Statistiska centralbyrån
Statistics Sweden

UNIVERSITY OF Southampton

The Leverhulme Trust

Simulation of adaptive survey design in a longitudinal survey

Peter Lundquist and Anton Johansson (Statistics Sweden)
Gabriele B. Durrant (University of Southampton)

5th Workshop on Advances in Adaptive and Responsive Survey Design,
University of Michigan, November 6th-7th, 2017

Introduction and background

- Large survey resources are being spent on making unproductive calls e.g. contact attempts that do not lead to an interview.
- Unstable data collection and unclear collection strategy
- We have used work by Durrant et al. (2013, 2015) who assess the prediction of nonresponse models using paradata from previous and current wave
- The ambition is to find a new data collection strategy for the Swedish Labour Force Survey
- A simulation tool to use in the fieldwork planning

Data used in Simulation

The Swedish Labour Force Survey (LFS)

- Longitudinal survey with 8 waves
- “A new sample” every week
- One data collection mode: telephone
- Two interviewer groups: field and call-center

- Data used in modeling: LFS in January 2016
  - Initial sample size 5,164; Week 4 sample
- Model used on next wave in April 2016:
  - 7/8 of the sample is the same

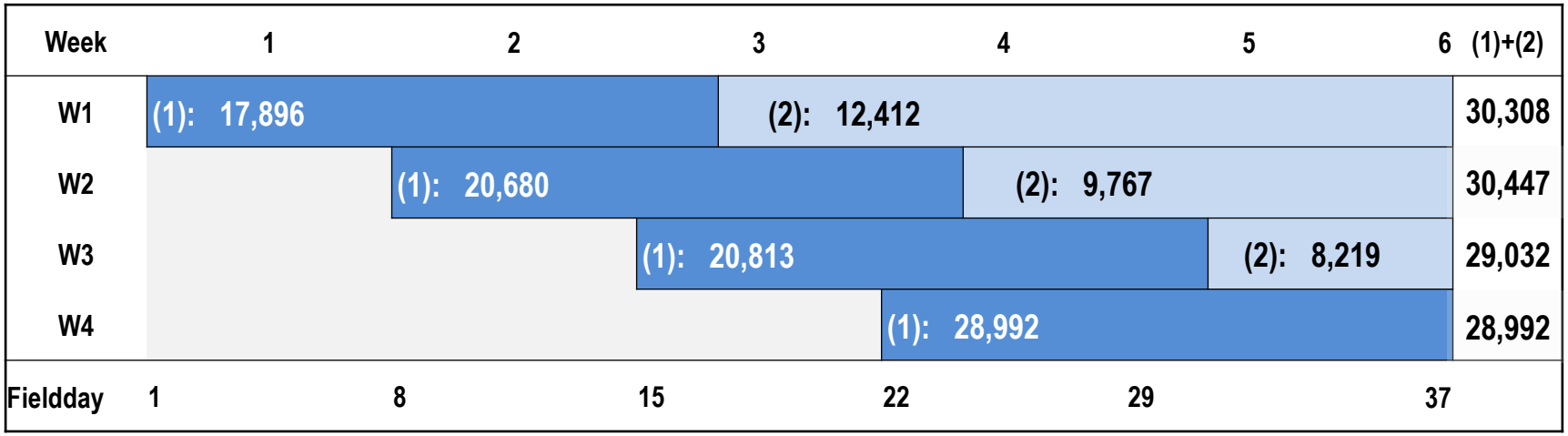
Data collection in LFS

Approximately 100 Field + 100 call center interviewers sign up for the following shifts:

Shift/Day	Mon-Thu	Friday	Saturday	Sunday
08:00-11:59				
12:00-16:59				
17:00-18:59				
19:00-21:59		Contact by agreement		

- We assume that the resources could be better allocated to the different subsamples (e.g. Week 1-Week 4 in January 2016)

