Health Care Manage Sci (2007) 10:25-45
DOI 10.1007/s10729-006-9006-3

The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta

Nabil Channouf · Pierre L'Ecuyer · Armann Ingolfsson · Athanassios N. Avramidis

Received: 20 June 2006 / Accepted: 1 October 2006 / Published online: 28 November 2006

Abstract We develop and evaluate time-series models of call volume to the emergency medical service of a major Canadian city. Our objective is to offer simple and effective models that could be used for realistic simulation of the system and for forecasting daily and hourly call volumes. Notable features of the analyzed time series are: a positive trend, daily, weekly, and yearly seasonal cycles, special-day effects, and positive autocorrelation. We estimate models of daily volumes via two approaches: (1) autoregressive models of data obtained after eliminating trend, seasonality, and special-day effects; and (2) doubly-seasonal ARIMA models with special-day effects. We compare the estimated models in terms of goodness-of-fit and forecasting accuracy. We also consider two possibilities for the hourly model: (3) a multinomial distribution for the vector of number of calls in each hour conditional on the total volume of calls during the day and (4) fitting a time series to the data at the hourly level. For our data, (1) and (3) are superior.

Keywords Emergency medical service · Arrivals · Time series · Simulation · Forecasting

N. Channouf · P. L'Ecuyer · A. N. Avramidis
DIRO, Université de Montréal, Montréal, Canada

A. Ingolfsson (✉)
School of Business, University of Alberta,
Edmonton, Alberta T6G 2R6, Canada
e-mail: armann.ingolfsson@ualberta.ca

1 Introduction

Most cities in the developed world have organizations that provide *Emergency Medical Service* (EMS), consisting of pre-hospital medical care and transport to a medical facility. Demand for such services is increasing throughout the developed world, in large part because of the aging of the population. In the U.S., EMS funding decreased following conversion of direct federal funding to block grants to states [11, 37] that have, in many cases, been used for purposes other than EMS. Tighter budgets make efficient use of resources increasingly important. Reliable demand forecasts are crucial input to resource use planning, and the focus of this paper is on how to generate such forecasts.

Almost all demand to EMS systems arrives by phone, through calls to an emergency number (911 in North America). Calls that arrive to 911 are initially routed to EMS, fire, or police. Calls routed to EMS are then *evaluated*, which involves obtaining an address, determining the nature and importance of the incident, and possibly providing instructions to a bystander on the use of CPR or other first-aid procedures. *Dispatching* an ambulance to the call, the next step, is a separate function that can occur partly in parallel with call evaluation. The crew of the dispatched vehicle(s) then begins traveling toward the scene of the call, where they assess the situation, provide on-site medical care, and determine whether transport to a medical facility is necessary (this is the case roughly 75% of the time). Once at the medical facility, EMS staff remain with the patient until they have transferred responsibility for her or his care to a nurse or physician. The crew may then need to complete various forms before it becomes available to take new calls.

The resource requirements per EMS call are on the order of a few minutes for call evaluation and dispatch, and on the order of an hour for an ambulance and its crew. The latter component is growing in many locations because of increased waiting times in hospital emergency rooms [13, 14, 33, 34].

The primary performance measure for an EMS system is typically the fraction of calls reached within some time standard, from the instant the call was made. In North America, a typical target is to reach 90% of the most urgent calls within 9 min. Although universal standards are lacking [28], the *response time* is typically considered to begin when call evaluation begins and end when an ambulance reaches the call address. Secondary performance measures include waiting times on the phone before reaching a 911 operator; (for example, 90% in 10 s [29] or 95% in 5 s [12]), average call evaluation times, average dispatch times, and average time spent at hospital.

The main decisions that require medium-term call volume forecasts (a few days to a few weeks into the future) are scheduling decisions for call evaluators, dispatchers, and, most importantly, ambulances and their crews. Longer-term call volume forecasts are needed for strategic planning of system expansion or reorganization. Shorter-term (intra-day) forecasts could be used to inform decisions about when to call in extra resources.

Service level standards for EMS systems are imperfect proxies for the real goal of such systems, namely to save lives and prevent suffering (see [15] for a discussion of models that attempt to quantify such goals more explicitly). Meeting these service-level standards is expensive, so it is a problem of substantial economic and social interest to manage EMS systems efficiently. Generally speaking, efficiency involves balancing quality of service against system costs. An important input to this operational problem is the call volume. Uncertainty in future call volume complicates the process of determining levels of EMS staffing and equipment. It is therefore important to correctly model the stochastic nature of call volumes, and in particular, to make predictions of future call volumes, including uncertainty estimates (via prediction intervals).

Operations researchers have been developing planning models for EMS systems, as well as police and fire services, since the 1970s. Green and Kolesar [17] provide a recent perspective on the impact of this work. Swersey [35] surveys the academic literature on this topic and [16] provides an EMS-practitioner-oriented literature survey. EMS planning models include simulation models ([20] and [22] are recent examples), analytical queueing models (notably the hypercube queueing model, see Larson [24, 25]), and optimization models for location of facilities and units. All of these models require estimates of demand as input. Typically, planning models assume that demand follows a Poisson process—an assumption that is supported by both theoretical arguments [e.g., 19] and empirical evidence [e.g., 18, 39]. However, empirical studies of demand for both EMS and other services (notably Brown et al. [9]) indicate that the rate of the Poisson arrival process varies with time and may be random. The work we report in this paper is aimed at estimating the arrival rate during day-long or hour-long periods. We elaborate in Section 3 on how our estimates can be used to support simulation and analytical studies that assume a Poisson arrival process.

Goldberg [16] mentions that "the ability to predict demand is of paramount importance" but that this area has seen little systematic study. The work that has been done can be divided in two categories: (1) models of the spatial distribution of demand, as a function of demographic variables and (2) models of how demand evolves over time. In the first category, Kamenetsky et al. [23] surveyed the literature before 1982 and presented regression models to predict EMS demand as a function of population, employment, and two other demographic variables. Their models successfully explained most of the variation in demand ($R^2 = 0.92$) among 200 spatial units in southwestern Pennsylvania. McConnell and Wilson [27] is a more recent article from this category which focuses on the increasingly important impact of the age distribution in a community on EMS demand. We refer the reader to Kamenetsky et al. [23] and McConnell and Wilson [27] for further relevant references.

This paper falls in the second category, of modeling and forecasting EMS demand over time. EMS demand varies strongly by time of day and day of week, for example see Zhu et al. [39] and Gunes and Szechtman [18]. Past related work that attempts to forecast daily EMS demand includes Mabert [26], who analyzed emergency call arrivals to the Indianapolis Police Department. He considered several simple methods based on de-seasonalized data and found that one of them outperforms a simple ARIMA model [7]. In a similar vein, Baker and Fitzpatrick [4] used Winter's exponential smoothing models to separately forecast the daily volume of emergency and "routine" EMS calls and used goal programming to choose the exponential smoothing parameters.

Recent work on forecasting arrivals to call centers from a variety of industries is also relevant. For the prediction of daily call volumes to a retailer's call center, Andrews and Cunningham [3] incorporate advertising

effects in an ARIMA model with transfer functions; their covariates are indicator variables of certain special days and catalog mailing days. Bianchi et al. [6] use an ARIMA model for forecasting daily arrivals at a telemarketing center, compare against the Holt–Winters model, and show the benefits of outlier elimination. Tych et al. [36] forecast hourly arrivals in a retail bank call center via a relatively complex model with unobserved components named "dynamic harmonic regression" and show that it outperforms seasonal ARIMA models; one unusual feature of their methodology is that estimation is done in the frequency domain. Brown et al. [9] develop methods for the prediction of the arrival rates over short intervals in a day, notably via linear regression on previous day's call volume.

In this paper, we study models of daily and hourly EMS call volumes and we demonstrate their application using historical observations from Calgary, Alberta. Although we focus on the Calgary data, we expect the models could be used to model EMS demand in other cities as well and we will comment on likely similarities and differences between cities.

We have 50 months (from 2000 to 2004) of data from the Calgary EMS system. Preliminary analysis reveals a positive trend, seasonality at the daily, weekly, and yearly cycle, special-day effects, and autocorrelation. In view of this, we consider two main approaches: (1) autoregressive models of the residual error of a model with trend, seasonality, and special-day effects; and (2) doubly-seasonal ARIMA models for the residuals of a model that captures only special-day effects. Within approach (1), we explore models whose effects are the day-of-week and month-of-year. We also consider a model with cross effects (interaction terms) and a more parsimonious model, also with cross effects, but where only the statistically significant effects are retained. The latter turns out to be the best performer in terms of both goodness-of-fit and forecasting accuracy. All the models are estimated with the first 36 months of data and the forecasting error is measured with the data from the last 14 months. We used the `R` and `SAS` statistical software for the analysis.

The remainder of the paper is organized as follows. Section 2 provides descriptive and preliminary data analysis. In Section 3 we present the different models of daily arrivals and compare them in terms of quality of fit (in-sample) and forecast accuracy (out-of-sample). In Section 4, we address the problem of predicting hourly call volumes. Section 5 offers conclusions.

2 Preliminary data analysis

We have data from January 1, 2000 to March 16, 2004, containing the time of occurrence of each ambulance call, the assessed call priority, and the geographical zone where the call originated. We work with the number of calls in each hour instead of their times of occurrence, to facilitate the application of time series models. We explain in the next section how such hourly counts can be related to a stochastic model of the times of individual arrivals. The average number of arrivals is about 174/day, or about 7/h.

Figure 1 provides a first view of the data; it shows the daily volume for year 2000. The figure suggests a positive trend, larger volume in July and December, and shows some unusually large values, e.g., on January 1, July 8, November 11, December 1; and low values, e.g., on January 26, September 9. Figure 2 shows monthly volume over the entire period. This plot reveals a clear

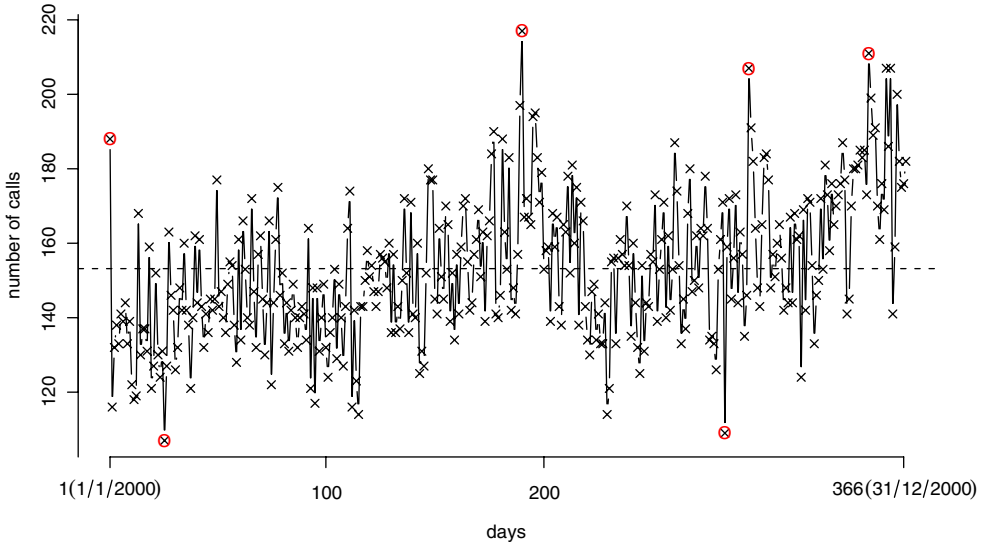

Fig. 1 The daily call volume for the year 2000. Some outliers appears in the plot

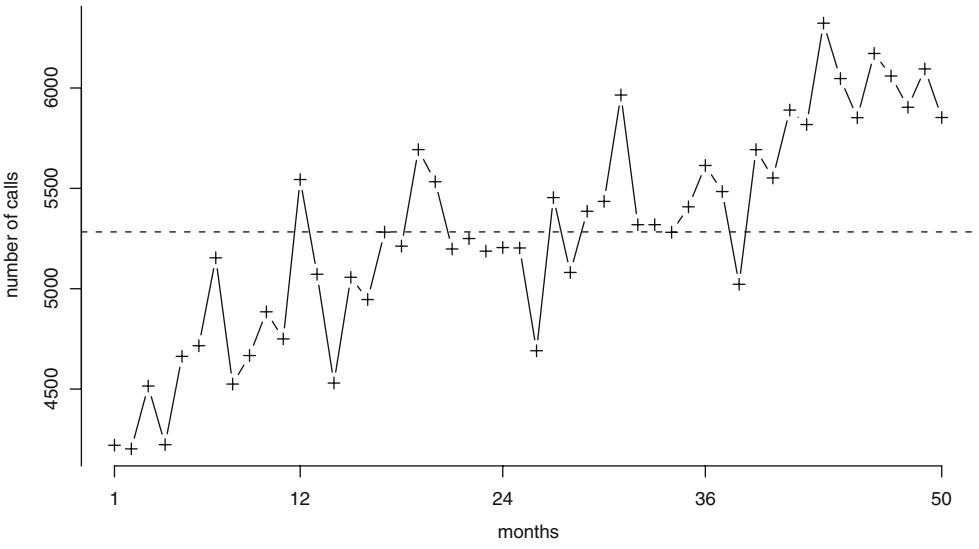

Fig. 2 The monthly call volume over the entire period

positive trend; the likely explanation is a combination of population growth and aging in the city. Figure 3 shows average volume by hour over the weekly cycle. The plot reveals a clear hour-of-day seasonality: over a 24-h cycle, higher call volumes are usually observed between 10 A.M. and 8 P.M.; substantially lower volumes are seen overnight. One also observes day-of-week effects. Closer inspection reveals, not surprisingly, increased activity during Friday and Saturday evening and early night. With respect to daily volume, larger values are observed over Friday and Saturday relative to the other days of the week. These observations would have to be taken into account when designing shift schedules for the ambulance crews.

