EMS performance evaluation with analytical stochastic models

Armann Ingolfsson,
armann.ingolfsson@ualberta.ca
University of Alberta School of Business

1st International Workshop on Planning of Emergency Services: Theory and Practice,
CWI, Amsterdam, 26 June 2014

The answer to Rob's question from yesterday …

CV Assumptions

- For *an average call*, travel time T_{avg} has: variance = $b_0 + b_1 \times$ mean ← Herman and Lam (1974)
- Random variable B (mean 1, variance b_2) captures *call-to-call variability* via $T = B \times T_{avg}$, where T is travel time for a randomly chosen call
- B and T_{avg} are independent
- Functional form of CV vs. median relation is the same as for CV vs. mean

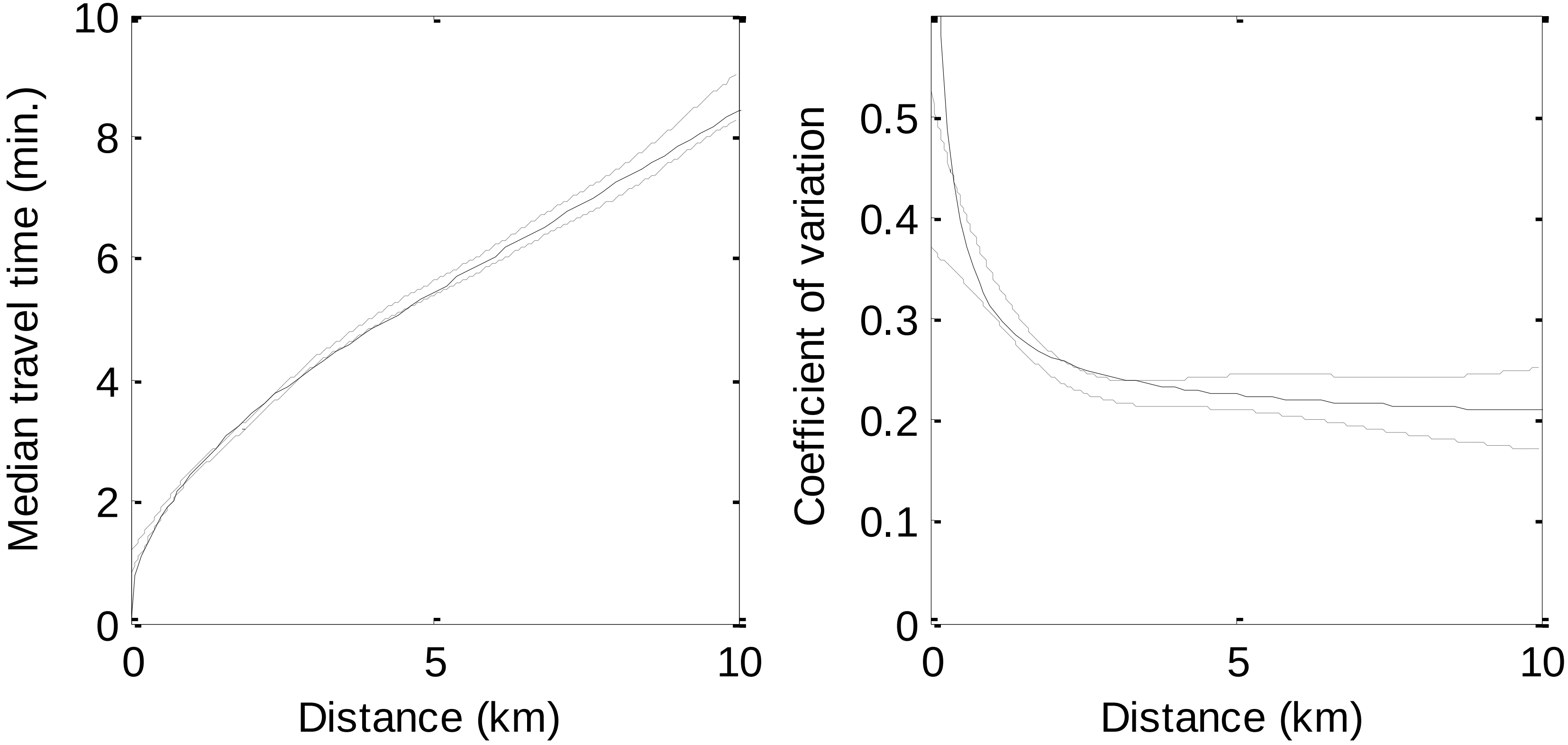
Budge, S., Ingolfsson, A., & Zerom, D. (2010). Empirical analysis of ambulance travel times: the case of Calgary emergency medical services. Management Science, 56(4), 716-723.

Herman, R., & Lam, T. (1974). Trip time characteristics of journeys to and from work. Transportation and traffic theory, 6, 57-86.

Parametric CV function

- Then: $$\mathrm{CV}(d) = \frac{\sqrt{b_0(b_2+1) + b_1(b_2+1)m(d) + b_2 m(d)^2}}{m(d)}$$
- Interpretation of parameters:
  - b_0: "fixed variability"—data recording errors, time spent finding an address, spatial aggregation, etc.
  - b_1: short-term variability in speed during a trip
  - b_2: long-term call-to-call variability, due to factors not included in the model
  - CV approaches $\sqrt{b_2}$ as distance goes to infinity
- CV has same breakpoint as median

Parametric Functions



Outline

- Performance Evaluation Models
- Using the Erlang B Performance Evaluation Model for Yellow and Red Alerts

Performance Evaluation

Repeated from yesterday …

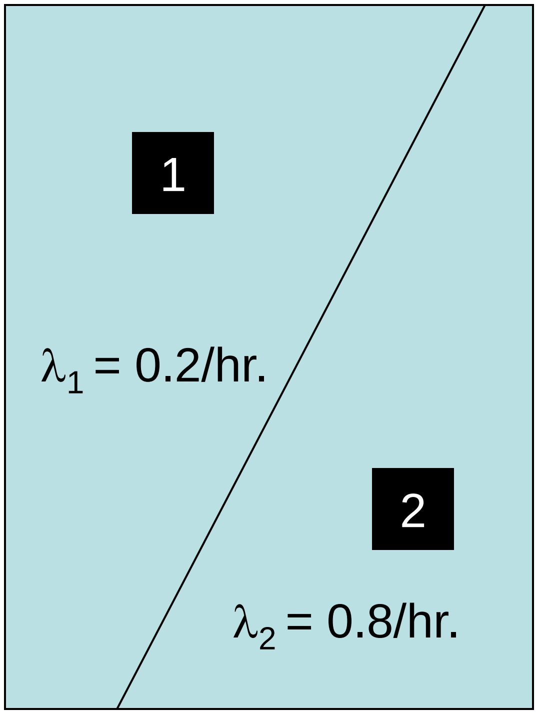
Decomposing Performance

- Performance estimates:
  - p_{ij} = estimated performance for calls from *j* if station *i* responds
  - "performance:" could be coverage probability / survival probability / average response time / …
- Dispatch probabilities:
  - f_{ij} = Pr{station *i* responds | call from *j*}
  - This is where queueing / service syste

