THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

GERAD
Groupe d'études et de recherche en analyse des décisions

Statistical Modeling of Call Centers

A Queueing-Science Perspective

Haipeng Shen and Noah Gans

Yong-Ping Zhou, and Genesys Telecommunications
Nan Liu, Avi Mandelbaum, Han Ye, and Genesys Telecommunications

NSF Grants DMS-0606577, CMMI-0800575 and CMMI-0800645

July 18, 2011

Outline

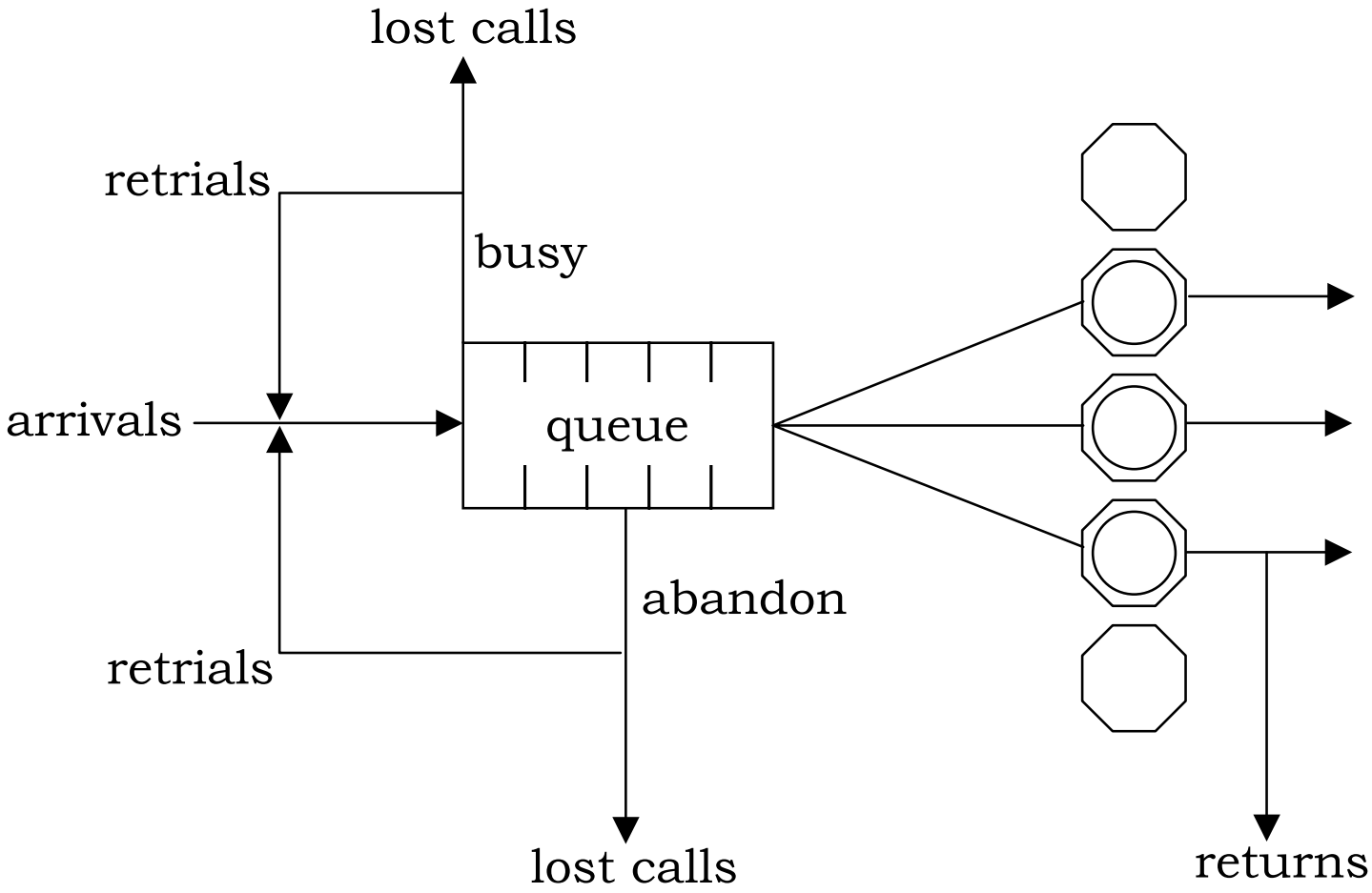
Outline

One of the early call centers ...

A more modern call center ...

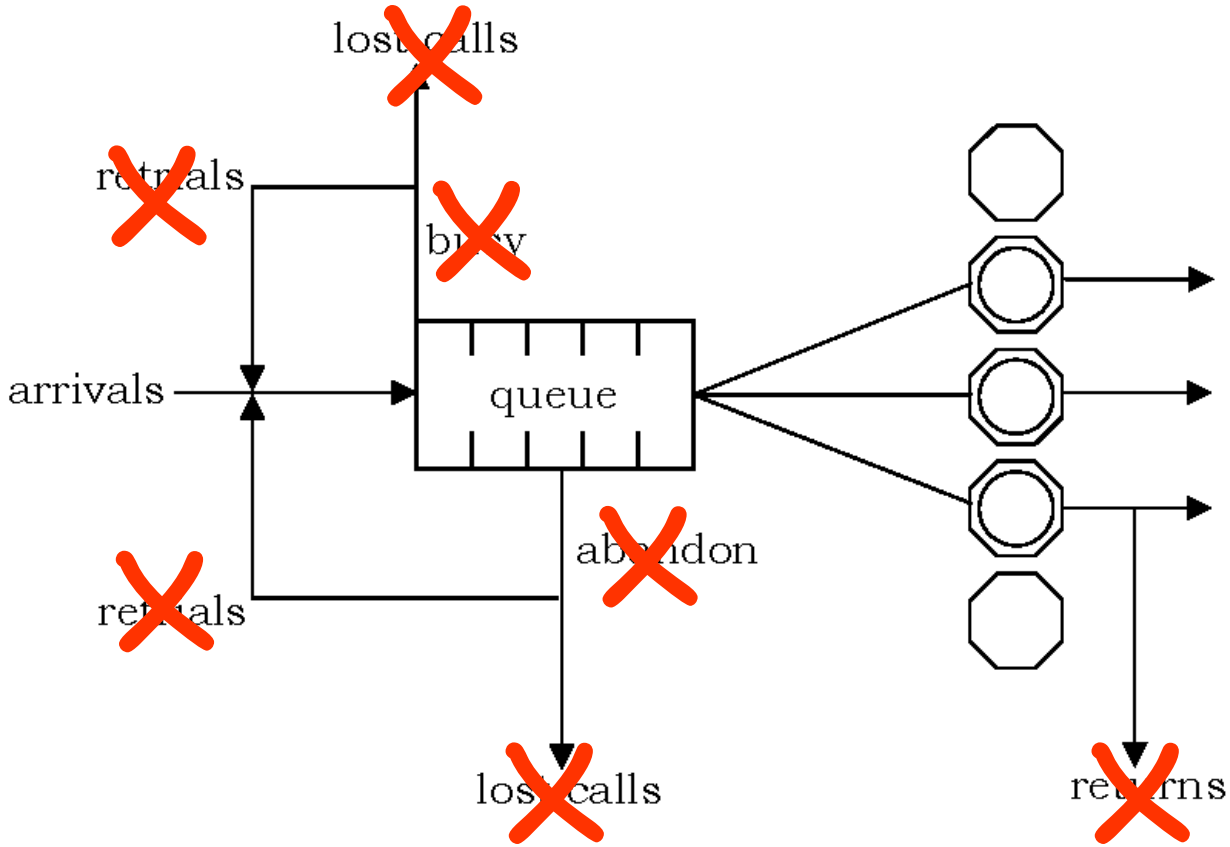
A sweatshop call center???

Queueing model for a single call center



Gans, Koole and Mandelbaum (2003)

The $M/M/N+\infty$ model or the Erlang-C model



- no blocking, abandonment, or retrials
- fixed arrival rate λ_i and service rate μ_i for time period i
- exponential inter-arrival and service times

Are the Erlang-C assumptions valid in call centers?

Brown, Gans, Mandelbaum, Sakov, Shen, Zeltyn and Zhao (2005)

One Israeli call center

- small: 15 seats at most
- multiple types of service:
  - regular banking service
  - stock-trading
  - IT support for online banking
- 450K agent-seeking calls in 1999

Are the arrivals Poisson?

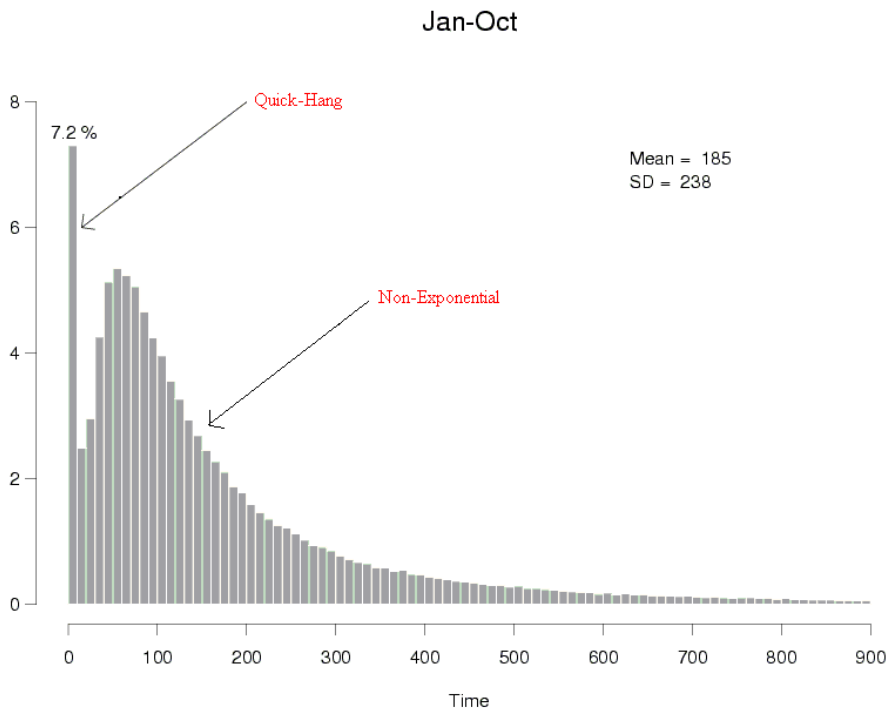
Consider short time intervals such as quarter hours,

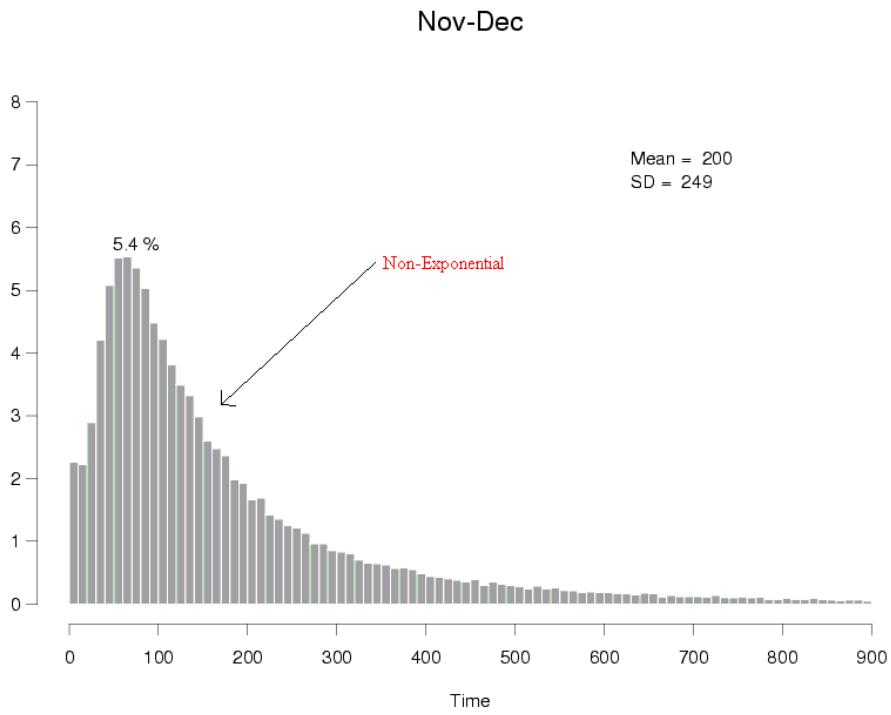
- Arrivals are Poisson with time-varying rates.
- The rates are still random given available covariates:
  - time-of-day, day-of-week, service types.
- Hence, a doubly stochastic Poisson process, or a Cox process.

