Functional regression-based monitoring of quality of service in hospital emergency departments

Devashish Das\textsuperscript{a}, Kalyan S. Pasupathy\textsuperscript{b,c}, Curtis B. Storlie\textsuperscript{b,c}, and Mustafa Y. Sir\textsuperscript{b,c}

\textsuperscript{a}Industrial and Management Systems Engineering, University of South Florida, Tampa, FL, USA; \textsuperscript{b}Health Science Research, Mayo Clinic, Rochester, MN, USA; \textsuperscript{c}Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, USA

\textbf{ABSTRACT}

This article focuses on building a statistical monitoring scheme for service systems that experience time-varying arrivals of customers and have time-varying service rates. There is a lack of research in the systematic statistical monitoring of large-scale service systems, which is critical for maintaining a high quality of service. Motivated by the emergency department at a major academic medical center, this article intends to fill this research gap and provide a practical statistical monitoring scheme capable of detecting changes in service using readily available time stamp data. The proposed method is focused on building a functional regression model based on customer arrival and departure time instances from an in-control system. The model finds the expected departure intensity function for an observed arrival intensity on any given day of operation. The mean squared difference between the expected departure intensity function and the observed departure intensity functions is used to generate an alarm indicating a significant change in service. This methodology is validated using simulation and real data case studies. The proposed method can identify patterns of inefficiency or delay in service that are hard to detect using traditional statistical monitoring algorithms. The method offers a practical approach for monitoring service systems and determining when staffing levels need to be re-optimized.

\textbf{1. Introduction}

Most Emergency Departments (EDs) in the United States are overburdened, due to surging patient volumes (Derlet, 2002). Therefore, there is a need to implement process improvement methods in order to improve the efficiency of EDs (Callahan and Griffen, 2003). Fortunately, vast amounts of operational data are available from Electronic Medical Records (EMR), which can provide valuable insight into the performance of the EDs. Traditionally, stochastic processes based on queuing theory have been used to model the performance of various service systems including hospital EDs. Such models are valuable in designing operational policies such as determining staffing level, admission control policies, and patient routing policies (Saghafian et al., 2015). In addition to modeling efficient EDs, it is also important to monitor the daily operation of EDs for continuous process improvement, as well as identifying the need to redesign operational policies. For example, a statistical monitoring scheme would help the ED administration to decide when the current level of staffing is no longer optimal. However, there is a significant lack of research in statistical monitoring methods for monitoring emergency service systems. To fill this research gap, we introduce and evaluate a Statistical Process Control (SPC) scheme. On any particular day of operation, the SPC scheme estimates the expected departure rates for the observed arrival rates and generates an alarm if the departure rate of patients has “significantly” deviated from the expected departure rate. Such an alarm would prompt operational managers to look into probable causes of the alarm and recommend necessary action to resolve it, including but not limited to, reevaluating current staffing levels, which in turn would help in maintaining a higher quality of service.

In the SPC literature, monitoring queuing systems has been studied only to a limited extent. Typically, the focus has been on monitoring the number of customers waiting for service (i.e., the queue length). (Bhat and Rao, 1972; Bhat, 1987; Chen et al., 2011). More recently, Chen and Zhou (2015) designed a cumulative sum chart to monitor queuing systems. The major limitation of all these approaches is that they focus on M/M/1, M/G/1, or G/G/1 queues, which do not accurately depict service systems such as hospital EDs because: (i) arrival rates to EDs is time-inhomogeneous with intensity varying from day to day; (ii) treatment includes multiple stages, the duration of which depends on the complexity of a patient’s condition, staffing levels and overall patient numbers in the ED, all of which vary over time. Therefore, developing stochastic models to study the effects of time-varying demand on service systems...
In the interest of making EDs more efficient, vast amounts of operational data, including arrival time, departure time and time a physician sees a patient, etc., are readily available in most hospitals; they are used to report Center for Medicare and Medicaid Services and Joint Commission performance metric data from the EMRs usually involves investing in additional information technology and personnel/resources, and is prone to inaccuracies.

The basic idea of the proposed SPC method is illustrated in Figure 2. We estimate the arrival intensity for day $i$ from the arrival time stamps and then use a novel functional regression method to estimate the expected departure rate for the observed arrival rate on day $i$. Then the mean squared difference between the estimated and expected departure intensity observed on day $i$ gives a test statistic that is used to design an SPC scheme for the ED. There are two key contributions in the current work: (i) introduction of an SPC scheme for monitoring time-varying service systems; and (ii) a functional regression scheme to calculate the expected departure intensity function from an observed arrival intensity function. To the best of our knowledge, the development of a functional regression scheme for a counting process and associated methodological development is a novel contribution of this research. Also, we derive properties of the estimation method using kernel functions, which ensures efficient estimation of the coefficient function. Functional data analysis has been shown to be an useful tool for manufacturing process control applications (Jin et al., 2008; Liao and Sun, 2011; Huang et al., 2017). However, to the best of our knowledge, the unique challenges involved in applying functional data analysis methods to service systems with stochastic arrivals and departures have not been reported in the literature.

In the remainder of this article, we discuss the proposed method in some detail in Section 2 followed by simulation studies in Section 3. A case study based on the data collected from the Mayo Clinic Emergency Department, a Level 1 Trauma Center for both adult and pediatric patients, is presented in Section 4 followed by conclusions in Section 5.

### 2. Functional regression-based monitoring scheme

Let $\tau_{i,1}, \tau_{i,2}, \ldots, \tau_{i,N_i}$ and $\nu_{i,1}, \nu_{i,2}, \ldots, \nu_{i,M_i}$ be the arrival times and departure times of patients to the ED on day $i$, where

![Figure 1](image.png)

**Figure 1.** (a) Arrival rates to an ED and (b) departure rates from an ED.
\( N_i \) and \( M_i \) are the number of arrivals and number of departures observed on day \( i \), respectively. Assume that \( \tau_{j,1}, \nu_{j,1} \in [0, T], \forall i, j = \{1, 2, ..., N_i\}, j_1 = \{1, 2, ..., M_i\} \). In this article, arrivals and departures occurring from midnight of day \( i \) to midnight of day \( i + 1 \) are considered as one sample, therefore \( T = 24 \) hours. The choice of a 24-hour time window is typical in systems engineering and operations research applications to an ED, as patient arrival and census patterns typically exhibit a 24-hour cycle (Sir et al., 2017). Assuming that the arrivals and departures on day \( i \) are realizations of inhomogeneous Poisson processes \( \lambda_i(t) \) and \( \delta_i(t) \), \( \lambda_i(t) \) and \( \delta_i(t) \) are estimated from \( \tau_{i,1}, \tau_{i,2}, ..., \tau_{i,N_i} \) and \( \nu_{i,1}, \nu_{i,2}, ..., \nu_{i,M_i} \) respectively, details of which are given in Section 2.1. A functional regression model is used to establish a relationship between \( \lambda_i(t) \) and \( \delta_i(t) \), which is given as:

\[
E(\delta_i(t)|\lambda_i(t)) = \int_{0 \vee (t-T)}^{t} \beta(s,t) \lambda_i(s) ds, \tag{1}
\]

where \( \beta(s,t) \) is the coefficient function and \( a \cap b = \max\{a, b\} \). The intuition behind this model is that the departure rate at time \( t \) is a function of arrival rates observed prior to \( t \), in this case from \( 0 \vee (t-T) \) to \( t \). \( \beta(s,t) \) can be interpreted as the effect the arrival rate at \( s \) has on the departure rate at \( t \) for \( s < t \). For practical purposes, \( T \) is chosen to be arbitrarily large. For example, if known, \( T \) can be set as the upper bound of the service time. This model belongs to the family of functional regression models known as historical functional regression models (Malafit and Ramsay, 2003).

