Microsimulating and Optimizing CATI Call Scheduling

Workshop on Microsimulation for Surveys

National Institute of Statistical Sciences (NISS)

Choudhry, G.H., **Hidiroglou, M.A.**, **Laflamme, F.,**
Bélanger Y., Neusy, E. and Couture, K.

Outline

- Background
  - Paradata sources, context and lessons learned
  - Cost-efficient framework
- Research projects
  - Microsimulation
  - Optimization
  - Relationship between the two projects
- Conclusion

Background

- Historically, paradata research focussed at improving the current data collection process and practices for CATI surveys
  - Identified several strategic opportunities for improvement
  - Implemented some of them (e.g. responsive collection design, time slice, cap on calls, etc.)

➢ *Active management used to monitor data collection suggested that resources were not always optimally used throughout collection period*

Paradata Sources

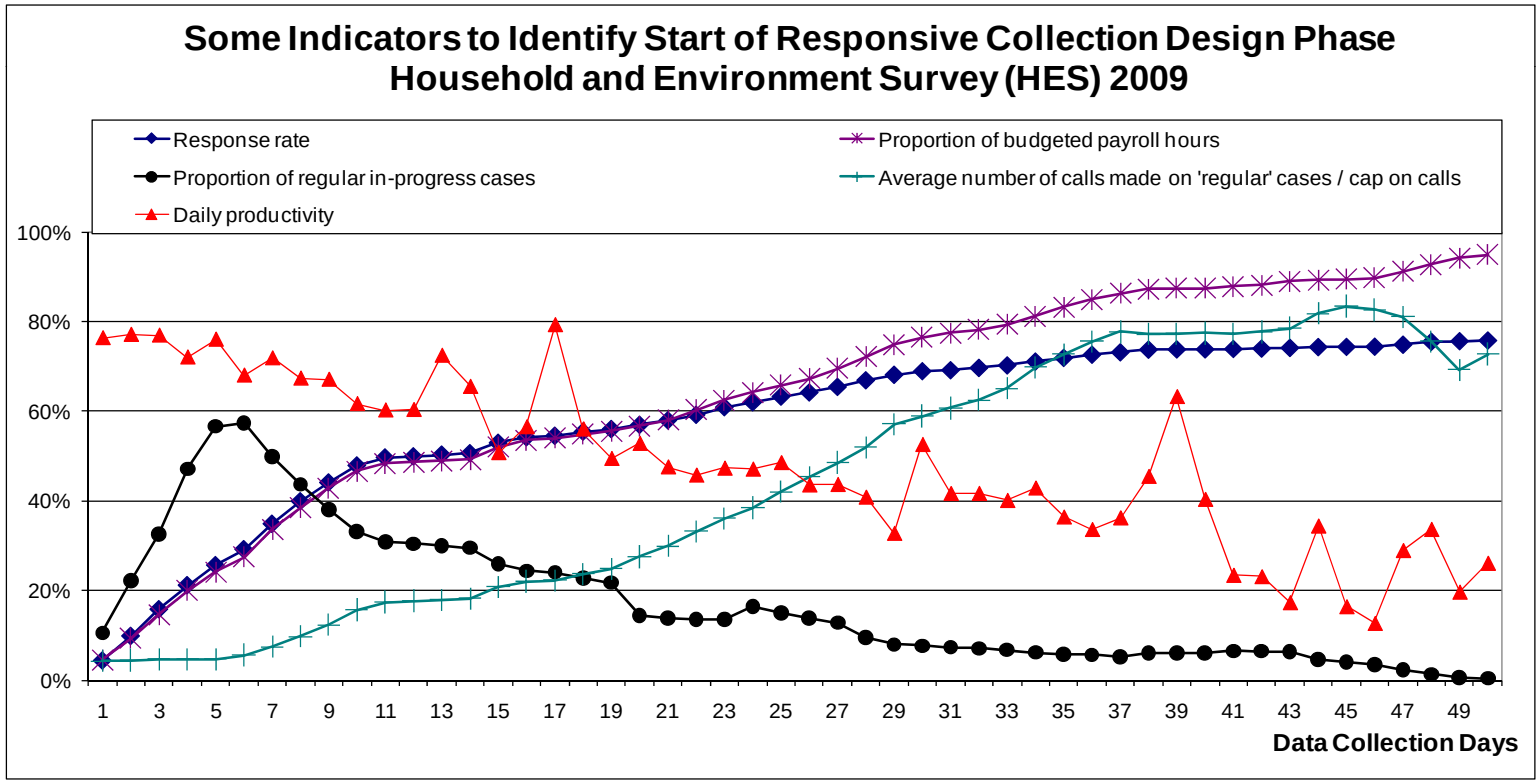
- Blaise Transaction History (BTH) file
  - A record is created each time an open case is closed, either for data collection or other purposes - a record is constructed for each call
  - Main variables
    - Survey, cycle, Regional Office (RO) ID, interviewer identification
    - Date, start time and end times of the call
    - Duration of the call and associated time slice
    - Outcome code (e.g. complete, appointment, no contact)
- Interview payroll information
  - Total payroll hours represents the hours charged to the survey

