Clustering of arrivals in queueing systems: autoregressive conditional duration approach

Petra Tomanová1 · Vladimír Holý1

Accepted: 18 March 2021 / Published online: 2 April 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Arrivals in a queueing system are typically assumed to be independent and exponentially distributed. Our analysis of an online bookshop, however, shows that there is an autocorrelation structure. First, we adjust the inter-arrival times for diurnal and seasonal patterns. Second, we model adjusted inter-arrival times by the generalized autoregressive score (GAS) model based on the generalized gamma distribution in the spirit of the autoregressive conditional duration (ACD) models. Third, in a simulation study, we investigate the effects of the dynamic arrival model on the number of customers, the busy period, and the response time in queueing systems with single and multiple servers. We find that ignoring the autocorrelation structure leads to significantly underestimated performance measures and consequently suboptimal decisions. The proposed approach serves as a general methodology for the treatment of arrivals clustering in practice.

Keywords Inter-Arrival Times · Queueing Theory · Autoregressive Conditional Duration Model · Generalized Autoregressive Score Model · Retail Business

1 Introduction

In various applications of operations research, it is undeniable that the characteristics of a model evolve over time. The parameters of interest can depend on the time of day and season as well as on their past values and other past indicators. In the present paper, we focus on the latter dependency in arrivals to queueing systems from the perspective

Preliminary results were presented in Tomanová (2018, 2019b, a).
of the autoregressive conditional duration models with the generalized autoregressive score dynamics.

Many standard queueing systems consider inter-arrival times to be independent, for the sake of analytical tractability. Some studies, however, explicitly consider autocorrelation and model arrivals using the Markovian arrival process (MAP) (see, e.g., Adan and Kulkarni 2003; Buchholz and Kriege 2017; Manafzadeh Dizbin and Tan 2019), the Markov renewal process (see, e.g., Tin 1985; Patuwo et al. 1993; Szekli et al. 1994), the moving average process (see, e.g., Finch 1963; Finch and Pearce 1965; Pearce 1967) or the discrete autoregressive process (see, e.g., Hwang and Sohraby 2003; Kamoun 2006; Miao and Lee 2013). Hwang and Sohraby (2003) argue that time series models with few parameters are more suitable in practice than the MAP models, which might be overparametrized. Simulation studies investigating the autocorrelation in arrivals include Livny et al. (1993), Resnick and Samorodnitsky (1997), Altiok and Melamed (2001), Nielsen (2007) and Civelek et al. (2009). Overall, these studies show that ignoring the autocorrelation structure in a queueing system, if one is present, leads to biased performance measures.

Arrival processes are also extensively studied in the financial high-frequency literature. In this field, the duration analysis deals with the modeling of the times between successive transactions (trade durations), times until the price reaches a certain level (price durations), and times until a certain volume is traded (volume durations). Typically, the autoregressive conditional duration (ACD) model of Engle and Russell (1998) is used. Its dynamics are analogous to the generalized autoregressive conditional heteroskedasticity (GARCH) model of Bollerslev (1986). In its basic version, the ACD model is based on the exponential distribution, but many other distributions are considered in the literature as well. Notably, Lunde (1999) introduces the generalized gamma distribution to the ACD model. Bauwens et al. (2004) and Fernandes and Grammig (2005) find that in financial applications, the generalized gamma distribution is more adequate than the exponential, Weibull, and Burr distributions. Hautsch (2003) further finds that the four-parameter generalized F-distribution reduces to the three-parameter generalized gamma distribution in most cases of financial durations. For a survey of financial duration analysis, see Pacurar (2008) and Saranjeet and Ramanathan (2018).

A modern approach to time series modeling is the general autoregressive score (GAS) model of Creal et al. (2013), also known as the dynamic conditional score (DCS) model by Harvey (2013). The GAS model is an observation-driven model providing a general framework for modeling time-varying parameters of any underlying probability distribution. It captures the dynamics of time-varying parameters by the autoregressive term and the score of the conditional density function using the shape of the density function. The theoretical properties of the GAS models together with their estimation by the maximum likelihood method are investigated, e.g., by Blasques et al. (2014) and Blasques et al. (2018). The empirical performance of the GAS models is studied, e.g., by Koopman et al. (2016) and Blazsek and Licht (2020). So far, there have been over 200 papers devoted to numerous models belonging to the GAS family, with various applications, see www.gasmodel.com for a comprehensive list.

The class of ACD models and the class of GAS models overlap. In the case of the exponential distribution, the ACD model is equivalent to the GAS model (see
Clustering of arrivals in queueing systems... Creal et al. 2013). For more complex distributions, however, they tend to differ, as the ACD models are driven by the lagged observation (or, when rewritten, the difference between the observation and the expected value) while the GAS models are driven by the lagged score. In general, the GAS models are very often superior than the alternatives (see, e.g., Blazsek and Villatoro 2015; Koopman et al. 2016; Chen and Xu 2019; Gorgi and Lit 2019; Harvey et al. 2019; Blazsek and Licht 2020). Concerning the GAS models for positive or non-negative values that are suitable for the duration analysis, Fonseca and Cribari-Neto (2018) use the Birnbaum–Saunders distribution, Blasques et al. (2020) use the zero-inflated negative binomial distribution as well as the generalized gamma distribution, and Harvey and Ito (2020) use the generalized beta distribution as well as the generalized gamma distribution.

