ecuyer, P. 2004):

- **A.** The call volumes vary significantly over the time (*arrival rate uncertainty*).
- **B.** The variance of the arrival volumes is much greater than their means (*data over-dispersion*).
- **C.** The dependence between arrival volumes in the successive periods of a day is very strong (*strong positive correlation*).

In this paper, both descriptive and theoretical modeling approaches are used to study and develop suitable stochastic data models for the arrival counts to inbound call centers. In the model development, the issues mentioned above are considered in order to create data models that are capable of representing real-life arrival data effectively.

2. Statistical Models for the Arrival Process

In literature, several statistical models have been considered to deal with the modeling issues discussed above as the following models.

2.1 The Poisson Process

According to the Palm-Khintchine theorem, the counting process of events occurring from a large number of independent sources – where anyone of the sources contribution to the total number of events is small – behaves asymptotically as a Poisson process (Heyman, D. and Sobel, M. 2004). In call centers context, when a large number of independent customers, each of whom has a comparatively small calling probability, are possible callers of a call center, the Palm-Khintchine theorem provides justification for using the Poisson process to model the arrival process to that call center. The Poisson process is the most elementary random process used in modeling the arrivals of customers to call centers. The standard definition of this process as it appeared in (Ross, Introduction to Probability Models 2007) is as follows.

Definition 1. The counting process $\{N(t), t \geq 0\}$ – where $N(t)$ is the total number of "events" that occur by time t – is said to be a Poisson process having rate λ, $\lambda > 0$, if the following assumptions hold:

- **i.** $N(0) = 0$.
- **ii.** The process has independent increments. That is, for all s, t, v, and $u \geq 0$, $N(t) - N(s)$ and $N(u) - N(v)$ are independent for any non-overlapping intervals $(s, t]$ and $(v, u]$.
- **iii.** The process has stationary increments. The number of events in any interval of length s is Poisson distributed with mean λs. That is, for all s, $t \geq 0$

$$P\{N(t+s) - N(t) = n\} = e^{-\lambda s}\frac{(\lambda s)^n}{n!}, \qquad where\ \ n = 0, 1, \ldots \tag{1}$$

Property ***A*** of the arrival process in a call center contradicts the assumption of the standard Poisson model that the arrival process has stationary increments with the same arrival rate as the arrival rate varies considerably during different daily periods. Moreover, the arrival rate – in the presence of property ***B*** – could not be modeled using a Poisson distribution as it assumes that the mean and variance are equal while the data experiences considerable over-dispersion. Additionally, the independent increments assumption is inconsistent with property ***C*** where the strong positive correlation refutes the independence between the non-overlapping periods. Consequently, these properties render the standard Poisson process model inadequate.

2.2 The Non-Homogeneous Poisson Process

In view of property ***A***, the problem of uncertain arrivals during different periods of the day could be solved by considering the arrival process during separate periods, for instance one-hour periods, as a Poisson process but with a rate varying from a time-period to another. This time sampling Poisson process generates a non-homogeneous process that, by definition, allows to model time-dependent arrivals (Ross, Introduction to Probability Models 2007).

Definition 2. The counting process $\{N(t), t \geq 0\}$ – where $N(t)$ is the total number of "events" that occur by time t – is said to be a *non-homogeneous Poisson process* (*NHPP*) with intensity function for time varying arrival rate $\lambda(t)$, $\lambda(t) > 0$, if the following assumptions hold:

- **i.** $N(0) = 0$.
- **ii.** $\{N(t), t \geq 0\}$ has independent increments. That is, for all s, t, v, and $u \geq 0$, $N(t) - N(s)$ and $N(u) - N(v)$ are independent for any non overlapping intervals $(s, t]$ and $(v, u]$.
- **iii.** The process has stationary increments. The number of events in any interval of length s is Poisson distributed with mean $s\lambda(t)$. That is, for all s, $t \geq 0$

$$P\{N(t+s) - N(t) = n\} = e^{-s\lambda(t)} \frac{\left(s\lambda(t)\right)^n}{n!}, \qquad where\ n = 0, 1, 2, \ldots \tag{2}$$

The main problem here is how to estimate the time-varying arrival rate. Jongbloed and Koole in (Henderson 2003) assume that the arrival rate function is piecewise constant over the subsequent time-intervals of the day and that could be estimated from the data, and show that this estimator could be a consistent estimator of the original arrival-rate function by performing an asymptotic analysis of this method. In (Massey, W., Parker, G., and Whitt, W. 1996), a piecewise linear rate function is proposed for the arrival rate and then several ways are investigated for the parameters estimation of this model such as ordinary least squares (OLS), iterative weighted least squares (IWLS) and maximum likelihood (ML) methods. A piecewise polynomial approximation is suggested in (Kao, E. and Chang, S. 1988) to represent the rate function using maximum likelihood estimators. Time-inhomogeneous Poisson processes successfully address the problem of daily period-to-period variability; the other levels of variability (day-to-day variability as well as weekly and monthly variability) could be handled by providing a separate model for each period in which the arrival patterns are consistent by means of clustering various periods.