← Now we focus on methods to calculate these

- Call arrival rates:
  - Neighborhood *j*: λ_j, system: λ
- System performance: $\sum_j \frac{\lambda_j}{\lambda} \sum_i f_{ij} p_{ij}$

(“Simplest Interesting”?) Example



2 stations, each with 1 unit

2 neighborhoods

$1/\mu$ = avg. service time = 1 hour

λ = call arrival rate = 1 / hour

Performance estimates:

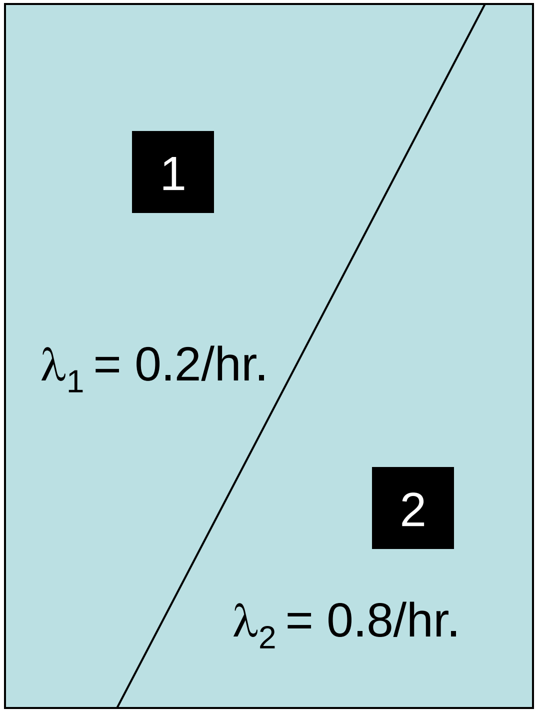
p_{11} = Pr{response time ≤ standard | call from 1, 1 responds}

= 0.95

$p_{12} = p_{21} = 0.5$

$p_{22} = 0.95$

Model 1: "Always Available"



Model	f_{11}	f_{21}	f_{12}	f_{22}	B	Performance
Always available	1.00	0.00	0.00	1.00	0.00	0.95

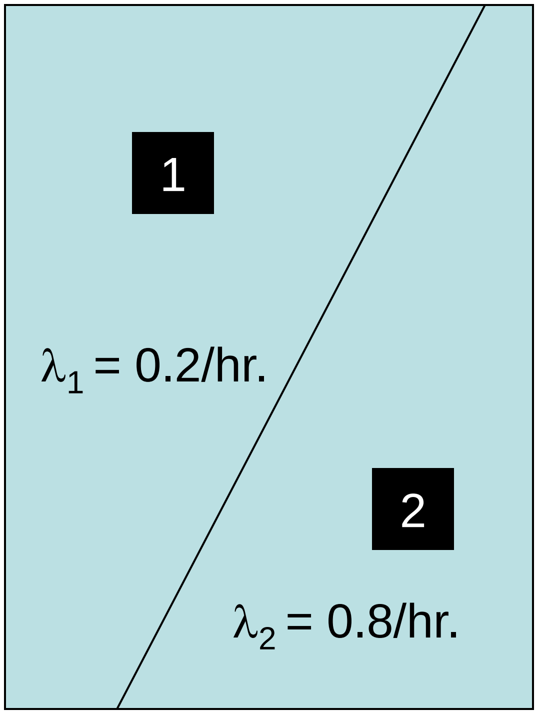
p_{11}	p_{21}	p_{12}	p_{22}
0.95	0.50	0.50	0.95

Assumes all stations have an available ambulance at all times

Provides upper bound on performance

Used in some station location optimization models

Model 2: Binomial



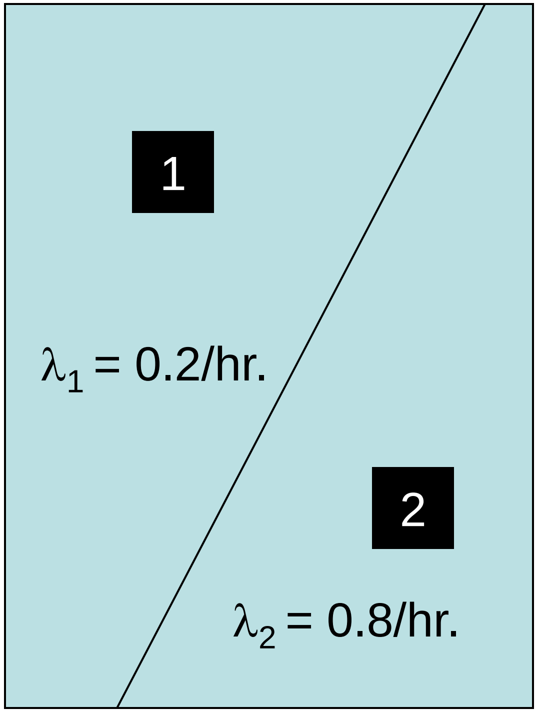
Model	f_{11}	f_{21}	f_{12}	f_{22}	B	Performance
Always available	1.00	0.00	0.00	1.00	0.00	0.95
Binomial	0.60	0.24	0.24	0.60	0.16	0.69

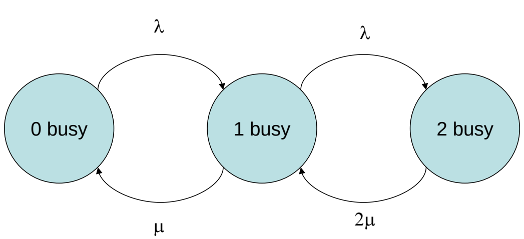
	p_{11}	p_{21}	p_{12}	p_{22}
	0.95	0.50	0.50	0.95

Input:
p = average busy fraction = 0.4 = probability that an ambulance is busy, independent of status of all other ambulances

Used in some ambulance allocation optimization models

Model 3: Erlang B





Model	f_{11}	f_{21}	f_{12}	f_{22}	B	Performance
Always available	1.00	0.00	0.00	1.00	0.00	0.95
Binomial	0.60	0.24	0.24	0.60	0.16	0.69
Erlang B	0.60	0.20	0.20	0.60	0.20	0.67