In the present paper, we put together three cornerstones – queueing theory, duration analysis, and the GAS models – and demonstrate that they fit together perfectly. The literature has already successfully incorporated GAS models with the duration analysis as discussed above, however, the perspective from queueing theory is our novel contribution. We analyze the inter-arrival times between orders from an online Czech bookshop. First, we adjust the arrivals for diurnal and seasonal patterns, using the cubic spline. Second, we find that the adjusted inter-arrival times exhibit strong clustering behavior: short inter-arrival times are usually followed by short times. To capture this autocorrelation, we use the dynamic model based on the generalized gamma distribution in the spirit of the ACD models. We confirm that the proposed specification is quite suitable for the observed data. Third, we investigate the effects of the proposed arrivals model on queueing systems with single and multiple servers and exponential services. In a simulation study, we show that various performance measures – the number of customers in the system, the busy period of servers, and the response time – have higher mean and variance as well as heavier tails for the proposed dynamic arrivals model than for the standard static model. Lastly, we illustrate how the misspecification of the arrivals model can lead to suboptimal decisions.

The rest of this paper is structured as follows. In Sect. 2, we present the model based on the generalized gamma distribution with the GAS dynamics for diurnally adjusted inter-arrival times. In Sect. 3, we show that real data of a retail store exhibit an autocorrelation structure that is well captured by our model. In Sect. 4, we investigate the impact of the proposed arrivals model on the performance measures using simulations. We conclude the paper in Sect. 5.

2 Dynamic model for arrivals

2.1 Diurnal and seasonal adjustment

Before we use the generalized autoregressive score (GAS) model to capture the autoregressive structure of inter-arrival times, we need to deal with diurnal, weekly, and monthly seasonality patterns. To model the non-linear behavior of the diurnal and seasonal patterns and to properly adjust the inter-arrival times, the cubic spline method is used. A cubic spline is a piecewise cubic polynomial with continuous derivatives
up to order two at each of the $K$ fixed points, called knots, $k = 1, \ldots, K$. Bruce and Bruce (2017) point out that the cubic spline method is often a superior approach to polynomial regression since the polynomial regression often leads to undesirable wiggliness in the regression equation.

To take into account the specifics of raw inter-arrival times $\{\tilde{y}_i\}_{i=1}^n$, we define the cubic spline with knots at $\{\xi_k\}_{k=1}^K$ as

$$\log \tilde{y}_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \gamma t_i + \varepsilon_i, \quad (1)$$

where $\{\beta_j\}_{j=1}^{K+3}$ and $\gamma$ are parameters to be estimated, $\varepsilon_i$ is a disturbance term, $t_i$ is the trend variable, $\{b_j\}_{j=1}^{K+3}$ are the basis functions, and $x_i$ is the time difference in minutes between the time-stamp of the $i$th observation and the beginning of the week (Monday 00:00) to which the $i$th observation belongs. Thus, $\{x_i\}_{i=1}^n$ is able to capture both diurnal and intra-week patterns. The basis functions are equal to (i) the variable $x_i$, $b_1(x_i) = x_i$; (ii) its square, $b_2(x_i) = x_i^2$; (iii) its cube, $b_3(x_i) = x_i^3$; and (iv) truncated power functions, $b_{K+3}(x_i) = \max\left\{0, (x_i - \xi_k)^3\right\}$, $k = 1, \ldots, K$.

The trend variable $t_i$ is linear in time (not linear in observations), $t_1 = 0$ and $t_i = \sum_{j=1}^{i-1} \tilde{y}_j$ for $i = 2, \ldots, n$, to take into account any irregularity in the spacing of the observations. Moreover, the logarithmic transformation of $\tilde{y}$ ensures the non-negativity of the adjusted inter-arrival times. Equidistant intervals are used for identifying the knots, since intervals based on quantiles might lead to too few knots being allocated to off-peak hours.

The regression parameters in (1) are estimated by the weighted least squares (WLS) method with weights being the inter-arrival times. The WLS naturally compensates for the possibility that during a particular time interval either a small number of long inter-arrival times or a higher number of shorter inter-arrival times is observed, i.e., the number of observed inter-arrival times within a time interval depends on the values of the inter-arrival times themselves. Unlike ordinary least squares, this approach properly weights the inter-arrival times during hours that exhibit a small median but a huge dispersion. Once the parameters are estimated, the diurnally and seasonally adjusted and detrended inter-arrival times $y_i$ are set to exponentiated residuals from regression (1).

### 2.2 Generalized gamma distribution

Next, we assume that the adjusted inter-arrival times $y_i$ follow the generalized gamma distribution. The generalized gamma distribution is a continuous probability distribution for non-negative variables proposed by Stacy (1962). It is a three-parameter generalization of the two-parameter gamma distribution and contains the exponential distribution and the Weibull distribution as special cases. The distribution has the scale parameter $\alpha$ and the shape parameters $\psi > 0$ and $\varphi > 0$. We use the parametrization allowing for arbitrary values of $\alpha$ which is quite suitable for modeling its dynamics.
The probability density function is

\[ f(y|\alpha, \psi, \varphi) = \frac{1}{\Gamma(\psi)} \frac{\varphi}{e^{\alpha}} \left(\frac{y}{e^{\alpha}}\right)^{\psi-1} e^{-\left(\frac{y}{e^{\alpha}}\right)^{\psi}} \quad \text{for } y \in (0, \infty), \]  

for \( y \in (0, \infty) \), where \( \Gamma(\cdot) \) is the gamma function. The expected value and variance is

\[
\begin{align*}
E[Y] &= e^{\alpha} \frac{\Gamma(\psi + \varphi^{-1})}{\Gamma(\psi)}, \\
\text{var}[Y] &= e^{2\alpha} \frac{\Gamma(\psi + 2\varphi^{-1})}{\Gamma(\psi)} - \left( e^{\alpha} \frac{\Gamma(\psi + \varphi^{-1})}{\Gamma(\psi)} \right)^2.
\end{align*}
\]