The model allows for the arrival intensity function to vary from day to day. Therefore, we use subscript \( i \) in the arrival intensity function \( \lambda_i(t) \). Thus, allowing the model to be applied to problems where the arrival intensity functions are doubly stochastic inhomogeneous Poisson processes.

As in any other SPC scheme, we first collect samples from the system when it is in-control, i.e., the data is collected when the system is known to be well-behaved. Let \( I_0 \) be the set of in-control samples and let \( \beta(s,t) \) be the coefficient function estimated from \( \hat{\lambda}_i(t) \) and \( \hat{\delta}_i(t) \) for \( i \in I_0 \). Estimation of \( \beta(s,t) \) is described in Section 2.2. Once \( \beta(s,t) \) has been estimated, we want to determine if a new sample \( i \) is in-control or out-of-control, where \( I_1 \) denotes the set of test samples. The test statistic to verify if \( i \in I_1 \) is in-control or out-of-control is given as:

\[
\epsilon_i = \int_0^T \left( \hat{\delta}_i(t) - \int_{0 \vee (t-T)}^{t} \beta(s,t) \hat{\lambda}_i(s) ds \right)^2 dt. \tag{2}
\]

If \( \epsilon_i \) is greater than a threshold \( h \), it indicates that on day \( i \), the system has significantly deviated from its normal performance. The threshold \( h \) is estimated such that the Type-I error rate is at a specified level \( \alpha \).

### 2.1. Estimation of \( \hat{\lambda}_i(t) \) and \( \hat{\delta}_i(t) \)

As mentioned earlier, \( \tau_{i,1}, \tau_{i,2}, ..., \tau_{i,N_i} \) are assumed to be arrival epochs of an inhomogeneous Poisson process with intensity function \( \lambda_i(t) \). We follow the method proposed by Wu et al. (2013) to estimate \( \lambda_i(t) \), which is briefly described in this section. Based on the assumption that \( \nu_{i,1}, \nu_{i,2}, ..., \nu_{i,M_i} \) are also departure epochs for an inhomogeneous Poisson process with intensity function \( \delta_i(t) \), we can use the same method to estimate \( \delta_i(t) \).

The likelihood of \( \lambda_i(t) \) for \( \tau_{i,1}, \tau_{i,2}, ..., \tau_{i,N_i} \) is given as:

\[
L(\lambda_i) \propto \prod_{j=1}^{N_i} \hat{\lambda}_i(\tau_{ij}) \exp \left(- \int_0^T \lambda_i(t) dt \right) dt, \tag{3}
\]

(1) the derivation of this expression can be found in Drazek (2013). The likelihood in Equation (3) can also be written as:

\[
L(C_i, f_i) \propto \prod_{j=1}^{N_i} f_i(\tau_{ij}) C_i^{N_i} \exp(-C_i),
\]

with \( \hat{\lambda}_i(t) = C_i f_i(t) \), for some \( C_i \geq 0 \), such that \( \int_0^T f_i(t) dt = 1 \). The maximum likelihood estimation of \( C_i \) is given as \( \hat{C}_i = N_i \) and estimation of \( f_i(t) \) is equivalent to fitting a density function to \( \tau_{i,1}, \tau_{i,2}, ..., \tau_{i,N_i} \). Wu et al. (2013) estimate \( f_i \) as a kernel density function and following the same idea, \( \hat{\lambda}_i(t) \) is estimated as:

\[
\hat{\lambda}_i(t) = \sum_{j=1}^{N_i} \frac{1}{w_j} K_i \left( \frac{t - \tau_{ij}}{w_j} \right), \tag{4}
\]

where \( K_i \) is the kernel function that satisfies the condition that \( K_i(u) = K_i(-u) \forall u, \int_{-\infty}^{\infty} K_i(u) du = 1 \). (Here the subscript \( s \) in \( K_i \) is used to differentiate from the kernel function \( K \) used estimating departure intensity functions.) In this article we use the rule of thumb, \( w_s = 1.06 \sigma_i N_i^{-1/5} \), where \( \sigma_i \) is the standard deviation of \( \tau_{i,1}, \tau_{i,2}, ..., \tau_{i,N_i} \) (Silverman, 1986). Also, we restrict the analysis to Epanechnikov’s kernel, \( K_i(u) = \frac{3}{4}(1 - u^2)e_{|u| \leq 1} \). Here, \( e_\cdot \) is the indicator function, which is equal to one if the condition within the parenthesis are satisfied and zero otherwise. We use Epanechnikov’s kernel because of its optimal bias-variance tradeoff properties (Benedetti, 1977). However, other kernel functions can also be used. In our study, the choice of kernel functions was not found to be significant in determining the overall performance of the proposed SPC scheme.
As the estimation procedure of \( \hat{\lambda}_i(t) \) from \( \tau_{i,1}, \tau_{i,2}, \cdots, \tau_{i,N_i} \) and \( \hat{\delta}_i(t) \) from \( \nu_{i,1}, \nu_{i,2}, \cdots, \nu_{i,M_i} \) are identical, the departure intensity estimated form \( \nu_{i,1}, \nu_{i,2}, \cdots, \nu_{i,M_i} \) is given as:

\[
\hat{\lambda}_i(t) = \sum_{j=1}^{M_i} \frac{1}{w_i} K_j \left( \frac{t - b_{ij}}{w_i} \right),
\]

where \( K_j \) is also a kernel function, \( w_i = 1.06 \hat{\sigma}_i M_i^{-1/5} \) and \( \hat{\sigma}_i \) is the standard deviation of \( \nu_{i,1}, \nu_{i,2}, \cdots, \nu_{i,M_i} \). For similar reasons, \( K_j \) is also selected as the Epanechnikov's kernel function.