2.3 Doubly Stochastic Poisson Model in Call Centers

According to the above discussion, ignoring any of the calls arrival process main properties will produce inaccurate queuing/simulation models that are to be used later in solving the staffing problem which may bring the validity of this work into question. If the variability in the real arrival process, for example, is higher than that obtained by the standard Poisson process, then the estimated service level is lower than it would be otherwise (Steckley, S., Henderson, S., and Mehrotra, V. 2005). In order to consider the aforementioned data properties in the data generation model, the conditional Poisson process is used to model the call arrival counts.

Definition 3. Let $\{\boldsymbol{X}(t), t \geq 0\}$ be a counting process and there is a positive random variable Λ such that conditional on $\Lambda = \lambda$ the counting process is a Poisson process with rate λ. That counting process is called a conditional Poisson process representing the doubly stochastic arrival model. This model stands for the call center context when the following hold:

- **i.** If the random vector of arrival counts is $\mathbf{X} = (X_1, X_2 \ldots, X_k)$, where X_i is the number of call arrivals in period i and k is the number of daily periods, then the random arrival counts X_i follow Poisson distributions with probability mass function

$$P(X_i = x) = e^{-\Lambda_i} \frac{\Lambda_i^x}{x!} \tag{3}$$

- **ii.** The rate Λ_i of a Poisson random variable X_i is a random variable generated randomly from a period-dependent distribution for Λ on $(0, \infty)$. This accounts for the over -dispersion problem (property ***C***) in the standard Poison model (Jongbloed, G. and Koole, G. 2001).
- **iii.** The Poisson random variable X_i with a rate Λ_i is generated separately for each daily period from a separate standard Poisson process and this accounts for the high level of variability in arrival counts data (property ***A***).

There are many ways to estimate the rate of the doubly stochastic Poisson process such as those proposed in (Jongbloed, G. and Koole, G. 2001, Avramidis, A., Deslauriers, A., and L'Ecuyer, P. 2004). In this study a new model for the rate function – daily proportion-based arrival rate – is developed, studied and compared to the gamma dependant arrival rate proposed by (Avramidis, A., Deslauriers, A., and L'Ecuyer, P. 2004).

According to (Avramidis, A., Deslauriers, A., and L'Ecuyer, P. 2004), the arrival rates Λ_i are modeled as dependent random variables randomized by a common gamma variable that accounts for the correlation between the number of arrivals in the subsequent periods (property ***C***) where

$$\Lambda_i = W\lambda_i, \qquad W \sim Gamma(\gamma, 1)$$

The proposed function of arrival rates could result in a *negative multinomial distribution* for the vector **X** with parameters ($\gamma, \lambda_1, \lambda_2 \ldots \lambda_k$) where probability mass function is given by

$$f(\mathbf{x}) = \frac{\Gamma\left(\gamma + \sum_{i=1}^{k} x_i\right)}{\Gamma(\gamma)\prod_{i=1}^{k} x_i!}\left(\frac{1}{1+\sum_{j=1}^{k}\lambda_j}\right)^{\gamma}\prod_{i=1}^{k}\left(\frac{\lambda_i}{1+\sum_{j=1}^{k}\lambda_j}\right)^{x_i} \tag{4}$$

The parameters of the *negative multinomial distribution* could be estimated by the maximum likelihood estimation method (MLE). If the vector $\left[\mathbf{X}_j = \left(X_{1,j}, X_{2,j}, \ldots, X_{k,j}\right)\right]_{j=1}^{n}$ represents the data set of arrival counts observations – where n is the number of similar days in the data set and k is the number of daily periods – then the maximum likelihood estimators (MLEs) could be obtained according to the following estimation algorithm:

$$Daily\ Call\ Volume\ Y_j = \sum_{i=1}^{k} X_{i,j}, \qquad for\ j = 1, 2, \ldots, n \tag{5}$$

$$F_l = \frac{1}{n}\sum_{j=1}^{n} I\{Y_j \geq l\}, for\ l = 1, 2, \ldots, M \qquad where\ M = max(Y_j) \tag{6}$$

$$\sum_{l=1}^{M} (\hat{\gamma} + l - 1)^{-1} F_l = log\left(1 + \frac{1}{n\hat{\gamma}}\sum_{j=1}^{n} Y_j\right) \tag{7}$$

$$\hat{\lambda}_i = \frac{\sum_{j=1}^{n} X_{i,j}}{n\hat{\gamma}}, \qquad for\ i = 1, 2, \ldots, k \tag{8}$$

