On the Accuracy of Delay-History-Based Predictors in Large Call Centers

Mor Armony, New York University

Joint work with:
Achal Bassamboo, Northwestern University
Rouba Ibrahim, Columbia University
Avi Mandelbaum, Technion

MSOM 2010
Motivating Application: Delay Announcements

Modern Call Centers

- Large
- Uncertain and time-varying demand \(\Rightarrow\) inappropriate staffing
- Long waiting times (e.g., service-oriented call centers)
- Uncertainty about length of wait (invisible queues)

Delay Announcements

- Inexpensive
- Relatively easy to implement
- Improve quality of service
- Control congestion: **impact customer behavior**
Delay-History-Based Predictors

Exploit the recent history of delays in the system.

Advantages

- Do not rely on system parameters
- Robust
- Easy to interpret

Last-to-Enter-Service Predictor (LES)

- $w(t) =$ waiting time of the customer arriving at time $t$
- $\tau_t =$ arrival time of the LES customer at time $t$

\[ \theta_{LES}(t) \equiv w(\tau_t) \]
Single class/Single pool Model

- Arrival rate: $\lambda$
- Prob. of balking: $b(w)$ depends on the announced delay $w$
- Service rate: $\mu$
- Number of servers: $N$
- Abandonment rate: $\alpha(w)$ depends on the announced delay $w$
Our Work

**Question: How accurate is the LES delay predictor?**

**Part 1: Asymptotic Accuracy of LES**
- Abandonment: independent of the announcement
- Abandonment: dependent on the announcement
- Balking: dependent on the announcement

**Part 2: Empirical Study**
- Real-life call center data
- Accuracy of LES
- New delay-history-based predictors
Literature Review

- Delay announcements and their effect on system dynamics: Hassin (1986), Whitt (1999a), Armony & Maglaras (2004), Guo & Zipkin (2007), Armony et al. (2009), Allon et al. (2010a,b)
Part 1

Asymptotic Accuracy of LES
A Sequence of Systems (QED Asymptotic Regime)

\(N^{th}\) System:

- Service rate: \(\mu\)
- Number of Servers: \(N\)
- Arrival rate: \(\lambda^N = N\mu + O(\sqrt{N})\)
- Abandonment rate: \(\alpha^N(w)\)
- Balking probability: \(b^N(w)\)
Consider the case where $\alpha^N(w) \equiv \alpha$, $b^N(w) \equiv 1$

Waiting times are “small” (Garnett et al.)

$$w^N(t) = O\left(\frac{1}{\sqrt{\lambda^N}}\right)$$

Scaled queue length is diffusion

Scaled queue length: almost constant between arrival and departure

Snapshot principle: $\sqrt{N}w^N \approx \frac{1}{\mu} \frac{Q^N}{\sqrt{N}}$ (Puhalskii).
Theorem

As the system size increases,

\[ \sqrt{N}|w(t^N) - w(\tau_t^N)| \Rightarrow 0, \text{ for all } t > 0. \]

WT based on LES \( \approx \) Actual WT

\[ \Uparrow \]

\( LES \) is asymptotically correct
Announcement dependent abandonment behavior

- Consider the case where $\alpha^N(w) \in [\alpha_1, \alpha_2]$, $b^N(w) \equiv 1$

- We show,

$$w^N(t) = O\left(\frac{1}{\sqrt{\lambda^N}}\right)$$

- Scaled queue length may/may not be a diffusion!

- Queue length: almost constant between arrival and departure
Theorem

As the system size increases,

$$\sqrt{N}|w(t^N) - w(\tau_t^N)| \Rightarrow 0, \text{ for all } t > 0.$$  

WT based on LES $\approx$ Actual WT

$LES$ is asymptotically correct
Sketch of the Proof

Bounding argument

- First show that

\[ |\tau_t^N - t| \to 0, \text{ as } N \to \infty. \]

- Consider two systems initialized at time \( \tau_t^N \):
  - System I: Abandonment rate is \( \alpha_1 \)
  - System II: Abandonment rate is \( \alpha_2 \).