- For one arrival stream, two-way dependence among the rates:
  - Inter-day dependence: today/tomorrow, weekly, . . .;
  - Intra-day dependence: morning/afternoon/night, ...
- Such dependence is crucial for inter-day rate forecasting and intra-day updating.

- For multiple arrival streams, inter-stream (i.e. call type) dependence.

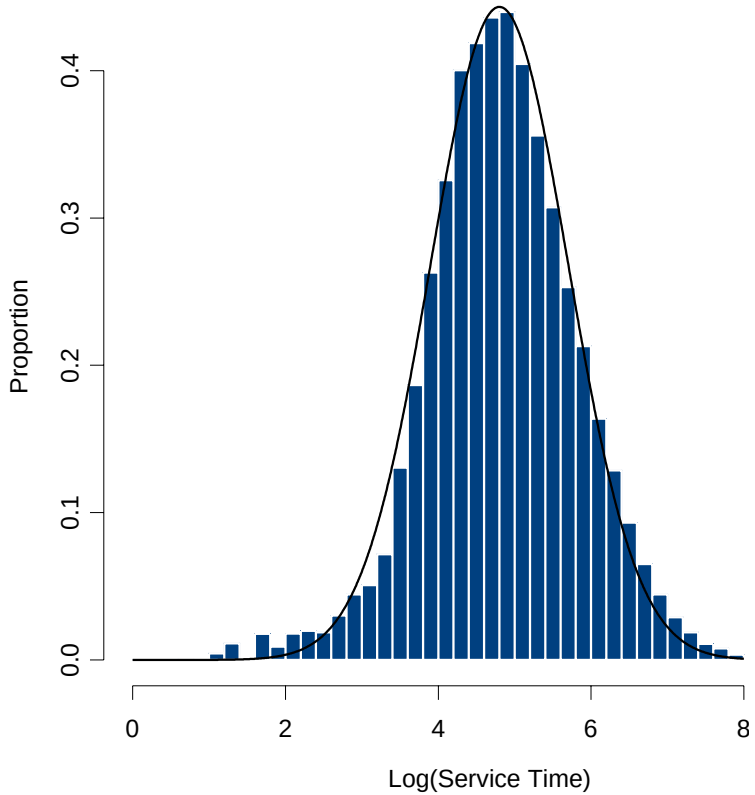
Are the service times exponential?





- quick-hang: agents hang up on customers
- non-exponential distribution

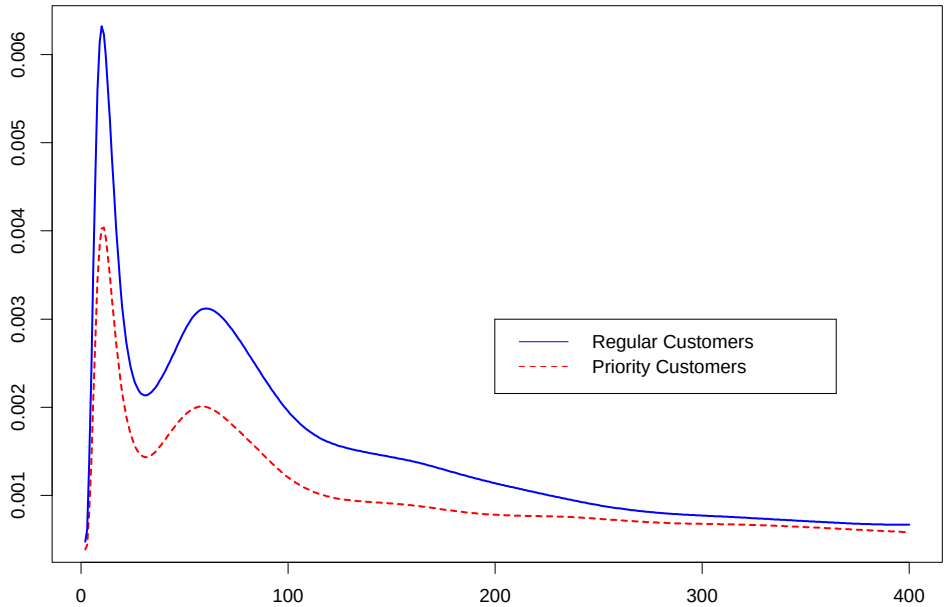
In fact, service times are lognormal



The lognormality

- exists at multiple levels: individual agents, different service types, time-of-day, ..., multiple companies (US/Israel)
- becomes handy later on for understanding agent heterogeneity

How about abandonment?



- x-axis: waiting time in queue
- y-axis: instantaneous probability of abandonment, (or hazard rate)

- bi-modal with an exponential tail
- priority customers are more patient
- Erlang-A: Garnett, Mandelbaum, Reiman (2002)
  - exponential time to abandonment, (or patience)

Outline

"Standard" model for call center capacity planning

1. Forecast offered load (e.g., by the 1/2-hour)

$$\{R_i = \lambda_i/\mu_i : i = 1, \ldots, m\}$$

2. Find minimum numbers of agents to make QoS constraint

$$s_i = \min\{s \mid \mathrm{P}\{\text{Delay} \leq \alpha\} \geq \beta\}$$

3. Find minimum cost assignment of agents to schedules

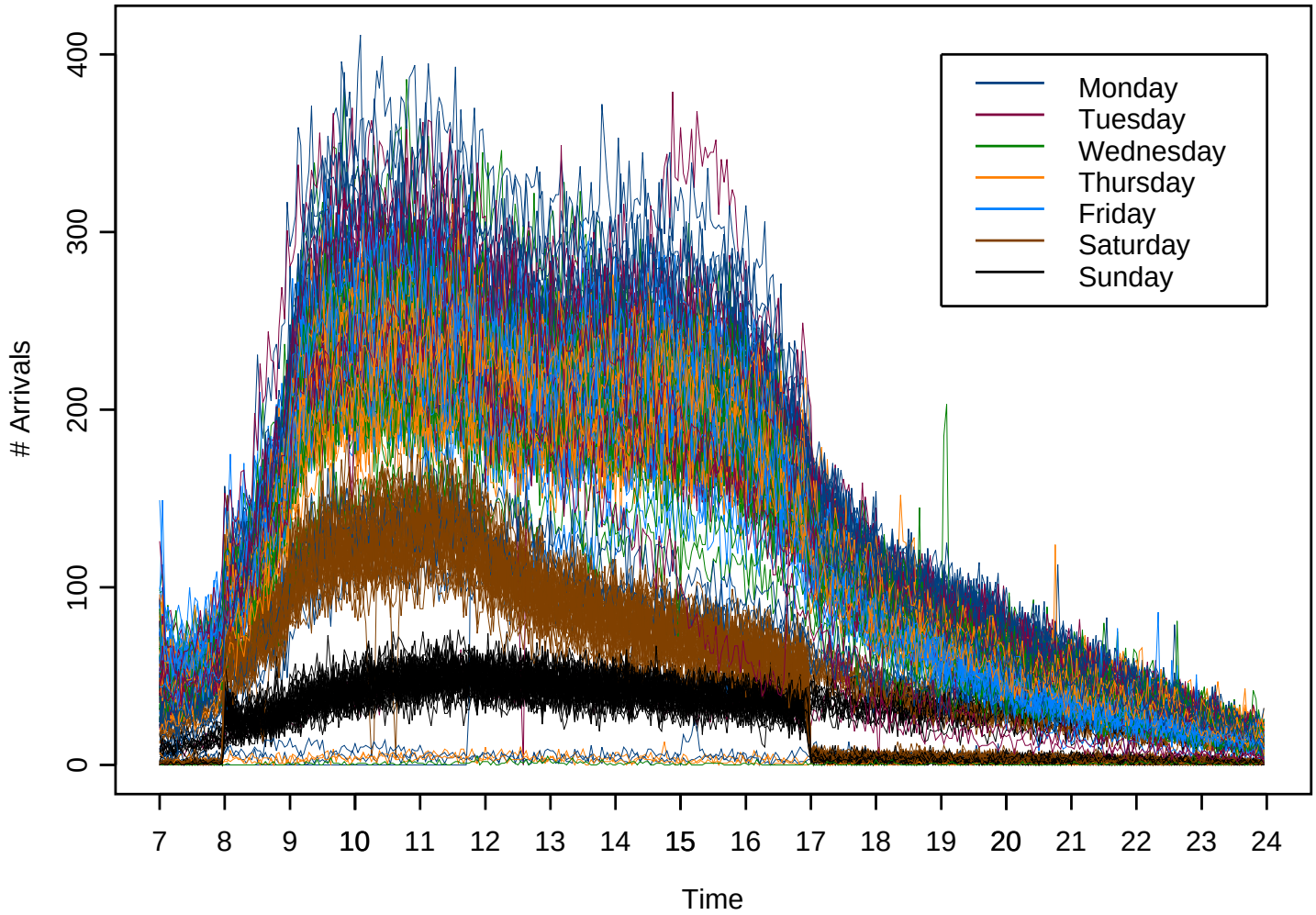
$$\min\{cy \mid Ay \geq s;\ y \geq 0;\ y \text{ integer}\}$$

Two complications we are currently working on

- The arrival rates are not certain: λ_i
  - we forecast them, and there are forecast errors
  - distributional forecasts may be better than point forecasts
  - solve stochastic programs that account for forecast distributions
  - one call type – Gans, Shen, and Zhou, with Genesys
  - multiple call types – Luedtke, Shen and Ye
- Service times are not i.i.d. random variables: μ_i
  - vary across agents
  - for a given agent, vary with experience
  - vary with other factors: type of call, time of day,...
  - preliminary data analysis – Gans, Liu, Mandelbaum, Shen, Ye
  - current analysis – Gans, Shen, and Ye, with Genesys
  - related hiring and retention problem – Arlotto, Chick, and Gans

