Can customer arrival rates be modelled by sine waves?

Ningyuan Chen, Donald K.K. Lee*, Haipeng Shen
HKUST, Yale University, HKU

Customer arrival patterns observed in the real world typically exhibit strong seasonal effects. It is therefore natural to ask: Can a nonhomogeneous Poisson process with a rate that is the simple sum of sinusoids provide an adequate description of reality? We empirically investigate this question in two settings of interest to operations scholars: Patient arrivals to an emergency department and customer calls to a call centre. Our study relies on a new sinusoidal rate model that generalizes the truncated Fourier series specifications of Green and Kolesar (1991) and Eick et al. (1993) while retaining their analytic tractability. We find that the model is consistent with arrivals data from the emergency department as well as from the call centre. Taken together, the flexibility and tractability of the sinusoidal specification suggest that it is a worthy workhorse model for time-varying arrival processes.

In fitting the specification to data, surprising pitfalls arise. To bring these issues to the attention of scholars interested in putting the specification to use, we use a real example to illustrate how intuitive estimation approaches can fail spectacularly. To provide researchers with a proper way to perform the estimation, we give a user friendly introduction to a statistical learning technique recently developed for queueing data, and explain intuitively how it addresses these pitfalls. Matlab code for the estimation is provided.

Version history: 20 Feb, 19 Apr 2018

Key words: arrival rate estimation; spectral estimation; nonhomogeneous Poisson process; emergency departments; call centres

1. Introduction

That real world arrival processes are consistent with a nonhomogeneous Poisson process (NHPP) was empirically verified in the works of Brown et al. (2005) and Kim and Whitt (2014a,b). A natural follow-on question is then whether a simple yet accurate functional form also exists for their arrival rates. Such a specification can potentially reconcile the aims of modelling with those

* Correspondence: dkklee@gmail.com

of simulation. In modelling, the goal is to obtain insights for the queuing system under study, hence analytic tractability takes precedence over exactness. In simulations, model resolution is paramount, so nonparametric approaches to estimating the intensity (e.g. piecewise constant fits) are used but are difficult to interpret (Massey et al. 1996, Green et al. 2006, Alizadeh et al. 2008, Zeltyn et al. 2011, Kim and Whitt 2014a, Shi et al. 2015).

To further characterize the functional form of the NHPP's arrival rate, in this paper we empirically examine whether a simple sum of sinusoids can be a satisfactory model for real world arrival rates. The model is motivated by the fact that customer arrival patterns typically exhibit cyclic, but not necessarily periodic, behaviour. For example, strong seasonal patterns have been observed in patient arrivals to emergency departments (Green et al. 2006, 2007, Zeltyn et al. 2011, Saghafian et al. 2012, Armony et al. 2015, Huang et al. 2015, Shi et al. 2015, Whitt and Zhang 2017). Indeed, numerous models of queuing systems already exploit periodicity to model time-varying arrivals (Green and Kolesar 1991, Eick et al. 1993, Jennings et al. 1996, Green et al. 2001, Feldman et al. 2008, Liu and Whitt 2012, Whitt 2014, Chan et al. 2016, Whitt 2016). Intuitively, sine waves are the simplest possible functional form for cyclic patterns, hence one might expect them to be the most tractable class of models. Therefore they hold promise as a workhorse model for time-varying arrival processes.

1.1. Sinusoidal arrival rates

The sinusoidal model we consider in this paper is introduced from recent statistical learning literature for queueing data, and it generalizes the existing sine wave models in queueing literature. The basic one proposed in Green and Kolesar (1991) posits a single sinusoid with period $1/\nu_1$ for the arrival rate function:

$$\lambda(t) = c_0 + c_1 \cos(2\pi\nu_1 t). \tag{1}$$

In practice, if the intensity is low enough at certain points to cause the estimate $\hat{\lambda}(t)$ to dip below zero, $\max\{0, \hat{\lambda}(t)\}$ can be used instead. In fact, doing so always improves accuracy since the true arrival rate must be non-negative.

A generalization of the basic model is given by Eick et al. (1993), which recognizes that any periodic arrival rate can be approximated arbitrarily well by its truncated Fourier series:

$$\lambda(t) = c_0 + \sum_{k=1}^{p} c_k \cos(2\pi k \nu_1 t + \phi_k). \tag{2}$$

Here, the k-th frequency $\nu_k = k\nu_1$ is an integer multiple of ν_1, and $\phi_k \in [0, 2\pi)$ are the phases. The model can also approximate any square-integrable function over a finite time interval[1]. Hence this functional form[2] has promise from a practical perspective, and moreover it is tractable enough to permit analysis. For example, Section 8 of Eick et al. (1993) demonstrates that the mean number of busy servers in a $M_t/G/\infty$ queue admits a closed form expression when the arrival rate takes on the form of (2). The results in Eick et al. (1993) can also be applied to obtain expressions for $M_t/M/s_t + M$ and $M_t/G/s_t + G$ queues whose arrival rates consist of a single sinusoid (Feldman et al. 2008, Liu and Whitt 2012, Liu 2018).

Yet another level of generalization exists in statistical literature that is essentially nonparametric (Shao 2010, Shao and Lii 2011, Chen et al. 2016), and it is the subject of our study. The model is obtained by allowing the frequencies $\{\nu_k\}_k$ to take on any values in a pre-specified band $[-B, +B]$:

$$\lambda(t) = c_0 + \sum_{k=1}^{p} c_k \cos(2\pi \nu_k t + \phi_k), \tag{3}$$

thereby resulting in an uncountably infinite degree of freedom. Unlike (2), the class above is not constrained to be periodic and p is free. Moreover, the model inherits the tractability of (2): The argument used to prove the closed form expressions in Section 8 of Eick et al. (1993) carries over to (3).

1.2. When might sine waves be preferred over piecewise smooth fits?

When fitting a piecewise smooth model to data, the usual practice is to average over, say, arrivals during Tuesdays 10-11am over successive weeks. This forces the fitted pattern to be periodic,

[1] For example if the true arrival rate contains a linear trend, its leading Fourier expansion terms can be used to model the trend over that interval. The associated coefficients can be refitted over time to extend the trend to future periods.

[2] A variant is the 'EPT' functional form (Kuhl et al. 1997, Lee et al. 1991) whereby the right hand side of (2) is exponentiated. However, the resulting function remains periodic and hence can be approximated by (2).

which can result in bias if non-weekly cycles are also present. For example, Figure 1 shows what a piecewise constant fit looks like when applied to a hypothetical arrival rate that has both a weekly as well as a 30-day cycle. In this case the fit is applied to the exact arrival rate (no statistical noise), and even then it is unable to capture the intra-monthly variations. The sinusoidal specification (3) does not suffer from this problem since it naturally adapts to the periodicities in the arrival pattern.

To illustrate the degree of improvement that the sinusoidal specification can achieve in practice, we fit both specifications to the arrivals data studied in Aldor-Noiman et al. (2009). The data came from the call centre of an Israeli telecom company, and the arrivals are aggregated into 30-minute interval buckets. The data were collected from mid-February to the end of December in 2004. It was found that the arrival pattern exhibits significant non-weekly cycles due to the lengths of the various billing cycles employed in the service contracts. The in-sample root mean squared error (RMSE) from using a 30-minute piecewise constant fit is 81 arrivals per hour$^{1/2}$. This uses $7 \times 24 \times 2 = 336$ piecewise constant intervals to cover the course of a week. By contrast, the in-sample RMSE from using a sinusoidal fit with just $336 \times 10\% = 33$ frequencies is 76 arrivals per hour$^{1/2}$, or a 6% improvement. These frequencies were selected from the most significant ones appearing in the discrete Fourier transform of the arrival counts, with 15 of them being non-weekly cycles.