$I\{\xi\}$ is the indicator function which is equal to one if ξ is true and zero otherwise. Y_j and F_l are calculated directly from the observed data. $\hat{\gamma}$-value is calculated by solving equation (7) above numerically using Newton–Raphson method and then $\hat{\lambda}_i$ is obtained for each period. This estimation process is repeated, separately, for each day within the week due to the high amount of day-to-day variability (e.g., $\hat{\lambda}_1$ on Saturdays is different from $\hat{\lambda}_1$ on Sundays and so on).

3. Daily Proportion-Based Arrival Rate Model

The model of *Gamma-dependent arrival rate* succeeds in finding remedies for all problems of call arrivals data related to its three properties. Reaching a better model validity and the reduction of required number of parameters to be estimated, however, are the main motive to propose a new model for arrival counts. In the new model, daily proportion-based arrival rate model, the arrival rate in each daily time period is modeled as a proportion P_i of the total daily volume of arrivals Y_j where

$$\Lambda_i = P_i.Y_j, \qquad where\ i = 1, 2, \ldots, k\ and\ j = 1, 2, \ldots, 7 \tag{9}$$

3.1 Total Daily Arrivals Model

Proposition 1. The total number of daily call volume arrived to call center Y_j is normally distributed with probability density function

$$f_{\mathrm{Y}} = \mathrm{N}(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}.e^{\frac{-(y-\mu)^2}{2\sigma^2}} \qquad where -\infty < y < \infty, -\infty < \mu < \infty, and\ \sigma > 0 \tag{10}$$

Proof. Given that $Y_j = \sum_{i=1}^{k} X_i$, the proof of *proposition* 1 follows directly from the central limit theorem. ■

The central limit theorem states that the sum of a large number of random variables has a distribution that is approximately normal. It also explains the remarkable fact that the empirical frequency of so many natural populations exhibit bell shaped (that is, normal) curves (Ross, A First Course in Probability 1998). Andrews in (Andrews 1991) extended the results of the central theorem to be applied to dependent non-identically distributed random variables. This is the case with call center total daily volume of arrivals which is a sum of dependent, but non-identically, distributed random variables (i.e., call arrivals in each daily period). Hence, Y_j is assumed to be normally distributed according to the central limit theorem. Furthermore, this assumption is supported empirically from the observed data by testing the goodness of fit of the total daily volume data to the normal distribution. Kolmogorov-Smirnov test is used to assess the goodness of fit revealing high p-values which means it yields a very good fit supporting the theoretical assumption of normally distributed total daily call volumes.

Due to the high level of variability in total daily call volume as described previously, a separate daily arrival normal model is estimated for each day of the week (i.e., there may be seven separate models for total daily arrivals). In order to verify the need to these separate models, the statistical significance of the difference between different samples of the daily call arrivals for different week days are studied. This could be done using the two samples t-test of hypothesis. The hypothesis testing on the difference between the means μ_1 and μ_2 of two normal populations is considered. Suppose that we are interested in testing whether the difference in means $\mu_1 - \mu_2$ is equal to a specified value δ_0. Thus, the null hypothesis will be stated as

$$H_0: \mu_1 - \mu_2 = \delta_0 \tag{13}$$

Obviously, in many cases, $\delta_0 = 0$ is specified to test the equality of two means (i.e., H_0: $\mu_1 = \mu_2$). Suppose that the alternative hypothesis is

$$H_1: \mu_1 - \mu_2 \neq \delta_0 \tag{14}$$

Now, a sample value of $\bar{x}_1 - \bar{x}_2$ that is considerably different from δ_0 is evidence that H_1 is true.

Remark 1. Performing this test on different pairs of week days yields an important result. If there are pairs that fail to reject the null hypothesis, they could be dealt as clusters of days that have the same total daily arrivals distribution which could be used to reduce the number of estimated parameters and distributions.

