Challenges in Modeling Arrival and Service Processes in Service Systems

Pierre L'Ecuyer

Université de Montréal, Canada
and
GERAD, CIRRELT

Thanks to Wyean Chan, Rouba Ibrahim, Boris Oreshkin, Nazim Régnard, Laure Leblanc, Delphine Réau, Mamadou Thiongane, Ger Koole

"Meet a GERAD Researcher" conference, November 2017

Simulation Challenges

Want to simulate large complex systems to study their behavior and improve decision making.

- Speed of execution of large simulations.
- "Modeling" methodology and tools for large and complex systems.

Simulation Challenges

Want to simulate large complex systems to study their behavior and improve decision making.

- Simulation-based **optimization** and control.
- Speed of execution of large simulations.
- "Modeling" methodology and tools for large and complex systems.

Simulation Challenges

Want to simulate large complex systems to study their behavior and improve decision making.

- Trustable (valid) **stochastic modeling** of complex systems. Taking account of various kinds of information.
- Simulation-based **optimization** and control.
- Speed of execution of large simulations.
- "Modeling" methodology and tools for large and complex systems.

Big Data

Sometimes huge amounts of data available to build stochastic models. How can we exploit this huge mass of data to build credible models?

How to effectively update the models in real time as new data comes in?

Strong links with data mining, machine learning, Bayesian statistics.

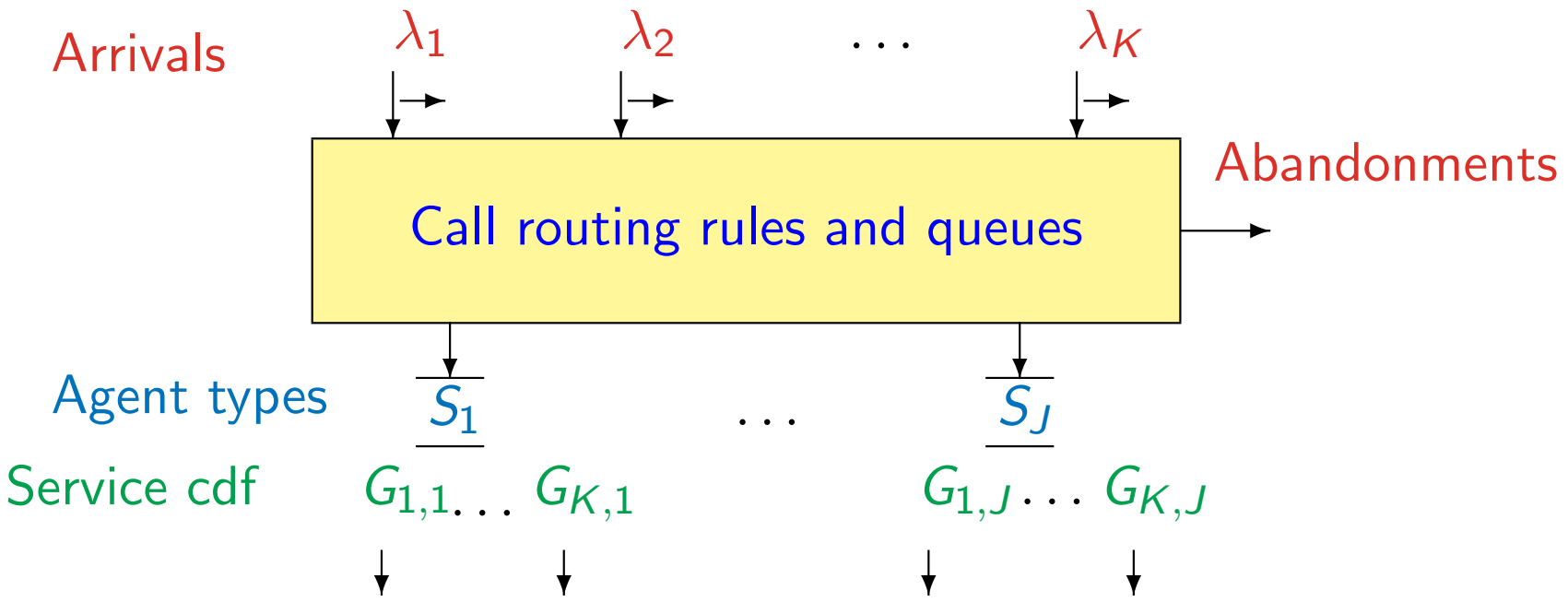
Generally much more complicated than selecting univariate distributions and estimating their parameters. Model inputs are often multivariate distributions and stochastic processes, with hard-to-model (but important) dependence between them, and parameters that are themselves stochastic.

Call centers (or contact centers)

Include sales by telephone, customer service, billing/recovery, public services, 911, taxis, pizza order, emergency services, etc.
Employ around 3% of workforce in North America.

Example: A Multiskill Call Center

Different call types. Depends on required skill, language, importance, etc.
Agent types (groups). Each has a set of skills to handle certain call types.
Service time distribution may depend on pair $\langle$call type, agent group$\rangle$.



Examples of common performance measures

Service level:
$\mathsf{SL}(\tau)$ = fraction of calls answered within acceptable waiting time τ.
(May exclude calls that abandon before τ.)
May consider its observed value over a fixed time period (a random variable), or its expectation, or the average in the long run (infinite horizon), or a tail probability $\mathbb{P}[\mathsf{SL}(\tau) \geq \ell]$.

Examples of common performance measures

Service level:
$\mathsf{SL}(\tau) =$ fraction of calls answered within acceptable waiting time τ.
(May exclude calls that abandon before τ.)
May consider its observed value over a fixed time period (a random variable), or its expectation, or the average in the long run (infinite horizon), or a tail probability $\mathbb{P}[\mathsf{SL}(\tau) \geq \ell]$.

Abandonment ratio: fraction of calls that abandon.

Average waiting time for each call type.

Agent occupancy: fraction of the time where each agent is busy.

Performance evaluation, single call type

Arrival rate λ, service rate μ, load λ/μ, s servers, waiting time W.

Assumes Poisson arrivals with constant rate (not realistic) + single type.

$M/M/s$ **queue (Erlang-C)**. CTMC model.
Approx. of $\mathbb{P}[W > 0]$, $\mathbb{P}[W > \tau]$, and $\mathbb{E}[W]$.

Performance evaluation, single call type

Arrival rate λ, service rate μ, load λ/μ, s servers, waiting time W.

Assumes Poisson arrivals with constant rate (not realistic) + single type.

$M/M/s$ **queue (Erlang-C)**. CTMC model.
Approx. of $\mathbb{P}[W > 0]$, $\mathbb{P}[W > \tau]$, and $\mathbb{E}[W]$.

Approximation under quality and efficiency driven (QED) regime:
$\lambda \to \infty$ and $s \to \infty$ with $\alpha = \mathbb{P}[W > 0] \in (0, 1)$ fixed.
Halfin and Whitt (1981).
Square root safety staffing: $s^* = \lceil \lambda/\mu + \beta\sqrt{\lambda/\mu} \rceil$.
Could make sense for some large call centers.

Performance evaluation, single call type

Arrival rate λ, service rate μ, load λ/μ, s servers, waiting time W.

Assumes Poisson arrivals with constant rate (not realistic) + single type.

$M/M/s$ **queue (Erlang-C)**. CTMC model.
Approx. of $\mathbb{P}[W > 0]$, $\mathbb{P}[W > \tau]$, and $\mathbb{E}[W]$.

Approximation under quality and efficiency driven (QED) regime:
$\lambda \to \infty$ and $s \to \infty$ with $\alpha = \mathbb{P}[W > 0] \in (0, 1)$ fixed.
Halfin and Whitt (1981).
Square root safety staffing: $s^* = \lceil \lambda/\mu + \beta\sqrt{\lambda/\mu} \rceil$.
Could make sense for some large call centers.

$M/M/s + M$ **queue (Erlang-A)**.

Approx. of $\gamma = \mathbb{P}[\text{abandon}]$, $\mathbb{P}[W > 0]$, and $\alpha = \mathbb{P}[W > \tau]$.
QED(τ): Fix τ, α, and $\gamma > 0$.
Modified square root rule: $s^* = \lceil (1-\gamma)\lambda/\mu + \delta\sqrt{(1-\gamma)\lambda/\mu} \rceil$.

Erlang formula calculators developed by Wyean Chan `http://www-ens.iro.umontreal.ca/~chanwyea/erlang/erlangC.html`

Multiple call types, multiskill agents

Much more difficult.
Call routing rules become important and can be complicated.
Approximations for service levels are not very good.

Must rely on simulation.

In my lab, we developed ContactCenters and CCOptim, Java simulation and optimization software libraries for contact centers.
Also some tools for model estimation from data.
Developed mostly by Eric Buist (simulation part) and Wyean Chan (optimization part).

Typical call center

Arrival process is nonstationary and much more complicated than Poisson.
Service times are not exponential and not really independent.
Abandonments (balking + reneging), retrials, returns, etc.

Typical call center

Arrival process is nonstationary and much more complicated than Poisson.
Service times are not exponential and not really independent.
Abandonments (balking + reneging), retrials, returns, etc.

Skill-based routing: Rules that control in real time the call-to-agent and agent-to-call assignments. Can be complex in general.
Static vs dynamic rules. (e.g., using weights).

Typical call center

Arrival process is nonstationary and much more complicated than Poisson.
Service times are not exponential and not really independent.
Abandonments (balking + reneging), retrials, returns, etc.

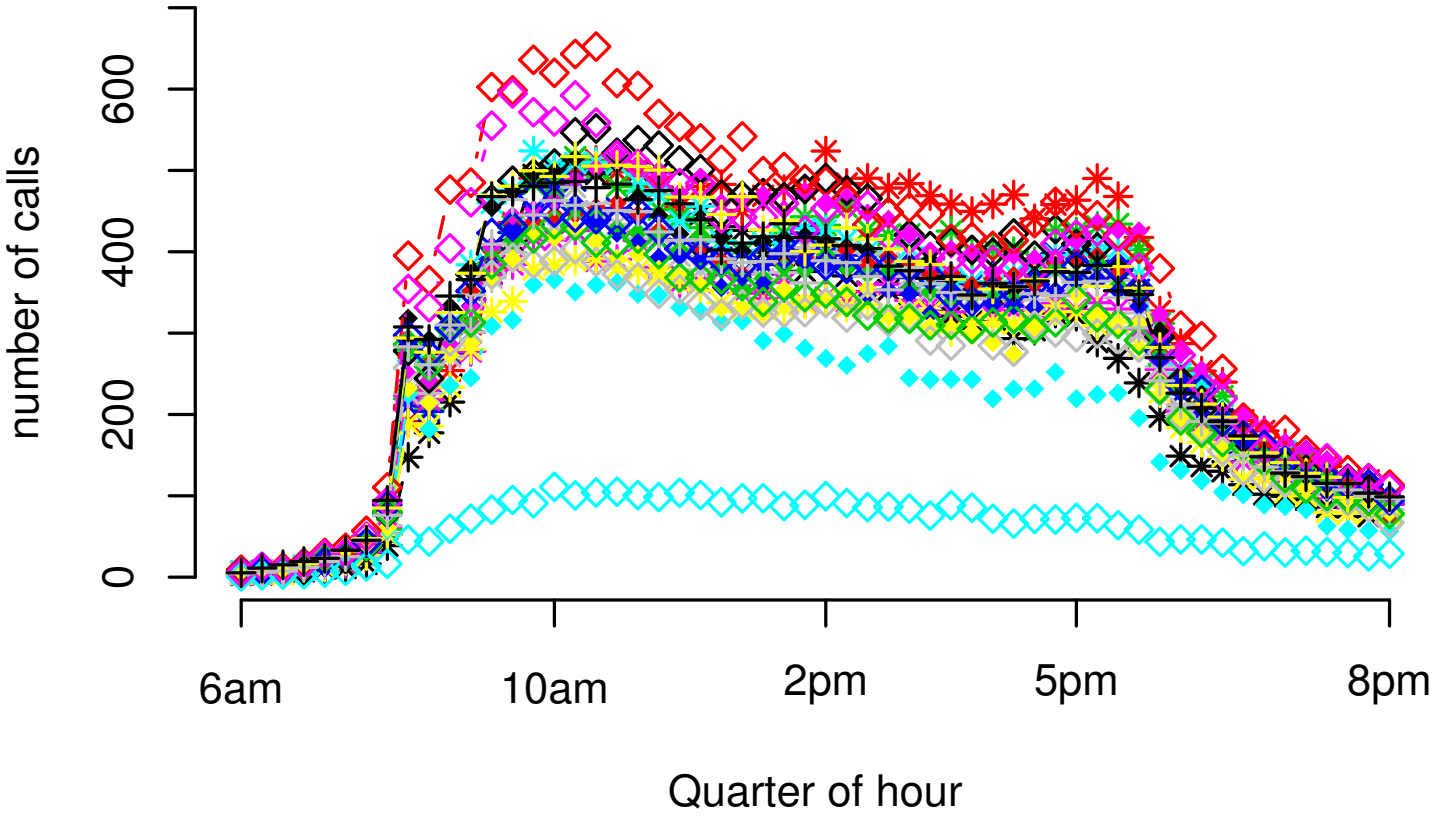
Skill-based routing: Rules that control in real time the call-to-agent and agent-to-call assignments. Can be complex in general.
Static vs dynamic rules. (e.g., using weights).

Agents using fewer skills tend to work faster. Also less expensive.
Compromise between single-skill agents (specialists) vs flexible multiskill agents (generalists).

Staffing/scheduling/routing optimization: objective function and constraints can account for cost of agents, service-level, expected excess waiting time, average wait, abandonment ratios, agent occupancy ratios, fairness in service levels and in agent occupancies, etc. Various constraints on work schedules.

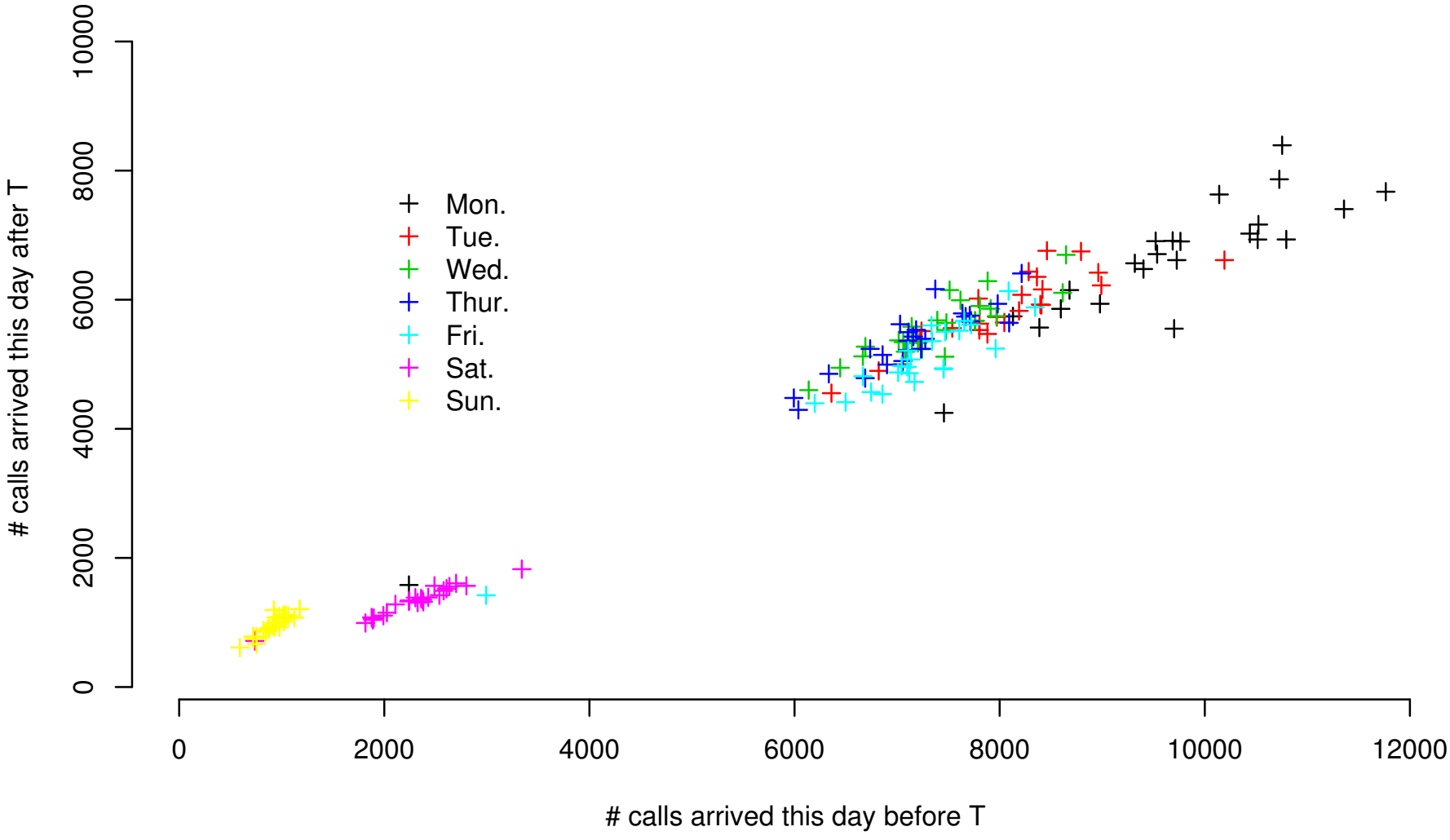
Data on call arrivals

Available observations (for each day): $\mathbf{X} = (X_1, \ldots, X_p)$, arrival counts over (15 or 30 minutes) successive time periods.



Ex.: Typical realizations of $\mathbf{X}$ for a Monday (15-min periods).
Non-stationary. Strong dependence between the X_j's.
Similar behavior in many other settings: customer arrivals at stores, incoming demands for a product, arrivals at hospital emergency, etc.

All days, call volumes before and after $T = 2$ p.m.



Modeling the arrivals

Stationary Poisson process as in Erlang formulas? No.

Modeling the arrivals

Stationary Poisson process as in Erlang formulas? No.

Poisson process with time-dependent arrival rate $\lambda(t)$?
Would imply that $\mathrm{Var}[X_j] = \mathbb{E}[X_j]$. Typically far from true.

Modeling the arrivals

Stationary Poisson process as in Erlang formulas? No.

Poisson process with time-dependent arrival rate $\lambda(t)$?
Would imply that $\mathrm{Var}[X_j] = \mathbb{E}[X_j]$. Typically far from true.

True arrival rate depends on several factors that are hard to predict. We can view it as stochastic, say

$$\Lambda_j = B_j \lambda_j \quad \text{and } X_j \sim \mathrm{Poisson}(\Lambda_j) \qquad \text{over period } j, \text{ where}$$

$\mathbf{B} = (B_1, \ldots, B_p) =$ vector of random busyness factors with $\mathbb{E}[B_j] = 1$,
$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p) =$ vector of constant base rates (scaling factors).

Modeling the arrivals

Stationary Poisson process as in Erlang formulas? No.

Poisson process with time-dependent arrival rate $\lambda(t)$?
Would imply that $\mathrm{Var}[X_j] = \mathbb{E}[X_j]$. Typically far from true.

True arrival rate depends on several factors that are hard to predict. We can view it as stochastic, say

$$\Lambda_j = B_j \lambda_j \quad \text{and } X_j \sim \mathrm{Poisson}(\Lambda_j) \qquad \text{over period } j, \text{ where}$$

$\mathbf{B} = (B_1, \ldots, B_p) =$ vector of random busyness factors with $\mathbb{E}[B_j] = 1$,
$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p) =$ vector of constant base rates (scaling factors).

$$\mathrm{Var}[X_j] = \mathbb{E}[\mathrm{Var}[X_j|B_j]] + \mathrm{Var}[\mathbb{E}[X_j|B_j]] = \lambda_j(1 + \lambda_j \mathrm{Var}[B_j]).$$

Dispersion index (DI) and its standardized version (SDI):

$$\begin{aligned} \mathrm{DI}(X_j) &= \mathrm{Var}[X_j]/\lambda_j = 1 + \lambda_j \mathrm{Var}[B_j] \geq 1, \\ \mathrm{SDI}(X_j) &= (\mathrm{DI}[X_j] - 1)/\lambda_j = \mathrm{Var}[B_j]. \end{aligned}$$

$$\mathrm{Corr}[X_j, X_k] = \frac{\mathrm{Corr}[B_j, B_k]}{[((1 + 1/(\mathrm{Var}[B_j]\lambda_j))(1 + 1/(\mathrm{Var}[B_k]\lambda_k))]^{1/2}}.$$

We expect:

$\mathrm{DI}(X_j) \gg 1$ and $\mathrm{Corr}[X_j, X_k] \approx \mathrm{Corr}[B_j, B_k]$ for "large" $\lambda_j \mathrm{Var}[B_j]$;
i.e., large periods or high traffic.

