Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

SHORT-TERM ADAPTIVE EMERGENCY CALL VOLUME PREDICTION

Elioth Sanabria
Henry Lam
Enrique Lelo de Larrea
Jay Sethuraman

Department of IE&OR
Columbia University
500 West 120th Street
New York, NY 10027, USA

Sevin Mohammadi
Audrey Olivier
Andrew W. Smyth

Department of Civil Eng. & Eng. Mechanics
Columbia University
500 West 120th Street
New York, NY 10027, USA

Edward M. Dolan
Nicholas E. Johnson
Timothy R. Kepler
Afsan Quayyum
Kathleen S. Thomson

Bureau of Management Analysis and Planning
Fire Department, City of New York
9 MetroTech Center
Brooklyn, NY 11201, USA

ABSTRACT

Sudden periods of extreme and persistent changes in the distribution of medical emergencies can trigger resource planning inefficiencies for Emergency Medical Services, causing delayed responses and increased waiting times. Predicting such changes and reacting adaptively can alleviate these adversarial impacts. In this paper, we propose a simple framework to enhance historically calibrated call volume models, the latter a focus of study in the arrival estimation literature, to give more accurate short-term prediction by refitting their residuals into time series. We discuss some justification of our framework from the perspective of doubly stochastic Poisson processes. We illustrate our methodology in predicting the hourly call volume to the 911 call center during the Covid-19 pandemic in NYC, showing how it could improve the performance of baseline historical estimators by close to 50% measured by the out-of-sample prediction error for the next hour.

1 INTRODUCTION

During the Covid-19 pandemic, Emergency Medical Services (EMS) suffered episodes of unusual increased demand that were severe both in magnitude and duration. EMS worldwide were overburdened to levels beyond maximum capacity during the peak periods of the Covid-19 waves. For example, during the worst months of the early Covid-19 pandemic in NYC, the call volume to the 911 ambulance call center was significantly above the historical average (see Figure 1).

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

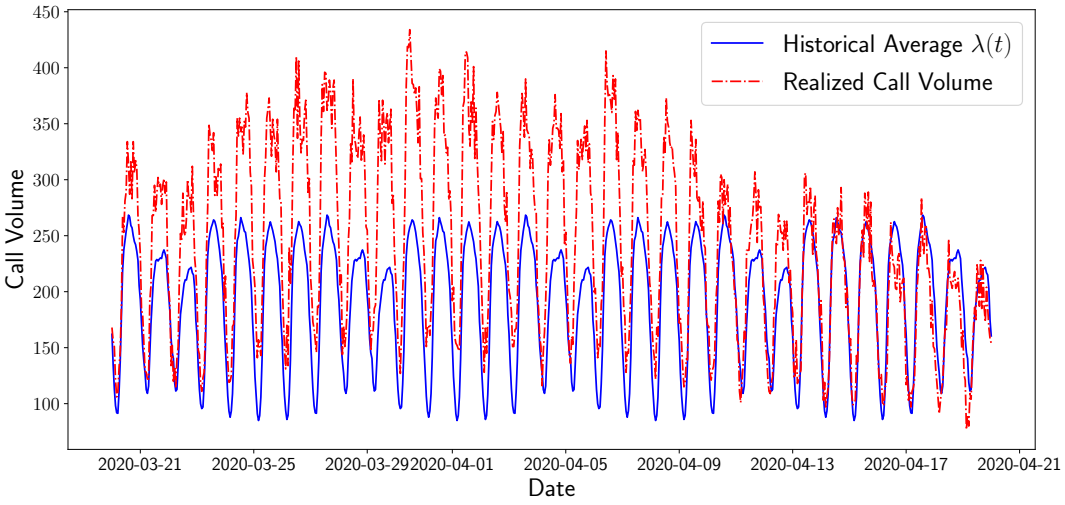

Figure 1: Realized hourly call volume (red line) vs baseline historical average $\lambda(t)$ for the month with highest call volume during 2020.

During unusual periods such as those encountered in Figure 1, using solely a historical estimation of the demand could suffer poor performances. On the other hand, much of the call volume or arrival process literature, at least within the operations research community, has focused on historical models. Typically, an arrival process is modeled by means of specifying an intensity function (either parametric or non-parametric) of a Non-homogeneous Poisson process, and fitted to historical data using Maximum Likelihood (ML). These include, for instance, piece-wise constant intensities in Morgan et al. (2016) and linear in Glynn and Zheng (2019), and splines in Morgan et al. (2019). Taylor and Letham (2018) study a "prophet" forecasting paradigm to estimate intensity functions by combining many popular models used widely by practitioners. Also related is Kim and Whitt (2014) that test the Poisson process assumption of call arrival processes with piece-wise constant or linear intensities, by dividing time into subintervals and testing the conditional order statistic distribution using the Kolmogorov-Smirnov test. Moreover, in the spirit of time series modeling, Ibrahim and L'Ecuyer (2013) study linear models for call center volume prediction, and Matteson et al. (2011) use a latent variable time series model to forecast based on seasonal covariates of medical emergency arrivals. Both of these works are similar to our approach, although we focus on studying the effect of the estimation window length for capturing the changes of distribution, as well as capturing the conditional changes of the arrival counts distribution. Finally, for a survey on call center arrival modeling, see Ibrahim et al. (2016).

To handle and react to unusual patterns that deviate from the historical, in this paper we focus on short-term prediction that corrects for the latter. In particular, we consider a simple framework to correct a given historical estimator (which could be built from any approaches described above) in order to adapt to recent changes in the call arrival pattern. More specifically, our framework models the residuals between the recent observed volume and the historical model prediction via an additional layer of time series, which gives a short-term adjustment on the historical model based on the most recent data. This allows us to make short-term forecast that is better than using a standalone historical model or a pure autoregression-type time series. Our framework resembles time series with seasonality (e.g., Brockwell and Davis (2016) Section 1.4), but the seasonality component can be flexibly modeled by any established historical call volume models in the arrival estimation literature. We discuss how our framework connects to the use of doubly stochastic Poisson processes.