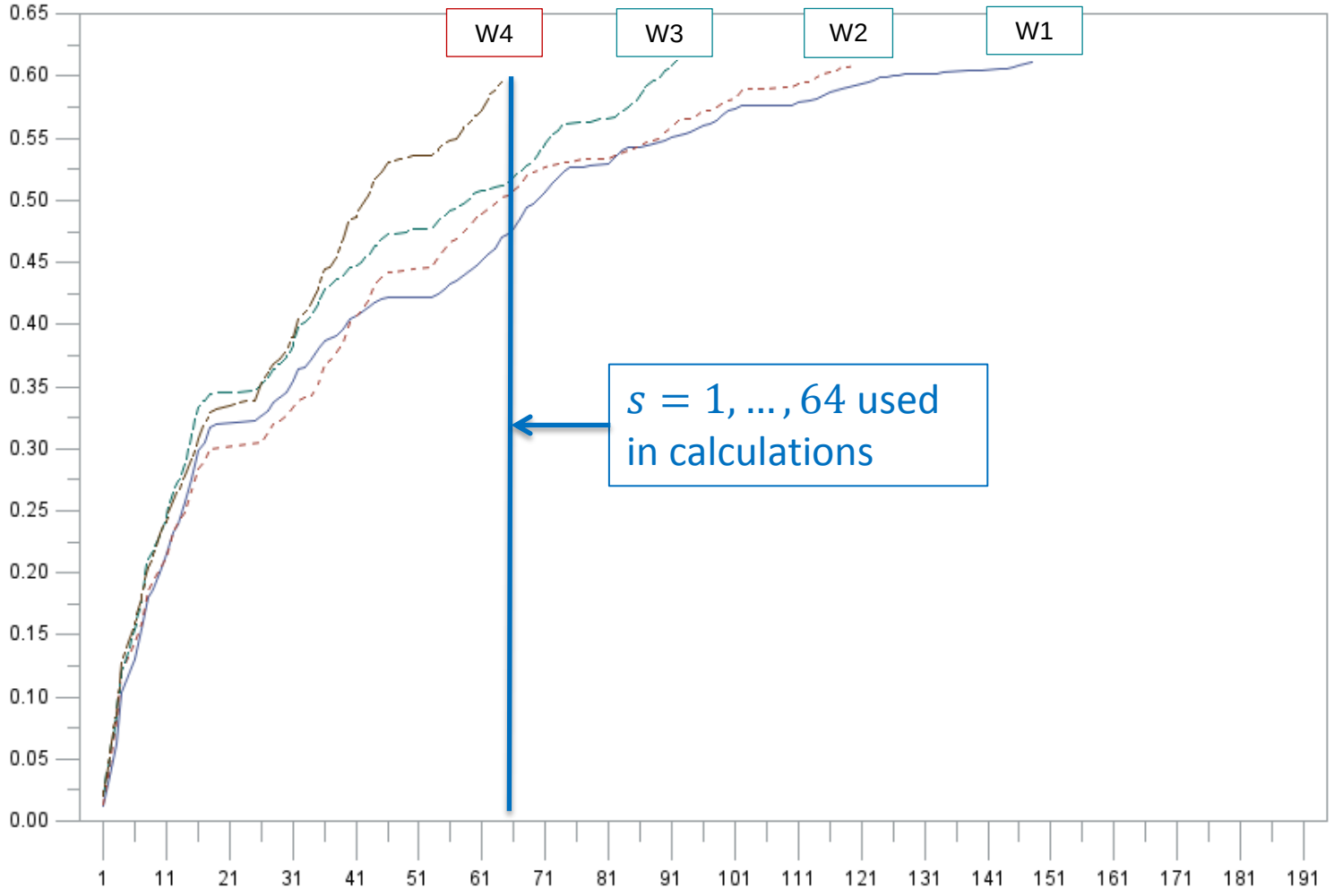
LFS – fieldwork January 2016
Number of call attempts



(1) Primary fieldwork 16 days (equal all samples W1-W4)

(2) Extended fieldwork

Response progression over time-slots for 4 LFS-weeks



LFS: Jan 2016

How to decide the three phases?