3.2 The Proportions Model

As for the proportion value P_i, it is assumed to be independent from the daily volume itself, and thus, it does not differ significantly from day-to-day during the week. This assumption could be verified empirically by showing that there is no significant statistical difference between different samples of the same daily period proportions for different week days. A two samples t-test could be used again to assess the statistical significance of that amount of difference. Pursuant to this assumption, the daily proportions of different daily periods could be defined by a single vector $\mathbf{P} = (P_1, P_2 \ldots P_k)$. P_i is the ratio between period *i* arrival rate and the total arrival rate of the day. Then,

$$P_i = \frac{\Lambda_i}{\sum_{i=1}^{k} \Lambda_i} \tag{15}$$

Proposition 2.

i. The daily proportions vector **P** has a *Dirichlet distribution* with parameters (α_1, $\alpha_2 \ldots \alpha_k$) with probability mass function

$$f(p_1, p_2 \ldots p_{k-1}) = \mathrm{D}(\alpha_1, \alpha_2 \ldots \alpha_k) = \frac{1}{c} . \prod_{i=1}^{k} p_i^{\alpha_i - 1} \tag{16}$$

Over the k-dimensional simplex $\boldsymbol{S_k}$ defined by inequalities $P_i > 0$ ($i = 1, 2 \ldots k\text{-}1$), $P_k = 1 - \sum_{i=1}^{k-1} P_i$ and $\sum_{i=1}^{k} P_i = 1$. Here, *c* (normalization constant) is the multinomial beta function with the following expression

$$c = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}, \qquad where\ \alpha_0 = \sum_{i=1}^{k} \alpha_i \tag{17}$$

ii. The marginal probability distribution of P_i is a *Beta distribution* defined on the interval (0, 1) having parameters (α_i, β_i) with probability mass function

$$f_{P_i} = \begin{cases} \dfrac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i).\Gamma(\beta_i)} . p_i^{\alpha_i - 1} . (1 - p_i)^{\beta_i - 1} & 0 < p_i < 1, \quad where\ \beta_i = \alpha_0 - \alpha_i \\ 0 & otherwise \end{cases} \tag{18}$$

Proof. Assuming that the arrival rates Λ_1, Λ_2...Λ_k are independent *Gamma random variables* with parameters $\alpha_i > 0$ respectively. The joint density function of the Λ_i's is

$$f(\lambda_1, \lambda_2, \dots, \lambda_k) = c. \prod_{i=1}^{k} \lambda_i^{\alpha_i - 1} . e^{-\sum_{i=1}^{k} \lambda_i} \tag{19}$$

Consider the transformation $Z = \sum_{i=1}^{k} \Lambda_i$, $P_i = \Lambda_i / Z$ $(i \leq k)$, which has a reverse transformation $\Lambda_i = Z.P_i$, and $\Lambda_k = Z\left(1 - \sum_{i=1}^{k-1} P_i\right)$. The Jacobian of the transformation is Z^k. Thus, the joint density function of (Z, P_1, P_2...P_k) is

$$f(z, p_1, p_2, \dots, p_{k-1}) = c. \prod_{i=1}^{k} p_i^{\alpha_i - 1} . z^{\sum_{i=1}^{k} \alpha_i - 1} . e^{-z} \tag{20}$$

$$Then, \qquad f(p_1, p_2 \dots p_{k-1}) = \mathrm{D}(\alpha_1, \alpha_2 \dots \alpha_k)\ and\ P_k = 1 - \sum_{i=1}^{k-1} P_i \tag{21}$$

and this proves result (i) of *proposition 2*. Result (ii) is a known result derived from the probability mass function of the *Dirichlet distribution*, which is mentioned and proven in (Ferguson 1973). Moreover, the assumption that the arrival rates Λ_1, Λ_2...Λ_k are *Gamma random variables* is verified empirically by testing the goodness of fit of arrival rates to *Gamma distribution*; in (Jongbloed, G. and Koole, G. 2001) the same assumption is employed to develop also a model of arrivals to call centers. ■

Remark2. Through the analysis of the observed data, the validity of result (ii) has been recognized in an attempt to find a suitable distribution to fit the proportions of daily periods. Applying goodness of fit tests for the *Beta distribution* supported the plausibility of the assumption. The suitability of the *Beta* distribution also follows from the fact that the beta distribution is used in modeling continuous random variables which take on values that lie between 0 and 1, such as proportions and percentages. The fact that *Beta distribution* is an appropriate model for the individual period's proportions motivated the use of its multivariate generalization, *Dirichlet distribution*, to achieve the correlated tuple of call arrivals.