(3)

The score for the parameter \( \alpha \) is

\[
\nabla_{\alpha}(y, \alpha, \psi, \varphi) = \frac{\partial}{\partial \alpha} \log f(y|\alpha, \psi, \varphi) = \varphi \left( y^\varphi e^{-\varphi \alpha} - \psi \right) \quad \text{for } y \in (0, \infty). \]

(4)

The Fisher information for the parameter \( \alpha \) is

\[
\mathcal{I}_{\alpha}(\alpha, \psi, \varphi) = E\left[ \nabla_{\alpha}(y, \alpha, \psi, \varphi)^2 | \alpha, \psi, \varphi \right] = \psi \varphi^2.
\]

(5)

Note that the Fisher information for \( \alpha \) is not dependent on \( \alpha \) itself. Special cases of the generalized gamma distribution include the gamma distribution for \( \varphi = 1 \), the Weibull distribution for \( \psi = 1 \), and the exponential distribution for \( \psi = 1 \) and \( \varphi = 1 \). The generalized gamma distribution itself is contained in a larger family – the generalized F-distribution with four parameters.

### 2.3 Generalized autoregressive score dynamics

We now consider the scale parameter to be time-varying. In the generalized autoregressive score (GAS) framework of Creal et al. (2013), the time-varying parameters are linearly dependent on their lagged values and the lagged values of the score of the conditional density. Typically, only the first lag is used. In our case, the parameter \( \alpha_i \) follows the recursion

\[
\alpha_{i+1} = c + b\alpha_i + a \nabla_{\alpha}(y_i, \alpha_i, \psi, \varphi) \\
= c + b\alpha_i + a \varphi \left( y_i^\varphi e^{-\varphi \alpha_i} - \psi \right),
\]

(6)

where \( c \) is the constant parameter, \( b \) is the autoregressive parameter, \( a \) is the score parameter, and \( \nabla_{\alpha}(y_i, \alpha_i, \psi, \varphi) \) is the score defined in (4). In the GAS framework, the score can be scaled by the inverse of the Fisher information or the square of the inverse of the Fisher information. In our case, however, both scaling functions, that based on the Fisher information and the unit scaling, lead to the same model, since the Fisher information does not depend on \( \alpha_i \). The score for a time-varying parameter
\( \alpha_i \) is the gradient of the log-likelihood with respect to \( \alpha_i \) and indicates how sensitive the log-likelihood is to \( \alpha_i \). In the GAS model, the score drives the time variation in \( \alpha_i \) based on the shape of the generalized gamma density function.

Let \( \theta = (c, b, a, \psi, \phi) \) denote the vector of parameters in the model. We can estimate \( \theta \) straightforwardly by the maximum likelihood method. The log-likelihood function is given by

\[
\ell(\theta) = \ln f(y_0|\alpha_0, \psi, \phi) + \sum_{i=1}^{n} \ln f(y_i|\alpha_i, \psi, \phi),
\]

where \( f(\cdot) \) is the generalized gamma density function given by (2). We deliberately set aside the first term as the time-varying parameter \( \alpha_i \) needs to be initialized at \( i = 0 \). We set the value of \( \alpha_0 \) to the long-term mean value \( c/(1-b) \). Subsequent values of \( \alpha_i, i = 1, \ldots, n \) then follow recursion (6). The parameter estimates \( \hat{\theta} \) are obtained as the answer to the non-linear optimization problem

\[
\hat{\theta} = \max_\theta \ell(\theta).
\]

3 Empirical evidence

3.1 Data overview and preparation

The data sample was obtained from the database of an online bookshop with one brick-and-mortar location in Prague, Czechia. The data cover the period from June 8 to December 20, 2018, resulting in 28 full weeks and 5,753 observations. The precision of the timestamp is one minute. Thus, zero inter-arrival times might occur in the data due to two or more orders that arrive within one minute. Since the generalized gamma distribution has strictly positive support, the zero inter-arrival times are set to a small positive number. Bauwens (2006) replaces the zero inter-arrival times with a value equal to one-half of the minimum positive inter-arrival time and argued that this is a more correct approach than discarding them. Hence, all 81 zero inter-arrival times are set to 0.5 minutes.

3.2 Diurnal and seasonal patterns

The median of the raw inter-arrival times is 24 minutes and the mean is 49 minutes—more than double, due to the long inter-arrival times at night (specifically, the hours between midnight and 9 AM, see Fig. 1). The hours between 9 and 11 AM exhibit many short inter-arrival times and several very long inter-arrival times, resulting in high dispersion (SD = 111.39). The rest of the rush hours (until 5 PM) shows a similar inter-arrival time median but much lower dispersion (SD = 35.98). Moreover, strong weekly and monthly seasonal patterns are observed. The highest order counts—and consequently lower inter-arrival time values—occur at the beginning of the week and decrease until Saturday, see Fig. 2. On Sundays, order counts increase again and exhibit
the highest dispersion. During the summer months, the order counts are rather low—resulting in higher inter-arrival times—and linearly increase until December.

To obtain the diurnally and seasonally adjusted and detrended inter-arrival times, the regression equation (1) with a selected number of knots is estimated. In practice, the selection of a suitable number of knots is an empirically-driven task. One must bear in mind that too many knots can result in overfitting (e.g., one knot for every hour results in too unnatural bumpy behavior), and, on the other hand, that too few knots can result in an inadequate fit (e.g., one knot for every two hours does not satisfactorily capture the nonlinear behavior of the data). After a little experimenting, we selected one knot for every 90 minutes, which captures all the important nonlinearities and does not produce overfitting. Note that weekly aggregation is used in (1), which results in the same daily seasonal component for Mondays, Tuesdays, etc. To ensure continuity between Sundays and Mondays, the sample is stacked three times consecutively and the adjusted inter-arrival times are computed based on the second sub-sample. Parameters are estimated by the WLS.

