Queues with Impatient Customers

Service Engineering

Queueing Science

Eurandom

September 8, 2003

e.mail : avim@tx.technion.ac.il

Website: http://ie.technion.ac.il/serveng

2. Supporting Material

Gans, Koole, and M.: “Telephone Call Centers: Tutorial, Review and Research Prospects.” *MSOM*.

Brown, Gans, M., Sakov, Shen, Zeltyn, Zhao: "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective." Submitted.

Garnett, M. and Reiman: "Designing a Call Center with Impatient Customers." *MSOM*.

M. and Zeltyn: "The Impact of Customer Patience on Delay and Abandonment: Some Empirically-Driven Experiments with the M/M/N+G Queue." Submitted.

Zeltyn: Ph.D. thesis, in progress.

Palm: "Intensitatsschwankungen im fernsprechverkehr," (In English) Ericsson Technics, 1943.

Palm: "Methods of judging the annoyance caused by congestion." Tele, 1953: Recommended.

Baccelli and Hebuterne: "On queues with impatient customers." In Performance '81, ed. Gelenbe, 1981.

Contents

Call Centers = Queue with Impatient Customers

12 Years History, or A Modelling Spectra

1. Kella, Meilijson: Practice ⇒ Abandonment important
2. Shimkin, Zohar: No data ⇒ Rational patience in Equilibrium
3. Carmon, Zakay: Cost of waiting ⇒ Psychological models
4. Garnett, Reiman: Palm/Erlang-A to replace Erlang-C/B
   as the standard Steady-state model
5. Massey, Reiman, Rider, Stolyar: Predictable variability ⇒
   Fluid models, Diffusion refinements
6. Ritov, Sakov, Zeltyn: Finally Data ⇒ Empirical models
7. Brown, Gans, Haipeng, Zhao: Statistics ⇒ Queueing Science
8. Garnett, Atar, Reiman: Skills-based routing ⇒ Control models
9. Nakibly, Meilijson, Pollatchek: Prediction of waiting ⇒
   Online Models and Real Time Simulation
10. Garnett: Practice ⇒ 4CallCenters.com
11. Zeltyn: Queueing Science ⇒ Empirically-Based Theory
12. Borst, Reiman, Zeltyn: Dimensioning M/M/N+G

QED Q's: State of Art (8/2003)

1. GI/**M**/N $\quad N \approx R + \beta\sqrt{R}, \quad \beta > 0$

   - Conceptual: Erlang; **Halfin-Whitt**

   - Dimensioning: Borst, Reiman

2. Abandonment (Erlang-A, with $-\infty < \beta < \infty$)

   - Conceptual: Garnett, Reiman; Zeltyn; **Whitt**

   - Dimensioning: (Borst, Reiman, Zeltyn) in progress

3 Time-Varying (Non-homogenous Poisson arrivals)

   - Infinite-server heuristics: Jennings, Massey, Whitt

   - Conceptual: (Massey, Rider) in progress

   - Dimensioning: ?

4 Skills-Based Routing:

   - Conceptual: Atar, Reiman; Gurvich (V-Model)

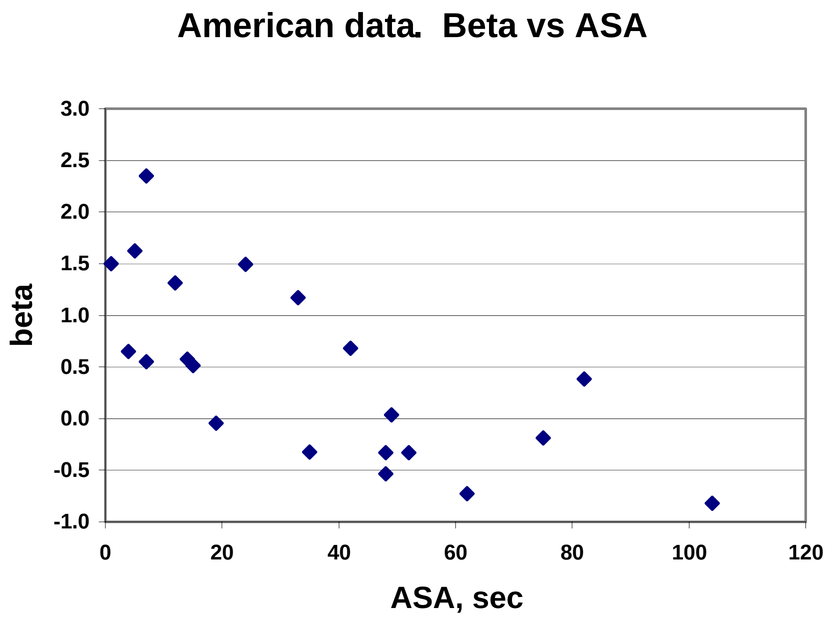
   - Dimensioning: **Borst, Seri** (General); Gurvich (V);