Definition 4: Based on the above discussion, the number of arrival counts to a call center at a certain time period *i* $\{X_i\}$ is a *Poisson* random variable with rate Λ_i where the following holds:

- **i.** $\mathbf{\Lambda} = (\Lambda_1, \Lambda_2 \dots \Lambda_k)$, where $\Lambda_i = P_i . Y_j$
- **ii.** $\mathbf{P} = (P_1, P_2 \dots P_{k-1})$ where $\mathbf{P} \sim \mathrm{D}(\alpha_1, \alpha_2 \dots \alpha_k)\ and\ P_k = 1 - \sum_{i=1}^{k-1} P_i$
- **iii.** $\mathbf{Y} = (Y_1, Y_2 \dots Y_7)$ where $Y_i \sim \mathrm{N}(\mu_j, \sigma_j)$

3.3 Parameter Estimation

In order to estimate the parameters of the new model, the parameters of both the *Normal distribution* and the *Dirichlet distribution* need to be estimated. Let $\left[\mathbf{Y_q} = (Y_{1,q}, Y_{2,q}, \dots, Y_{7,q})\right]_{q=1}^{m}$ represents the data set of total daily arrivals observations – where m is the number of weeks in the data set – then both parameters of the normal distribution μ_j and σ_j for each week day could be estimated directly from the observed data set through using the following couple of formulas

$$\mu_j = \bar{Y}_m = \frac{1}{m} \sum_{q=1}^{m} Y_{j,q} , \qquad \text{where j} = 1, 2, \dots, 7 \tag{22}$$

$$\sigma_j = S_m^2 = \frac{1}{m-1} \sum_{q=1}^{m} \left(Y_{j,q} - \mu_j\right)^2 \tag{23}$$

Unlike the *Normal* distribution, the *Dirichlet distribution* is defined with parameters that do not correspond directly to either the mean or variance of the distribution. Rather, the mean and variance of the *Dirichlet distribution* are functions of its parameters α_i. Thus, the parameters of the *Dirichlet distribution* are represented by the maximum likelihood estimators MLEs. If the vector $\left[\mathbf{P}_z = \left(P_{1,r}, P_{2,r}, \dots, P_{k,r}\right)\right]_{r=1}^{V}$ represents the data set of observed daily period proportions – where V is the number of days in the data set and k is the number of daily periods – then the parameters for a Dirichlet distribution could be estimated by maximizing the log-likelihood function of the data, which is given by:

$$F(\alpha) = log \prod_i \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_{ik}^{\alpha_k - 1} = N\left(log\Gamma\left(\sum_k \alpha_k\right) - \sum_k log\Gamma(\alpha_k) + \sum_k (\alpha_k - 1) log\ \bar{p}_k\right) \quad (24)$$

where $log\ \bar{p}_k = \frac{1}{N}\sum_i log\ p_{ik}$. Newton-Raphson method is traditionally used to find the unknown parameters.

3.4 Random Number Generation

Regarding the generation of the random arrival rates, the total daily arrivals is generated using a generator of Normal random numbers. After that, the daily period proportions of each daily period will be generated from the *Dirichlet distribution* using either the *marginal beta distribution* utilizing the result (ii) of *proposition* 2 so that $P_i \sim B(\alpha_i, \beta_i)$ or using the Gamma distribution by finding $\Lambda_i \sim Gamma(\alpha_i, 1)$, then $P_i = \Lambda_i / \sum_{i=1}^{k-1} \Lambda_i$ and $P_k = 1 - \sum_{i=1}^{k-1} P_i$. The rate of the arrivals in each period is the product of the two generated random variables. Finally, a Poisson arrival count is generated using that rate of arrivals.

4. Case Study (Egyptian Phone Directory Call Center)

The studied data is obtained from the *telephone directory call center* of the national Terrestrial Communications Network Company in Egypt. This call center is a single-skill inbound call center that operates in two different locations, Cairo and Alexandria. The operations in both centers are similar and independent from each other so this study focuses only on Cairo site. The call center provides *twenty-four hours a day/seven days a week* support to its customers. The available data to this work – three months of call center operations – is aggregated as explained above over sixty-minute time periods. That is, the twenty-four hours are partitioned into twenty-four periods in each of which the model parameters remains constant and then changes when the new period starts on each of seven week days.

4.1 Prelude Data Analysis

The analysis of the call center arrival process reveals the existence of the three main properties of arrivals data: high level of uncertainty, over-dispersion, and strong correlation. These properties are investigated here in details for the data obtained from the Egyptian Phone Directory call center.

4.1.1 High Uncertainty Property

The uncertain arrival pattern is a key property in the call center arrivals and presents several problems in modeling call centers. Considering this issue, we find different levels of variability: monthly, weekly, daily, and periodically within the same day. Considering the monthly arrival volumes in year 2010 reveals high seasonality in the six month-period between May and November with a mean of 3,885,064 arrived calls per month, a standard deviation of 396,789 calls, and the coefficient of variation of 0.102. Analyzing the arrivals data at the weekly level shows a considerable variation from week-to-week. The degree of variability, at the weekly level, also differs considerably from period to period along the year. In May-July period, the weekly arrivals coefficient of variation is 1.5 times that of February-April period. The daily call volumes vary also from day-to-day over the course of a week. This variation may extend to find that the daily arrivals pattern differ from week to week. In eight weeks period, for example, the variation coefficient of the daily call volumes over the course of each week varies from 0.19 to 0.25. Call volumes exhibit also a strong seasonality pattern over the course of a day with coefficient of variation varying between 0.22 and 0.9 over the different days. In addition to the strong period-to-period variability during the same day, call arrivals in a certain period experiences a significant variation from day to day and that is concluded from the significant variation of the coefficient of variation of the different days.