To give some motivational illustration, suppose we simply use the historical average of the same time of the year as the historical model. Figure 2 shows that the higher call volume days of the Covid-19 pandemic in NYC bore a sizable residual (i.e., gap) between the observed volume and the historical average very early in the morning (the figure shows that after 3 a.m. the call volume was already outside the historical confidence interval). If we capture the trend of the residual as a time series as in our proposed framework early in the morning (for example at 6a.m. in Figure 2, when the residual gap was already sizable), the hours of higher volume later in the day (after 10 a.m.) could have been forecasted, and action could have been preemptively taken to deal with the increased demand. Indeed, the primary motivation for studying our approach is to enhance downstream decision-making tasks including adaptive staffing and call shedding to more accurately react to the real-time call volume trend.

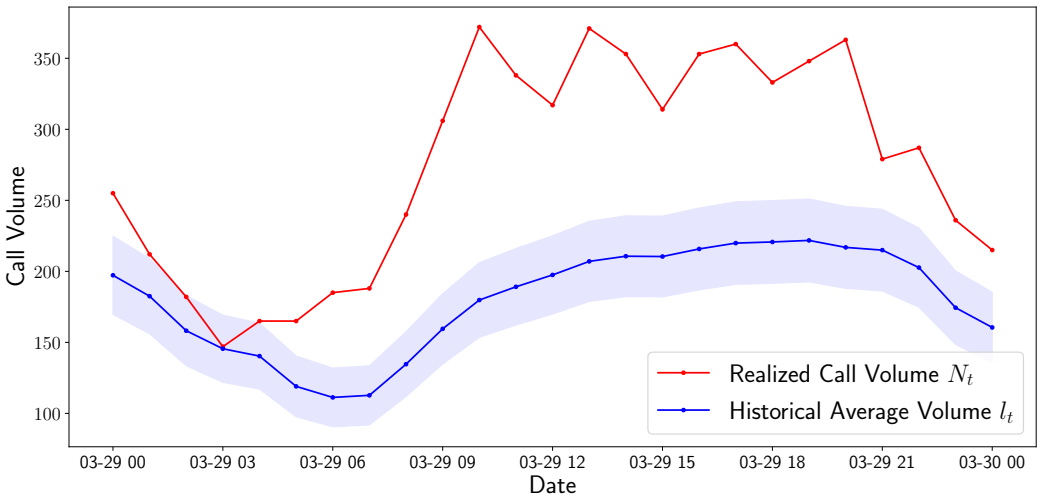

Figure 2: Realized hourly call volume vs historical average estimated using hourly arrivals (by hour of the week with data from 2018-2019), for a day with previously unseen high call volume during the Covid-19 crisis.

We close this introduction by discussing a small number of works on short-term prediction for call volumes to our knowledge. L'Ecuyer et al. (2018) specify a doubly stochastic intensity function to characterize short lived call bursts in a call center (depending on 3 random parameters, denoting the amplitude, decay and duration of the burst). Along the same vein, Oreshkin et al. (2016) study a doubly stochastic intensity, specified as a deterministic intensity multiplied by a time-varying stochastic factor. These approaches have in common the need of specifying an underlying doubly stochastic intensity to capture the periods when the call distribution changes. Our fitting approach can also be viewed as modeling a doubly stochastic process, but instead of fitting an explicit model for the intensity, we formulate a data-driven time series estimation of the prediction error of a given historical prediction model which is then related to an implicit underlying doubly stochastic intensity.

The rest of this paper is as follows. Section 2 presents our algorithm and analysis. Section 3 presents a numerical example of our methodology to predict 911 call volume during the Covid-19 pandemic in NYC. Section 4 gives some concluding remarks.

2 METHODOLOGY

We consider discrete time intervals of size Δ, say 1 hour. Time is indexed as $\Delta t = t$ for $t \in \mathbb{Z}$. We set the following notation:

- $N_t :=$ Random arrival count in the time interval $[t-1,t)$. Think of this as the arrival counts for a call center.
- $l_t :=$ Historical estimator for $\mathsf{E}N_t$. This estimator can be a seasonal historical average of the arrivals or be associated to a calibrated intensity function $\lambda(u)$ such that $l_t = \int_{t-1}^{t} \lambda(u)du$.
- $\varepsilon_t = N_t - l_t :=$ Residual (error) between the random unobserved volume N_t and the historically calibrated expected volume l_t.
- $\mathscr{F}_{(t-k):t} :=$ Information from time $t-k$ to t, namely $\{N_j\}_{j=t-k}^{t}$. This is used to represent the recently observed arrival counts.

We assume the historical model or estimator $\{l_t\}$ is given and is independent of the most recent observations. We aim to model the series $\{\varepsilon_t\}$ which is the residual between recent volume N_t and the historical model l_t. Then, suppose we want to predict N_{t+1} for the next hour $t+1$, we would first compute $\mathsf{E}(\hat{\varepsilon}_{t+1}|\mathscr{F}_{(t-k):t})$, i.e., the expected residual conditional on the most recent information, and then the short-term estimator for N_{t+1} is $\hat{s}_{t+1} := l_{t+1} + \mathsf{E}(\hat{\varepsilon}_{t+1}|\mathscr{F}_{(t-k):t})$. In other words, $\hat{s}_{t+1}$ corrects the historical estimator l_{t+1} shifting it by the amount implied by the expected residual $\mathsf{E}(\hat{\varepsilon}_{t+1}|\mathscr{F}_{(t-k):t})$. We summarize our proposed procedure in Algorithm 1, where we use an AR-GARCH (Brockwell and Davis (2016)) to fit the residuals. The AR component is used to model a time-varying mean of the residual ε_t, while the GARCH is used to model the variance. We could choose other time series models, but the one in Algorithm 1 appears to work reasonably in our experimental tests, and we can also make some intuitive (though not rigorous) argument to support its use, which we discuss next.

Algorithm 1: Short-term estimator with estimation window of size k at time t.
Inputs: Estimation window k, historical arrival counts $\{N_j\}_{j=t-k}^{t}$, given historical model $\{l_j\}_{j=t-k}^{t}$.