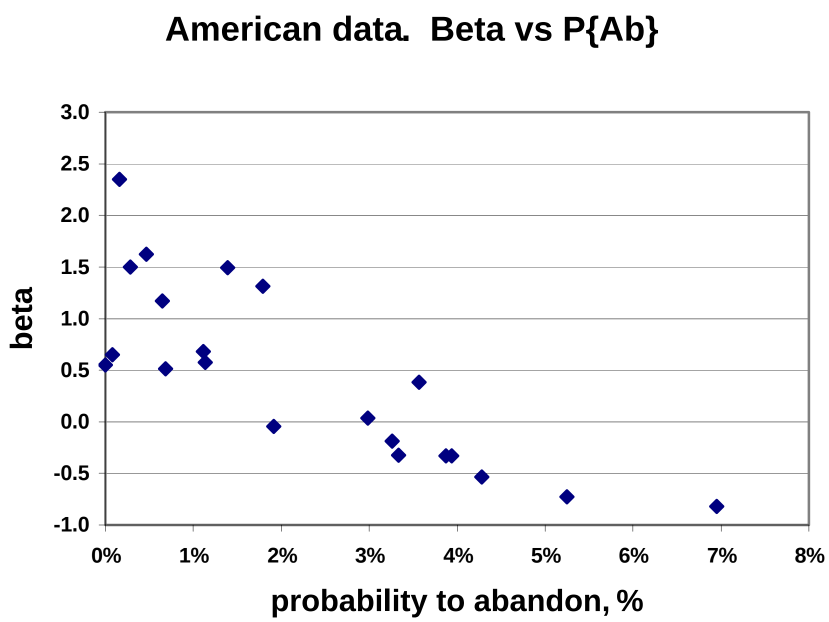
     Armony (Reversed-V)**;**

5. Service-Time Duration:

   - Conceptual: Whitt H2*/G; Jelenkovic, Momcilovic D

Empirical Service Grade (Beta)





QED Relevance: Wharton CC-Forum

6/13/00 - Tue

Time	Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
Total	**20,577**	**19,860**	**~3.0%**	**30**	**307**	**95.1%**	**85.4%**	**222.7**	**234.6**	**95.0%**
8:00	332	308	7.2%	27	302	87.1%	79.5%	59.3	66.9	88.5%
8:30	653	615	5.8%	58	293	96.1%	81.1%	104.1	111.7	93.2%
9:00	866	796	8.1%	63	308	97.1%	84.7%	140.4	145.3	96.6%
9:30	1,152	1,138	1.2%	21	303	90.8%	81.6%	211.1	221.3	95.4%
10:00	1,330	1.286	3.3%	22	307	98.4%	84.3%	223.1	229.0	97.4%
10:30	**1,364**	**1,338**	**1.9%**	**33**	**296**	**99.0%**	84.1%	**222.5**	227.9	97.6%
11:00	1,380	1,280	7.2%	34	306	98.2%	84.0%	222.0	223.9	99.2%
11:30	1,272	1,247	2.0%	44	298	94.6%	82.8%	218.0	233.2	93.5%
12:00	1,179	1,177	0.2%	1	306	91.6%	88.6%	218.3	222.5	98.1%
12:30	**1,174**	**1,160**	**1.2%**	**10**	**302**	**95.5%**	93.6%	**203.8**	209.8	97.1%
13:00	**1,018**	**999**	**1.9%**	**9**	**314**	**95.4%**	91.2%	**182.9**	187.0	97.8%
13:30	1,061	961	9.4%	67	306	100.0%	88.9%	163.4	182.5	89.5%
14:00	1,173	1,082	7.8%	78	313	99.5%	85.7%	188.9	213.0	88.7%
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	86.0%	**206.1**	220.9	93.3%
15:00	1,137	1,122	1.3%	15	320	96.9%	83.5%	205.8	222.1	92.7%
15:30	1,169	1,137	2.7%	17	311	97.1%	84.6%	202.2	207.0	97.7%
16:00	1,107	1,059	4.3%	46	315	99.2%	79.4%	187.1	192.9	97.0%
16:30	914	892	2.4%	22	307	95.2%	81.8%	160.0	172.3	92.8%
17:00	615	615	0.0%	2	328	83.0%	93.6%	135.0	146.2	92.3%
17:30	420	420	0.0%	0	328	73.8%	95.4%	103.5	116.1	89.2%
18:00	49	49	0.0%	14	180	84.2%	89.1%	5.8	1.4	416.2%

Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios (E(S) = 3:45):

λ/hr	N	OCC	ASA	% Wait ≤ 2 sec
1599	100	99.9%	59:33	1%
1599	105	95.2%	0:23	51%
1600	100	100%	infinity	0%
		BUT	with	Impatience
				%Abandonment
1600	100	97.3%	0:23	2.7 %
1600	95	98.4%	0:23	6.5%
1800	105	97.7%	0:23	3.4%

QED with **(Im)patient** Customers:

The "fittest" survive and wait less – much less!

Erlang-A: Erlang-C with Exponential Patience / **A**bandonment

Downloadable implementation: 4CallCenters(.com)

Theorem (with Garnett and Reiman '02; Zeltyn '03):

Consider a sequence of M/M/**N**+G models, **N=1,2,3,…**

Then the following **points of view** are equivalent:

- **QED** $\quad \%\{\text{Wait} > 0\} \approx \alpha, \qquad 0 < \alpha < 1\,;$

- **Customers** $\quad \%\{\text{Abandon}\} \approx \dfrac{\gamma}{\sqrt{N}}\,, \qquad 0 < \gamma;$

- **Agents** $\quad \text{OCC} \approx 1 - \dfrac{\beta}{\sqrt{N}} \qquad -\infty < \beta < \infty;$

- **Managers** $\quad N \approx R + \beta\sqrt{R}\ , \quad R = \lambda \times \text{E(S)}$ not small;

QED performance (ASA, ...) is easily computable, all in terms of β (the square-root safety staffing level) – see later.

Covers also the Extremes:

$\alpha = 1\ :\ \ \text{N} = \text{R} - \gamma\ \text{R}$ **Efficiency-driven**

$\alpha = 0\ :\ \ \text{N} = \text{R} + \gamma\ \text{R}$ **Quality-driven**

Economics: $\sqrt{\cdot}$ Safety-Staffing

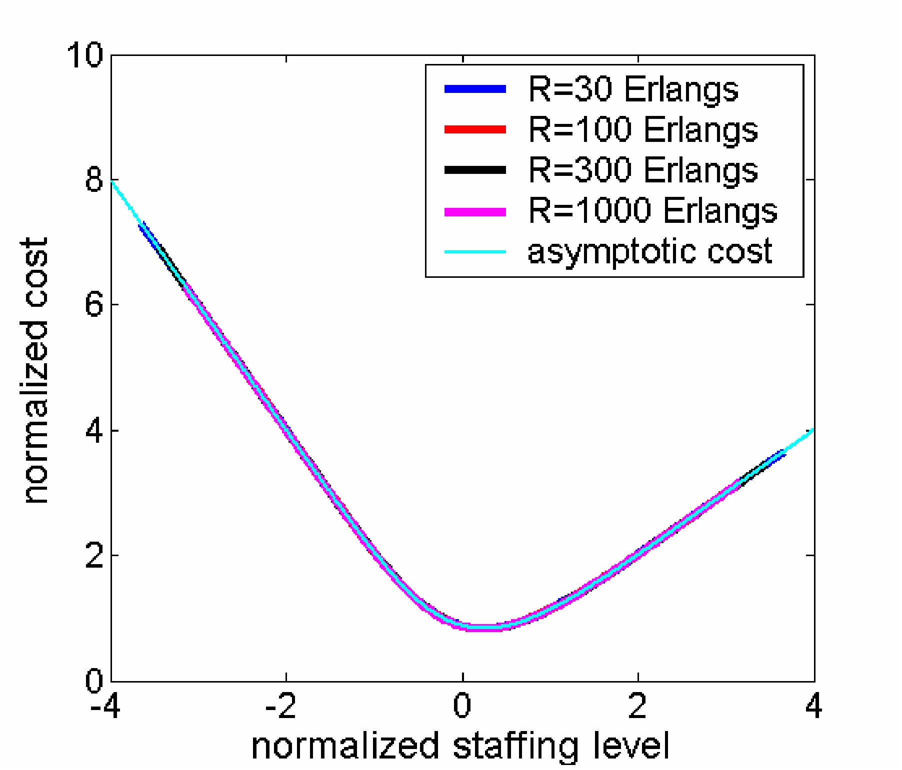
Cost $= c \cdot N + d \cdot \lambda E W_q$ (costs: c-staffing, d-delay).

