Modeling the Customer Role in Services: Examples from Call Centers

Zeynep Akşin, Koç University
B. Ata, C. Su, University of Chicago
S. Emadi, University of North Carolina
O. Jouini, Y. Dallery, Ecole Centrale Paris
E. Güneş, L. Örmeci, Koç University
H. Özden, Amazon
Services ..

• ... are settings where the process is the product
• ... are provided to customers and cannot be produced independently of them
• ... are produced, distributed and consumed simultaneously
• ... mismatch between demand and capacity is inherent
Traditional queuing models

Passive and mostly homogeneous
Traditional queuing models versus queues where customers react

Passive and mostly homogeneous
Traditional queuing models versus queues where customers react

Passive and mostly homogeneous

Strategic
Makes choices
Reacts to system features
Can be heterogeneous
Overview

• Customer reactions to waiting
  – Patience in queues
  – Delay announcement practices aimed to control patience

• Customer reactions to sales attempts
  – Customer purchase
  – Cross-selling practices aimed to control customer purchase
Understanding and Controlling Customer Patience

Setting: Call Centers
A bank call center: daily number of calls and abandonments
M/M/N+M: Erlang-A Model

Arrivals

Agents

Abandonments
Phone queues with abandonments / reneging

- Baccelli and Hebuterne (1981)
- Brandt and Brandt (1997, 1999)
- Akşin and Harker (1999)
- Garnett, Mandelbaum, Reiman (2002)
Understanding abandonments correctly makes a difference!

From Table 2 Garnett, Mandelbaum and Reiman, 2002

<table>
<thead>
<tr>
<th></th>
<th>$M/M/N$</th>
<th>$M/M/N + M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction abandoning</td>
<td>—</td>
<td>3.1%</td>
</tr>
<tr>
<td>Average speed of answer</td>
<td>20.8 sec.</td>
<td>3.6 sec.</td>
</tr>
<tr>
<td>Waiting time’s 90th percentile</td>
<td>58.1 sec.</td>
<td>12.5 sec.</td>
</tr>
<tr>
<td>Average queue percentile</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Agents’ utilization</td>
<td>96%</td>
<td>93%</td>
</tr>
</tbody>
</table>
Patience modeling

- Assume a patience time distribution
- Estimate distribution parameter(s) using data on observed abandonments
- Issues:
  - Data is heavily censored
  - Call center data is aggregated in 15 min intervals
  - Observed patience for a given system
First documented use of individual call data
Brown et al. (2005)- hazard rate
Abandonment is not independent from queue design

Firm Decisions:
- Staffing
- Priority Policy

......

Callers' chances of receiving service

Performance measures

Callers' abandonment behavior
The traditional approach to treating abandonments

Data under the current policy

Waiting time distribution

Patience Time Distribution

The new policy

Prediction

Exogenous model
The new approach which enables one to find the impact of policy changes

Data under the current policy

Endogenous model

Callers’ parameters

Waiting time distribution

The new policy

Callers’ decision making model

Callers’ abandonment behavior

Prediction

Aksin, Ata, Emadi and Su (2013)
Decision Making Rational Customers in Observable Queues (Naor, 1969)

Service has a value $R$

Waiting has a cost $c$

Utility = $R - c \times \text{Wait}$

Join

Don’t Join

Service rate $\mu$

Waiting has a cost $c$

Arrival rate $\lambda$
Rational Customers in Queues

• Huge literature-Hassin and Haviv (2003) To queue or not to queue
  – Seminal paper by Naor (1969)
  – Observable or unobservable
  – Homogenous vs heterogeneous customers
  – Linear versus non-linear waiting costs
  – Etc.

• Rational abandonments
  – Hassin (1995)- reward drops to zero after a threshold wait
  – Mandelbaum and Shimkin (2000)- a queue with a fault state
  – Shimkin and Mandelbaum (2004)-non-linear waiting costs
  – Etc.
The typical OR/OM framework

ASSUME

\[ R, c \]

FORMULATE

Maximize Utility(\( R, c \))

Maximize E(Profit(\( Cu, Co \)))

ANALYZE

Rational Join, Balk, Wait, Abandon decisions

Optimal order quantity decisions
Structural Estimation

OBSERVE FROM DATA

Join, Balk, Wait, Abandon decisions
Order quantity decisions

FORMULATE A DECISION MAKING MODEL

Maximize Utility(R, c, others)
Maximize E(Profit(Cu, Co, others))

INFER

R, c
Cu, Co

Callers are forward looking and make decision dynamically.
Callers’ Parameters

- $r_i$ : Reward from receiving service
- $c_i$ : Cost incurred by waiting for a unit of time

Heterogeneity of the callers:

$$r_i = \exp(m_r + \sigma_r y_{1i})$$
$$c_i = \exp(m_c + \sigma_c y_{2i})$$

$y_{1i}$ and $y_{2i}$ have independent normal distributions.