- We can construct such that

\[ Q_{\text{Sys II}}(t + s) \leq Q(t + s) \leq Q_{\text{Sys I}}(t + s) \text{ for all } s \geq 0. \]

- \( Q_{\text{Sys}} \) converge to diffusion processes
General distribution for abandonment time

- If announcement is \( w \), then the abandonment time of the customer has distribution \( F_w \).
- We assume that the hazard rate of \( F_w \) is uniformly bounded from above and below.
- Similar bounding argument holds.
- \( LES \) announcements are asymptotically accurate.
System with Balking

- Customer balks with probability $b(w)$ if given announcement $w$

- If $b(w)$ is fixed and is of $O\left(\frac{1}{\sqrt{N}}\right)$
  - Diffusion limit holds
  - Snapshot principle implies asymptotic accuracy

- For general $b(w)$, under technical conditions LES is asymptotically accurate
Part 2

Statistical Analysis of Call Center Data
Description of the Data

Call Center of a US Bank

- Large call center:
  - 900-1200 agents on weekdays
  - 200-500 agents on weekends
- Multiple sites: NY, PA, RI, and MA
- Routing: skill-based, across sites
- Up to 300,000 calls/day
- Types of services: Retail, Premier, Business, Consumer Loans, Online Banking, and Telesales

Data Set

- Single customer class: Telesales
- 7769 calls registered over two weekdays:
  - 05/22/2003 (3654 calls)
  - 05/28/2003 (4115 calls)
- Working hours: 7 AM - midnight
- Around 50 agents (time-varying)
Summary Statistics

Wait = time until either entry to service or abandonment.

<table>
<thead>
<tr>
<th>05/22/2003 (in secs)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average wait</td>
<td>17</td>
</tr>
<tr>
<td>Std dev. of wait</td>
<td>62</td>
</tr>
<tr>
<td>Average positive wait</td>
<td>43</td>
</tr>
<tr>
<td>Proportion of delayed customers</td>
<td>38%</td>
</tr>
<tr>
<td>Proportion of abandonment</td>
<td>3.0%</td>
</tr>
<tr>
<td>75th percentile</td>
<td>5.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>05/28/2003 (in secs)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average wait</td>
<td>24</td>
</tr>
<tr>
<td>Std dev. of wait</td>
<td>83</td>
</tr>
<tr>
<td>Average positive wait</td>
<td>54</td>
</tr>
<tr>
<td>Proportion of delayed customers</td>
<td>45%</td>
</tr>
<tr>
<td>Proportion of abandonment</td>
<td>4.0%</td>
</tr>
<tr>
<td>75th percentile</td>
<td>14</td>
</tr>
</tbody>
</table>

- Large variance ⇒ overall average is not a reliable predictor
- Need to use information about current system state
Peaks correspond to a decrease in the number of available agents (e.g., lunch break around 1pm)

- Fluctuations $\Rightarrow$ errors in delay-history-based predictions
Quantifying Accuracy

Sample Bias ($B$)

$$B \equiv \frac{1}{k} \sum_{i=1}^{k} (p_i - d_i)$$

- $p_i =$ delay prediction for customer $i$
- $d_i =$ measured delay for customer $i$ ($d_i > 0$)
- $k =$ sample size

Average Squared Error ($ASE$)

$$ASE \equiv \frac{1}{k} \sum_{i=1}^{k} (p_i - d_i)^2 .$$

We consider $\sqrt{ASE}$.
### Biased LES Prediction

Time unit = 1 second.

<table>
<thead>
<tr>
<th>Date</th>
<th>$B(\text{LES})$</th>
<th>$\sqrt{\text{ASE(LES)}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>05/22</td>
<td>-18</td>
<td>99</td>
</tr>
<tr>
<td>05/28</td>
<td>-7.5</td>
<td>160</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date</th>
<th>$\frac{B(\text{LES})}{\text{Avg. Wait}}$</th>
<th>$\frac{\sqrt{\text{ASE(LES)}}}{(\text{Avg. Wait})}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>05/22</td>
<td>-0.42</td>
<td>2.3</td>
</tr>
<tr>
<td>05/28</td>
<td>-0.15</td>
<td>3.0</td>
</tr>
</tbody>
</table>