$\mathrm{DI}(X_j) \approx 1$ and $\mathrm{Corr}[X_j, X_k] \approx 0$ for small $\lambda_j \mathrm{Var}[B_j]$.

Approximately a Poisson process when $\lambda_j \mathrm{Var}[B_j]$ is small.

One good theorem often tells you much more than a bunch of experiments!!!

Do we see this in real data?

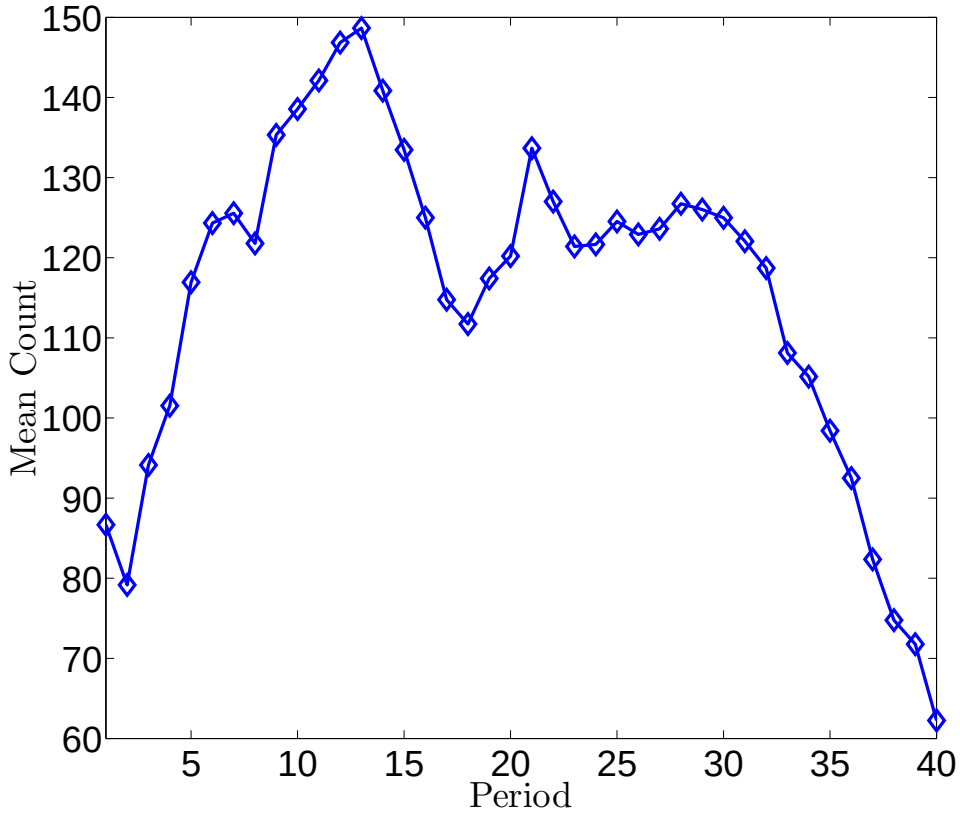
In a simulation, we want to generate the B_j's, then generate the arrivals one by one conditional on the piecewise-constant rates Λ_j.

Another approach (less convenient) is to model and directly generate the X_j's, then randomize the arrival times.

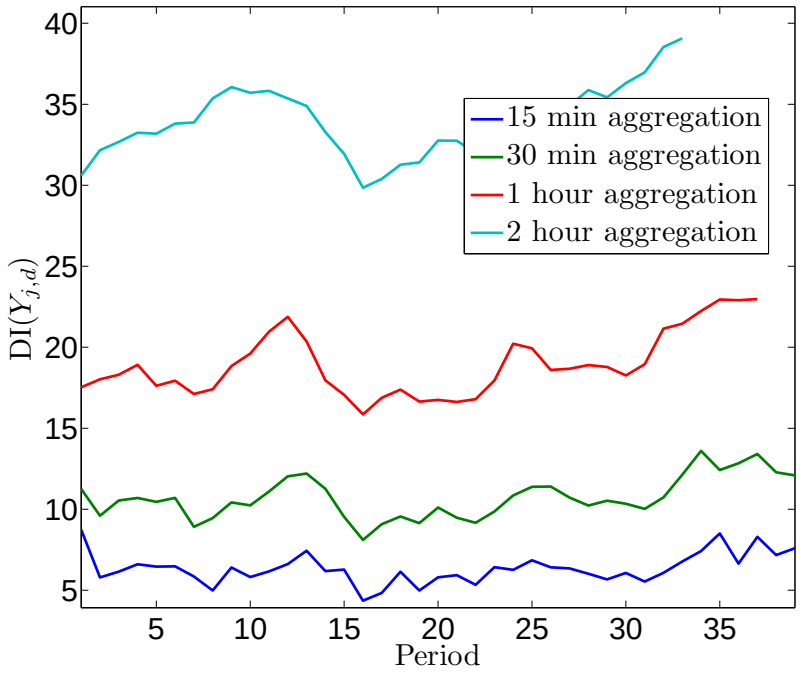
Modeling the rates is harder because they are not observed!

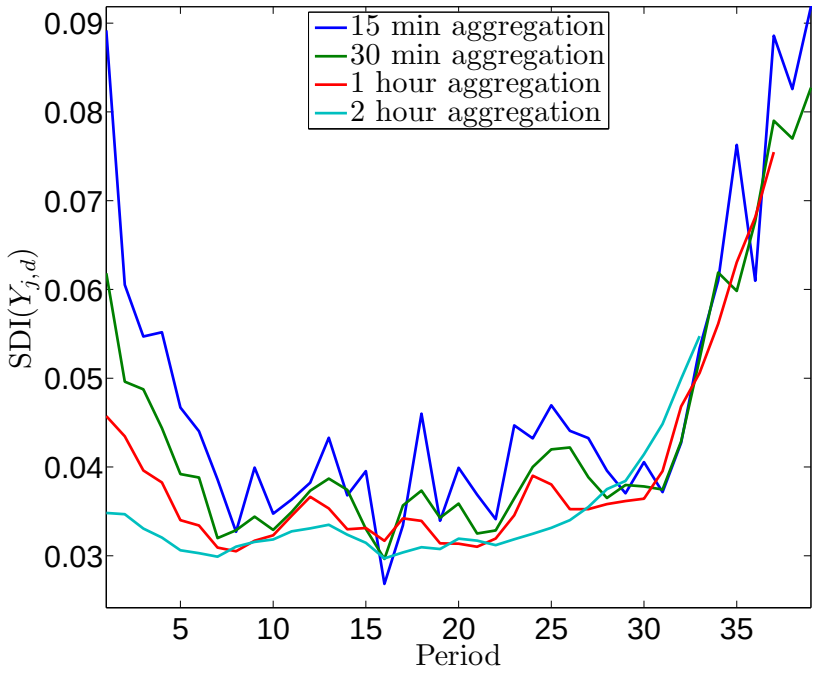
Data from a public utility call center (U)

One call type, data aggregated over 40 15-minute periods per day, from 8:00 to 18:00, Monday to Friday, after removing special days.



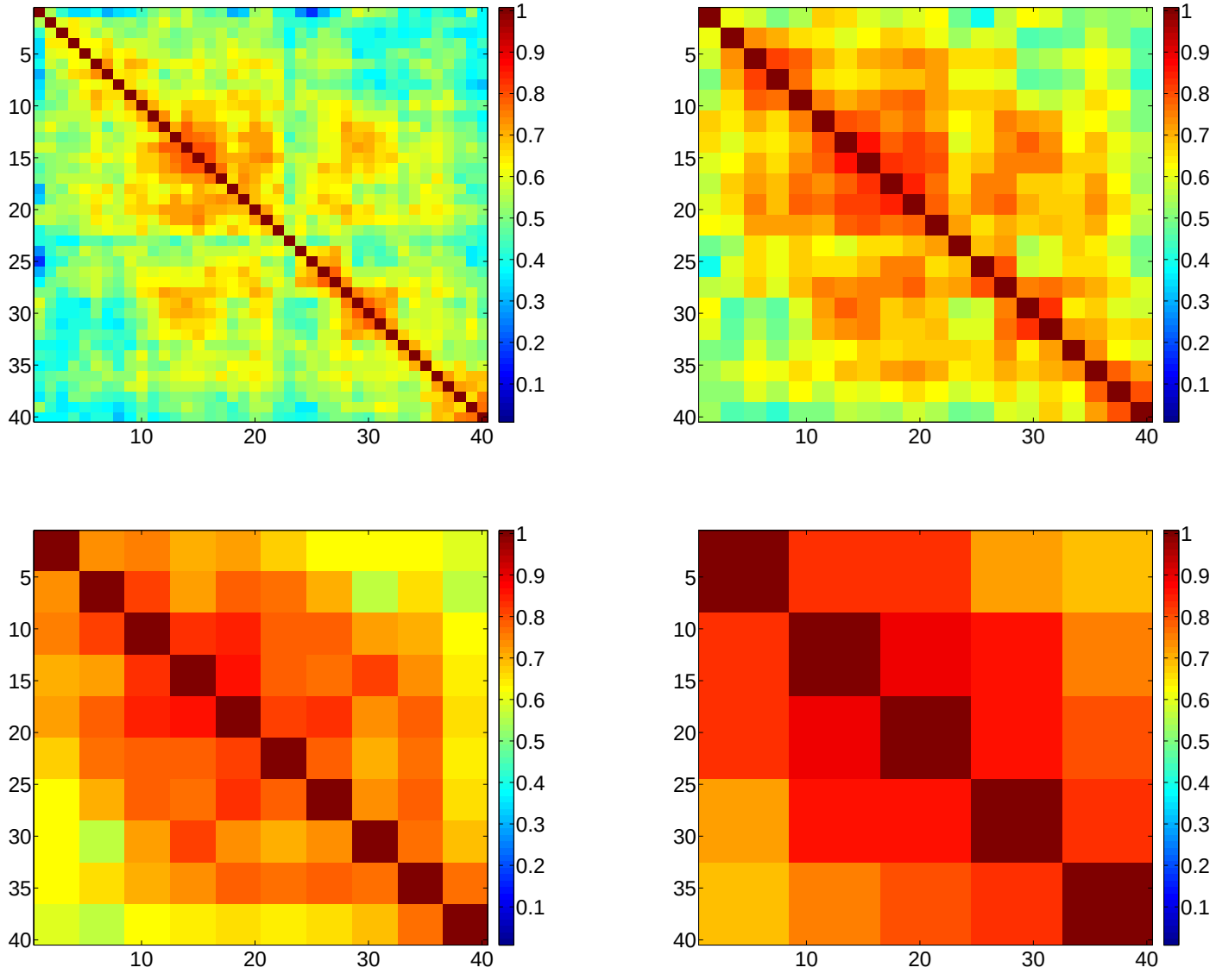
Call center U





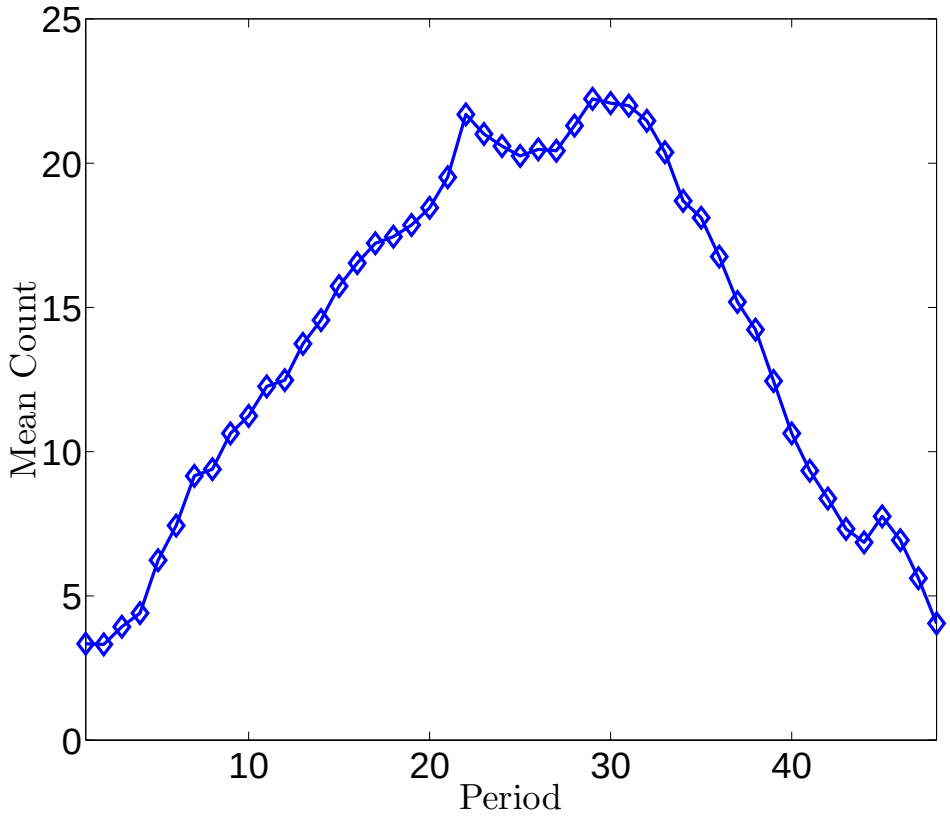
DI (left) and SDI (right) as a function of j for different period lengths.

$\mathrm{Corr}[X_j, X_k]$ in call center U, for 30 min to 4 hour data aggregations.

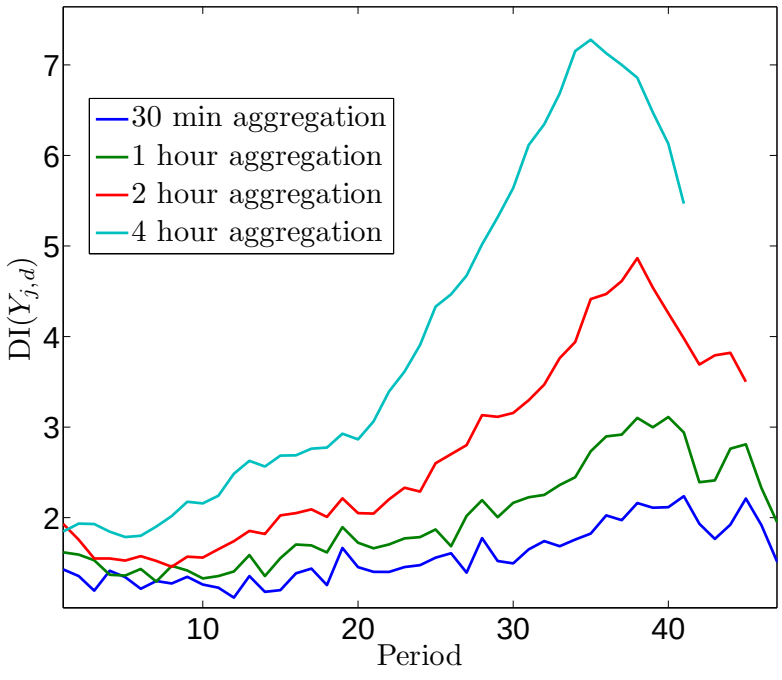


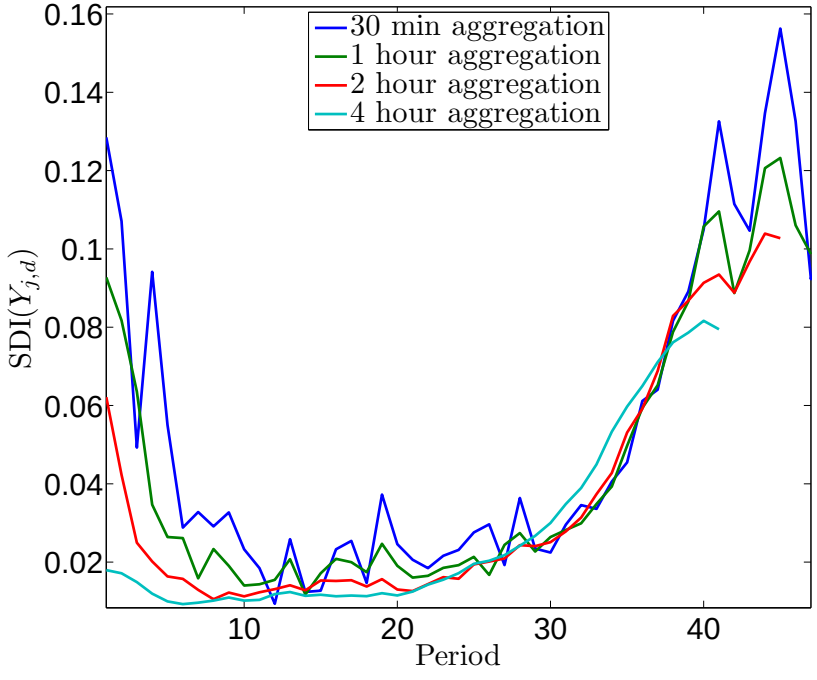
Data from an emergency call center (E)

Take one call type, Monday to Thursday (similar days), after removing special days (holidays, etc.). Other days have different arrival patterns. Day starts at 5 a.m. and is divided into 48 periods of 30 minutes. Mean counts per period, $\approx \lambda_j$:



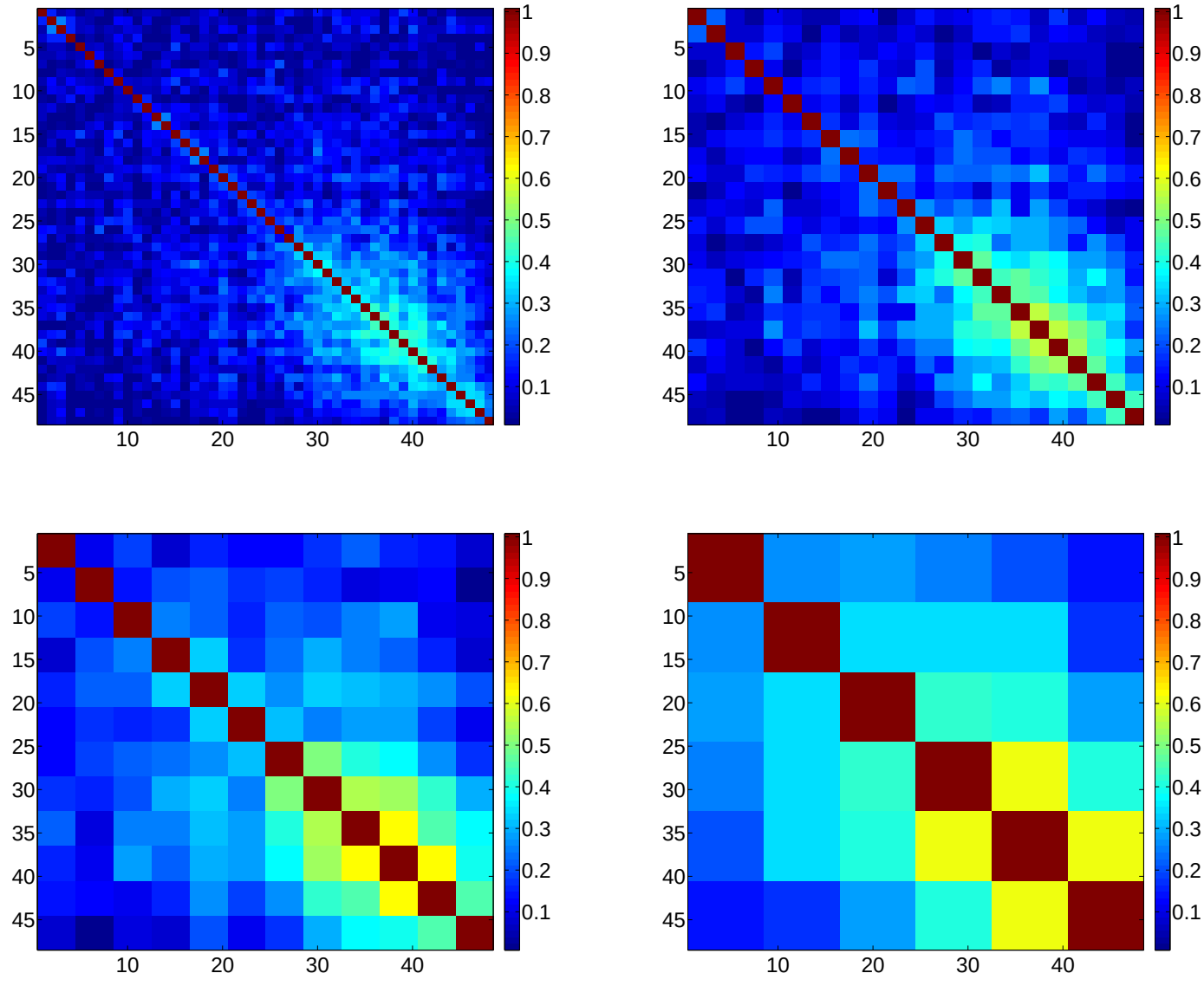
Emergency call center





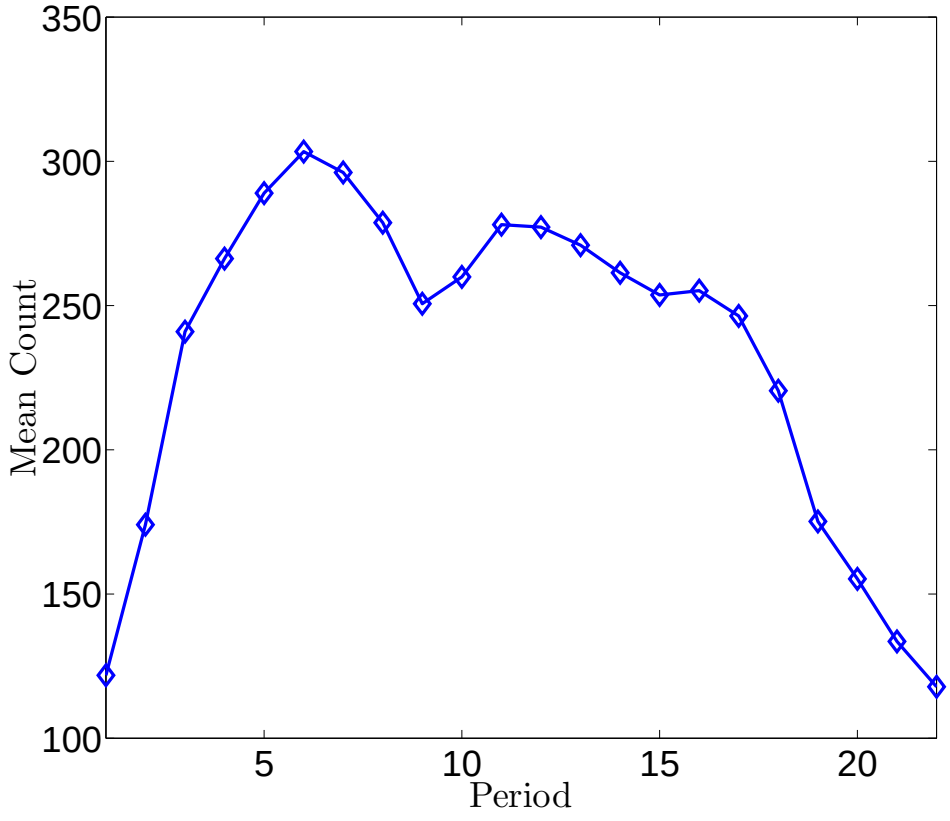
DI (left) and SDI (right) as a function of j for different period lengths.