Another desirable property of the sinusoidal specification is that it can be used to detect hidden periodicities that could yield unexpected insights. For example, Prof. E.H. Kaplan studied patient arrivals to psychiatric wards and verified the existence of a lunar (28 day) cycle, which suggests a link between the phases of the moon and mental health (personal communication). Such insights are not readily available from piecewise constant fits.

1.3. Aims of present work

The first goal of this paper is to empirically test the specification (3) in two settings of interest to operations scholars. In Section 5 we examine arrivals data from an academic Emergency Department (ED) in the United States, and in Section 6 we examine data from a bank call centre in

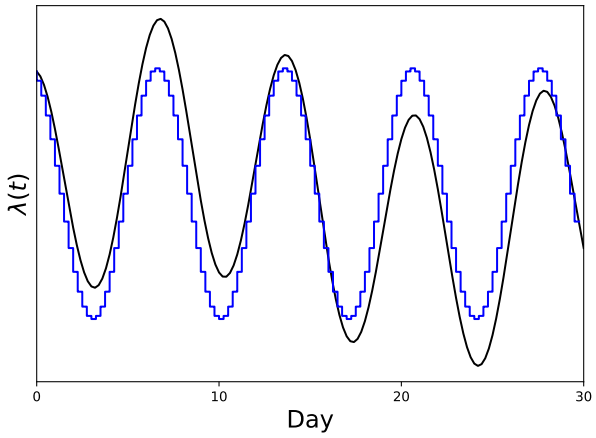

Figure 1 Piecewise constant fit to a hypothetical arrival rate with both a weekly and a 30-day cycle.

the United States that is kindly made available by the SEELab at Technion (Mandelbaum 2017). Using the statistical tests we develop in Section 4, we show that both settings are consistent with a NHPP with a rate of the form (3). Our findings provide researchers with empirical support for the use of a sinusoidal rate specification in practice and in modelling (for example, Green and Kolesar 1991 and Eick et al. 1993). However, echoing Kim and Whitt (2014a), we refrain from drawing blanket conclusions and advise that each setting be tested using the approach developed here.

The second goal of this work is to introduce into the operations literature statistical learning ideas recently developed for queueing data (Shao 2010, Shao and Lii 2011, Chen et al. 2016) for fitting the specification (3). The arrivals data we are interested in consists of customer arrival timestamps $\{t_j\}_{j=1}^N$ in the observation window $[0,T]$. In Section 3 we provide a user friendly description of the procedure in Chen et al. (2016) that researchers can use to estimate and implement queuing models for prescriptive purposes. The method extends the classic periodogram for point processes used in Shao and Lii (2011), and a Matlab implementation can be found at `github.com/chenny888/arrival_sinepoisson`.

Our third goal is to highlight the surprising pitfalls that can arise[3] if (3) is not estimated with care. These pitfalls are not well known outside of specialized spectral analysis literature, thus it is

[3] These estimation issues are not present in the special case of the gridded frequency model (2) where the locations of the frequencies are assumed to be known apriori. Estimation under this setting is studied in Lee et al. (1991) and Kuhl et al. (1997).

important to bring them to the attention of researchers interested in modelling customer arrival patterns. To illustrate the need for a carefully designed estimation method, in Section 2 we present a real dataset on which seemingly reasonable estimation approaches fail spectacularly. Although existing works have shown that these issues are theoretically possible, to our knowledge this is the first real arrivals example to show that these issues do matter in practice. In Section 3 we provide an intuitive explanation for how the estimation procedure in Chen et al. (2016) addresses these pitfalls.

2. Estimation challenges

We present two counterexamples to illustrate surprising potential pitfalls one encounters when attempting to fit the sinusoidal specification (3) to data. The first example shows that the precision of the frequency estimates needs to be extremely small for the amplitude/phase estimators to be consistent. The second example reveals a fundamental limit on the frequency resolution that any method can attain, and is a manifestation of Heisenberg's uncertainty principle. These counterexamples illustrate the need for a carefully designed estimation procedure, and one will be presented in Section 3.

2.1. Precision of frequency estimates

We consider a subset of 66,240 patient arrivals to the ED of an academic hospital in the United States from 2014 to Q3 of 2015 ($T = 652$ days of observations). The subset under consideration corresponds to patients assigned Emergency Severity Index (ESI) level 2, a measure of emergency severity that will be described in more detail in Section 3. The original motivation for analyzing this data stems from a capacity planning question posed by the ED management: To reduce patient wait times during the day, management considered adding an auxiliary ward to treat non-trauma patients like those in ESI level 2. The ward would only be open for part of the day when demand for services is high, and they wished to determine the appropriate capacity (number of beds) to build into it. Unlike staffing decisions which can be adjusted dynamically, this is a one time decision.

To identify the optimal capacity level via a queuing simulation, it is necessary to first estimate the arrival rate of patients. Suppose we wish to fit the sinusoidal specification (3) to data. The immediate approach that comes to mind for estimating the periodicities in the arrival pattern is to employ discrete Fourier transform, which requires bucketing the arrivals into time bins and then analyzing the bin counts. The logic behind this is that the mean count in the bucket $(t, t+\Delta]$ is

$$\begin{aligned}\int_t^{t+\Delta} \lambda(t)dt &= c_0\Delta + \sum_{k=1}^{p} c'_k \cos(2\pi\nu_k t + \phi_k - \frac{\pi}{2}) + c''_k \cos(2\pi\nu_k t + \phi_k) \\ &= c_0\Delta + \sum_{k=1}^{p} c'''_k \cos(2\pi\nu_k t + \phi'_k),\end{aligned}$$

which has the same periodocities as $\lambda(t)$. The left panel of Figure 2 displays the power spectrum of the centred arrival counts bucketed by three hour intervals. The horizontal axis displays a range of frequencies present in the count data, and the height of a spike is the squared amplitude of the cosine wave with the corresponding frequency. The plot reveals the presence of eight dominant frequencies in the data, the largest one being 1.0015 cycles/day.

To fit the arrival rate to (3) using just these eight cosines, we re-estimate their amplitudes c_k and phases ϕ_k using least squares. The dashed line in the right panel of Figure 2 shows the resulting fit over the course of a Wednesday. Compared to the solid line representing the average hourly arrival count, the estimated arrival rate is essentially flat and fails to capture the intraday variation in arrivals. Simulations based on this would effectively recommend that no auxilary capacity be added during the busy periods of the day, thus leading to substantial overcrowding during the day. It is rather unsettling that using the discrete Fourier transform to estimate the dominant frequencies would lead to such discrepancies.

