APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY
Appl. Stochastic Models Bus. Ind., 2006; **22**:297–311
Published online 13 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asmb.618

Non-parametric modelling of time-varying customer service times at a bank call centre

Haipeng Shen[1,*,†] and Lawrence D. Brown[2,‡]

[1] *Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, U.S.A.*
[2] *Department of Statistics, The Wharton School, University of Pennsylvania, U.S.A.*

SUMMARY

Call centres are becoming increasingly important in our modern commerce. We are interested in modelling the time-varying pattern of average customer service times at a bank call centre. Understanding such a pattern is essential for efficient operation of a call centre. The call service times are shown to be lognormally distributed. Motivated by this observation and the important application, we propose a new method for inference about non-parametric regression curves when the errors are lognormally distributed. Estimates and pointwise confidence bands are developed. The method builds upon the special relationship between the lognormal distribution and the normal distribution, and improves upon a naive estimation procedure that ignores this distributional structure. Our approach includes local non-parametric estimation for both the mean function and the *heteroscedastic* variance function of the logged data, and uses local polynomial regression as a fitting tool. A simulation study is performed to illustrate the method. We then apply the method to model the time-varying patterns of mean service times for different types of customer calls. Several operationally interesting findings are obtained and discussed. Copyright © 2006 John Wiley & Sons, Ltd.

Received 3 October 2005; Revised 23 December 2005; Accepted 6 January 2006

KEY WORDS: service engineering; queueing theory; local polynomial regression; variance estimation; heteroscedasticity; bandwidth selection

1. INTRODUCTION

Call centres are modern service networks in which customer service agents provide services to customers via telephones. They have become a primary communication channel between service providers and their customers. Thus, managing call centre operations efficiently is playing an increasingly important role in our modern business world [1]. Call centres are mathematically

*Correspondence to: Haipeng Shen, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.
† E-mail: haipeng@email.unc.edu
‡ E-mail: lbrown@wharton.upenn.edu

modelled as queueing systems and analysed using queueing theory. During the last decade, considerable research has been devoted to the call centre industry as documented in Mandelbaum [2]. However, relatively few statistical papers are listed. The current paper is part of a larger research project aiming at reducing the gap between the current practice of statistics and the prevalent needs in call centre modelling.

In this paper, the problem of interest is to model the time-varying pattern of call (or *customer*) service times at a bank call centre. The motivating application is described below in Section 2. Call service times are defined as times needed to serve individual customer calls. For a call centre system, the mean service time is one essential quantity for calculating several basic performance measures, such as average waiting time in the system or average delay in the queue as shown in Reference [3]. When combined with a prediction of future arrival rates, it can also be used to predict the future workload that will arrive to the system, which can then be used for agent staffing and capacity planning. Consequently, understanding the time-varying pattern of the mean service time is necessary for understanding the time-varying operational environment of a call centre, and also for dynamically forecasting future workload.

The importance of the mean service time necessitates the use of more precise tools for statistical inference about the mean. Common call centre analyses usually assume that the customer service times follow an exponential distribution. However, in an Israeli banking call centre (Section 2), the service times are approximately lognormally distributed instead. This empirical finding has potentially important implications for call centre system modelling [4].

The lognormal nature of the service times, as well as the specific interest in their mean, motivate us to develop a new method for non-parametric estimation of regression models involving lognormal errors, as well as for the generation of accompanying confidence bands. Although the motivating application is to model customer service times at a call centre, the same methodology can be applied in other contexts involving non-parametric regression problems with lognormal errors.

Our approach builds upon the special connection between the lognormal distribution and the normal distribution. Suppose $\{X_i, Z_i\}_{i=1}^n \sim^{\text{i.i.d.}} \{X, Z\}$ where $Z|X = x$ has a *conditional* lognormal distribution with mean $v(x) = E(Z|X = x)$. We are interested in providing a simple non-parametric estimator for $v(x)$, along with a reasonable pointwise confidence band. Let $Y = \ln(Z)$ and $Y_i = \ln(Z_i)$ for $i = 1, \ldots, n$. Then $Y_1|X_1, \ldots, Y_n|X_n$ have the same conditional distribution as $Y|X = x$, which is normal with mean $\mu(x)$ and variance $\sigma^2(x)$. Of particular interest are scenarios where the variance $\sigma^2(x)$ is a function of x (*instead of a constant*). In the call centre application, X is the time-of-day when a call begins its service, Z is the corresponding service time and $Y = \ln(Z)$ is the natural-logged service time of the call, which is normally distributed conditional on X. The current paper deals with single covariate cases. Possible extensions to multiple covariate problems are discussed in Section 6.

A simple calculation reveals that

$$v(x) = \exp\left[\mu(x) + \sigma^2(x)/2\right] \tag{1}$$

The relation (1) suggests a *simple* plug-in approach to derive the regression curve $\hat{v}(x)$ and the corresponding confidence band for $v(x)$. The basic idea is stated here while the estimation details are relegated to Section 3. From the transformed data $\{X_i, Y_i\}_{i=1}^n$, we derive estimates for $\mu(x)$ and $\sigma^2(x)$ with their corresponding confidence bands. The inference results are then back-transformed to the original scale to obtain the estimated mean curve, $\hat{v}(x)$, along with its confidence band. The above plug-in principle has been used in one-population lognormal mean

DOI: 10.1002/asmb

estimation [5]. By using the lognormal distributional structure, our method has a better performance than a naive alternative that ignores this knowledge (Section 4).

The rest of the paper is organized as follows. In Section 2, we describe the motivating call centre service time data. Our modelling approach is proposed in Section 3. We illustrate the proposed approach via a simulation study in Section 4, and show that it improves over the common naive approach. Section 5 reports the modelling results from the call centre service time data. We conclude the paper and discuss possible future work in Section 6.

2. CALL CENTRE SERVICE TIME DATA

In this section, we first describe the data that motivate our research. Then we empirically show that the service times are lognormally distributed. This observation has also been confirmed in two other call centre service time data sets. Mandelbaum and Schwartz [4] explore various implications of this distributional finding for modelling call centres.