$\mathrm{Corr}[X_j, X_k]$ in call center E, for 30 min to 4 hour data aggregations.

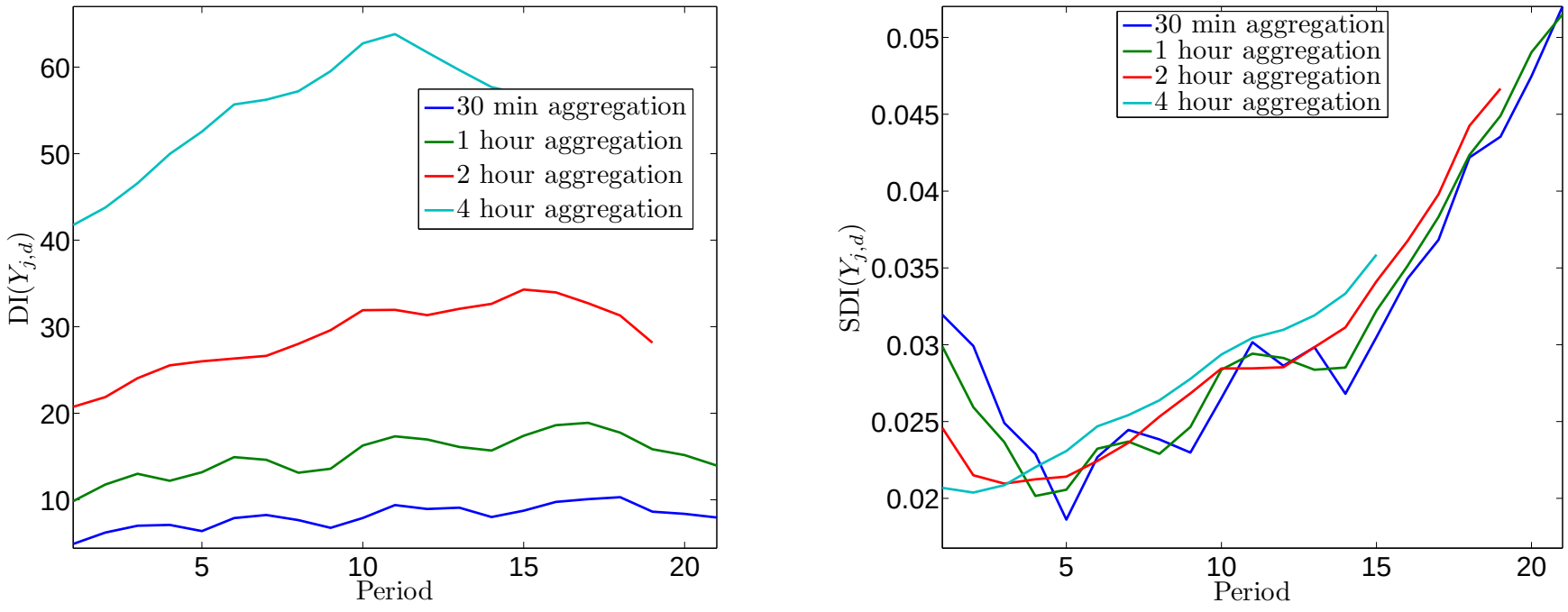


Data from a business call center (B)

One call type, Tuesday to Friday, after removing special days.
Opening hours (8:00 to 19:00) divided into 22 periods of 30 minutes.
Monday and Saturday have different patterns. Mean counts per period:

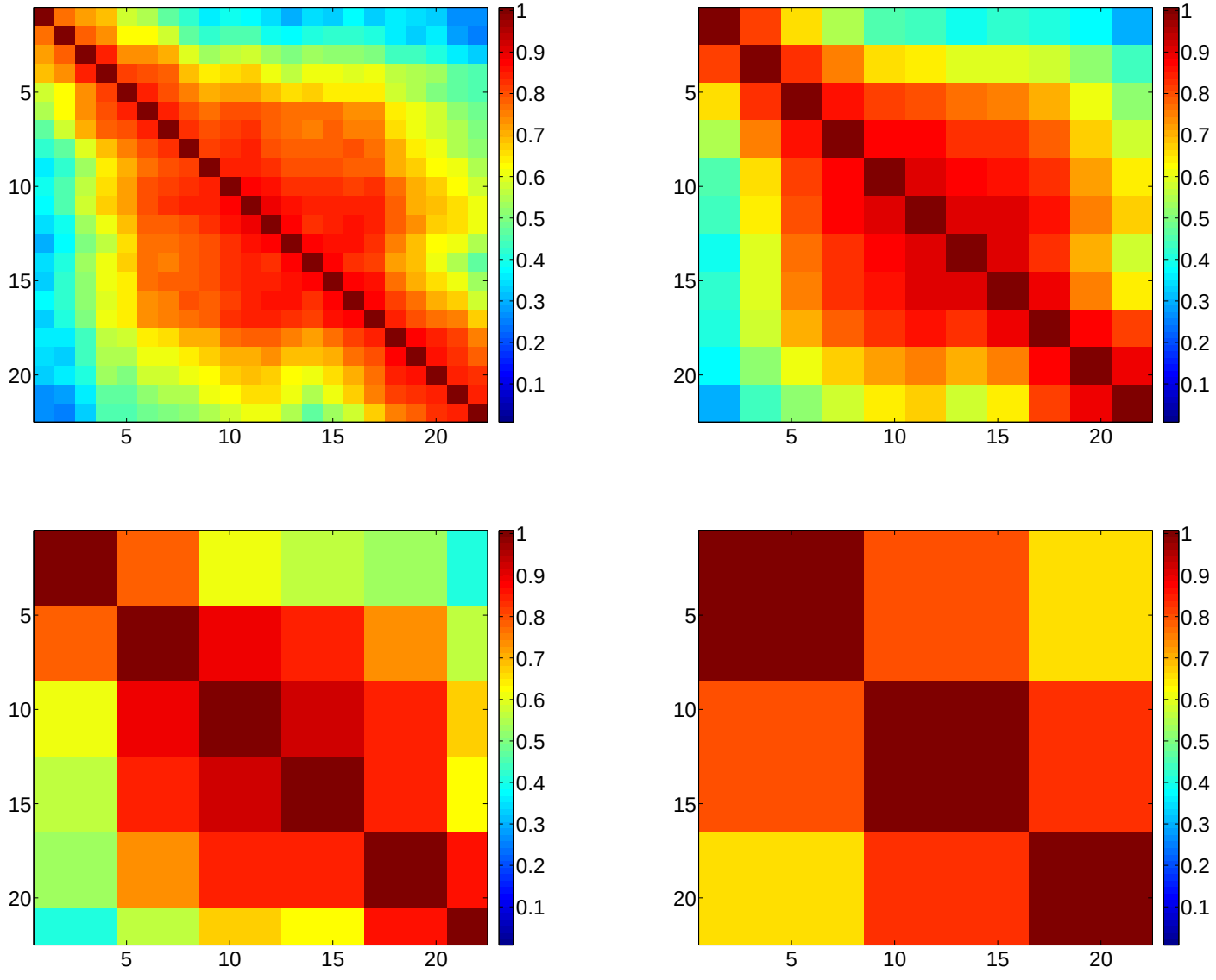


Call center B



DI (left) and SDI (right) as a function of j for different period lengths.
SDI is not as large as for center E, but DI is much larger.

$\mathrm{Corr}[X_j, X_k]$ in call center B, for 30 min to 4 hour data aggregations.



Rate models

$$\Lambda_j = B_j \lambda_j \quad \text{over period } j.$$

Poisson	$B_j = 1$ for all j.
PGsingle	$B_j = B$ for all j, where $B \sim \text{Gamma}(\alpha, \alpha)$.
PGindep	B_j's are independent, $B_j \sim \text{Gamma}(\alpha_j, \alpha_j)$.
PG2	$B_j = \tilde{B}_j B$, combines common B and independent $\tilde{B}_j$'s.
PG2pow	$B_j = \tilde{B}_j B^{p_j} / \mathbb{E}[B^{p_j}]$.
PGnorta	**B** has gamma marginals B_j and dependence specified by a normal copula (we fit all Spearman correlations). $B_j = G_j^{-1}(\Phi(Z_j))$ where $\mathbf{Z} = (Z_1, \ldots, Z_p) \sim N(0, \mathbf{R})$.
PGnortaAR1	Normal copula with $\text{Corr}[Z_j, Z_k] = \rho^{\lvert j-k \rvert}$.
PGnortaARM	Normal copula with $\text{Corr}[Z_j, Z_k] = a\rho^{\lvert j-k \rvert} + c$.

Difficulty: We want to model the B_j's, but they are not observed, only the X_j's are observed. This makes parameter estimation by maximum likelihood (ML) much more challenging, because we have no closed form expression for the likelihood.
Moment matching is often possible, but much less robust and reliable.
We use Monte Carlo-based methods for ML estimation.

Example: Likelihood Function for PG2 Model

$\tilde{B}_{i,j}$ = busyness factor for day i, period j.
$\bar{B}_i$ = busyness factor for day i.

$$p(\mathbb{X}|\mathbb{B},\beta,\boldsymbol{\alpha},\boldsymbol{\lambda}) = \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{I}\prod_{j=1}^{p} \frac{(\lambda_j \tilde{B}_{i,j}\bar{B}_i)^{X_{i,j}} e^{-\lambda_j \tilde{B}_{i,j}\bar{B}_i}}{X_{i,j}!} \frac{\alpha_j^{\alpha_j} \tilde{B}_{i,j}^{\alpha_j - 1} e^{-\alpha_j \tilde{B}_{i,j}}}{\Gamma(\alpha_j)} d\tilde{B}_{i,j}$$

$$= \prod_{i=1}^{I}\prod_{j=1}^{p} \int_0^\infty \frac{(\lambda_j \tilde{B}_{i,j}\bar{B}_i)^{X_{i,j}} e^{-\lambda_j \tilde{B}_{i,j}\bar{B}_i}}{X_{i,j}!} \frac{\alpha_j{}^{\alpha_j} \tilde{B}_{i,j}^{\alpha_j - 1} e^{-\alpha_j \tilde{B}_{i,j}}}{\Gamma(\alpha_j)} d\tilde{B}_{i,j}$$

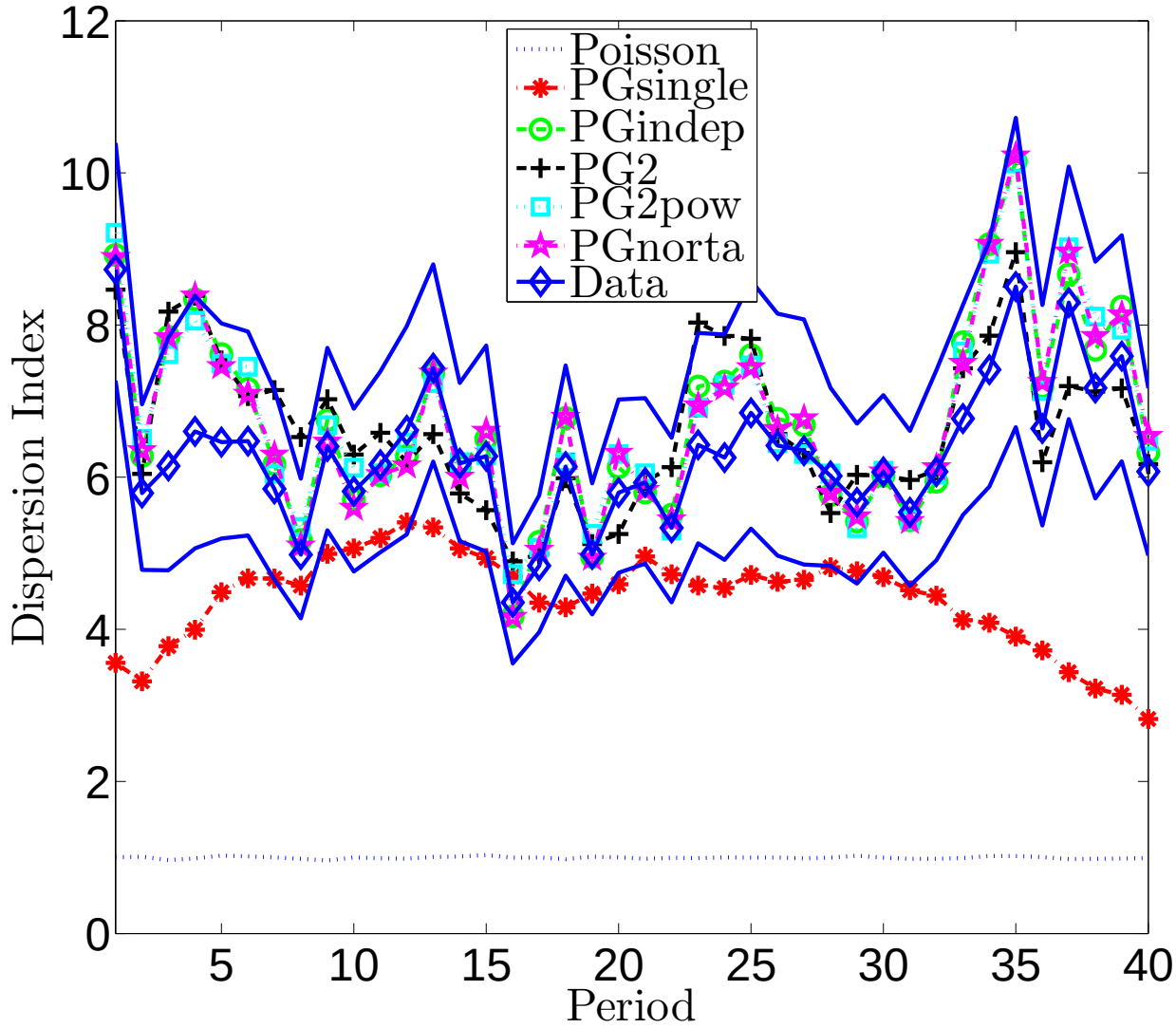
$$= \left[\prod_{j=1}^{p} \frac{\alpha_j{}^{I\alpha_j}}{\Gamma(\alpha_j)^I}\right] \prod_{i=1}^{I}\prod_{j=1}^{p} \frac{\Gamma(\alpha_j + X_{i,j})}{X_{i,j}!} \frac{(\bar{B}_i\lambda_j)^{X_{i,j}}}{(\alpha_j + \bar{B}_i\lambda_j)^{X_{i,j}+\alpha_j}}$$

$$p(\mathbb{X}|\beta,\boldsymbol{\alpha},\boldsymbol{\lambda}) = \left[\prod_{j=1}^{p} \frac{\alpha_j{}^{I\alpha_j}}{\Gamma(\alpha_j)^I}\right] \left[\prod_{i=1}^{I}\prod_{j=1}^{p} \frac{\Gamma(\alpha_j + X_{i,j})}{X_{i,j}!}\right]$$

$$\cdot \prod_{i=1}^{I} \int_0^\infty \left[\prod_{j=1}^{p} \frac{(\bar{B}_i\lambda_j)^{X_{i,j}}}{(\alpha_j + \bar{B}_i\lambda_j)^{X_{i,j}+\alpha_j}}\right] \frac{\beta^\beta \bar{B}_i^{\beta-1} e^{-\bar{B}_i\beta}}{\Gamma(\beta)} d\bar{B}_i.$$

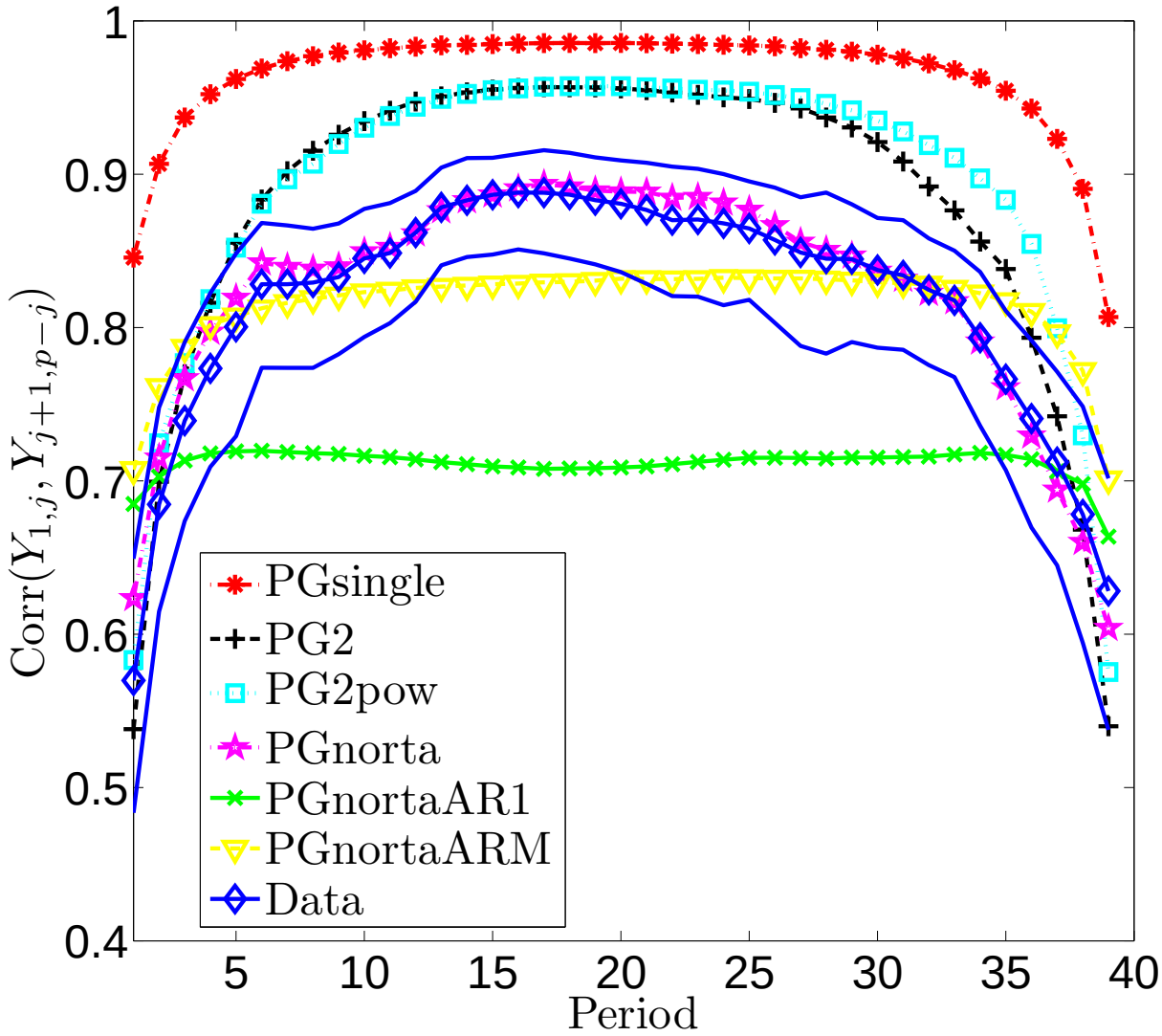
Want to maximize this. No closed form expression for the last integral.