Figures 4 and 5 give box-plots of the daily volume for each day of the week and monthly volume for each month of the year, respectively. Each box plot gives the median, the first and third quartiles (the bottom and top of the central box), the interquartile range (the height of the box), and two bars located at a distance of 1.5 times the interquartile range below the first quartile and above the fourth quartile, respectively. The small circles denote the individual observations that fall above or below these two bars. We see again that Friday and Saturday have more volume than the average. July, December, and November are the busiest months (in this order) while April is the most quiet month.

3 Models for daily arrivals

We now consider five different time-series models for the arrival volumes over successive days. Although in the end we conclude that one of these models fits

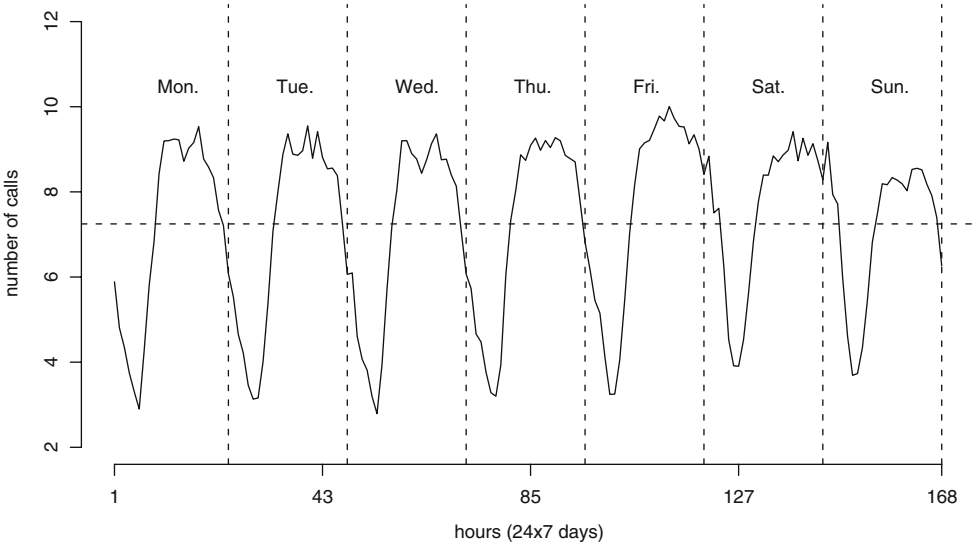

Fig. 3 The average hourly call volume over the weekly cycle

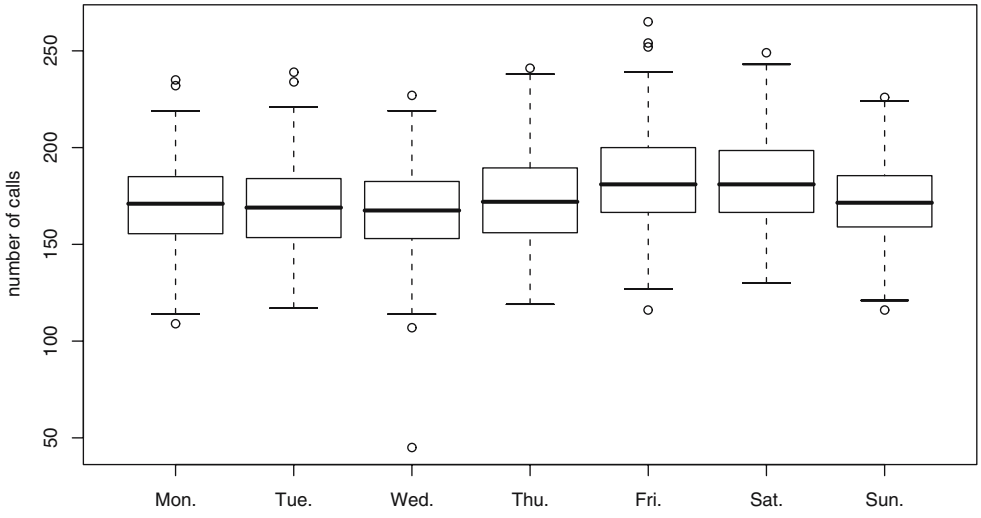

Fig. 4 Box plots of arrival volumes per day for each day of the week

the Calgary data best, we discuss all of them because different models from the collection that we present may be appropriate depending on the city being studied and the purpose of the analysis. These models are defined and studied in Sections 3.1 to 3.5. In Section 3.6, we compare these models in terms of both quality of fit (in-sample) and forecast accuracy (out-of-sample). Throughout the paper, t denotes the time index in days and the number of arrivals on day t is denoted Y_t, for $t = 1, 2, \ldots, n$, where $n = 1{,}537$. The models are fitted to the first 1,096 observations (January 1, 2000 through December 31, 2002), and the remaining 441 observations are used for prediction (January 1, 2003 through March 16, 2004).

There are compelling theoretical reasons to assume that call arrivals follow a nonhomogeneous Poisson process (NHPP). The *Palm–Khintchine* theorem [e.g., 10] states, approximately, that the superposition of arrival processes from many small and independent "sources" (patients, in an EMS context) is well-approximated by a Poisson process. The rate of this process will vary with time (because medical emergencies are more likely to occur at certain times) and the rate may not be known with certainty (because it may be influenced by factors other than time).

For purposes of illustration, suppose that arrivals during hour h follow a Poisson process with a random rate that remains constant during the hour. Conditional on the number of calls during the hour, call it Z_h, the arrival times of individual calls within the hour are independently and uniformly distributed between 0 and 1. This is the "order statistic property" for a Poisson process and it holds regardless of whether the arrival rate is deterministic or random [see 32, Sections 4.5–4.6]. Our models in this and the next section quantify the distribution of the daily arrival counts Y_t and the

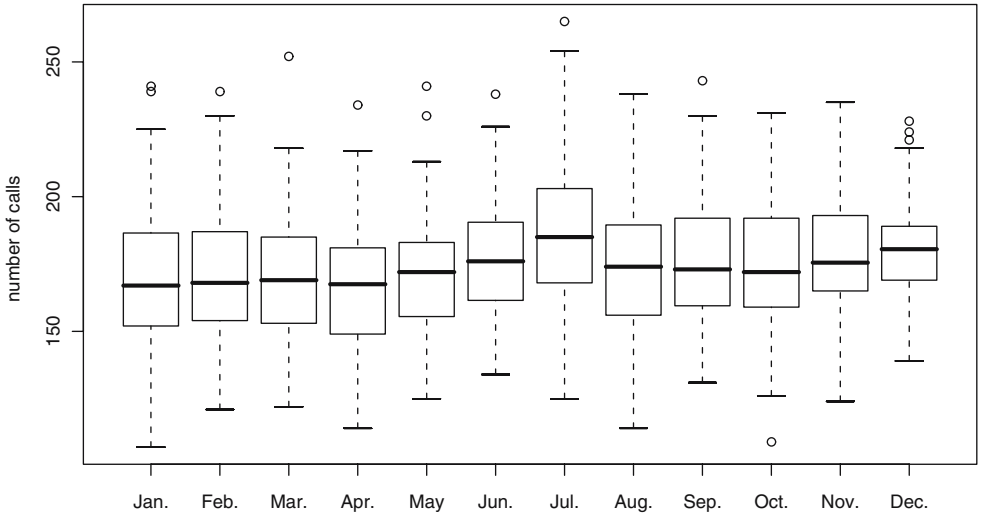

Fig. 5 Box plots of mean daily arrival volumes per month

hourly counts Z_h. One can use the following procedure to simulate call arrival times on day t:

1. Simulate the daily count Y_t. As we will see in this section, this involves simulating the residual from a standard autoregressive process.
2. Given Y_t, generate the vector $\mathbf{Z}_t$ of hourly counts on day t. As we will see in the next section, this involves simulating a multinomial random vector.
3. Use the order statistic property to distribute the simulated number of arrivals in each hour.

If the arrival rate varies too rapidly to be approximated as constant over hour-long periods, then it is straightforward to modify our models to use shorter periods, for example half-hours. Thus, if one limits attention to this general and plausible NHPP model, then each of our models of arrival counts by period yield corresponding stochastic models of all the arrival times, which can support analytical and simulation studies.

3.1 Model 1: fixed-effect model with independent residuals

One would expect to see month-of-year and day-of-week effects in EMS demand in most cities. Our preliminary analysis of the Calgary data indicates a positive trend and confirms the presence of month-of-year and day-of-week effects. This suggests the following linear model as a first approximation:

$$Y_{j,k,l} = a + \tilde{\beta}_j + \tilde{\gamma}_k + \tilde{\alpha}_l + \tilde{\epsilon}_{j,k,l}, \tag{1}$$

where $Y_{j,k,l}$ is the number of calls on a day of type j in month k of year l, the parameters a, $\tilde{\beta}_j$, $\tilde{\gamma}_k$, and $\tilde{\alpha}_l$, are real-valued constants, and the residuals $\epsilon_{j,k,l}$ are independent and identically distributed (i.i.d.) normal random variables with mean 0. The preliminary analysis suggests that for Calgary, the yearly effect is approximately a linear function of l, which allows us to express the model more conveniently as

$$Y_t = a + bt + \sum_{j=1}^{7} \beta_j C_{t,j} + \sum_{k=1}^{12} \gamma_k S_{t,k} + E_t, \tag{2}$$

where a, b, the β_j, and the γ_k are constants, the indicator $C_{t,j}$ is 1 if observation t is on the jth day of the week and 0 otherwise, the indicator $S_{t,k}$ is 1 if observation t is in the kth month of the year and 0 otherwise. In other cities, it might be more appropriate to model the yearly effect as a nonlinear function of t. We assume that the residuals E_t are i.i.d. normal with mean 0 and variance $\sigma^2_{\mathrm{E},0}$, i.e., a Gaussian white noise process. Given the presence of the constant parameter a, we impose the standard identifiability constraints:

$$\sum_{j=1}^{7} \beta_j = \sum_{k=1}^{12} \gamma_k = 0. \tag{3}$$

(Without these constraints, there would be redundant parameters; for example, adding a constant κ to all the β_j's and subtracting κ from a would give the same model.) We estimated the parameters for the regression model (2) using least squares and obtained the residuals displayed in Fig. 6, in which the circled points are at a distance larger than $3\hat{\sigma}_{\mathrm{E},0}$ from zero, where $\hat{\sigma}^2_{\mathrm{E},0}$ is the empirical variance of the residuals. There is a single residual larger than $4\hat{\sigma}_{\mathrm{E},0}$, which corresponds to January 1, 2002, and seven other residuals larger than $3\hat{\sigma}_{\mathrm{E},0}$: December 1, 2000; January 1, 2001; May 27, 2001; August 2, 2001; September 8, 2001; June 27, 2002; July 12, 2002. The single residual smaller than $-3\hat{\sigma}_{\mathrm{E},0}$ is on July 30, 2001. January 1 appears to be