In this example, one might guess that the dominant frequency 1.0015 cycles/day is really just a noisy estimate of the daily cycle. Suppose we refit (3) to the seven other frequencies in addition to the daily cycle, that is we replace 1.0015 cycles/day with 1 and keep all other frequencies the same. The dotted line in Figure 2 shows the resulting fit, which is a significant improvement over the previous one. This suggests that the frequency estimation error of 0.0015, which is $1/T = 1/652$ in this case, is too large. In fact, replacing $1.0015 = 1 + 1/T$ with a less noisy estimate of the daily

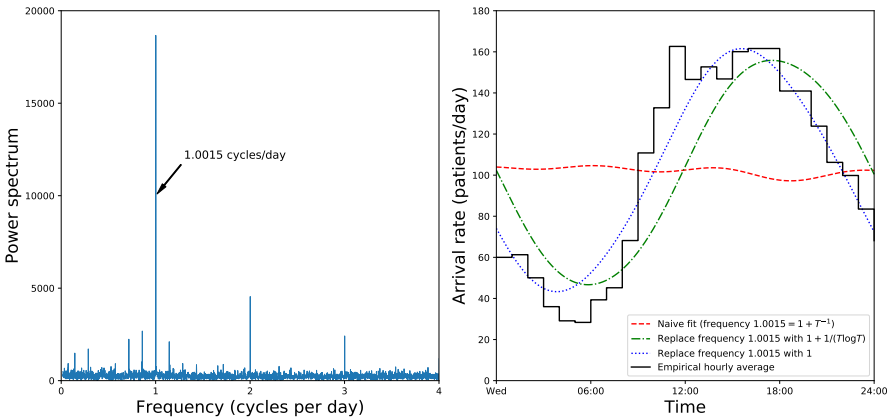

Figure 2 *Left panel*: Power spectrum of ED arrivals data for 66,240 ESI level 2 patients, centred to remove the zero frequency component (constant intercept). Eight dominant frequencies stand out, the largest one being 1.0015 cycles/day. *Right panel*: Estimated arrival rates over the course of a Wednesday. The solid line represents the average arrival count over each hour of Wednesday; the dashed line represents the sinusoidal rate (3) fitted to the eight dominant frequencies identified in the left panel; the dash-dot line represents the sinusoidal fit obtained from replacing 1.0015 with $1+1/(T\log T)$; the dotted line represents the sinusoidal fit obtained from replacing 1.0015 with 1.

cycle, say $1+1/(T\log T)$, already provides a more sensible fit (dash-dot line in the figure) that is closer to the accuracy of the dotted line. This suggests that the estimation error has to be smaller than order $1/T$ for consistent estimation of the rate function. Thus seemingly reasonable methods for fitting the specification (3) may not actually be precise enough for frequency estimation.

Why? The amplitude estimation involves regressing the arrival counts onto the explanatory variables $\{\cos(2\pi\hat{\nu}_k t)\}_{k=1}^{8}$ where $\hat{\nu}_1,\cdots,\hat{\nu}_8$ are the estimated frequencies. Since in this case $\hat{\nu}_1 = 1.0015 = 1+T^{-1}$ is likely a noisy estimate of $\nu_1 = 1$, using $\cos(2\pi\hat{\nu}_1 t)$ as an explanatory variable in lieu of $\cos(2\pi\nu_1 t)$ would result in model misspecification. For example at time $t = T/2$, the bias

$$\begin{aligned}\cos(2\pi\hat{\nu}_1 t) - \cos(2\pi\nu_1 t) &= \cos\left\{2\pi\left(1+\frac{1}{T}\right)\frac{T}{2}\right\} - \cos\pi T \\ &= -2\cos\pi T\end{aligned}$$

does not vanish as $T\to\infty$, hence no estimation approach (least squares or otherwise) will be able to consistently recover the amplitude. On the other hand if the precision of the frequency estimate is sharper than order $1/T$, e.g. $\hat{\nu}_1$ is $1+o(1)/T$ instead, then the misspecification bias is asymptotically negligible:

$$\cos(2\pi\hat{\nu}_1 t) - \cos(2\pi\nu_1 t) = \cos\left\{2\pi\left(1+\frac{o(1)}{T}\right)\frac{T}{2}\right\} - \cos\pi T$$
$$= \cos\{\pi T + o(1)\} - \cos\pi T \to 0.$$

In the literature for ordinary time series, Rice and Rosenblatt (1988) formally shows that consistent amplitude estimation requires the frequency estimation error $|\nu - \hat{\nu}|$ to be less than order $1/T$. The analogue for arrival processes is provided in Shao and Lii (2011) and Chen et al. (2016). Although the required level of precision is known theoretically, this is the first real arrivals example to show that this issue matters in practice.

In this particular case one might argue that 1.0015 cycles/day is clearly a noisy estimate for the daily cycle, and so no theory is needed for rounding 1.0015 to 1. However, in general, the 'obvious true' frequency may be wrong if the frequency estimate is, say, 1.93 cycles/day: As recounted in Parker (2011), the Allies' planning of the D-Day invasion of the Normandy beaches hinged on understanding the periodicities in tidal heights, which represent the arrival rate of fluid parcels. In a prescient act of patriotism, Laplace demonstrated mathematically in 1776 that the dominant frequencies are 1.93 cycles/day and 2.00 cycles/day. Without Laplace's genius, the Allies would have had to determine these empirically instead, and mistakingly rounding 1.93 to 2 might well have changed the course of history.

2.2. Frequency resolution

Inspired by Prof. E.H. Kaplan's study of patient arrivals to psychiatric wards, consider the following arrival rate model with two frequencies, a dominant monthly cycle and a weaker lunar one:

$$\lambda(t) = 8\cos\left(2\pi\frac{t}{30}\right) - \cos\left(2\pi\frac{t-2}{28}\right). \tag{4}$$

Suppose we are able to continuously measure $\lambda(t)$ up to time $T = 180$ days perfectly without noise. One might intuitively expect the power spectrum plot of the noiseless data to consist of two spikes, a large one at 1/30 (monthly cycle) and a smaller one at 1/28 (lunar cycle): The left panel of Figure 3 illustrates this idealized scenario for the square root of the power spectrum. In reality, the spectrum of the data is the one in the right panel. Instead of two frequency components, a

continuum of them are present in the data. While the peak corresponding to the monthly cycle is still visible, the lunar one is completely masked by the distortion. This distortion makes it impossible to detect frequencies that are close to the dominant monthly cycle from the plot.

Why? Although there is no statistical noise in the data, there is 'leakage noise' which is due to the fact that we only observe $\lambda(t)$ up to finite time T: Intuitively, to tease apart two similar frequencies, a large time window is needed for their cycles to desynchronize. Hence, for a finite T, we cannot expect to be able to resolve frequencies that are too close together. The fundamental limit on the frequency resolution that any method can attain is governed by the mathematical manifestation of Heisenberg's uncertainty principle. To explain this in an intuitive way, we use ideas from Section 2 of Chen et al. (2016): Suppose a function $f(t)$ can be written as a sum of complex exponentials $\sum_\nu c_\nu e^{2\pi i\nu t}$, potentially over a continuum of frequencies. Informally, the coefficient c_ν is given by the Fourier transform of $f(t)$:

$$\tilde{f}(\nu) = \int f(t) e^{-2\pi i\nu t} dt.$$

In the case of (4) we have

$$\tilde{\lambda}(\nu) = \begin{cases} c_{1/30} & \nu = 1/30 \\ c_{1/28} & \nu = 1/28 \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

representing the superposition of two spikes, one at the frequency of the 30-day cycle, and the other at the frequency of the 28-day cycle (left panel of Figure 3). Letting $I_{(0,T]}(t)$ be the indicator function for $t \in (0, T]$, our data is $\lambda(t) I_{(0,T]}(t)$, the result of truncating $\lambda(t)$ due to T being finite. It follows from the duality between multiplication and convolution that the frequency content in the data is

$$\widetilde{(\lambda \cdot I_{(0,T]})}(\nu) = c_{1/30} \tilde{I}_{(0,T]}\left(\nu - \frac{1}{30}\right) + c_{1/28} \tilde{I}_{(0,T]}\left(\nu - \frac{1}{28}\right). \tag{6}$$

The modulus of $\tilde{I}_{(0,T]}(\nu)$ is $|\sin(\pi T\nu)|/(\pi\nu)$, which resembles the shape of the right panel in Figure 3 but with the main lobe centred at $\nu = 0$ instead. Thus (6) replaces each localized spike in the left panel of Figure 3 with a multiple of $\tilde{I}_{(0,T]}$ centred at the same frequency location, and the side lobes of $\tilde{I}_{(0,T]}$ represent leakage of the spectral energy from a spike to neighbouring frequencies. In other