The fitted values are shown in Figs. 1 and 2. Note that they do not coincide with the smooth cubic spline function due to a linear trend which makes the corresponding fitted line saw-toothed. The diurnally and seasonally adjusted and detrended inter-arrival times are computed as the exponentiated residuals of estimated equation (1) and for convenience, they are standardized to have unit mean. Their values range from 0.002 to 11.23 minutes.

### 3.3 Fit of the dynamic model

Even after the seasonal and diurnal adjustment, the inter-arrival times still tend to cluster over time—long (short) inter-arrival times are likely to be followed by long (short) inter-arrival times. This dependence is not particularly strong, but nevertheless it is statistically significant, as illustrated in Fig. 3. To capture the autocorrelation, we use the dynamic model based on the generalized gamma distribution with the GAS dynamics in (6). The parameters are estimated by the maximum likelihood method determined by the non-linear optimization problem in (8) and the log-likelihood function in (7).
Fig. 2 Intra-week view of raw inter-arrival times and their fitted diurnal/seasonal pattern

Fig. 3 The autocorrelation function (ACF) and the partial autocorrelation function (PACF) of adjusted inter-arrival times. Red dashed lines indicate 5% confidence bounds

For comparison, we also present the results for static and dynamic models based on special cases of the generalized gamma distribution (G.G.), namely, the exponential (Exp.), Weibull, and gamma distributions.

Parameter estimates and the performance evaluation in terms of the Akaike information criterion (AIC) of both static and dynamic inter-arrival time models are shown in Table 1. The AIC values are at least 43.59 lower for the dynamic models than for their static counterparts. However, the differences between the dynamic models are not so striking: the highest difference is between the exponential and generalized gamma distributions (by 5.94). The best performing model is the most general one, the dynamic GAS model using the generalized gamma distribution. The dynamic models based on either the exponential or generalized gamma distributions in comparison with their static counterparts are further analyzed in the simulation study of queueing systems.
Clustering of arrivals in queueing systems...

Table 1 Parameter estimates of the inter-arrival time models with the log-likelihood value (Lik.) and the Akaike information criterion (AIC)

| Model     | Dist.      | Parameter Estimates | Model Fit |
|-----------|------------|---------------------|-----------|
| Spec.     |            | $c$  | $b$  | $a$  | $\psi$ | $\varphi$ | Lik.    | AIC     |
| Static    | Exp.       | 0.00 | 0.00 | 0.00 | 1.00  | 1.00  | $-5753.00$ | 11508.00 |
| Static    | Weibull    | $-0.01$ | 0.00 | 0.00 | 1.00  | 0.97  | $-5748.93$ | 11501.86 |
| Static    | Gamma      | 0.04 | 0.00 | 0.00 | 0.96  | 1.00  | $-5749.77$ | 11503.54 |
| Static    | G. G.      | $-0.12$ | 0.00 | 0.00 | 1.08  | 0.93  | $-5748.37$ | 11502.75 |
| Dyn.      | Exp.       | 0.00 | 0.76 | 0.06 | 1.00  | 1.00  | $-5728.28$ | 11462.56 |
| Dyn.      | Weibull    | 0.00 | 0.75 | 0.06 | 1.00  | 0.97  | $-5724.89$ | 11457.79 |
| Dyn.      | Gamma      | 0.01 | 0.76 | 0.06 | 0.97  | 1.00  | $-5725.97$ | 11459.95 |
| Dyn.      | G. G.      | $-0.06$ | 0.72 | 0.07 | 1.15  | 0.90  | $-5723.31$ | 11456.62 |

4 Impact on queueing systems

4.1 System with single server

Using simulations, we now investigate the effects of various arrival models on performance measures in queueing systems. We consider models based on the exponential and generalized gamma distributions with the static and dynamic specifications. The coefficients of the models are taken from Table 1. In all models, the rate of arrivals is $\lambda = 1$ job per minute. First, we focus on the queueing system with a single server only. We consider the service times to be independent and exponentially distributed with the rate $\mu$ ranging from 1.1 to 1.5 jobs per minute. We simulate the arrival and service processes and measure the number of customers in the system, the busy period of the server, and the response time. The number of simulation runs is equal to $10^9$, which seems to be sufficient for the reported precision of one decimal place as the results are in line with the theoretical performance measures for the static exponential scenario as well as Little’s law for all scenarios.

The results are presented in Table 2. For all values of $\mu$, the systems based on the generalized gamma distribution have higher values of performance measures than the systems based on the exponential distribution in terms of the mean, standard deviation, and 95 percent quantile. Similarly, systems with the dynamic specification have higher values of performance measures than the systems with the static specification. The left plot of Fig. 4 shows how the probability mass function of the number of customers differs for the static and dynamic models. The dynamic model has a higher probability of an empty system as there is a tendency to have longer periods of low activity. It has also higher probabilities of large numbers of customers in the system, as arrivals tend to cluster. The right plot of Fig. 4 shows how the density functions of the response times for the static and dynamic models differ. In the dynamic model, customers simply have to wait longer. The differences between the static and dynamic models are naturally weaker for larger $\mu$. 