1. Compute residual series $\{\varepsilon_j := N_j - l_j\}_{j=t-k}^{t}$.
2. Fit an AR-GARCH model $\varepsilon_t = \varphi + \sum_{i=1}^{p} \varphi_i \varepsilon_{t-i} + Z_t\sigma_t$, $\sigma_t^2 = \alpha + \sum_{j=1}^{q} \alpha_j \eta_{t-j}^2 + \sum_{s=1}^{r} \beta_s \sigma_{t-s}^2$, $\eta_t = Z_t\sigma_t$, by Maximum Likelihood.
3. Output the short-term predictor $\hat{s}_{t+1} = l_{t+1} + \mathsf{E}(\hat{\varepsilon}_{t+1}|\mathscr{F}_{(t-k):t}) = l_{t+1} + \hat{\varphi} + \sum_{i=1}^{p} \hat{\varphi}_i \varepsilon_{t+1-i}$.

2.1 Intuitive Justification of AR-GARCH via Doubly Stochastic Poisson Processes

We provide some intuitive justification to using Algorithm 1, based on the perspective that the arrivals naturally arise from a doubly stochastic Poisson process (DSPP), a model that is widely used in the arrival modeling literature. First, supposing the arrivals follow a non-homogeneous Poisson Process (NHPP) with deterministic intensity $\lambda(u)$, then the arrival count in time interval $[t-1,t]$ is Poisson with rate $l_t = \mathsf{E}N_t = \int_{t-1}^{t} \lambda(u)du$. In this setting, the residual series $\varepsilon_t = N_t - l_t$ is a sequence of independent mean-zero random variables (with variance l_t). If we use Algorithm 1, then when k increases, the AR-GARCH model would simply converge to a constant 0 mean for the residual and time dependent variance l_t. In this case, our approach does not do any enhancement to the historical model, but we also don't harm it materially either.

For a general DSPP, the intensity function λ is itself random, and given the intensity realization the arrivals are Poisson like in NHPP. In this case, the ultimate arrival counts possess both the *intensity randomness* and the *arrival randomness*. Given a realization of the intensity, suppose it remains fixed for some amount of time (reducing the intensity randomness in the short-term) and that the realized intensity can be approximated by an autoregressive recursion (more on this later), then, the residual is composed of a deterministic autoregressive recursion (from the intensity) plus the randomness of the arrivals, leading to an autoregressive process (AR) for the residual. Likewise, as the variance of the Poisson is equal to its mean, the variance also has an autoregresive structure in the short-term, captured by the GARCH specification in our algorithm. Thus, both the mean and variance of the residual depend in some sense on the observations in the recent past periods. Intuitively, estimating the mean of the residual (AR part), requires accounting (controlling) for the time-varying variance of the arrivals (GARCH) whenever present.

2.2 Regime-Switching Models: A Simple Example

Consider the following regime-switching model with two regimes: Arrivals are generated by a doubly stochastic Poisson intensity $\lambda(u) = \ell_t \lambda_1(u) + (1-\ell_t)\lambda_2(u)$ where $\ell_t \in \{0,1\}$ is an indicator fixed for a deterministic amount of time T, and after that, it is sampled as a Bernoulli random variable with success rate p. In this example, ℓ_t represents an unobservable random mechanism that selects one of the two intensities λ_1, λ_2 and fixes either for a deterministic amount of time T. Moreover, $p \nearrow 1$, meaning that with high probability only the intensity $\lambda_1(u)$ is observed most of the time. Calibrating a (discrete) non-homogeneous process on this doubly stochastic process gives $l_{t+1} = \mathsf{E}N_{t+1} = \int_t^{t+1}[p\lambda_1(u) + (1-p)\lambda_2(u)]du \approx \int_t^{t+1}\lambda_1(u)du$. In this case, the distribution of the residual ε_{t+1} conditional on ℓ_{t+1} can be written as mixture of two independent normals as:

$$\varepsilon_{t+1}|\ell_{t+1} \overset{\mathrm{d}}{=} \ell_{t+1}\left[(l_{t+1}-l_{t+1}) + Z_{t+1}\sqrt{l_{t+1}}\right] + (1-\ell_{t+1})\left[(s_{t+1}-l_{t+1}) + Z'_{t+1}\sqrt{s_{t+1}}\right], \tag{1}$$

where Z_{t+1}, Z'_{t+1} are independent standard normals. $s_{t+1} := \int_t^{t+1}\lambda_2(u)du$ is the Poisson mean when $\ell_{t+1} = 0$ (in this case the distribution is centered at $s_{t+1} - l_{t+1}$ as opposed to 0 when $\ell_{t+1} = 1$). This expression comes from a CLT approximation of the Poisson distribution to the normal assuming the rates l_{t+1}, s_{t+1} are large enough.

Without loss of generality the residual can be written as a normal with random mean μ_{t+1} and random variance σ^2_{t+1}. Where the randomness is conditional on ℓ_{t+1}:

$$\varepsilon_{t+1} \overset{\mathrm{d}}{=} \mu_{t+1}(\ell_{t+1}) + Z_{t+1}\sigma_{t+1}(\ell_{t+1}). \tag{2}$$

While μ_{t+1}, σ_{t+1} are random (depending on ℓ_{t+1}), assume I_{t+1} has been fixed at least for k amount of time (with $k << T$), the path defined by $\mu_{t-k}, \sigma_{t-k} \cdots, \mu_{t-1}, \sigma_{t-1}, \mu_t, \sigma_t$ has the same intensity selected, that is, $\ell_{t-k} = \cdots = \ell_{t-1} = \ell_t = i$. Thus, if μ_{t+1}, σ_{t+1} can be written as functions f, g of the previous k observations $\mu_{t-k}, \sigma_{t-k} \cdots, \mu_t, \sigma_t$, then, the distribution of ε_{t+1} can be estimated at time t and be used to correct the bias of the deterministic predictor l_{t+1}. As:

$$\varepsilon_{t+1} \overset{\mathrm{d}}{=} \mu_{t+1}(i) + Z_{t+1}\sigma_{t+1}(i) = f(\mu_{t-k}, \sigma_{t-k} \cdots, \mu_t, \sigma_t) + Z_{t+1}g(\mu_{t-k}, \sigma_{t-k} \cdots, \mu_t, \sigma_t), \tag{3}$$

which in turn can be used to correct the prediction l_{t+1} as $\mathsf{E}(\varepsilon_{t+1}|\mathscr{F}_{(t-k):t})$ is equal to $\mathsf{E}(\mu_{t+1}|\mathscr{F}_{(t-k):t}) = f(\mu_{t-k}, \sigma_{t-k} \cdots, \mu_t, \sigma_t)$, that is, the expected bias (residual) of the estimator l_{t+1} conditional on $\mathscr{F}_{(t-k):t}$. This follows from:

$$l_{t+1} + \mathsf{E}(\mu_{t+1}|\mathscr{F}_{(t-k):t}) = l_{t+1} + \mathsf{E}(\varepsilon_{t+1}|\mathscr{F}_{(t-k):t}) = l_{t+1} + \mathsf{E}(N_{t+1} - l_{t+1}|\mathscr{F}_{(t-k):t}) = \mathsf{E}(N_{t+1}|\mathscr{F}_{(t-k):t}). \tag{4}$$