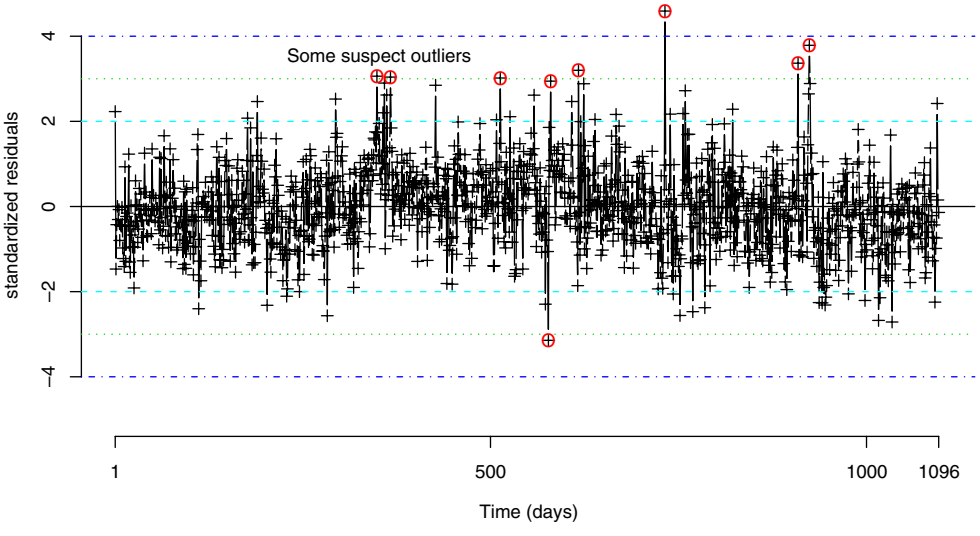

Fig. 6 The residuals E_t for the simple linear model of Eq. 2

a special day, with a call volume systematically larger than average. The month of July also has a larger volume per day than the other months (in the data). One potential explanation that we decided to consider is the Calgary Stampede, held every year in July. The Stampede includes one of the largest rodeos in the world and it is the most important annual festival in Calgary (http://calgarystampede.com). The dates for this event are: July 7–16, 2000; July 6–15, 2001; July 5–14, 2002; and July 4–13, 2003. To account for those two types of special days, we add two indicator variables $H_{t,1}$ and $H_{t,2}$ to our model, where $H_{t,1}$ is 1 if observation t is on January 1 and 0 otherwise, whereas $H_{t,2}$ is 1 if observation t is on one of the 40 Stampede days enumerated above, and 0 otherwise. This gives the model

$$Y_t = a + bt + \sum_{j=1}^{7} \beta_j C_{t,j} + \sum_{k=1}^{12} \gamma_k S_{t,k} + \omega_1 H_{t,1} + \omega_2 H_{t,2} + E_t, \quad (4)$$

in which we now have two additional real-valued parameters ω_1 and ω_2, and the residuals now have variance σ_E^2. The timing, nature, and number of such special events will vary between cities but the same general approach can be used if the dates of the special events are known. We estimate all the parameters of this linear regression model by standard least-squares, using the first $n = 1,096$ observations. If we denote the parameter estimates by $\hat{a}$, $\hat{b}$, $\hat{\beta}_j$, $\hat{\gamma}_k$, $\hat{\omega}_1$ and $\hat{\omega}_2$, then the estimates of Y_t and E_t are given by

$$\hat{Y}_t = \hat{a} + \hat{b}t + \sum_{j=1}^{7} \hat{\beta}_j C_{t,j} + \sum_{k=1}^{12} \hat{\gamma}_k S_{t,k} + \hat{\omega}_1 H_{t,1} + \hat{\omega}_2 H_{t,2} \quad (5)$$

and

$$\hat{E}_t = Y_t - \hat{Y}_t. \quad (6)$$

A naive estimator of σ_E^2 would be the empirical variance

$$\hat{\sigma}_E^2 = \frac{1}{n-s} \sum_{t=1}^{n} \hat{E}_t^2, \quad (7)$$

where $s = 21$ is the number of parameters estimated in the model. However, this variance estimator is biased if the residuals are correlated [5], and we will see in a moment that they are.

We must test the hypothesis that the residuals are a white-noise process, i.e., normally distributed and uncorrelated with zero mean and constant variance. Stationarity and normality of the residuals is plausible, based on Fig. 6 and on Q–Q (quantile–quantile) plots not shown here. To test for autocorrelation, we use the Ljung-Box test statistic, defined by

$$Q = n(n+2) \sum_{i=1}^{l} \frac{\hat{r}_i^2}{n-i},$$

where n is the number of residuals, $\hat{r}_i$ is the lag-i sample autocorrelation in the sequence of residuals, and l is the maximum lag up to which we want to test the autocorrelations. Under the null hypothesis that the residuals are uncorrelated and $n \gg s$, Q has approximately a chi-square distribution with l degrees of freedom. Here we have $n = 1,096$ and $s = 21$. We apply the test with $l = 30$ and obtain $Q = 154.8$. The corresponding p-value is smaller than 2.2×10^{-16}, so the null hypothesis is clearly rejected. This strong evidence of the presence of correlation between the residuals motivates our next model.

3.2 Model 2: an autoregressive process for the errors of Model 1

We improve Model 1 by fitting a time-series model to the residuals E_t. Since the E_t process appears to be normal and stationary, it suffices to capture the autocorrelation structure. We do this with an autoregressive process of order p (an AR(p) process), defined by

$$E_t = \phi_1 E_{t-1} + \cdots + \phi_p E_{t-p} + a_t, \quad (8)$$

where the a_t are i.i.d. normal with mean zero and variance σ_a^2. Based on the residuals defined by Eq. 6, and using standard tools of model identification [7, 38], we find that $p = 3$ is adequate (different values of p will be appropriate for different cities). When estimating the coefficients ϕ_l in a model with $p > 3$, we find that the coefficients ϕ_l for $l > 3$ are non-significant at the 5% level. For example, the p-value of the t-test for ϕ_4 is about 0.153.

The model obtained by combining Eqs. 4 and 8 with $p = 3$ can be written alternatively as

$$\phi(B)\left[Y_t - a - bt - \sum_{j=1}^{7} \beta_j C_{t,j} - \sum_{k=1}^{12} \gamma_k S_{t,k} - \omega_1 H_{t,1} - \omega_2 H_{t,2}\right] = a_t, \quad (9)$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3$, B is the backshift operator defined by $B^p E_t = E_{t-p}$, and ϕ_1, ϕ_2, ϕ_3 are the autoregressive parameters. We estimate the parameters $(a, b, \beta_1, ..., \beta_7, \gamma_1, ..., \gamma_{12}, \omega_1, \omega_2, \phi_1, \phi_2, \phi_3)$ by (nonlinear) least squares [1, page 67], based on the observations Y_t for $t = 4, ..., n$, where $n = 1096$.

Table 1 Parameter estimates for Model 2

Parameter	a	b	ω_1	ω_2			
	Intercept	Trend/month	Jan. 1	Stampede			
Estimate	**149.3**	**0.031**	**60.5**	2.7			
St. error	3.3	0.003	11.0	4.5			
p-val. of t-test	< 0.001	< 0.001	< 0.001	0.544			
Parameter	β_1	β_2	β_3	β_4	β_5	β_6	β_7
	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Estimate	**−4.2**	**−5.0**	**−5.3**	−0.9	**8.5**	**7.6**	−0.8
St. error	2.5	2.5	2.5	2.5	2.5	2.5	2.5
p-val. of t-test	0.095	0.046	0.034	0.719	0.001	0.002	0.747
Parameter	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.
Estimate	**−5.6**	−4.1	−2.5	−4.0	0.5	4.1	**13.2**
St. error	2.9	2.8	2.8	2.8	2.7	2.8	2.9
p-val. of t-test	0.048	0.152	0.358	0.149	0.849	0.138	< 0.001
Parameter	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}		
	Aug.	Sep.	Oct.	Nov.	Dec.		
Estimate	−1.3	−0.3	−4.3	0.3	4.1		
St. error	2.8	2.8	2.8	2.8	2.8		
p-val. of t-test	0.627	0.928	0.121	0.916	0.150		
Parameter	ϕ_1	ϕ_2	ϕ_3	σ_a^2			
Estimate	**0.192**	**0.108**	**0.083**	250.1			
St. error	0.030	0.031	0.030	–			
p-val. of t-test	< 0.001	< 0.001	0.006				

The parameter estimates are given in Table 1, together with their standard errors and the p-value of a t-test of the null hypothesis that the given parameter is zero, for each parameter. We then compute the residuals $\hat{a}_t = \hat{\phi}(B)(Y_t - \hat{Y}_t)$ in a similar manner as for Model 1 and we estimate σ_a^2 by

$$\hat{\sigma}_a^2 = \frac{1}{n-s}\sum_{t=4}^{n} \hat{a}_t^2, \tag{10}$$

where $n = 1{,}096$ and $s = 24$. This gives $\hat{\sigma}_a^2 = 250.1$. Figure 7 presents visual diagnostics for residual normality: we see the estimated residual density and a normal Q–Q plot, i.e., the empirical quantiles of normalized residuals plotted versus the corresponding quantiles of the standard normal distribution (with mean 0 and variance 1). Figure 8 is a diagnostic for (lack of) residual autocorrelation: it shows the standardized residuals, the sample autocorrelations up to lag 30, and the p-values of the Ljung-Box test for each lag. We conclude that the residuals a_t appear to be white noise. Thus, Model 2 is a much better fit than Model 1.

The most significant parameters in Table 1 are a (the mean), b (the positive trend), ω_1 (the positive January 1 effect), ϕ_1, ϕ_2, and ϕ_3 (the positive AR parameters), γ_7 (the positive July effect), and β_5 and β_6 (the positive Friday and Saturday effects). Other parameters significant at the 10% level are β_1 to β_3 (the negative effects of Monday–Wednesday) and γ_1 (the negative effect of January). (Since April has the lowest average in Fig. 5, one may find it surprising that January has significant negative coefficient and not April. But the average for January becomes smaller after removing the January 1 effect. Also, this estimation uses only the first 1,096 days of data, whereas Fig. 5 combines all 1,537 days). This gives a total of 13 significant parameters. We could eliminate the other ones; we will do that in Section 3.4. Observe that ω_2 (the Stampede effect) is not significant; most of the increased volume during the Stampede days is captured by the July effect. In fact, the average volume per day is about 186 during the Stampede days compared with 180 during the other days of July and 174 on average during the year.

We can also use this model to estimate the variance of the residuals E_t in Model 1. Their sample variance (7) *underestimates* $\sigma_E^2 = \mathrm{Var}[E_t]$ because they are positively correlated. If we multiply both sides of Eq. 8 by E_t and take the expectation, we get

$$\begin{aligned}\sigma_E^2 = \mathbb{E}[E_t^2] &= \phi_1\gamma_1 + \phi_2\gamma_2 + \phi_3\gamma_3 + \sigma_a^2 \\ &= (\phi_1\rho_1 + \phi_2\rho_2 + \phi_3\rho_3)\sigma_E^2 + \sigma_a^2,\end{aligned}$$

where $\gamma_i = \mathrm{Cov}(E_t, E_{t-i})$ and $\rho_i = \mathrm{Corr}(E_t, E_{t-i})$ for each i. Replacing all quantities in this last expression by their estimates and resolving for σ_E^2, we obtain $\hat{\sigma}_E^2 = 291.8$ as an estimate of σ_E^2. By comparing with the estimate $\sigma_a^2 = 250.1$, we see that Model 1 has about 17% more variance than Model 2.