2.2. Estimation of \( \beta(s, t) \)

Given the estimates \( \hat{\lambda}_i(t) \) and \( \hat{\delta}_i(t) \), the functional model regression involves estimation of \( \beta(s, t) \) by minimizing the following:

\[
\sum_{i \in \mathcal{D}} \int_0^T \left( \hat{\delta}_i(t) - \int_0^t \beta(s, t) \hat{\lambda}_i(s) \, ds \right)^2 \, dt.
\]

A few nonrestrictive assumptions are made in solving this minimization problem. First, we constrain \( \beta(s, t) \geq 0 \) to avoid having negative estimates of the departure rates. Also, in most scenarios, a higher arrival rate would not result in a decrease in departure rates. Thus, a negative value of \( \beta(s, t) \) may be unrealistic. Therefore, \( \beta(s, t) \) can be approximated using kernel functions as follows:

\[
\sum_{p=1}^P b_p B_p(s, t),
\]

with \( b_p \geq 0 \) and

\[
B_p(s, t) = \frac{1}{\omega_{p_1} \omega_{p_2}} K_{s_1} \left( \frac{s - s_p}{\omega_{p_1}} \right) \frac{1}{\omega_{p_2}} K_{s_2} \left( \frac{t - t_p}{\omega_{p_2}} \right).
\]

Here \( K_{s_1}, K_{s_2} \) are also kernel functions. The centers of the kernels, \( (s_p, t_p) \) are typically spread uniformly over the region \( \mathcal{D} = \{ (s, t) : t - T \leq s \leq t, s, t \geq 0 \} \). Again we ignore the boundary effect of the kernels and assume \( \beta(s, t) \) is negligibly small in the complement of \( \mathcal{D} \). This is illustrated in Figure 3.

A feasibly large number of kernels is recommended for approximating an irregular \( \beta(s, t) \). To avoid over-fitting, a penalty is added on the curvature of \( \beta(s, t) \). Thus, estimating \( \beta(s, t) \) is given as follows:

\[
\min_b \sum_{i \in \mathcal{D}} \int_0^T \left( \hat{\delta}_i(t) - \int_0^t \beta(s, t) \hat{\lambda}_i(s) \, ds \right)^2 \, dt
+
\phi \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left( \frac{\partial^2 \beta(s, t)}{\partial s^2} \right)^2 + \left( \frac{\partial^2 \beta(s, t)}{\partial t^2} \right)^2 \, ds \, dt,
\]

where \( b = [b_1, \ldots, b_P]^T \), \( b_p \geq 0, \forall p = 1, \ldots, P \) and \( \phi > 0 \) is a smoothing penalty.

2.3. Choice of kernel function in estimation of the coefficient function

In this article, the kernel used in the estimator is restricted to Epanechnikov’s kernel. That is
\[ U_{i,j,p}(t) = \min \{ t, \tau_{i,j} + w_s, s_p + \omega_s \} \]
\[ L_{i,j,p}(t) = \max \{ t - T, \tau_{i,j} - w_s, s_p - \omega_s \} \]

and

\[
\Phi_{p_i,p_j} = \frac{81}{256} \pi^2 (\lambda_{i_1,j_1} \leq 1) \pi^2 (\lambda_{i_2,j_2} \leq 1) \left[ (1 - \Delta_{t,p_i,p_j})(7\Delta^5_{i,p_i,p_j} - \frac{15}{8} \Delta^4_{i,p_i,p_j} + \frac{5}{3} \Delta^3_{i,p_i,p_j} + \frac{8}{15} ) \right] + (1 - \Delta_{t,p_i,p_j})(7\Delta^5_{i,p_i,p_j} - \frac{15}{8} \Delta^4_{i,p_i,p_j} + \frac{5}{3} \Delta^3_{i,p_i,p_j} + \frac{8}{15} ) .
\]

Here \( \Delta_{s,p_1,p_2} = \frac{|s_p - p_1|}{2 \omega_s} \) and \( \Delta_{t,p_1,p_2} = \frac{|t_p - p_2|}{2 \omega_t} \).

**Proof.** See Appendix A.1.

Rather than performing a numerical integration of \( \int_{0}^{T} \beta(s,t)\hat{\lambda}(s)ds \) in Equation (8), using Result 1 is relatively more convenient. A quadrature approximation is then used for the integral in Equation (9) as follows:

\[
\sum_{k=1}^{K} (\tilde{\delta}(u_k) - \sum_{p} b_p \Psi_{i,p}(u_k))^2,
\]

for a dense grid of points \( 0 = u_1 \leq u_2 \leq \cdots \leq u_L = T \). Then the optimization problem in Equation (9) reduces to:

\[
\min_{y} \sum_{i \in I_{\omega}} b_i'(X_i'y_i)X_i - 2y_i'X_i b + y_i'y_i + b_i'y_i + b_i'y_i \Phi b,
\]
such that

\[
b_i \geq 0, \forall p = 1, 2, \ldots, P.
\]

(11)

where \( y_i = [\tilde{\delta}(u_i)]_{1 \times 4} \) and \( X_i = [\Psi_{i,p}(u_i)]_{4 \times P} \).

The term \( \| \tilde{\lambda}_{i_1,j_1} \|_{1 \times 1} \) and \( \| \tilde{\lambda}_{i_2,j_2} \|_{1 \times 1} \) makes \( \Phi_{p_1,p_2} = 0 \) if \( s_{p_1} - s_{p_2} > 2 \omega_s \) or \( t_{p_1} - t_{p_2} > 2 \omega_t \). Thus, for suitable choices of \( \Delta_{s,p_1,p_2} \) and \( \Delta_{t,p_1,p_2} \), \( \Phi \) can be positive definite. Therefore, besides avoiding over-fitting, the additional benefit of including the penalty matrix \( \Phi \) it that it would ensure that Equation (11) is a strictly convex quadratic programming problem. However, in this process, there is a chance that \( \Phi \) might end up having all non-diagonal elements equal to zero. As a result, it would only penalize the sum of squares of \( b_p \). It might be better to penalize coefficient \( b_p \) not to be different for \( (s_{p_1}, t_{p_1}) \) that are close to each other, which would further ensure the smoothness of \( \beta(s,t) \).

As an illustration, consider a simple case where there are only five kernel centers in \( \beta(s,t) \), i.e., \( \beta(s,t) = \frac{5}{4} b_p B_p(s,t) \). Letting \( (s_{p_1}, t_{p_1}) \) to be equispaced as shown in Figure 4(a) and (b), with \( s_2 - s_1 = s_4 - s_3 = s_5 - s_4 = s_5 \) and \( t_2 - t_1 = t_4 - t_3 = t_5 - t_3 = t_4 = t_5 \). In Figure 4(a), \( \omega_s = (s_2 - s_1)/2, \omega_t = (t_2 - t_1)/2 \) and in Figure 4(b), \( \omega_s = (s_4 - s_1)/4, \omega_t = (t_2 - t_1)/4 \), which are shown as blue rectangles. When \( \omega_s = (s_2 - s_1)/2 \) and \( \omega_t = (t_2 - t_1)/2 \). \( \Phi \) is a diagonal matrix with diagonal elements given as 27/80. For \( \omega_s = (s_2 - s_1)/4 \) and \( \omega_t = (t_2 - t_1)/4 \):

\[
\Phi_{p_1,p_2} = \begin{bmatrix} 0.3375 & \forall p \in \{ 1, 2, 3, 4, 5 \} \\
0.125 & \Phi_{3,4} = \Phi_{4,3} = 0.013 & 8194 \\
0.031 & 712 & 96 \\
\Phi_{1,3} = \Phi_{5,5} = 0.031 & 712 & 96 \\
\Phi_{p_1,p_2} = \Phi_{p_2,p_1} & \forall p_1, p_2 \in \{ 1, 2, 3, 4, 5 \}
\end{bmatrix}
\]

for which \( \Phi \) is also a positive definite matrix. However, it should be noted that when \( \omega_s = (s_2 - s_1)/4, \omega_t = (t_2 - t_1)/4 \) rather than \( \omega_s = (s_2 - s_1)/2, \omega_t = (t_2 - t_1)/2 \), we penalize coefficient \( b_p \) not to be different for \( (s_{p_1}, t_{p_1}) \) that are adjacent to each other.