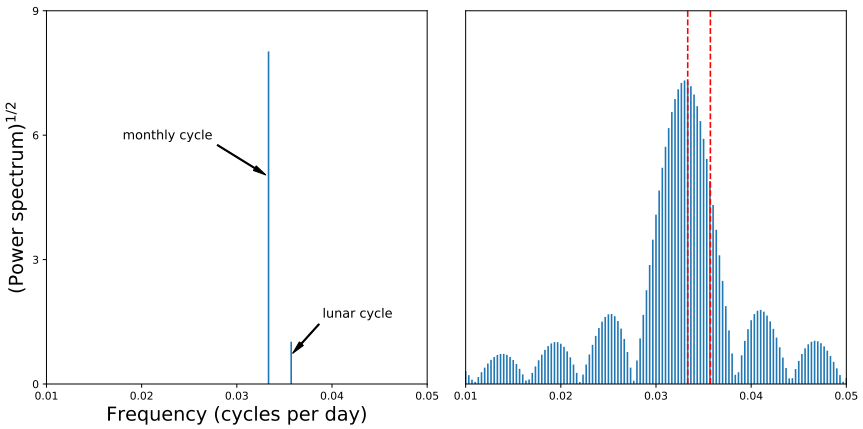

Figure 3 *Left panel*: Idealized (square-root) power spectrum for (4) over $T = 180$ days. *Right panel*: Actual square-root spectrum. A continuum of spurious frequency components are present. The lunar cycle is completely masked by the distortion.

words, truncating $\lambda(t)$ to $\lambda(t)I_{(0,T]}(t)$ results in smearing the frequency spikes into a continuum: For values of ν not $1/30$ or $1/28$, $\left|\widetilde{(\lambda \cdot I_{(0,T]})}(\nu)\right|$ can now be positive, creating an artificial leakage noise floor. As seen in the right panel of Figure 3 the leakage around the stronger monthly cycle completely masks the weaker lunar one nearby. We see from (6) that perfect frequency recovery is only possible if the support of $\tilde{I}_{(0,T]}(\nu)$ does not extend beyond $1/28 - 1/30$ from the origin, so that the two terms have no overlap. However an infinite length of observation ($T = \infty$) would then be needed because a non-zero function and its Fourier transform cannot both have finite support. This follows from a version of the uncertainty principle due to Amrein and Berthier (1977) and Benedicks (1985).

3. Estimation

Related to the takeaway from Section 2.2, the limit to how close the signal frequencies $\{\nu_k\}_k$ in (3) can be to one another is quantified in Moitra (2015), which shows that no estimator can generally distinguish between frequencies that are less than $1/T$ apart in noisy time series data. In this paper we will operate under a slightly coarser resolution[4] where the frequency gap is

[4] It is theoretically possible to also deal with frequency gaps of order $1/T$, which is known as the *super-resolution* regime. The corresponding exposition becomes more involved however, and the interested reader is referred to Chen et al. (2016) for details.

$$\min_{k \neq k'} |\nu_k - \nu_{k'}| \geq \frac{g(T)}{T} \tag{7}$$

for any function $g(T) \geq 4$ for which $g(T) \to \infty$, e.g. $g(T) = \mathcal{O}(\log T)$.

We follow the graphical method proposed in Chen et al. (2016) to analyze the arrivals data. The method extends the classic point process periodogram (Bartlett 1963, Vere-Jones 1982, Shao and Lii 2011), is conceptually simple, and does not require bucketing arrivals into discrete time bins like the naive approach in Section 2.1. Before describing the method in Section 3.1, it will be instructive to briefly overview how the approach handles the challenges outlined in the previous section. The method invokes a condition that informally requires the ratio of the largest to the smallest amplitude $\max_k |c_k| / \min_k |c_k|$ to be no greater than 14.5, and is referred to as the dynamic range of the amplitudes.[5] The intuition for requiring this can be seen from the example in Section 2.2, where the amplitude of the lunar cycle is so small relative to the monthly one that it is washed out by the leakage from the larger monthly cycle.

Under the prescribed setting, with high probability the method will recover all p frequencies with precision

$$\max_{k=1,\cdots,p} |\nu_k - \hat{\nu}_k| < \min\left\{\frac{2}{T}, C \max\left(\frac{1}{Tg(T)^3}, \frac{\sqrt{\log T}}{T^{3/2}}\right)\right\},$$

for a constant C that is specified in Chen et al. (2016). Since $g(T) \to \infty$, the frequency estimates $\{\hat{\nu}_k\}_k$ have error less than order $1/T$. Then, as alluded to in Section 2.1, the resulting amplitude and phase estimates can be shown to be consistent with error up to a multiple of

$$\max\left(\frac{1}{g(T)^3}, \frac{\sqrt{\log T}}{T^{1/2}}\right),$$

so the rate function $\lambda(t)$ can be consistently estimated by directly plugging all the estimates into (3).

[5] The lower bound of 4 for $g(T)$ and the dynamic range 14.5 can both be adjusted by adopting variations of the method described in this paper. The interested reader is referred to Chen et al. (2016) for details.

3.1. Estimation of rate function (3)

In essence, the approach in Chen et al. (2016) is to examine the power spectrum/periodogram of the *weighted* arrivals data and remove frequencies whose height is below a certain threshold τ. The rationale is that such frequencies are most likely noise and not part of the signal frequency set $\{\nu_k\}_{k=1}^p$ in (3). Thresholding therefore leaves behind a band of frequencies around each ν_k, so that the frequency corresponding to the largest amplitude in each neighbourhood provides an estimate for one of the ν_k's. Figure 4 is reproduced from Chen et al. (2016) and illustrates the idea for $p=2$.

Since the uncertainty principle prevents us from resolving frequencies that are too close together, after selecting a frequency estimate we remove all neighbouring ones within radius r before seeking the next highest peak. The height of the *centralized* and *windowed* periodogram is given by

$$|H_c(\nu)| = \frac{1}{T}\left|\sum_j w(t_j)e^{-2\pi i\nu t_j} - N\tilde{w}(\nu)/T\right|, \tag{8}$$

where $i=\sqrt{-1}$, N is the total number of arrivals in the time window $[0,T]$, and $w(t)$ is the Hann window function supported on $(0,T]$:

$$w(t) = \left(\sin^2\frac{\pi t}{T}\right) I_{[0,T]}(t) \leftrightarrow \tilde{w}(\nu) = \begin{cases} T/2 & \nu = 0 \\ -T/4 & \nu = \pm\frac{1}{T} \\ \frac{T}{2}e^{-i\pi T\nu}\frac{\sin(\pi T\nu)}{(\pi T\nu)\{1-(T\nu)^2\}} & \text{else} \end{cases}. \tag{9}$$

By design the peak at $\nu=0$ is removed from the centralized periodogram since this is the frequency corresponding to the constant c_0 in (3). The presence of the window $w(t)$ in $|H_c(\nu)|$ effectively assigns $w(t_j)$ units of customers to the j-th arrival, and does not have to be a whole number.

From a queuing perspective it may appear counterintuitive to bias the data by reweighting. However, from the Fourier perspective of Section 2.2 we see that the leakage from the signal frequencies in (6) is more muted when the tails of $\tilde{I}_{(0,T]}$ decay quickly. Thus replacing the data $\lambda(t)I_{(0,T]}(t)$ with $\lambda(t)w(t)$ improves the precision of the frequency estimation because the tails of $\tilde{w}$ are by design much lighter than those of $\tilde{I}_{(0,T]}$. This is in fact a key insight of Chen et al. (2016) that extends the classic point process periodogram. Algorithm 1 below in Section 4 describes a particular implementation of the one in Chen et al. (2016) that is used for the analyses in this paper.