The data motivating our study were collected at a small call centre from an Israeli bank in 1999. The centre provides several types of services such as *Regular Services*, *Stock Transactions*, *New Customer* and *Internet Assistance*. The data of interest here are the service records of those served service-request calls to the centre. These are the calls in which the caller requests service from an agent and actually gets the service before leaving the centre. The data include the starting and ending times of the service in addition to the agent names and the service types. See Reference [6] for more information about the data. Different features of the data are investigated rather broadly in Reference [7]. Shen [8] reports a thorough analysis of the service time data.

Figure 1 plots the lognormal quantile–quantile (Q–Q) plot of the service times for calls served in November and December. The two dashed curves are the simulated 95% band for the

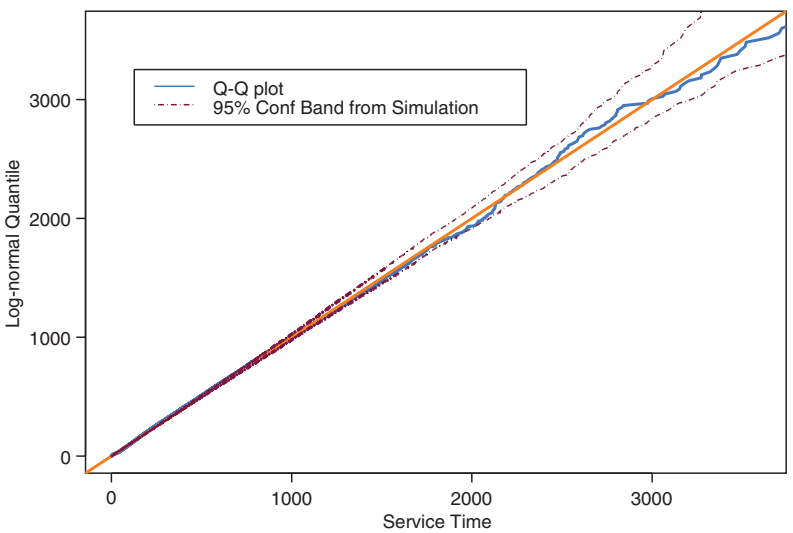

Figure 1. Lognormal Q–Q plot of service times (Nov + Dec).

DOI: 10.1002/asmb

estimated lognormal distribution in order to incorporate random variation. The quantile plot is very close to linear, and also lies well within the band. This suggests that the distribution of the service times is very nearly lognormal.

The lognormality also holds for data from other months. We note the exception that, for January to October, special care needs to be taken to separate out a group of very-short calls. These are due to several ill-behaved agents who simply hung up on customers to obtain 'extra' rest-times. In addition, the lognormality seems to hold well for different types of calls, for individual agents, and especially when conditioning on time-of-day. The lognormal structure will motivate our modelling approach. Lognormality of processing times has been previously recognized by researchers in telecommunication and psychology [9–11].

3. METHODOLOGY

The lognormal service times enable us to use powerful statistical machinery to model the mean service times as functions of various covariates, for example, time-of-day. In this section, we propose a simple method to model the mean service time as a continuous function of time-of-day.

As indicated by (1), in order to estimate the mean curve $v(x)$ non-parametrically, we need to estimate both $\mu(x)$ and $\sigma^2(x)$ non-parametrically. Several methods for estimating $\mu(x)$ are available, for example, kernel regression, local polynomial regression, smoothing splines and basis expansion methods such as polynomial splines and wavelets. Our basic idea should be amenable to any of these methods. For illustration purposes, we employ a local polynomial regression method [12, 13]. Other non-parametric regression methods could be used instead and would give generally similar results. We also adopt the same method to estimate $\sigma^2(x)$.

Below, in Section 3.1, we briefly review the local polynomial regression method. See References [12, 13] for details. A data-driven bandwidth selection method is described in Section 3.2. Our estimation procedure is then proposed in detail in Section 3.3.

3.1. Local polynomial regression

Suppose locally around a point x_0, the regression function $\mu(x)$ can be well approximated with a polynomial of order p according to Taylor's expansion, i.e.

$$\mu(x) \approx \sum_{j=0}^{p} a_j (x - x_0)^j$$

Then, the local polynomial estimator of $\mu(\cdot)$ at x_0 is defined as $\hat{\mu}(x_0) = \hat{a}_0$ where $(\hat{a}_0, \ldots, \hat{a}_p)$ minimizes the locally weighted sum of squares

$$\sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{p} a_j (X_i - x_0)^j \right]^2 K_h(X_i - x_0)$$

Here $K_h(\cdot) = h^{-1} K(\cdot/h)$ with $K(\cdot)$ being a kernel function on $\mathbb{R}^1$ and $h > 0$ is a bandwidth. The form of the kernel function K has a minor effect on the estimator. One popular choice, and the one we use below, is the *tricube* kernel function

$$K(x) = (1 - |x|^3)^3, \quad |x| \leqslant 1$$

DOI: 10.1002/asmb

As for the polynomial order p, the most common choices are $p = 1$ and $p = 2$, which correspond to local linear regression and local quadratic regression, respectively.

3.2. Data-driven bandwidth selection

The bandwidth h is one of the critical components for local polynomial regression. It controls the amount of smoothing applied to the data, which affects the bias–variance trade-off. To make the bandwidth adaptive, we employ a *nearest neighbour* bandwidth [13]. At a particular fitting point x_0, a nearest neighbour bandwidth $h(x_0)$ is chosen so that the local neighbourhood always contains a pre-specified number of points. For a smoothing parameter $\beta \in (0, 1)$, the nearest neighbour bandwidth $h(x_0)$ is computed using the following two steps:

1. Compute the distances $d(x_0, X_i) = |x_0 - X_i|$ between the fitting point x_0 and the data points X_i;
2. Choose $h(x_0)$ to be the $\lfloor n\beta \rfloor$th smallest distance.

Each local neighbourhood then contains approximately $100\beta\%$ of the data.

Sometimes, it suffices to choose a bandwidth subjectively, but there are occasions where the bandwidth needs to be selected automatically from the data. In our problem, there are two separate bandwidths to choose, one for estimating the mean function and the other for estimating the variance function.