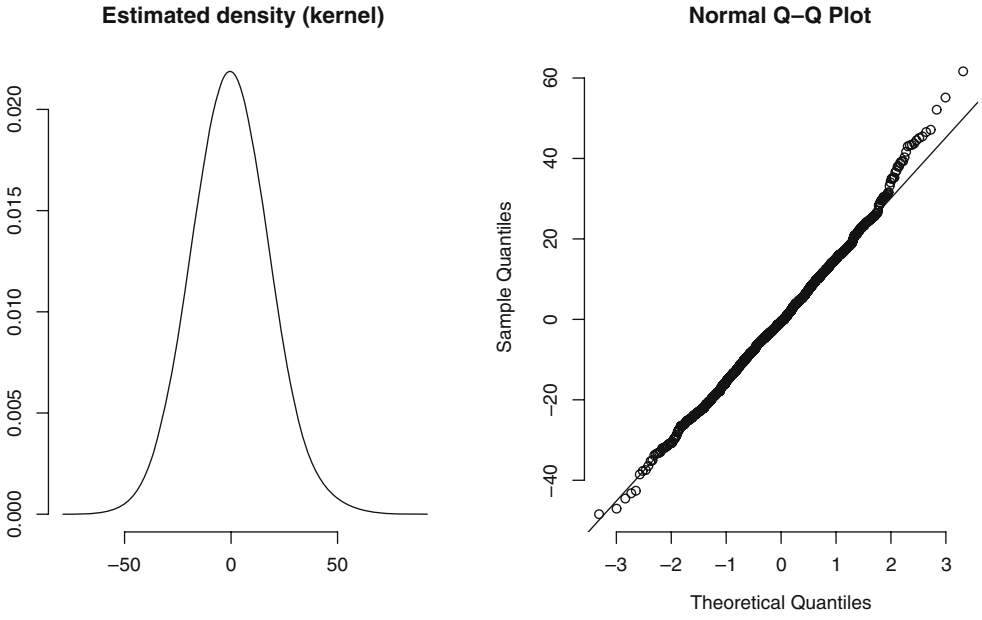

Fig. 7 Diagnostic for normality of residuals for Model 2

3.3 Model 3: adding cross effects

We now extend Model 2 by adding second-order terms to capture the interaction between the day-of-week and month-of-year factors. We simply add the term

$$\sum_{j=1}^{7}\sum_{k=1}^{12}\delta_{j,k}M_{t,j,k} \tag{11}$$

to the right side of Eq. 4 and subtract the same term inside the brackets in Eq. 9, where the indicator variable $M_{t,j,k}$ is 1 if observation t is on the jth day of the week and kth month of the year. This introduces the additional model parameters $\delta_{j,k}$, which must satisfy the identifiability constraints $\sum_{k=1}^{12}\delta_{j,k}=0$ for each j and $\sum_{j=1}^{7}\delta_{j,k}=0$ for each k.

We found that the estimates for the parameters β_j, γ_k, and ω_i in this model were almost the same as in Model 2. Table 2 gives the estimated values of the parameters that differ from those of Model 2, together with the p-value of a t-test that the given parameter is zero. The estimated variance of the residuals has

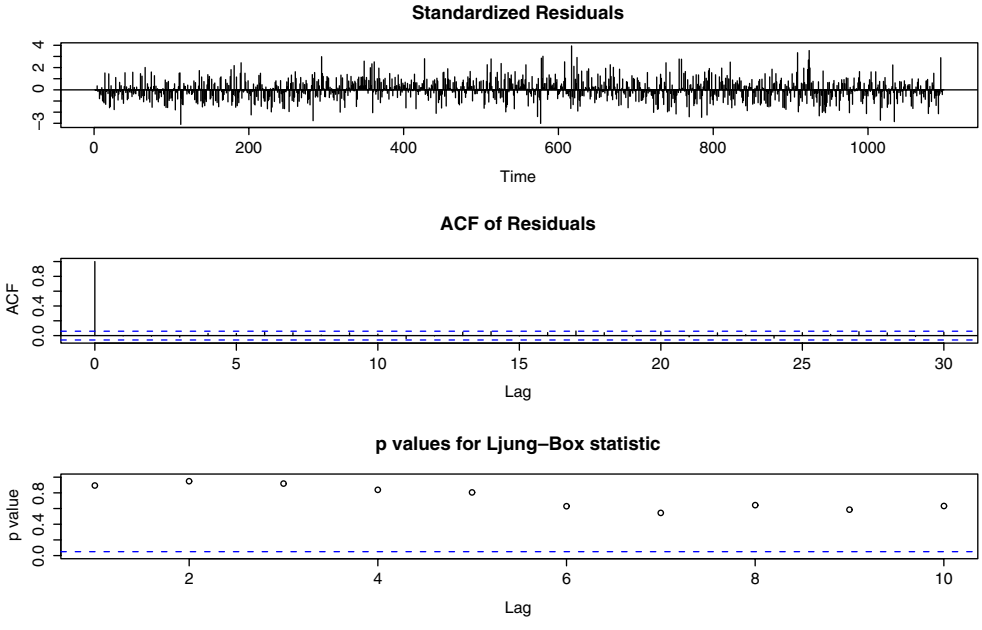

Fig. 8 Diagnostic for (lack of) residual autocorrelation for Model 2

Table 2 Parameter estimates for Model 3

	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.	
Parameter	$\delta_{1,1}$	$\delta_{2,1}$	$\delta_{3,1}$	$\delta_{4,1}$	$\delta_{5,1}$	$\delta_{6,1}$	$\delta_{7,1}$	Jan.
Est.	−0.3	5.0	4.6	−6.9	4.0	−6.3	−0.2	
St. error	3.9	4.0	4.0	4.2	4.2	4.2	4.2	
p-val. of *t*-test	0.942	0.210	0.256	0.102	0.339	0.137	0.965	
Parameter	$\delta_{1,2}$	$\delta_{2,2}$	$\delta_{3,2}$	$\delta_{4,2}$	$\delta_{5,2}$	$\delta_{6,2}$	$\delta_{7,2}$	Feb.
Est.	−5.4	5.6	3.7	−3.6	1.0	2.9	−4.2	
St. error	4.2	4.1	4.2	4.2	4.2	4.2	4.2	
p-val. of *t*-test	0.193	0.173	0.382	0.386	0.805	0.486	0.321	
Parameter	$\delta_{1,3}$	$\delta_{2,3}$	$\delta_{3,3}$	$\delta_{4,3}$	$\delta_{5,3}$	$\delta_{6,3}$	$\delta_{7,3}$	Mar.
Est.	**8.1**	6.2	−3.7	6.2	0.5	**−12.3**	−4.9	
St. error	4.1	4.2	4.1	3.9	3.9	4.2	4.2	
p-val. of *t*-test	0.052	0.137	0.364	0.113	0.891	0.003	0.240	
Parameter	$\delta_{1,4}$	$\delta_{2,4}$	$\delta_{3,4}$	$\delta_{4,4}$	$\delta_{5,4}$	$\delta_{6,4}$	$\delta_{7,4}$	Apr.
Est.	1.7	−3.7	0.4	4.2	−3.3	4.1	−3.4	
St. error	4.2	4.2	4.2	4.2	4.2	3.9	3.9	
p-val. of *t*-test	0.686	0.369	0.927	0.314	0.434	0.287	0.393	
Parameter	$\delta_{1,5}$	$\delta_{2,5}$	$\delta_{3,5}$	$\delta_{4,5}$	$\delta_{5,5}$	$\delta_{6,5}$	$\delta_{7,5}$	May
Est.	5.6	−0.8	−0.5	0.1	−6.6	−1.8	4.0	
St. error	3.8	3.9	3.9	4.2	4.2	4.2	4.2	
p-val. of *t*-test	0.143	0.840	0.908	0.986	0.113	0.663	0.345	
Parameter	$\delta_{1,6}$	$\delta_{2,6}$	$\delta_{3,6}$	$\delta_{4,6}$	$\delta_{5,6}$	$\delta_{6,6}$	$\delta_{7,6}$	Jun.
Est.	−2.6	−4.5	−1.3	**11.5**	**−6.7**	3.0	0.7	
St. error	4.1	4.1	4.2	3.8	3.9	4.2	4.2	
p-val. of *t*-test	0.525	0.276	0.747	0.003	0.083	0.468	0.875	
Parameter	$\delta_{1,7}$	$\delta_{2,7}$	$\delta_{3,7}$	$\delta_{4,7}$	$\delta_{5,7}$	$\delta_{6,7}$	$\delta_{7,7}$	Jul.
Est.	−3.9	−4.6	1.5	0.2	6.6	1.8	−1.6	
St. error	3.9	4.2	4.2	4.2	4.2	3.8	3.9	
p-val. of *t*-test	0.324	0.269	0.728	0.954	0.112	0.649	0.685	
Parameter	$\delta_{1,8}$	$\delta_{2,8}$	$\delta_{3,8}$	$\delta_{4,8}$	$\delta_{5,8}$	$\delta_{6,8}$	$\delta_{7,8}$	Aug.
Est.	6.1	4.7	−1.4	−0.8	−3.1	0.1	−5.4	
St. error	4.2	3.9	3.9	3.9	4.2	4.2	4.2	
p-val. of *t*-test	0.142	0.227	0.711	0.833	0.452	0.986	0.195	
Parameter	$\delta_{1,9}$	$\delta_{2,9}$	$\delta_{3,9}$	$\delta_{4,9}$	$\delta_{5,9}$	$\delta_{6,9}$	$\delta_{7,9}$	Sep.
Est.	−1.6	−5.3	−4.5	−2.3	2.9	4.9	5.9	
St. error	7.1	6.8	6.7	6.7	5.4	5.8	7.9	
p-val. of *t*-test	0.819	0.441	0.502	0.732	0.594	0.399	0.457	
Parameter	$\delta_{1,10}$	$\delta_{2,10}$	$\delta_{3,10}$	$\delta_{4,10}$	$\delta_{5,10}$	$\delta_{6,10}$	$\delta_{7,10}$	Oct.
Est.	0.9	−1.5	−3.7	−6.8	3.2	0.8	**7.1**	
St. error	4.0	4.0	4.2	4.2	4.2	4.2	3.9	
p-val. of *t*-test	0.822	0.702	0.371	0.105	0.436	0.839	0.073	
Parameter	$\delta_{1,11}$	$\delta_{2,11}$	$\delta_{3,11}$	$\delta_{4,11}$	$\delta_{5,11}$	$\delta_{6,11}$	$\delta_{7,11}$	Nov.
Est.	−1.2	0.1	**7.8**	−2.2	4.5	−5.4	−3.5	
St. error	4.3	3.9	3.9	3.9	4.1	4.1	4.6	
p-val. of *t*-test	0.775	0.982	0.044	0.578	0.280	0.187	0.444	
Parameter	$\delta_{1,12}$	$\delta_{2,12}$	$\delta_{3,12}$	$\delta_{4,12}$	$\delta_{5,12}$	$\delta_{6,12}$	$\delta_{7,12}$	Dec.
Est.	**−7.3**	−1.1	−2.7	0.4	−3.1	**8.2**	5.7	
St. error	4.3	4.3	4.2	4.2	3.9	3.9	4.0	
p-val. of *t*-test	0.092	0.797	0.518	0.924	0.417	0.036	0.154	
Parameter	ϕ_1	ϕ_2	ϕ_3	σ_a^2				
Est.	0.213	0.126	0.085	241.5				
St. error	0.030	0.031	0.031	–				
p-val. of *t*-test	< 0.001	< 0.001	0.006					

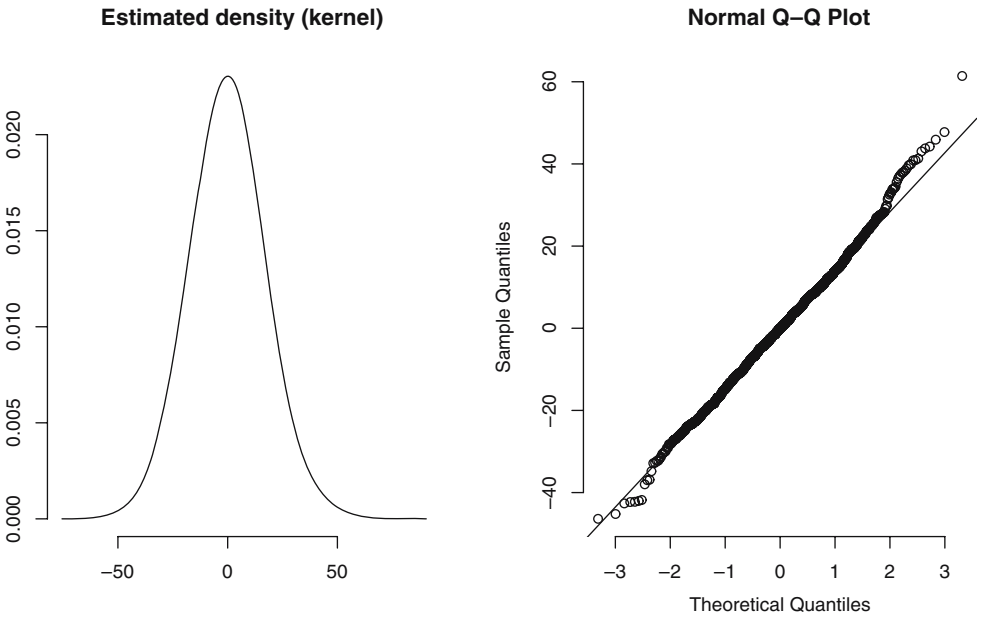

Fig. 9 Diagnostic for the normality of residuals, Model 3

been reduced to $\sigma_a^2 = 241.5$, about 4% less than for Model 2. The diagnostics for the residuals are in Fig. 9. The Ljung-Box test does not detect correlation in the residuals (we have $n = 1{,}096$, get $Q = 5.394$, and the p-value of the test is 0.944) (Fig. 10).