Optimal staffing level:

$$N^* \approx R + y^*(r;s) \cdot \sqrt{R}, \qquad r = \frac{d}{c}, \quad s = \sqrt{\frac{\mu}{\theta}} .$$

$$y^* = \underset{-\infty<y<\infty}{\arg\min} \left\{ c \cdot y + d \cdot \frac{P(y;s)}{y} \cdot ys \cdot [h(ys) - ys] \right\}.$$

Numerical tests exhibit **remarkable** accuracy:
Actual cost function "coincides" with asymptotic cost.



Normalized staffing level = $(N - R) / \sqrt{R}$,

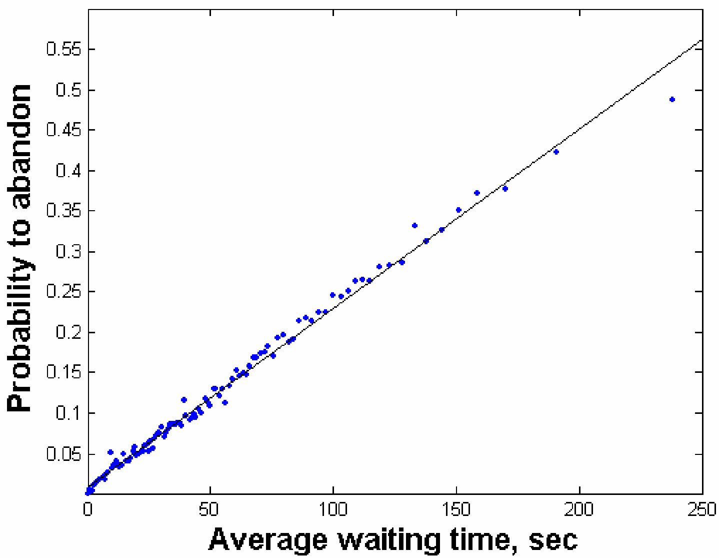
Normalized cost $= (\text{Cost} - cR) / \sqrt{R}$.

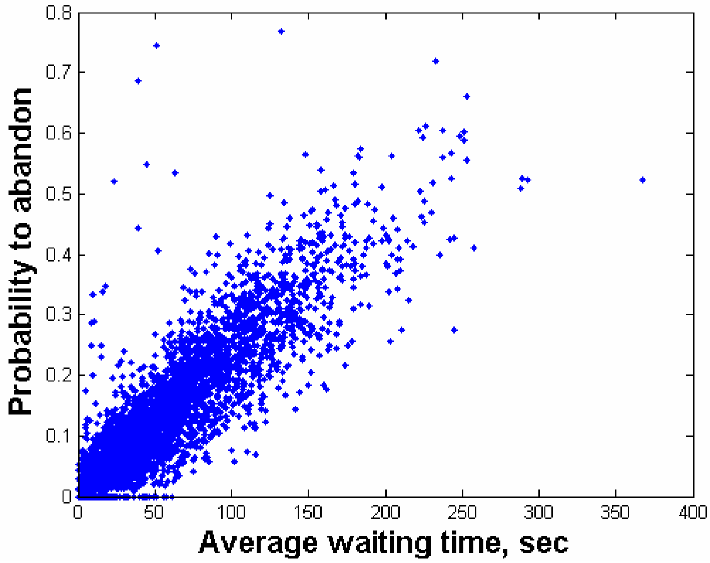
Asymptotic cost: $c \cdot y + d \cdot \frac{P(y;s)}{y} \cdot ys \cdot [h(ys) - ys]$.