Week 4 has a short data collection: 16 days

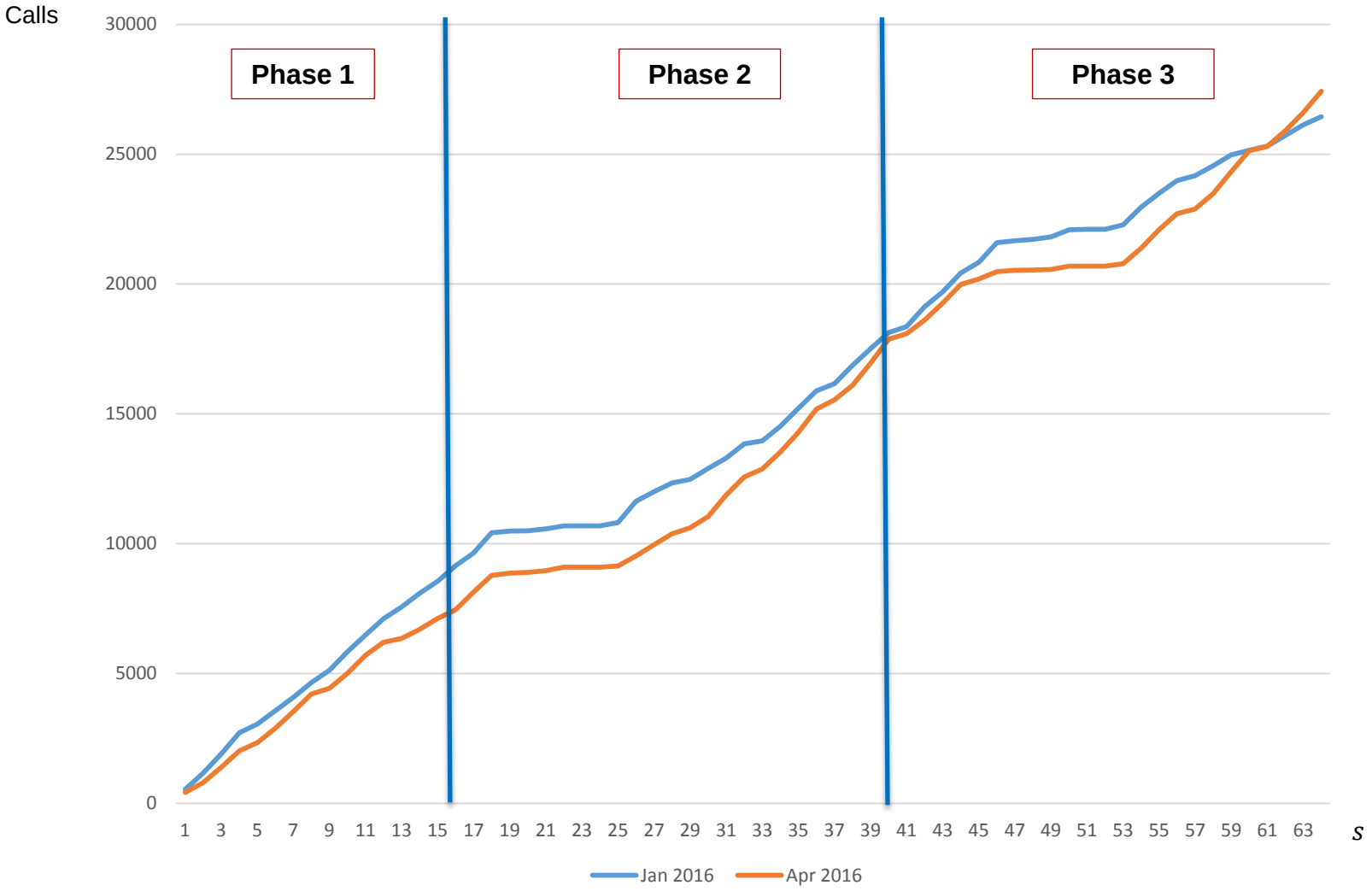
Phase 1: All units should have been contacted two times within the first four days.

Phase 2: Active units should have seven calls (day 5 to 10).

Phase 3: Active units should have 13 calls (day 11 to 16).

- Every day is divided in four shifts $(s = 1, 2, \ldots, 64)$

Comparison January and April 2016
Cumulative number of calls for week 4



Data collection strategies

Two different data collection strategies are studied in the simulation:

1. Maximize the response rate (comparison strategy)
2. Use assigned time slots in the first phase (desired strategy)

The data collection use the same three phases, i.e.
Phase 1 (day 1-4), Phase 2 (day 5-10) and
Phase 3 (day 11-16).

Assigned time slots

The LFS-interview (in January) ends with questions about a suitable time for the next interview (in April).

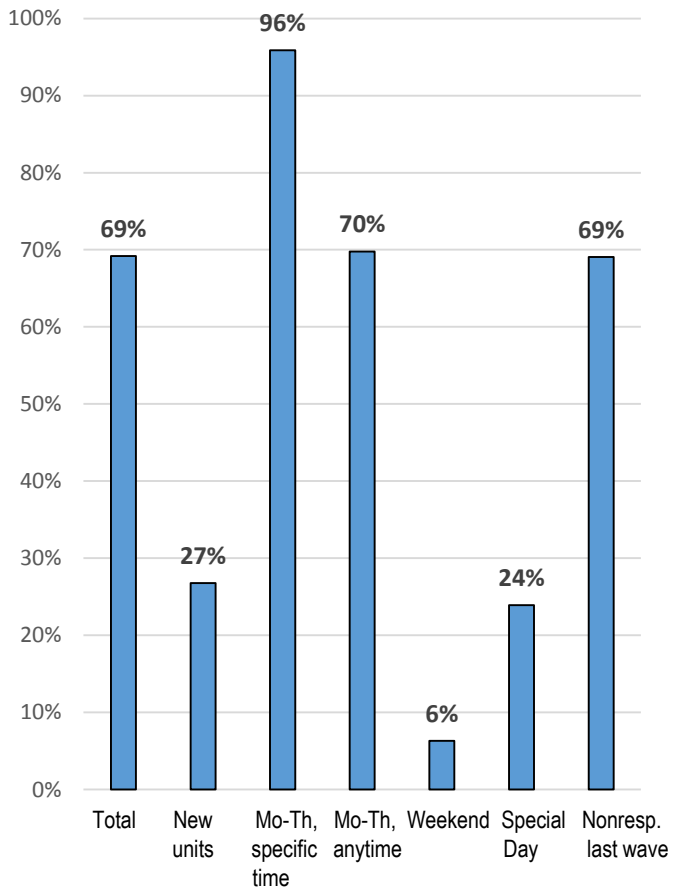
- For the respondents, it's then possible to place the first calls in appropriate time intervals (and days) in April.