The slightly better fit of this model compared with Model 2 is obtained at the expense of a much larger number of parameters and several of these parameters do not appear to be significant. The next step is to remove them.

3.4 Model 4: considering only the significant parameters

This model is a stripped-down version of Model 3, in which we keep only the parameters that are significant at the 10% level (i.e., for which the p-value of the t-test in Table 1 or 2 is less than 0.10). In Table 2, eight parameters $\delta_{j,k}$ and three parameters ϕ_i are significant at the 90% level. There are ten other significant parameters in Table 1, for a total of 20. With the identifiability

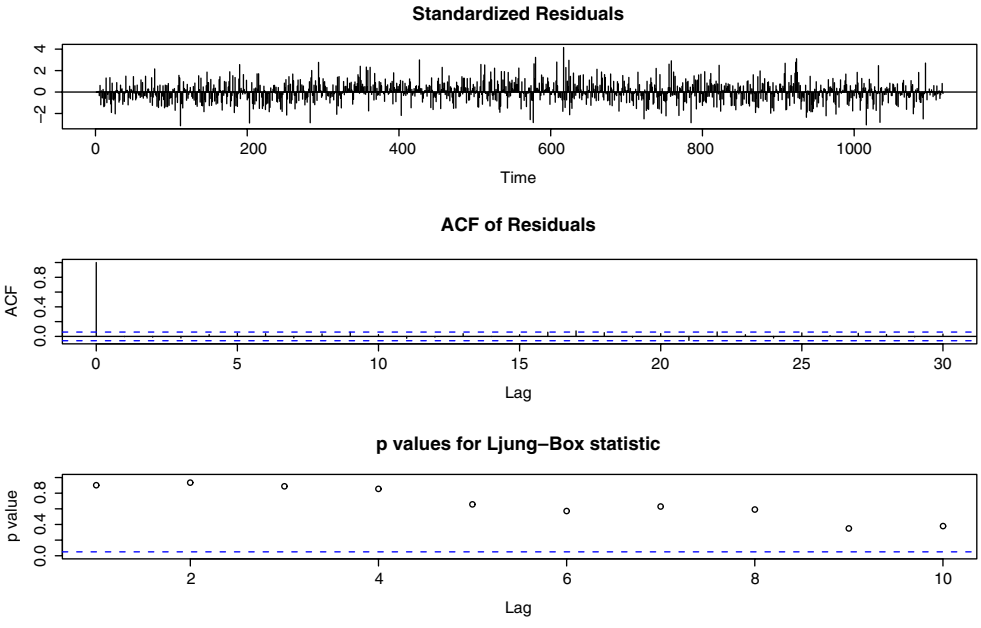

Fig. 10 Diagnostic for the correlation between residuals, Model 3

Table 3 Parameter estimates for Model 4

Parameter	a	b	ω_1					
	Intercept	Trend/month	Jan. 1					
Estimate	149.6	0.032	57.8					
St. error	1.9	0.003	10.7					
p-val. of t-test	< 0.001	< 0.001	< 0.001					
Parameter	β_1	β_2	β_3	β_5	β_6			
	Mon.	Tue.	Wed.	Fri.	Sat.			
Estimate	−4.3	−5.3	−6.2	7.8	7.9			
St. error	1.5	1.4	1.4	1.4	1.5			
p-val. of t-test	0.004	< 0.001	< 0.001	< 0.001	< 0.001			
Parameter	γ_1	γ_7						
	Jan.	Jul.						
Estimate	−5.7	12.4						
St. error	2.9	2.7						
p-val. of t-test	0.050	< 0.001						
Parameter	$\delta_{1,3}$	$\delta_{6,3}$	$\delta_{4,6}$	$\delta_{5,6}$	$\delta_{7,10}$	$\delta_{3,11}$	$\delta_{1,12}$	$\delta_{6,12}$
	Mon.	Sat.	Thu.	Fri.	Sun.	Wed.	Mon.	Sat.
	Mar.	Mar.	Jun.	Jun.	Oct.	Nov.	Dec.	Dec.
Estimate	7.9	−11.1	12.0	−7.1	8.3	8.5	−7.8	8.2
St. error	4.2	4.3	3.9	3.9	3.8	4.0	4.5	4.1
p-val. of t-test	0.060	0.010	0.002	0.069	0.029	0.034	0.083	0.046
Parameter	ϕ_1	ϕ_2	ϕ_3	σ_a^2				
Estimate	0.212	0.132	0.094	241.6				
St. error	0.030	0.031	0.031	–				
p-val. of t-test	< 0.001	< 0.001	0.002					

constraints, there remain $s = 15$ independent parameters out of those 20. The same strategy of including only the significant parameters from Model 3 could be used in other cities, but the set of significant parameters will vary between cities, of course.

We reestimate the model with those parameters only (all other parameters are set at zero) and obtain the values given in Table 3. All these parameters are significant. The most significant interaction parameters $\delta_{j,k}$ are for Saturday in March (negative interaction) and

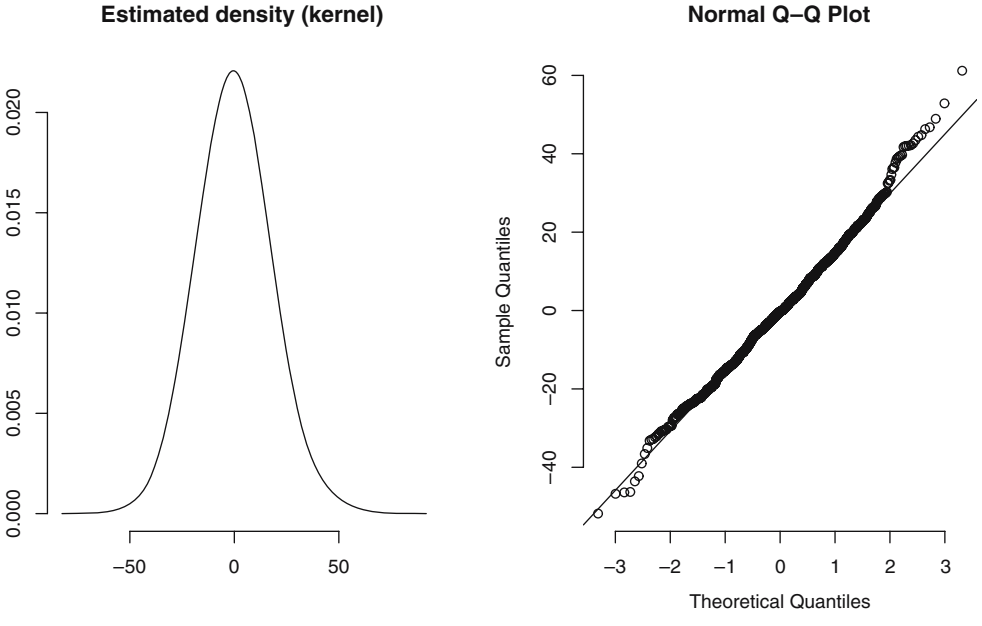

Fig. 11 Diagnostic for the normality of residuals, Model 4

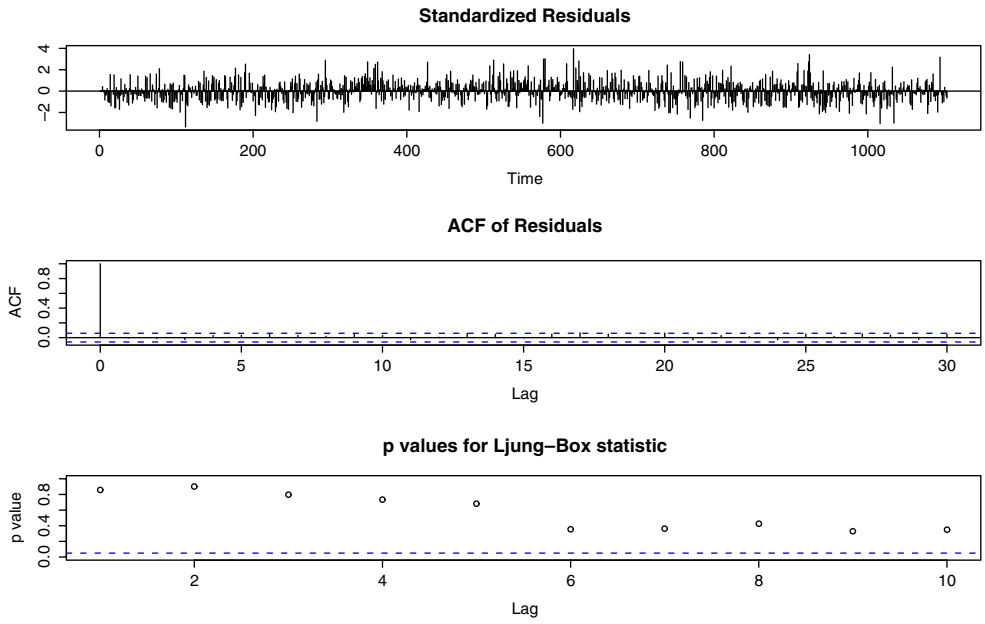

Fig. 12 Diagnostic for the correlation between residuals, Model 4

Thursday in June (positive interaction). Other mildly significant interactions, at the 10% level, are (by order of significance) Saturday in December, Wednesday in November, Monday in March, Sunday in October, Friday in June, and Monday in December.

The estimated variance of the residuals is $\sigma_a^2 = 241.9$. The diagnostics for the residuals are in Figs. 11 and 12. The Ljung-Box test does not detect correlation in the residuals (we have $n = 1{,}096$, get $Q = 11.581$, and the p-value of the test is 0.48).

3.5 Model 5: a doubly-seasonal ARIMA process

We now consider a different model: an ARIMA model with two seasonal cycles. We decompose our time series as

$$Y_t = N_t + \omega_1 H_{t,1} + \omega_2 H_{t,2}, \tag{12}$$

where $\{N_t\}$ is modeled as a doubly-seasonal ARIMA process and the other components capture the special days (January 1 and Stampede days). Given the seasonality patterns across the weekly and yearly cycle implied by the analysis in Section 2, we propose an ARIMA model with two seasonal cycles: a weekly cycle, with period $s_1 = 7$, and an approximate annual cycle, with period $s_2 = 365$. This choice of periodicities means that the conditional mean of N_t is regressed on N_{t-365}; for example, January 1, 2004 is regressed on January 1, 2003. In other words, after eliminating February 29 (which we did), this regression "aligns" the same dates across years.

The general form of a doubly-seasonal ARIMA model with periods s_1 and s_2 is [7, 8, 38]:

$$\begin{aligned}&\phi(B)\Phi_{s_1}(B^{s_1})\Phi_{s_2}(B^{s_2})\nabla^d\nabla_{s_1}^{d_1}\nabla_{s_2}^{d_2}N_t \\ &\quad = \theta(B)\Theta_{s_1}(B^{s_1})\Theta_{s_2}(B^{s_2})a_t,\end{aligned} \tag{13}$$

where $\nabla_s^d = (1 - B^s)^d$, ϕ, Φ_{s_1}, Φ_{s_2}, θ, Θ_{s_1}, and Θ_{s_2} are polynomial functions of order p, p_1, p_2, q, q_1, and q_2, respectively, and $\{a_t\}$ is a Gaussian white noise process. This model is referred to as an ARIMA$(p, d, q) \times (p_1, d_1, q_1)_{s_1} \times (p_2, d_2, q_2)_{s_2}$ process.