The following Lemma gives a rule of thumb for selecting \( \Delta_{s,p_1,p_2} \) and \( \Delta_{t,p_1,p_2} \): It guarantees that the optimization problem in Equation (11) is strictly convex.

**Result 2.** If \( (s_{p_1}, t_{p_1}) \) are equispaced, i.e., \( |s_{p_1} - s_{p_2}| = \min_{p_1 \neq p_2} |s_{p_1} - s_{p_2}| \) \( \forall p_1, p_2 \in \{ 1, 2, \ldots, P \} \) and \( |t_{p_1} - t_{p_2}| = \min_{p_1 \neq p_2} |t_{p_1} - t_{p_2}| \) \( \forall p_1, p_2 \in \{ 1, 2, \ldots, P \} \), then \( n = 0, 1, 2, \ldots, 1 \) and for \( \Delta_{s,p_1,p_2} = \Delta_{t,p_1,p_2} > 0.85 \), \( \Phi \) is positive definite.

**Proof.** See Appendix A.2.

It should be noted that this is a rough rule of thumb and indeed, for the cases studied in this paper, \( \Phi \) was found to be positive definite for \( \Delta_{s,p_1,p_2} = \Delta_{t,p_1,p_2} = 0.5 \).
2.4. Relationship with infinite server queues

An infinite server queue is an approximation of a service system with a large number of processing units working in parallel (Eick et al., 1993). This model has been widely studied and is a good approximation for a well-staffed service system where the wait time is small compared with service times. A typical application of this model is to find optimal staffing in service systems experiencing time-varying arrival rates (Massey, 2002). $M_t/G_t/\infty$ is used to denote an infinite server queue that experiences an inhomogeneous Poisson arrival rate and the service times follow some arbitrary probability distribution, which could also vary based on time. The following proposition establishes a relationship between the functional regression model in Equation (1) and $M_t/G_t/\infty$ queues.

Lemma 3. Let $\lambda(t)$ be the arrival rate to a $M_t/G_t/\infty$ queue and $g(s, t)$ be the density function of the random variable $S(t)$, where $S(t)$ is the service time for arrival occurring at time $t$. Then the departure process is a Poisson process with intensity $\delta(t) = \int_{-\infty}^{t} \beta(s, t) \lambda(s)ds$, where $\beta(s, t) = g(t-s, s)$.

Proof. See Appendix A.3

Assuming the service system is well-staffed during the in-control periods, infinite server queues are a good approximation of the queuing dynamics. Thus, the SPC scheme based on this functional regression model can be applied to a wide variety of service systems, as long as they are known to be adequately staffed under normal working conditions. Given the similarity between the functional regression method and the infinite server queues, we shall study $M_t/G_t/\infty$ queues in greater detail in Section 3.

If $\lambda(t)$ for a infinite server queue with exponential service time is estimated from the arrival time epochs using Equation (4), using the following result the expected departure intensity function $E(\delta_i(t)|\lambda(t))$ can be readily calculated.

Result 4. If

$$\lambda(t) = \sum_{j=1}^{N} \frac{1}{w_j} K_j \left( \frac{t - \tau_j}{w_j} \right)$$

and if the service time $s$ for an arrival at time $t$ has a density function:

$$g(s, t) = \sum_{q=1}^{Q} \prod_{i=1}^{N} \frac{1}{\mu_q} \exp \left( -\frac{s}{\mu_q} \right),$$

for $0 \leq t_1 \leq t_2 \leq \cdots \leq t_Q \leq T$:

$$E(\delta(t)|\lambda(t)) = \frac{3}{4w_s^2} \sum_{q=1}^{Q} \sum_{i=1}^{N} \int_{0}^{T} \left( \exp \left( -\frac{t}{\mu_q} \right) \left( -w_s^2 + \mu^2 + (w_t + \mu - (t - (0\nu(t_t - w_t)))^2) \right) \right. \left. - e^{-(0\nu(t_t - w_t))} \left( -w_s^2 + \mu^2 + (w_t + \mu - (t \wedge (t_t + w_t) / \mu_q))^2) \right) \right) dx.$$ (12)

Proof. See Appendix A.4.

Based on this result, for the special case of infinite server queues with exponential service times, the analytical expression (12) can be used instead of the functional regression model in Equation (1). We use this result to develop an alternate SPC scheme and compare it to the performance of the functional regression scheme. In Section 3, it will be shown that the performance of the functional regression based SPC scheme and the SPC scheme based on Result 4 are similar.

These SPC schemes are summarized in Section 2.5.

2.5. SPC algorithms

The method of fitting a functional regression model to the “in-control” data is described in earlier sections. The SPC scheme based on this functional regression model is described in Algorithm 1. In the remainder of this article we refer to this SPC scheme as FR chart (functional regression-based control chart).

Algorithm 1. Function regression based SPC scheme (FR-chart)

1: procedure PHASE I OF SPC SCHEME
2: Collect the arrival and departure time instance, $\tau_1, \tau_2, \cdots, \tau_i, N_i$ and $\nu_1, \nu_2, \cdots, \nu_i, M_i$, when the system was in-control.
3: Let the index set of in-control samples be $I_0$.
4: Use the procedure described in Section 2 to estimate $\beta(s, t)$.
5: Monitoring statistic

$$e_i = \int_{0}^{T} \left( \hat{\beta}(t) - \int \hat{\beta}(s, t) \lambda(s) ds \right)^2 dt$$

for $i \in I_0$ is calculated.
Table 1. Type II error rate - infinite servers (\(\mu_0 = 0.5\) hours).

| \(\mu\) | FR chart | IS chart | Shewhart chart |
|---|---|---|---|
| 1.1\(\mu_0\) | 0.86 | 0.87 | 0.84 |
| 1.2\(\mu_0\) | 0.82 | 0.82 | 0.72 |
| 1.3\(\mu_0\) | 0.75 | 0.77 | 0.57 |
| 1.4\(\mu_0\) | 0.68 | 0.72 | 0.41 |
| 1.5\(\mu_0\) | 0.62 | 0.65 | 0.28 |

Table 2. Type II error rate - infinite servers (\(\mu_0 = 1.0\) hours).

| \(\mu\) | FR chart | IS chart | Shewhart chart |
|---|---|---|---|
| 1.1\(\mu_0\) | 0.85 | 0.86 | 0.85 |
| 1.2\(\mu_0\) | 0.80 | 0.81 | 0.74 |
| 1.3\(\mu_0\) | 0.74 | 0.77 | 0.60 |
| 1.4\(\mu_0\) | 0.68 | 0.71 | 0.44 |
| 1.5\(\mu_0\) | 0.62 | 0.65 | 0.33 |

6: Either use simulation or bootstrapping to estimate \(h\) such that \(P(e_i > h) = \alpha\), given that the data is “in-control”. Here \(\alpha\) is a specified false alarm rate.