The first equality follows from the above discussion, the second one follows by definition, while for the last one note that l_{t+1} is constant and $\mathscr{F}_t$ measurable (known). Then, denote our corrected short-term estimator by $s_{t+1} = l_{t+1} + \mathsf{E}(\mu_{t+1}|\mathscr{F}_{(t-k):t})$.

We summarize the intuition of this section in the following points:

- The residual $\varepsilon_{t+1} = N_{t+1} - l_{t+1}$ can be written as a normal random variable with random (time-dependent) mean and variance $\mu_{t+1}, \sigma^2_{t+1}$ dependent on a unobservable random variable ℓ_{t+1} that is persistent over time (meaning that the unobservable variable remains constant for some period of time).
- If μ_{t+1}, σ_{t+1} can be written as functions f, g of k previous observations $\mu_{t-k}, \sigma_{t-k} \cdots, \mu_t, \sigma_t$. Then, the distribution of the residual ε_{t+1} can be estimated at time t as $\mu_{t-k}, \sigma_{t-k} \cdots, \mu_t, \sigma_t$ and μ_{t+1}, σ_{t+1} share the same unobservable driving random variable.
- With an estimator of μ_{t+1} at time t we get a conditional prediction of the unobserved volume $\mathsf{E}(N_{t+1}|\mathscr{F}_{(t-k):t})$ given by $l_{t+1} + \mathsf{E}(\mu_{t+1}|\mathscr{F}_{(t-k):t})$.

In this paper we take the functional form of f, g as a time series model with time-varying mean/variance (modeling the variance as a GARCH model, see Bollerslev (1986)) given by $\varepsilon_t = \varphi + \sum_{i=1}^{p} \varphi_i \varepsilon_{t-i} + Z_t \sigma_t$ in **Algorithm 1**. Letting $\mu_t := \varphi + \sum_{i=1}^{p} \varphi_i \varepsilon_{t-i}$ leads to an equivalent expression to Equation (2) with $\varepsilon_t = \mu_t + Z_t \sigma_t$. In this case, the predicted correction is given by $\mathsf{E}(\varepsilon_{t+1}|\mathscr{F}_{(t-k):t}) = \mathsf{E}(\mu_{t+1}|\mathscr{F}_{(t-k):t}) = \varphi + \sum_{i=1}^{p} \varphi_i \varepsilon_{t+1-i}$.

In the next subsection we generalize the unobservable mechanism to change distributions over time beyond two distributions and a fixed amount of time to a case with countably many distributions and random time for switching them. We show that the time series model described in **Algorithm 1**, while simple, is flexible enough to accommodate cases of interest for changes of regime.

2.3 General Regime-Switching Models

For a generalization of the previous case, consider a countable collection of γ-periodic, deterministic intensities $\{\lambda_i(u)\}_{i=1}^{\infty}$. As time is discrete, let $s_t(i) := \int_{t-1}^{t} \lambda_i(u)du$, the periodicity of λ_i implies $s_t(i) = s_{t+\gamma}(i)$ for all t. Let I be an independent discrete random variable with support $\{1, \ldots, \infty\}$ sampled with probabilities $\{p_i\}_{i=1}^{\infty}$ denoting which regime is sampled when there is a regime change. Denote τ as a geometric random variable with mean $\mathsf{E}\tau$. Let ℓ_t be a stochastic process written as:

$$\ell_t = \ell_{t-1}(1-B) + IB, \tag{5}$$

where B is a Bernoulli random variable with success probability $1/\mathsf{E}\tau$. ℓ_t can be interpreted as the indicator to select one of the intensities $\{\lambda_i(u)\}_{i=1}^{\infty}$ that will be set for a geometric time until ℓ_t changes (switches) to another intensity (once B is 1 and I is sampled). That is, the arrivals are distributed Poisson conditional on the intensity selected by ℓ_t (for a geometric amount of time) as:

$$N_t \sim \text{Poisson}(s_t(\ell_t)|\ell_t). \tag{6}$$

Likewise, denote $l_t := \mathsf{E}N_t = \mathsf{E}[s_t(\ell_t)] = \sum_{i=1}^{\infty} p_i s_t(i)$ and $\varepsilon_{t+1} = N_{t+1} - l_{t+1}$. As in the previous section the residual can be written as $\varepsilon_{t+1} \overset{\text{d}}{=} \mu_{t+1} + Z_{t+1}\sigma_{t+1}$ with $\mu_{t+1} = s_{t+1}(\ell_{t+1}) - l_{t+1}$ and $\sigma_{t+1} = \sqrt{s_{t+1}(\ell_{t+1})}$.

In the previous section and in **Algorithm 1** we propose an AR-GARCH $\mathscr{F}_{(t-k):t}$ measurable approximation of ε_{t+1}. In general, for such approximations to be correct further assumptions are needed on the mean of the geometric time $\mathsf{E}\tau$, the functional form of the intensities $\{\lambda_i(u)\}_{i=1}^{\infty}$ and how concentrated are the sampling probabilities $\{p_i\}_{i=1}^{\infty}$. Next, we discuss some cases of interest when the mean μ_{t+1} and variance σ_{t+1} can be written as time series model and estimated using our procedure in **Algorithm 1**. The first, is the most common way of thinking of shifts in the trend of the residual. The second shows that continuous and periodic intensity rates have an autoregressive structure representation that leads to a time series model for the residual. And in the last example a totally general change in the distribution of the arrival counts (intensity) is considered where the estimation price is that the regime changes should be longer.