The structural parameters to be estimated:

$$\Theta = (m_r, m_c, \sigma_r, \sigma_c)$$
Callers’ Actions and Utilities

\[ u(t, r_i, c_i, \varepsilon_{it}, d_{it}) = v(t, r_i, c_i, d_{it}) + \varepsilon_{it}(d_{it}) \]

\[ d_{it} = \begin{cases} 
1 & \text{Abandoning} \\
0 & \text{Waiting} 
\end{cases} \]

\( \varepsilon_{it}(d_{it}) \) random shock to the utility with type-I extreme value distribution.
Callers’ Nominal Utilities

- Nominal utility from abandoning:

\[ v(t, r_i, c_i, 1) = 0 \]

- Nominal utility from waiting:

\[ v(t, r_i, c_i, 0) = -c_i + \pi(t) r_i + (1 - \pi(t)) \mathbb{E} \left[ \max_{d \in \{0, 1\}} u(t + 1, r_i, c_i, \varepsilon_{i(t+1)}(d), d) \right] \]

- Probability of receiving service
- The integrated value function

\[ V(t, r_i, c_i) \]
The model

- The optimal action of a caller:

\[
d_{it} = \arg \max_{d \in \{0,1\}} u(t, r_i, c_i, \varepsilon_{it}, d)
\]

- Probability of abandoning:

\[
P_{it}(1; r_i, c_i) = \frac{1}{1 + \exp(-c_i + \pi(t)r_i + (1 - \pi(t))V(t, r_i, c_i))}
\]

- The integrated value function:

\[
V(t, r_i, c_i) = \log \left(1 + \exp(-c_i + \pi(t+1)r_i + (1 - \pi(t+1))V(t+1, r_i, c_i))\right)
\]

\[
V(T, r_i, c_i) = 0
\]
Data

- The focus of our analysis is on calls
  - In the Retail service group.
  - Received during the working days during weeks without holidays between 9 a.m. and 2 p.m.
  - Entered the system through VRU and proceeded to wait in the queue
  - Having normal termination, transfer or abandonment as an outcome.
  - Their maximum waiting time is less than 960 seconds.

<table>
<thead>
<tr>
<th>Priority group</th>
<th>Number of observations</th>
<th>Abandonment rate</th>
<th>Average waiting time(s)</th>
<th>Maximum waiting time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High priority</td>
<td>184,722</td>
<td>2.12 %</td>
<td>18.83</td>
<td>857</td>
</tr>
<tr>
<td>Medium priority</td>
<td>516,685</td>
<td>3.68 %</td>
<td>42.19</td>
<td>958</td>
</tr>
<tr>
<td>Low priority</td>
<td>253,963</td>
<td>6.66 %</td>
<td>72.02</td>
<td>949</td>
</tr>
<tr>
<td>No priority</td>
<td>367,701</td>
<td>24.65 %</td>
<td>96.20</td>
<td>960</td>
</tr>
<tr>
<td>Sum</td>
<td>1,323,071</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Estimation

- Kaplan Meier to estimate the probability of service from the data
- MLE to estimate model primitives

<table>
<thead>
<tr>
<th>Priority group</th>
<th>$r$-Mean ($)</th>
<th>$c$-Mean ($$/\text{minute})$</th>
<th>$r$-St.Dev.</th>
<th>$c$-St.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Priority</td>
<td>6.309</td>
<td>1.067</td>
<td>4.52E-05</td>
<td>3.09E-05</td>
</tr>
<tr>
<td>Medium Priority</td>
<td>6.175</td>
<td>0.506</td>
<td>4.56E-05</td>
<td>2.76E-05</td>
</tr>
<tr>
<td>Low Priority</td>
<td>5.299</td>
<td>5.45E-04</td>
<td>3.02E-05</td>
<td>5.91E-07</td>
</tr>
<tr>
<td>No Priority</td>
<td>4.211</td>
<td>0.122</td>
<td>0.645</td>
<td>2.057</td>
</tr>
</tbody>
</table>
Counterfactual illustrating the importance of modeling patience endogenously
Reversed strict priority policy

1. No priority callers
2. Low priority callers
3. Medium priority callers
4. High priority callers
Reversed strict priority policy