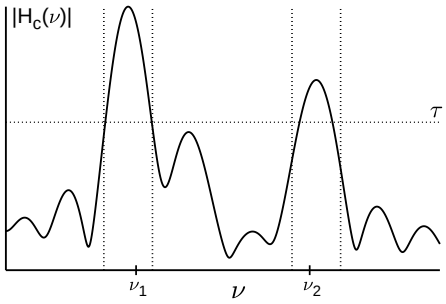

Figure 4 Illustration of the estimation procedure in Algorithm 1 of Section 4. In the depicted periodogram there are two signal frequencies ν_1 and ν_2. Setting τ (horizontal line) above the ambient noise level leaves behind two bands (between the pairs of vertical lines) that contain ν_1 and ν_2.

4. Statistical tests for NHPP with sinusoidal rates

Having described a method from the statistical literature for obtaining sinusoidal fits in Section 3, we now develop a statistical test for whether an arrivals dataset is generated by a NHPP with a sinusoidal rate. In the spirit of Brown et al. (2005) and Kim and Whitt (2014a), we assess model fit to each time subinterval of the arrivals window $[0, T]$. Whereas the previous works test whether the arrivals in each subinterval follow a NHPP, we go one step further in pinning down the form of the arrival rate function. Since no model is perfect, we do not expect the parsimonious sinusoidal specification (3) to fit well to all subintervals. Therefore, if it passes a goodness-of-fit test for a large majority of subintervals then we deem (3) to be a reasonable model. We first propose a test for a single time interval, and then describe multiple testing corrections for applying the test to all subintervals.

4.1. Single subinterval

If the arrivals $t_1, \cdots, t_N$ in an interval $[t_0, t_0 + L)$ are generated from a NHPP with the sinusoidal rate (3), the interarrival times $\tau_j = t_j - t_{j-1}$ $(j \geq 1)$ are distributed as

$$\mathbb{P}(\tau_j > \tau) = \exp\left(-\int_{t_{j-1}}^{t_{j-1}+\tau} \lambda(t)dt\right) = \exp\{\Lambda(t_{j-1}) - \Lambda(t_{j-1} + \tau)\}. \tag{11}$$

Algorithm 1 An implementation of the estimation procedure in Chen et al. (2016)

1 Noting that $|H_c(\nu)|$ is symmetric in ν, compute its value over a range $[-B,+B]$ of frequencies of interest.

2 Set $r=2/T$ and the threshold according to (3.8) of Chen et al. 2016:

$$\tau = 0.0574 \sup_{\nu\in[0,B]} |H_c(\nu)| + 1.06 \min\left(\hat{\chi}_T, 4\frac{\sqrt{N\log T}}{T}\right), \tag{10}$$

where $\hat{\chi}_T$ is $\sup_{\nu\in[0,B]}|H_c(\nu)|$ for simulated arrivals from a homogeneous Poisson process with rate N/T over $[0,T]$. That is, simulate the arrival times $\hat{t}_1,\hat{t}_2,\cdots$ from the process and compute $\sup_{\nu\in[0,B]}|H_c(\nu)|$ for it.

3 Identify the frequency region $R=\{\nu : r\le|\nu|\le B, |H(\nu)|>\tau\}$ where the value of periodogram exceeds τ.

4 Set $\nu_0=0$, $k=1$ and repeat the following steps:

- Find the highest stationary peak of the periodogram in R and set $\hat{\nu}_k$ as the corresponding frequency location. If no peaks exist then exit loop.
- Perform the updates $k\leftarrow k+1$ and $R\leftarrow R\backslash(\hat{\nu}_k-r,\hat{\nu}_k+r)$. This removes a neighbourhood of radius r centred at $\hat{\nu}_k$ from R.

5 The estimate $\hat{c}=\{\hat{c}_k\}_{k=0}^p$ for the coefficients is given by $\hat{\Gamma}^{-1}y$, where the (j,k)-entry of the $(p+1)\times(p+1)$ matrix $\hat{\Gamma}$ is

$$\hat{\Gamma}_{jk} = e^{-i\pi T(\hat{\nu}_j-\hat{\nu}_k)}\frac{\sin\{\pi T(\hat{\nu}_j-\hat{\nu}_k)\}}{\pi T(\hat{\nu}_j-\hat{\nu}_k)},$$

and the k-th entry of the $(p+1)$-vector y is $\frac{1}{T}\sum_j e^{-2\pi i\hat{\nu}_k t_j}$.

Since a NHPP has independent increments, $\{u_j=\exp[\Lambda(t_{j-1})-\Lambda(t_j)]\}_j$ are independent and identically distributed uniformly on $[0,1]$. The Kolmogorov-Smirnov statistic

$$\sup_{0\le t\le 1}\left|\frac{1}{N}\sum_{k=1}^{N} I(u_j\le t)-t\right| \tag{12}$$

then provides a way for testing the null hypothesis that the arrival process in $[t_0,t_0+L)$ is a NHPP with rate $\lambda(t)$.

On the other hand if there are no arrivals in $[t_0, t_0 + L)$, then the probability that a Poisson distribution with mean $\Lambda(t_0 + L) - \Lambda(t_0)$ is equal to zero,

$$\exp[-\{\Lambda(t_0 + L) - \Lambda(t_0)\}], \tag{13}$$

serves as the p-value for the null hypothesis.

In practice, substituting the estimated $\hat{\Lambda}(t)$ for $\Lambda(t)$ means that the null hypothesis is that the arrival process is a NHPP with the deterministic rate $\hat{\lambda}(t)$, which is a more stringent test than whether the rate function follows any sinusoidal pattern at all. Our null hypothesis is also stronger than the one in Kim and Whitt (2014a) which does not specify a rate function, but only that the arrival process is NHPP. Furthermore, their null hypothesis does not distinguish between deterministic or stochastic rate functions (Cox processes).

As discussed in Brown et al. (2005) and Kim and Whitt (2014a), arrival times which have been rounded (e.g. to the nearest second) will require unrounding before tests on the interarrival times can be performed. This removes interarrival times of zero if multiple arrivals are rounded down to the same second. We follow the suggestion to add uniform random numbers between $[0, 1]$ seconds to the timestamps to mitigate the problem.

4.2. All subintervals

If each subinterval is of length L, then applying the test above to each subinterval will yield T/L hypothesis tests. A multiple testing correction is then needed to control the number of false rejections at the 5% level. For a single test (i.e. $T/L = 1$), this corresponds to controlling the type I error, and a generalization of the type I error to multiple tests is the familywise error rate

$$FWE = \mathbb{P}(\#\text{false positives} > 0). \tag{14}$$

To see why it is not enough to just control the type I error of each individual test (i.e. reject a test whenever its p-value is less than 0.05), suppose for simplicity that we have 10 independent tests, and that all null hypotheses are true. Thus, the probability that each test results in a false

rejection is 0.05, but the probability that there is at least one false rejection among the tests is $FWE = 1 - 0.95^{10} > 0.4$.

We will use the procedure of Holm (1979) to ensure that $FWE \leq 0.05$. The procedure is uniformly more powerful than the better known Bonferroni correction, and works in the following way: Let $P_{(1)}, \cdots, P_{(T/L)}$ be the ordered p-values from the smallest to the largest, and compute the smallest k such that

$$P_{(k)} > \frac{0.05}{T/L + 1 - k}. \tag{15}$$

The null hypotheses associated with $P_{(1)}, \cdots, P_{(k-1)}$ are then rejected, while the ones associated with the larger p-values are not.