Shifted Intensities: For example, the time series model in **Algorithm 1** corrects the bias of the historical estimator l_t if the intensities are just "shifted" from some baseline intensity, that is $\lambda_i(u) = c_i + \lambda(u)$ where c_i is the shift from the baseline intensity. Here, $l_t = \mathsf{E}c_i + \int_{t-1}^{t} \lambda(u)du$. In this case, μ_{t+1} is constant and equal to $(c_{\ell_t} - \mathsf{E}c_i)\Delta$ while ℓ_t remains fixed (that is, for the geometric amount of time τ). Conditional that the intensity selected remains the same, that is, $\ell_{t+1} = \ell_t = \cdots$. We have μ_{t+1} could be written as a AR(1) process with $\varphi + \varphi_1 \varepsilon_t$ such that $\mu_t = \frac{\varphi}{1-\varphi_1} = (c_{\ell_t} - \mathsf{E}c_i)\Delta = \mu_{t+1}$ (recalling the mean of an AR process is constant, then taking expected values on both sides of $\varepsilon_{t+1} = \varphi + \varphi_1 \varepsilon_t$ yields $\mu_{t+1} = \varphi + \varphi_1 \mu_t$, as $\mu_{t+1} = \mu_t$ implies $\mu_t = \mu_{t+1} = \frac{\varphi}{1-\varphi_1}$).

To see why $(c_{\ell_t} - \mathsf{E}c_i)\Delta = \mu_{t+1}$, conditional ℓ_{t+1} note that $\mathsf{E}(N_{t+1}|\ell_{t+1}) = \Delta c_{\ell_{t+1}} + \int_{t}^{t+1} \lambda(u)du$, additionally, $\mathsf{E}(\varepsilon_{t+1}|\ell_{t+1}) = \mu_{t+1} = \mathsf{E}(N_{t+1} - l_{t+1}|\ell_{t+1}) = \mathsf{E}(N_{t+1}|\ell_{t+1}) - [\Delta \mathsf{E}c_i + \int_{t}^{t+1} \lambda(u)du]$. Putting the two expressions together yields $\mu_{t+1} = (c_{\ell_{t+1}} - \mathsf{E}c_i)\Delta$, which as long as the intensity does not change, that is, $\ell_{t+1} = \ell_t = \cdots$, the means $\mu_{t+1} = \mu_t$ will remain constant (for a geometric amount of time).

On the other hand, the variance is time dependent. That is $\sigma_{t+1} = \sqrt{c_{\ell_{t+1}} + \int_t^{t+1} \lambda(u)du}$, thus the need to specify the GARCH component of the autoregressive process (with a constant intercept $\alpha = c_{\ell_{t+1}}$ and a seasonal time-varying term $\int_t^{t+1} \lambda(u)du$). Estimating the model with a rolling window of k observations (as seen later on a condition for $k << \mathsf{E}\tau$, such that the intensity indicator ℓ_t remains constant with high probability for the estimation window) allows to estimate the mean and variance of the residual using **Algorithm 1** for large enough k.

Continuous intensities/rates: As another key use case of the generality of our approach, here we show the intuition that continuous intensities/rates (in a closed interval) can be written as autorregresive processes specified by our **Algorithm 1**: By Fourier analysis it is a well known fact that any continuous function in a compact interval can be approximated by a linear combination of sines and cosines. Then, as the rates considered are periodic, they can be approximated as a Fourier series. Next, it can be shown that a sine function can be written recursively as $\sin(t+1) = 2\cos(\Delta)\sin(t) - \sin(t-1)$ (recalling time is indexed $t = \Delta t$. The identity follows from elementary trigonometric identities, and the same can be done for cosines). For example, letting the rates be $s_t(i) = \sin(t) + c_i$, we have that they follow an AR(2) recursion as $s_{t+1}(i) = \phi + \phi_1 s_t(i) - s_{t-1}(i)$ with $\phi = 2c_i(1-\cos\Delta)$ and $\phi_1 = 2\cos\Delta$. As this procedure can be done for sines and cosines with arbitrary amplitudes, by Fourier analysis this can be generalized to approximate any continuous and periodic rate function (including l_t), that is, as every sine and cosine has an AR representation, a linear combination of them (the Fourier series) is also an AR model. In summary, continuous and periodic rate functions have an associated AR representation.

Recall the residual is equal to $\varepsilon_{t+1} = N_{t+1} - l_{t+1} \stackrel{\mathrm{d}}{=} (s_{t+1}(\ell_{t+1}) - l_{t+1}) + Z_{t+1}\sqrt{s_{t+1}(\ell_{t+1})}$. As discussed in the previous paragraph $s_{t+1}(\ell_{t+1})$ has an autoregressive recursion (and so does $s_{t+1}(\ell_{t+1}) - l_{t+1}$, as a difference of periodic functions is also periodic), the residual follows an AR-GARCH functional form as in **Algorithm 1** while $\ell_{t+1} = \ell_{t-k}$, i.e. the regime has not switched. Intuitively, the mean of the residual $s_{t+1}(\ell_{t+1}) - l_{t+1}$ has a different autoregressive representation than its variance $s_{t+1}(\ell_{t+1})$, which reflects the need to specify the GARCH component when fitting the time series model in addition to the AR component for the mean.

General Intensities and Seasonality: Exploiting the γ-periodicity of the intensities leads to another interesting case: Recall that the point of doing prediction in this setting amounts to find $\mathscr{F}_{(t-k):t}$ measurable approximation/estimation of μ_{t+1} and σ_{t+1} while the regime (intensity) does not change. As the intensities are γ periodic, μ_{t+1} and σ_{t+1} could be estimated by looking at the observations $\{\varepsilon_{t+1-\gamma k}\}_{k=1}^{K}$, that, as long as the indicator for intensities does not change, these are i.i.d. observations with the same distribution as ε_{t+1} which is the same as the NHPP case except in the sense that the estimation is done in a window of size K instead of taking all history. Intuitively, this could be stated as: if the periodicity of the intensity is daily and there has not been a change in the distribution of the arrivals, then, a good estimator of the prediction error (residual) for the next hour is the average error of the same hour yesterday and the days before (while the intensity has not changed). This can be stated as a seasonal time series model by:

$$\varepsilon_{t+1} = \alpha + \beta\varepsilon_{t+1-\gamma k} + Z\sigma_{t+1}. \tag{7}$$

As long as the regime has not changed interval considered (of size γK), the observations are i.i.d and the mean and variance of the error are constant (modulo γ). In practice, the price to pay for the rolling window estimation is that the time τ for switching the intensity distribution should be in average longer, that is, $\gamma K << \mathsf{E}\tau$ (more on this later). As long as that is the case, estimating in a rolling window of size $K\gamma$ yields consistent estimates for the residual ε_{t+1}. Note that in this case there are no restrictions on the functional form of the intensities nor their sampling probabilities, moreover, this regression can be estimated using our procedure in **Algorithm 1** with only an AR specification of the model (modulo γ).