The literature on automatic bandwidth selection for non-parametric mean function estimation is extensive. See Fan and Gijbels [14], Ruppert *et al.* [15], Jones *et al.* [16, 17], just to name a few. On the other hand, the bandwidth selection literature for non-parametric variance function estimation is rather sparse. Ruppert *et al.* [18] use a data-driven bandwidth selector proposed by Ruppert [19] to select bandwidths for estimating both the mean and variance functions using local polynomial regression. Levine [20] proposes a cross-validation-type bandwidth selector for difference-based estimators for non-parametric variance function estimation, and argues that it works better than a plug-in alternative.

For the current paper, we do not intend to compare existing bandwidth selectors or propose a new one. Instead, we use the traditional K-fold cross-validation to choose bandwidths. Part of the following description is cited from Section 7.10 of Hastie *et al.* [21]. The method is simple and appears to work well empirically in the simulation study and the call centre application.

K-fold cross-validation usually works as follows. It first randomly splits the data into K roughly equal-sized parts. Let $\kappa : \{1, \ldots, n\} \mapsto \{1, \ldots, K\}$ be an indexing function that indicates the partition set to which the ith observation x_i is allocated by the randomization. Let $\hat{f}_{\beta}^{-k}(\cdot)$ denote the fitted function computed with the kth part of the data removed. Here the subscript β emphasizes the fact that the fitted function depends on the bandwidth parameter, β. Then the K-fold cross-validation estimate of the prediction error is

$$\mathrm{CV}(\beta) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}_{\beta}^{-\kappa(i)}(x_i))$$

where L is a loss function, and is chosen to be the squared error loss in the current paper. Finally, we select a bandwidth h corresponding to $\hat{\beta}$ that minimizes $\mathrm{CV}(\beta)$. This then leads to the final chosen model $\hat{f}_{\hat{\beta}}(\cdot)$, which is fitted using the entire data. In practice, K is usually chosen to be 5 or 10. The case where $K = n$ is known as *leave-one-out* cross-validation, which can be computationally expensive for moderate to large data.

DOI: 10.1002/asmb

3.3. The estimation procedure

In this subsection, we go over the estimation procedure step by step, using the local polynomial regression method for illustration. In particular, local quadratic regression is employed. As we go through the details below, one can see that it is fairly easy to generalize the proposed method to other smoothers.

We first sort the original data $\{X_i, Z_i\}_{i=1}^n$ in increasing order of the X_i's. Such an ordering is necessary for estimating the variance function $\sigma^2(x)$ as shown below in Section 3.3.2. One has to be careful with possible ties within the X_i's. If these occur, one solution is to take a random order of the tied observations. To be more rigorous, one can take all the permutations of the tied observations, apply the following procedure, and then take the average.

Then, we transform the sorted data $\{X_i, Z_i\}_{i=1}^n$ to $\{X_i, Y_i\}_{i=1}^n$ by taking the natural logarithm of the responses, i.e. $Y_i = \ln(Z_i)$. On the transformed scale, our model is

$$Y_i = \mu(X_i) + \sigma(X_i)\varepsilon_i \tag{2}$$

where $\varepsilon_i | X_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$. Both the mean function $\mu(x)$ and the variance function $\sigma^2(x)$ are unknown and need to be estimated.

3.3.1. Estimation of $\mu(x)$. We apply the local quadratic regression to estimate the mean function $\mu(x)$. According to the general introduction of local polynomial regression in Section 3.1, the local quadratic estimator of $\mu(x)$ is $\hat{\mu}(x) = \hat{a}_0$ where $(\hat{a}_0, \hat{a}_1, \hat{a}_2)$ minimizes the following weighted sum of squares:

$$\sum_{i=1}^{n} [Y_i - a_0 - a_1(X_i - x) - a_2(X_i - x)^2]^2 K_h(X_i - x)$$

It follows from weighted least squares theory that

$$\hat{\mu}(x) = \hat{a}_0 = e_{1,3}^{\mathrm{T}} (\widetilde{X}^{\mathrm{T}} W \widetilde{X})^{-1} \widetilde{X}^{\mathrm{T}} W Y \tag{3}$$

where $e_{1,3} = (1,0,0)^{\mathrm{T}}$, $\widetilde{X}$ denotes an $n \times 3$ matrix with $(1, X_i - x, (X_i - x)^2)$ as its ith row, and $W = \mathrm{diag}\{K_h(X_1 - x), \ldots, K_h(X_n - x)\}$, and Y is the column response vector. Furthermore, the variance of $\hat{\mu}(x)$ is

$$\sigma_\mu^2(x) = \mathrm{var}(\hat{\mu}(x)) = e_{1,3}^{\mathrm{T}} (\widetilde{X}^{\mathrm{T}} W \widetilde{X})^{-1} \widetilde{X}^{\mathrm{T}} W \Sigma_Y W \widetilde{X} (\widetilde{X}^{\mathrm{T}} W \widetilde{X})^{-1} e_{1,3} \tag{4}$$

where $\Sigma_Y = \mathrm{var}(Y) = \mathrm{diag}\{\sigma^2(X_1), \ldots, \sigma^2(X_n)\}$.

The variance expression (4) suggests that, in order to estimate $\sigma_\mu^2(x)$, one needs to first estimate the $\sigma^2(X_i)$'s. The variance estimation problem is addressed below in Section 3.3.2. Suppose we obtain an estimate $\hat{\sigma}_\mu^2(x)$. Then, one approximate $100(1-\alpha)\%$ confidence interval for $\mu(x)$ is

$$\hat{\mu}(x) \pm z_{\alpha/2} \hat{\sigma}_\mu(x) \tag{5}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal. Note that this confidence interval is a variance band. A common practice is to *under-smooth* the data so that the resulting local quadratic estimate $\hat{\mu}(x)$ is approximately unbiased.

The errors are normally distributed, which justifies our use of a local regression method instead of a local likelihood method. (See Section 6 for more discussion of local likelihood.) The nearest neighbour bandwidth can be chosen subjectively, or be selected automatically by a data-driven method as pointed out in Section 3.2.

DOI: 10.1002/asmb

3.3.2. Estimation of $\sigma^2(x)$. The variance function $\sigma^2(x)$ arises in $v(x)$ and $\sigma_\mu^2(x)$. We propose to estimate $\sigma^2(x)$ using the following two-step procedure, a *simple* difference-based variance estimator plus local quadratic regression.