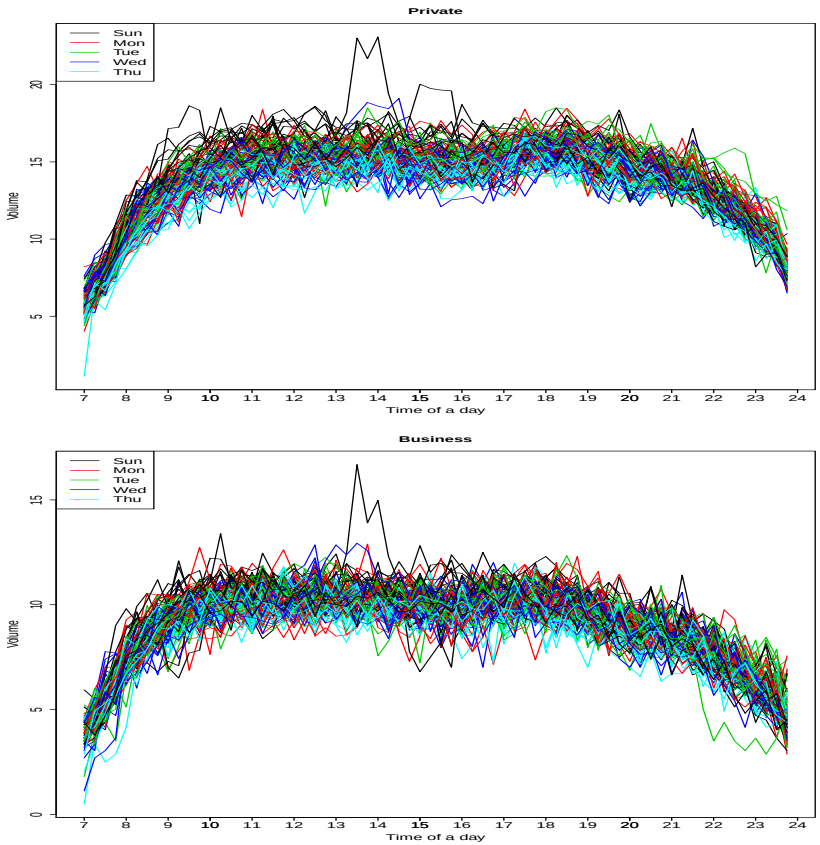
Outline

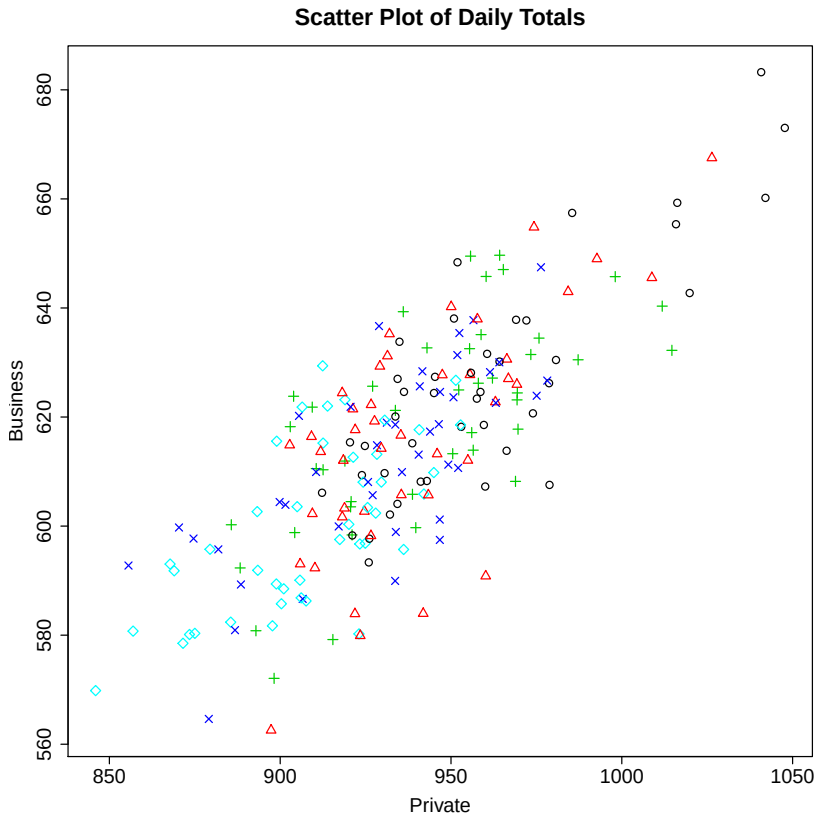
The arrival rate is not known with certainty



- Bank with a network of 4 call centers in northeast US
- 300K calls/day, 60K/day seeking agents, 1K agents in peak hours

Two arrival streams





- Israeli telecom company
- Two major arrival streams: Private (30%), Business (18%)

Work addressing arrival-rate uncertainty

- Acknowledgement that arrival-rate uncertainty affects performance
  - Whitt (99), Chen and Henderson (01), Jongbloed and Koole (01)
- Arrival-rate forecasting and updating methods
  - Avramides et al. (04), Brown et al. (05), Shen and Huang (05, 08), Steckley et al. (07), Taylor (07), Weinberg et al. (08), Aldor-Noiman et al. (09)
- Scheduling that accounts for arrival-rate uncertainty
  - Harrison and Zeevi (05), Whitt (06), Bassamboo, Harrison and Zeevi (05, 06, 07), Mehrotra and Ozluk (06), Bassamboo and Zeevi (07), Robbins and Harrison (07), Bertsimas and Doan (08), Gurvich et al. (09),

Our goal

- Develop distributional forecasts for arrival rates
  - Updating given additional information
- Perform stochastic scheduling using the distributional forecasts
  - Recourse actions after forecast updating
  - Changing staffing assignments
    - ⋆ send agents home early ... $\rightarrow$ reduce cost
    - ⋆ call in part-time agents ... $\rightarrow$ better achieve QoS measure
- Test the approach in large-scale real systems

Forecasting arrival rates using low-dimensional time series factor models

- Array of arrival counts N_{ij}
  - intervals within days, $i = 1, \ldots, m$
  - days $j = 1, \ldots, n$
  - $x_{ij} = \sqrt{N_{ij} + 1/4}$
- Intra-day feature vectors $f_1, \ldots, f_K \in R^m$ $(K \ll m)$
  - summarize intra-day call arrival patterns
  - reveal dominant intra-day arrival features
- View x_j as the composition of the K factors,

$$x_j = \beta_{j1} f_1 + \ldots + \beta_{jK} f_K + \epsilon_j$$

The factor model reduces the dimensionality of the forecasts

- Use in-sample data to find the best f_k's and β_j's

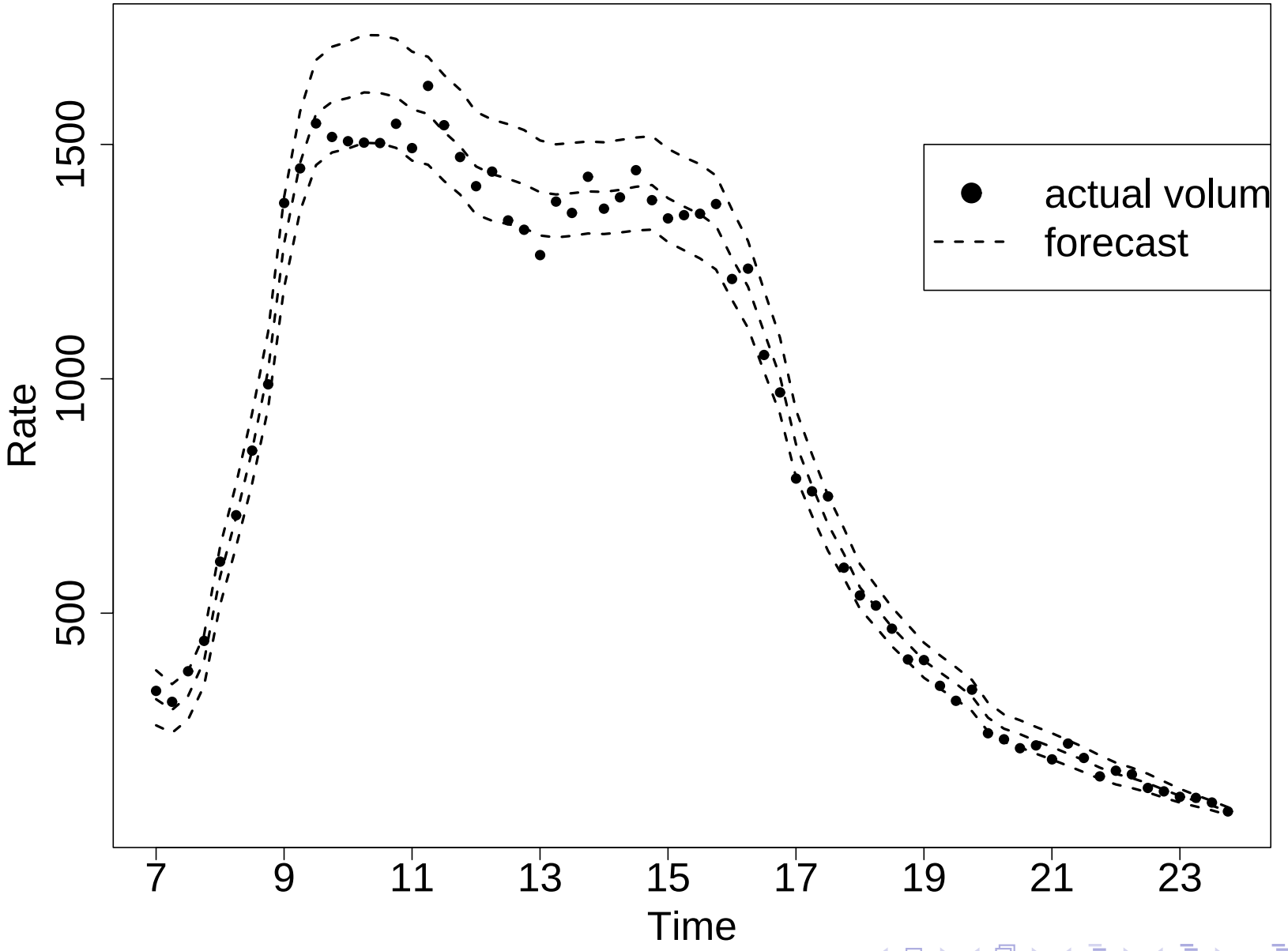
$$\min_{\substack{\beta_{j1},\ldots,\beta_{jK} \\ f_1,\ldots,f_K}} \sum_{j=1}^{n} \|\epsilon_j\|^2 = \min_{\substack{\beta_{j1},\ldots,\beta_{jK} \\ f_1,\ldots,f_K}} \sum_{j=1}^{n} \left\| x_j - (\beta_{j1} f_1 + \cdots + \beta_{jK} f_K) \right\|^2$$

  - Require $f_k' f_\ell = \delta_{k\ell}$ for identifiability

- Use in-sample β_j's to forecast out-of-sample β_j's
  - β_j's capture inter-day time series dependence

- Combine (forecasted) out-of-sample β_j's with in-sample f_k's to forecast future arrival rates/volumes

Use β_j's for distributional arrival-rate forecasts



Stochastic program for scheduling agents

- Distribution of the Λ_i's determined from the forecast
  - may be driven by a parametric forecast on the β_j's
  - may be non-parametric forecast using bootstrap