Springer
Table 2  Mean values (M), standard deviations (SD) and 95%-quantiles (95%) of the number of customers in the system, the busy period of the server and the response time in various queueing systems with a single server

| Queueing system | No. of customers | Busy period | Response time |
|-----------------|------------------|-------------|---------------|
| μ              | Spec. Dist. | M  | SD | 95% | M  | SD | 95% | M  | SD | 95% |
| 1.1 Static      | Exp.          | 10.0 | 10.5 | 31.0 | 10.0 | 45.8 | 39.8 | 10.0 | 10.0 | 30.0 |
| 1.1 Static      | G. G.         | 10.4 | 10.9 | 32.0 | 10.4 | 47.6 | 41.4 | 10.4 | 10.4 | 31.1 |
| 1.1 Dyn.        | Exp.          | 12.4 | 13.4 | 39.0 | 10.8 | 54.4 | 41.2 | 12.4 | 12.6 | 37.6 |
| 1.1 Dyn.        | G. G.         | 12.8 | 13.8 | 41.0 | 11.2 | 56.1 | 43.0 | 12.8 | 13.1 | 39.0 |
| 1.2 Static      | Exp.          | 5.0  | 5.5  | 16.0 | 5.0  | 16.6 | 22.1 | 5.0  | 5.0  | 15.0 |
| 1.2 Static      | G. G.         | 5.2  | 5.7  | 17.0 | 5.2  | 17.2 | 22.9 | 5.2  | 5.2  | 15.5 |
| 1.2 Dyn.        | Exp.          | 6.0  | 6.8  | 20.0 | 5.4  | 19.5 | 23.3 | 6.0  | 6.1  | 18.3 |
| 1.2 Dyn.        | G. G.         | 6.2  | 7.1  | 20.0 | 5.6  | 20.2 | 24.4 | 6.2  | 6.4  | 19.0 |
| 1.3 Static      | Exp.          | 3.3  | 3.8  | 11.0 | 3.3  | 9.2  | 14.8 | 3.3  | 3.3  | 10.0 |
| 1.3 Static      | G. G.         | 3.4  | 3.9  | 11.0 | 3.4  | 9.6  | 15.3 | 3.4  | 3.4  | 10.3 |
| 1.3 Dyn.        | Exp.          | 3.9  | 4.6  | 13.0 | 3.5  | 10.7 | 15.7 | 3.9  | 4.0  | 11.9 |
| 1.3 Dyn.        | G. G.         | 4.0  | 4.8  | 14.0 | 3.7  | 11.2 | 16.4 | 4.0  | 4.2  | 12.4 |
| 1.4 Static      | Exp.          | 2.5  | 3.0  | 8.0  | 2.5  | 6.1  | 11.0 | 2.5  | 2.5  | 7.5  |
| 1.4 Static      | G. G.         | 2.6  | 3.1  | 9.0  | 2.6  | 6.3  | 11.3 | 2.6  | 2.6  | 7.7  |
| 1.4 Dyn.        | Exp.          | 2.8  | 3.5  | 10.0 | 2.6  | 7.0  | 11.5 | 2.8  | 2.9  | 8.7  |
| 1.4 Dyn.        | G. G.         | 3.0  | 3.7  | 10.0 | 2.7  | 7.3  | 12.1 | 3.0  | 3.1  | 9.1  |
| 1.5 Static      | Exp.          | 2.0  | 2.4  | 7.0  | 2.0  | 4.5  | 8.6  | 2.0  | 2.0  | 6.0  |
| 1.5 Static      | G. G.         | 2.1  | 2.5  | 7.0  | 2.1  | 4.6  | 8.9  | 2.1  | 2.1  | 6.2  |
| 1.5 Dyn.        | Exp.          | 2.2  | 2.9  | 8.0  | 2.1  | 5.1  | 9.0  | 2.2  | 2.3  | 6.8  |
| 1.5 Dyn.        | G. G.         | 2.3  | 3.0  | 8.0  | 2.2  | 5.3  | 9.4  | 2.3  | 2.4  | 7.1  |

These results carry a warning for practice. When the standard M/M/1 system is assumed but the arrivals actually follow the GAS model based on the generalized gamma distribution, the performance measures are significantly underestimated. For example, the mean number of customers and the mean response time are 22 percent lower than the actual value for $\mu = 1.1$ jobs per minute. It is therefore crucial to correctly specify the model for arrivals.

4.2 System with multiple servers

Next, we consider queueing systems with multiple servers. We base the simulations on the same setting as in the previous section. The only difference lies in the service structure. We let the number of servers $c$ range from 11 to 15 and take the individual service rate to be $\mu = 0.1$ jobs per minute. Such values result in the same server utilizations $\rho = \lambda/(c\mu)$ as in the previous section. Again, we measure the number of customers in the system, the busy period of the servers, and the response time. By the busy period, we mean the full busy period, i.e., the duration of the state in which all servers are busy.
The results are presented in Table 3. They are very similar to those for a system with a single server: the generalized gamma distribution and the dynamic specification increase all performance measures. When incorrectly assuming an M/M/c system, the specification error is distinct but not as high as in the case of a single server. For example, when assuming an M/M/11 system, the mean number of customers and the mean response time are 14 percent lower than the actual value for arrivals based on the generalized gamma distribution with the dynamic specification.

In the following toy example, we illustrate how the misspecification of the arrival model can affect decision making. Let us assume that there are two types of costs associated with the operation of the system: the cost of running one server per unit of time $C_1 = 10$ euro per minute, and the cost of having a queue longer than 30 customers per unit of time $C_2 = 3000$ euro per minute. The analytic department is faced with the question of how many servers to operate. The composition of costs for different numbers of servers is shown in Fig. 5. The optimal number of servers according to the static model is 12 while it is 13 for the dynamic model. An analyst employing the static model believes that the total optimal costs are 127,13 euro per minute while they actually are 142,87 euro per minute for the suboptimal choice of 12 servers. An analyst correctly specifying the dynamic model finds that the lowest possible costs are 132,32 euro per minute for the optimal choice of 13 servers. The decision based on the misspecified arrival model therefore results in a total cost increase of 8 percent.