How the models match the DI for Center U



Comparison of the DI for the models and data.

How the models match the correlations for Center U



Comparison of sample correlation between past and future demand.

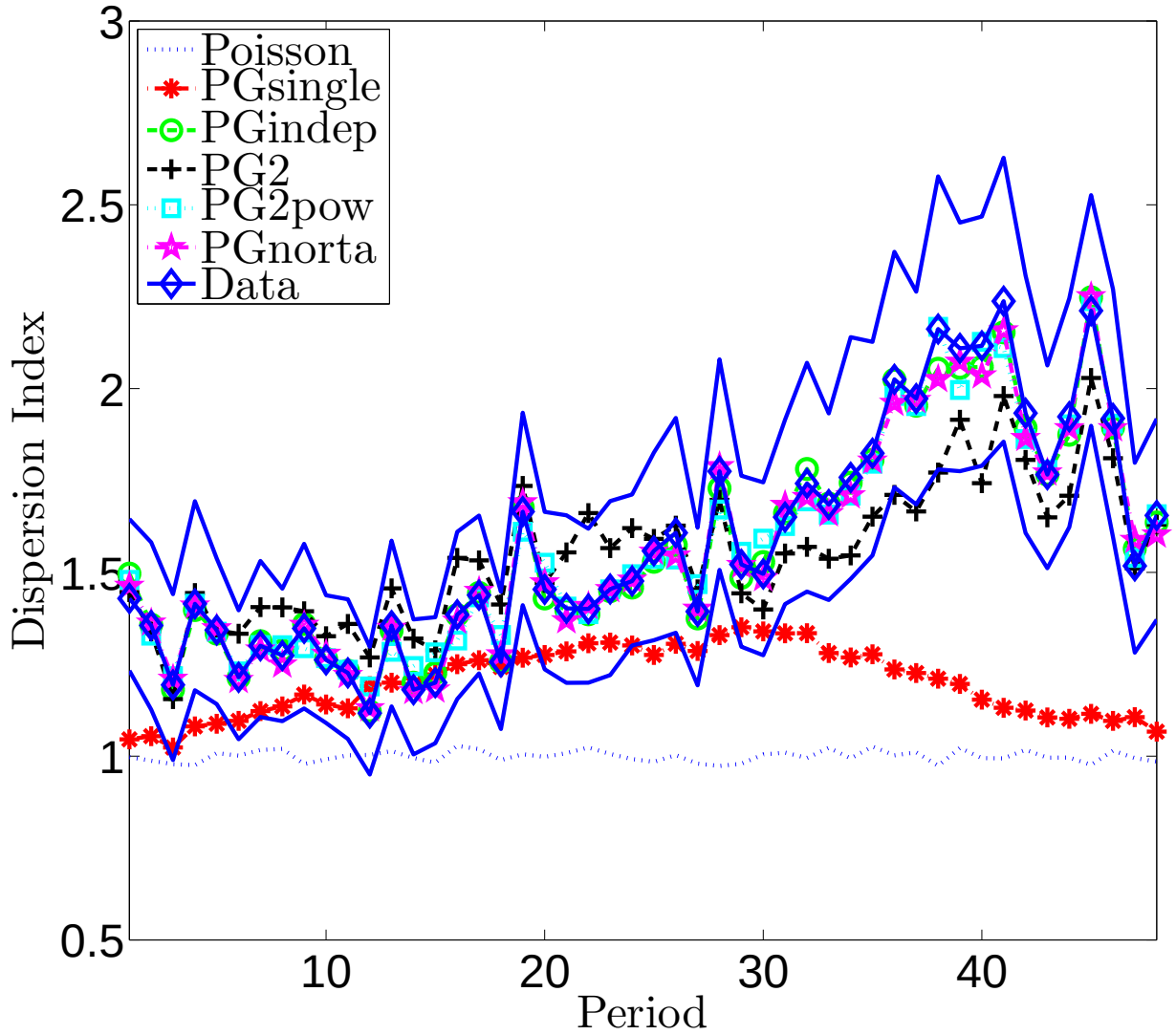
How the distribution predicted by the model fits the data out-of-sample

For each observation i (one day), estimate the model without that day, then for each period j (or block of successive periods) compute interval $[L_{i,j}, U_{i,j}]$ such that $\mathbb{P}[X_{i,j} \in [L_{i,j}, U_{i,j}]] \approx p$ (desired coverage) according to model, then compute the proportion of days where $X_{i,j} \in [L_{i,j}, U_{i,j}]$ and compare with p via sum of squares.

RMS Deviation of out-of-sample coverage probability, for call center U.

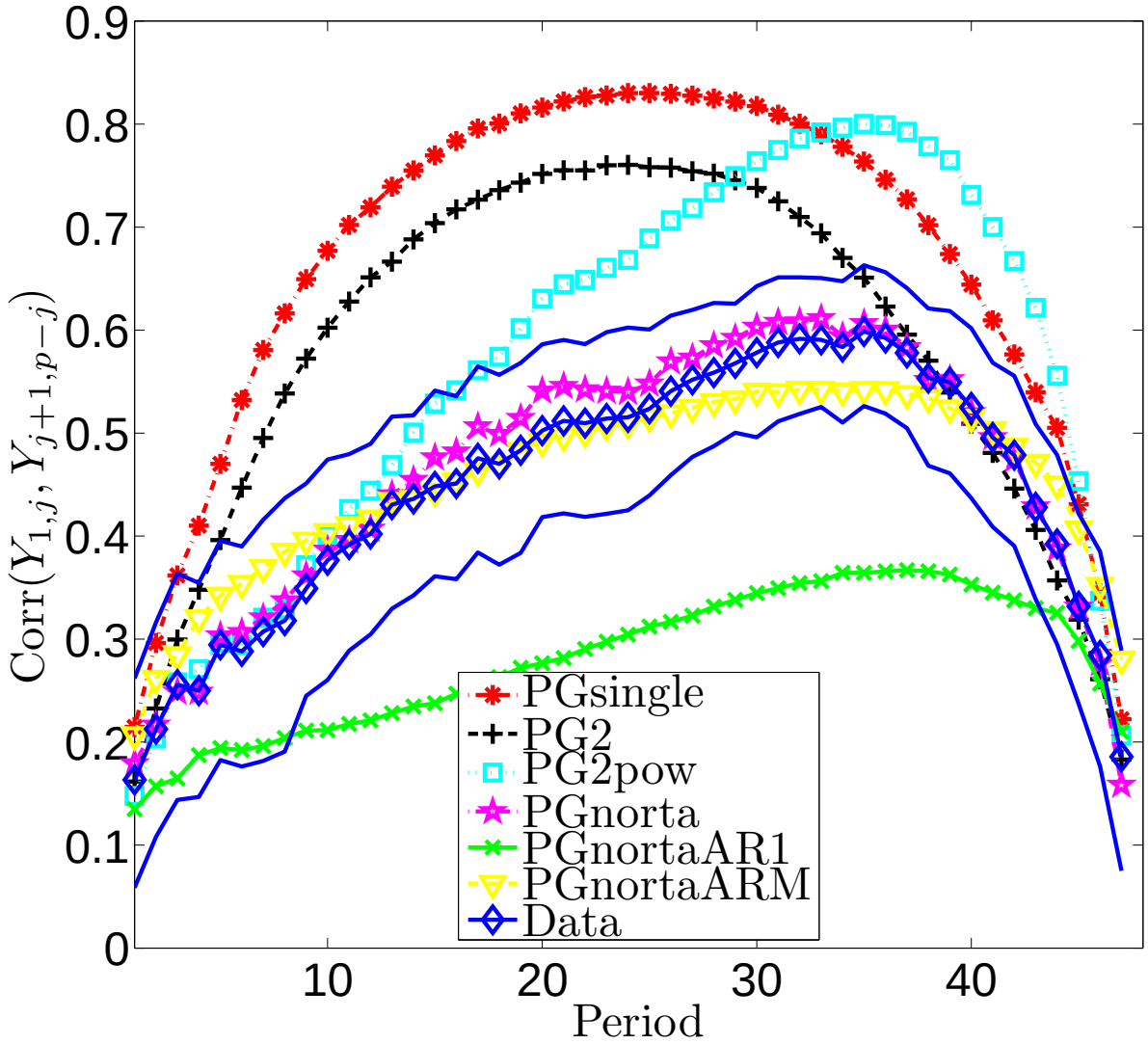
	75% target cover					90% target cover				
	1/4 h	1/2 h	1 h	2 h	4 h	1/4 h	1/2 h	1 h	2 h	4 h
Poisson	38.9	47.1	53.9	59.6	64.6	39.2	50.9	59.7	67.5	74.4
PGsingle	8.6	8.0	6.9	4.0	1.7	7.3	7.0	5.4	3.1	1.4
PGindep	4.5	10.5	24.6	36.5	46.3	1.8	8.4	22.3	37.8	51.2
PG2	4.4	3.4	3.8	3.3	2.2	2.0	3.0	3.5	2.5	1.7
PG2pow	4.0	2.3	2.4	2.7	2.0	1.5	1.7	1.6	1.1	1.1
PGnorta	4.4	4.1	3.9	3.4	2.7	1.8	2.2	2.4	2.3	2.4
PGnortaARM	4.4	4.0	4.2	4.0	3.2	1.8	2.3	2.5	2.7	2.2

How the models match the DI, for Center E



The DI for the models and data.

How the models match the correlations, for Center E

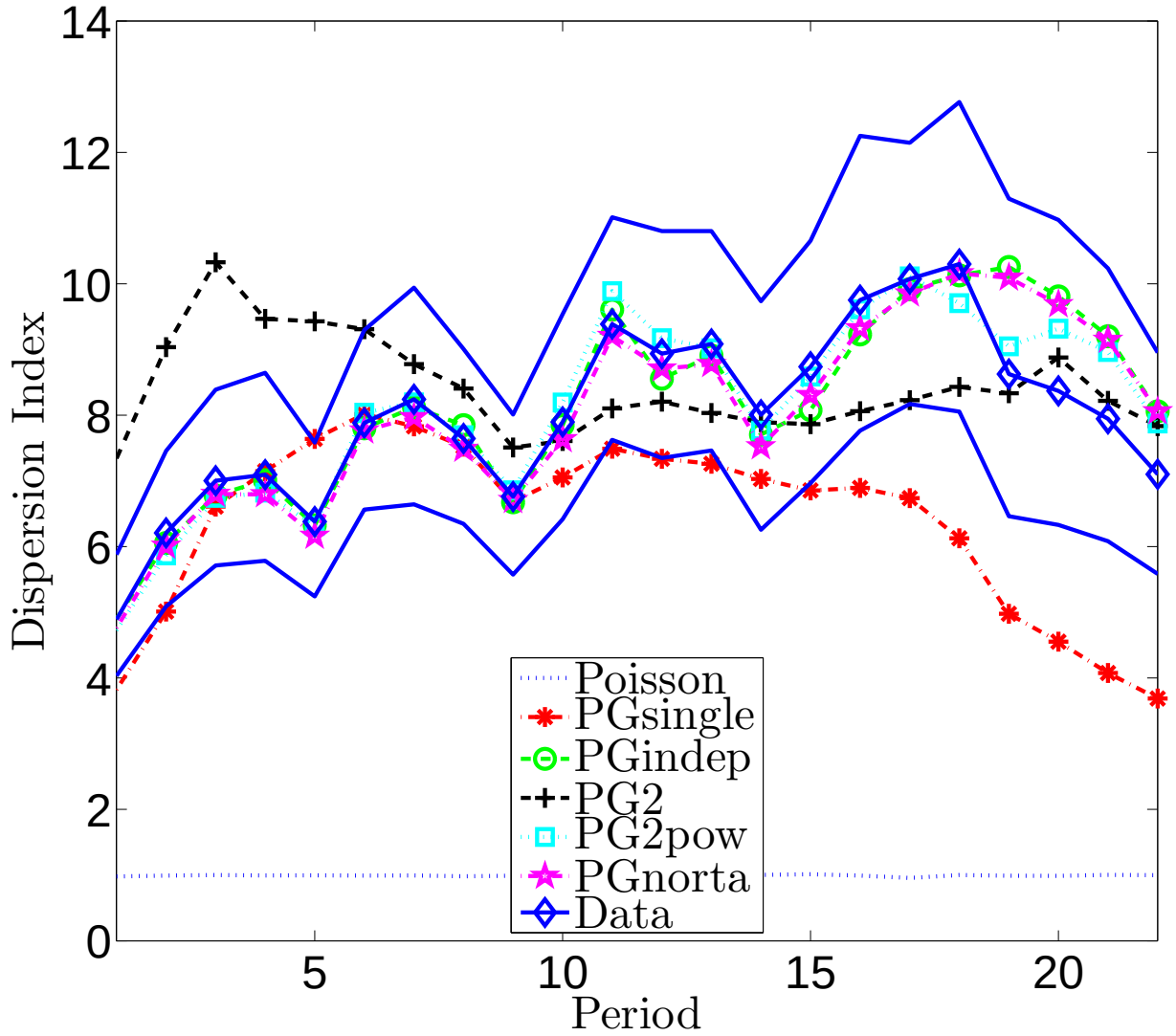


Sample correlation between past and future demand.

RMS Deviation of out-of-sample coverage probability, for call center E:

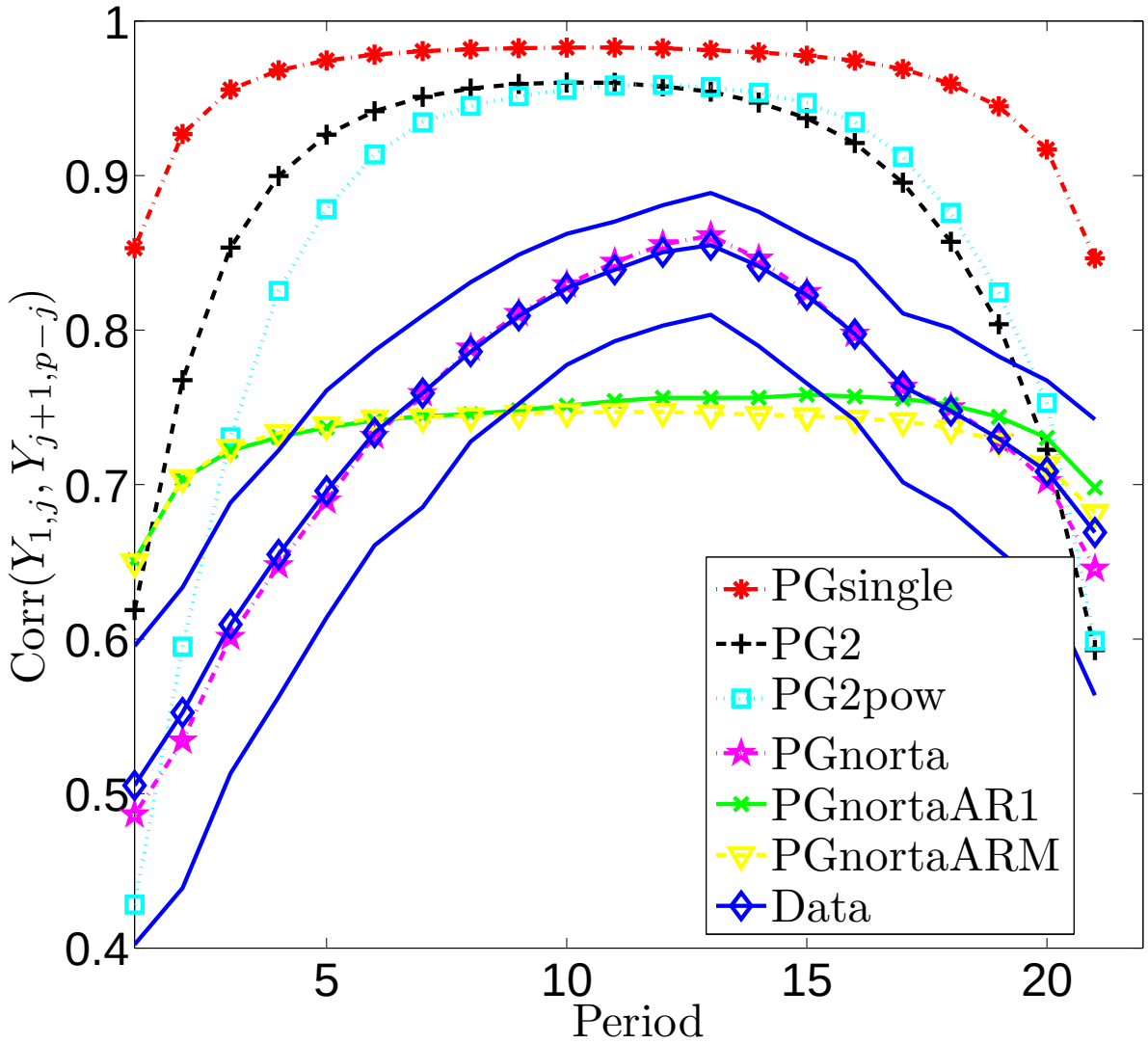
	75% target cover					90% target cover				
	0.5 h	1 h	2 h	4 h	8 h	0.5 h	1 h	2 h	4 h	8 h
Poisson	10.7	16.6	23.5	31.3	37.5	8.5	13.8	21.0	30.1	38.7
PGsingle	7.2	10.0	12.5	13.5	12.0	5.3	7.8	10.1	11.4	9.0
PGindep	1.3	5.3	12.7	21.4	29.9	0.8	4.1	10.2	18.7	29.1
PG2	2.1	4.9	8.7	11.4	11.6	1.6	3.8	6.8	9.4	8.8
PG2pow	1.5	2.9	4.4	5.1	5.0	1.0	2.0	3.1	3.4	3.0
PGnorta	1.3	1.7	1.7	1.7	1.3	0.8	1.1	1.2	1.3	0.8
PGnortaARM	1.3	2.4	3.4	4.3	4.5	0.9	1.5	2.2	2.7	2.8

How the models match the DI for Center B



The DI for the models and data.

How the models match the correlations for Center B



Sample correlation between past and future demand.

RMS Deviation of out-of-sample coverage probability, for call center B.