The time slots are: 08:00-11:59, 12:00-16:59, 17:00-18:59 and 19:00-21:59.

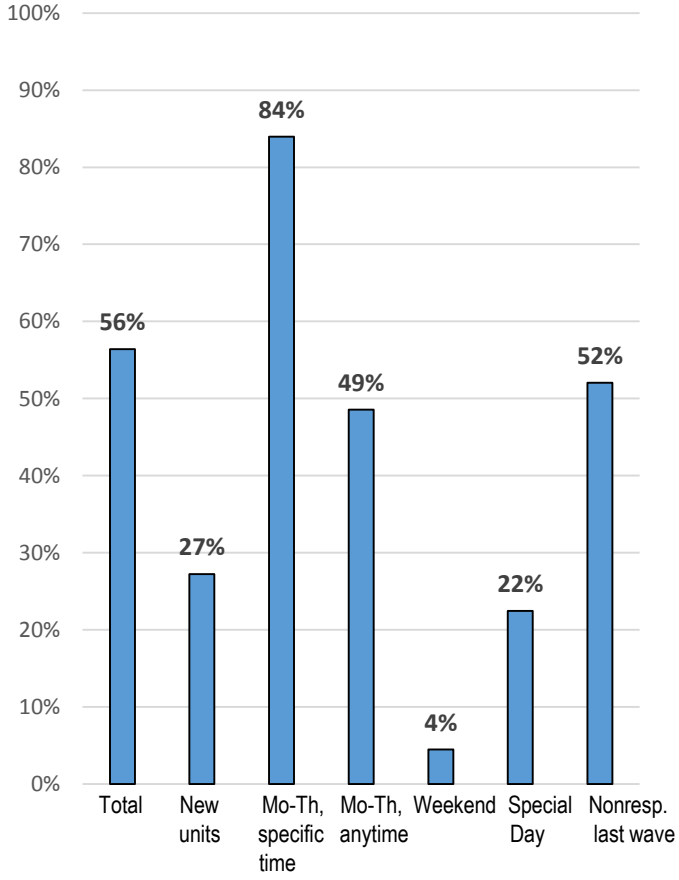
- For the nonrespondets register information, process data and information from earlier waves are used.
- For wave 1 only register information is used.