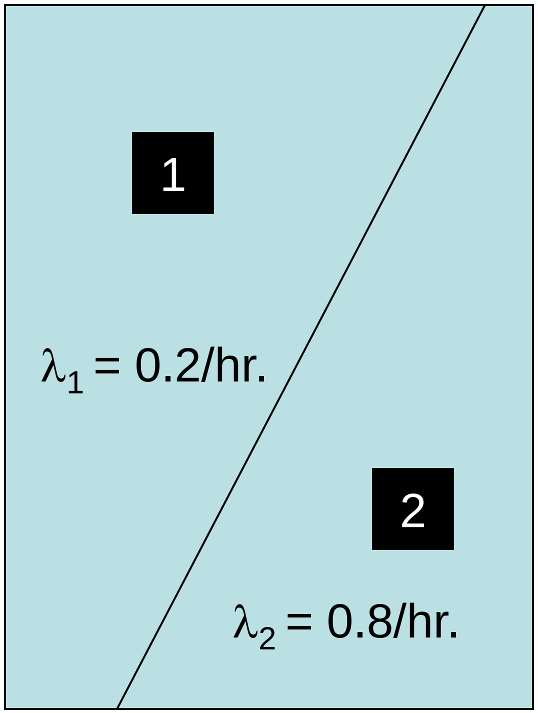
p_{11}	p_{21}	p_{12}	p_{22}
0.95	0.50	0.50	0.95

λ Was chosen so that ambulance utilization = p = 0.4

Probability that closest ambulance responds is the same as in binomial model

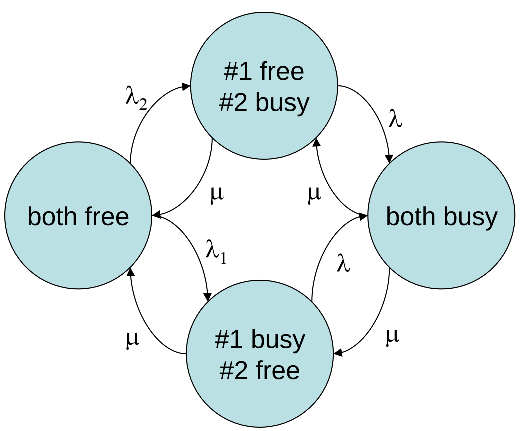
Probability that 2nd-closest ambulance responds is lower, because Pr{2nd-closest ambulance is busy | closest ambulance is busy} > p

Model 4: Hypercube Queueing Model



Model	f_{11}	f_{21}	f_{12}	f_{22}	B	Performance
Always available	1.00	0.00	0.00	1.00	0.00	0.95
Binomial	0.60	0.24	0.24	0.60	0.16	0.69
Erlang B	0.60	0.20	0.20	0.60	0.20	0.67
HQM	0.66	0.14	0.26	0.54	0.20	0.65

	p_{11}	p_{21}	p_{12}	p_{22}
	0.95	0.50	0.50	0.95

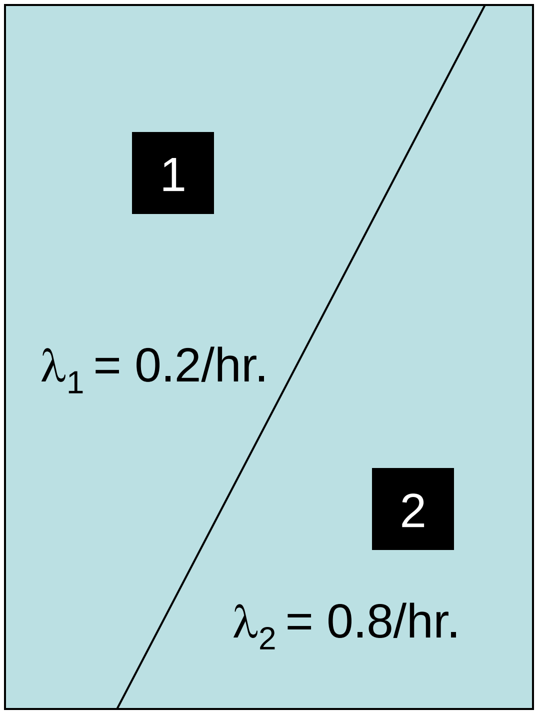


In this model, the two ambulances are distinguishable

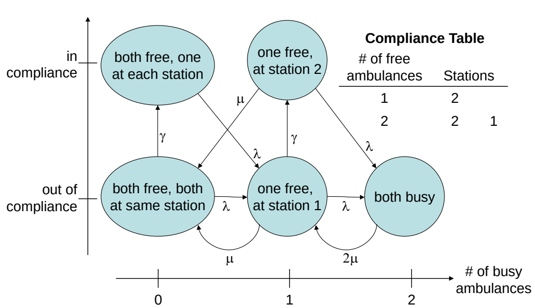
➔ Ambulance 2 is busier

➔ Neighborhood 2 has a lower probability of closest station responding

Model 5: Repositioning



Model	f_{11}	f_{21}	f_{12}	f_{22}	B	Performance
Always available	1.00	0.00	0.00	1.00	0.00	0.95
Binomial	0.60	0.24	0.24	0.60	0.16	0.69
Erlang B	0.60	0.20	0.20	0.60	0.20	0.67
HQM	0.66	0.14	0.26	0.54	0.20	0.65
Repositioning	0.45	0.35	0.09	0.72	0.20	0.70
	p_{11}	p_{21}	p_{12}	p_{22}		
	0.95	0.50	0.50	0.95		



Models 1 – 4 assume an ambulance always returns to its home station

Model 5: If only one ambulance is available and it is at Station 1, then move it to Station 2 (avg. move time = 6 min.)