We follow a standard model-building protocol to identify the model (choice of the polynomial orders and exponents d, d_1, and d_2), estimate the parameters (ω_1, ω_2, and the polynomial coefficients), and perform diagnostic checks [7, 38]. ARIMA models with more than one seasonal cycle are difficult to estimate in general, because the multiple seasonalities complicate Eq. 13 with several operators, due to the multiplicative nature of the expressions involved. A concrete selection criterion must be adopted for model selection. Here, we used Akaike's information criterion (AIC), discussed in Section 3.6. We keep the model with minimum AIC, subject to non-rejection of the null hypothesis that model residuals are a white-noise process [7]. Based on this criterion, we identify the following model for N_t:

$$\begin{aligned}&(1 - \phi_7 B^7 - \phi_{14} B^{14} - \phi_{28} B^{28})(1 - \phi_{365} B^{365})(1 - B)N_t \\ &\quad = (1 - \theta_1 B)a_t.\end{aligned} \tag{14}$$

Table 4 Parameter estimates, Model 5

Parameter	ω_1 Jan. 1	ω_2 Stampede				
Estimate	43.8	16.6				
St. error	12.0	3.8				
p-val. of t-test	< 0.001	< 0.001				
Parameter	ϕ_7	ϕ_{14}	ϕ_{28}	ϕ_{365}	θ_1	σ_a^2
Estimate	0.064	0.103	0.082	0.128	0.905	251.7
St. error	0.03	0.03	0.03	0.04	0.01	–
p-val. of t-test	0.038	0.001	0.007	0.001	< 0.0001	

The parameters are estimated jointly via least squares based on Eqs. 12 and 14, i.e., we find the parameter values that minimize the sum of squares of the estimated residuals. The estimates are given in Table 4, together with their p-values. Note that Model 5 has considerably fewer parameters than the other models. It is also interesting to observe that for this model, the parameter ω_2 (Stampede days effect) is highly significant, in contrast with Models 2 to 4. The explanation is that there is no "July effect" term in the model.

3.6 Model comparison: Goodness of fit and forecast performance

In this section, we compare the five models in terms of their quality of fit and forecasting performance. The results are in Table 5.

With respect to quality of fit, we report the standard error of model residuals, $\hat{\sigma}_a$, the number s of parameters estimated, and Akaike's information criterion (AIC, see Akaike [2] and Wei [38, page 153]). The AIC has the advantage of taking into account both the mean-square error of the residuals and the number of estimated parameters in the model. It is designed to be an approximately unbiased estimator of the *Kullback–Leibler distance* (or *cross-entropy* or *relative entropy*) between the true model and fitted model. It is defined by

$$\mathrm{AIC}(s) = n \ln(\hat{\sigma}_a^2) + 2s, \tag{15}$$

where n is the number of observations, s is the number of estimated parameters in the model, and $\hat{\sigma}_a^2$ is the maximum likelihood estimator of the variance of residuals, which is approximately the same as the sample variance (10) under the assumption that the residuals are i.i.d. normal [30]. Bias-reduced variants known as the AICC are discussed, e.g., in [8, pages 301–304]. A model with minimal AIC is a good compromise between parsimony and small (empirical) variance of the residuals.

The models of Sections 3.1–3.5 were fitted to the first 1,096 days of data. We then used the estimated models to forecast for the remaining 441 days ($t = 1{,}097, \ldots, 1{,}537$), at *forecast lag* ranging from 1 day ahead to 21 days ahead. The lag-ℓ forecast error at day t is defined as

$$e_t(\ell) = Y_{t+\ell} - \hat{Y}_t(\ell),$$

where $\hat{Y}_t(\ell)$ is the forecast of $Y_{t+\ell}$ based on the information available on day t. Forecasts for doubly-seasonal ARIMA processes obey fairly complicated recursive formulas; see, for example, Brockwell and Davis

Table 5 Comparison of models for daily arrivals

	Model 1	Model 2	Model 3	Model 4	Model 5
$\hat{\sigma}_a^2$	291.8	250.1	241.5	241.6	251.7
St. error of fit $\hat{\sigma}_a$	17.08	15.81	15.54	15.55	15.87
s	21	24	90	15	7
Degrees of freedom	1075	1072	1006	1081	1088
AIC(s)		6099	6194	6045	6068
RMSE(1)	17.82	15.38	15.31	13.91	15.68
MRAE(1) (in %)	7.58	6.14	6.14	5.72	6.91

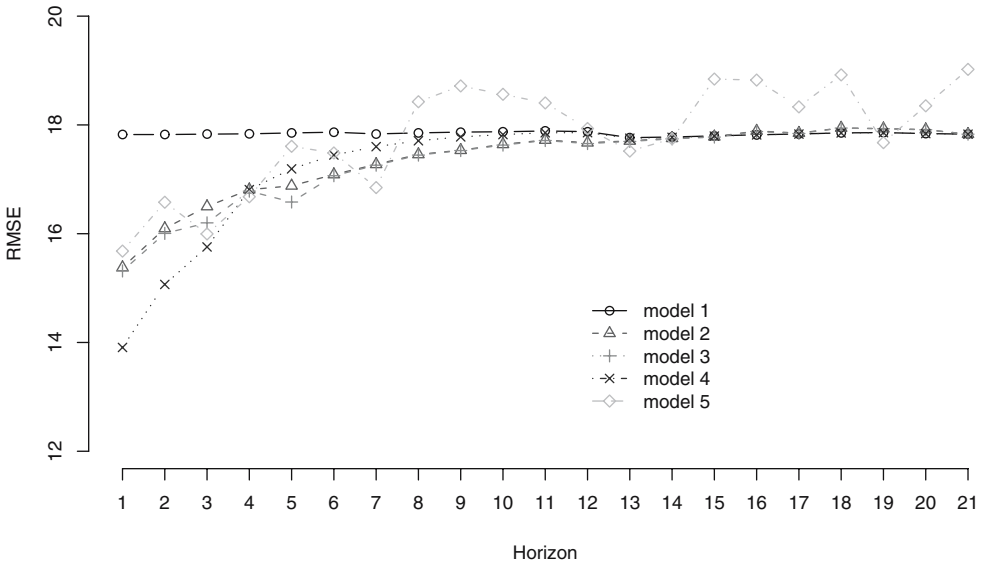

Fig. 13 Forecast RMSE(s) for Models 1–5, for forecast lags $\ell = 1, \ldots, 21$

[8, pages 175-182], but forecasting software facilitates their computation.

Commonly used forecast-accuracy metrics are the *Root Mean Square Error* (RMSE) and the *Mean Relative Absolute Error* (MRAE) at various forecast lags, defined in our case as

$$\text{RMSE}(\ell) = \sqrt{\frac{1}{442-\ell}\sum_{t=1097}^{1538-\ell} e_t^2(\ell)} \qquad \text{and}$$

$$\text{MRAE}(\ell) = \frac{1}{442-\ell}\sum_{t=1097}^{1538-\ell} \frac{|e_t(\ell)|}{Y_{t+\ell}}.$$

for lag ℓ. The MRAE standardizes each forecasting error term by the corresponding process value $Y_{t+\ell}$, to reflect the idea that larger numbers usually require less absolute accuracy; it must be used with caution because it may be inflated substantially by a few moderate absolute errors that correspond to very small values $|Y_{t+\ell}|$.

Table 5 summarizes the model evaluation. The upper part of the table collects information on the fit with the data used for the estimation (the first 36 months). It recalls the estimated variance of the residuals, $\hat{\sigma}_{\text{a}}^2$, then its square root $\hat{\sigma}_{\text{a}}$ called the *standard error of fit*, the number s of independent estimated parameters, the number $n - s$ of degrees of freedom, and the AIC(s) criterion (for Model 1, $\hat{\sigma}_{\text{a}}^2$ is replaced by $\hat{\sigma}_{\text{E}}^2$, defined at the end of Section 3.2). According to the AIC criterion, Model 4 is the winner, followed by Model 5 and then Model 2. It must be underlined, however, that Model 5 was selected by minimizing the AIC over a class of ARIMA models, so the AIC measure is biased to its advantage.

The second part of the table gives the RMSE and MRAE for the forecasts of lag 1. The RMSE for lags 1 to 21 are displayed in Fig. 13. For small lags, Model 4 is clearly the best model in terms of forecast accuracy, followed by Models 2 and 3. For lags $s \geq 13$ (approximately), RMSE(s) is about the same for all models except Model 5, whose forecasts are much noisier. Encouragingly, the AIC measure at the estimation stage has successfully identified the best model.

In interpreting the standard error of fit and the RMSE, it is helpful to recall from our preliminary data analysis that the average number of calls per day was about 174. If calls were generated by a stationary Poisson process with a rate of 174/day, then the standard deviation of the number of calls per day would be $\sqrt{174} = 13.2$. The Model 4 RMSE with a lag of 1 comes close to this value. This suggests that, given knowledge of call volumes up to a certain point in time, the Poisson arrival rate for the next 24 h is almost deterministic. The RMSE for longer lags is higher, suggesting that when modeling arrivals more than one day into the future, one should view them as being generated by a Poisson process with a random arrival rate. The discussion at the beginning of this Section outlines how one can quantify the distribution for the arrival rate.

4 Modeling hourly arrivals

Now that we have a good model of day-by-day call volumes, we turn to the modeling of hour-by-hour call volumes. We will denote the number of calls during hour h by Z_h, where $h = 1, \ldots, 24n$ and $n = 1{,}537$. We investigate two modeling and forecasting approaches.

Both build on a model for the daily call volume Y_t and add to that model a component that divides the daily volume across the 24 h of the day. Our first approach is via the *conditional distribution* of the vector of number of calls in each hour, given the total daily call volume. The second approach fits a time-series model directly to the data at the hourly level.

4.1 Model 6: modeling the conditional distribution

Here we use Model 4 for the daily arrival volumes, then assume that on day t, the conditional distribution of the vector $\mathbf{Z}_t = (Z_{24(t-1)+1}, \ldots, Z_{24t})$, given Y_t, is independent of what happens on days other than t. A simple candidate for this conditional distribution is a multinomial distribution with parameters $(N, p_1, \ldots, p_{24})$, where $N = Y_t$. Each p_i represents the probability that a randomly selected call arriving during the day arrives in hour i. The vector $(p_1, \ldots, p_{24})$ is called the *daily profile*. Use of the multinomial distribution implies that the hours of occurrence of different calls on day t are independent, conditional on Y_t.

Figure 3 suggests that different days of the week should have different daily profiles; for instance, Fridays and Saturdays have a very different profile than the other days. Based on a more detailed analysis of our data, we regrouped the days of the week into four daily profile categories: (1) Monday–Wednesday, (2) Thursday and Sunday, (3) Friday, and (4) Saturday. Each category c has a different daily profile vector $(p_{c,1}, \ldots, p_{c,24})$ for category c, for $c = 1, 2, 3, 4$. The probability $p_{c,i}$ is estimated as the fraction of calls in category c that occur in hour i, i.e.,

$$\hat{p}_{c,i} = \frac{\sum_{t=1}^{n} Z_{24(t-1)+i} P_{t,c}}{\sum_{t=1}^{n} \sum_{h=1}^{24} Z_{24(t-1)+h} P_{t,c}}, \tag{16}$$

for $i = 1, 2, \ldots, 24$, where the indicator variable $P_{t,c}$ is 1 if day t is in category c and 0 otherwise. A positive aspect of this model is that the model for Y_t remains exactly the same as before. We could also use other distributions than the multinomial for the conditional distribution of $\mathbf{Z}_t$.

One way of testing the goodness-of-fit of this model is as follows. Under the multinomial assumption, conditional on Y_t and if day t is in category c, the chi-square statistic

$$Q_t = \sum_{i=1}^{24} \frac{(Z_{24(t-1)+i} - Y_t p_{c,i})^2}{Y_t p_{c,i}}$$

should have approximately the chi-square distribution with 23 degrees of freedom if $Y_t p_{c,i}$ is large enough (e.g., larger than 5) for all i [31]. So we could compute these Q_t's for all days and compare their empirical distribution to the chi-square distribution. But it turns out that $Y_t p_{c,i}$ is often rather small (less than 5) for the night hours. For this reason, before applying the test we regrouped four early morning hours, from 3:00 A.M. to 7:00 A.M., in a single period. All other hours count for one period each. This gives $m = 21$ periods and the expected number of calls in each period is at least five under the multinomial model. The probability that a call is in period i on a day of category c is then

$$\tilde{p}_{c,i} = \begin{cases} p_{c,i} & \text{for } i = 1, 2, 3, \\ p_{c,4} + p_{c,5} + p_{c,6} + p_{c,7} & \text{for } i = 4, \\ p_{c,i+3} & \text{for } i = 5, \ldots, 21. \end{cases} \tag{17}$$

Let $\tilde{Z}_{t,i}$ be the number of calls in period i of day t. We have $\sum_{i=1}^{m} \tilde{Z}_{t,i} = Y_t$. For a day of category c, conditional on Y_t, our multinomial (null) hypothesis now states that $\tilde{\mathbf{Z}}_t = (\tilde{Z}_{t,1}, \ldots, \tilde{Z}_{t,m})$ has the multinomial distribution with parameters $(Y_t, \tilde{p}_{c,1}, \ldots, \tilde{p}_{c,m})$. To estimate the parameters $\tilde{p}_{c,i}$, we use the consistent estimators $\hat{\tilde{p}}_{c,i}$ obtained simply by summing the $\hat{p}_{c,i}$'s of category 4 appropriately and using the correct hour-to-period correspondence as in Eq. 17. Then, for large enough n, the Pearson test statistic

$$Q_{m-1,t} = \sum_{i=1}^{m} \frac{(\tilde{Z}_{t,i} - Y_t \hat{\tilde{p}}_{c,i})^2}{Y_t \hat{\tilde{p}}_{c,i}}, \tag{18}$$

should have approximately the chi-square distribution with $m - 1 = 20$ degrees of freedom under the multinomial model.