The proportion in initially selected time slot

1st Call attempt



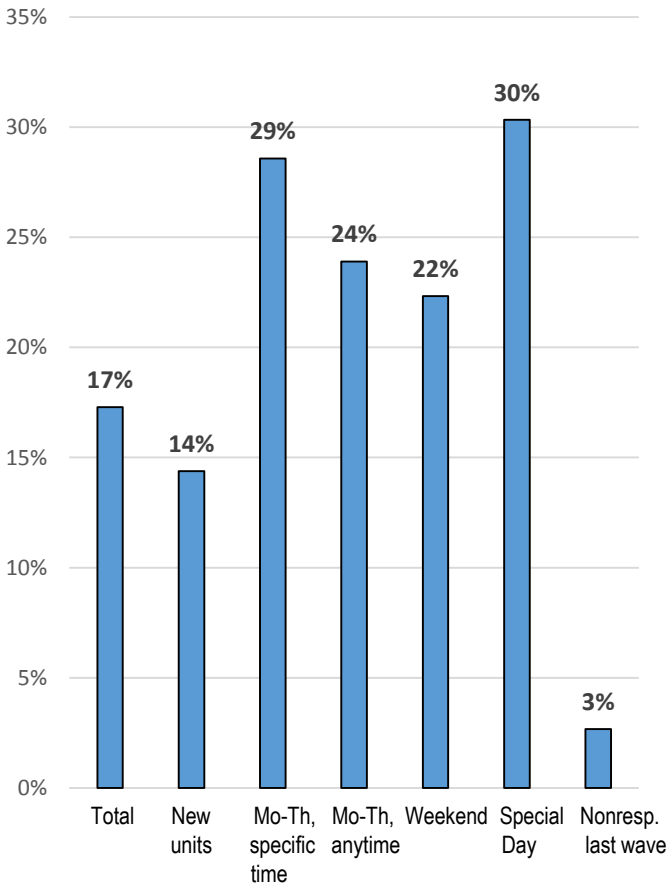
2nd Call attempt



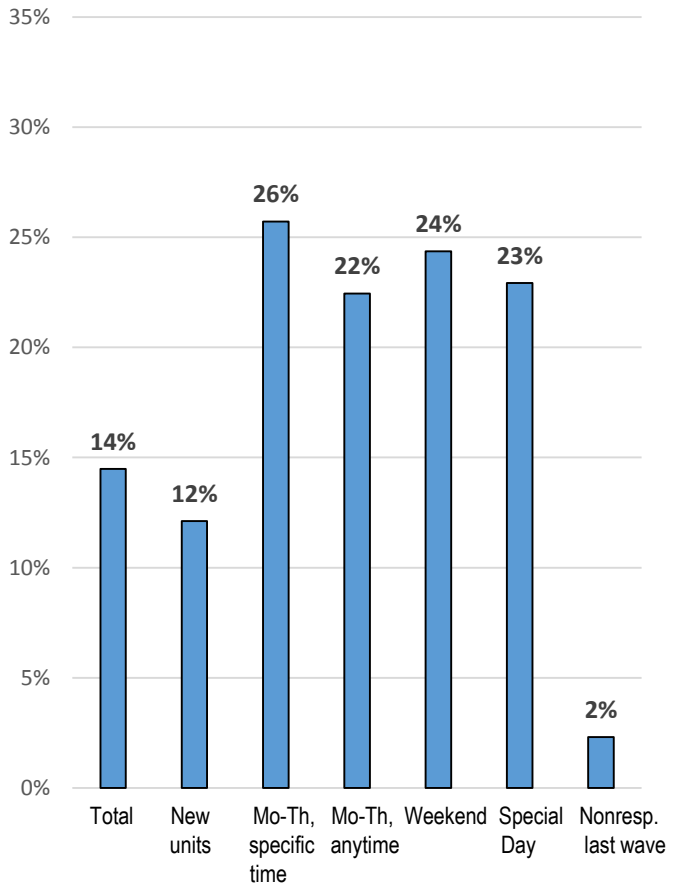
Adapted from Johansson and Olsson internal document 2017

The proportion interviews

1st Call attempt



2nd Call attempt



Adapted from Johansson and Olsson internal document 2017

We now turn to the simulations…

- Logistic regression
- Register data and process data
- Number of calls in time slots
- Modelling the strategies
- Indicators

Models for Phase 1, 2 and 3

Logistic regression (binary), three different models

Model 1
Phase 1 time slots: $s = 1,\ 2, \ldots, 16$ (day 1 to 4)

Model 2

Phase 2 time slots: $s = 17,\ \ldots, 40$ (day 5 to 10)

Model 2

Phase 3 time slots: $s = 41,\ \ldots, 64$ (day 11 to 16)

Dependent variable:

response (interview or **not** interview)

- Objects are individuals (not households)
- Within each s is only one call attempt allowed

Models for Phase 1, 2 and 3

Logistic regression (binary), explanatory variables

	Explanatory variables	Model 1	Model 2	Model 3
Register data	Age (16-54 or 55-74 yrs)	X	X	X
	Born in Sweden or not	X	X	
	Education (high) or not	X	X	X
	Married or not	X	X	X
	Employed or not	X		X
Process data	More than 2 calls	X		
	1st wave	X	X	X
	2nd-8th wave and interview last wave	X	X	X
	2nd-8th wave and **no** interview last wave	X	X	X
	2nd-8th wave and LONG -1 (more than 6 call attempts or not **last wave**)		X	X
	LONG (more than 6 call attempts or not)		X	X
	Day shift (8-12, 12-17, 17-19, 19-22)	X	X	X
	Time slot (1, 2, …, 64)	X	X	X

Models January 2016

Logistic regression, response as dependent variable

	Pseudo R^2	ROC
Phase 1: Model 1	0.13	0.705
Phase 2: Model 2	0.15	0.750
Phase 3: Model 3	0.13	0.757

Pseudo-R2 Statistic : the proportion of variation in the dependent variable that is explained by the model.