- With distributions for Λ_i's, solve the stochastic program

$$\min \{cy\}$$

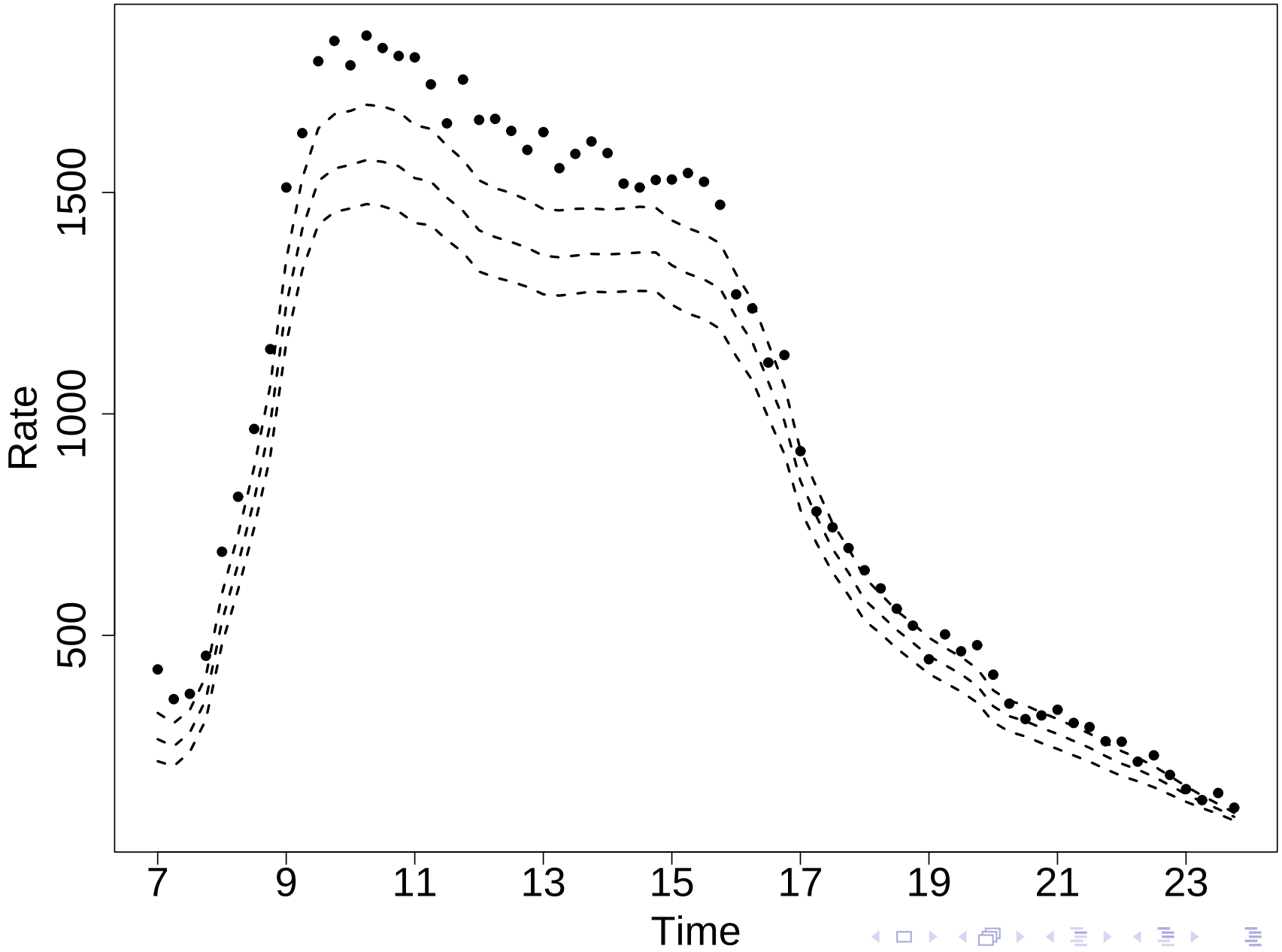
s.t.

$$\sum_{i=1}^{m} \mathsf{E}_{\Lambda_i} [f(\Lambda_i, a_i y)] \leq \alpha^* \sum_{i=1}^{m} \mathsf{E} [\Lambda_i] \qquad i = 1, \ldots, m$$

$$y \geq 0; \quad y \text{ integer},$$

where $f(\Lambda_i, a_i y)$ is the expected abandonment count in period i

Night-before forecasts can sometimes be off



Setup for intra-day forecast updates

- Let N_i be the call volume during the ith upcoming time interval
- Example: one-dimensional model with factor $f \in R^m$

$$\begin{aligned} N_i &\sim \text{Poisson}(\Lambda_i), \\ X_i &\equiv \sqrt{N_i + 1/4} \sim \text{N}(\Theta_i, \sigma_0^2) \\ \Theta_i &= \sqrt{\Lambda_i} = wf_i, \text{where } w \sim \text{N}(\mu, \sigma^2). \end{aligned}$$

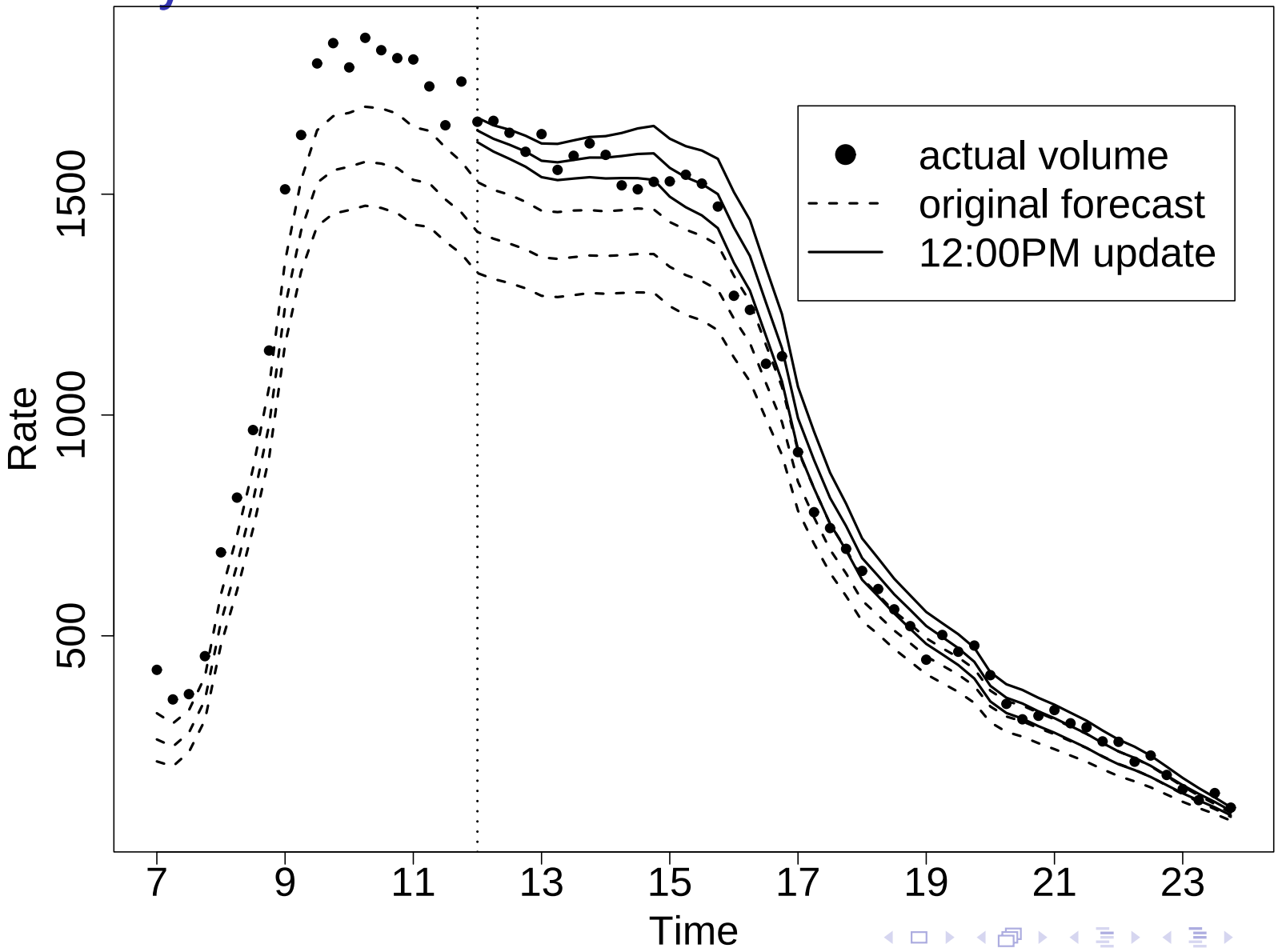
- Suppose each day consists of an early and late part
  - early intervals $1, \ldots, m_0 \Rightarrow$ counts X^e and unobserved rates Θ^e
  - late intervals $m_0 + 1, \ldots, m \Rightarrow$ unobserved rates Θ^l
  - forecast (hence, schedule) update performed after time m_0.

Generating forecast updates

- Task: update the distribution of the unobservable rate, Θ^l
  - based on some observed early volume X^e
- We know that $X^e \sim N(\Theta^e, \sigma_0^2 I_{m_0})$
- We can forecast in two steps
  - First use X^e to update the distribution of the unobservable rate Θ^e.
  - Then use the updated Θ^e to update the unobservable rate Θ^l.

The approach has both Bayesian and Ridge Regression interpretations.

Forecast updates can significantly reduce error and uncertainty



Stochastic programming with recourse

- Stage 1:
  - Solve the same stochastic program as before
  - Calculate the expected abandonment rate during the latte part of the planning horizon, α_l
- Stage 2:
  - Based on early arrival counts, generate updated arrival rate forecast
  - Updating staffing from y to z would require additional cost $d(y - z)$
  - Solve stochastic program with recourse, with α_l as the QoS target in the latter part

Recourse program that uses 2-stage forecast

- Idea: account for recourse actions in initial schedule
- Example: suppose the cost structure is such that
  - it costs little to send agents home after m_0
  - it costs a lot to increase staffing after m_0
  - Should initially staff high and send people home, if necessary
- In two-stage program
  - First-stage periods as before: initial staffing y fixed across scenarios
  - Second-stage periods more complex: for each initial scenario, second-stage action z varies

We test six scheduling schemes

- Two schemes with no updating
  - one scenario = IP ◊
  - 100 scenarios = SP100 ⧫
- Two schemes with an afternoon update of the original schedule
  - one scenario = UP □
  - 100 scenarios = UP100 ■
- Two schemes that update an original schedule with recourse
  - one scenario = RP ○
  - 100 scenarios = RP100 ●

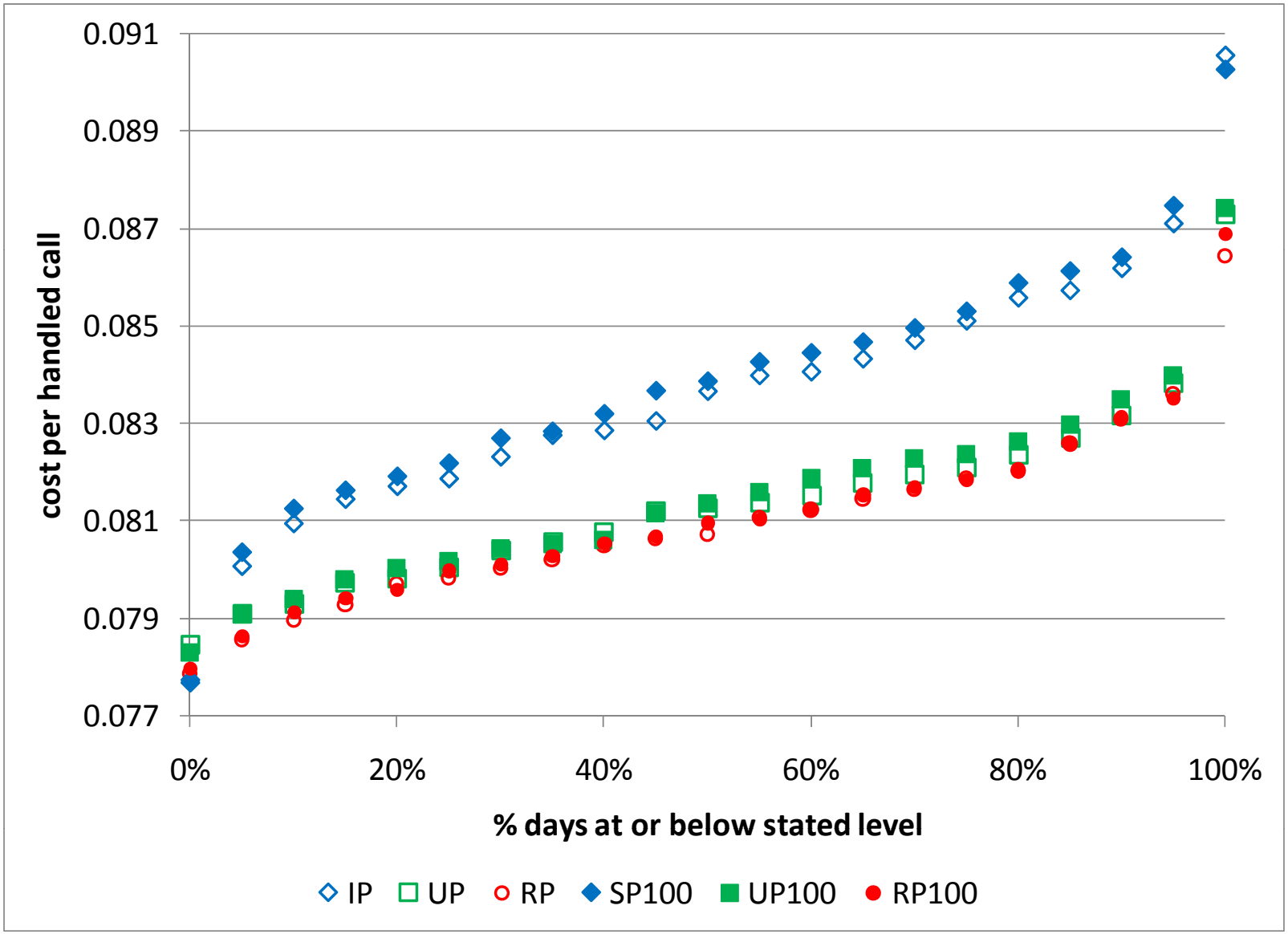
Testing the value of the scheduling schemes