<table>
<thead>
<tr>
<th>Reversed strict priority policy</th>
<th>High priority</th>
<th>Medium priority</th>
<th>Low priority</th>
<th>No priority</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sec.</td>
<td>% Ab.</td>
<td>sec.</td>
<td>% Ab.</td>
</tr>
<tr>
<td>Endogenous model</td>
<td>89.06</td>
<td>62.98</td>
<td>41.08</td>
<td>2.45</td>
</tr>
<tr>
<td>Exogenous model</td>
<td>397.79</td>
<td>57.66</td>
<td>39.55</td>
<td>4.12</td>
</tr>
</tbody>
</table>

- Exogenous model overestimates the waiting time for the high priority group.
- In the endogenous model: for callers with large waiting cost, service quality ↓ → abandonment probability ↑
- In the exogenous model: for callers with large waiting cost, service quality ↓ → abandonment probability (no change)
Patience Control: Delay announcement with Customer Reaction

Firm Decisions:
  - Staffing
  - Priority Policy

Callers' chances of receiving service

Performance measures

Callers' abandonment behavior
Delay announcement problem challenges

How to model individual behavior?

Performing transient queueing analysis when individual level model is embedded in a queue with abandonment

Estimating parameters

Announcement → Individual Patience Behavior → Balking Abandonment → System Performance

estimation
Early solution by Whitt (1999)

Assume random patience threshold

Simplify by ignoring abandonment reactions; threshold only affects balking decisions

Estimation not considered
Guo and Zipkin (2007)

- Assume utility maximizing customers
- Simplify by ignoring abandonment reactions

- Announcement
- Individual Patience Behavior
- Balking Abandonment
- System Performance

- Estimation
- Estimation not considered
Armony, Shimkin and Whitt (2009): Include customer reactions in the form of abandonments

- Assume an exogenously specified different abandonment rate
- Simplify through heavy traffic analysis

- Announcement → Individual Patience Behavior → Balking → Abandonment → System Performance

- Estimation not considered

- No modeling of individual behavior
Jouini, Aksin, Dallery (2011): individual customer patience reaction is modeled

Assume random patience threshold adjusts after announcement

Simplify by assuming aggregate abandonment remains exponential

Announcement → Individual Patience Behavior → Balking Abandonment → System Performance

estimation

Estimation not considered
Modeling reneging behavior

New arrival

\[ p^B(n) = p(T < d_n) \]

\[ 1 - p^B(n) \]

\[ t_i \]

\[ d_n = F_{Xn}^{-1}(\beta) \]

patience threshold = \( \theta t_i + (1 - \theta) d_n \)
New reneging behavior: analysis

\[ p^B(n) = p(T < d_n) \]

\[ d_n = F_{X_n}^{-1}(\beta) \]

- \( r_n = r_n(\beta) \), probability that a customer who elects to wait initially will renege

- \( r_n \) is the conditional probability that the realization of his waiting time \( D_n \) exceeds his patience threshold, given that he elects to join the queue

\[ r_n = P(D_n > \text{patience threshold} \mid T > d_n) \]

patience threshold = \( \theta t_i + (1 - \theta) d_n \)
New reneging behavior: analysis

- We denote by $\lambda^R$ the stationary flow of abandoning customers.

- On the one hand, we have due to exponential reneging: $\lambda^R = \gamma' L_q$.

- On the other hand, we have using PASTA: $\lambda^R = \sum_{n=0}^{\infty} \lambda (1 - \alpha_0) (1 - p^B(n)) p(s + n) r_n$.

- All quantities are functions of $\gamma'$. Then, $\gamma'$ is a point mapped to itself by a given continuous function.

- Using a fixed-point algorithm we can compute $\gamma'$. 

\[ \text{New arrival} \quad \begin{array}{c} p^B(n) \quad \text{1 - } p^B(n) \\ t_i \end{array} \rightarrow \lambda^R \]
Aksin, Ata, Emadi, Su (2014): estimation, reaction, queueing combined

Utility maximization: optimal stopping problem under delay announcement

Simplify by making Markovian approximations

Announcement → Individual Patience Behavior

estimation

Balking Abandonment → System Performance

Estimation from data
Understanding and Controlling Customer Relationships

Setting: Cross-selling in bank call centers
The motivating idea

• Customers have relationships with service firms
• Relationship dynamics can be modeled as Markov chains
• Firm actions may influence customer behavior
• Changes in customer behavior may in turn affect firm decisions
• Idea: modeling firm decisions in the presence of customer reactions
Related Literature

Descriptive models of customer relationship
- Schmittlein, Morrison and Colombo, 1987
- Schmittlein and Peterson, 1994
- Netzer, 2004

Optimizing customer equity
- Ho, Park and Zhou, 2005
- Rust, Lemon and Zeithaml, 2004
- Venkatesan and Kumar, 2004
- Ching, Wong, Altman, 2004
- Sun and Li, 2005
- Sun, Li and Zhou, 2006

Markovian Life Time and Arrival Process

Endogeneity problem (Rust and Chung, 2006)
Customer Reactions to Cross-selling

Try to Cross-sell?