**Problem:** Announce $LES = 0$ to delayed customers.

- **05/22:** 51% of $LES$ announcements $= 0$
- **05/28:** 58% of $LES$ announcements $= 0$
Head-of-Line Predictor (HOL)

- $w_H = \text{elapsed delay of HOL customer}$

$\theta_{HOL}(w_H) \equiv w_H$

The HOL announcement is positive if there is an HOL customer. Otherwise, announce LES.
Accuracy of LES Predictor
M. Armony
A. Bassamboo
R. Ibrahim
A. Mandelbaum

Introduction
Asymptotic Results
Independent abandonment
Dependent abandonment
Balking
Empirical Study
Description of Data
Accuracy of LES
New Predictors
Conclusions

Performance of Predictors

Time unit = 1 second.

<table>
<thead>
<tr>
<th></th>
<th>HOL</th>
<th>LES</th>
<th>(LES + HOL)/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B$</td>
<td>-0.34</td>
<td>-18</td>
<td>-9.1</td>
</tr>
<tr>
<td>$\sqrt{ASE}$</td>
<td>120</td>
<td>99</td>
<td>98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>HOL</th>
<th>LES</th>
<th>(LES + HOL)/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B$/ Avg. Wait</td>
<td>-0.0078</td>
<td>-0.42</td>
<td>-0.21</td>
</tr>
<tr>
<td>$\sqrt{ASE}$/Avg. Wait</td>
<td>2.7</td>
<td>2.3</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Estimates on 05/22/2003

We corrected for the bias but large variance remains.
New Predictor Based on the Traffic Intensity

Refined LES Predictor ($\text{LES}_r$)

- $t_L = \text{arrival time of LES customer}$
- $t_C = \text{arrival time of current customer}$
- $w_H = \text{elapsed delay of HOL customer}$
- $w_L = \text{delay of LES customer}$
- $\rho(t) = \text{estimate of traffic intensity at time } t$

$$\theta_{\text{LES}_r} \equiv \frac{\rho(t_C)}{\rho(t_L)} \times \frac{w_H + w_L}{2}$$
Frame of Reference

No-Information Predictor (NI)
- Uses no information about current system state
- Announces average waiting time

Queue-Length-Based Predictor (QL)
- $n =$ queue length upon arrival
- $s =$ number of agents upon arrival
- $m =$ average service time upon arrival

\[ \theta_{QL}(n) \equiv (n + 1) \times \frac{m}{s} \]
## Performance Conditional on the Level of Delay

Time unit = 1 second.

<table>
<thead>
<tr>
<th>Delays</th>
<th>( B )</th>
<th>( \sqrt{ASE} )</th>
<th>( NI )</th>
<th>( QL )</th>
<th>( LES )</th>
<th>( LES_r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>smaller than 30 (71%)</td>
<td>0</td>
<td>7.7</td>
<td>0</td>
<td>11</td>
<td>6.8</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \sqrt{ASE} )</td>
<td></td>
<td></td>
<td>52</td>
</tr>
<tr>
<td>in (30, 120) (20%)</td>
<td>0</td>
<td>25</td>
<td>-66</td>
<td>70</td>
<td>-26</td>
<td>-23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \sqrt{ASE} )</td>
<td></td>
<td></td>
<td>69</td>
</tr>
<tr>
<td>larger than 120 (9%)</td>
<td>0</td>
<td>62</td>
<td>-274</td>
<td>336</td>
<td>-205</td>
<td>-185</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \sqrt{ASE} )</td>
<td></td>
<td></td>
<td>270</td>
</tr>
</tbody>
</table>

- \( LES_r \) is more accurate than \( LES \) for long delays
- Error in \( LES_r \) prediction: large bias remains
Autocorrelation Function

Autocorrelation Function (ACF) for waits on 05/28/03

Suggests averaging over several past delays.
Concluding Remarks

We studied the accuracy of LES

▶ Easy to implement
▶ Needs no info. regarding system parameters

Asymptotic Results

▶ It is asymptotically accurate in the QED regime with abandonments/balking
▶ The result holds even if there is no diffusion approximation!

Empirical Results

▶ Problem: Large variance of delays, significant time-variability
▶ LES has significant prediction error: bias + variance
▶ New delay-history-based predictors: significant bias remains
▶ Time-series analysis approach seems promising