7: **procedure Phase II of SPC scheme**
8: Collect test sample \(i \in I_1\).
9: if \(e_i \geq h\) then
10: sample \(i\) is out of control

Result 4 can be used to estimate the expected departure intensity function \(E(\delta_i(t)|\lambda_i(t))\) for infinite server queues with exponentially distributed service times. Therefore, in order to compare the performance of the FR-chart in numerical experiments involving infinite server queues with exponentially distributed service times, another SPC scheme is used, where \(E(\delta_i(t)|\lambda_i(t))\) is calculated using Result 4 rather than the functional regression model. More specifically, in Step 5 of Algorithm 1, Result 4 is used to estimate 
\(e_i = \int_0^T (\delta_i(t) - E(\delta_i(t)|\lambda_i(t)))^2 dt\) for \(i \in I_0\). This SPC scheme is referred to as the IS chart (infinite server - control chart).

The performance of the SPC schemes described above is also compared with a Shewhart chart designed for detecting an increase in mean service time. If \(\gamma_1, \gamma_2, \ldots, \gamma_M\) are the service time for each entity processed in the service system for \(i \in I_0\), \(\overline{\gamma}\) is estimated as
\[
\bar{\gamma} = \frac{1}{N(I_0)} \sum_{i \in I_0} \sum_{j=1}^{M_i} \gamma_{ij}
\]
where \(N(I_0)\) is the cardinality of set \(I_0\). A test sample \(i \in I_1\) is classified as out-control by the Shewhart chart for service time if
\[
\left( \frac{1}{M} \sum_{j=1}^{M} \gamma_{ij} - \overline{\gamma} \right)^2 \geq h^2.
\]

\(h^2\) is selected such that this chart has the same false alarm rate as the functional regression-based SPC scheme.

### 3. Simulation Study

As mentioned in Section 1, most real service systems experience arrival rates that vary from day to day. In the simulation studies presented in this section, the arrival intensity of day \(i\) is given as:
\[
\lambda_i(t) = \exp(\mu(t) + \varepsilon_{i,1}\xi_1(t) + \varepsilon_{i,2}\xi_2(t)),
\]
where,
\[
\mu(t) = 0.125 - 0.125 \cos\left(\frac{2\pi}{24}t - 0.500 \sin\left(\frac{2\pi}{24}t\right)\right) + 0.125 \cos\left(\frac{4\pi}{24}t\right)
\]
\[
\xi_1(t) = -0.25 - 0.125 \cos\left(\frac{2\pi}{24}t\right) + 0.125 \sin\left(\frac{2\pi}{24}t\right)
\]
\[
\xi_2(t) = 0.10 \cos\left(\frac{2\pi}{24}t\right) - 0.25 \sin\left(\frac{2\pi}{24}t\right) - 0.25 \cos\left(\frac{4\pi}{24}t\right)
\]

and \(\varepsilon_{i,1}\) and \(\varepsilon_{i,2}\) are normally distributed random variables with mean zero and standard deviation one. This stochastic process is shown in Figure 5. The solid blue line in Figure 5 shows the mean arrival intensity and the shaded region around it represents the variation (one standard deviation) in intensity for each time of the day. In this example the variation in arrival intensity is modeled by \(\varepsilon_{i,1}\) and \(\varepsilon_{i,2}\). The process follows a cyclic pattern repeating in each 24-hour period. The rates increase during the morning hours and decrease during the evening. This is an arrival process that varies within a day, which is described by functions \(\lambda(t), \xi_1(t), \text{and } \xi_2(t)\), as well as it varies from day to day, which is due to the random variables \(\varepsilon_{i,1}\) and \(\varepsilon_{i,2}\). This arrival process was selected because it is simple to specify and yet the magnitude and variation of the intensity function are similar to a typical arrival process observed in an ED that motivated this study (see Figure 1(a)).

We study two categories of service systems with infinite servers and with finite servers. For each case, the system experiences arrivals with intensities that are i.i.d samples of the intensity process in Equation (13) and has exponentially distributed service times. For simplicity we start the simulation with an empty service system, i.e., the number of entities being processed is zero, and we ignore the departures from the system that occur after \(T = 24\) hours. Various types of queues are simulated, and time stamp data on various arrival and departure events \(\tau_{i,1}, \tau_{i,2}, \ldots, \tau_{i,N_i}\) and \(\nu_{i,1}, \nu_{i,2}, \ldots, \nu_{i,M_i}\) for \(i \in I_0\) are collected. \(\beta(s, t)\) is estimated based on these simulated values.

The centers of the bivariate kernel function \(B_p(s, t)\) that approximate \(\beta(s, t) = \sum_{p=1}^{P} b_p B_p(s, t)\) is chosen such that \(\{(s_p, t_p)\} = \{(n/2, \frac{m}{2}) : 0 < (m/2 - 10) \leq n/2 \leq m/2; 0 \leq n/2, m/2 \leq 24\}\) for \(n, m = 0, 1, 2, \ldots\) and \(p = 1, 2, \ldots, P\). This results in \(P = 484\) kernels. Thus, an equispaced grid of kernel centers \((s_p, t_p)\) with \(\min_p(s_p, t_p) = \min_p s_p = -s_p, \max p(s_p, t_p) = \min_p s_p = -s_p, \Delta p(s_p, t_p) = 0.5\) is obtained. Here, \(\omega_x = \omega_y = 0.5\) is selected and thus \(\min_p(s_p, t_p)\Delta p(s_p, t_p) = \min_p(s_p, t_p)\Delta p = 0.5\). The resulting \(\Phi\) is positive definite. The penalty coefficient \(\phi > 0\) is selected by cross-validation. From \(\phi \in \{10^{-6}, 10^{-5}, 10^{-4}\}, \phi = 10^{-2}\) was found to have minimum cross-validation error in a sample size of 1000, for all the models studied in this section. Also, for the estimation
of signaling thresholds $h$, $h_{\text{in}}$, and $h_{\text{out}}$ the false alarm rate was set to 0.1 using a simulated sample of size 10,000.

### 3.1. Infinite server queues

In this section, infinite server queues with arrival intensity function as in Equation (13) are considered. For simplicity, it is assumed that the service time distributions follow an exponential distribution. Please see Appendix B.1 for an illustration comparing the predicted departure rate and true departure rate for this simulation model. To compare the various SPC methods, consider the case were service times are exponentially distributed with mean $\mu$, which is the same for all entities arriving at any time. The service rate is decreased and the performance of the FR chart, IS chart and the Shewhart chart for detecting these changes are compared.