Empirically-Based Theory

1999 Call Center Data – individual calls.

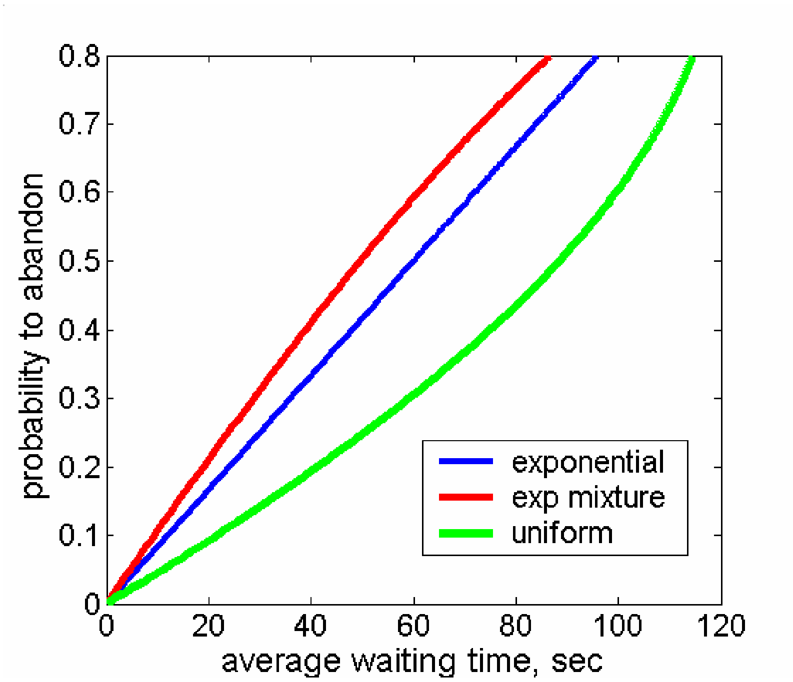
Linear pattern observed: P{Abandon} = $C \bullet$ E[Wait]



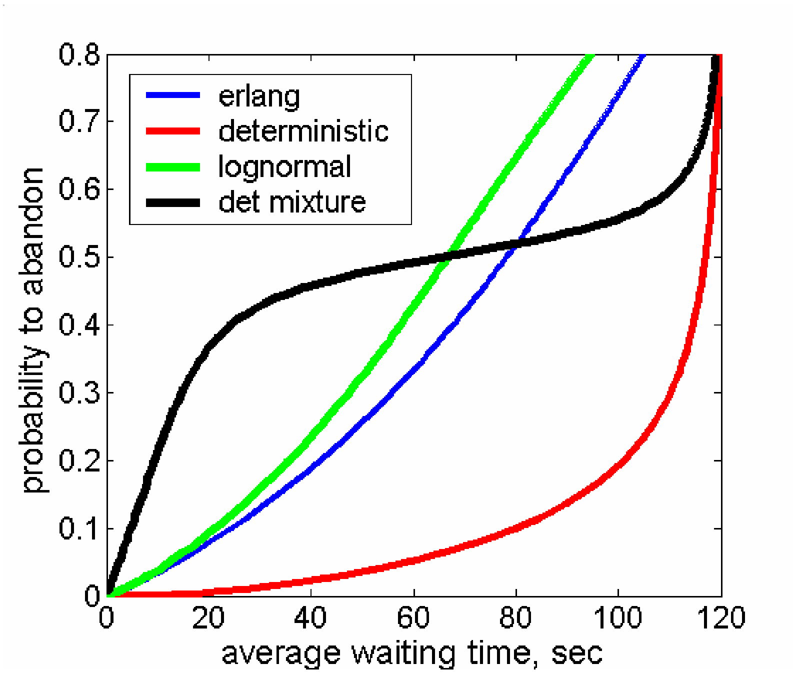


Theoretical Relations

Linear (practically speaking)