Note that the FWE is the probability that *any* of rejected nulls is a false rejection. An alternative is to control the false discovery rate

$$FDR = \mathbb{E}\left(\frac{\#\text{false rejections}}{\#\text{rejections}}\right) \tag{16}$$

instead, which is the expected proportion of false rejections among all the rejected nulls. This criterion is less stringent and provides more statistical power because $FDR \leq FWE$, hence any test that controls FWE automatically controls FDR as well. As a separate check we will use the procedure given by Theorem 1.3 of Benjamini and Yekutieli (2001) to control the FDR at the 5% level. The procedure seeks the largest k such that

$$P_{(k)} \leq \frac{0.05k}{(T/L)\sum_{j=1}^{T/L}(1/j)}, \tag{17}$$

and rejects the null hypotheses associated with $P_{(1)}, \cdots, P_{(k)}$.

5. Patient arrivals to an ED

We apply Algorithm 1 to analyze arrivals data from the emergency department (ED) of an academic hospital in the United States. The dataset contains timestamps for 168,392 patient arrivals from 2014 to Q3 of 2015 ($T = 652$ days). In addition, the Emergency Severity Index (ESI) of each patient is also recorded, with level 1 being the most severe (e.g. cardiac arrest) and level 5 the least (e.g.

rash). Information on demand for ED services is a critical input to staffing and other operational decisions that influence the efficiency of health care delivery, and it enables the construction of high resolution queuing simulations. For the ED in question, ESI level 1 trauma patients ($< 1\%$ of arrivals) are assigned to dedicated bedspaces on arrival and are therefore excluded from the analysis. Since level 2 patients are treated in a ward separate from the other ESI levels, from a capacity management perspective we will analyze the two groups separately.

5.1. ESI level 2

In our dataset, 66,240 patient arrivals were assigned an ESI level of 2. A preliminary analysis of the ESI level 2 data was given in Chen et al. (2016) using a variant of Algorithm 1, but no goodness-of-fit tests were performed. For completeness, we will re-analyze the data using Algorithm 1 and perform the tests in Section 4.

As shown in the left panel of Figure 5, Algorithm 1 selected three intraday frequencies and three week-based ones. The intraday frequencies include a daily cycle ($\hat{\nu}_1 = 1.00$), a 12 hour cycle ($\hat{\nu}_2 = 2.00$), and an 8 hour cycle ($\hat{\nu}_3 = 3.00$). The week-based ones include a 1/5 week cycle ($\hat{\nu}_4 = 0.714$), a 1/6 week cycle ($\hat{\nu}_5 = 0.857$), and a 1/8 week cycle ($\hat{\nu}_6 = 1.143$).

Given that the fitted rate has a weekly period, we can compare it to the empirical average arrival rate for each hour of the week (right panel of Figure 5). The estimate reveals two intraday peaks, the first at around 11am and the second at around 5pm. We also see that the intensity of arrivals fade steadily into the weekend. For most parts of the week, the sinusoidal fit does a very good job in capturing the variation in the empirical arrival rate, at least from a visual viewpoint.

To use the statistical tests described in Section 4 to quantify the goodness of fit, we divide the arrivals data into time subintervals of length $L = 2$ hours[6]. Of the resulting $T/L = 7,824$ subintervals, the null hypotheses for only 7% of them are rejected, and this is before applying any multiple testing corrections (which will reject even fewer hypotheses). Hence the sinusoidal specification (3) is a reasonable model for ESI level 2 arrivals.

[6] Given that the highest frequency detected in the data is an 8 hour cycle, Nyquist-Shannon sampling considerations suggest using a bin width of not more than 4 hours.

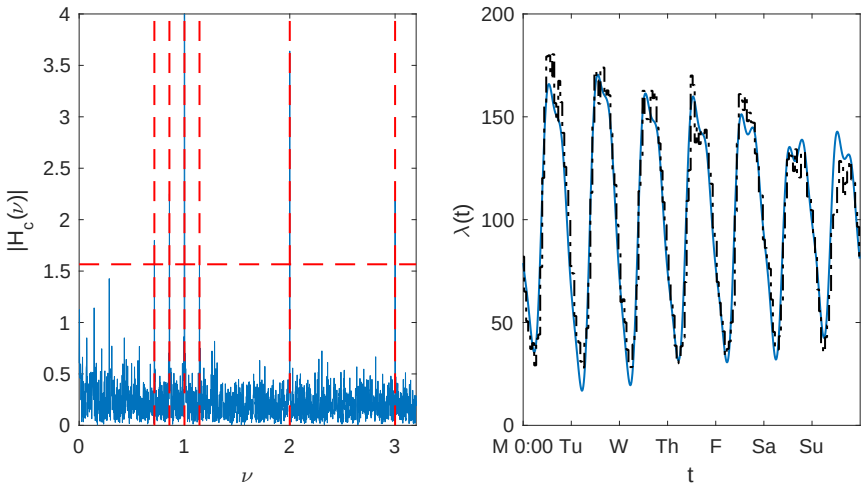

Figure 5 ESI level 2 arrivals. *Left panel:* The centralized windowed periodogram (8). The selected threshold (10) is represented by the dashed horizontal line, and the location of the frequency estimates $\hat{\nu}_k$ are given by the vertical ones. *Right panel:* The estimated arrival rate (arrivals per day) over the course of a week is given by the solid line. The dash-dot line represents the empirical average arrival rate for each hour of the week.

5.2. ESI levels 3 to 5

There were in aggregate 99,205 arrivals assigned to ESI levels 3 to 5. Unlike the ESI level 2 data, this dataset has not been analyzed before. Figure 6 shows that the same intraday frequencies from the level 2 case are selected, but only the 1/6 week cycle is selected from the week-based ones. The estimated arrival rate exhibits a more subdued day-of-week effect when compared to level 2, but the midday peak is now more pronounced relative to the evening one. On the other hand, the empirical average arrival rate exhibits a substantial spike in arrivals on Monday. We know from Fourier theory that capturing these time-localized effects will require more sine waves than the sparse specification (3) is designed for[7], which explains the sinusoidal model's underfit to Mondays. For the other days of the week, (3) provides a good visual fit to the average rate.

Similar to ESI level 2 arrivals, the null hypotheses for only 8% of the 7,824 subintervals are rejected. Therefore the sinusoidal specification (3) is also a reasonable model for ESI levels 3-5

[7] The decay of the Fourier series coefficients of a function is slow when the function is not smooth, for example when time-localized effects are present. Thus such functions require more terms of its Fourier expansion to approximate it. A possible future refinement to the sinusoidal specification could be to augment the sine waves with time-localized basis functions like wavelets.

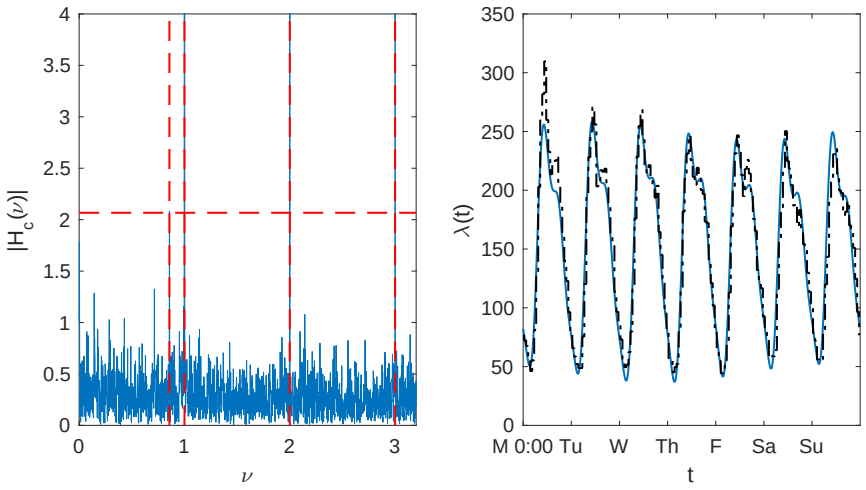

Figure 6 ESI levels 3 to 5. *Left panel:* The centralized windowed periodogram (8), selected threshold (10), and the locations of the frequency estimates $\hat{\nu}_k$. *Right panel:* The estimated (solid line) and empirical (dash-dot line) arrival rates over the course of a week.

arrivals.