Limitations: The success on the convergence of the parameters of the window estimation in **Algorithm 1** (of size γK or k) hinges on the assumption that the changes of regime are not very frequent, that is, $\gamma K << \mathsf{E}\tau$ or $k << \mathsf{E}\tau$, such that the distribution of the arrivals remains the same in the estimation and forecasting window, and the mean and variance of the residual (modeled as time series) corresponds to a

single regime. If $\mathsf{E}\tau$ is small, the changes of arrival distribution are frequent and the window estimation will intuitively fail as the likelihood of having a regime change in the estimation window is equal to $1-(1-\mathsf{E}\tau^{-1})^k$ (one minus the probability there are no regime changes in k hours). For a window of size k to have only one regime (with probability a) it should be true that $\mathsf{E}\tau$ is at least $\frac{1}{1-a^{1/k}}$ (this comes from solving $(1-\mathsf{E}\tau^{-1})^k \geq a$). For example, for a window of size $k=100$ hours for the estimation to have only one regime (w.p. $a=95\%$), $\mathsf{E}\tau$ should be at least 2000 hours in average.

Nonetheless, the case when the regimes change too often highlights a key limitation of the window estimation method and the need to use different methods in this case (or restricting the assumptions on the changes of the intensities when the regime changes). In general, having larger windows is better for the asymptotic behavior of the parameters of the time series model with the caveat the estimation is correct (only has one regime) with probability $(1-\mathsf{E}\tau^{-1})^k$, which is a decreasing function on k and dependent on $\mathsf{E}\tau$.

3 NUMERICAL EXAMPLES

In this section we start by calibrating a piecewise constant intensity $\lambda(t)$ as described in Section 2 for the call arrivals to the FDNY's 911 EMS call center with historical data from January 2018 to December 2019. The calibrated intensity is plotted in Figure 3. We use this calibrated intensity as the baseline prediction $l_{t+1} = \int_t^{t+1} \lambda(u)du = \mathsf{E}N_{t+1}$. This calibrated intensity can be simply seen as the historical average weekly incidents for every hour of the week.

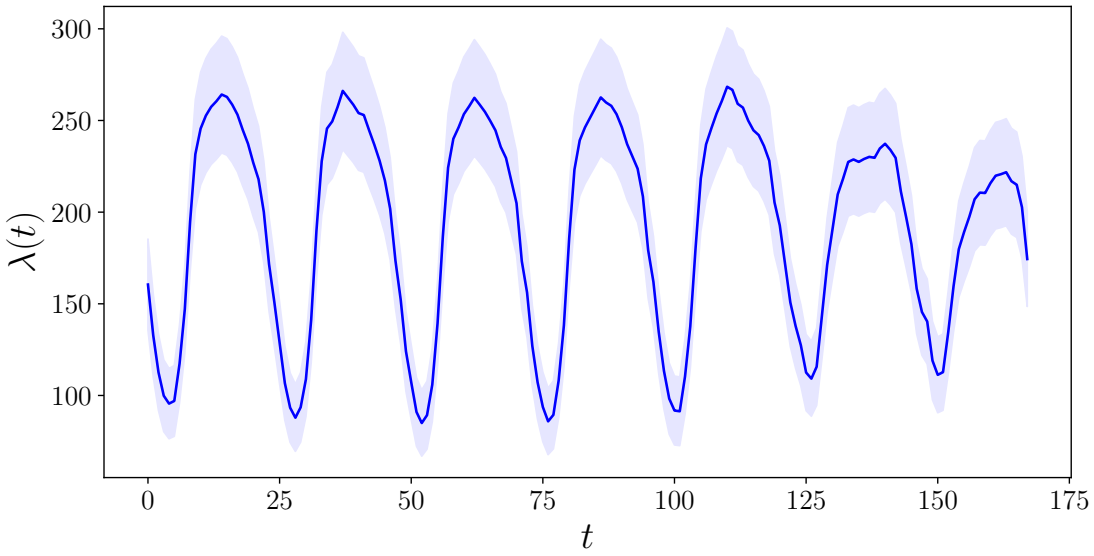

Figure 3: Example of a calibrated weekly arrival intensity $\lambda(t)$ for the 168 hours of the week with confidence intervals.

For testing our model vs the baseline prediction we use the hourly call arrivals from January to December 2020. In Figure 1 we plot a period where the call volume was higher than previously seen. In particular, the dates around April 1st showed historical highs in total call volumes at the FDNY 911 call center.

We predict the hourly call volume using **Algorithm 1** using an estimation window of $k=150$ hours. That is, to predict the call volume next j-th hour $t+j$ at time t, the previous $k=150,200,250$ hours are used to estimate the parameters of the AR-GARCH model, which are used to compute the short-term prediction $\hat{s}_{t+j} = l_{t+j} + \mathsf{E}(\hat{\varepsilon}_{t+j}|\mathscr{F}_{(t-k):t})$. We measure the performance by estimating the average absolute

error, that is, $\mathsf{E}(|N_{t+j} - \hat{s}_{t+j}|)$ out of sample. As a reference benchmark we compare with the mean absolute error of the historical prediction, that is, $\mathsf{E}(|N_t - l_t|)$ for the same testing period. In Table (1) we report the performance of the model. For the AR-GARCH model we use order $(p = 2, q = 1, r = 1)$ in reference to **Algorithm 1** (this order of the model was the most numerically stable in our data. Nonetheless, testing different order combinations is encouraged on a case by case basis).