The observations $\{X_i, Y_i\}_{i=1}^n$ are first regrouped into consecutive non-overlapping pairs

$$\{X_{2i-1}, Y_{2i-1}; X_{2i}, Y_{2i}\}_{i=1}^{\lfloor n/2 \rfloor}$$

Define a squared pseudo-residual D_{2i} to be of the form $(Y_{2i} - Y_{2i-1})^2/2$, which naturally estimates $\sigma^2(X_{2i})$, the local variance at X_{2i}. It can be shown that D_{2i} is approximately a multiple of a χ_1^2 random variable with the multiplier being $\sigma^2(X_{2i})$; hence $E(D_{2i}) = \sigma^2(X_{2i})$ and $\text{var}(D_{2i}) = 2\sigma^4(X_{2i})$.

The estimator D_{2i} is a special difference-based estimator. There are many other difference-based estimators in the literature as discussed in References [22–24]. More recently, Levine [25] studies the theoretical properties of a class of difference-based estimators for variance functions. Our choice of D_{2i} is a simple one that suffices for our purpose. More efficient estimators might slightly improve results, especially for problems with small sample sizes. For example, one could opt for a more efficient estimator by using adjacent overlapping pairs [25]. In addition to the difference-based estimators, there are other types of variance function estimators. See References [18, 26, 27] and references within for more details. Hall *et al*. [23] and Levine [25] show that difference-based estimators, in general, reduce the bias caused by the unknown mean function.

After obtaining the D_{2i}'s, we treat $\{X_{2i}, D_{2i}\}_{i=1}^{\lfloor n/2 \rfloor}$ as our observed data points and apply local quadratic regression to obtain $\hat{\sigma}^2(x)$. The following model is assumed:

$$D_{2i} = \sigma^2(X_{2i}) + \sqrt{2}\sigma^2(X_{2i})\varepsilon'_{2i}, \quad i = 1, \ldots, \lfloor n/2 \rfloor \tag{6}$$

where ε'_{2i} have mean 0 and variance 1, and are independent for varying i. Part of our justification is that, under (2), the D_{2i}'s are (conditionally) independent given the X_{2i}'s, because the D_{2i}'s are generated from non-overlapping pairs.

Similarly, the local quadratic estimate of $\sigma^2(x)$ is

$$\hat{\sigma}^2(x) = e_{1,3}^{\mathrm{T}}(\widetilde{X}_D^{\mathrm{T}} W_D \widetilde{X}_D)^{-1} \widetilde{X}_D^{\mathrm{T}} W_D D \tag{7}$$

where $e_{1,3} = (1, 0, 0)^{\mathrm{T}}$, $\widetilde{X}_D$ is an $\lfloor n/2 \rfloor \times 3$ matrix with $(1, X_{2i} - x, (X_{2i} - x)^2)$ as its ith row, and $W_D = \text{diag}\{K_h(X_2 - x), \ldots, K_h(X_{2\lfloor n/2 \rfloor} - x)\}$, and $D = (D_2, \ldots, D_{2\lfloor n/2 \rfloor})^{\mathrm{T}}$.

The derived $\hat{\sigma}^2(x)$ can then be plugged into the variance formula (4) to obtain an estimate for $\sigma_\mu^2(x)$, which then leads to a confidence interval for $\mu(x)$ according to (5). Furthermore, the variance of $\hat{\sigma}^2(x)$ is given by

$$\sigma_\sigma^2(x) = e_{1,3}^{\mathrm{T}}(\widetilde{X}_D^{\mathrm{T}} W_D \widetilde{X}_D)^{-1} \widetilde{X}_D^{\mathrm{T}} W_D \Sigma_D W_D \widetilde{X}_D (\widetilde{X}_D^{\mathrm{T}} W_D \widetilde{X}_D)^{-1} e_{1,3} \tag{8}$$

where $\Sigma_D = \text{var}(D) = \text{diag}\{2\sigma^4(X_2), \ldots, 2\sigma^4(X_{2\lfloor n/2 \rfloor})\}$. Expression (8) suggests that we can estimate $\sigma_\sigma^2(x)$ by plugging in an estimated Σ_D based on estimate (7).

A $100(1 - \alpha)\%$ confidence interval for $\sigma^2(x)$ is approximately

$$\hat{\sigma}^2(x) \pm z_{\alpha/2}\hat{\sigma}_\sigma(x)$$

Note that we use $z_{\alpha/2}$ as the cut-off value when deriving the above confidence interval, rather than a quantile from a Chi-square distribution. This approximation works fine with a moderate to large sample size, which is the case for our call centre application.

We want to comment on three things here. First, since the $\{D_{2i}\}$'s have a Chi-squared distribution, $\sigma^2(x)$ can also be estimated via a local likelihood approach. With a large sample

DOI: 10.1002/asmb

size, the two approaches give similar results as shown in Reference [8]. Second, we propose to estimate the variance of $\hat{\sigma}^2(x)$ by $2\hat{\sigma}^4(x)$. This is due to the Chi-square nature of the $\{D_{2i}\}$'s. Alternatively, one can use the squared differences of the $\{D_{2i}\}$'s to estimate the variance. Shen [8] applies both methods to the call centre data and obtains similar results. It might be of interest to compare the two approaches in a simulation study. Finally, a separate bandwidth needs to be selected here, which will in general differ from the bandwidth selected for estimating the mean function.

3.3.3. Estimation of $v(x)$. Finally, we can back-transform the inference results obtained above to the original scale, and derive the following plug-in estimator for $v(x)$:

$$\hat{v}(x) = \exp[\hat{\mu}(x) + \hat{\sigma}^2(x)/2]$$

The plug-in principle has been used before to obtain estimators for one-population lognormal means as reviewed in Shen *et al*. [5]. For example, the maximum likelihood estimator (MLE) for the mean v is obtained by plugging in the MLEs of μ and σ^2.