The motivation behind comparing the FR chart with the IS chart is to compare the functional regression-based estimation of departure intensity function with true estimation of expected departure intensity function. It provides a benchmark for comparing the accuracy with which $\hat{\beta}(s,t)$ approximates the true queuing process. The Shewhart chart uses the service time information. It is expected to be most sensitive to changes in the mean. Nevertheless, the Shewhart chart is also compared with other charts to provide a benchmark.

Table 1 shows the Type II error rate, i.e., the probability of not detecting the change, when the mean of service times increases from $\mu_0 = 0.5$ hours to $\mu$. Table 2 shows the same results for $\mu_0 = 1.0$ hours.

Comparison with the Shewhart chart provides a measure of loss of statistical power for detecting an increase in mean service times. The Shewhart chart, which uses the mean service time information directly, has the best performance of the three monitoring schemes. This is indicated by lower Type II error rates of the Shewhart chart. Considering the fact all $N_i$ customers exhibit an increase in service rate, a Shewhart chart would be expected to be more sensitive to such simple increase in mean service rate. This is because the other monitoring schemes use the estimated arrival and departure rates to study the effect of an increase in mean service time. However, for a small increase in $\mu$, the performances of the FR chart and the IS chart are comparable to that of the Shewhart chart.

From Table 1 and Table 2, we can see that the FR chart and IS chart have comparable performances. In fact, the FR chart seems to perform slightly better than the IS chart. Since there is a bias in estimation of $\hat{\lambda}_i(t)$, due to sample size ($N_i$), it results in an estimation error for $\hat{\delta}_i(t)$ based on Result 4. However, fitting the functional regression model compensates for this bias. Thus, it gives a slightly more accurate estimation of $\hat{\delta}_i(t)$ for these queuing models, which results in smaller Type II error rates. The results presented in this section are based on a 24-hour time window. We also performed additional simulations studies with 6-hour, 12-hour, or 18-hour time windows. The simulation results show that the performance of the FR chart does not deteriorate for different time windows. On the other hand, the performance of the Shewhart chart deteriorates as the number of samples per time unit of simulation decreases when smaller time windows are used.

These case studies provide a benchmark for the functional regression-based monitoring scheme, using the simplest case where each arriving entity has an exponentially distributed service system. However, most real-world service systems not only experience time-varying arrivals, but also the service time distribution varies over time. Hospital EDs provide a classic example, due to since the number of staff in the ED varying...
Table 5. Type II error rate - infinite servers for service time distribution as in Equation (17).

| D (hours) | FR chart | IS chart | Shewhart chart |
|-----------|----------|----------|---------------|
| 0.1       | 0.89     | 0.89     | 0.90          |
| 0.2       | 0.87     | 0.86     | 0.89          |
| 0.3       | 0.82     | 0.82     | 0.88          |
| 0.4       | 0.77     | 0.76     | 0.85          |
| 0.5       | 0.69     | 0.69     | 0.83          |

Figure 8. Shift in optimal staffing.

Table 6. Type II error rate for detecting shifts in staffing level.

| D (hours) | FR chart | Shewhart chart |
|-----------|----------|---------------|
| 1         | 0.89     | 0.90          |
| 2         | 0.85     | 0.91          |
| 3         | 0.80     | 0.90          |
| 4         | 0.76     | 0.90          |
| 5         | 0.72     | 0.90          |

during the day. In order to study service systems that exhibit time-varying service time distributions, we perform three other case studies. First, consider a case where the in-control system has service time density given as:

$$g_0(s, t) = 2e^{-2s}$$

and out-of-control service time distribution is given by

$$g_1(s, t) = \mathbb{I}_{(0 \leq t \leq 10)}2e^{-2s} + \mathbb{I}_{(10 \leq t \leq 15)}e^{-s} + \mathbb{I}_{(15 \leq t \leq 24)}2e^{-2s}$$

These out-of-control scenarios represent cases where the service time density changes due to some operational changes:

When the delays in service tend to occur during the hours when the system is suppose to be faster (with respect to the in-control characteristic of the system), it can indicate that certain measures need to be taken to address this issue.

For an infinite server queue with in-control service time distribution as in Equation (16) and out-of-control distribution (17), we compare the Type II error rate of the FR chart, the IS chart and the Shewhart chart for different magnitudes of shift D (See Figure 6). These results are presented in Table 4. Again, we see that the FR chart and IS chart have similar Type II error rates. More importantly, the Shewhart chart is unable to detect these changes.

Lastly, consider another service scenario where the in-control service time distribution is given by

$$g_0(s, t) = e^{-s}$$

and out-of-control service time distribution is given by

$$g_1(s, t) = \mathbb{I}_{(0 \leq t \leq 12)} \frac{1}{1 - D} e^{-\frac{t}{D}} + \mathbb{I}_{(12 \leq t \leq 24)} \frac{1}{1 + D} e^{-\frac{t}{D}}$$

The mean service time for in-control and out-of-control scenarios are illustrated in Figure 7. Here, the mean service times are lower than expected during the morning hours and the system gets slower during the afternoon. We compare the Type II error rate of the FR chart, the IS chart and the Shewhart chart for an infinite server queue where this phenomenon is simulated. For different values of D, the Type II error rates are given in Table 5. As can be seen from the results, this is another type of change that is difficult to detect using a simple Shewhart chart. The proposed FR chart, on the other hand, can better detect the structural changes in service time, and therefore, can be extremely useful for operations managers of complex service systems.

Various insights obtained from the simulation studies about the effectiveness of the functional regression based SPC are now summarized:

1. For infinite server queues with constant, as well time-varying, arrival rates, the FR chart and IS chart perform
similarly. This validates the model fitting procedure, i.e., estimation of $\beta(s, t)$. Since the IS chart is based on the exact estimation of $\tilde{\delta}_i(t)$ from $\tilde{\lambda}_i(t)$, the similar performances of the FR and IS charts shows that $\beta(s, t)$ is a good approximation of the queuing dynamics for infinite server queues.

2. A Shewhart chart that directly uses the service time information is more sensitive than the FR and IS charts when the overall mean of the service time increases. This is due to information lost by indirectly studying the queuing process using arrival and departure information rather than service times.

3. Simulation studies show that the FR chart is more effective in detecting structural changes in service time distribution compared with the Shewhart chart. The Shewhart and FR charts have different uses. Therefore, it is recommended that a Shewhart chart is used for mean length of stay data, which is already widely used in EDs, and that it is also used along with the FR chart because it is more effective in detecting scenarios such as when staffing levels are no longer optimal. This will be further discussed in Section 3.2. These changes do occur in real-world service systems and will be discussed using a case study in Section 4.