We computed the 1,537 values of $Q_{m-1,t}$ for our data, and compared their empirical distribution (distribution A) to the chi-square distribution (distribution C) via a Q–Q plot. Having found a bit of discrepancy in the right tail, we thought that perhaps the chi-square distribution is not a good enough approximation of the exact distribution of $Q_{m-1,t}$ under the multinomial model, so we also generated (by simulation) a times series of 1,537 successive realizations of Y_t under Model 4, then a sample of vectors $\tilde{\mathbf{Z}}_t$ conditional on Y_t under the multinomial distribution hypothesis, and computed the corresponding values of $Q_{m-1,t}$, for $t = 1, ..., 1{,}537$. The empirical distribution of this sample is called distribution B. Figure 14 shows a Q–Q plot of distribution A against distribution B. The fit is excellent except in the right tail. The observations in the right tail correspond to days where the observed daily profile differed significantly from the usual daily profile for that type of day. This could be due to unusual (perhaps unpredictable) events that happened on those days. This has occurred on Sunday October 29, 2000, Monday March 19, 2001, Friday April 13, 2001, Monday September 30, 2001,

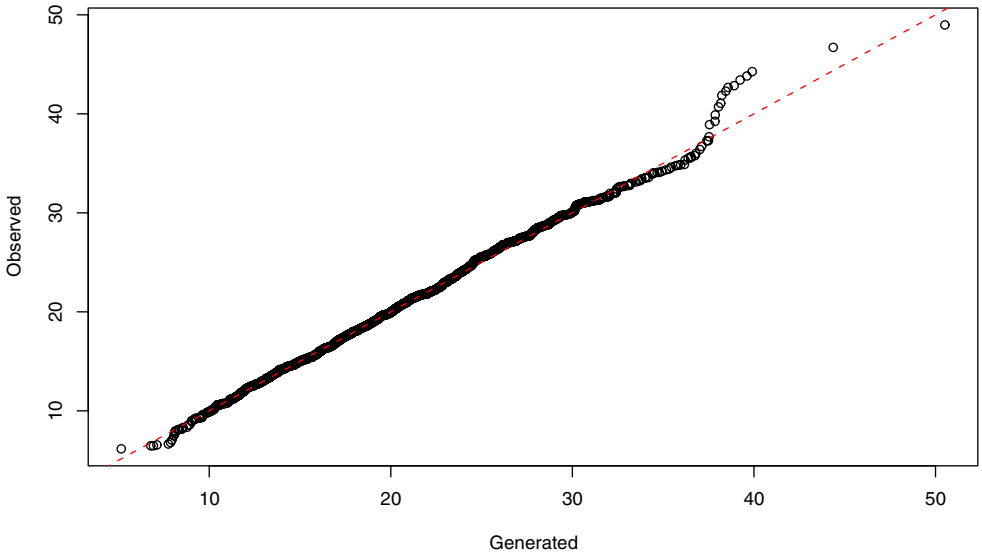

Fig. 14 Q–Q plot of the empirical distribution A against the empirical distribution B of a sample of $Q_{m-1,t}$ generated from Model 6 under the multinomial assumption

Sunday October 28, 2001, Saturday October 28, 2002, Saturday May 21, 2003, and on January 1 of each year. For January 1, the different profile could be predicted because it happens every year. We conclude that even though the fit is not perfect in the right tail, it is generally good enough to justify the use of the multinomial model in practice, in particular for purposes of forecasting.

We now turn to forecasting hourly volumes with this model. If h is the ith hour of day t, a forecast of Z_h made ℓ days before day t (at the end of day $t-\ell$, so $\ell = 1$ means at the beginning of day t) is simply $\hat{p}_{c,i}\hat{Y}_{t-\ell}(\ell)$, where $\hat{Y}_{t-\ell}(\ell)$ is the forecast of Y_t as defined in Section 3.6 and c is the category for day t.

To forecast Z_h on the day that hour h occurs, we could use $\hat{p}_{c,i}\hat{Y}_{t-1}(1)$, but we should be able to do better by taking into account the information we have in addition to the call volume of the previous days, i.e., the call volume on day t, up to hour $h-1$. For example, suppose that at 11:00 A.M. we want call volume forecasts for each of the next 13 h. We assume we already know the number of calls during the first 11 h of the day. If $W_{t,11}$ is that number, then a naive idea would be to estimate the remaining call volume on day t as $\hat{Y}_{t-1}(1) - W_{t,11}$ and then use

$$\frac{\hat{p}_{c,i}(\hat{Y}_{t-1}(1) - W_{t,11})}{\hat{p}_{c,12} + \cdots + \hat{p}_{c,24}} \tag{19}$$

as a forecast for the ith hour, for $i > 11$. This is a bad idea because a larger $W_{t,11}$ results in a *smaller* forecast for the rest of the day, suggesting a negative correlation between the volumes over the different hours of the day. In reality, the correlation is typically positive.

Let $W_{t,i} = Z_{24(t-1)+1} + \cdots + Z_{24(t-1)+i}$ be the number of calls during the first i hours of day t. Under our model, $\tilde{Y}_t = \hat{Y}_{t-1}(1)$ is a sufficient statistic for the information from previous days. Using Bayes' formula, the conditional distribution of Y_t given $\tilde{Y}_t$ and $W_{t,i}$ is

$$\begin{aligned}\mathbb{P}\big[Y_t = y \mid \tilde{Y}_t,\ W_{t,i} = w\big] &= \frac{\mathbb{P}\big[W_{t,i} = w \mid Y_t = y\big]\ \mathbb{P}\Big[Y_t = y \mid \tilde{Y}_t\Big]}{\mathbb{P}\Big[W_{t,i} = w \mid \tilde{Y}_t\Big]} \\ &= \frac{\mathbb{P}\big[W_{t,i} = w \mid Y_t = y\big]\ \mathbb{P}\big[Y_t = y\big]}{\mathbb{P}\Big[W_{t,i} = w \mid \tilde{Y}_t\Big]}\end{aligned} \tag{20}$$

for all integers $0 \le w \le y$. The distribution of $W_{t,i}$ conditional on $Y_t = y$ is binomial with parameters $(y,\ p_{c,1:i})$, where $p_{c,1:i} = \sum_{\ell=1}^{i} p_{c,\ell}$. Even though Y_t can only take integer values, Model 4 approximates its distribution conditional on $\tilde{Y}_t$ by a normal with mean $\tilde{Y}_t$ and variance $\sigma_Y^2 = \mathrm{Var}[\phi^{-1}(B)(a_t)]$. This could be used to write down a specific expression for the probabilities in Eq. 20 and computing them numerically. For Y_t, the probability of any integer value y can be approximated by integrating the normal density over the interval $[y - 1/2,\ y + 1/2]$. Note that conditional on $W_{t,i}$ and Y_t, the vector $(Z_{t,i+1}, \ldots, Z_{t,24})$ has a multinomial distribution with parameters $(Y_t - W_{t,i},\ p_{c,i+1}/(1 - p_{c,1:i}),\ \ldots,\ p_{c,24}/(1 - p_{c,1:i}))$.

For forecasting, we may only need the conditional expectation $\mathbb{E}[Y_t \mid \tilde{Y}_t,\ W_{t,i}]$ instead of the entire distribution (20). If we assume that the pair $(Y_t,\ W_{t,i})$ has approximately a bivariate normal distribution, which is close to the truth under our model when $p_{c,1:i}$ is not to

close to 0 or 1 and Y_t has a large enough expectation, then we have [21, page 93]:

$$\mathbb{E}\big[Y_t \mid \tilde{Y}_t, W_{t,i} = w\big] = \tilde{Y}_t + \frac{\mathrm{Cov}[Y_t, W_{t,i} \mid \tilde{Y}_t]}{\mathrm{Var}\,[W_{t,i} \mid \tilde{Y}_t]} \times (w - \mathbb{E}[W_{t,i} \mid \tilde{Y}_t]). \quad (21)$$

But $\mathbb{E}[W_{t,i} \mid \tilde{Y}_t] = p_{c,1:i}\tilde{Y}_t$, $\mathrm{Cov}[Y_t, W_{t,i} \mid \tilde{Y}_t] = p_{c,1:i}\sigma_a^2$, and

$$\begin{aligned}\mathrm{Var}[W_{t,i} \mid \tilde{Y}_t] &= \mathbb{E}_{Y_t}[\mathrm{Var}[W_{t,i} \mid \tilde{Y}_t,\ Y_t]] \\ &\quad + \mathrm{Var}_{Y_t}[\mathbb{E}[W_{t,i} \mid \tilde{Y}_t,\ Y_t]] \\ &= \tilde{Y}_t p_{c,1:i}(1 - p_{c,1:i}) + p_{c,1:i}^2\sigma_a^2.\end{aligned}$$

Combining this with Eq. 21, we obtain

$$\mathbb{E}[Y_t \mid \tilde{Y}_t,\ W_{t,i} = w] = \tilde{Y}_t + \frac{w - p_{c,1:i}\tilde{Y}_t}{(1 - p_{c,1:i})\tilde{Y}_t/\sigma_a^2 + p_{c,1:i}}. \quad (22)$$

We will see the results of applying this formula later in this section.

4.2 Model 7: an extension of Model 4 with an hour-of-day effect

We write this model as

$$Z_h = p_{c,i}Y_t + W_h$$

if h is the ith hour of day t and c is the category for day t, where the process Y_t obeys one of the previous day-to-day models and the process W_h is AR(q) for some q.

If Y_t obeys Model 4, for instance, then the variance σ_a^2 of the residuals in that model would have to be reduced, to account for the additional variance coming from the W_h's. This gives the following:

$$Z_h = p_{c,i}\left[\hat{Y}_t + E_t\right] + \sum_{l=1}^{24} \alpha_l D_{h,l} + W_h, \quad (23)$$

where $D_{h,l} = 1$ if h is the l-th hour of the day and 0 otherwise, and $\{E_t\}$ and $\{W_h\}$ are AR processes.

An important distinction between Models 6 and 7 is the following. With Model 6, there is positive correlation between the arrivals counts in different hours of the same day, regardless of the distance between those hours, and the only correlation between the hours of two successive days is due to the autocorrelations in the process E_t. With Model 7, on the other hand, the correlation between hours on the same day decreases with the distance between them and there is also an additional correlation between hours on two successive days but that are close in time (e.g., Friday evening and Saturday morning hours).

When we estimate the model, we add the two constraints $\sum_{l=1}^{24} \alpha_l = 0$ and $\sum_{h=24(t-1)+1}^{24t} W_h = 0$, for $t = 1, ..., n$, and we force $\{E_t\}$ to be an AR(3) process as in models for days. We estimated the process $\{W_h\}$. The largest (observed) lag for which the autoregressive parameter was significant at the 5% level is lag 44, but few autoregressive parameters for lags larger than 25 were significant at the 1% level. For this reason, we decided to retain an AR(25) model for W_h and an AR(3) model for E_t. We reestimated the model in Eq. 23 with these constraints. This is our Model 7. With it, we obtained a standard error of fit of 3.6. The residuals diagnostic is shown on Fig. 15.