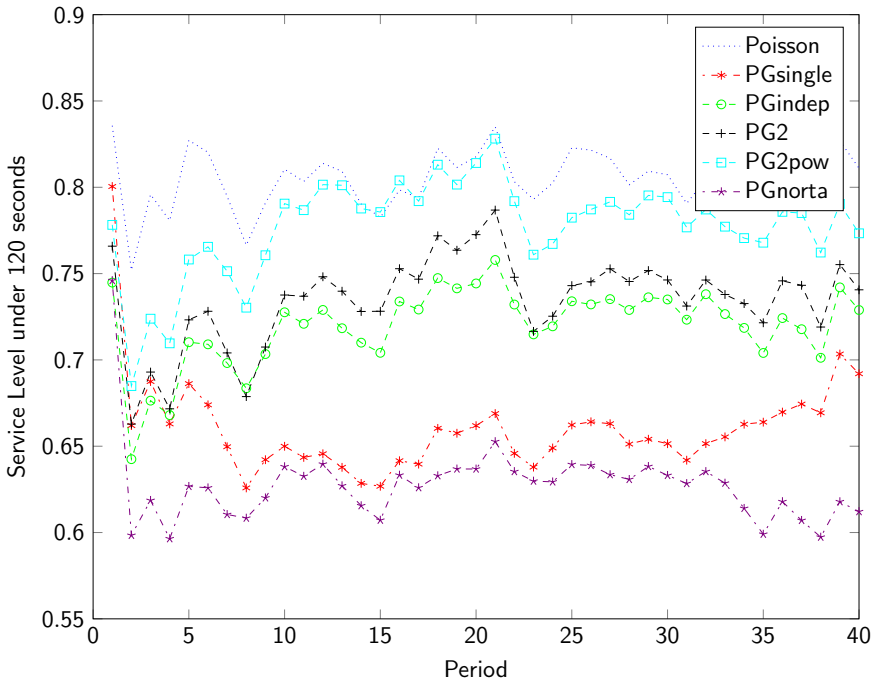
	75% target cover					90% target cover				
	0.5 h	1 h	2 h	4 h	8 h	0.5 h	1 h	2 h	4 h	8 h
Poisson	43.1	50.9	57.5	61.9	66.7	44.7	55.8	64.4	71.3	77.7
PGsingle	7.6	7.1	6.1	4.0	2.3	5.8	6.1	5.4	4.0	3.4
PGindep	3.1	13.2	27.3	39.3	48.7	2.0	12.1	26.4	41.2	51.9
PG2	4.8	4.1	5.1	4.3	2.6	3.0	2.9	3.3	2.9	2.3
PG2pow	2.5	3.3	4.1	2.0	0.8	1.7	3.3	3.7	2.8	2.7
PGnorta	3.2	3.0	2.7	1.2	1.3	2.0	2.4	2.2	1.7	2.0
PGnortaARM	3.2	3.1	2.8	1.9	0.5	2.0	2.4	2.3	2.2	2.8

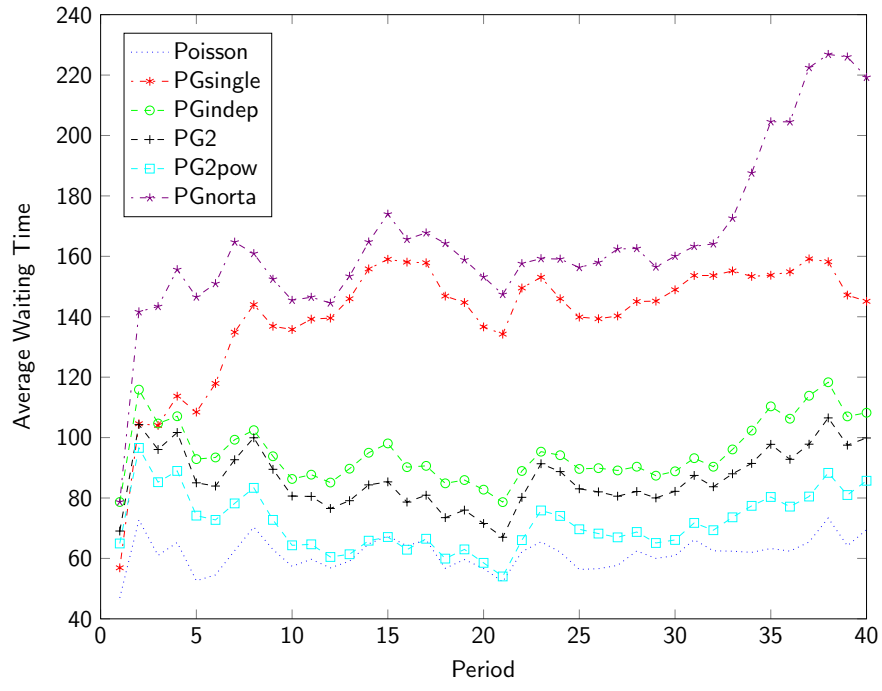
Impact of choice of arrival model

Take call center U on a week day. Single call type.
Lognormal service times with mean 206.4 and variance 23 667 (seconds).

Abandonment at rate $1/2443$ per second.
Staffing in each period: (16, 24, 31, 36, 43, 48, 51, 52, 56, 60, 62, 65, 67, 67, 66, 65, 62, 61, 60, 61, 64, 64, 63, 63, 64, 64, 64, 64, 65, 65, 64, 64, 62, 60, 58, 56, 53, 49, 48, 44).
Performance measures:
average waiting time (AWT);
service level (SL) with threshold $\tau = 120$ seconds.

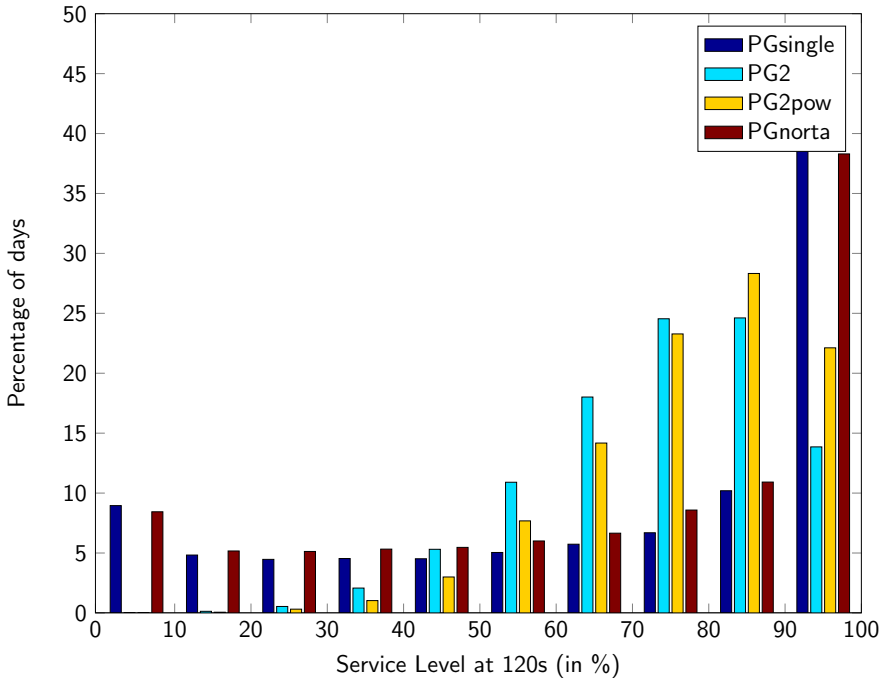
We simulated 10,000 days with each arrival model.

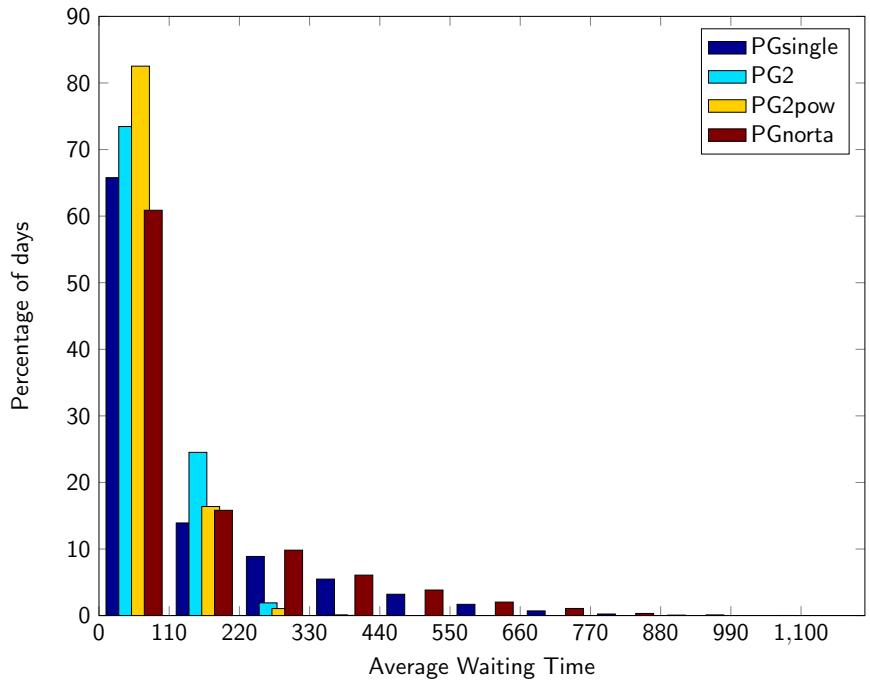




Evolution of the SL (left) and AWT in seconds (right) during the day for the Quebec utility society.

SL = proportion of calls answered within 120 seconds in the long run.





Histogram of the distribution of the daily SL (left) and daily AWT (right), with different models, for the Quebec utility society.

SL = proportion of calls answered within 120 seconds during the day.

More on arrival process modeling

Modeling the arrival rates over successive days.
Dependence between the days.
Seasonal effects (day of the week, period of the year).
Special days (holidays, special events, etc.).
External effects (weather, marketing campaigns, etc.).

More on arrival process modeling

Modeling the arrival rates over successive days.
Dependence between the days.
Seasonal effects (day of the week, period of the year).
Special days (holidays, special events, etc.).
External effects (weather, marketing campaigns, etc.).

Dependence between call types: the arrival rate should in fact be a multivariate process. Modeling via copulas.

More on arrival process modeling

Modeling the arrival rates over successive days.
Dependence between the days.
Seasonal effects (day of the week, period of the year).
Special days (holidays, special events, etc.).
External effects (weather, marketing campaigns, etc.).

Dependence between call types: the arrival rate should in fact be a multivariate process. Modeling via copulas.

Arrival bursts in emergency call center.

Modeling the service times

In call center U, the available data for service times is the number of calls of each type handled by each agent on each day, and the average duration of these calls. From this, we can estimate the mean and variance of a service times and match those to the mean and variance of a distribution such as lognormal or gamma.

Service times are usually not exponential.

Modeling the service times

In call center U, the available data for service times is the number of calls of each type handled by each agent on each day, and the average duration of these calls. From this, we can estimate the mean and variance of a service times and match those to the mean and variance of a distribution such as lognormal or gamma.

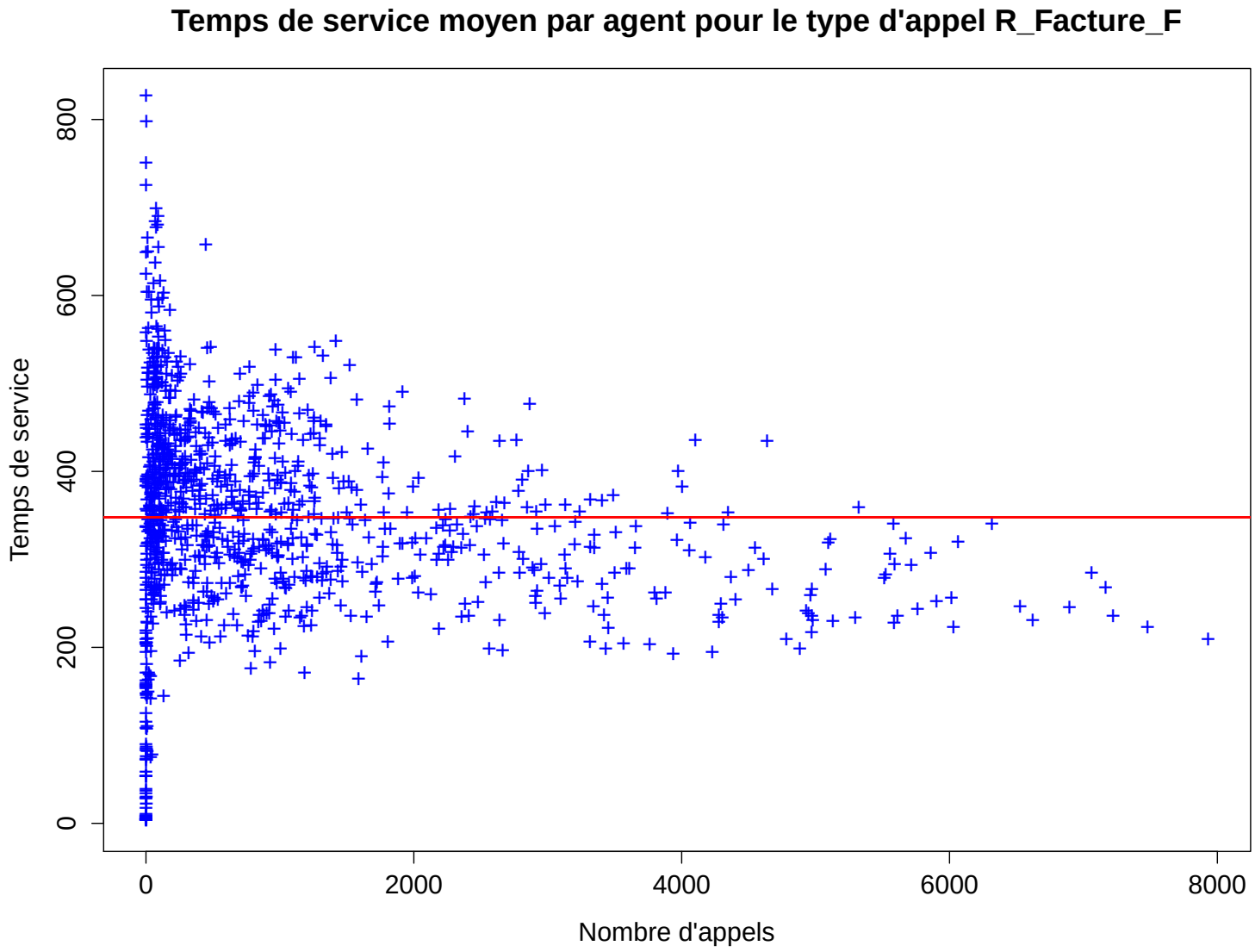
Service times are usually not exponential.

Common assumption: the distribution depends only on the call type.

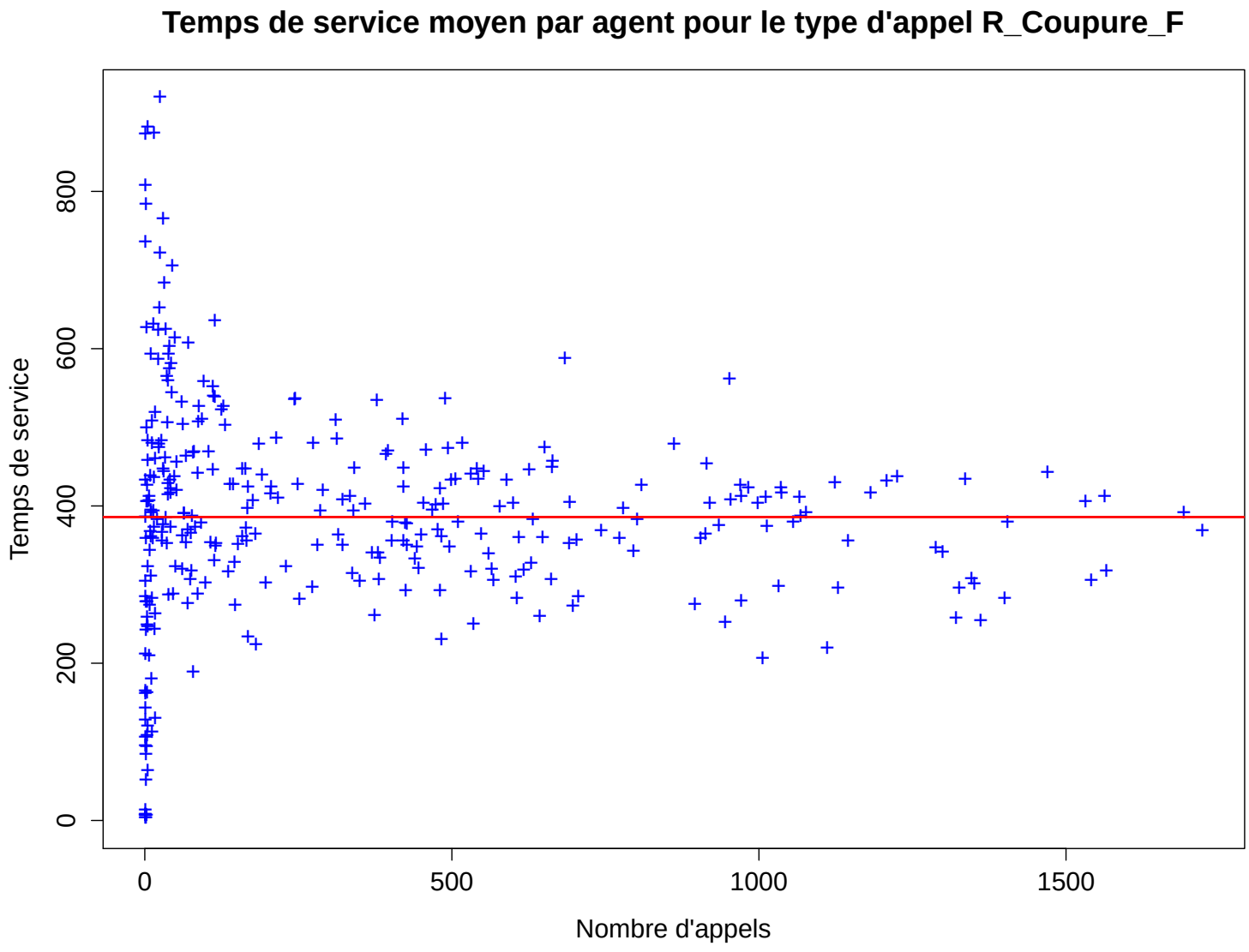
But on closer examination, we find that it depends on the individual agent, on the number of call types that the agent is handling, and may change with time (learning effect, motivation and mood of agent, etc.).

This is an important fact to consider when making work schedules!

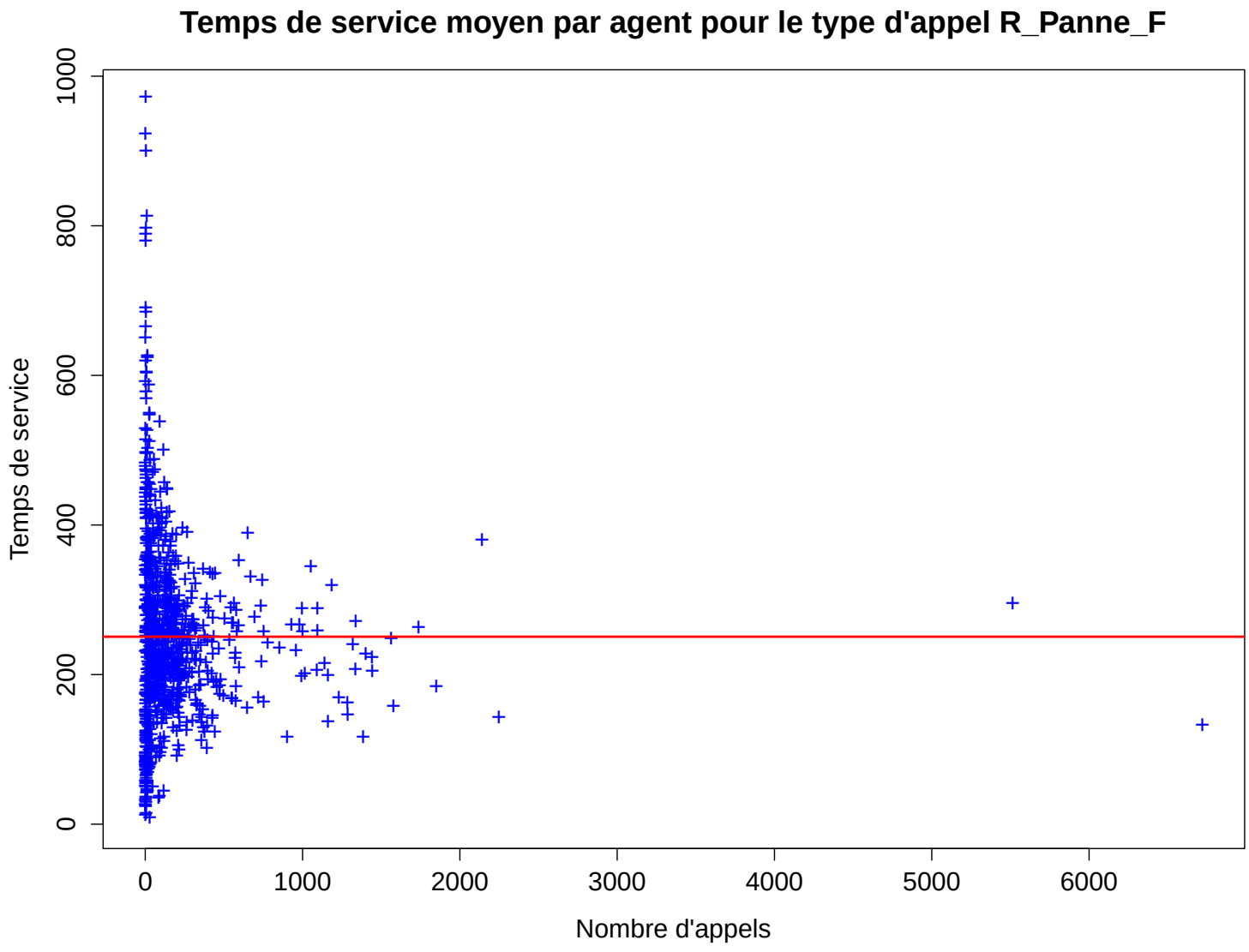
Average service time per agent for one call type, in center U (more than 1000 agents)