ROC (Receiver Operating Curve): the greater area under the curve (AUC) the greater predictive power. AUC values range from 0.5 (no discrimination) to 1 (perfect discrimination).

Predicted response propensities

For unit $k \in sample$ in phase j $(j = 1, 2, 3)$ and $s = 1,2, \ldots 64$

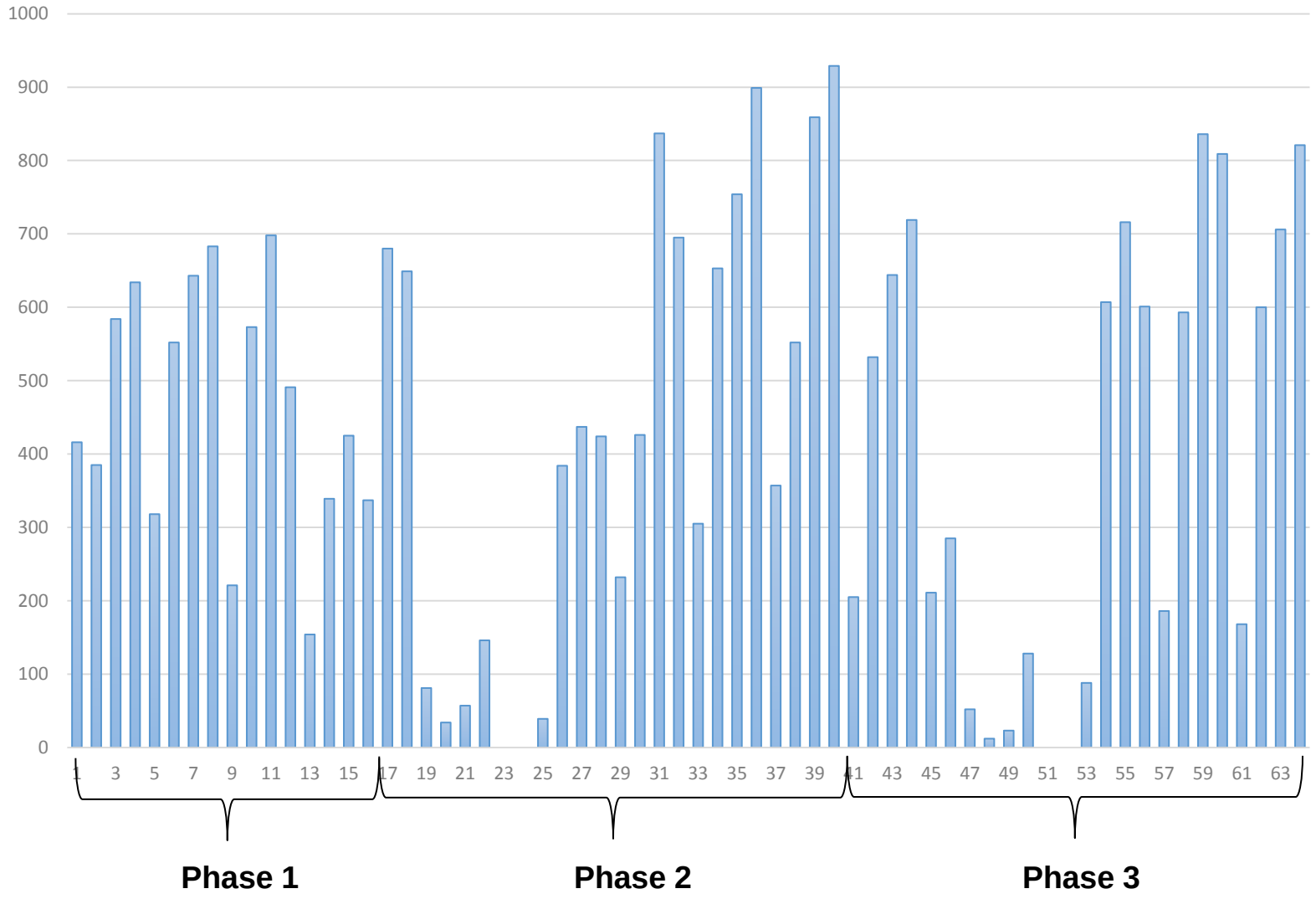
$$\hat{p}_k = \frac{\exp(\widehat{\boldsymbol{\beta}}_j \mathbf{x}_{k,j(s)})}{1 + \exp(\widehat{\boldsymbol{\beta}}_j \mathbf{x}_{k,j(s)})}$$

are used in the simulations.