1. Preliminary forecast using previous n days of data
2. Solve 4 scheduling problems based on initial forecast
   - IP ◊ and SP100 ⧫
   - 1st phase of RP ○ and RP100 ●
3. Update forecast based on 1st part of day
4. Update solutions based on revised forecast
   - IP $\Rightarrow$ UP □ and SP100 $\Rightarrow$ UP100 ■
   - 2nd phase of RP ○ and RP100 ●
5. Simulate using schedules and actual arrival counts

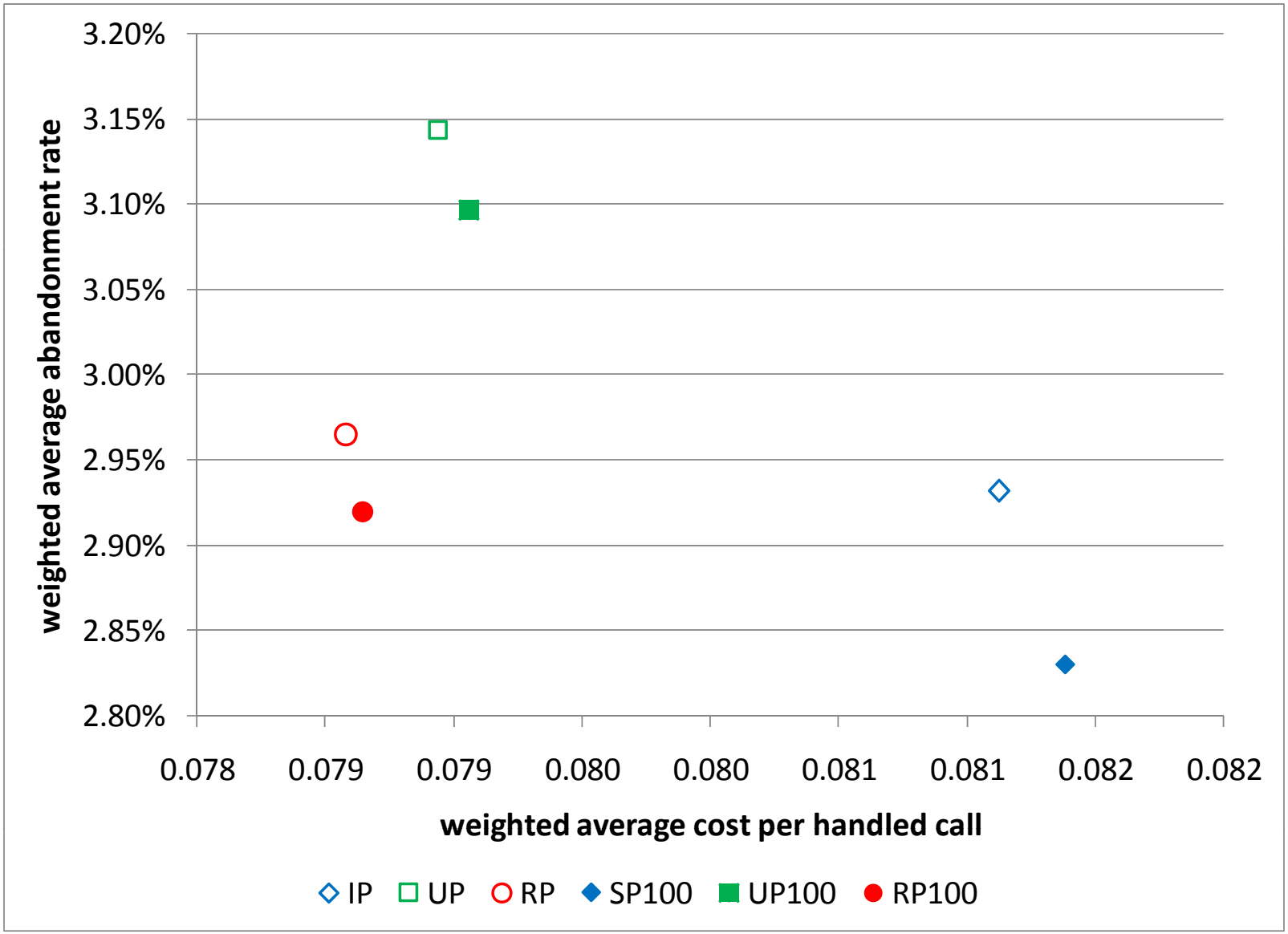
One set of empirical tests

- The same network of four large retail-banking call centers in US
  - Schedule updates at 11am
- Shift structure and costs
  - 262 feasible daily schedules (7 and 9-hour shifts, with breaks)
    - ★ cost of 1 per agent per 1/2-hour interval
  - 4,973 potential recourse actions (with 1/2-hour costs)
    - ★ send home (-0.75), overtime (1.5), call in (2.0)
- Arrival data, forecasts, and QoS target
  - Last 100 days as testing set
  - Forecasts based on previous (rolling) 110 days of data
  - Target expected abandonment rate of 3% across scenarios

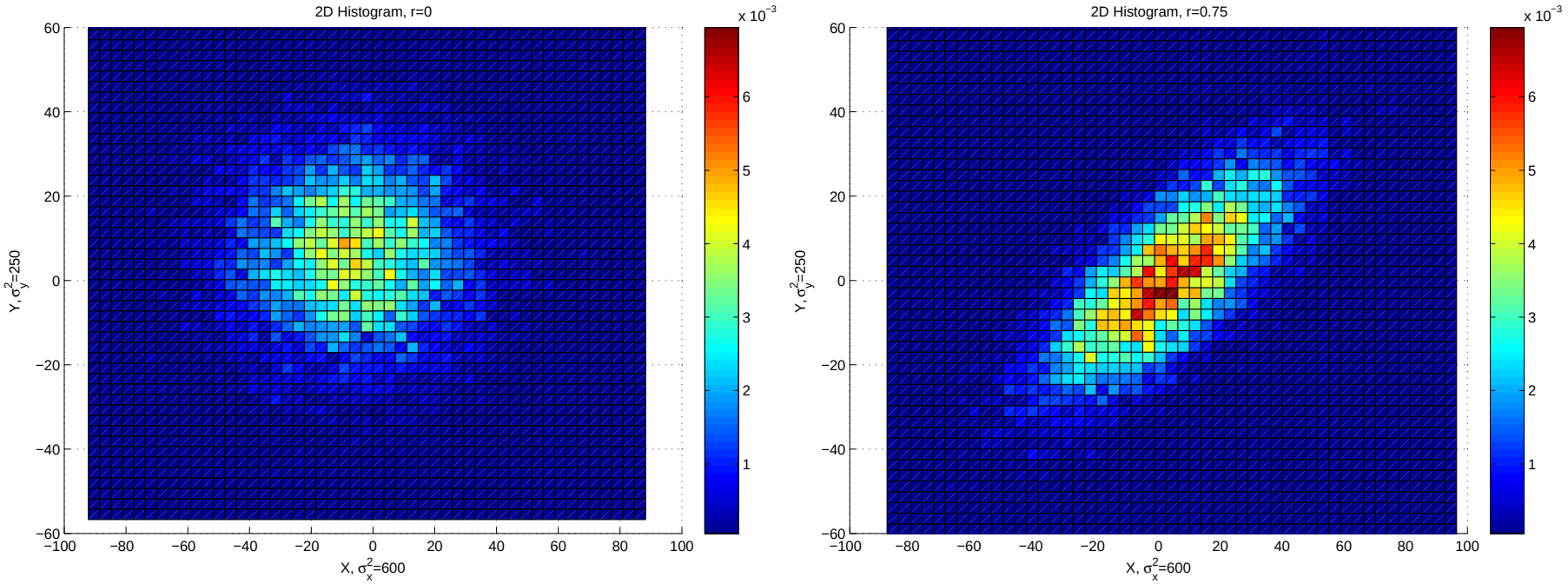
Updating systematically lowers cost per call



RP and RP100 saving 3.2%–3.5% vs IP and SP100



Work in progress: Israeli telecom, Private and Business



Inter-type correlation: 0.7 to 0.8

Luedtke, Shen and Ye (2011)

Outline

Motivation for studying service time

- They are economically important
  - represent the bulk of costs in call centers
- They are operationally important
  - drive queueing performance / QoS in call centers
- They are not well understood empirically
- Findings for them may carry over to other labor-intensive services

Preliminary analysis of call-by-call data

- Two years of call-by-call data
  - large US bank with set of 4 retail-banking call centers
- Data for about 5,000 agents
  - 900-1,200 on weekdays and 200-500 on weekends
- Busy agents take 50-100 calls per day
- Access to a number of operations covariates
  - type of call, order of call, time of call, congestion level

Gans, Liu, Mandelbaum, Shen, and Ye (2010)

Linear regressions for a cohort of 21 agents

- Log-linear learning curve model for each individual agent
  - where ε_j are i.i.d. and N(0,σ^2)