Another call type

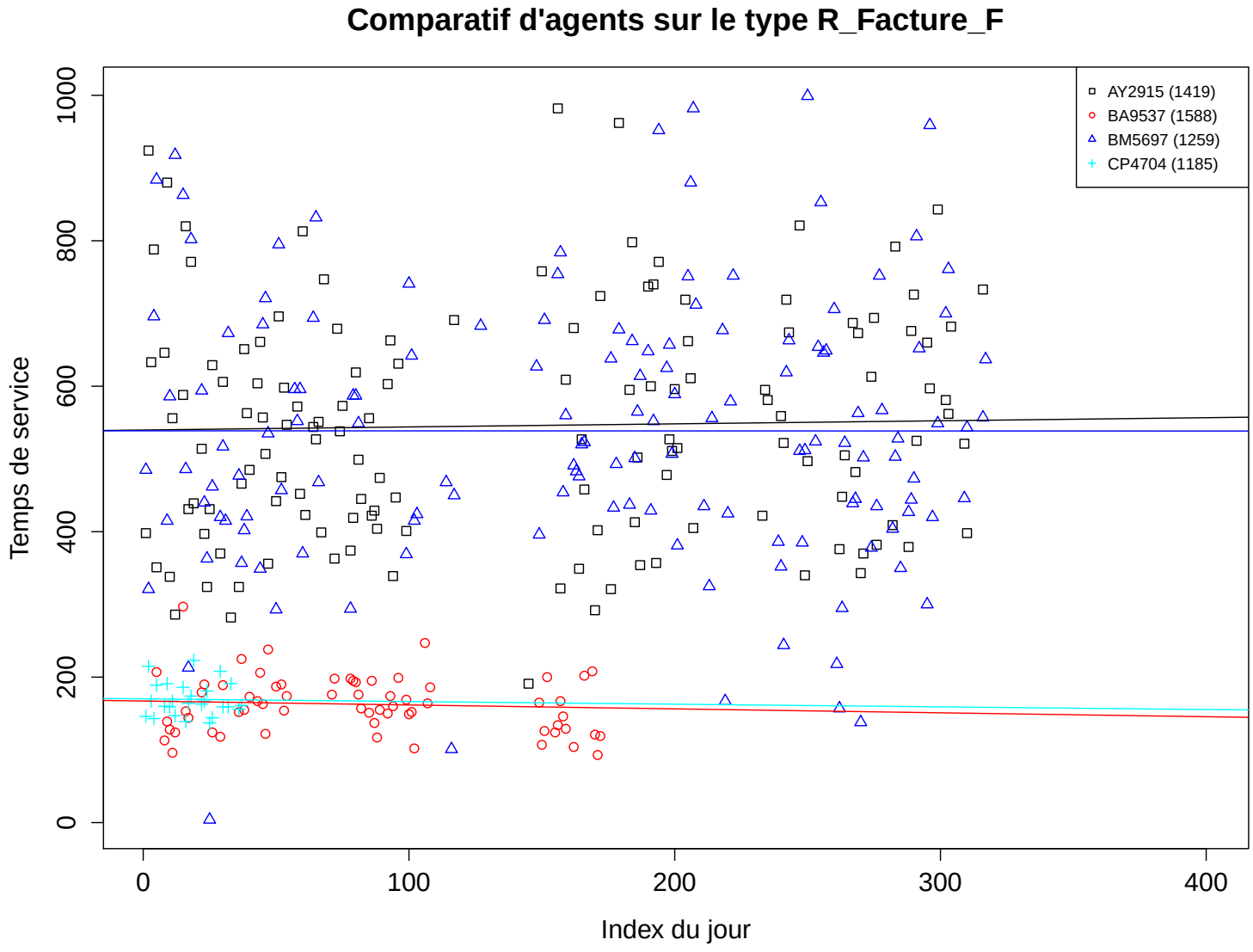


Another call type

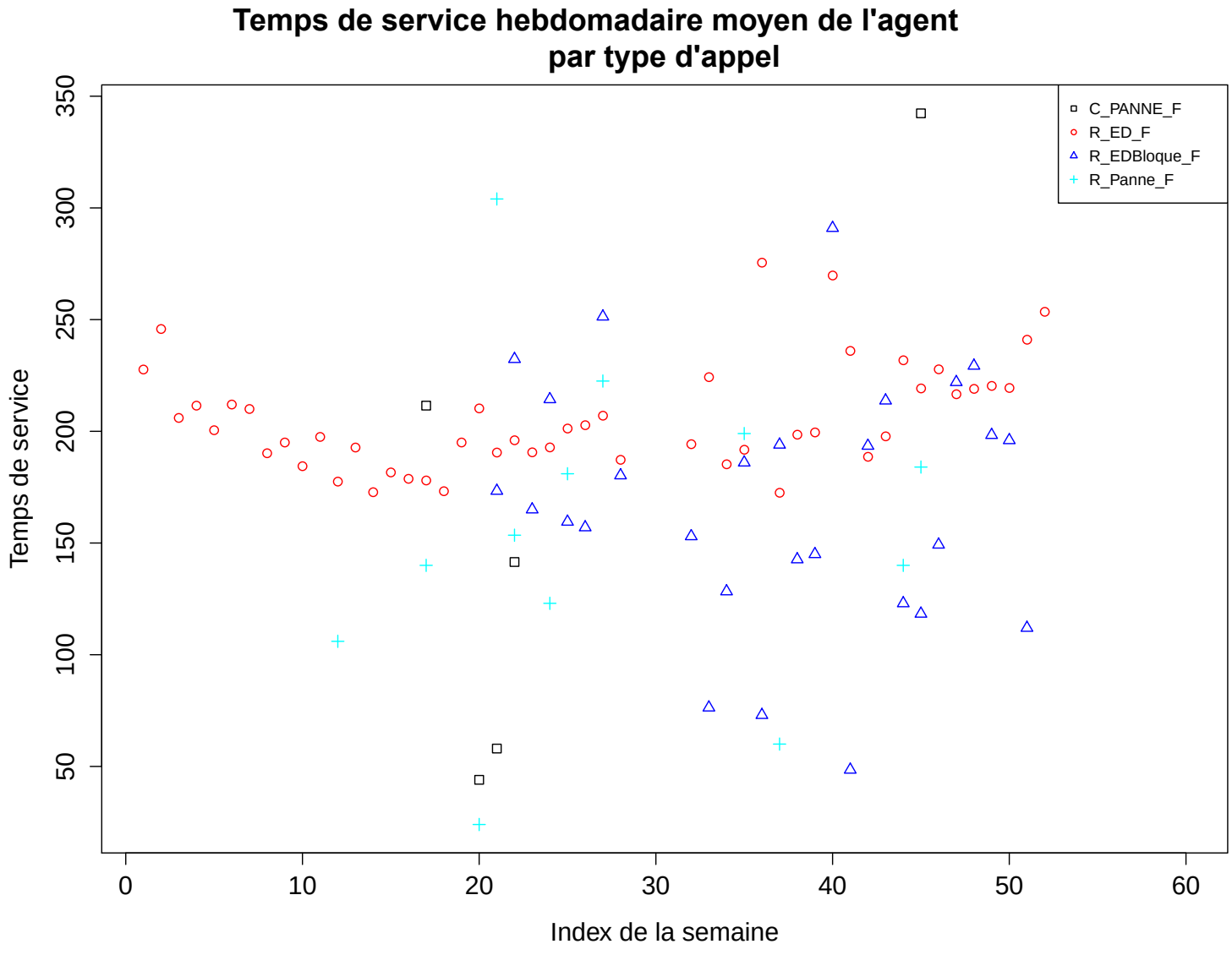


Four different agents, same call type

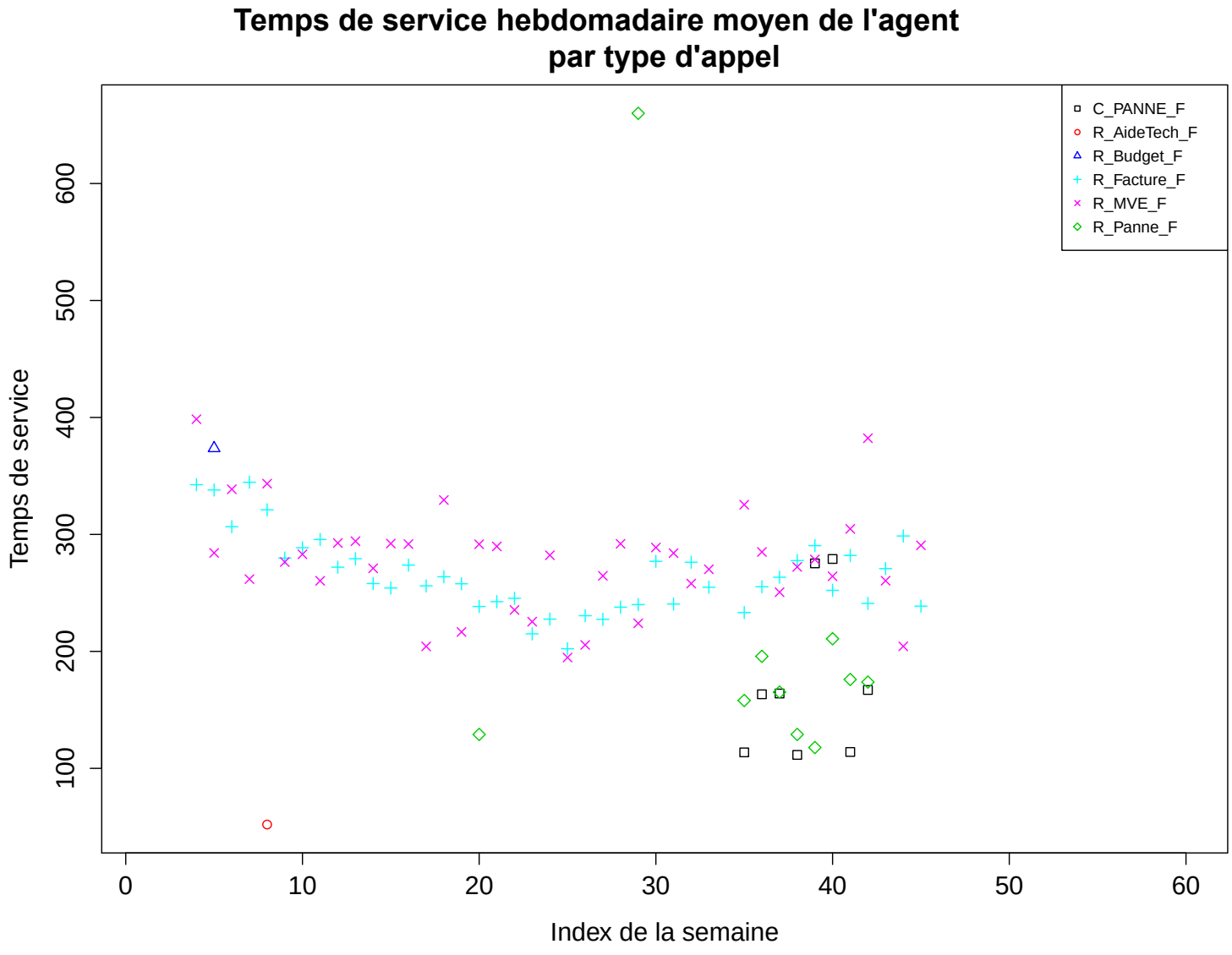
All have handled more than 1000 calls. Daily averages:



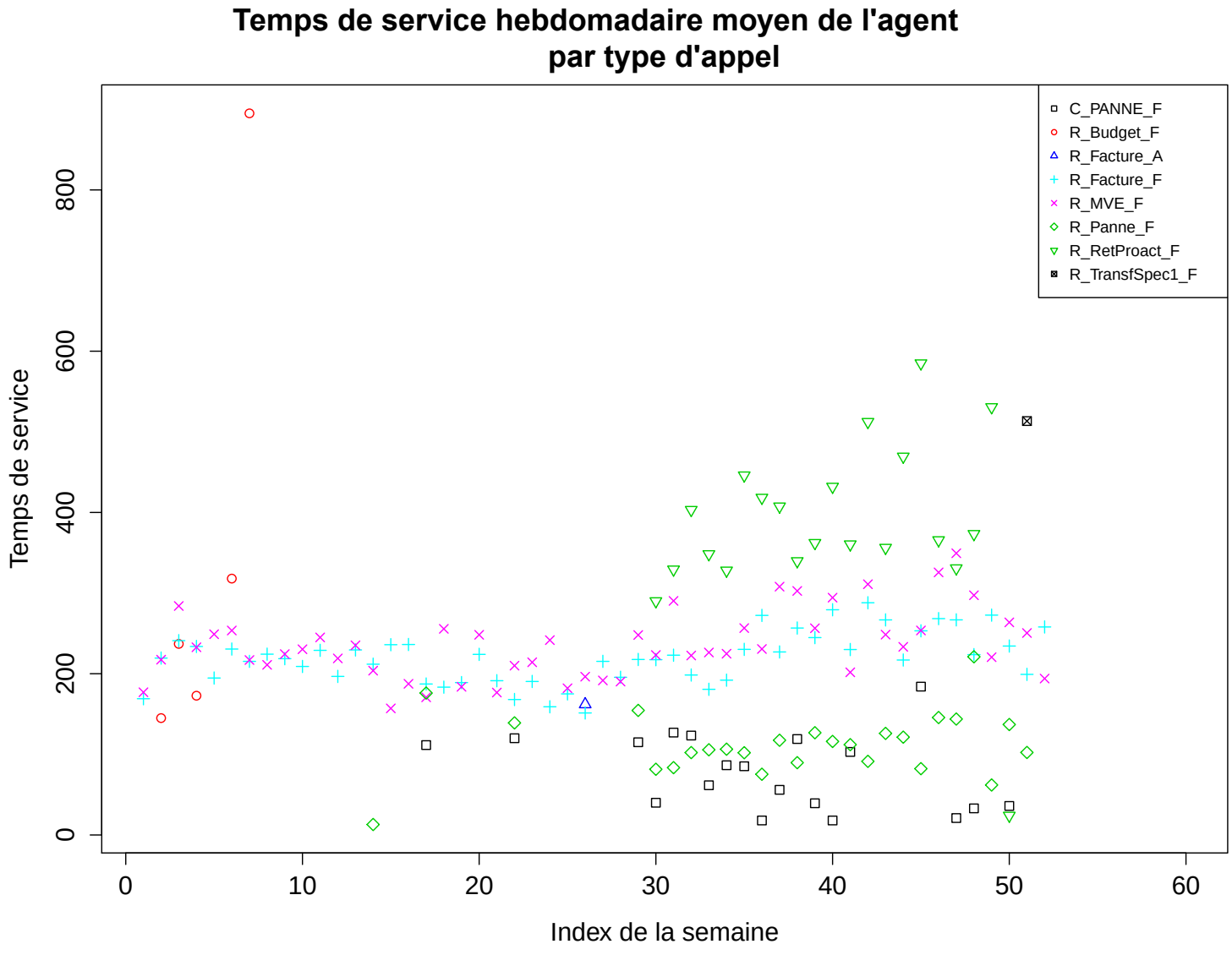
Same agent, 4 call types, weekly averages



Same agent, 6 call types



Same agent, 8 call types



Modeling the evolution of service time averages

For given agent and call type, day i:

$$M_i = \beta_{d_i} + \Gamma_{w_i} + \epsilon_i,$$

where d_i = type of day i, w_i = week of day i, and Γ_w is a random effect that may follow, e.g., an AR process:

$$\Gamma_w = \rho\Gamma_{w-1} + \psi_w.$$

The ϵ_i and ψ_w are residuals (noise).
Gives better predictions than just taking overall average for each agent.

For multiple call types, there can be a different Γ_w for each call type, or a single Γ_w for all call types (does better for our data set).

There could also be common effects across agents.

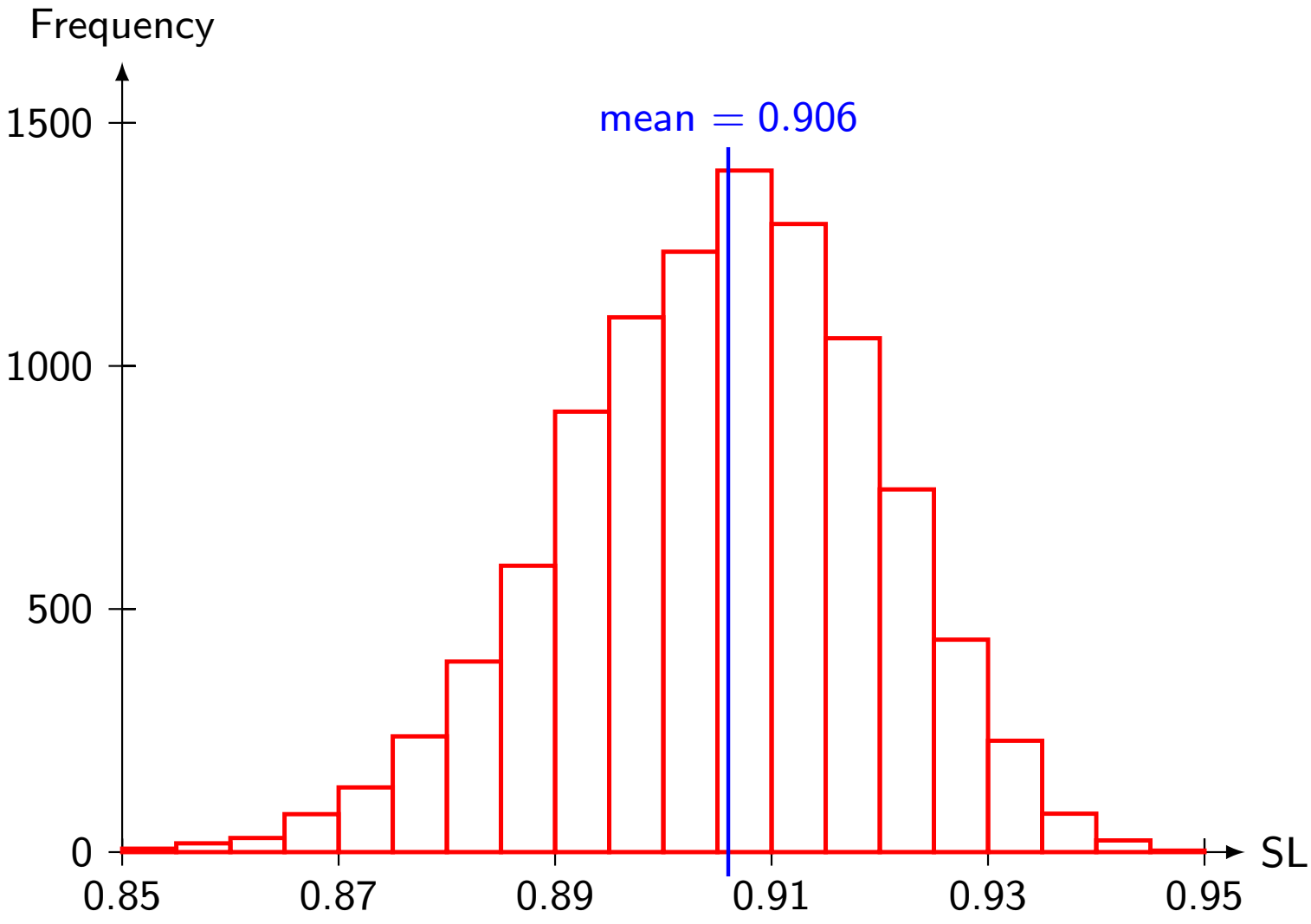
Better: model the evolution of all distribution parameters.