4.3 Discussion of more complex systems

We have focused on rather simple queueing systems in order to get transparent results. The M/M/1 system is as straightforward as can be, and therefore the best choice for an illustration of the impact of autocorrelated arrivals. The M/M/c system is used as a robustness check to show that the behavior observed for the M/M/1 system is present...
Table 3  Mean values ($M$), standard deviations (SD) and 95%-quantiles (95%) of the number of customers in the system, the full busy period of servers and the response time in various queueing systems with multiple servers and $\mu = 0.1$ jobs per minute

| Queueing System | No. of customers | Busy period | Response time |
|-----------------|------------------|-------------|---------------|
| $c$ | Spec. Dist. | $M$ | SD | 95% | $M$ | SD | 95% | $M$ | SD | 95% |
| 11 | Static Exp. | 16.8 | 10.7 | 38.0 | 10.0 | 45.8 | 39.9 | 16.8 | 13.8 | 43.8 |
| 11 | Static G. G. | 17.2 | 11.1 | 39.0 | 10.4 | 47.6 | 41.4 | 17.2 | 14.0 | 44.6 |
| 11 | Dyn. Exp. | 19.1 | 13.5 | 46.0 | 12.4 | 58.6 | 49.7 | 19.1 | 15.7 | 49.9 |
| 11 | Dyn. G. G. | 19.5 | 14.0 | 47.0 | 12.8 | 60.3 | 51.8 | 19.5 | 16.1 | 50.9 |
| 12 | Static Exp. | 12.2 | 5.8 | 24.0 | 5.0 | 16.6 | 22.1 | 12.2 | 10.8 | 33.6 |
| 12 | Static G. G. | 12.4 | 6.1 | 24.0 | 5.2 | 17.2 | 22.9 | 12.4 | 10.9 | 33.8 |
| 12 | Dyn. Exp. | 13.1 | 7.2 | 27.0 | 6.1 | 21.1 | 27.5 | 13.1 | 11.3 | 35.4 |
| 12 | Dyn. G. G. | 13.3 | 7.5 | 28.0 | 6.3 | 21.9 | 28.7 | 13.3 | 11.5 | 35.8 |
| 13 | Static Exp. | 11.0 | 4.4 | 19.0 | 3.3 | 9.2 | 14.8 | 11.0 | 10.3 | 31.3 |
| 13 | Static G. G. | 11.0 | 4.5 | 19.0 | 3.4 | 9.6 | 15.3 | 11.0 | 10.3 | 31.4 |
| 13 | Dyn. Exp. | 11.4 | 5.2 | 21.0 | 4.0 | 11.7 | 18.4 | 11.4 | 10.5 | 32.0 |
| 13 | Dyn. G. G. | 11.5 | 5.4 | 22.0 | 4.2 | 12.1 | 19.1 | 11.5 | 10.5 | 32.2 |
| 14 | Static Exp. | 10.4 | 3.8 | 17.0 | 2.5 | 6.1 | 11.0 | 10.4 | 10.1 | 30.5 |
| 14 | Static G. G. | 10.5 | 3.9 | 17.0 | 2.6 | 6.3 | 11.3 | 10.5 | 10.1 | 30.6 |
| 14 | Dyn. Exp. | 10.7 | 4.4 | 19.0 | 3.0 | 7.7 | 13.5 | 10.7 | 10.2 | 30.9 |
| 14 | Dyn. G. G. | 10.7 | 4.5 | 19.0 | 3.1 | 8.0 | 14.0 | 10.7 | 10.2 | 30.9 |
| 15 | Static Exp. | 10.2 | 3.5 | 16.0 | 2.0 | 4.5 | 8.6 | 10.2 | 10.0 | 30.2 |
| 15 | Static G. G. | 10.2 | 3.6 | 16.0 | 2.1 | 4.6 | 8.9 | 10.2 | 10.0 | 30.2 |
| 15 | Dyn. Exp. | 10.3 | 3.9 | 17.0 | 2.3 | 5.6 | 10.5 | 10.3 | 10.1 | 30.4 |
| 15 | Dyn. G. G. | 10.3 | 4.1 | 18.0 | 2.4 | 5.8 | 10.9 | 10.3 | 10.1 | 30.4 |

Fig. 5  Costs related to the number of servers and long queues for the static and dynamic arrival models based on the generalized gamma distribution in queueing systems with multiple servers and $\mu = 0.10$ jobs per minute
even for different specifications. As for the toy example of decision making in the M/M/c system, it is meant just as a simplistic illustration revealing a potential source of suboptimal decisions.

On the other hand, Tomanová (2018, 2019b, a) explore a much more realistic and complex queueing system specific to this case of an online bookshop. As this queueing system is tailored just for this specific application and cannot be easily transferred to others, we only summarize the main findings. Tomanová (2018) performs a process quality assessment based on process simulation and reports that the key quality target is not satisfied in almost twice as many cases when the dynamic model is considered (the target is not satisfied in 6.16 percent of them) than when the static model is considered (for which the target is not satisfied 3.23 percent of the time). The common approach—a static model which assumes that times between arrivals follow the exponential distribution with a constant rate—underestimates the probability of extreme values and thus significantly skews the basis for process quality assessment and leads to suboptimal decisions. Tomanová (2019b) also demonstrates that the clustering of arrivals increases the probability of weeks with an extreme number of arrivals, something which has a negative effect on the fulfillment of targets. Tomanová (2019a) further extends that work to making final recommendations for the management of the online bookshop. The main finding is that 21 percent of the orders are not satisfied within a working day due to insufficiently allocated resources for the first stage (pre-processing of arrivals).