Our short-term predictor $\hat{s}_{t+j}$ improves the performance of the long-term estimation l_{t+j} close to 50% every hour when predicting the next hour ($j = 1$). As the forecasting horizon j increases, the performance deteriorates but plateaus around a mean absolute error of 20 calls per hour, which is still a 25% improvement over the historical estimator l_t. The plateau phenomenon is expected due to the mean reverting nature of ARIMA models as time goes to infinity, making the prediction of the residual time series constant for long prediction horizons, thus, the emphasis on the short-term prediction use of the model and the need to re-calibrate the parameters on a rolling window. It can also be seen that changing the prediction window k does not give a significant performance improvement (lower values of k created numerical instabilities in the maximum likelihood estimation of the GARCH component and are omitted).

Table 1: Out-of-sample mean absolute error $\mathsf{E}(|N_{t+j} - \hat{s}_{t+j}|)$ of our short-term estimator $\hat{s}_{t+j}$ in **Algorithm 1** for different estimation windows k and prediction horizons j. As a benchmark, the mean absolute error of the historical predictor $\mathsf{E}(|N_t - l_t|)$ is equal to 27.48 calls per hour.

	Hours predicted into the future $t + j$					
	$j = 1$	$j = 2$	$j = 5$	$j = 10$	$j = 24$	$j = 36$
$k = 150$	14.63	16.04	18.56	19.52	18.79	19.96
$k = 200$	14.64	16.04	18.59	19.51	19.47	20.08
$k = 250$	14.64	16.04	18.58	19.56	20.01	20.46

In Figure 4 we plot our short-term predictor $\hat{s}_{t+1}$ against the long-term estimator l_{t+1} for the week with highest observed volume in 2020. In Figure 5 the distribution of the error of the long-term estimator l_{t+1} vs our predictor $\hat{s}_{t+1}$ is plotted. Note that with our method the error is centered at 0 with less variance and bias w.r.t. the long-term estimator, which is a desirable property.

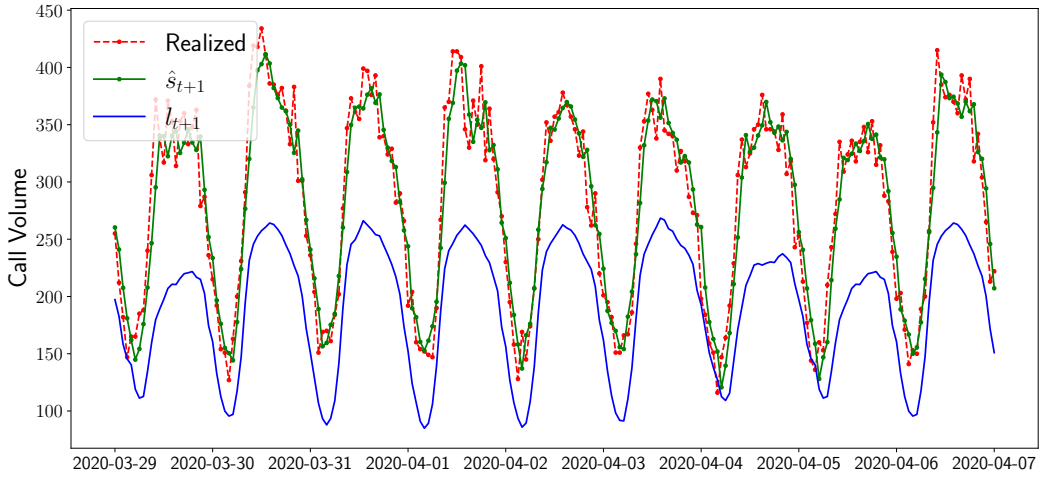

Figure 4: Comparison between our proposed short-term prediction $\hat{s}_{t+1}$ vs long-term prediction l_{t+1}, actual call arrival is the red line.

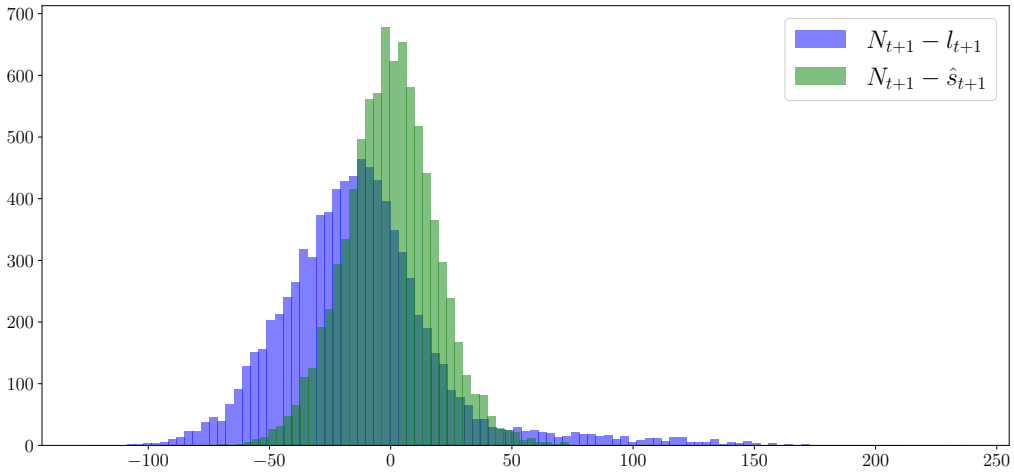

Figure 5: Histogram of the long-term prediction error ($N_{t+1} - l_{t+1}$ with mean -12.40 and standard deviation 34.23) compared with the error of our method ($N_{t+1} - \hat{s}_{t+1}$ with mean -0.04 and standard deviation 18.84).

4 CONCLUDING REMARKS

In this paper we characterized a model and an algorithm to improve a long-term historical prediction of arrivals, that is both tractable and easy to implement. This model can be used in multiple settings to forecast demand near real-time. Moreover, the model provides a complete distribution of the prediction error that can be used to build confidence intervals or do robust allocation of resources based on the distribution of the prediction error $N_{t+1} - l_{t+1} = \varepsilon_{t+1} = \mu_{t+1} + \sigma_{t+1} Z_{t+1}$.

An important use case of the model is the prediction of call arrivals to an EMS call center, where such model could direct efforts in mitigating the effects of increases in the call volume and ensuring the reliability of operations during such events (by either increasing staffing or changing the processing of calls adaptively relative to the demand).