Performance measures and optimization

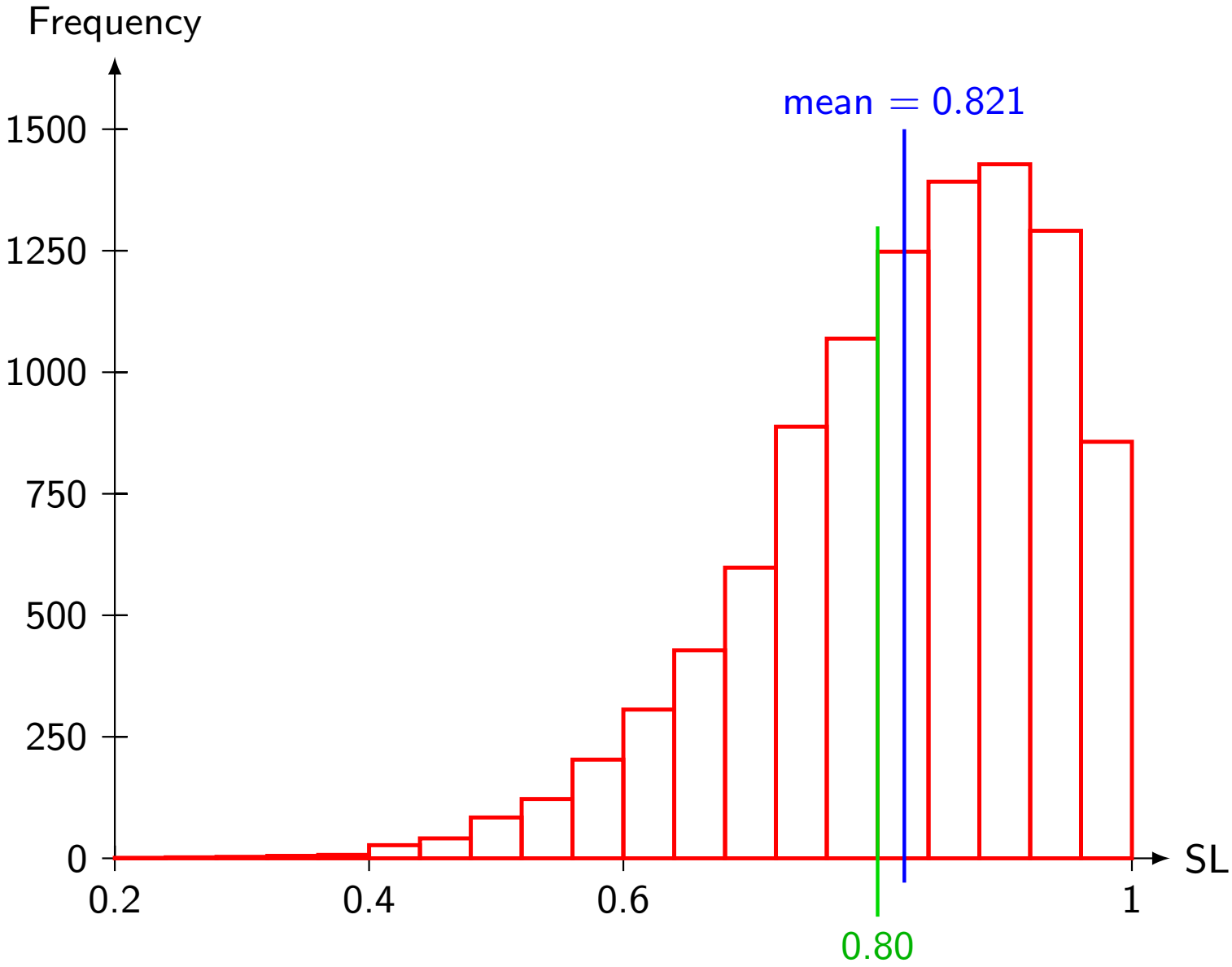
For a given staffing and routing strategy, the SL on a given day (or given period) is a random variable. We may be interested in its distribution.

What if we pay a penalty iff the SL is below a given number today?

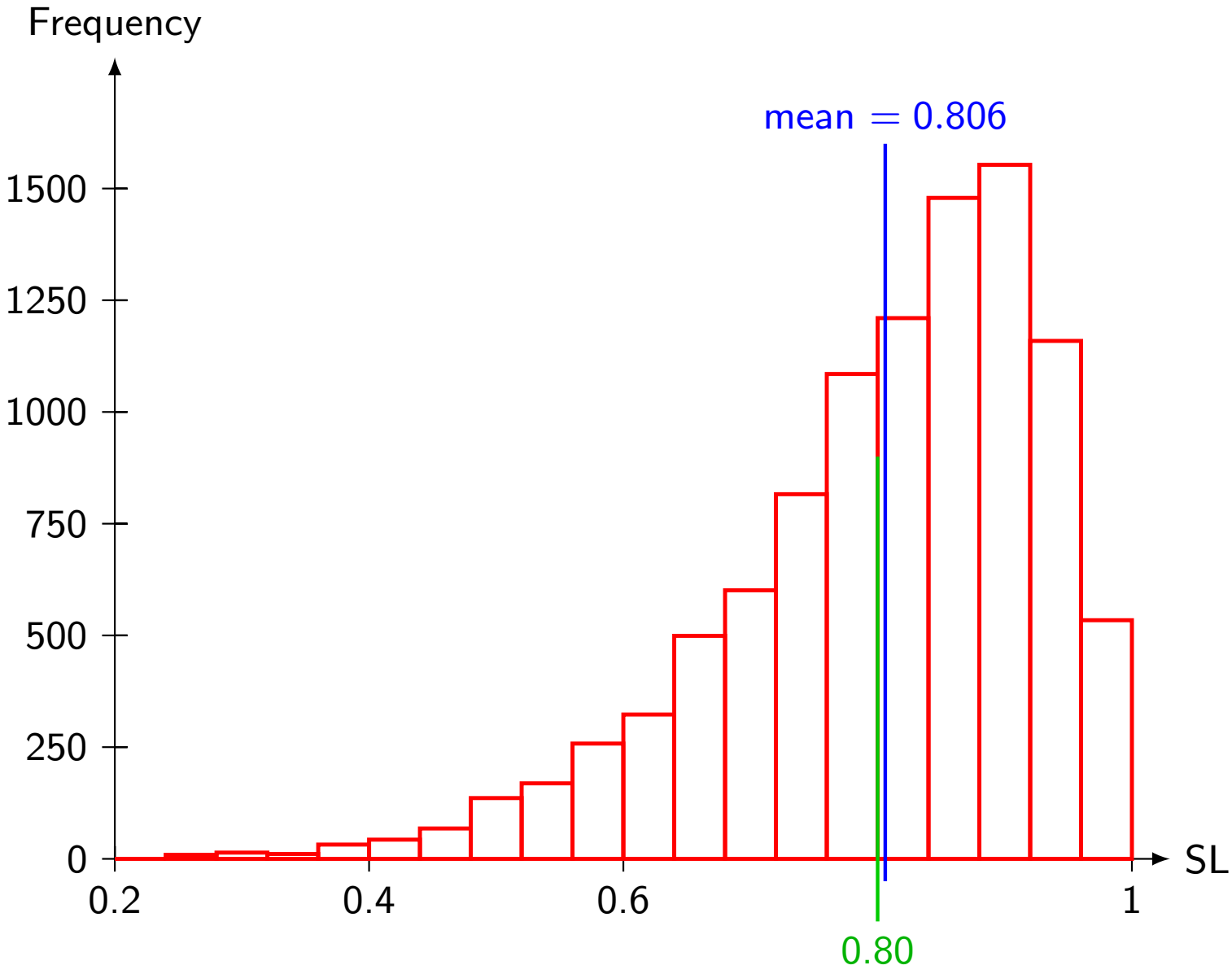
After solving some work-schedule optimization problem in some call center, we re-simulated with our best feasible solution for 10000 days, and computed the empirical distribution of the SL.