5 Conclusion

We have analyzed the dependence of inter-arrival times in queueing systems and demonstrated the negative effect of misspecifying the arrival model on decision making. To capture the autocorrelation structure of the inter-arrival times, we have proposed using a dynamic model based on the generalized gamma distribution with the GAS dynamics. We have found that this approach is superior to the standard model that uses the exponential distribution with a constant rate, since it leads to a more faithful representation of the mean and extreme values of the arrival process. Our study has carried out three steps.

1. We have constructed a suitable model for capturing the diurnal and seasonal dependencies which takes into account a specific time-structure of inter-arrival times. It uses a cubic spline approach and estimates the parameters by the weighted ordinary least square method to properly adjust inter-arrival times during hours that exhibit a small median but a huge dispersion.

2. We have found that the GAS models based on the generalized gamma distribution and its special cases fit the data better than do their static counterparts. This is due to the fact that the static models ignore the autocorrelation structure, which is still present even after the proper diurnal and seasonal adjustments.

3. We have compared both static and dynamic models in a simulation study of queueing systems with single and multiple servers and exponential services. We have shown that ignoring the autocorrelation structure leads to biased performance mea-
sures. The number of customers in the system, the busy periods of the servers, and the response times, have higher means and variances as well as heavier tails for the proposed dynamic arrivals model than for the standard static model. We have also shown how a trust in the standard static model for inter-arrival times leads to suboptimal decisions and consequently to a loss of profits.

A proper treatment of arrival dependence is of great importance since its ignorance generates extra costs. Our approach is useful for process simulations and consequently for process optimization and process quality assessment.

The main limitation of this paper and a topic for future research is the theoretical treatment of queueing systems with inter-arrival times following the GAS model. In the paper, we have resorted to simulations to determine the moments, quantiles, and density functions of the performance measures. Theoretical derivations of these quantities and functions is undoubtedly challenging but perhaps possible in some cases. Another topic for future research, which would be easier to achieve, is the use of the proposed approach in other applications. Besides retail order processing, these may include customer service, project management, manufacturing engineering, emergency services, logistics, transportation, telecommunications, computing, and others.

Acknowledgements We would like to thank the organizers and participants of the 7th International Conference on Management (Nový Smokovec, September 26–29, 2018), the 30th European Conference on Operational Research (Dublin, June 23–26, 2019), the 15th International Symposium on Operations Research in Slovenia (Bled, September 25–27, 2019) and the 3rd International Conference on Advances in Business and Law (Dubai, November 23–24, 2019) for fruitful discussions.

Funding The work on this paper was supported by the Internal Grant Agency of the Prague University of Economics and Business under project F4/27/2020, the Czech Science Foundation under project 19-08985S, and the Institutional Support Funds for the long-term conceptual development of the Faculty of Informatics, Prague University of Economics and Business.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Adan IJBF, Kulkarni VG (2003) Single-server queue with Markov-dependent inter-arrival and service times. Queueing Syst 45(2):113–134

Altioň T, Melamed B (2001) The case for modeling correlation in manufacturing systems. IIE Trans 33(9):779–791

Bauwens L (2006) Econometric Analysis of Intra-Daily Trading Activity on the Tokyo Stock Exchange. Monetary and Economic Studies 24(1):1–24, URL http://www.imes.boj.or.jp/research/abstracts/english/me24-1-1.html

Bauwens L, Giot P, Grammig J, Veredas D (2004) A comparison of financial duration models via density forecasts. Int J Forecast 20(4):589–609

Blasques F, Koopman SJ, Lucas A (2014) Stationarity and ergodicity of univariate generalized autoregressive score processes. Electr J Stat 8(1):1088–1112

Blasques F, Gorgi P, Koopman SJ, Wintenberger O (2018) Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. Electr J Stat 12(1):1019–1052
Clustering of arrivals in queueing systems...

Blasques F, Holý V, Tomanová P (2020) Zero-inflated autoregressive conditional duration model for discrete trade durations with excessive zeros, 1812.07318

Blazsek S, Licht A (2020) Dynamic conditional score models: a review of their applications. Appl Econ 52(11):1181–1199

Blazsek S, Villatoro M (2015) Is beta-t-EGARCH(1,1) superior to GARCH(1,1)? Appl Econ 47(17):1764–1774

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econ 31(3):307–327

Bruce P, Bruce A (2017) Practical statistics for data scientists: 50 essential concepts. O’Reilly Media, Sebastopol. URL: https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/

Buchholz P, Kriege J (2017) Fitting correlated arrival and service times and related queueing performance. Queueing Syst 85(3–4):337–359

Chen R, Xu J (2019) Forecasting volatility and correlation between oil and gold prices using a novel multivariate GAS model. Energy Econ 78:379–391. https://doi.org/10.1016/j.eneco.2018.11.011

Civelek I, Biller B, Scheller-Wolf A (2009) The impact of dependence on queueing systems. URL: https://www.researchgate.net/publication/228814043

Creal D, Koopman SJ, Lucas A (2013) Generalized autoregressive score models with applications. J Appl Econ 28(5):777–795. https://doi.org/10.1002/jae.1279

Engle RF, Russell JR (1998) Autoregressive conditional duration: a new model for irregularly spaced transaction data. Econometrica 66(5):1127–1162. https://doi.org/10.2307/2999632