Given the methods used for estimating $\mu(x)$ and $\sigma^2(x)$, $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$ are asymptotically independent, which suggests that

$$\text{var}\,(\hat{\mu}(x) + \hat{\sigma}^2(x)/2) \approx \sigma_{\mu}^2(x) + \sigma_{\sigma}^2(x)/4$$

Then, the corresponding $100(1-\alpha)\%$ large sample confidence interval for $v(x)$ is

$$\exp\left[\hat{\mu}(x) + \hat{\sigma}^2(x)/2 \pm z_{\alpha/2}\sqrt{\hat{\sigma}_{\mu}^2(x) + \hat{\sigma}_{\sigma}^2(x)/4}\right]$$

The use of $z_{\alpha/2}$ in the above confidence interval is supported by the nice finite-sample coverage property of Cox's interval [28] for lognormal means. Shen [8] shows that this approximation works fine as long as the sample size is not too small and the variance is not too large. To be exact, we should derive the confidence interval based on the cut-off values of the exact distribution of $\hat{\mu}(x) + \hat{\sigma}^2(x)/2$. A parametric bootstrap approach seems to be a reasonable route to go. Shen and Zhu [29] describe one such approach for a lognormal linear model setup. We intend to follow this reasoning in a future manuscript.

4. A SIMULATION STUDY

In this section, we use Monte Carlo simulation to investigate the performance of the proposed approach, and also compare it with an alternative direct estimation approach, which ignores the lognormal nature of the errors. For both approaches, the 'optimal' bandwidth is selected using the 5-fold cross-validation as described in Section 3.2.

To gauge the performance of an estimator $\hat{v}(\cdot)$ of $v(\cdot)$, we define a criterion, the square-root of average squared error (RASE), as

$$\text{RASE} = \sqrt{\sum_{k=1}^{n_{\text{grid}}} [\hat{v}(u_k) - v(u_k)]^2 / n_{\text{grid}}} \tag{9}$$

where $\{u_k, k = 1, \ldots, n_{\text{grid}}\}$ are grid points that are chosen to be equally spaced over a certain interval within the data range.

We study the model in (2) where $\mu(x) = 3 + 6(x+0.3)e^{-8x^2} + 2(x+0.3)e^{-4(x-0.7)^2}$, $\sigma(x) = a + (x-0.5)^2$ with a chosen to be 0.5, 0.75, 1.5 and 2, $\{X_i\}$ are i.i.d. U[0, 1] and $\{\varepsilon_i\}$

DOI: 10.1002/asmb

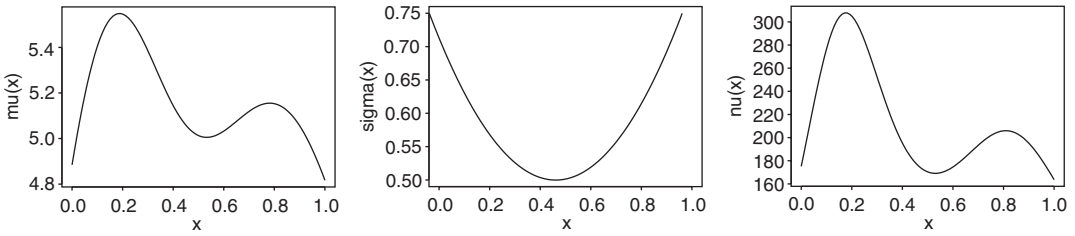

Figure 2. Plot of $\mu(\cdot)$, $\sigma(\cdot)$ and $v(\cdot)$ for $a = 0.5$.

Table I. Mean (SE) of the RASE ratios (Model (2)/Model (10)).

n	$a = 0.50$	$a = 0.75$	$a = 1.50$	$a = 2.00$
2000	0.902 (0.0225)	0.948 (0.0329)	0.924 (0.0389)	0.804 (0.0441)
5000	0.901 (0.0213)	0.897 (0.0274)	0.903 (0.0417)	0.717 (0.0445)

are i.i.d. N(0, 1). The sample size n is chosen to be 2000 and 5000. For each simulation setup, the simulation is replicated 100 times. Model (2) suggests that $z_i|X_i$ is conditionally lognormally distributed with mean

$$v(X_i) = \exp\left[\mu(X_i) + \sigma^2(X_i)/2\right]$$

For illustration purposes, Figure 2 plots the functions $\mu(\cdot)$, $\sigma(\cdot)$ and $v(\cdot)$ for $a = 0.5$. Different a shifts $\sigma(\cdot)$ vertically, and consequently changes $v(\cdot)$.

The naive alternative approach ignores lognormality of the z_i's and assume the standard model

$$z_i = v(X_i) + \varepsilon_i^*, \quad i = 1, \ldots, n \tag{10}$$

where $\{\varepsilon_i^*\}$ are normally distributed. Under this model, v can be estimated non-parametrically by regressing z_i on X_i using local polynomial regression.

When calculating the RASEs from each model, we choose 100 equally spaced points between 0 and 1 as the grid points $\{u_k\}$ to be used in (9). For each simulation setup, Table I reports the mean and the standard error of the ratios between the 100 RASEs from the proposed lognormal approach and the naive normal approach. The summaries clearly suggest that our proposed estimation procedure, which takes into account the lognormality, gives much more accurate results than the direct estimation procedure. The improvement of Model (2) over Model (10) increases as the variance function σ^2 gets larger. This is consistent with results from comparing various estimators of one-population lognormal means [5].

Figure 3 plots the real function and the average fitted function over the 100 runs for $a = 1.5$ and $n = 5000$. For each grid point, we also plot the 5%- and 95%- quantiles of the corresponding 100 fitted values. The two panels correspond to Models (2) and (10), respectively. As one can see, our proposed procedure can estimate the real function with smaller bias and also smaller variability than the direct approach. The ratio between the average absolute biases is

DOI: 10.1002/asmb

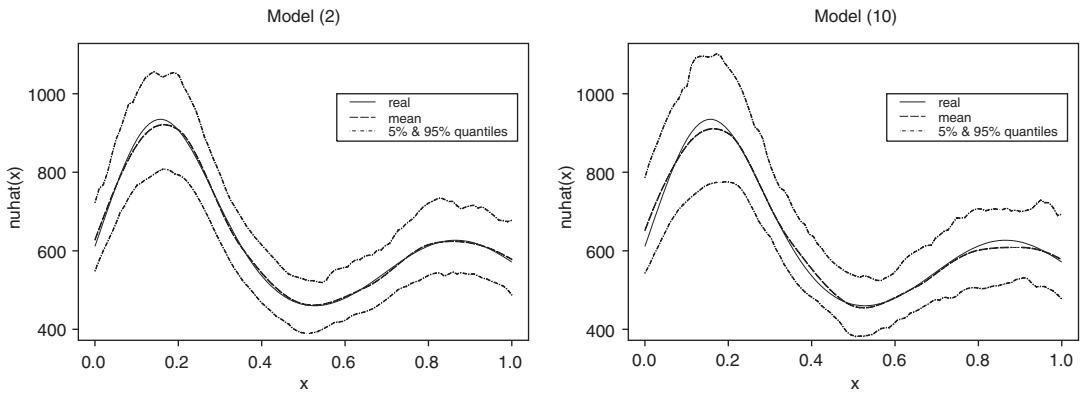

Figure 3. Real, mean, 5% and 95% functions of $\hat{v}(\cdot)$ ($a = 1.5, n = 2000$).