4.1.2 Over-dispersion Property

The arrival volumes to the call center have a considerable over-dispersion relative to the Poisson distribution. The variance of the number of arrival counts is much greater than the mean of the same arrival counts. This property is

verified by studying the mean and variances of the twenty-four hour periods arrivals over the course of a month, and the mentioned over-dispersion exists in all of them. The least ration between the variance and mean during that month was 18 not one at all.

4.1.3 Strong Correlation Property

The careful analysis of call arrivals data shows that there is a strong dependence between the arrival volumes in the subsequent time periods during the day. One reason for this correlation may be due to the retrial phenomenon that occurs when calls – that abandoned in a period – retry to call again but in the subsequent period. This property is verified by calculating the *Pearson Correlation Coefficient* ***r*** for the successive daily-period pairs over the course of a month. The calculated values show a strong positive correlation between the subsequent periods, with values as low as 0.27 and as high as 0.98; 56% of the twenty-four pairs possess correlation coefficient greater than (0.7).

4.2 Model Estimation

According to remark 1, the test of hypothesis on the significance of the statistical difference between total daily volumes showed that the week days could be divided to four clusters in each of which the parameters of the normally distributed total daily arrivals are constant. This could be concluded from the statistical comparisons, between the total daily arrivals of week days' different pairs, made using two samples t-test. Knowing that the pair failing to reject the null hypothesis obtains high p-value greater than the significance level $\alpha = 0.05$, then Sunday, Monday, and Tuesday form a separate cluster as the p-values of different pairs of these days vary between 0.45 and 0.7. Additionally, Wednesday and Thursday give another cluster with P-value of 0.56; while each of Friday and Saturday form separate clusters as they show very low P-values for all pairs with other weekdays. Likewise, refereeing to remark 2, all week days except Friday have the same structure of the daily period proportions. Friday shows a different structure of the proportions.

4.2.1 Gamma Dependent Random Arrival Rate Model Estimation

Following the parameter estimation algorithm obtained previously, the values of γ for each cluster is estimated separately from equation (7), and λ is also estimated for each period of the twenty-four daily periods in the four clusters from equation (8). This results in estimating twenty-five parameters for each cluster and totaling one hundred parameters. A sample of estimated parameters for the first cluster is shown in Table 1.

Table 1: Sample of estimated parameters for the first cluster

γ	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
107.4	29.314	20.788	11.017	6.0346	3.7631	2.2377	2.5676	4.9082	16.26	46.644

4.2.2 Daily Proportion-Based Arrival Rate Model Estimation

In this model, there are actually two models to be estimated, the model of total daily arrivals and the model of period proportion. In order to estimate the total daily arrivals model, the mean and standard deviation for each cluster is calculated from equations (22 & 23) resulting in eight parameters (e.g., for cluster 1, μ = 127,249 and σ = 12,663). The parameters of the period proportion model are estimated according to the algorithm mentioned in section 3.3 and the number of estimated parameters is twenty-four parameters for all weekdays except Friday, shown in, and other twenty-four parameters for Friday only. The total number of estimated parameters here is fifty-six parameters.

Table 2: Sample of estimated parameters α_i for weekdays Saturday – Thursday

α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}
6.69806	**4.78286**	**2.63384**	**1.46477**	**0.89055**	**0.51233**	**0.58453**	**1.11749**	**3.55878**	**10.1227**	**17.7973**

4.3 Models Validation

The main properties of the studied/developed models are investigated and compared to the real-data main properties to quantify how far they are valid and represents the real data. This is done by developing simulation models and running them using the developed data models to produce simulated data sets similar to the real one. The following results are obtained from the analysis of the first cluster data and the same results are the same for all other clusters. The first thing to be compared is the estimated and real means of arrival counts during the different periods. The means comparison shows, as shown in Figure 1, that the developed models succeed in producing accurate estimates for the arrival counts; the difference from the mean is, nearly, neglected. This result is also supported by measuring the statistical difference using the test of hypothesis of the means equality which results in high P-values for both models. The mean comparison does not show tangible difference between both models.