Neighborhood 1 is better off, Neigbhorhood 2 is worse off

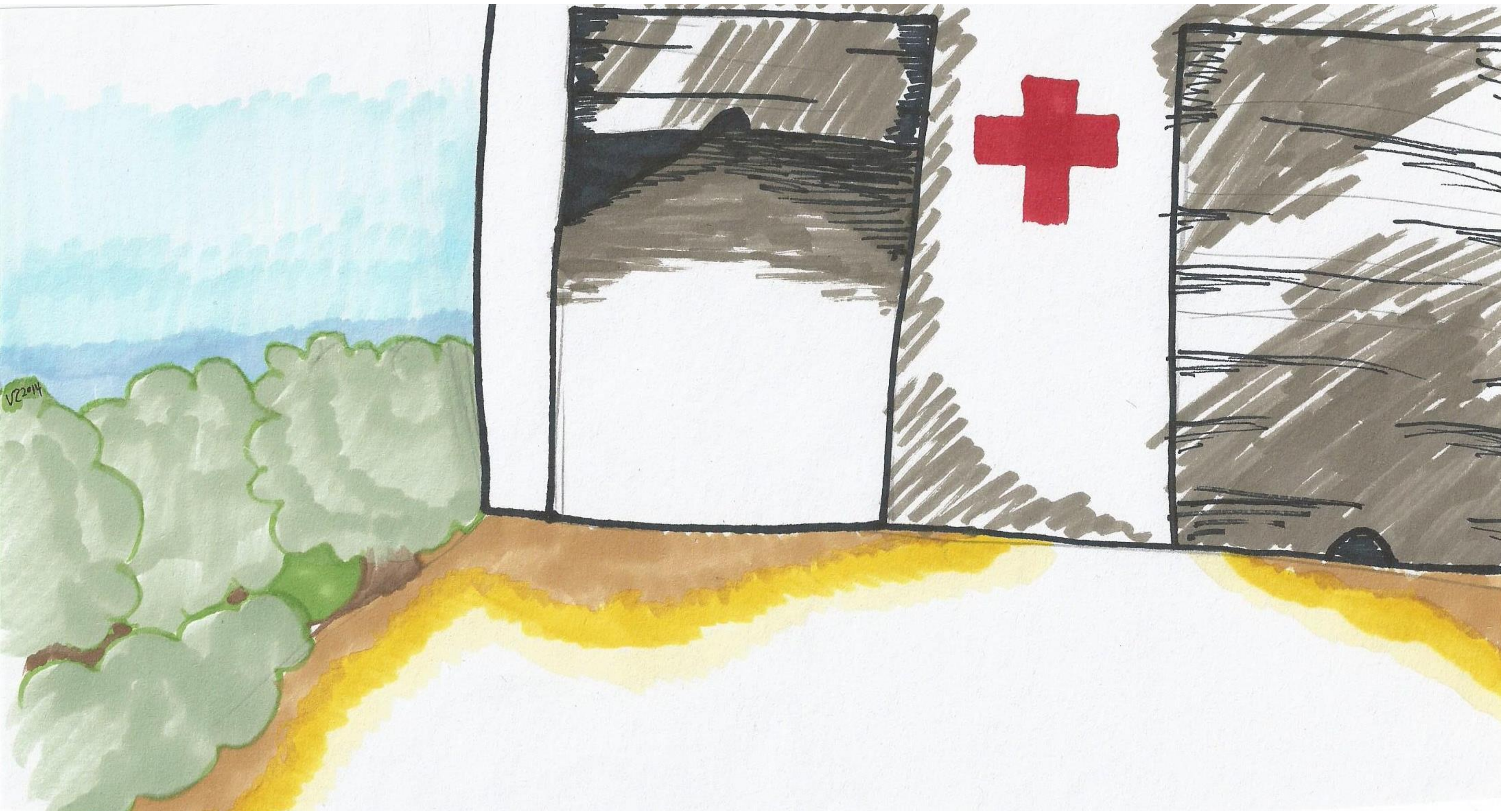
Comparison of Models

Model	Performance	Increased realism	Repositioning	Incorporated in math programs	Scaling issues
Always available	0.95			✓	
Binomial	0.69			✓	
Erlang B	0.67				
HQM	0.65			?	✓
Repositioning	0.70		✓	?	

Managing red and yellow alerts and the consequences of calling in additional units or expediting hospital turnaround

Amir Rastpour, Bora Kolfal, Armann Ingolfsson
School of Business, University of Alberta

Managing red and yellow alerts and the consequences of calling in additional units or expediting hospital turnaround



Ambulance shortage periods

‘Busy’ ambulance system causes concern for paramedics

During the first nine months of 2010, the city of Edmonton had no ambulances to cover medical emergencies for almost 10 hours in total.

- Edmonton Journal, Jan. 20, 2012

Too few paramedics to answer call: Union official

- Toronto Sun, May 13, 2012

Opposition demands EMS wait time review

- Calgary Sun, Feb. 24, 2012

Alert periods

Periods during which:

- Most ambulances are busy

  Yellow Alert

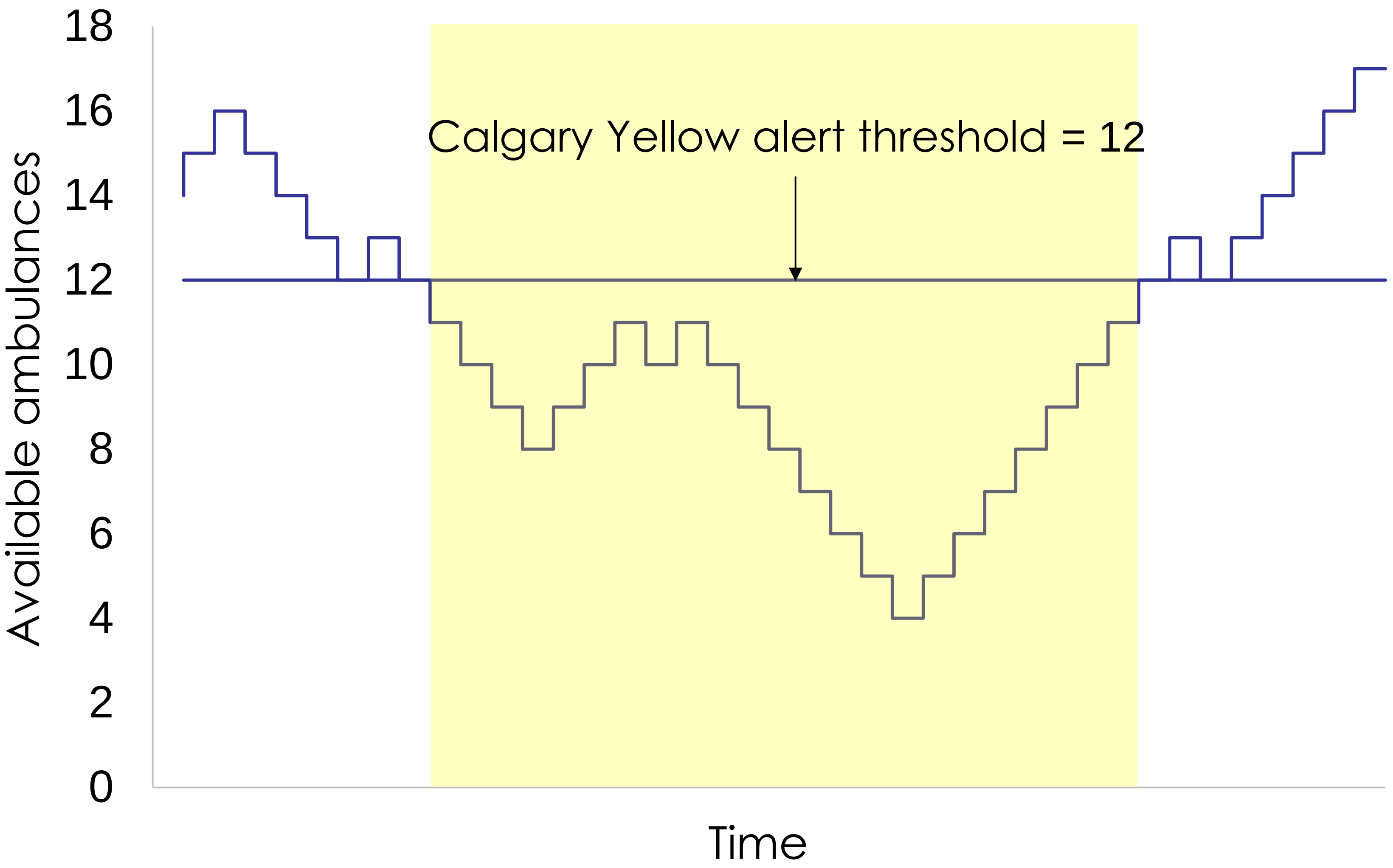
  Available ambulances below a threshold.