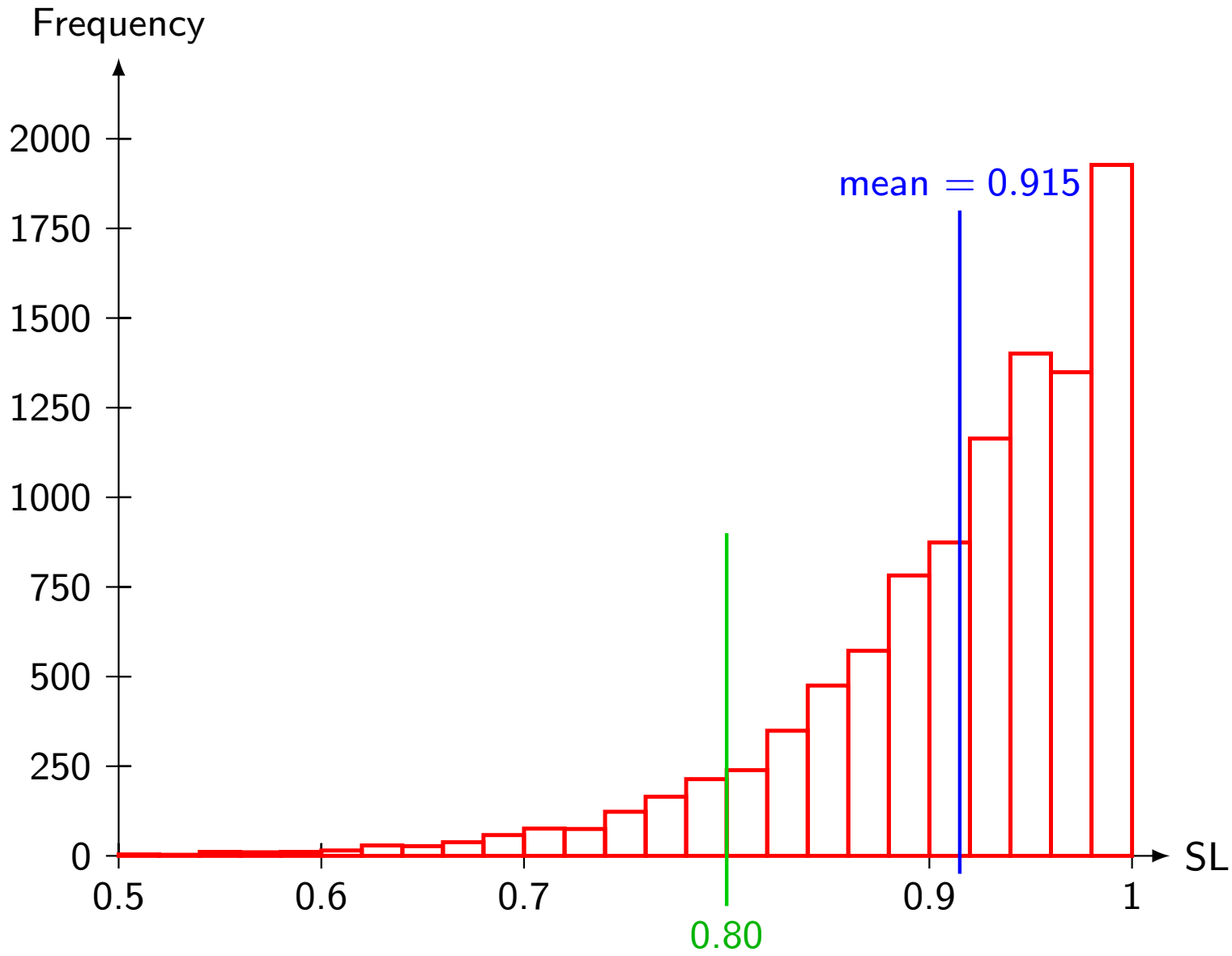
Distribution of Global SL



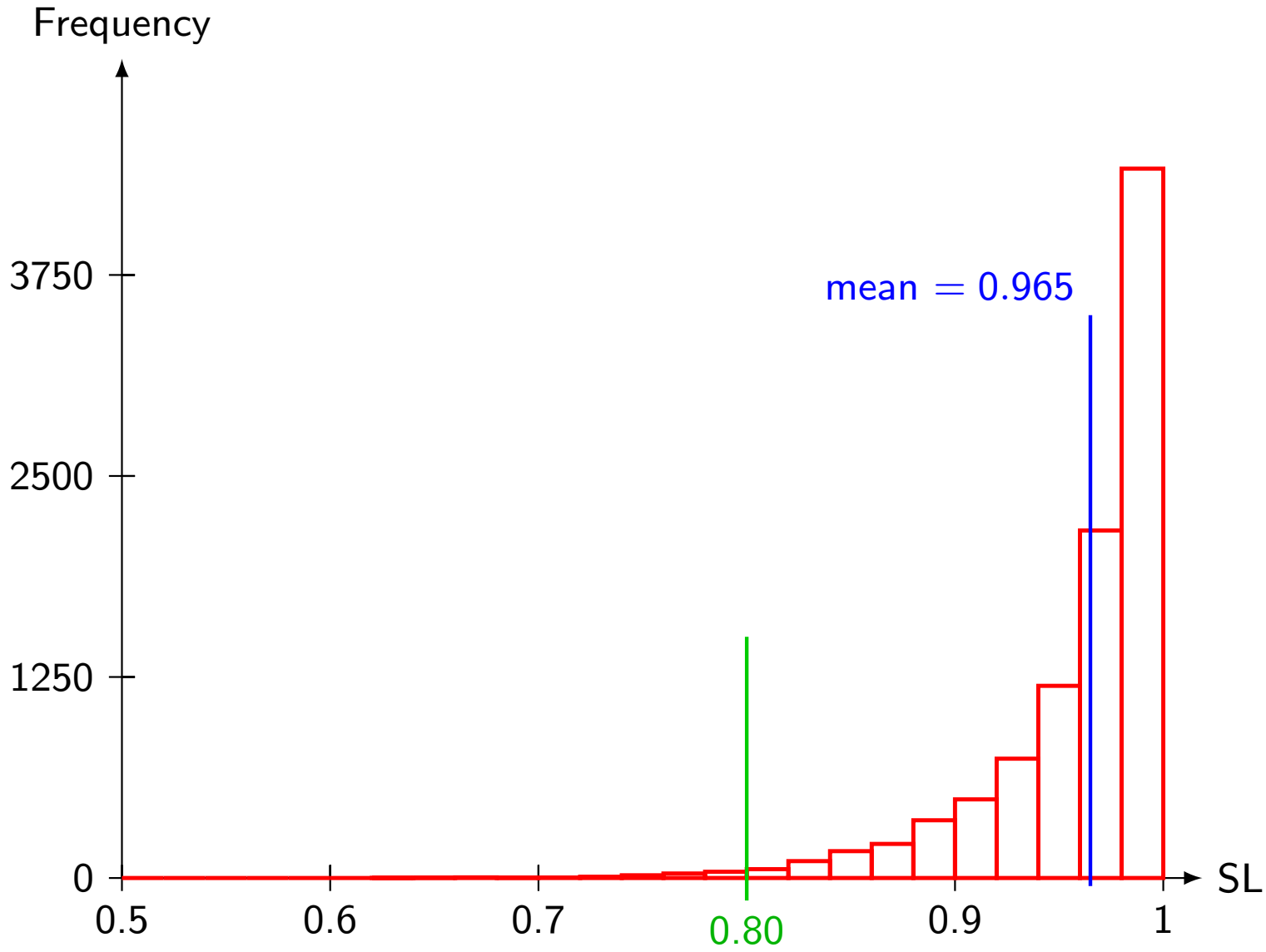
Distribution of SL Period: 7



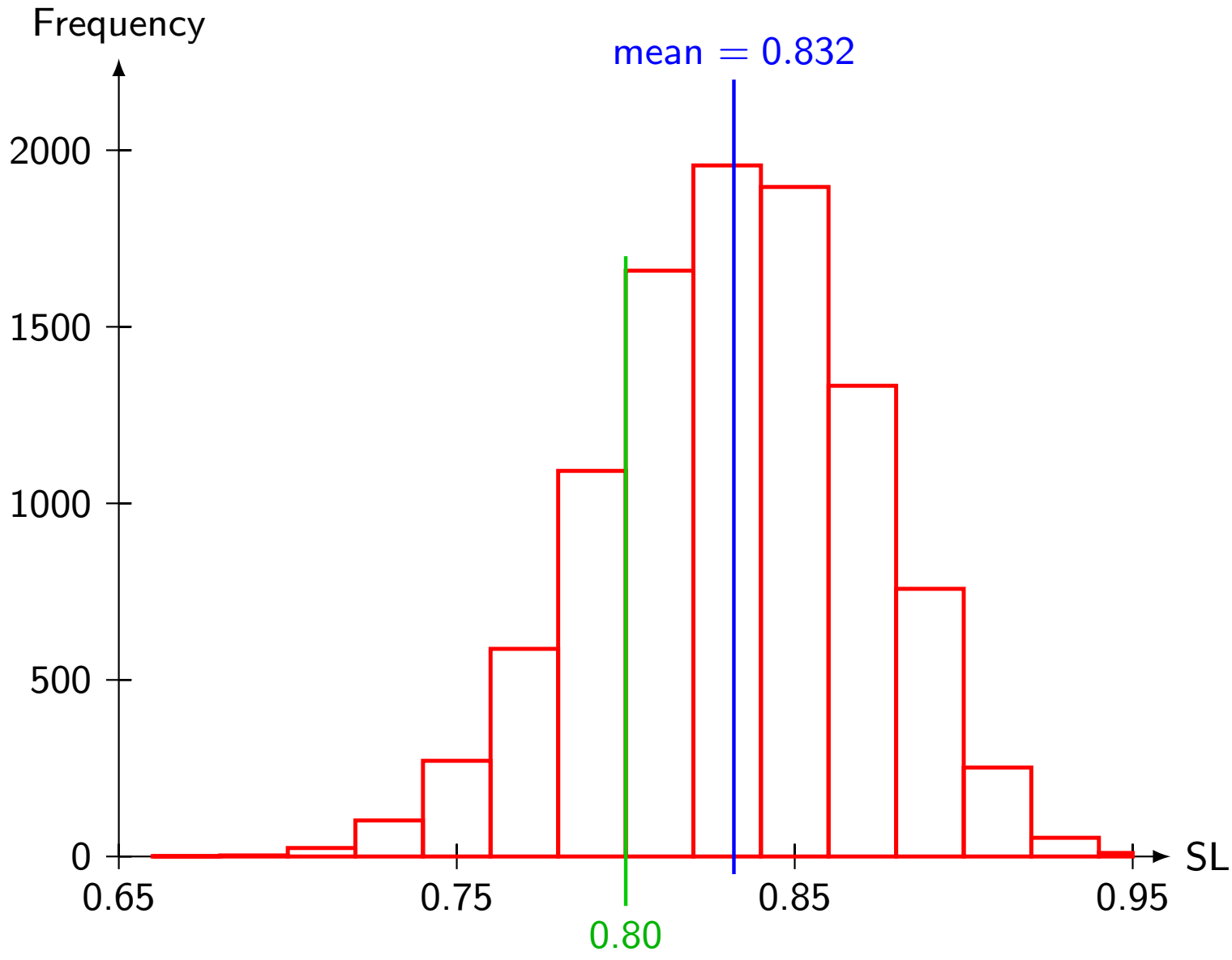
Distribution of SL Period: 17



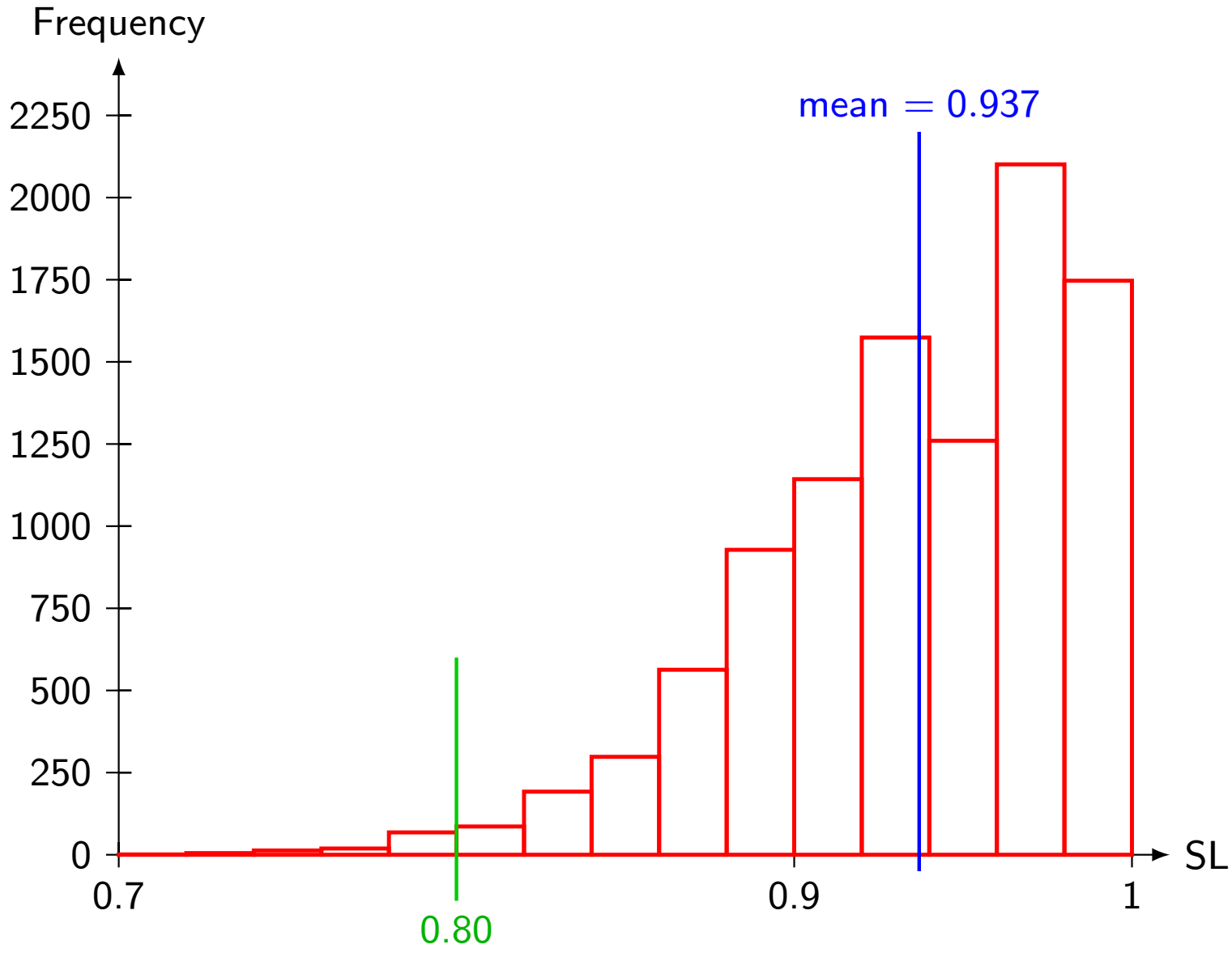
Distribution of SL Period: 11



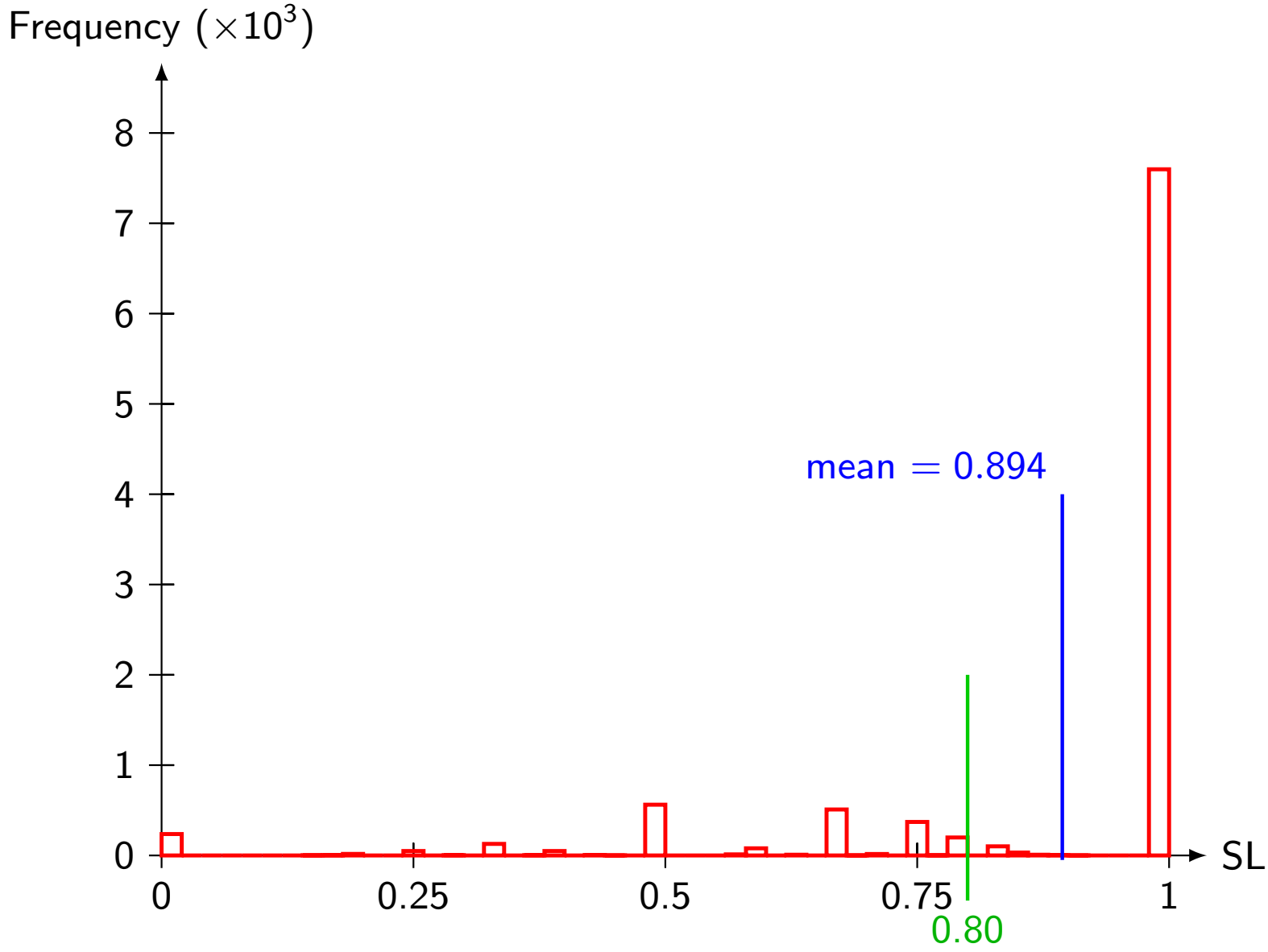
Distribution of SL Period: 49



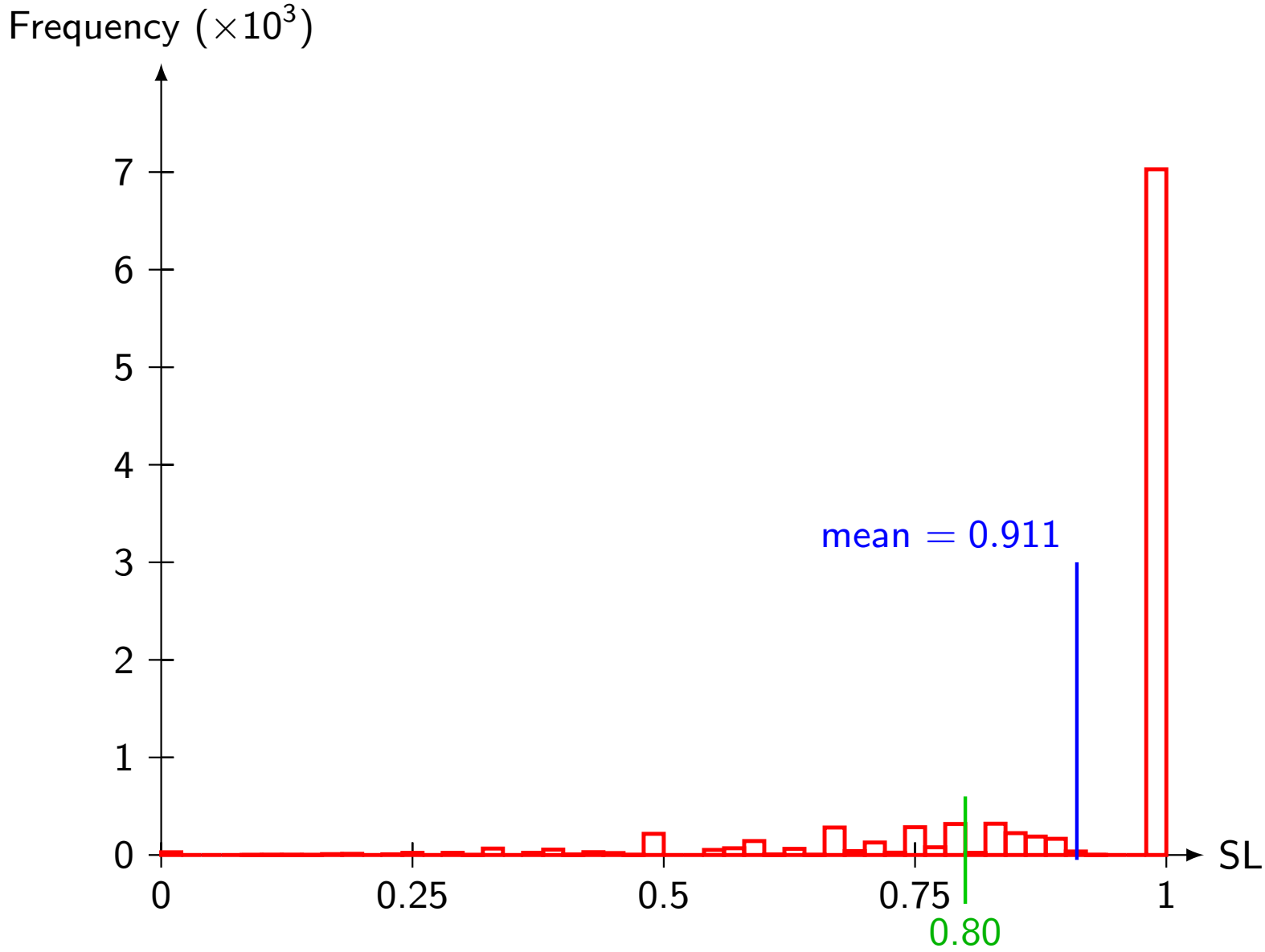
Distribution of SL for Call Type 3



Distribution of SL for Call Type 17



Distribution of SL for Call Type 2 in Period 20



Distribution of SL for Call Type 12 in Period 30

Example of scheduling optimization problem

Suppose the routing rules are fixed.

Several call types, several agent types, several time periods.

Example of scheduling optimization problem

Suppose the routing rules are fixed.

Several call types, several agent types, several time periods.

A shift type specifies the time when the agent starts working, when he/she finishes, and all the lunch and coffee breaks.
$c_{s,q}$ = cost of an agent of type s having shift type q.

Example of scheduling optimization problem

Suppose the routing rules are fixed.

Several call types, several agent types, several time periods.

A shift type specifies the time when the agent starts working, when he/she finishes, and all the lunch and coffee breaks.
$c_{s,q}$ = cost of an agent of type s having shift type q.

The decision variables $\mathbf{x}$ and $\mathbf{z}$ are:
(i) $x_{s,q}$ = number of agents of type i having shift type q;
(ii) $z_{\ell,s,j}$ = number of agents of type ℓ that work as type-s agents in period j, with $S_s \subset S_\ell$ (they use only part of their skills).

This determines indirectly the staffing vector $\mathbf{y}$, where $y_{s,j}$ = num. agents of type s in period j, and $a_{j,q} = 1$ iff shift q covers period j:

$$y_{s,j} = \sum_q a_{j,q} x_{s,q} + \sum_{l \in \mathcal{S}_s^+} z_{l,s,j} - \sum_{l \in \mathcal{S}_s^-} z_{s,l,j} \qquad \text{for all } s, j.$$

Scheduling Optimization Problem

$\mathbf{x}$ = vector of shifts; $\mathbf{c}$ = their costs; $\mathbf{y}$ = staffing vector;

(Long-run) service level for type k in period j (depends on entire vector $\mathbf{y}$):

$$g_{k,j}(\mathbf{y}) = \frac{\mathbb{E}[\text{num. calls type } k \text{ in period } j \text{ answered within time limit}]}{\mathbb{E}[\text{num. calls type } k \text{ in period } j\text{, ans., or abandon. after limit}]}.$$

(P0) : [Scheduling problem]

$$\begin{aligned}
\min \quad & \mathbf{c}^t\mathbf{x} = \textstyle\sum_{s=1}^{I}\sum_{q=1}^{Q} c_{s,q}x_{s,q} \\
\text{subject to} \quad & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{y}, \\
& g_{k,j}(\mathbf{y}) \geq l_{k,j} \qquad \text{for all } k, j, \\
& g_j(\mathbf{y}) \geq l_j \qquad \text{for all } j, \\
& g_k(\mathbf{y}) \geq l_k \qquad \text{for all } k, \\
& g(\mathbf{y}) \geq l, \\
& \mathbf{x} \geq 0, \mathbf{z} \geq 0, \mathbf{y} \geq 0, \text{ and integer.}
\end{aligned}$$

Sample-path optimization via simulation

We simulate n independent operating days of the center, to estimate the functions g.

Let ω represent the source of randomness, i.e., the sequence of independent uniform r.v.'s underlying the entire simulation (n runs).

The empirical SL's over the n simulation runs are:

$\hat{g}_{n,k,j}(\mathbf{y}, \omega)$ for call type k in period j;
$\hat{g}_{n,j}(\mathbf{y}, \omega)$ aggregated over period j;
$\hat{g}_{n,k}(\mathbf{y}, \omega)$ aggregated for call type k;
$\hat{g}_{n}(\mathbf{y}, \omega)$ aggregated overall.

For a fixed ω, these are deterministic functions of $\mathbf{y}$.

We replace the (unknown) functions $g(\cdot)$ by $\hat{g}(\cdot, \omega)$ and optimize.

To compute them at different values of $\mathbf{y}$, we use simulation with well-synchronized common random numbers. Discuss.

Empirical (sample) scheduling optimization problem

$(SP0_n)$: [Sample scheduling problem]

$$\min \quad \mathbf{c}^t\mathbf{x} = \sum_s \sum_{q=1}^{Q} c_{s,q} x_{s,q}$$

$$\text{subject to} \quad \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{y},$$
$$\hat{g}_{n,k,j}(\mathbf{y}) \geq l_{k,j} \qquad \text{for all } k, j,$$
$$\hat{g}_{n,j}(\mathbf{y}) \geq l_j \qquad \text{for all } j,$$
$$\hat{g}_{n,k}(\mathbf{y}) \geq l_k \qquad \text{for all } k,$$
$$\hat{g}_n(\mathbf{y}) \geq l,$$
$$\mathbf{x} \geq 0, \mathbf{z} \geq 0, \text{ and integer.}$$

Theorem: When $n \to \infty$, the optimal solution of $SP0_n$ converges w.p.1 to that of P0. Moreover, if a standard large deviation principle holds for $\hat{g}$ (which is typical), the probability that the two solutions differ converges to 0 exponentially with n. [Adaptation of Vogel 1994, for example.]

Solving the sample optimization problem

Integer programming with cutting planes.

[Atlason, Epelman, and Henderson, 2004; Cezik and L'Ecuyer 2005]

Replace the nonlinear constraints in $\mathrm{SP0}_n$ by a set of linear constraints.
This gives an integer program (IP).

We start with a relaxation of the IP problem (fewer constraints).
Then, at each step, use simulation to compute the service levels in $\mathrm{SP0}_n$ for the optimal solution $\bar{\mathbf{y}}$ of the current IP.
For each SL constraint that is not satisfied, add a cut based on estimated subgradient.
Stop when all SL constraints of $\mathrm{SP0}_n$ are satisfied.

Solving the sample optimization problem

Integer programming with cutting planes.
[Atlason, Epelman, and Henderson, 2004; Cezik and L'Ecuyer 2005]
Replace the nonlinear constraints in SP0_n by a set of linear constraints.
This gives an integer program (IP).

We start with a relaxation of the IP problem (fewer constraints).
Then, at each step, use simulation to compute the service levels in SP0_n for the optimal solution $\bar{\mathbf{y}}$ of the current IP.
For each SL constraint that is not satisfied, add a cut based on estimated subgradient.
Stop when all SL constraints of SP0_n are satisfied.

In practice, for large problems, we solve the IP as an LP and round the solution (at each step, to be able to simulate). We select a rounding threshold δ (usually around 0.5 or 0.6). Heuristic!

Solving the sample optimization problem

Integer programming with cutting planes.
[Atlason, Epelman, and Henderson, 2004; Cezik and L'Ecuyer 2005]
Replace the nonlinear constraints in $SP0_n$ by a set of linear constraints.
This gives an integer program (IP).

We start with a relaxation of the IP problem (fewer constraints).
Then, at each step, use simulation to compute the service levels in $SP0_n$ for the optimal solution $\bar{\mathbf{y}}$ of the current IP.
For each SL constraint that is not satisfied, add a cut based on estimated subgradient.
Stop when all SL constraints of $SP0_n$ are satisfied.

In practice, for large problems, we solve the IP as an LP and round the solution (at each step, to be able to simulate). We select a rounding threshold δ (usually around 0.5 or 0.6). Heuristic!

Phase II: run longer simulation to perform a local adjustment to the final solution, using heuristics (add, remove, switch).

Other objectives and constraints (alternative formulations)

Chance constraints: Replace long-term average $g_{k,j}(\mathbf{y})$ by a tail probability of the service level, e.g.:

$$\mathbb{P}[\mathsf{SL}_{k,j}(\tau) \geq l_{k,j}] \geq \alpha_{k,j} \qquad \text{for all } k, j.$$

Can use sample average approximation (SAA).
Not easy to solve the SAA. Cutting planes, Benders decomposition, ...
Ongoing PhD thesis of Anh Thuy Ta.

Other objectives and constraints (alternative formulations)

Chance constraints: Replace long-term average $g_{k,j}(\mathbf{y})$ by a tail probability of the service level, e.g.:

$$\mathbb{P}[\mathsf{SL}_{k,j}(\tau) \geq l_{k,j}] \geq \alpha_{k,j} \qquad \text{for all } k, j.$$

Can use sample average approximation (SAA).
Not easy to solve the SAA. Cutting planes, Benders decomposition, ...
Ongoing PhD thesis of Anh Thuy Ta.

Optimizing call routing rules. Chan, Koole, L'Ecuyer, in Manufacturing and Service Operations Management (2014).

Replace constraints by penalties.

Etc.

Conclusion

Simulation and optimization can be useful only to the extent that we can trust the model.

We can do more and more simulation runs and compute arbitrarily tight confidence intervals on certain unknown quantities, but this can be meaningless if the simulation model is not representative.

Huge masses of data are becoming available, at a rate never seen before. Exploiting this data to build credible and valid stochastic models of complex systems is in my opinion the biggest challenge that we now face for simulation.

References for the material of this talk:

- B. N. Oreshkin, N. Régnard, and P. L'Ecuyer, "Rate-Based Daily Arrival Process Models with Application to Call Centers", Operations Research, **64**, 2, 510–527, 2016.
- R. Ibrahim, P. L'Ecuyer, H. Shen, and M. Thiongane, "Inter-Dependent, Heterogeneous, and Time-Varying Service-Time Distributions in Call Centers", European Journal of Operational Research, **250** (2016), 480–492
- R. Ibrahim, H. Ye, P. L'Ecuyer, and H. Shen, "Modeling and Forecasting Call Center Arrivals: A Literature Survey and a Case Study", International Journal of Forecasting, **32**, 3, 865–874, 2016.
- R. Ibrahim and P. L'Ecuyer, "Forecasting Call Center Arrivals: Fixed-Effects, Mixed-Effects, and Bivariate Models," Manufacturing and Service Operations Management, **15**, 1 (2013), 72–85.
- A. Jaoua, P. L'Ecuyer and L. Delorme, "Call-Type Dependence in Multiskill Call Centers", Simulation: Transactions of the Society for Modeling and Simulation International, **89**, 6 (2013), 722–734.
- R. Ibrahim, P. L'Ecuyer, N. Régnard, and H. Shen, "On the Modeling and Forecasting of Call Center Arrivals", Proceedings of the 2012 Winter Simulation Conference, IEEE Press, 2012, 256–267.

- N. Channouf and P. L'Ecuyer, "A Normal Copula Model for the Arrival Process in Call Centers," International Transactions in Operational Research, **19** (2012), 771–787.
- A. N. Avramidis, W. Chan, M. Gendreau, P. L'Ecuyer, and O. Pisacane, "Agent Scheduling in a Multiskill Call Center," European J. of Operations Research, 200, 3 (2010) 822–832.
- T. Cezik and P. L'Ecuyer, "Staffing Multiskill Call Centers via Linear Programming and Simulation", Management Science, **54**, 2 (2008), 310–323.
- A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer, "Modeling Daily Arrivals to a Telephone Call center", Management Science, **50**, 7 (2004), 896–908.
- A. N. Avramidis and P. L'Ecuyer, "Modeling and Simulation of Call Centers", Proceedings of the 2005 Winter Simulation Conference, IEEE Press, 2005, 144–152.
- W. Chan, T. A. Ta, P. L'Ecuyer, and F. Bastin, "Two-stage chance-constrainted staffing with agent recourse for multi-skill call centers," Proceedings of the 2016 Winter Simulation Conference, IEEE Press, 2016, 3189–3200.
- T. A. Ta, P. L'Ecuyer, and F. Bastin, "Staffing Optimization with Chance Constraints for Emergency Call Centers," MOSIM 2016: the 11th International Conference on Modeling, Optimization and Simulation, Montreal, 2016.