5.3. Seasonal effects

No year-based seasonal cycles (365/12 (monthly), 365/4 (quarterly), 365/2, 365) are selected from the data. Given our frequency resolution of $4/T$, it is possible that we missed the semi-annual and annual cycles whose frequencies are within 4/652 of zero (frequency of the constant c_0 in (3)). This can be remedied by adding them to the set of selected frequencies, and then estimating their coefficients using step 4 of Algorithm 1. For both ESI groups we find that the semi-annual cycle is negligible, and that the annual cycle is weaker than the weakest frequency signal selected by our procedure. For example, for ESI levels 3 to 5, the amplitudes of the selected frequencies range from 4.0 to 44. The amplitudes of the semi-annual and annual cycles are 0.82 and 3.3 respectively.

6. Bank call centre arrivals data

We now examine a dataset from the call centre of a bank in the United States. The dataset is kindly made available by the SEELab at Technion (Mandelbaum 2017). The call centre consists of sites across the Northeast that are integrated into one virtual centre. About a fifth of the callers seek to speak with a live agent, and the rest conduct self-service transactions at the Voice Response Unit

#Cycles per day	#Cycles per week	Other cycles
1,2,3,4	1,2,3,4,5,6,8,9,15	1 cycle per 15.2 days 1 cycle per 31.1 days

Table 1 Frequencies selected for the arrival rate of VRU-Premier calls

(VRU), Announcement, or Message stages. Section 5.4 of Kim and Whitt (2014a) study whether the call arrivals between 7am-10pm in the 30 days of April 2001 can be modelled by NHPPs. Among the eight call types in the data, six are found to pass the NHPP test in more than 75% of days. These call types are VRU-Premier, VRU-Business, VRU-Loan, VRU-Summit, Business, and Message.

To further test whether the arrival rates of these call types can be modelled by the sinusoidal specification (3), we use calls arriving between April 2001 and March 2002 inclusive to fit the arrival rate. Arrivals of the type VRU-Summit appeared in only 140 days of record during the period (up until August 2001), therefore we only consider the other five types of calls.

Table 1 shows the 15 frequencies selected by Algorithm 1 for the arrival rate of VRU-Premier calls. In addition to the intraday and the week-based cycles, frequencies with periods close to 15 days and 31 days are also selected. Since these are not divisible by 7, the arrival rate for a specific hour of a particular day of week will vary from week to week. This explains a portion of the overdispersion discovered in Kim and Whitt 2014a.

The first column of Table 2 displays the number of calls of each type from April 2001 to March 2002 inclusive. We see that VRU-Loan has the lightest call volume, but on a per unit time basis, even this has roughly four times as many arrivals as ESI levels 3-5 in Section 5.2. This allows us to use a finer time subinterval for the tests in Section 4. The last two columns of Table 2 show the results of applying the tests to the April 2001 period studied in Kim and Whitt (2014a). Whether we correct for multiple testing issues using FWE or FDR, we see that the null hypotheses for most of the 30 minute subintervals are not rejected. Applying the test out-of-sample to call arrivals in April 2002 yields similar results (Table 3). Taken together, our results extend the finding in Kim and Whitt (2014a) by further showing that the sinusoidal specification (3) is a reasonable model for the arrival rates of the NHPP call types.

	#Calls btwn Apr '01 and Mar '02	#Calls in Apr '01	#Subintervals not rejected	
			$FDR \leq 0.05$	$FWE \leq 0.05$
VRU-Premier	1,071,481	96,187	845 (94%)	872 (97%)
VRU-Business	2,571,511	218,163	731 (81%)	807 (90%)
VRU-Loan	208,328	13,782	885 (98%)	885 (98%)
Business	987,606	71,010	745 (83%)	810 (90%)
Message	977,579	78,914	628 (70%)	721 (80%)

Table 2 In-sample results from applying the tests in Section 4 to calls to the bank in April 2001. 30 minute subintervals are used to group calls between 7am-10pm, resulting in a total of 900 subintervals.

	#Calls in Apr '02	#Subintervals not rejected	
		$FDR \leq 0.05$	$FWE \leq 0.05$
VRU-Premier	98,910	855 (95%)	872 (97%)
VRU-Business	225,862	700 (78%)	783 (87%)
VRU-Loan	16,610	882 (98%)	884 (98%)
Business	88,988	762 (85%)	806 (90%)
Message	101,619	628 (70%)	721 (80%)

Table 3 Out-of-sample results from applying the tests in Section 4 to calls to the bank in April 2002. 30 minute subintervals are used to group calls between 7am-10pm, resulting in a total of 900 subintervals.

7. Discussion

By introducing ideas from recent statistical learning literature on queueing data, we find that emergency department and call centre arrival patterns are consistent with a NHPP with a sinusoidal rate of the form (3). Given the simplicity and flexibility of the sinusoidal specification, our findings provide researchers with empirical support for its use in practice. Moreover, as discussed in Section 1.2, the specification has certain practical advantages over piecewise-smooth fits. From a modelling standpoint, we hope our empirical findings will also encourage queuing theorists to revisit the sinusoidal specification as a workhorse model for time-varying arrivals. Such efforts could generate further analytic results that add to the existing ones for cyclic arrival rates (Eick et al. 1993, Feldman et al. 2008, Liu and Whitt 2012).

The sinusoidal model considered here is of course not perfect. For example, it is well known that sinusoids are not efficient at capturing time-localized demand surges, like the spike in arrivals of ESI levels 3-5 patients on Mondays. The fact that holidays have an impact on arrivals on the day before and on the day after is another example of a time-localized demand surge. A future direction for further refinement is to augment the sinusoids with time-localized basis functions like

wavelets to capture these effects. However, this increase in flexibility needs to be traded off against the resulting decrease in tractability.

The specification considered here implicitly assumes that the arrival rate is deterministic. We note that there is a stream of literature surveyed in Ibrahim et al. (2016) that treats the arrival process as a NHPP but with a stochastic arrival rate, i.e. a Cox process. As discussed in (Jongbloed and Koole 2001, Steckley et al. 2005, Aksin et al. 2007, Steckley et al. 2009, Bassamboo et al. 2010) and the references therein, this can happen if $\lambda(t)$ depends on unobserved variables which introduce additional uncertainty into its estimate. These more complex models represent refinements that are particularly suited to supporting decision making over short time scales, where information on past arrivals can be used to forecast future demand (Jongbloed and Koole 2001, Avramidis et al. 2004, Weinberg et al. 2007, Shen and Huang 2008a,b, Soyer and Tarimcilar 2008, Aldor-Noiman et al. 2009, Mehrotra et al. 2010, Aktekin and Soyer 2011, Taylor 2012, Ibrahim and L'Ecuyer 2013, Zhang et al. 2014, Gans et al. 2015, Ibrahim et al. 2016, Oreshkin et al. 2016). It may be interesting in the future to explore how to estimate the sinusoidal functional form in the framework of a Cox process. This is likely to be technically challenging given the more complicated noise structure, which will depend explicitly on the specific model under consideration.