### 3.2. Detecting when staffing levels are no longer optimal

Infinite server models have been used to optimize staffing in service systems that experience time-varying arrivals (Jennings et al., 1996). Following Jennings et al. (1996), we studied the number of busy servers in an infinite server queue with exponentially distributed service times having a mean of $\mu = 1.0$ in a simulation. The number of busy servers at the different times of the day gives an estimate of what would be the ideal number of servers needed to make sure that the service system has minimal waiting. An illustration of the method used to obtain the number of servers available at different times of the day is discussed in Appendix B.2.

To compare the statistical monitoring algorithms, the number of servers available at different times of the day shift is varied (See Figure 8). The blue lines indicate the “in-control” staffing levels and the blue shaded region indicate “out-of-control” or suboptimal staffing levels. The overall shift in staffing levels, for different magnitudes of $D$, would be representative of possible suboptimal staffing levels.

Table 6 shows the Type II error rate for the FR chart and Shewhart charts in the context of detecting different suboptimal staffing levels. Again, this shows that the Shewhart chart is unable to detect suboptimal staffing levels, whereas the functional regression-based monitoring scheme can detect it. This simulation study quantifies the sensitivity of the proposed monitoring scheme to detect suboptimal staffing levels, which is one of the motivations for developing this statistical monitoring framework. The ability of the FR chart to detect changes in staffing levels has many practical applications. If an optimized staffing schedule was utilized during the in-control phase, out-of-control signals would indicate that management needs to reevaluate staffing levels. Further discussions and practitioner guidelines are given in Section 4.

### 3.3. Robustness of functional regression-based monitoring scheme

In Section 3.1, the FR chart and IS chart were shown to have similar performances in detecting various kinds of changes in service time distribution. In this section, we simulate a single server queue with exponentially distributed service times and mean $\mu_0 = 1$. The arrival rates are as in Equation (13). Then the mean service times are increased, and the Type II error rate for IS and FR charts are compared. The results are given in Table 7 and the FR chart has lower Type II error rates than the IS chart.

In the simulation study in Section 3.1, the departure process was an inhomogeneous Poisson process (because the simulation was based on infinite server queues). For single server queues, this assumption no longer applies. This result shows that although the performances of the FR and IS charts are quite similar when the system has an infinite number of servers, the FR chart is more robust when the assumption that the departure process is inhomogeneous Poisson process is no longer satisfied. This result implies that the functional regression-based scheme is more effective for real-world applications that deviate from the infinite server assumptions.

The purpose of the proposed functional regression-based SPC scheme is to identify patterns of change in mean service times, which may or may not manifest as a change in the magnitude of mean service times. For example, if changes in service times as shown in Figure 6 are recurrent, it could imply that the peak rush hours have moved, which might imply that the
staffing levels need modification. These insights are valuable in a many service system. In the next section, we illustrate this using a case study based on patient arrival and departure data from the ED of a major academic medical center.

4. Case study: The ED of a major academic medical center

In this section, the FR chart is evaluated using data collected from the Mayo Clinic ED in Rochester, MN, an academic Level I trauma center. Through discussion with clinicians and analysis, the data from January 1, 2013, to December 31, 2013, was used to design the in-control sample. First, the arrival and departure time stamps of patients staying less than 5 hours were used and then out of these the days with average length of stay less than the 0.95 percentile and greater than 0.95 percentile are dropped. The arrival and departure time stamps of the remaining patients were used as the in-control sample. In the process, data from 40 days was dropped. This process of selecting the in-control sample is commonly used in the design of control charts (Montgomery, 2001).

Figure 1(a) and 1(b) show the arrival and departure intensity for this in-control data. Both Figure 1(a) and 1(b) exhibit a clear pattern over the duration of a day, i.e., the intensity is low during early hours of the day and high during the afternoons, before it finally decreases in the evening hours. Figure 9 shows the length of stay of patients in the ED, arriving at different times of the day. Again, the solid blue line represents the mean length of stay and the shaded region around it shows the variation. From Figure 9, we can see that there is some evidence to indicate that the patients arriving during the afternoon hours tend to have a longer length of stay. However, the overall variability in patient length of stay is to large to allow the identification of a clear temporal pattern of length of stay. As mentioned in Section 1, the temporal correlation and care delivery process of individual patients makes it harder to build statistical models for monitoring length of stay. As with the simulation studies, the initial analysis compares the proposed functional regression-based monitoring scheme with a Shewhart chart for monitoring the average length of stay per day. All SPC schemes were designed to have a false alarm rate of 0.1.

The same functional regression model as in Section 2 is used to estimate $\beta(s, t)$. Again, the centers of the bivariate kernel function $B_p(s, t)$ that approximate $\beta(s, t) = \sum_{p=1}^{P} b_p B_p(s, t)$ is chosen such that $\{(s_p, t_p)\} = \{(n/2, m/2) : 0 \leq (m/2 - 10) \leq n/2 \leq m/2; 0 \leq n/2, m/2 \leq 24\}$ for $n, m = 1, 2, \ldots$ and $p = 1, 2, \ldots, P$. Since the cross-validation error is non-decreasing, it can be said that the current choice of 484 bivariate kernels as a basis for $\beta(s, t)$ does result in overfitting the in-control data. Although this number could be increased to estimate a more accurate $\beta(s, t)$, here the analysis is restricted to the current selection of basis functions, as the focus is on detecting changes in service times.

We present results for the case where the test data set, $X_1$, is the data collected from March 2014. Please see Appendix B.3 for a figure of the test statistic for days in March 2014. Samples in the shaded blue regions are the ones that are classified as in-control. It can be seen that March 8, 12, 31 are classified out-of-control by the SH chart. From Section 3, it is known that the SH-chart is sensitive to changes in an overall increase in the length of stay of patients in the ED. The mean length of stay of patients in 2013 was 3.5 hours. On March 8 and March 12 the length of stay was 2.82 hours and 4.45 hours. Therefore, March 8 and March 12 are signaled out-of-control by the SH-chart but not the FR chart. However, the FR chart signals March 14, 21, 22, 23, and 31 as out-of-control. To get further insight into the reason for this, we also plot the expected and true departure rate for March 21 in Figure 10. It can be seen that the departure rate was close to the expected departure rate in the early hours of March 21, but it decreased for a short period during noon. However, the average length of stay of patients on this day, which was 3.8 hours, was not very large compared with the overall mean length of stay in the in-control data, which was 3.5 hours.

The average length of stay for patients coming to an ED of a hospital is an important metric that indicates the level of patient satisfaction, as well as the efficiency of day-to-day operations, and a Shewhart chart designed for detecting changes to the average length of stay can be useful for monitoring changes in patient length of stay. However, as shown in the data from March 2014, the average length of stay may not always indicate a change in the day-to-day operation, as observed on March 21, 2014. Sometimes fluctuation in the service rate can be detrimental to the quality of service. For example, if the pattern observed on March 21 is recurrent, it might indicate the need to increase staffing levels around midday.