  Calgary EMS threshold = 12 ambulances

- All ambulances are busy

  Red Alert

Yellow alert example



Descriptive Statistics

Table 1 EMS configuration in Edmonton in 2008 and in Calgary in 2009.

Parameter	Edmonton	Calgary
Yellow Alert threshold (θ)	8	12
Minimum number of scheduled ambulances	19	28
Maximum number of scheduled ambulances	36	54
Average number of scheduled ambulances	25	41

Table 2 Descriptive statistics for the duration of alert periods in Edmonton in 2008 and in Calgary in 2009.

	Yellow Alert		Red Alert	
Statistic	Edmonton	Calgary	Edmonton	Calgary
Sample Size	1349	703	587	9
Mean (min.)	106.41	7.09	7.20	1.37
Standard Deviation (min.)	120.26	11.53	11.32	1.32
Maximum (min.)	1012.02	127.28	138.93	4.53
Squared Coefficient of Variation	1.28	2.64	2.47	0.94

Decision faced by dispatchers

- Ride out the alert… or act?

- Possible actions:
  - Reposition ambulances*
  - Call in additional ambulances
  - Free up busy ambulances in EDs

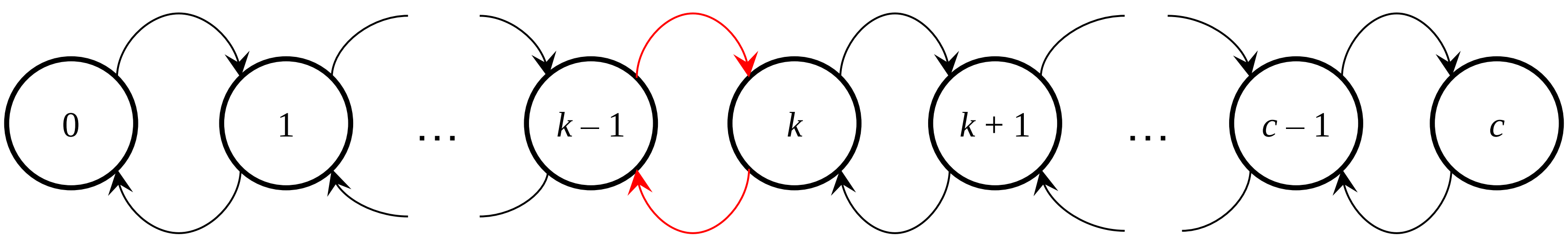
  - ?

*Alanis et al. 2012, Maxwell et al. 2010, Schmid 2011

Mathematical Model: "Erlang B loss model"

Analogy:
phone lines = ambulances
busy signal = red alert

k-partial busy period:
k or more of *c* servers are busy



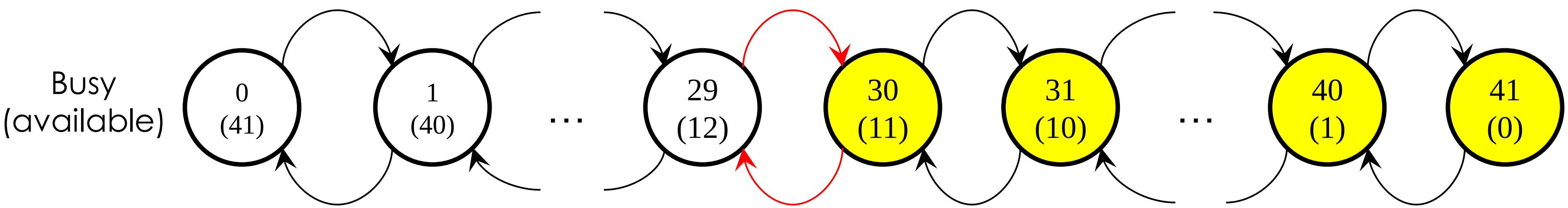
Yellow Alert = (c − threshold +1)-partial busy period

Red Alert = c-partial busy period

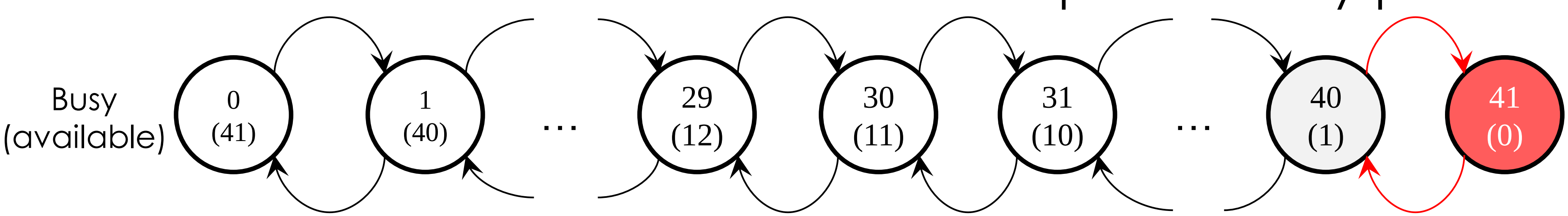
Relationship between alert periods and partial busy periods