Table II. Mean (Median) of the empirical coverage probability of model (2).

n	$a = 0.50$	$a = 0.75$	$a = 1.50$	$a = 2.00$
2000	0.95 (0.96)	0.93 (0.94)	0.94 (0.94)	0.94 (0.96)
5000	0.93 (0.93)	0.95 (0.96)	0.94 (0.95)	0.95 (0.95)

0.537, and the ratio between the average interval widths (i.e. the distances between the 5%- and 95%-quantiles) is 0.855. Similar results are obtained for the other simulation setups. Again, the improvement of the lognormal approach over the normal approach increases as a gets larger, which makes intuitive sense.

We also investigate the coverage property of the pointwise confidence band for $v(\cdot)$ as proposed in Section 3.3. For each simulation setup, Table II reports the mean and median of the pointwise empirical coverage probabilities of the 95% confidence bands calculated for the 100 simulations. These summaries are all very close to the nominal level, which suggests that our proposed confidence band has good finite-sample coverage.

5. THE CALL CENTRE APPLICATION

In this section, we apply the Section 3 methodology to the call centre data and model the time-of-day pattern of mean service times. Out of the six major types of calls handled in the centre, we consider two specific types of calls, *Regular Service* (PS) and *Internet Assistance* (IN). PS calls constitute the majority of all the calls while IN calls are handled by a separate pool of service agents beginning in August. It is of interest to perform separate analyses for these two call types and compare the results. As it turns out, these two types of calls have very different mean service time patterns across time-of-day. This observation is very important for call centre staffing, especially for call centres using skill-based routing, where different types of calls are routed to agents with different skills for service.

DOI: 10.1002/asmb

Skill-based routing is a newly developed technology that allows for distinctions to be made among different types of calls and different skills of agents. The separate agent pool for the IN calls is one simple example of skill-based routing. Another example would be to group agents into regular and premium agents and let them handle different types of calls. See Reference [1] and the references within for more on skill-based routing and associated capacity-planning problems.

For the following analyses, we apply the local quadratic regression method with the tricube weight function. The nearest neighbour bandwidths are chosen automatically using the 5-fold cross-validation approach described in Section 3.2. As pointed out earlier, two separate bandwidths need to be selected.

Due to the *very-short-call* phenomenon mentioned in Section 2, the analyses below involve only the served calls in November and December. Thus, a mixture model analysis to separate out those very-short calls is not needed. Furthermore, we focus on those calls arriving during the normal business hours (7:00AM to midnight) of weekdays; the call centre does not operate fully during weekends and call volumes are much smaller then. In total there are 62 303 calls in our data: 42 613 PS calls and 5066 IN calls.

5.1. PS calls

For mean function estimation, the bandwidth parameter β is searched between 0.01 and 0.61 with a step size of 0.02, and the 'optimal' choice is 0.07, as shown in the first panel of Figure 4. The second panel of Figure 4 plots the corresponding fitted mean curve (*solid line*) with the 95% confidence band (*dashed lines*) attached. As for the variance function, the search range for the bandwidth parameter is between 0.2 and 0.6 with the same step size. The cross-validation selects a bandwidth of 0.28 as shown in the first panel of Figure 5. The fitted variance function with the corresponding confidence band is plotted in the second panel of Figure 5.

From Figure 5, one sees variance heteroscedasticity, with the variances increasing towards 9:30PM. The difference is significant as indicated by the non-overlapping confidence intervals at 6:00PM and 9:30PM. This difference has a more significant effect on the final estimate of the mean service time because it is exponentiated and multiplied to the exponential of the estimated mean Log(Service Time).

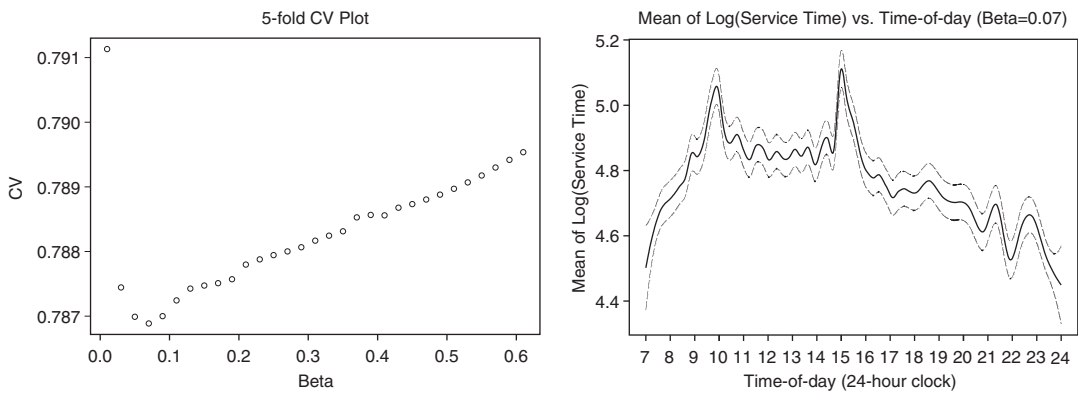

Figure 4. 5-fold Cross-validation for mean of Log(Service Time) (PS).