How to decide the workload?

- In each time slot is a fixed number of calls possible.
- We will in the modelling use the **workload** for week 4 in January, but in the simulation then **workload** for week 4 in April
- It is however clear that this will not always work.
- In Choudhry et al. (2011) the **workload** was optimized to minimize the data collection costs

Strategy 1 and 2: Number of calls April 2016



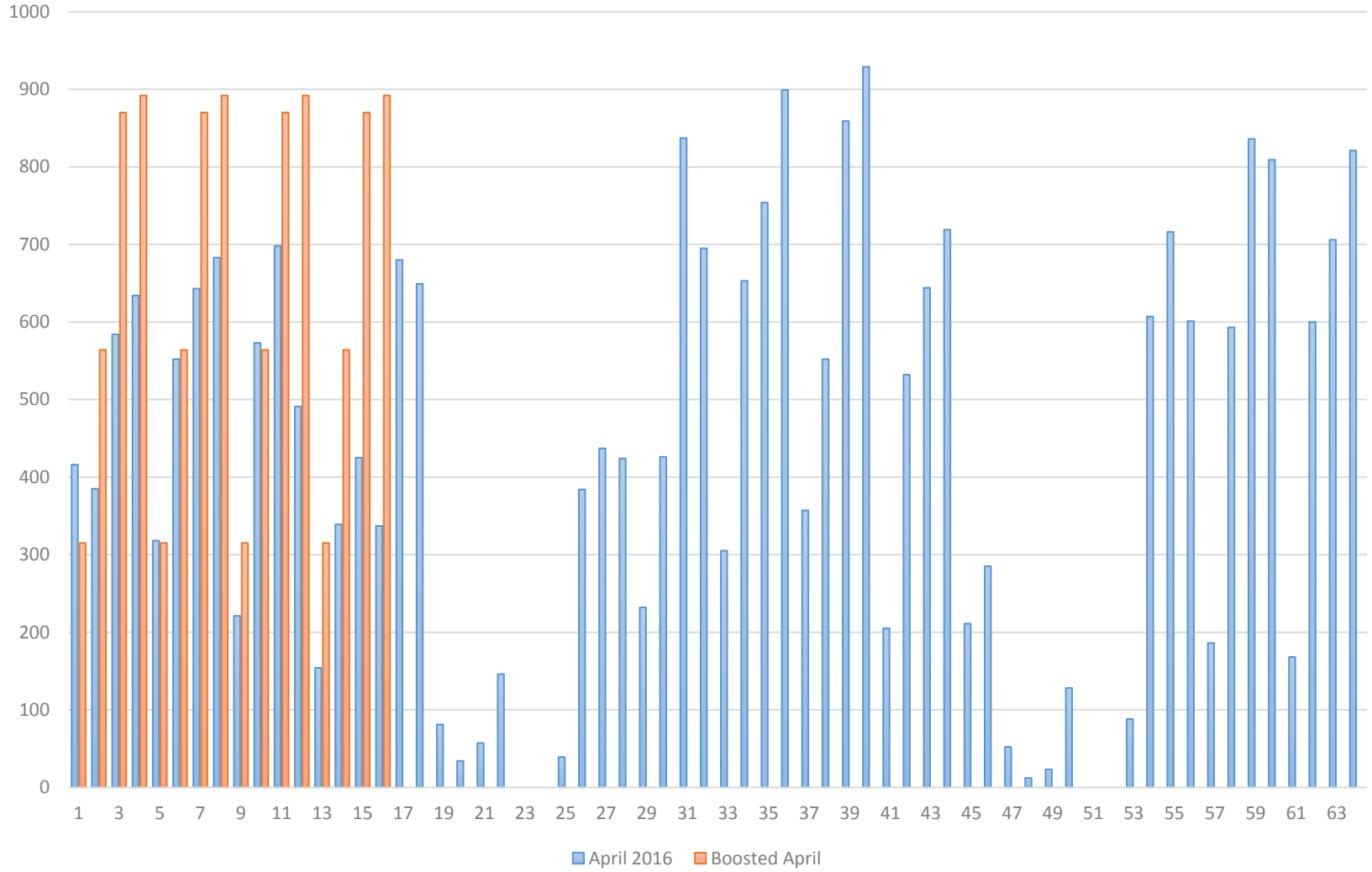
Data collection: Strategy 1

Phase	Maximize the response rate (comparison strategy)
1) Day 1-4 s=1-16 Model 1	• For each s the $\hat{p}_k$ objects in treatment are sorted (descending) • A maximum of 3 calls
2) Day 5-10 s=17-40 Model 2	• For each s the $\hat{p}_k$ objects in treatment are sorted (descending) • A maximum of 7 calls • After 4 calls: stop individuals in wave 2-8 that refused to participate last wave
3) Day 11-16 s =41-64 Model 3	• For each s the $\hat{p}_k$ objects in treatment are sorted (descending) • A maximum of 13 calls