Calgary: 41 servers

Yellow Alert = 30-partial busy period



Red Alert =
41-partial busy period



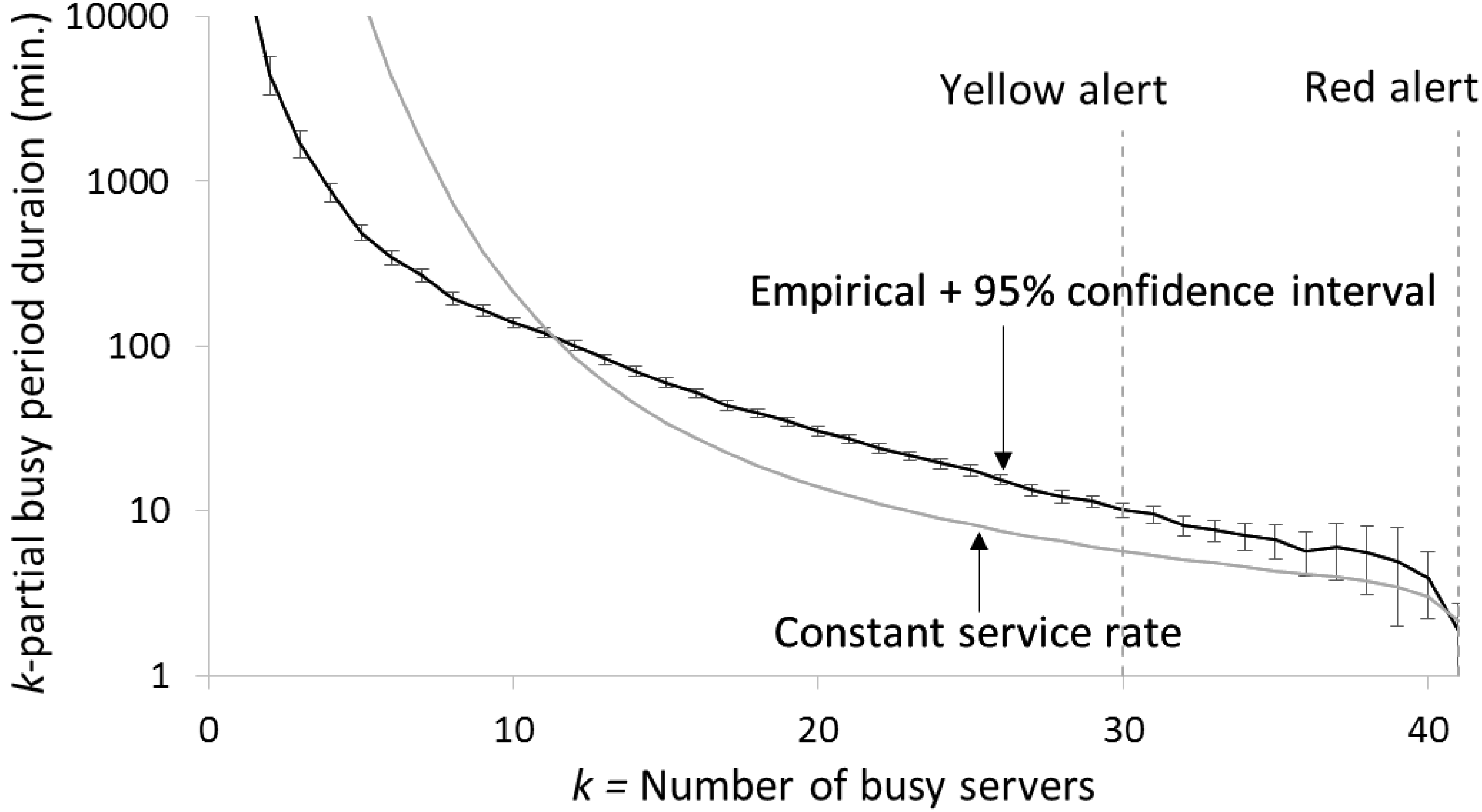
Main Result

Equations to calculate average busy period durations:

$$\mathrm{E}(B_c) = \frac{1}{c\,\mu}, \mathrm{E}(B_k) = \frac{\lambda \mathrm{E}(B_{k+1})}{k\,\mu} + \frac{1}{k\,\mu}, k = c - 1, \ldots, 1.$$