Service systems that have to respond to the stochastic nature of demand are forced to make decisions that often deviate from long-term tactical decisions, e.g., staffing decisions in EDs. EDs study patient demand patterns and determine staffing levels for times of day, days of the week and seasonal variation to meet certain service level thresholds.
However, demand fluctuates, and there are times when an ED fails to meet these service thresholds. Such variations are often perceived by the front-line staff, namely the physicians and nurses that are working shifts. However, the perceptions vary across the individuals, and ED operations managers typically lack a systematic and standardized data-driven evidence base to make informed changes to staffing levels. Although the methodology described in this article focuses on 24-hour monitoring, different time windows can be selected to monitor the system.

The SPC scheme developed here provides a mechanism to continuously monitor the performance of the ED and detect when it becomes out-of-control. The proposed SPC only requires simple arrival and departure level time stamps. Though overall patient population was considered in the numerical experiment, the scheme can be easily adapted for specific patient subpopulations, for example, for different Emergency Severity Index level patients. Detection of such recurrent patterns of deviation from expected departure rates indicates the need to change staffing levels. Hence, any ED could use the SPC scheme and determine if there is a need to perform staffing redesign.

5. Conclusions

In this article, a functional regression-based SPC scheme is developed using in-control data; a functional regression model is used to estimate the expected departure rate for observed arrivals and then measure the difference between the expected and observed departure intensities. This metric lets operations managers identify specific aspects of the service process that can be improved. Although other SPC schemes can be designed to identify changes such as an increase in mean service times, they may not be as insightful as compared with the proposed functional regression-based SPC scheme. The proposed scheme is based on the conditional expectation of departure intensities, as detecting a delay in service times due to a surge in patient registrations is not very useful. It is critical to detect structural changes in service times that may not manifest as an increase in mean service times. The efficacy of the proposed method is demonstrated via a simulation study as well as a real-world case study concerning the ED of a hospital.

There are many possible extensions of the current work. Further research is needed to establish the asymptotic properties of the functional regression scheme, such as convergence of \( \beta(s, t) \) to \( g(s, t) \) and to develop SPC schemes where, instead of a fixed \( \beta(s, t) \), different \( \beta(s, t) \) are estimated based on patient complexity or arrival patterns. In future, it would be interesting to apply this nonparametric data-driven method to traditional applications such as staffing optimization and other queuing networks. We would also like to extend the proposed method to design exponentially weighted moving average and cumulative sum control charts. Lastly, a real-time monitoring scheme that provides a continuous performance measure and does not just alarm at the end of a 24-hour period would be an important extension of the current work.

Funding

This work is funded in part by the Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery.

References

Ahmad, N., Ghani, N.A., Kamil, A.A. and Tahar, R.M. (2015) Modelling the complexity of emergency department operations using hybrid simulation. *International Journal of Simulation and Process Modelling*, 10(4), 360–371.

Benedetti, J.K. (1977) On the nonparametric estimation of regression functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 248–253.

Bertsimas, D. and Vinh Doan, X. (2010) Robust and data-driven approaches to call centers. *European Journal of Operational Research*, 207(2), 1072–1085.

Bhat, U.N. (1987) A sequential technique for the control of traffic intensity in Markovian queues. *Annals of Operations Research*, 8(1), 151–164.

Bhat, U.N. and Subba Rao, S. (1972) A statistical technique for the control of traffic intensity in the queuing systems m/g/1 and gi/m/1. *Operations Research*, 20(5), 955–966.

Callahan, C.D. and Griffen, D.L. (2003) Advanced statistics: Applying statistical process control techniques to emergency medicine: A primer for providers. *Academic Emergency Medicine*, 10(8), 883–890.

Chen, N., Yuan, Y. and Zhou, S. (2011) Performance analysis of queue length monitoring of m/g/1 systems. *Naval Research Logistics*, 58(8), 782–794.

Chen, N. and Zhou, S. (2015) CUSUM statistical monitoring of m/m/1 queues and extensions. *Technometrics*, 57(2), 245–256.

Defraeye, M. and van Nieuwenhuyse, T. (2016) Staffing and scheduling under nonstationary demand for service: A literature review. *Omega*, 58, 4–25.

Derlet, R.W. (2002) Overcrowding in emergency departments: Increased demand and decreased capacity. *Annals of Emergency Medicine*, 39(4), 430–432.

Drazeck, L.C. (2013) Intensity estimation for Poisson processes. PhD thesis, The University of Leeds, UK.

Eick, S.G., Massey, W.A. and Whitt, W. (1993) The physics of the m†g/∞ queue. *Operations Research*, 11(4), 731–742.

Goodwin, S., Lin, Y.L., Singh, S. and Kuo, Y.F. (2013) Variation in length of stay and outcomes among hospitalized patients attributable to hospitals and hospitalists. *Journal of General Internal Medicine*, 28(3), 370–376.

Huang, X., Zhou, Q., Zeng, L. and Li, X. (2017) Monitoring spatial uniformity of particle distributions in manufacturing processes using the k function. *IEEE Transactions on Automation Science and Engineering*, 14(2), 1031–1041.

Ibrahim, M.R. and Whitt, W. (2009) Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science*, 55(10), 1729–1742.

Jennings, O.B., Mandelbaum, A., Massey, W.A. and Whitt, W. (1996) Server staffing to meet time-varying demand. *Management Science*, 42(10), 1383–1394.

Jin, N., Zhou, S., Chang, T.S. and Huang, H.H.H. (2008) Identification of influential functional process variables for surface quality control in hot rolling processes. *IEEE Transactions on Automation Science and Engineering*, 5(3), 557–562.

Joint Commission. (2019) Specifications manual for national hospital inpatient quality measures, version 5.5a. https://www.jointcommission.org/specifications_manual_for_national_hospital_inpatient_quality_measures.aspx.

Liao, H. and Sun, J. (2011) Nonparametric and semi-parametric sensor recovery in multichannel condition monitoring systems. *IEEE Transactions on Automation Science and Engineering*, 8(4), 744–753.
Malfait, N. and Ramsay, J.O. (2003) The historical functional linear model. *Canadian Journal of Statistics, 31*(2), 115–128.
Massey, W.A. (2002) The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems, 21*(2–4), 173–204.
Matteson, D.S., McLean, M.W., Woodard, D.B. and Henderson, S.G. (2011) Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics, 5*, 1379–1406.
Montgomery, D.C. (2001) *Introduction to Statistical Process Control*. John Wiley, New York, NY.
Ramsay, J., Hooker, G. and Graves, S. (2009) *Functional data analysis with R and MATLAB*. Springer Science & Business Media, New York, NY.
Saghañian, S., Austin, G. and Traub, S.J. (2015) Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering, 5*(2), 101–123.
Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, CRC Press, Boca Raton, FL.
Sir, M.Y., Nestler, D., Hellmich, T., Das, D., Laughlin Jr, M.J., Dohlman, M.C. and Pasupathy, K. (2017) Optimization of multidisciplinary staffing improves patient experiences at the Mayo clinic. *Interfaces, 47*(5), 425–441.
Wu, S., Müller, H. and Zhang, Z. (2013) Functional data analysis for point processes with rare events. *Statistica Sinica, 23*(1), 1–23.