Data collection: Strategy 2

Phase	Use time slots (desired strategy)
1) Day 1-4 s=1-16 Model 1	• For each s the $\hat{p}_k$ objects in treatment are sorted (randomly) • Use time slots • A maximum of 3 calls
2) Day 5-10 s=17-40 Model 2	• For each s the $\hat{p}_k$ objects in treatment are sorted (descending) • A maximum of 7 calls
3) Day 11-16 s =41-64 Model 3	• For each s the $\hat{p}_k$ objects in treatment are sorted (descending) • A maximum of 13 calls • Select easy objects still in treatment and wave 1-2

Strategy 2B:
Boosted number of calls in phase 1



Evaluation of the new strategy

$P =$ the weighted response rate in per cent

IMB = the imbalance measure measures the difference between the response set r and the selected sample s for a chosen **x**-vector.

It could be demonstrated* that IMB is equal to the variance for the response propensities for the chosen **x**-vector

$$CV_s = \frac{\sqrt{IMB}}{P}$$

***Särndal & Lundquist 2014**

Note: auxiliary variables (register data) depends on available variables and the indicators depends on the sample.

Simulation results, averages and standard errors based on 1,000 simulations

				Time (min)	Number of calls	
LFS-April 2016	***P***	***IMB***	***CV_s***	***Per case***	***Total***	***Per case***
Strategy 1	64.8 (0.5)	1.67 (0.11)	19.9 (0.7)	17.8	27,424	5.19
Strategy 2	61.2 (0.5)	1.43 (0.10)	19.5 (0.7)	16.5	25,166	4.76
Strategy 2B*	62.6 (0.5)	1.37 (0.10)	18.7 (0.7)	17.2	26,573	5.03

The response rate *P* is weighted in percent, *IMB*, CV_s, are multiplied with 100.

*) Increased number of calls in phase 1

x-vector used in computations: Age, High Education, Origin, Civil, Gender
(3) (2) (2) (2) (2)

Discussion

- **Work with the logistic regression models:**
  - **Further elaboration with factors**
  - **Bayesian models** were not necessary in this situation, however in other kind of simulations this could be of interest (as in a real data collection)
  - **Development of the tool:** Indicators for the different phases and maximum interviewer resources
- **The simulation tool:**
  - Makes it possible to find better strategies
  - Generates new ideas how to develop and control the data collection planning
- **Experiments** -This possibility should be noted!
- **BADEN:** use tool by Schouten et al. 2017.

Thank you!

peter.lundquist@scb.se
anton.johansson@scb.se

Simulation on April 2016

EXAMPLE Strategy 1

- The response propensities are assumed to be $Be(\hat{p}_k)$ distributed in the logistic regression models.
- For time slot 1: $5{,}282$ random selections were made from Model 1. The randomization corresponds to the outcome if all the individuals in the sample were contacted for $s = 1$. The n_1 highest response propensities are inspected and those who "respond" are set aside.
- The "nonresponders" continue to the 2nd time slot, $s = 2$ The n_2 highest response propensities are inspected and those who "respond" are set aside…
- The procedure continues until time slot 64, where the "data collection" ends.

The data collection is replicated 1,000 times for the described strategies.

References

Choudhry, Hidiroglou & Laflamme (2011). ”Optimizing CATI workload to minimize data collection cost.” Proceedings of the Survey Research Methods Section, 1904-1913, ASA.

Durrant, D’Arrigo & Müller (2013). “Modeling Call Record Data: Examples from Cross-Sectional and Longitudinal Surveys,” in Improving Surveys with Paradata: Analytic Use of Process Information, ed. Kreuter, F., 281–308, Hoboken, NJ: John Wiley and Sons.

Durrant, Maslovskaya & Smith (2015). “Modeling final outcome and length of call sequence to improve efficiency in interviewer call scheduling.” Journal of Survey Statistics and Methodology, 3, 397–424.

Johansson, Lundquist & Durrant (2016). ”Stopping rules in a longitudinal survey – impact on cost and survey quality.” Presented at AAPOR.

Särndal & Lundquist (2014). “Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation.” Journal of Survey Statistics and Methodology, 2, 361-387.

Schouten, Mushkudiani et al. (2017). “A Bayesian analysis of design parameters in survey data collection.” Paper submitted.