DOI: 10.1002/asmb

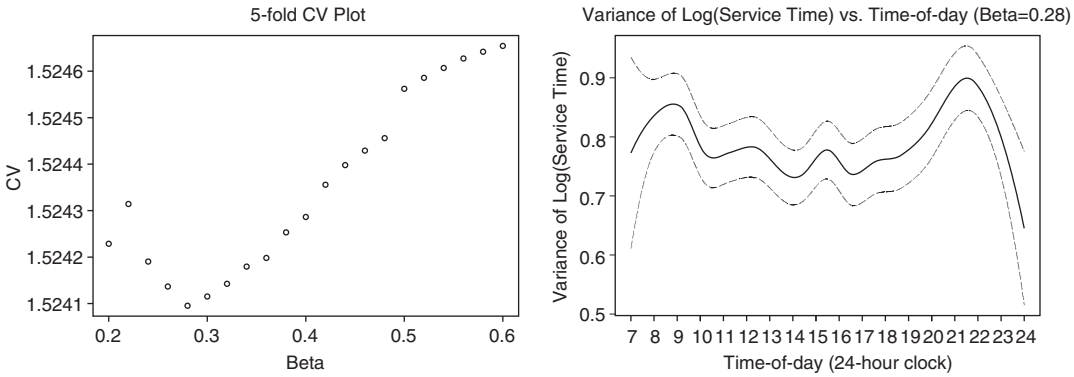

Figure 5. 5-fold Cross-validation for variance of Log(Service Time) (PS).

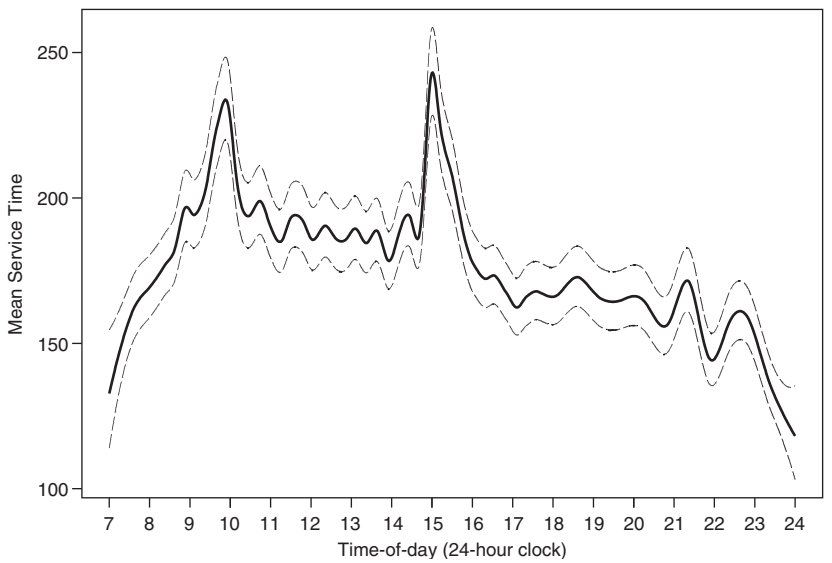

Figure 6. Mean Service Time (PS) vs. time-of-day.

Figures 4 and 5 are the building blocks for Figure 6, which graphs the mean service times of PS calls across time-of-day. Since PS calls constitute 68.4% of all the calls, the pattern is very similar to the one for all calls.

From Figure 6, we can see that the mean service times are not constant across time-of-day, but range between 140 and 230 seconds and peak around 10:00AM and 3:00PM. Starting from the beginning of a day, they increase until their first peak just prior to 10:00AM, then decrease until 1:30PM, when they begin to increase again and reach the second peak around 3:00PM before they decrease again until 5:00PM. After that, they increase again until 6:00PM,

DOI: 10.1002/asmb

stay relatively constant afterward until 10:30PM, then decrease until the end of the day. Given the accompanying 95% confidence band, the bimodal pattern is statistically very significant. Call centre managers should take this into account while arranging staffing for the call centre.

We now empirically validate the lognormal assumption of the service times, conditioning on time-of-day, by looking at the residuals from the regression of Log(Service Time) on Time-of-day for these PS calls. The normal Q–Q plot of the residuals (*unshown here*) suggests that they are very close to being normal, validating our assumption of lognormality of the service times. Bandwidths selected from 10-fold cross-validation yield very similar results. One can also subjectively choose the bandwidths to generate interesting curves that are nearly free of extraneous wiggles, which turn out to be close to the automatically chosen bandwidths in this case.

5.2. *IN calls*

Due to the special nature of *Internet Assistance*, IN calls require special skills from the service agents. As such, the call centre provides a separate pool of agents to handle IN calls beginning in August. Figure 7 plots the mean service times across time-of-day along with a 95% confidence band. The bandwidths were chosen via 5-fold cross-validation.

The pattern of the mean service times is significantly different from the PS calls. The mean service times range between 350 and 450 seconds, much longer than the PS mean service times. There are some fluctuations within the day, but they may not be significant given the wide confidence band. We thus conclude that the IN mean service times do not change much over the course of a day. This might be an effect of the separate agent pool or the special service nature of the IN calls.

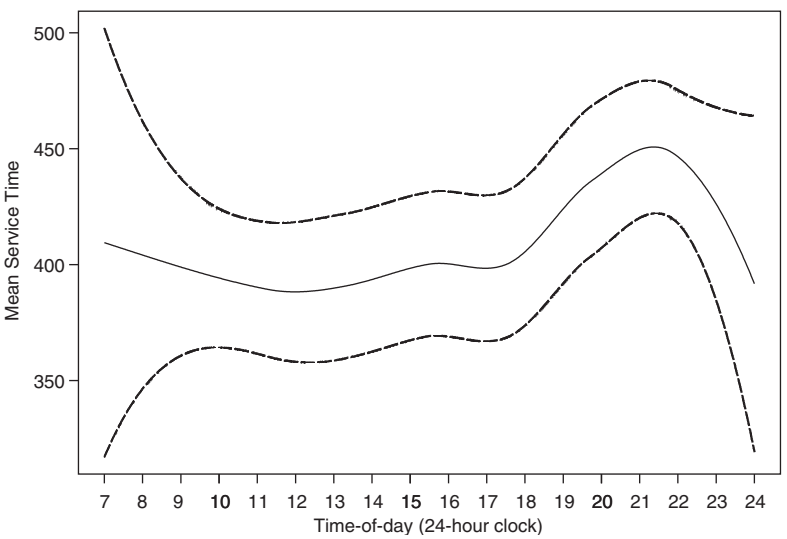

Figure 7. Mean Service Time (IN) vs. time-of-day.