Also have equations for variance and other quantities

Validation—the whole year

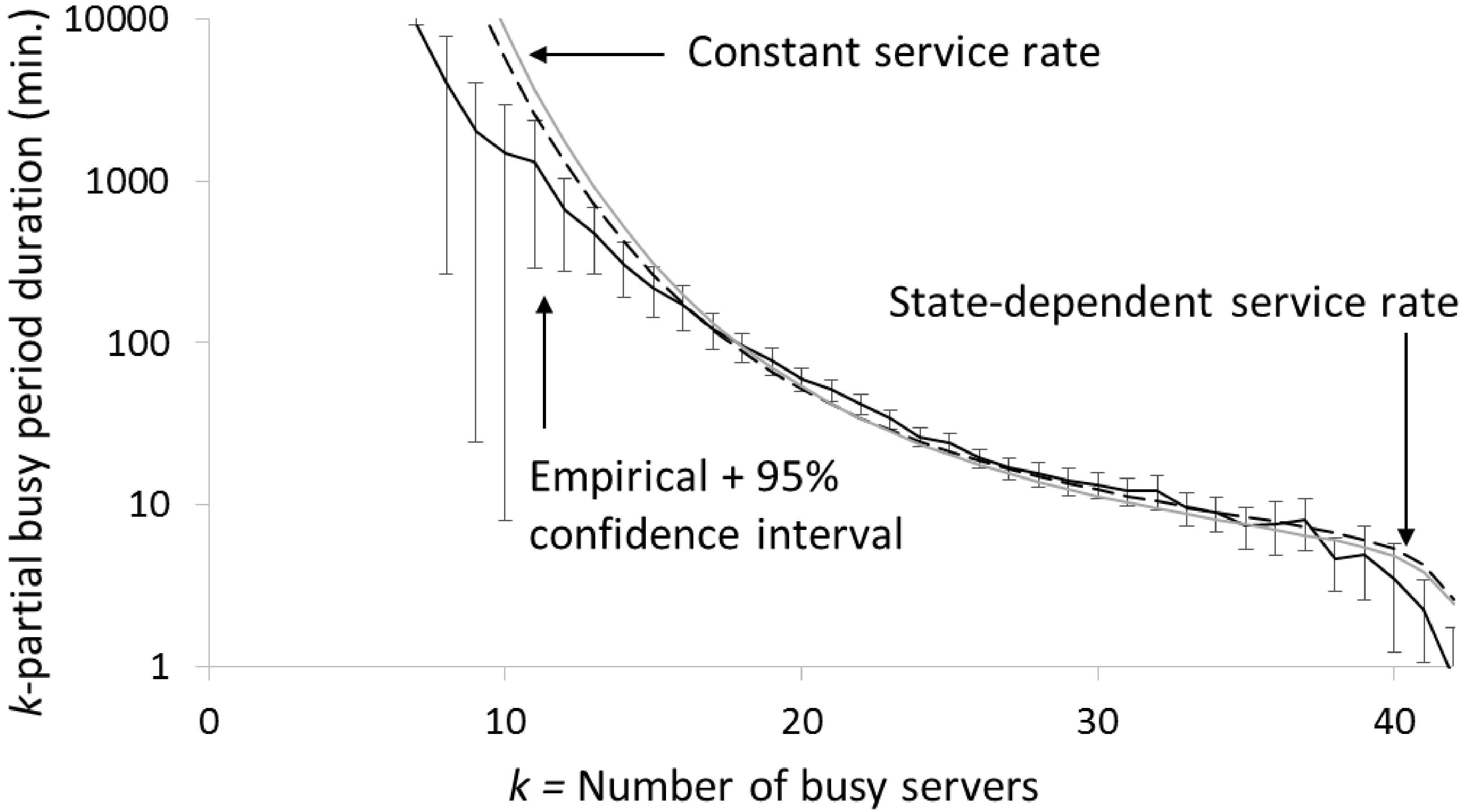


Reasons for poor fit

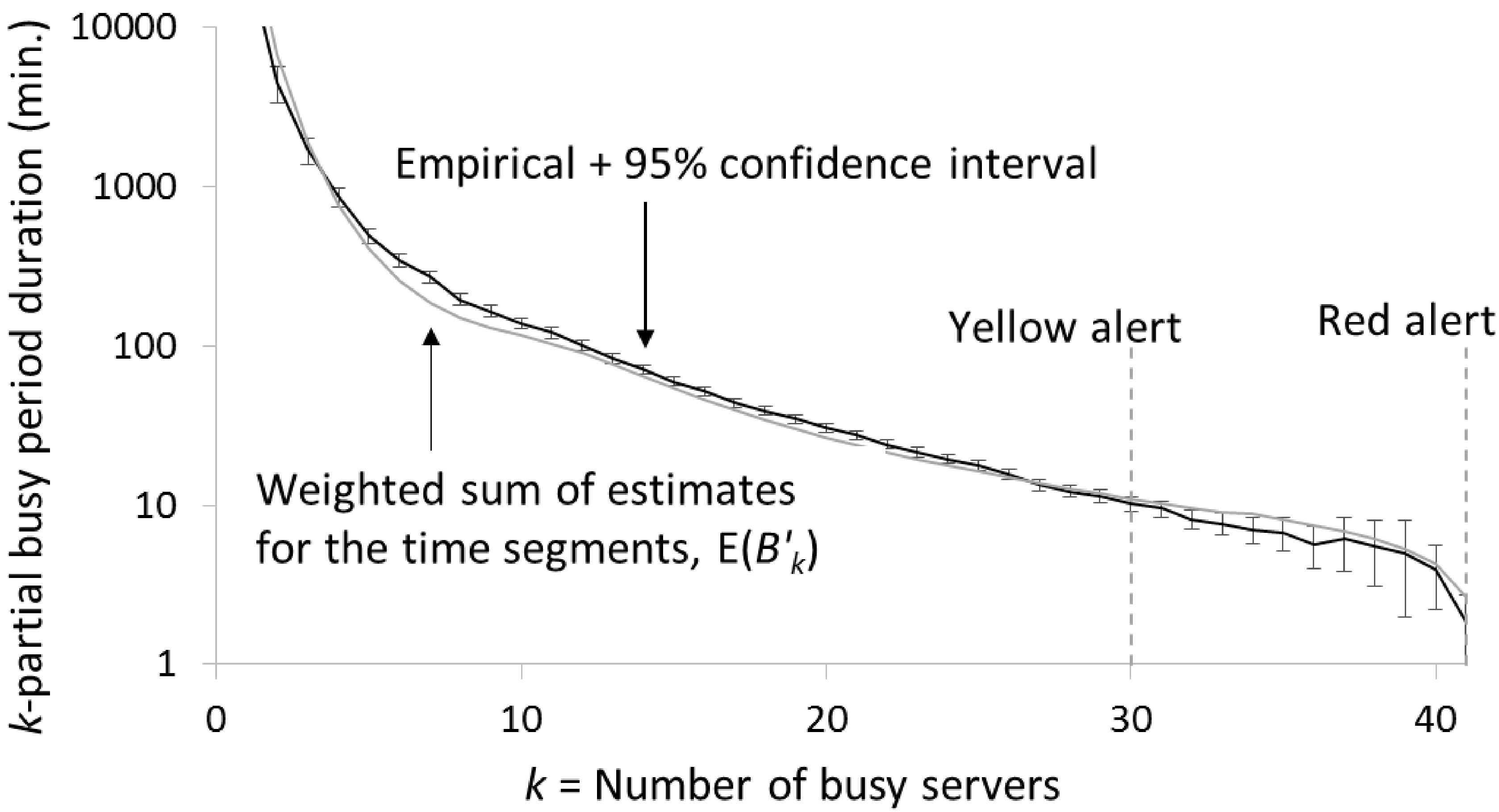
- Number of units varies with time
- Call rates vary with time
- “Service speed” varies with number of busy units