Fernandes M, Grammig J (2005) Nonparametric specification tests for conditional duration models. J Econ 127(1):35–68. https://doi.org/10.1016/j.jeconom.2004.06.003

Finch PD (1963) The single server queueing system with non-recurrent input-process and erlang service time. J Aust Math Soc 3(2):220–236. https://doi.org/10.1017/s1446788700027968

Finch PD, Pearce C (1965) A second look at a queueing system with moving average input process. J Aust Math Soc 5(1):100–106. https://doi.org/10.1017/s144678870002591x

Fonseca RV, Cribari-Neto F (2018) Bimodal Birnbaum-Saunders generalized autoregressive score model. J Appl Stat 45(14):2585–2606. https://doi.org/10.1080/02664763.2018.1428734

Gorgi Koopman SJ, Lit R (2019) The analysis and forecasting of tennis matches by using a high dimensional dynamic model. J R Stat Soc Ser A (Statistics in Society) 182(4):1393–1409. https://doi.org/10.1111/jrssa.12464

Harvey A, Hurn S, Thiele S (2019) Modeling directional (circular) time series. https://doi.org/10.17863/cam.43915

Harvey AC (2013) Dynamic models for volatility and heavy tails: with applications to financial and economic time series. Cambridge University Press, New York. https://doi.org/10.1017/cbo9781139540933

Harvey AC, Ito R (2020) Modeling time series when some observations are zero. J Econ 214(1):33–45. https://doi.org/10.1016/j.jeconom.2019.05.003

Hautsch N (2003) Assessing the risk of liquidity suppliers on the basis of excess demand intensities. J Financ Econ 1(2):189–215. https://doi.org/10.1093/jjfinec/nbh010

Hwang GU, Sohraby K (2003) On the exact analysis of a discrete-time queueing system with autoregressive inputs. Queueing Syst 43(1–2):29–41. https://doi.org/10.1023/a:1021848330183

Kamoun F (2006) The discrete-time queue with autoregressive inputs revisited. Queueing Syst 54(3):185–192. https://doi.org/10.1007/s11134-006-9591-3

Koopman SJ, Lucas A, Scharth M (2016) Predicting time-varying parameters with parameter-driven and observation-driven models. Rev Econ Stat 98(1):97–110. https://doi.org/10.1162/rest_a_00553

Livny M, Melamed B, Tsolias AK (1993) The impact of autocorrelation on queueing systems. Manag Sci 39(3):322–339. https://doi.org/10.1287/mnsc.39.3.322

Lunde A (1999) A Generalized Gamma Autoregressive Conditional Duration Model, URL: https://www.researchgate.net/publication/228464216

Manafzadeh Dizbin N, Tan B (2019) Modelling and analysis of the impact of correlated inter-event data on production control using Markovian arrival Processes. Flex Serv Manuf J 31(4):1042–1076. https://doi.org/10.1007/s11134-018-9329-7

Miao DWC, Lee HC (2013) Second-order performance analysis of discrete-time queues fed by DAR(2) sources with a focus on the marginal effect of the additional traffic parameter. Appl Stoch Models Bus Ind 29(1):45–60. https://doi.org/10.1002/asmb.939

Nielsen EH (2007) Autocorrelation in queueing network-type production systems - revisited. Int J Prod Econ 110(1–2):138–146. https://doi.org/10.1016/j.ijpe.2007.02.014
Pacurar M (2008) Autoregressive conditional duration models in finance: a survey of the theoretical and empirical literature. J Econ Surv 22(4):711–751. https://doi.org/10.1111/j.1467-6419.2007.00547.x

Patuwo BE, Disney RL, McNickle DC (1993) The effect of correlated arrivals on queues. IIE Trans 25(3):105–110. https://doi.org/10.1080/07408179308964296

Pearce C (1967) An imbedded chain approach to a queue with moving average input. Op Res 15(6):1117–1130. https://doi.org/10.1287/opre.15.6.1117

Resnick S, Samorodnitsky G (1997) Performance decay in a single server exponential queueing model with long range dependence. Op Res 45(2):235–243. https://doi.org/10.1287/opre.45.2.235

Saranjeet KB, Ramanathan TV (2018) Conditional duration models for high-frequency data: a review on recent developments. J Econ Surv 33(1):252–273. https://doi.org/10.1111/joes.12261

Stacy EW (1962) A generalization of the gamma distribution. Ann Math Stat 33(3):1187–1192. https://doi.org/10.2307/2237889

Szekli R, Disney RL, Hur S (1994) MR/GI/1 queues by positively correlated arrival stream. J Appl Probab 31(2):497–514. https://doi.org/10.1017/s0021900200045009

Tin P (1985) A queueing system with Markov-dependent arrivals. J Appl Prob 22(3):668–677. https://doi.org/10.1017/s0021900200029417

Tomanová P (2018) Measuring intensity of order arrivals and process quality assessment of an online bookshop: a case study from the Czech Republic. In: Frankovský M, Dobrovič J, Fedorko R (eds) Proceedings of the 7th international conference on management, Bookman s.r.o., Prešov, pp 768–773, URL http://www.managerconf.com/

Tomanová P (2019a) Business process simulation and process quality assessment of Czech online bookshop. In: Proceedings of the 3rd international conference on advances in business and law, Dubai Business School, Dubai, pp 1–4, URL http://publications.ud.ac.ae/index.php/ICABML-CP/article/view/403/127

Tomanová P (2019b) Clustering of arrivals and its impact on process simulation. In: proceedings of the 15th international symposium on operations research in Slovenia, Slovenian society informatika, Bled, pp 314–319, URL http://fgg-web.fgg.uni-lj.si/~sdrobne/sor/SOR’19-Proceedings.pdf

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.