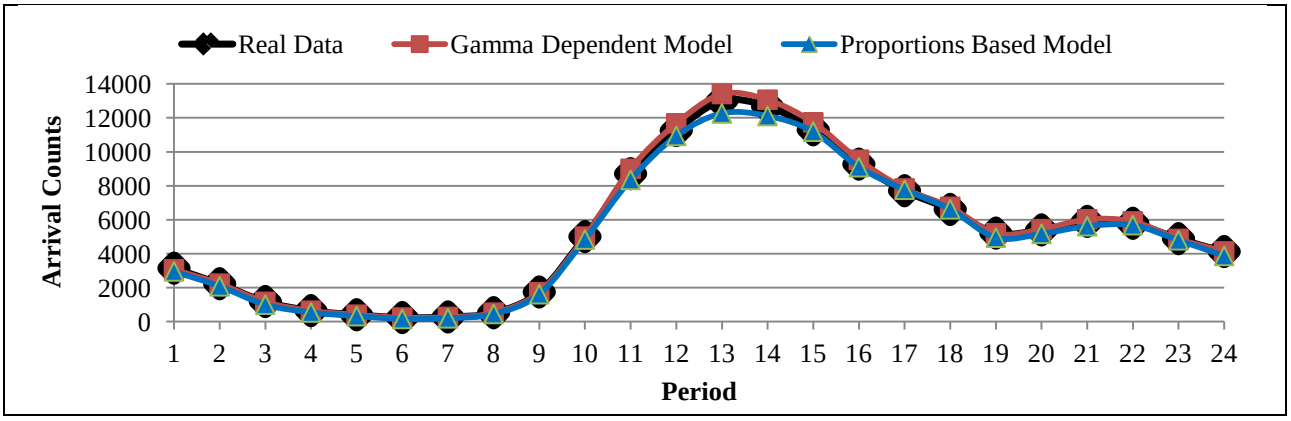

Figure 1: Comparison of Estimated and Real Means of Call Volumes/Period

Comparing the coefficient of variation (CV), as shown in Figure 2, reveals that both models obtain, nearly, similar constant CV for all periods. The estimated CV for both models match that of real data in the periods that experience high arrival counts (periods between 10:00 am and 09:00 pm) but fail to do that in the less-dense periods.

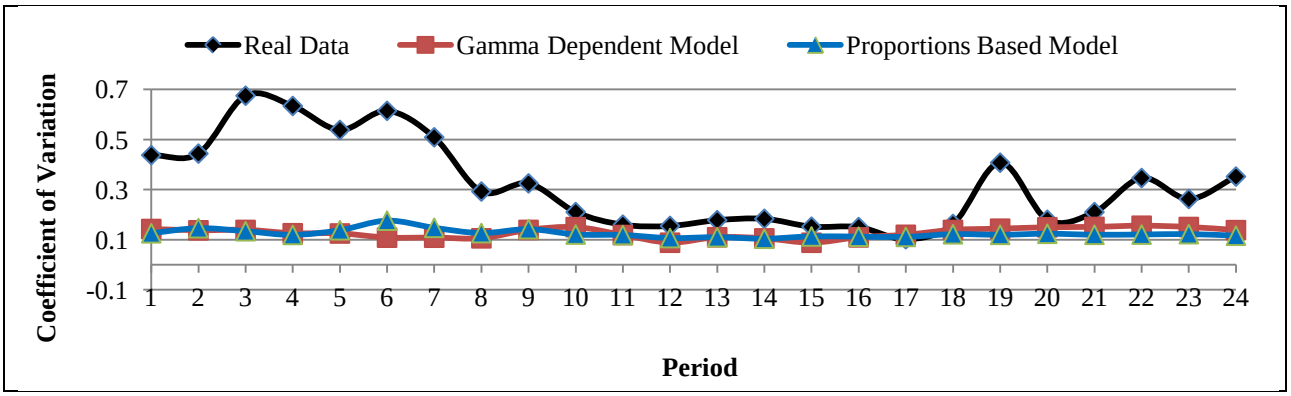

Figure 2: Coefficient of Variation Comparison

As shown in Figure 3, the developed models produce, nearly, similar correlation structures that overestimate the correlation values in some periods and underestimate it in the other periods. The most important result, however, that there is a correlation structure that reflects the effect of each period on its successor. The other discussed models in literature almost neglect that effect and assume zero-correlation.