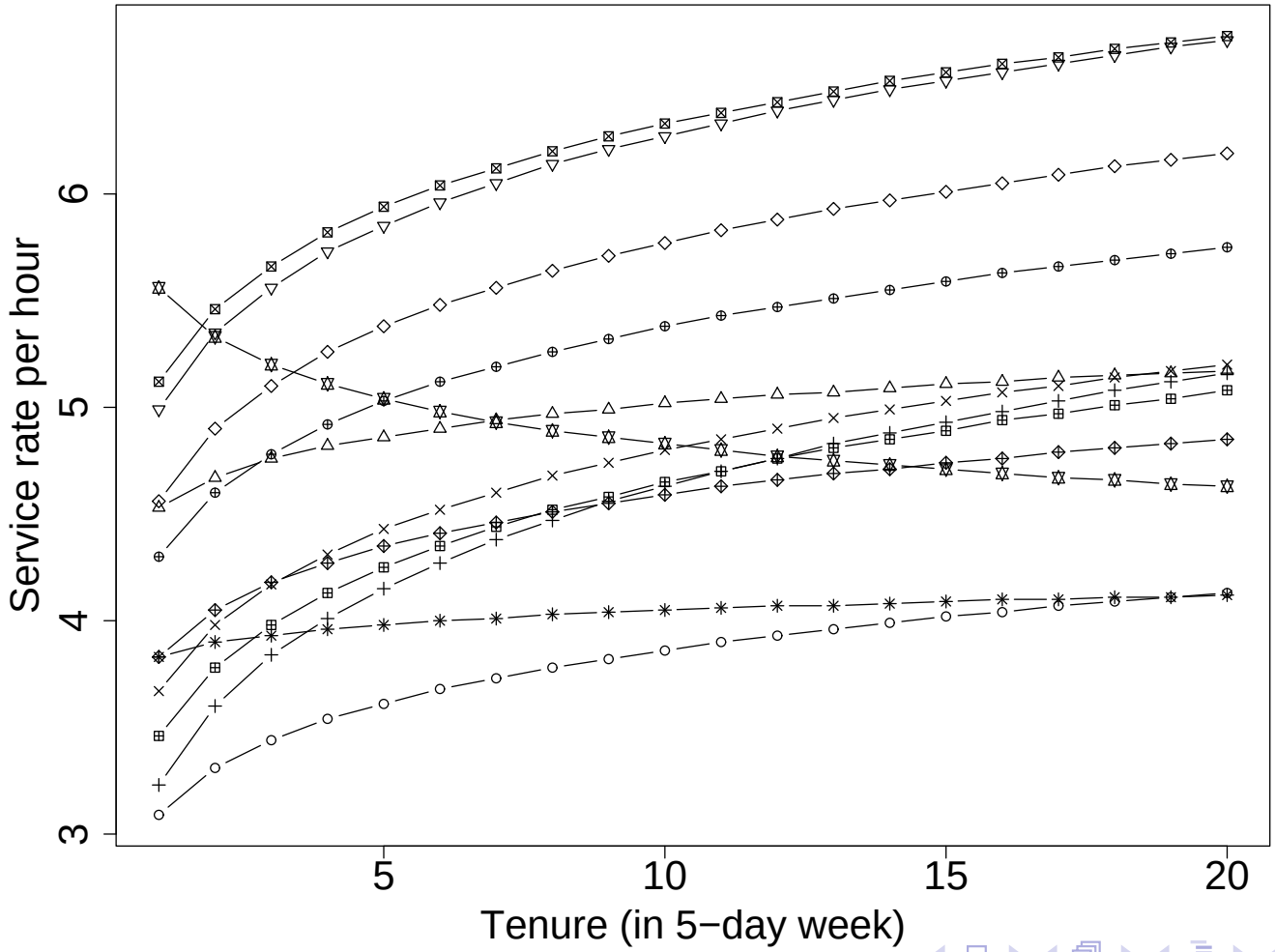
intercept	p-value	log (recnum)	p-value	log (run_len)	p-value	ab_rate	p-value	agent term	p-value	unknown term	p-value	transfer term	p-value	later segment	p-value
5.460	0.000	-0.087	0.000	0.034	0.001	0.019	0.000	-0.365	0.000	0.872	0.000	0.273	0.000	0.053	0.650
5.062	0.000	-0.016	0.128	-0.014	0.204	0.018	0.000	-0.450	0.000	1.069	0.000	0.300	0.000	0.109	0.379
5.763	0.000	-0.075	0.000	-0.048	0.030	0.019	0.018	-0.538	0.000	0.795	0.000	0.047	0.553	-0.368	0.091
5.405	0.000	-0.066	0.000	0.028	0.000	0.018	0.000	-0.775	0.000	0.869	0.000	0.256	0.000	0.302	0.002
5.251	0.000	-0.059	0.000	0.026	0.002	0.018	0.000	-1.020	0.000	0.803	0.000	-0.032	0.179	0.126	0.165
5.126	0.000	-0.034	0.001	0.061	0.000	0.017	0.000	-1.258	0.000	0.699	0.000	0.091	0.004	0.220	0.068
5.866	0.000	-0.058	0.000	-0.027	0.005	-0.003	0.571	-0.659	0.000	0.365	0.000	-0.190	0.000	0.132	0.000
5.592	0.000	-0.044	0.000	-0.030	0.000	0.013	0.003	-0.645	0.000	0.460	0.000	-0.110	0.000	0.126	0.000
5.364	0.000	-0.011	0.142	-0.040	0.000	0.001	0.817	-0.750	0.000	0.376	0.000	-0.131	0.000	0.107	0.000
4.362	0.000	-0.043	0.000	0.018	0.030	0.007	0.143	1.087	0.000	1.624	0.000	1.357	0.000	0.562	0.000
4.906	0.000	0.002	0.848	0.021	0.015	0.012	0.021	-1.102	0.000	0.862	0.000	0.144	0.000	0.131	0.114
5.335	0.000	-0.046	0.000	0.005	0.561	0.017	0.006	-0.609	0.000	0.952	0.000	0.346	0.000	0.306	0.001
5.416	0.000	-0.048	0.000	0.004	0.655	0.015	0.000	-1.227	0.000	0.751	0.000	0.036	0.084	0.023	0.782
5.203	0.000	-0.046	0.000	-0.006	0.414	0.031	0.000	-0.377	0.000	1.029	0.000	0.213	0.000	0.213	0.010
5.147	0.000	-0.052	0.000	0.028	0.000	0.020	0.000	-1.088	0.000	0.855	0.000	0.302	0.000	0.457	0.000
5.209	0.000	-0.054	0.000	0.024	0.000	0.015	0.000	-1.197	0.000	0.858	0.000	0.101	0.000	0.376	0.000
4.950	0.000	0.023	0.003	0.027	0.001	0.020	0.000	-0.703	0.000	0.721	0.000	0.167	0.000	0.219	0.015
5.456	0.000	-0.040	0.000	-0.009	0.400	0.012	0.045	-1.145	0.000	0.774	0.000	0.122	0.000	0.343	0.002
5.711	0.000	-0.091	0.000	0.005	0.439	0.019	0.000	-0.449	0.000	0.648	0.000	0.098	0.000	0.286	0.000
4.369	0.000	0.087	0.000	-0.022	0.057	0.024	0.000	-0.560	0.000	0.820	0.000	0.241	0.000	0.636	0.000
5.703	0.000	-0.078	0.000	0.008	0.212	0.018	0.000	-1.127	0.000	0.666	0.000	0.053	0.006	0.307	0.000

- Significant variation across agents (fixed, random effects)
- Other significant covariates
  - learning, congestion, who handles call, who ends call

Daily learning curves of 12 agents at site S

A cohort of 12 out of the 21 agents above

- handle only “retail banking” calls
- work at site S



What the regression results tell us:

- The scale of the learning curve appears to be
  - significant for the long run – over days or weeks
  - not significant in the short run – within a day
- There appear to be significant differences among agents
  - intercept $\Rightarrow$ initial speed; overall speed
  - slope $\Rightarrow$ how quickly they move down the learning curve
- What is the effect of agent heterogeneity on system performance?

Staffing that ignores agent-to-agent variation...

- The cohort of 12 retail agents at site S
  - at week 12 have the estimated service rates (per hour) of

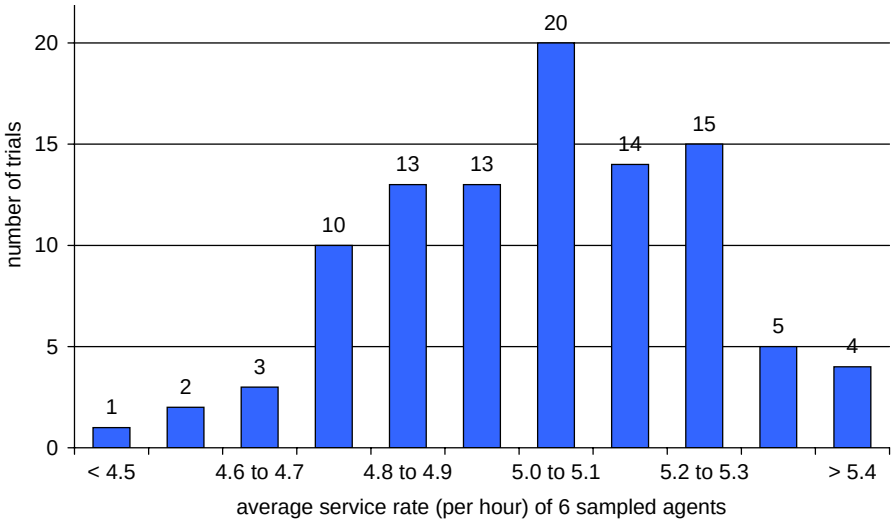
    3.86, 4.05, 4.59, 4.63, 4.65, 4.80, 4.83, 5.02, 5.38, 5.77, 6.27, and 6.33

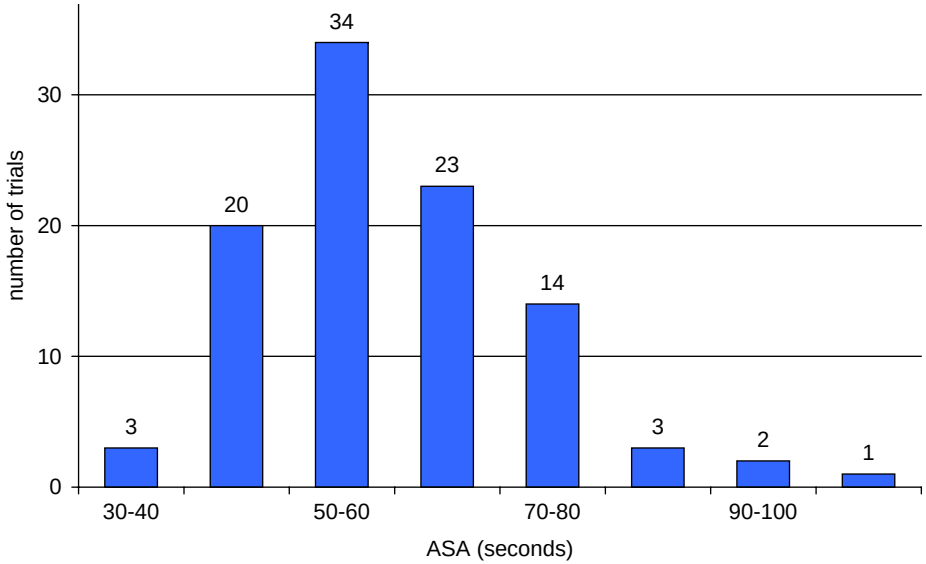
  - an across-agent average of 5.015 calls per hour

- The following M/M/n + M system has an ASA of 58.8 seconds
  - $\lambda = 21$ calls per hour
  - $\mu = 5.015$ calls per hour
  - $n = 6$ agents
  - $\theta^{-1} = 0.5$ hours to abandonment

Ends up with large, unanticipated swings in QoS

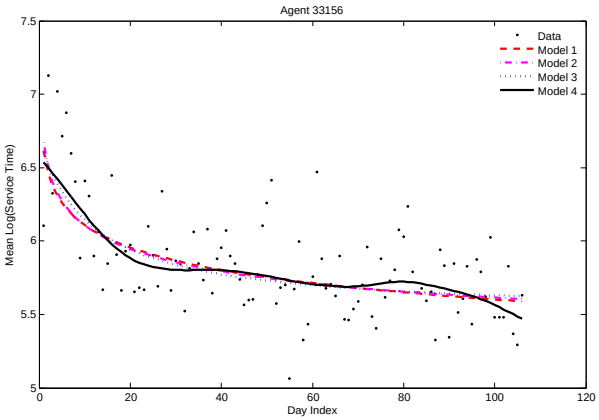
- If we “schedule” 6 of the 12 agents to work by random trial
- Aggregate service rate and ASA vary widely across 100 trials