Acknowledgments

The emergency department arrivals data was kindly provided by Dr. Kito Lord. We are also grateful to Noa Zychlinski and the SEELab at Technion for providing us with access to the call centre arrivals data.

References

Aksin, Z, M Armony, V Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.

Aktekin, T, R Soyer. 2011. Call center arrival modeling: A bayesian state-space approach. *Naval Research Logistics (NRL)* **58**(1) 28–42.

Aldor-Noiman, S, PD Feigin, A Mandelbaum. 2009. Workload forecasting for a call center: Methodology and a case study. *Annals of Applied Statistics* 1403–1447.

Alizadeh, F, J Eckstein, N Noyan, G Rudolf. 2008. Arrival rate approximation by nonnegative cubic splines. *Operations Research* **56**(1) 140–156.

Amrein, WO, AM Berthier. 1977. On support properties of L^p-functions and their Fourier transforms. *Journal of Functional Analysis* **24**(3) 258–267.

Armony, M, S Israelit, A Mandelbaum, YN Marmor, Y Tseytlin, GB Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* **5**(1) 146–194.

Avramidis, AN, A Deslauriers, P L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Management Science* **50**(7) 896–908.

Bartlett, MS. 1963. The spectral analysis of point processes. *J R Statist. Soc. B* 264–296.

Bassamboo, A, RS Randhawa, A Zeevi. 2010. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* **56**(10) 1668–1686.

Benedicks, M. 1985. On Fourier transforms of functions supported on sets of finite Lebesgue measure. *Journal of Mathematical Analysis and Applications* **106**(1) 180–183.

Benjamini, Y, D Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 1165–1188.

Brown, L, N Gans, A Mandelbaum, A Sakov, H Shen, S Zeltyn, L Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* **100**(469) 36–50.

Chan, CW, J Dong, LV Green. 2016. Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research* **65**(2) 469–495.

Chen, NY, DKK Lee, SN Negahban. 2016. Super-resolution estimation of cyclic arrival rates. *Working paper* .

Eick, SG, WA Massey, W Whitt. 1993. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science* **39**(2) 241–252.

Feldman, Z, A Mandelbaum, WA Massey, W Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Science* **54**(2) 324–338.

Gans, N, H Shen, YP Zhou, N Korolev, A McCord, H Ristock. 2015. Parametric forecasting and stochastic programming models for call-center workforce scheduling. *Manufacturing & Service Operations Management* **17**(4) 571–588.

Green, LV, PJ Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* **37**(1) 84–97.

Green, LV, PJ Kolesar, J Soares. 2001. Improving the sipp approach for staffing service systems that have cyclic demands. *Operations Research* **49**(4) 549–564.

Green, LV, PJ Kolesar, W Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16**(1) 13–39.

Green, LV, J Soares, JF Giglio, RA Green. 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13**(1) 61–68.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 65–70.

Huang, J, B Carmeli, A Mandelbaum. 2015. Control of patient flow in Emergency Departments, or multiclass queues with deadlines and feedback. *Operations Research* **63**(4) 892–908.

Ibrahim, R, P L'Ecuyer. 2013. Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management* **15**(1) 72–85.

Ibrahim, R, H Ye, P L'Ecuyer, H Shen. 2016. Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting* **32**(3) 865–874.

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

Jongbloed, G, G Koole. 2001. Managing uncertainty in call centres using poisson mixtures. *Applied Stochastic Models in Business and Industry* **17**(4) 307–318.

Kim, SH, W Whitt. 2014a. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing & Service Operations Management* **16**(3) 464–480.

Kim, SH, W Whitt. 2014b. Choosing arrival process models for service systems: Tests of a nonhomogeneous poisson process. *Naval Research Logistics* **61**(1) 66–90.

Kuhl, ME, JR Wilson, MA Johnson. 1997. Estimating and simulating Poisson processes having trends or multiple periodicities. *IIE Transactions* **29**(3) 201–211.

Lee, S, JR Wilson, MM Crawford. 1991. Modeling and simulation of a nonhomogeneous Poisson process having cyclic behavior. *Communications in Statistics-Simulation and Computation* **20**(2-3) 777–809.

Liu, Y. 2018. Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Operations Research* **66**(2) 514–534.

Liu, Y, W Whitt. 2012. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research* **60**(6) 1551–1564.

Mandelbaum, A. 2017. Service Enterprise Engineering (SEE) lab. Assessed Nov 28 2017. *web.iem.technion.ac.il/en/service-enterprise-engineering-see-lab/general-information.html* .

Massey, WA, GA Parker, W Whitt. 1996. Estimating the parameters of a nonhomogeneous Poisson process with linear rate. *Telecommunication Systems* **5**(2) 361–388.

Mehrotra, V, O Ozlük, R Saltzman. 2010. Intelligent procedures for intra-day updating of call center agent schedules. *Production and Operations Management* **19**(3) 353–367.

Moitra, A. 2015. Super-resolution, extremal functions and the condition number of vandermonde matrices. *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*. ACM, 821–830.

Oreshkin, BN, N Réegnard, P L'Ecuyer. 2016. Rate-based daily arrival process models with application to call centers. *Operations Research* **64**(2) 510–527.

Parker, B. 2011. The tide predictions for D-Day. *Physics Today* **64**(9) 35–40.

Rice, JA, M Rosenblatt. 1988. On frequency estimation. *Biometrika* **75**(3) 477–484.

Saghafian, S, WJ Hopp, MP van Oyen, JS Desmond, SL Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* **60**(5) 1080–1097.

Shao, N. 2010. Modeling almost periodicity in point processes .

Shao, N, KS Lii. 2011. Modelling non-homogeneous Poisson processes with almost periodic intensity functions. *J. R. Statist. Soc. B* **73**(1) 99–122.

Shen, H, JZ Huang. 2008a. Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Annals of Applied Statistics* 601–623.

Shen, H, JZ Huang. 2008b. Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management* **10**(3) 391–410.

Shi, P, MC Chou, JG Dai, D Ding, J Sim. 2015. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science* **62**(1) 1–28.

Soyer, R, MM Tarimcilar. 2008. Modeling and analysis of call center arrival data: A bayesian approach. *Management Science* **54**(2) 266–278.

Steckley, SG, SG Henderson, V Mehrotra. 2005. Performance measures for service systems with a random arrival rate. *Proceedings of the 37th conference on Winter simulation*. Winter Simulation Conference, 566–575.

Steckley, SG, SG Henderson, V Mehrotra. 2009. Forecast errors in service systems. *Probability in the Engineering and Informational Sciences* **23**(2) 305–332.

Taylor, JW. 2012. Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science* **58**(3) 534–549.

Vere-Jones, D. 1982. On the estimation of frequency in point-process data. *J. of Appl. Probab.* 383–394.

Weinberg, J, LD Brown, JR Stroud. 2007. Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association* **102**(480) 1185–1198.

Whitt, W. 2014. Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters* **42**(6) 458–461.

Whitt, W. 2016. Heavy-traffic fluid limits for periodic infinite-server queues. *Queueing Systems* **84**(1-2) 111–143.

Whitt, W, X Zhang. 2017. A data-driven model of an Emergency Department. *Operations Research for Health Care* **12** 1–15.

Zeltyn, S, YN Marmor, A Mandelbaum, B Carmeli, O Greenshpan, Y Mesika, S Wasserkrug, P Vortman, A Shtub, T Lauterman, Schwartz D, Moskovitch K, Tzafrir S, Basis F. 2011. Simulation-based models of emergency departments:: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation* **21**(4) 24.

Zhang, X, LJ Hong, J Zhang. 2014. Scaling and modeling of call center arrivals. *Proceedings of the Winter Simulation Conference*. IEEE, 476–485.