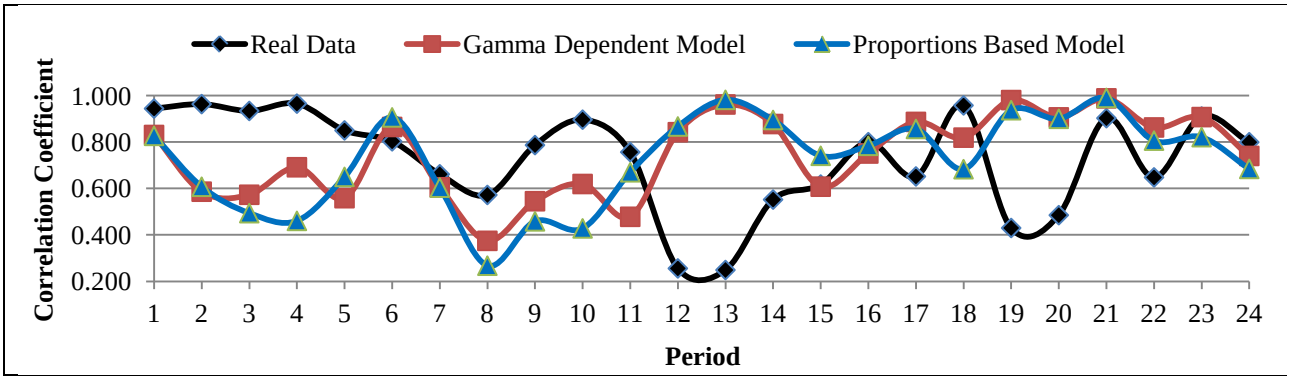

Figure 3: Sample Correlation Comparison

5. Conclusions

Studying the problem of arrival data modeling in call centers shows three important issues on arrival data, each of which has a great influence on the validity of any model representing the arrival data. Several models from literature were studied and one of those models tried to account for the three issues. In this paper, a new model is presented to capture also the effects of the different modeling issues but with different modeling approach seeking better fit to the real data and less parameter. Both of models is studied and applied on the arrival process of the Egyptian Phone Directory Call Center. Through this case study, the three modeling issues were verified and both models were compared showing that they succeeded in dealing with these issues with nearly similar performance. The developed model in this study, however, is still better as it requires the half number of parameters that other model requires.

Acknowledgment

The authors would like to thank the workforce management staff at the Egyptian Phone Directory Call Center for providing the data sets for the case study presented through this work.

References

Aksin, Z., Armony, M., and Mehrotra, V. "The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research." *Production And Operations Management* 16, no. 6 (2007): 665–688.

Andrews, D. "An Empirical Process Central Limit Theorem for Dependent Non-Identically Distributed Random Variables." *Journal of Multivariate Analysis* 38, no. 2 (1991): 188-203.

Avramidis, A., Deslauriers, A., and L'Ecuyer, P. "Modeling Daily Arrivals to a Telephone Call Center." *Management Science* 50, no. 7 (2004): 896-908.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. "Statistical Analysis of a Telephone Call Center." *Journal of the American Statistical Association* 100, no. 469 (2005): 36-50.

Ferguson, T. "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics* 1, no. 2 (1973): 209-230.

Henderson, S. "Estimation for nonhomogeneous Poisson processes from aggregated data." *Operations Research Letters* 31 (2003): 375-382.

Heyman, D. and Sobel, M. *Stochastic Models in Operations Research:Stochastic Processes and Operating Characteristics.* 1st Edition. Vol. I. Mineola, NY: Dover Publications, 2004.

Jongbloed, G. and Koole, G. "Managing uncertainty in call centres using Poisson mixtures." *Applied Stochastic Models in Business and Industry* 17, no. 4 (2001): 307-318.

Kao, E. and Chang, S. "Modeling Time-Dependent Arrivals to Service Systems: A Case in Using a Piecewise-Polynomial Rate Function in a Nonhomogeneous Poisson Process." *Management Science* 34, no. 11 (1988): 1367-1379.

Koole, G. and Mandelbaum, A. "Queueing Models of Call Centers: An Introduction." *Annals of Operations Research* 113 (2002): 41-59.

Mandelbaum, A., Sakov, A., and Zeltyn, S. "Empirical Analysis of a Call Center." Working Paper, 2002.

Massey, W., Parker, G., and Whitt, W. "Estimating the parameters of a nonhomogeneous Poisson process with linear rate." *Telecommunications Systems* 5, no. 2 (1996): 361-388.

Ross, S. *A First Course in Probability.* Fifth Edition. Prentice Hall, 1998.

Ross, S. *Introduction to Probability Models.* Ninth Edition. Academic Press, 2007.

Steckley, S., Henderson, S., and Mehrotra, V. "Performance Measures for Service Systems With a Random Arrival Rate." *Winter Simulation Conference.* 2005. 566-575.