Further research effort is needed in finding $\mathscr{F}_{(t-k):t}$ measurable approximations of μ_{t+1}, σ_{t+1} and characterizing the convergence of data-driven procedures in their calibration relative to the expected regime change time $\mathsf{E}\tau$, the assortment of intensities $\{\lambda_i\}$ and their sampling probabilities beyond the time series model presented.

ACKNOWLEDGMENTS

We are grateful to the members of the Bureau of Management Analysis and Planning at the FDNY for the many enlightening conversations and for providing their expert knowledge and data on the EMS system. We appreciate the guidance of Chief Denise Werner and Chief Ian Swords. We gratefully acknowledge support from Google and the Tides Foundation under the grant "EMS Resource Deployment Modeling" and the Columbia University Urban Technology Pilot Award.

REFERENCES

Bollerslev, T. 1986. "Generalized autoregressive conditional heteroskedasticity". *Journal of econometrics* 31(3):307–327.

Brockwell, P. J., and R. A. Davis. 2016. *Introduction to time series and forecasting*. Springer.

Glynn, P. W., and Z. Zheng. 2019. "Estimation and inference for non-stationary arrival models with a linear trend". In *2019 Winter Simulation Conference (WSC)*, 3764–3773. IEEE.

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

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

L'Ecuyer, P., K. Gustavsson, and L. Olsson. 2018. "Modeling bursts in the arrival process to an emergency call center". In *2018 Winter Simulation Conference (WSC)*, 525–536. IEEE.

Matteson, D. S., M. W. McLean, D. B. Woodard, S. G. Henderson et al. 2011. "Forecasting emergency medical service call arrival rates". *Annals of Applied Statistics* 5(2B):1379–1406.

Morgan, L. E., B. L. Nelson, A. C. Titman, and D. J. Worthington. 2019. "A spline-based method for modelling and generating a nonhomogeneous Poisson process". In *2019 Winter Simulation Conference (WSC)*, 356–367. IEEE.

Morgan, L. E., A. C. Titman, D. J. Worthington, and B. L. Nelson. 2016. "Input uncertainty quantification for simulation models with piecewise-constant non-stationary Poisson arrival processes". In *2016 Winter Simulation Conference (WSC)*, 370–381. IEEE.

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

Taylor, S. J., and B. Letham. 2018. "Forecasting at scale". *The American Statistician* 72(1):37–45.

AUTHOR BIOGRAPHIES

ELIOTH SANABRIA is a Ph.D. student in Operations Research at Columbia University. He is primarily interested in the interplay between simulation, machine learning and optimization from a probabilistic point of view, as well as their broad applications in healthcare to improve patients outcomes. His email address is m.elioth@columbia.edu.

EDWARD DOLAN is the Deputy Commissioner for Strategic Initiatives and Policy at the Fire Department of the City of New York (FDNY). His email address is edward.dolan@fdny.nyc.gov.

NICHOLAS JOHNSON is the Director of Operations Research for the FDNY. Prior to joining the FDNY, he was a Postdoctoral Associate at NYU's Marron Institute of Urban Management where his research focused on modeling real-time urban populations using mobility data. He earned a Ph.D. in Urban Science from the University of Warwick in the United Kingdom and holds a Master's degree from NYU's Interactive Telecommunications Program where he used physical computing and interaction design to explore the impact and pervasiveness of waste streams in urban environments. His email address is nicholas.johnson@fdny.nyc.gov.

TIMOTHY KEPLER is the Director of Data Quality for the FDNY. He has a Master's degree in Public Administration from the City University of New York at Baruch College. For the past five years he has worked as an analyst for the FDNY, working intensively with data from computer aided dispatch systems. His email address is timothy.kepler@fdny.nyc.gov.

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. His research focuses on simulation and optimization under uncertainty. His email address is khl2114@columbia.edu.

ENRIQUE LELO DE LARREA is a Ph.D. student in Operations Research at Columbia University. Before joining Columbia, he worked as a credit risk analyst at BBVA Mexico. He holds a double Bachelor's degree in Applied Mathematics and Actuarial Science from ITAM and a Master's degree in Operations Research from Columbia. His research interests include stochastic simulation, applied probability, and financial engineering. His email address is enrique.lelodelarrea@columbia.edu.

SEVIN MOHAMMADI is currently pursuing a Ph.D. degree in Civil Engineering and Engineering Mechanics at Columbia University. Prior to her current position, she obtained a Master's degree in Civil Engineering, majoring in Transportation, from the University of Tennessee Knoxville. Her research interests include data-driven and machine learning approaches in Transportation Engineering. Her email address is sm4894@columbia.edu.

AUDREY OLIVIER is currently an Associate Research Scientist in Civil Engineering at Columbia University, and will be joining the University of Southern California as an Assistant Professor in fall 2021. She holds a Ph.D. in Civil Engineering and Engineering Mechanics from Columbia University and a Diplôme d'Ingénieur from Ecole Centrale de Nantes, France. Her research interests revolve around probabilistic data analytics and physics-based modeling for civil engineering applications. Her email address is audreyol@usc.edu.

AFSAN QUAYYUM is a Data Scientist at the Bureau of Management Analysis and Planning at the FDNY. He has an M.S. in Mathematics and a B.S. in Mathematics with a minor concentration in Applied Physics from New York University. His work focuses on implementing statistical learning techniques for inference and prediction to help EMS and Fire Operations. His email address is afsan.quayyum@fdny.nyc.gov.

JAY SETHURAMAN is a Professor of Industrial Engineering and Operations Research at Columbia University. Currently, he serves as the chair of the IEOR department at Columbia. His research interests are in discrete optimization and applications, game theory, mechanism design, and applied probability. His email address is jay@ieor.columbia.edu.

ANDREW SMYTH is the Carleton Professor of Civil Engineering & Engineering Mechanics and also serves as the Co-Chair of the Smart Cities Center of the Data Science Institute at Columbia University. His research focuses on infrastructure monitoring, dynamic system identification and modeling. He received his Ph.D. in Civil Engineering from the University of Southern California as well as an M.S. in Electrical Engineering, an M.S. from Rice University, and a Sc.B. and A.B. from Brown University. His email address is smyth@civil.columbia.edu.

KATHLEEN THOMSON is the Assistant Commissioner for the Bureau of Management Analysis and Planning at the FDNY. Her email address is kat.thomson@fdny.nyc.gov.