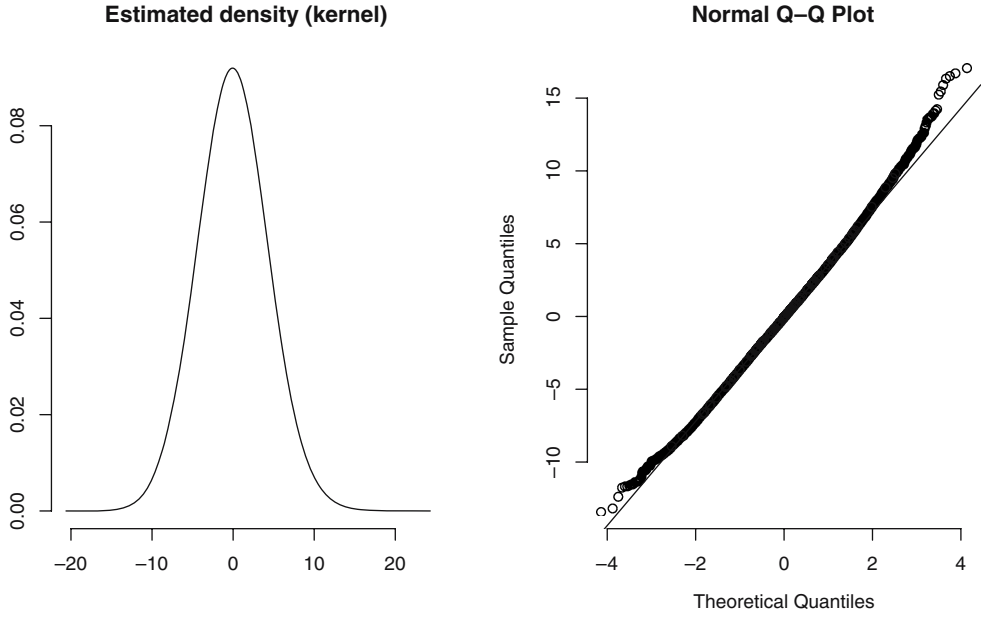

Fig. 15 Diagnostic for residual normality, hourly Model 7

4.3 Comparison of Models 6 and 7

We compare the forecasting performance of Models 6 and 7 by measuring the *root mean square error* (RMSE) at different lags in the 441 days for $t = 1{,}097, ..., 1{,}537$. For $h = 24(t-1)+1, ..., 24t$, we define the lag-r error at hour h by $e_h(r) = Z_{h+r} - \hat{Z}_h(r)$, where $\hat{Z}_h(r)$ is the forecast of Z_{h+r} based on the selected model. We consider two cases:

(1) For the forecasts of the 24 coming hours of day t when we are at the beginning of day t and have $\tilde{Y}_t$ as a forecast of Y_t from Model 4, we measure the error by

$$\mathrm{RMSE}(r) = \sqrt{\frac{1}{441}\sum_{t=1096}^{1536} e^2_{24(t-1)}(r)};$$

(2) For the forecasts of the hours $i = 12, \ldots, 24$ of day t after having observed the first 11 hours, using the formula (22) to update the forecast of Y_t for Model 6, we measure the error by

$$\mathrm{RMSE}(r) = \sqrt{\frac{1}{441}\sum_{t=1097}^{1537} e^2_{24(t-1)+11}(r)}.$$

Table 6 gives a representative subset of the results. Overall, Model 6 outperforms Model 7 for both Cases (1) and (2); its RMSE is never larger and it is often clearly smaller. For a very short horizon of 1 h, it is not surprising that the two models perform about the same, because they both catch the relatively strong correlation between two successive hours. For longer horizons, they also perform about the same, presumably because the true correlation is not very strong in that case. But for the values of r in between (horizons of a few hours), Model 6 clearly performs better. For example, if we are at 11:00 A.M. (we have observed $W_{t,11}$) and want to predict the volume of calls between 4:00 and 5:00 P.M. on the same day (6 h ahead), the RMSE is 3.7 for Model 7 compared to 2.3 for Model 6. We also see the benefit of using information about the call volume during the early hours of the day when forecasting for the latter part of the day. For example, using Model 6 at 11:00 A.M. to forecast the call volume between 4:00 and 5:00 P.M., the RMSE is 3.5 if we ignore the call volume from midnight to 11:00 A.M. but it drops to 2.3 if we incorporate this information.

5 Conclusion

We have considered a variety of time series models for estimating and forecasting daily and hourly EMS call volumes. EMS demand is influenced by when people work, commute, sleep, and celebrate, and our models attempt to capture these influences at least in part via yearly, weekly, and daily seasonal cycles, as well as special treatment of New Year's day and the Stampede, the most important festival in the city we studied.

We used three basic approaches for daily call volumes: standard regression ignoring dependencies, regression models with correlated residuals, and a third approach (doubly-seasonal ARIMA) that takes into account a specific cross-effect dependency structure at the start and captures the correlations between residuals as well. The usual interpretation of these models is that the first deterministic part captures the seasonal and non-seasonal components, and the second stochastic part (errors) captures the effect of omitted or non-observable effects such as serial correlation. We find that a model from the second category, that includes a selected subset of significant day-of-week main effects, month-of-year main effects, and interaction terms, performs best when forecasting 1 or 2 days into the future. The advantage of this model over the standard regression model decreases as the length of the forecast horizon increases, and disappears at around 2 weeks.

Table 6 RMSEs by origin and horizon for the two hourly models

	Horizon r	Model 6	Model 7
Case (1):	12 (11:00–12:00 A.M.)	3.1	3.5
Forecasts of $Z_{24(t-1)+r}$ to Z_{24t},	14 (1:00–2:00 P.M.)	3.1	3.9
at time $h = 24(t-1)$	17 (4:00–5:00 P.M.)	3.5	3.9
	23 (10:00–11:00 P.M.)	3.1	3.2
	24 (11:00–12:00 P.M.)	3.2	3.3
Case (2):	1 (11:00–12:00 A.M.)	3.1	3.1
Forecasts of $Z_{24(t-1)+11+r}$ to Z_{24t},	3 (1:00–2:00 P.M.)	2.9	3.5
at time $h = 24(t-1)+11$	6 (4:00–5:00 P.M.)	2.3	3.7
	12 (10:00–11:00 P.M.)	3.1	3.2
	13 (11:00–12:00 P.M.)	3.1	3.1

The doubly-seasonal ARIMA model performed poorly when forecasting more than a week into the future.

For hourly call volumes, we used two approaches: one built around the conditional distribution of hourly volumes, conditional on the daily volume, and another that fits a time-series model to the hourly data. Both approaches can be combined with any of the daily call volume models that we investigated, and we illustrated its use with the best-performing daily call volume model. We also showed how one could compute intraday forecast updates, which could be useful for real-time staffing decisions. We found that the conditional distribution approach generally worked better. We demonstrated that updating hourly forecasts using call volume from the early part of the day can improve forecast accuracy considerably, at least for certain hours of the day.

The models we present are simple and practical, and could be used for routine forecasting for an EMS system, as well as in simulation models of such systems. Our models that combine regression and ARMA processes showed an improvement over pure seasonal ARIMA. We also demonstrated the importance of modeling the effects of special days, day-of-week, and month-of-year.

Although we expect that the general approach described in this paper should be applicable in other cities, it is important to investigate whether this is the case. In future research, it would be interesting and useful to develop models that forecast the spatial distribution of demand, not only based on time but also on demographic variables.

Acknowledgements This work has been supported by NSERC-Canada Grants no. ODGP-0110050 and CRDPJ-251320, a grant from Bell Canada via the Bell University Laboratories, and a Canada Research Chair to P. L'Ecuyer as well as NSERC-Canada Grant no. 203534 to A. Ingolfsson. We thank the Calgary EMS department (particularly Heather Klein-Swormink and Tom Sampson) for making this work possible, J. Cheng, E. Erkut, D. Haight, and T. Riehl for useful discussions and assistance with data preparation, and the anonymous referees for useful comments.

References

1. Abraham B, Ledolter J (1983) Statistical methods for forecasting. Wiley, Toronto
2. Akaike H (1973) Information theory and an extention of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Proc. 2nd international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281
3. Andrews B, Cunningham SM (1995) L.L. Bean improves call-center forecasting. Interfaces 25:1–13
4. Baker JR, Fitzpatrick KE (1986) Determination of an optimal forecast model for ambulance demand using goal programming. J Oper Res Soc 37(11):1047–1059
5. Bell WR, Hillmer SC (1983) Modeling time series with calendar variation. J Am Stat Assoc 78(383):526–534
6. Bianchi L, Jarrett J, Hanumara RC (1998) Improving forecasting for telemarketing centers by ARIMA modeling with intervention. Int J Forecast 14:497–504
7. Box GEP, Jenkins GM, Reinsel GC (1994) Time series analysis, forecasting and control, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ
8. Brockwell PJ, Davis RA (1991) Time series: theory and methods, 1991 edn. Springer, Berlin Heidelberg New York
9. Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: a queueing-science perspective. J Am Stat Assoc 100:36–50
10. Cinlar E (1972) Superposition of point processes. In: Lewis PAW (ed) Stochastic point processes: statistical analysis, theory, and applications. Wiley, New York pp 549–606
11. Committee on the future of emergency care in the United States health system (2006) Emergency medical services: at the crossroads. Washington, DC
12. E-Comm (2005) Quarterly service report. http://www.ecomm.bc.cacorporatepublicationsesr_final.pdf
13. Eckstein M, Chan LS (2004) The effect of emergency department crowding on paramedic ambulance availability. Ann Emerg Med 43:100–105
14. Eckstein M, Isaacs SM, Slovis CM, Kaufman BJ, Loflin JR, O'Connor RE, Pepe PE (2005) Facilitating EMS turnaround intervals at hospitals in the face of receiving facility overcrowding. Prehosp Emerg Care 9:267–275
15. Erdogan G, Erkut E, Ingolfsson A (2006) Ambulance deployment for maximum survival. Technical report. http://www.bus.ualberta.ca/aingolfsson/publications
16. Goldberg JB (2004) Operations research models for the deployment of emergency service vehicles. EMS Mngt J 1:20–39
17. Green LV, Kolesar PJ (2004) Improving emergency responsiveness with management science. Manage Sci 50:1001–1014
18. Gunes E, Szechtman R (2005) A simulation model of a helicopter ambulance service. In: Proceedings of the 2005 winter simulation conference. IEEE Press, Piscataway, NJ pp 951–957
19. Henderson SG (2005) Should we model dependence and nonstationarity, and if so, how? In: Proceedings of the 2005 winter simulation conference. IEEE Press, Piscataway, NJ pp 120–129
20. Henderson SG, Mason AJ (2004) Ambulance service planning: simulation and data visualization. In: Sainfort F, Brandeau ML, Pierskalla WP (eds) Handbook of operations research and health care methods and applications. Kluwer, Boston pp 77–102
21. Hogg RV, Craig AF (1995) Introduction to mathematical statistics, 5th edn. Prentice-Hall, Englewood Cliffs, NJ
22. Ingolfsson A, Erkut E, Budge S (2003) Simulation of single start station for Edmonton EMS. J Oper Res Soc 54:736–746
23. Kamenetsky R, Shuman L, Wolfe H (1982) Estimating need and demand for prehospital care. Oper Res 30:1148–1167
24. Larson RC (1974) A hypercube queuing model for facility location and redistricting in urban emergency services. Comput Oper Res 1(1):67–95
25. Larson RC (1975) Approximating the performance of urban emergency service systems. Oper Res 23(5):845–868

26. Mabert VA (1985) Short interval forecasting of emergency phone call (911) work loads. J Oper Manag 5(3):259–271
27. McConnell CE, Wilson RW (1998) The demand for prehospital emergency services in an aging society. Soc Sci Med 46(8):1027–1031
28. Moeller A (2004) Obstacles to measuring emergency medical service performance. EMS Mngt J 1:8–15
29. National Fire Protection Association (2002) NFPA 1221: standard for the installation, maintenance, and use of emergency service communication systems.
30. Pierce DA (1971) Least squares estimation in the regression model with autoregressive-moving average errors. Biometrika 58(2):299–312
31. Read TRC, Cressie NAC (1988) Goodness-of-fit statistics for discrete multivariate data. Springer series in statistics. Springer, Berlin Heidelberg New York
32. Resnick S (1992) Adventures in stochastic processes. Birkhauser, Boston
33. Segal W, Verter V, Colacone A, Afilalo M (2006) The in-hospital interval: a description of EMT time spent in the emergency department. Prehosp Emerg Care 10:378–382
34. Sprivulis P, Gerrard B (2005) Internet-accessible emergency department workload information reduces ambulance diversion. Prehosp Emerg Care 9:285–291
35. Swersey AJ (1994) The deployment of police, fire and emergency medical units. In: Pollock SM, Rothkopf M, Barnett A (eds) Handbooks in operations research and management science, vol 6. North-Holland, New York pp 151–190
36. Tych W, Pedregal DJ, Young PC, Davies J (2002) An unobserved component model for multi-rate forecasting of telephone call demand: the design of a forecasting support system. Int J Forecast 18:673–695
37. US Congress, O.o.T.A. (1989) Rural emergency medical services—special report. Washington, DC
38. Wei WW-S (1990) Time series analysis: univariate and multivariate methods. Addison-Wesley, New York
39. Zhu Z, McKnew MA, Lee J (1992) Effects of time-varied arrival rates: an investigation in emergency ambulance service systems. In: Proceedings of the 1992 winter simulation conference. IEEE Press, Piscataway, NJ pp 1180–1186