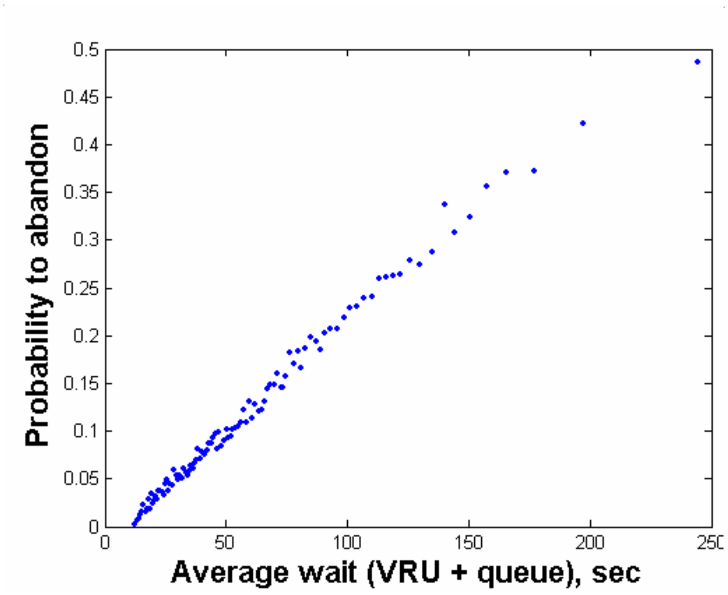


Non-linear

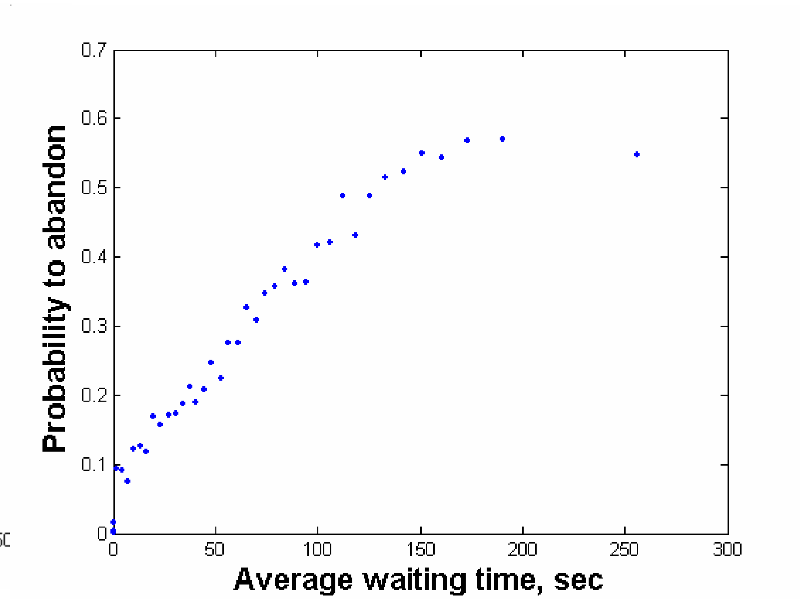


Human Behavior

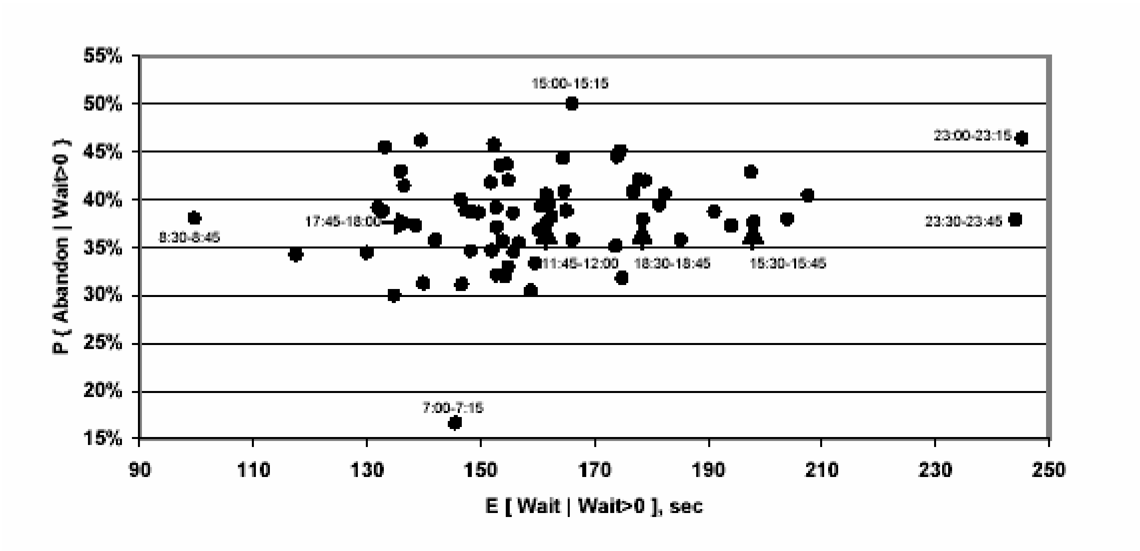
Delayed Abandons (IVR)