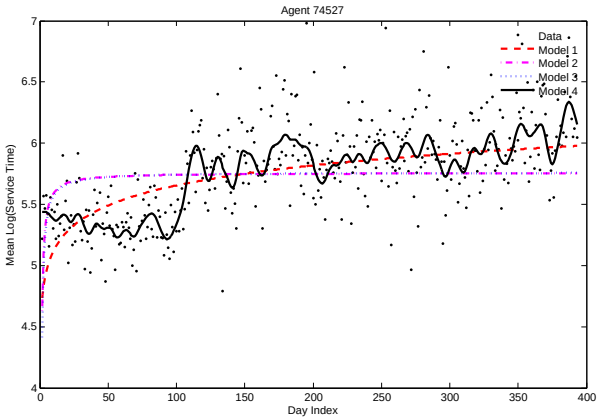


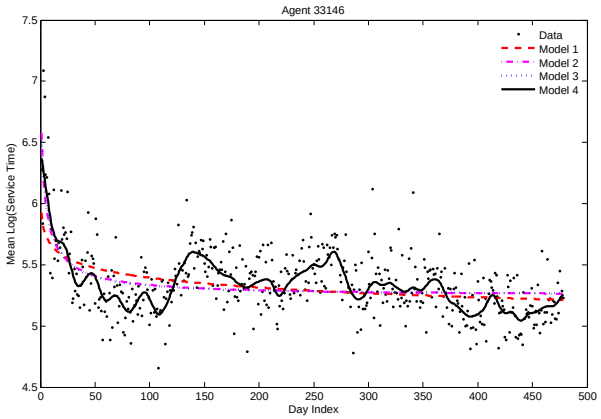


- Ideal: forecast individual agent rates and schedule “accordingly”

Forecasting average log(service time)s of 129 agents







- Typical learning patterns: improving, deteriorating, mixed
- Local splines produced better forecasts than parametric models
- Unable to return to the US bank to investigate anomalies

Ongoing analysis of a second dataset from an outsourcer

- About 20 months of call data from one site serving two clients
- Data for more than 300 agents
  - 75% worked for one client and 25% for the other
  - only 3 agents had worked for both clients
- Busy, experienced agents
  - took about 75 calls / day for one client
  - took about 120 calls / day for the other client
- Currently, access only to time-stamps
  - arrival time, delay, start time, call duration, end time, call-parts

Gans, Shen, and Ye with Genesys

Fitting hyperbolic curves to learning episodes

- Service time of the jth call on the ith day of an episode

$$s_{ij} = k \cdot \left(\frac{i+r}{i}\right)^{b} \cdot e^{\varepsilon_{ij}} \quad \Rightarrow \quad \log s_{ij} = \log k + b \log\left(\frac{i+r}{i}\right) + \varepsilon_{ij}$$

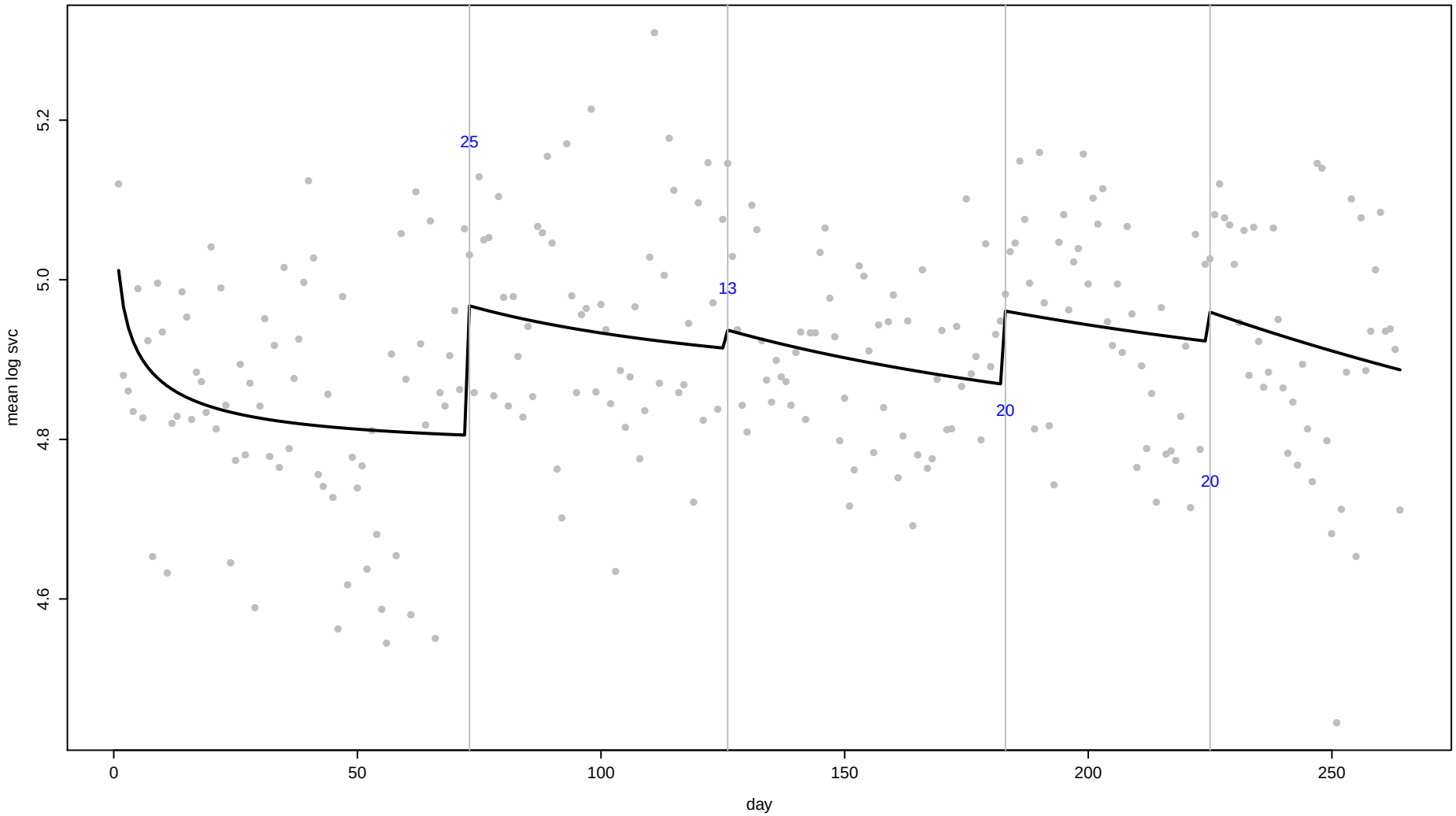
  where

  k – asymptote
  r and b – shape
  ε_{ij} are i.i.d. and N(0,σ^2)

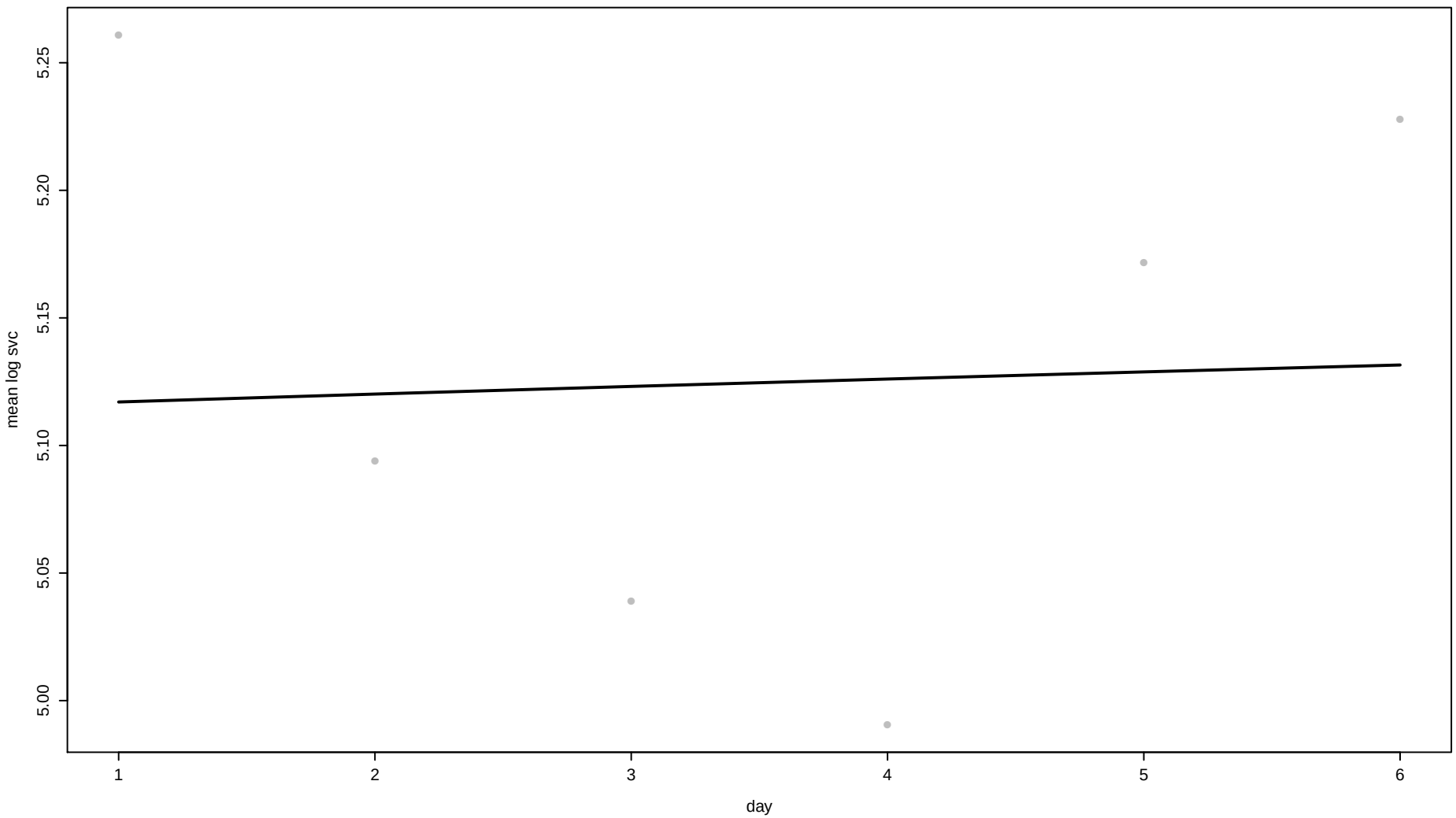
- Generalizes hyperbolic curves used to fit repetitive learning
  - Mazur and Hastie (1978), Nembhard and Uzumari (2000)

- Learning episodes separated by breaks of n or more days
  - as of now, $n = 7$, but this is ad hoc