- Check how much fit improves after controlling for these factors

Validation for weekday 9 am – 1 pm



Aggregation over 16 time segments

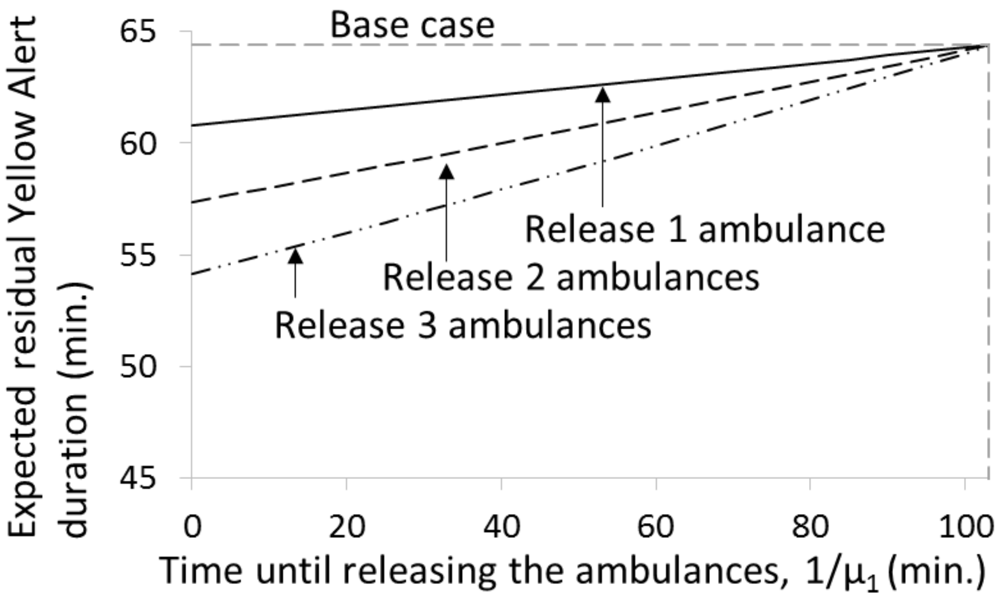


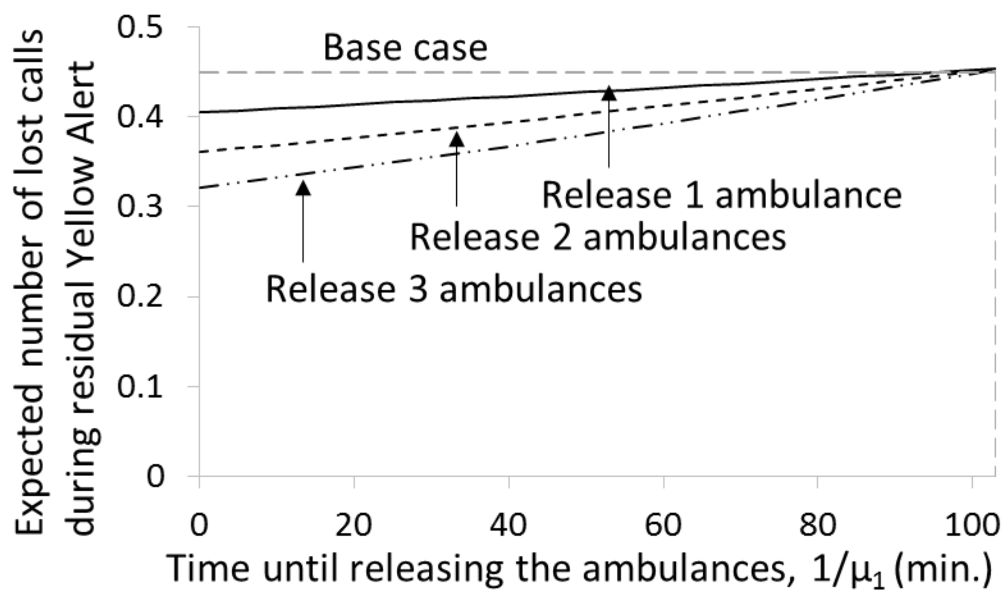
Actions and Performance Measures

- Actions
  - Call in additional units
  - Free up units in EDs
    Modeled as "increase service rate"

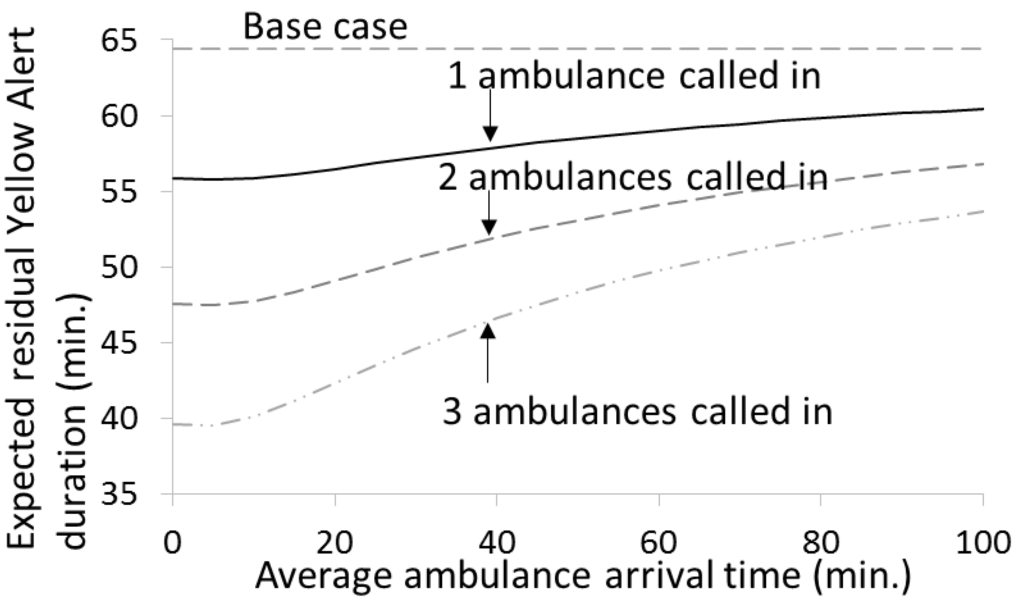
- Performance measures
  - Average remaining Yellow Alert duration
  - Average number of "missed" calls
    (because of red alert)

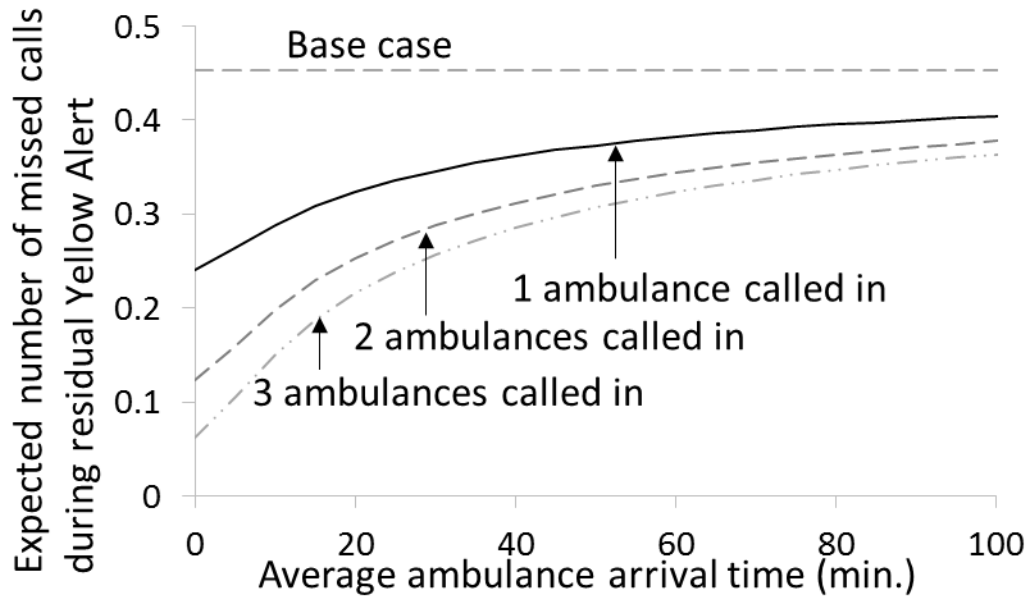
Expediting Hospital Turnaround





Calling in Additional Units





“Optimal” Combination of Actions

Released ambulances				
3	55			
2	58	50		
1	61	53	41	
0	64	56	48	40
	0	1	2	3
	Called in ambulances			

(a) Expected residual Yellow Alert duration.

Released ambulances				
3	0.33			
2	0.37	0.23		
1	0.41	0.26	0.18	
0	0.45	0.29	0.20	0.15
	0	1	2	3
	Called in ambulances			

(b) Expected number of lost calls.

The Optimal Combination can Depend on the Performance Measure

Released ambulances				
3	54			
2	57	53		
1	61	56	51	
0	64	59	54	50
	0	1	2	3
	Called in ambulances			

(a) Expected residual Yellow Alert duration.

Released ambulances				
3	0.32			
2	0.36	0.31		
1	0.40	0.34	0.31	
0	0.45	0.38	0.34	0.32
	0	1	2	3
	Called in ambulances			

(b) Expected number of lost calls.