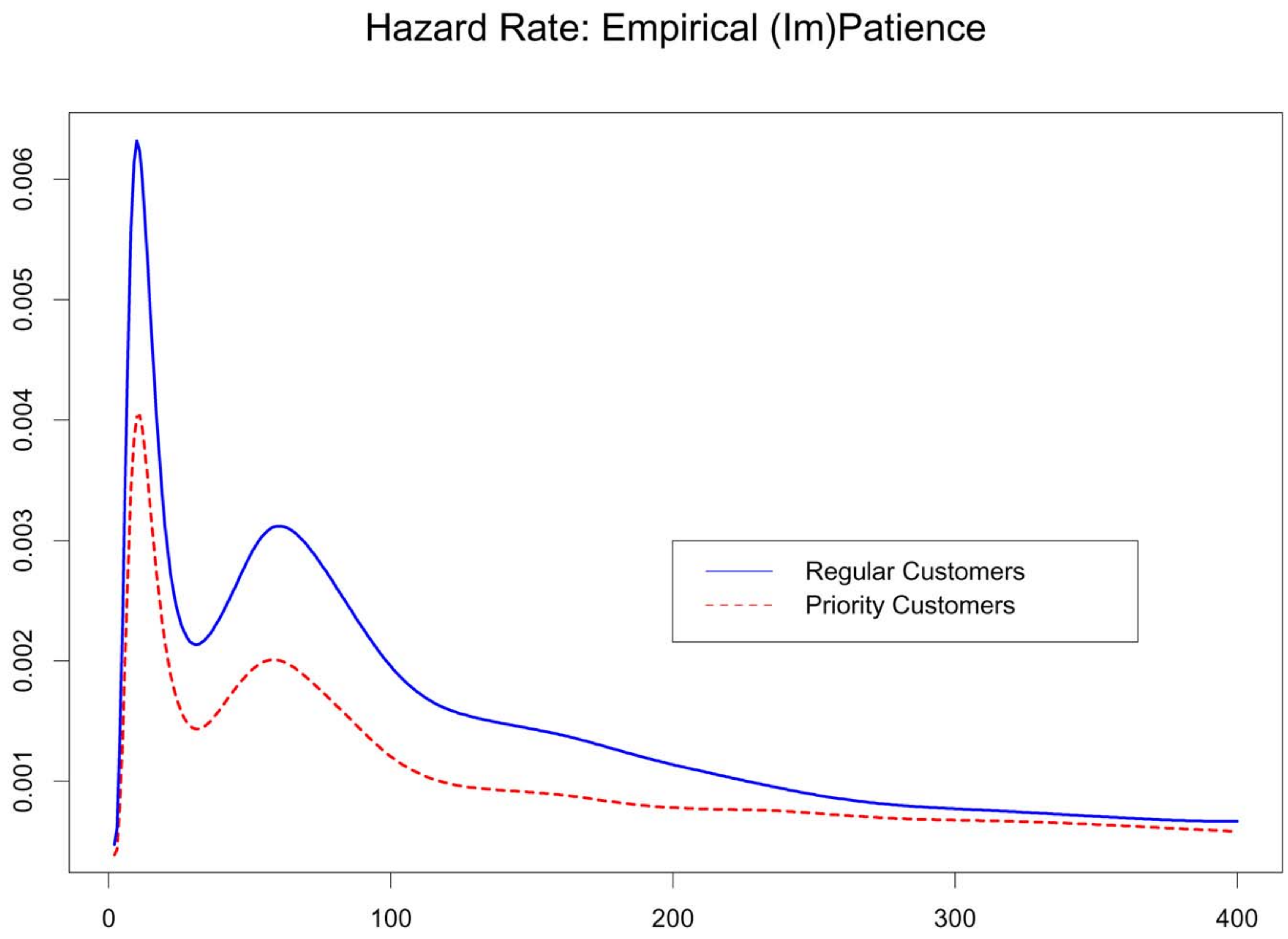


Balking (New Customers)



Learning (Internet Customers)




Hazard Rate: Empirical (Im)Patience
Regular Customers
Priority Customers
0.001
0.002
0.003
0.004
0.005
0.006
0
100
200
300
400

13. Fitting a Simple Model to a Complex Reality