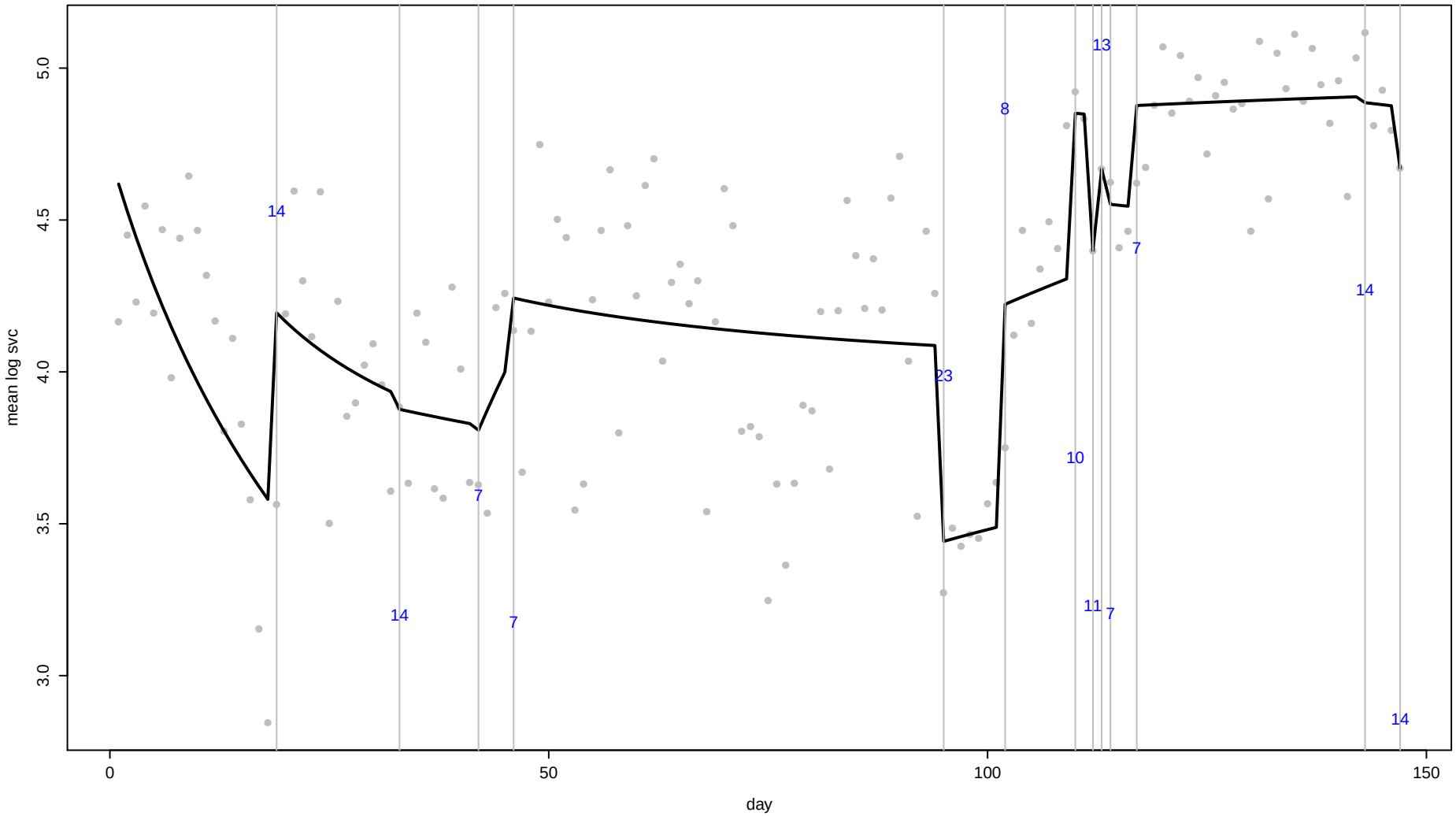
More than 200 of the agents' plots look well behaved



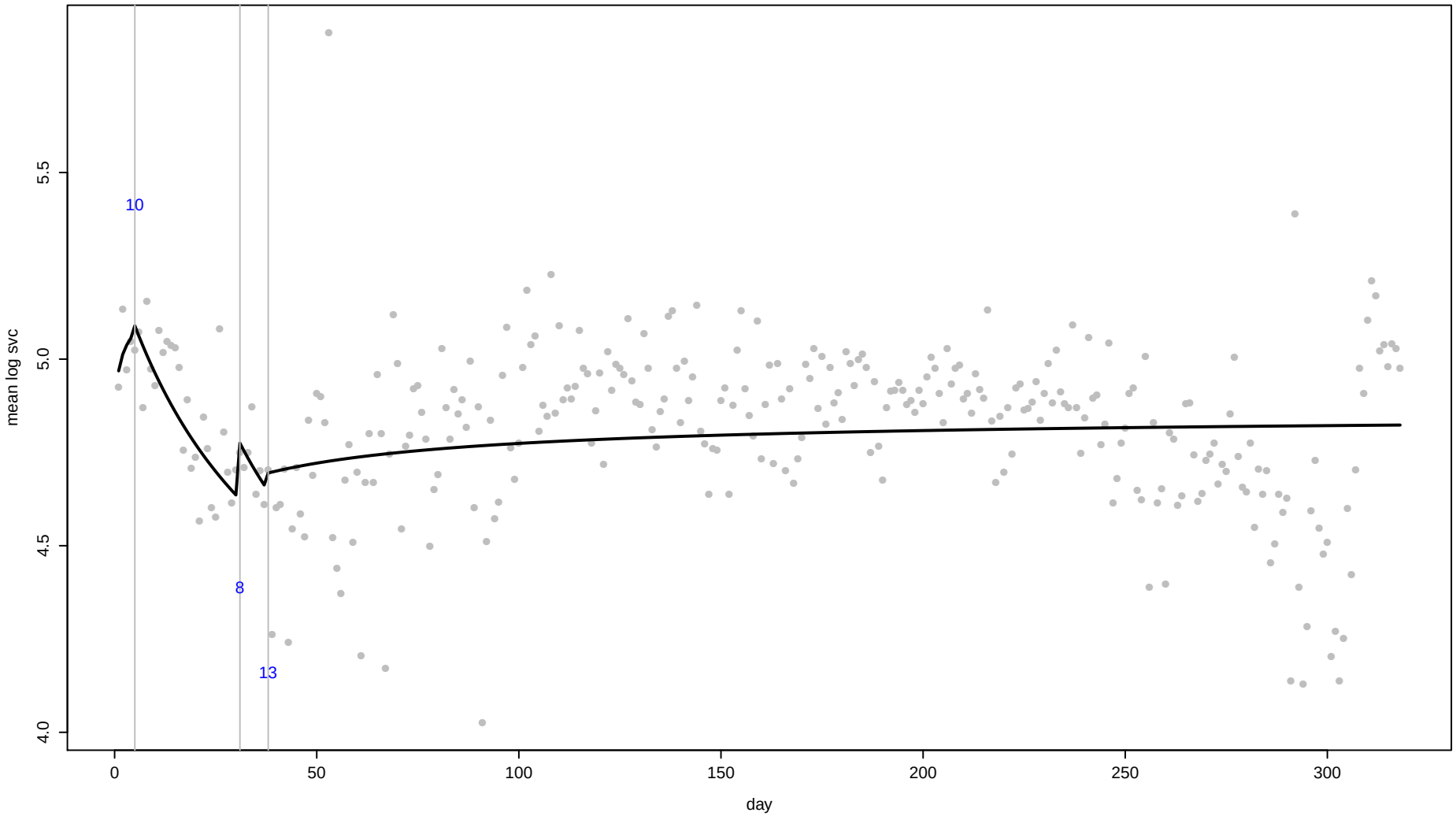
About 40 have very few days of calls



About 30 have unusually many breaks



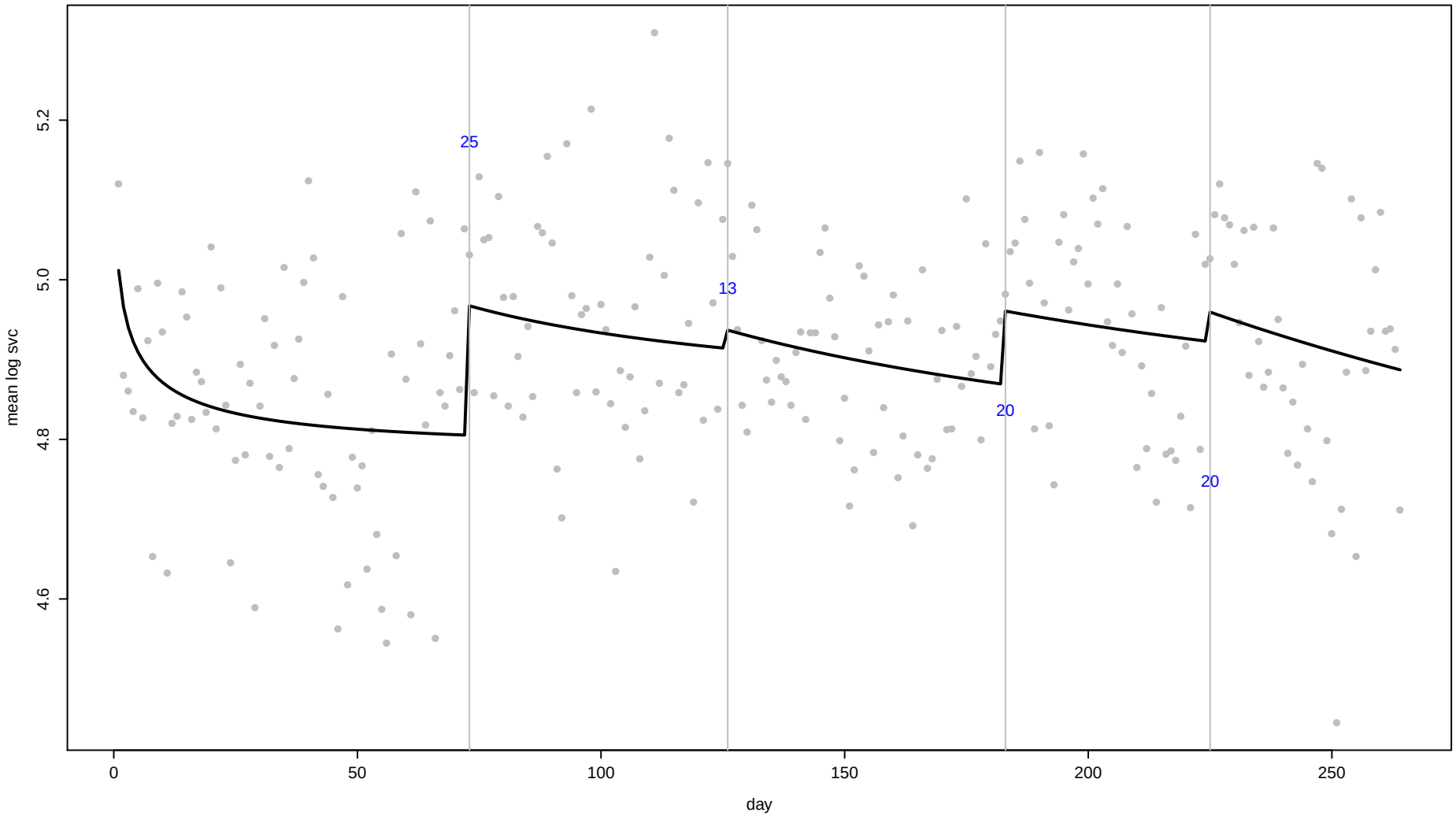
About 25 do not seem to fit the hyperbolic curve



Next steps for the analysis

- Continue to talk with the outsourcer about the data
  - check agents whose data display anomalies
  - try to obtain other covariates (HR, other data)
- Analyze forgetting during breaks
  - see positive, negative, and apparently insignificant instances
  - Nembhard and Uzumari (2000) consider function of agent experience
  - Bailey (1989) estimates forgetting on an episode-by-episode basis
- Analyze call-by-call patterns within days
  - once day-by-day analysis is stable
  - a two-way model, similar to arrival rate models

Forecasting the effects of experience and breaks



Outline

What we've learned

- Fruitful marriage: queueing theory + statistical analysis
- Common framework: a doubly stochastic process
  - rates (and 2nd moments) are inputs to queueing models
  - forecasting generates random distributions of rates
- Pretty good grip on arrival rates
  - arrival-rate realization generates poisson process
  - large call centers have high rates with relatively low CV's
  - scheduling problems relatively straightforward
- Service rates harder to characterize
  - natural object of analysis: a single agent – less data
  - many highly significant covariates; potentially complex interactions
  - but *something* better than current characterization (nothing)
- Larger impact of these analysis
  - while can be cost benefits (e.g., in scheduling)
  - main impact is to better manage / stabilize QoS

Future problems

- Data availability
  - mainly operational, human resource data very difficult to obtain
  - tremendous amount of efforts involved, hard to sustain due to administrator turnover
- Workload involves both arrival rate and service rate
  - prediction of service rate: little work, lack of meaningful covariates
  - forecasting needs to be combined with staffing, scheduling, routing (for multiple types)
- Most often, need simulation for evaluation of the impact