DOI: 10.1002/asmb

6. CONCLUSIONS AND FUTURE WORK

This paper estimated the time-varying mean service times of calls at a bank call centre. The service times were shown to be approximately lognormally distributed. Motivated from this structure, we propose a new methodology for non-parametric (*heteroscedastic*) regression with lognormal errors, which also provides a pointwise confidence band. Local polynomial regression is employed in the procedure. The methodology is shown via a simulation study to have better performance than a naive approach. The method is applied to model the mean service time patterns of two types of calls served at the call centre. The results show that the mean service times may dramatically depend on time-of-day, and they differ significantly between different types of calls. These findings have important operational implications for call centre managers in terms of agent staffing and call routing.

The same methodology can be applied to other regression contexts where lognormal errors are involved. For example, Ingolfsson *et al*. [30] describe a data set of individual ambulance calls, where the ambulance travel time is shown to be lognormally distributed, conditioning on the distance between the ambulance station and the destination. The problem of interest is to estimate the mean travel time as a function of the travel distance. Such quantification is a necessary input for various planning models for ambulance deployment as well as for pricing ambulance service. Here, it is reasonable to assume that the mean function is monotonically increasing with distance, which is different from our call centre application. We intend to look into this in the future.

Our current method estimates $\mu(\cdot)$ and $\sigma^2(\cdot)$ separately. An alternative is to estimate them simultaneously. One idea is to approximate $\mu(\cdot)$ and $\ln(\sigma^2(\cdot))$ by spline functions. Since the responses are lognormally distributed, the corresponding likelihood function can be written down in a closed form, and maximum likelihood estimates for $\mu(\cdot)$ and $\sigma^2(\cdot)$ can be obtained using *Newton–Raphson* or *Fisher-scoring* methods. The accompanying confidence bands can be derived based on the asymptotic normality of the MLEs. Conceptually, there is no problem in extending this approach into a multivariate scenario. An investigation of this approach is currently under way. Alternatively, one could approximate $\mu(\cdot)$ and $\ln(\sigma^2(\cdot))$ locally using polynomials according to Taylor expansion, and maximize the corresponding likelihood function to obtain local likelihood estimates of $\mu(\cdot)$ and $\sigma^2(\cdot)$.

ACKNOWLEDGEMENTS

Thanks are due to the editor and a referee whose comments greatly improved the paper.

REFERENCES

1. Gans N, Koole G, Mandelbaum A. Telephone call centers: tutorial, review and research prospects. *Manufacturing and Service Operations Management* 2003; **5**:79–141.
2. Mandelbaum A. Call centers: research bibliography with abstracts. *Technical Report*, Technion, 2004.
3. Wolff RW. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall: Englewood Cliffs, NJ, 1989.
4. Mandelbaum A, Schwartz R. Simulation experiments with M/G/100 queues in the Halfin-Whitt (Q.E.D.) regime. *Technical Report*, Technion, 2002.
5. Shen H, Brown LD, Zhi H. Efficient estimation of log-normal means with application to pharmacokinetic data. *Statistics in Medicine* 2005 (in press).

DOI: 10.1002/asmb

6. Mandelbaum A, Sakov A, Zeltyn S. Empirical analysis of a call center. *Technical Report*, Technion, 2000.
7. Brown LD, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L. Statistical analysis of a telephone call center: a queueing science perspective. *Journal of the American Statistical Association* 2005; **100**:36–50.
8. Shen H. Nonparametric regression for problems involving log-normal distribution. *Ph.D. Dissertation*, University of Pennsylvania, 2003.
9. Bolotin VA. Telephone circuit holding time distributions. *Proceedings of the 14th International Teletraffic Congress Antibes Juan-Les-Pins, France*, 1994; 125–134.
10. Ulrich R, Miller J. Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology* 1993; **37**:513–525.
11. Van Breukelen GJP. Parallel information processing models compatible with lognormally distributed response times. *Journal of Mathematical Psychology* 1995; **39**:396–399.
12. Fan J, Gijbels I. *Local Polynomial Modelling and Its Applications*. Chapman & Hall: London, 1996.
13. Loader C. *Local Regression and Likelihood*. Springer: New York, 1999.
14. Fan J, Gijbels I. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaption. *Journal of the Royal Statistical Society, Series B* 1995; **57**:371–394.
15. Ruppert D, Sheather SJ, Wand MP. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* 1995; **90**:1257–1270.
16. Jones MC, Marron JS, Sheather SJ. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* 1996; **91**:401–407.
17. Jones MC, Marron JS, Sheather SJ. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* 1996; **11**:337–381.
18. Ruppert D, Wand MP, Holst U, Hössjer O. Local polynomial variance-function estimation. *Technometrics* 1997; **39**:262–273.
19. Ruppert D. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association* 1997; **92**:1049–1062.
20. Levine M. Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: a possible approach. *Computational Statistics and Data Analysis* 2005 (in press).
21. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. Springer: New York, 2001.
22. Müller HG, Stadtmüller U. Estimation of heteroscedasticity in regression analysis. *Annals of Statistics* 1987; **15**: 610–625.
23. Hall P, Kay JW, Titterington DM. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* 1990; **77**:521–528.
24. Dette H, Munk A, Wagner T. Estimating the variance in nonparametric regression—what is a reasonable choice? *Journal of the Royal Statistical Society, Series B* 1998; **60**:751–764.
25. Levine M. Variance estimation for nonparametric regression and its applications. *Ph.D. Dissertation*, *University of Pennsylvania*, 2003.
26. Fan J, Yao Q. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* 1998; **85**:645–660.
27. Yu K, Jones MC. Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association* 2004; **99**:139–144.
28. Land CE. An evaluation of approximate confidence interval estimation methods for lognormal means. *Technometrics* 1972; **14**:145–158.
29. Shen H, Zhu Z. Efficient mean estimation in lognormal linear models. 2005 (in press).
30. Ingolfsson A, Budge S, Erkut E. Optimal ambulance location with random delays and travel times. *Management Science*, 2005 (in press).

DOI: 10.1002/asmb