Customer reaction

“This new loan option is exactly what I need!”
+ $$€€$

• (+) Retention: Marple and Zimmerman, 1999
• (+) Reduce churn: Kamakura et al. 2003

“I don’t want another sales pitch, just transfer the money!”
Lost time, annoyance

• (-) Switch: Kamakura et al. 2003
## Cross selling in the literature

<table>
<thead>
<tr>
<th>When to cross-sell, Which product? - relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Paas and Kuijlien (2001)</td>
</tr>
<tr>
<td>- Li, Sun and Wilcox (2005)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>When to cross-sell, Which product? - system congestion</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Aksin and Harker (1999)</td>
</tr>
<tr>
<td>- Ormeci and Aksin (2006)</td>
</tr>
<tr>
<td>- Byers and So (2007)</td>
</tr>
<tr>
<td>- Armony and Gurvich (2006)</td>
</tr>
<tr>
<td>- Gurvich et al. (2006)</td>
</tr>
</tbody>
</table>

**Sequential offers**

**congestion vs. revenue**
Gunes, Aksin, Ormeci, Ozden (2010)
Two Features of a Customer Relationship

Customer Evolution:
Maturity increases the chances of buying the next product (Li et al. 2005, Kamakura et al. 1991)

Customer Reaction:
Excessive x-selling may irritate customers (Kamakura et al. 2003, Eichfeld 2006)

Customer reacts:
- lowers utilization of service? ($\lambda$)
- quits relationship earlier? ($\mu$)
- less inclined to accept future attempts? ($P_f$)
Overview of results from MDP analysis

• Optimal cross-sell policy is of threshold type with no customer evolution or reaction
• Optimal cross-sell policies are of state-dependent threshold type when only evolution or reaction is present
• When both evolution and reaction is present, policies can take complex dynamic forms
Modeling $P_f(i, j)$

Customers’ utility of not buying the proposed product relative to buying is $U(i,j)$

$$U(i,j) = \beta_0 + \beta_1 i + \beta_2 j + \varepsilon$$

$$P_f(i, j) = \frac{1}{1 + e^{-U(i,j)}}$$
How can $P_f(i,j)$ be estimated in practice?

- Using data from a bank
- 149 customers over a 2 year period

<table>
<thead>
<tr>
<th>Grand Total: Over All Customers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Grand total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Responses per Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Average</td>
</tr>
<tr>
<td>SD</td>
</tr>
<tr>
<td>Min.</td>
</tr>
<tr>
<td>Max.</td>
</tr>
</tbody>
</table>
A clustering analysis based on success/attempt ratio

<table>
<thead>
<tr>
<th>Cluster</th>
<th>N</th>
<th># of Successes (accept)</th>
<th># of Failures (reject or maybe later)</th>
<th>Total # of Attempts</th>
<th>Success/Attempt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>37</td>
<td>435</td>
<td>62</td>
<td>497</td>
<td>0.87</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>54</td>
<td>373</td>
<td>381</td>
<td>754</td>
<td>0.49</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>58</td>
<td>78</td>
<td>691</td>
<td>769</td>
<td>0.11</td>
</tr>
</tbody>
</table>
Estimation using random coefficient logit

- Dataset contains cross-sell attempt dates and outcomes
- We use time as a proxy for $j$
- The outcomes are used to determine $i$ - we count both reject and maybe later as failure

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter M</th>
<th>Parameter SD</th>
<th>Share &lt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
<td>0.4248 (243.2061)</td>
<td>0.0102 (0.0000)</td>
<td>0.0000</td>
</tr>
<tr>
<td>$j$</td>
<td>-0.0002 (0.0005)</td>
<td>0.0023 (0.0008)</td>
<td>0.5338</td>
</tr>
<tr>
<td>Constant</td>
<td>0.3482 (0.1883)</td>
<td>1.9292 (0.1801)</td>
<td>0.4299</td>
</tr>
</tbody>
</table>
What makes all of this exciting?

• Modeling of customer role in services is an interdisciplinary area
  – Economics framework
  – Behavioral framework: New models motivated by behavioral findings

• Choice models and structural estimation
  – Allows capturing strategic customers
  – Combines theoretical modeling with practice through estimation

• Data availability
  – Better field data
  – Lab experiments
Thank You