➢ ***Historical information since 2003 for all surveys***

➢ ***Updated on a daily basis***

Key Indicators throughout Data Collection for Responsive Collection Design (RCD)

- Key indicators: quality (response rate), cost (% of the budget spent), productivity and responding potential of the remaining in-progress cases.



Lessons Learned

- Substantive efforts spent close to the end of data collection yield relatively small marginal returns

- Lots of effort (calls and time) is put on average on cases for which an interview is not conducted at the first contact

- Other types of measures showed the same trend
  - Average number of days between the n^{th} and $(n+1)^{th}$ calls
  - Proportion of cases called on 2 consecutive days
  - Proportion of cases called more than x times on a given day

- Interviewer staffing levels not always optimally allocated with respect to the workload sample and the expected productivity

➢ *Develop a draft framework to improve the cost-efficiency of data collection*

Cost-Efficient Framework

- "Collection process and practices" is not the only dimension to take into account to improve the cost-efficiency of data collection

- The Cost-Efficiency Framework has 5 dimensions:

1. Metrics used for costing and budgeting surveys
2. Resources allocation within surveys
3. Resources allocation between surveys
4. Collection process and practices
5. Operational constraints

Two Research Projects

- Simulation project
  - Modeling and simulation of survey collection using paradata
  - Canada Survey of Giving, Volunteering and Participating (CSGVP)

    RDD Survey, initial sample ~ 90,000, number of calls ~ 500,000
    Response rate 54%
    No cap on calls

- Optimization project
  - Optimizing CATI call scheduling to minimize data collection costs
  - Survey of Labour and Income Dynamics (SLID)

    Longitudinal survey: sample ~ 35,000, number of calls ~ 400,000
    Response rate: 72%
    Cap on calls: 40

Microsimulation Project

- Recreate CATI collection environment in the call centers
  - Model that operates at the case level
  - Every call is simulated
  - Two parts: modeling and simulation
- Advantages of microsimulation
  - Allows manipulation of collection parameters (i.e. different scenarios) in a controlled environment
  - Test the impact of each scenario prior to collection: Many strategies can be tested - Not possible in the field (and more costly)
  - Can compare the results of many strategies and identify the most promising ones
  - Can take into account some operational constraints (e.g. capacity)

➢ ***Ultimate goal: Make CATI survey collection more efficient***

- Collection process and resources allocation

Modeling using Paradata

- Use existing BTH record - CSGVP
  - Call outcome: Multinomial logistic regression
  - Call duration: Create histograms and fit distributions for each of the outcomes
- Output model parameters
  - Estimated parameters from logistic regression model
  - Fitted distribution and parameters
- Input into simulation model
  - Create a ‘simulated’ BTH file

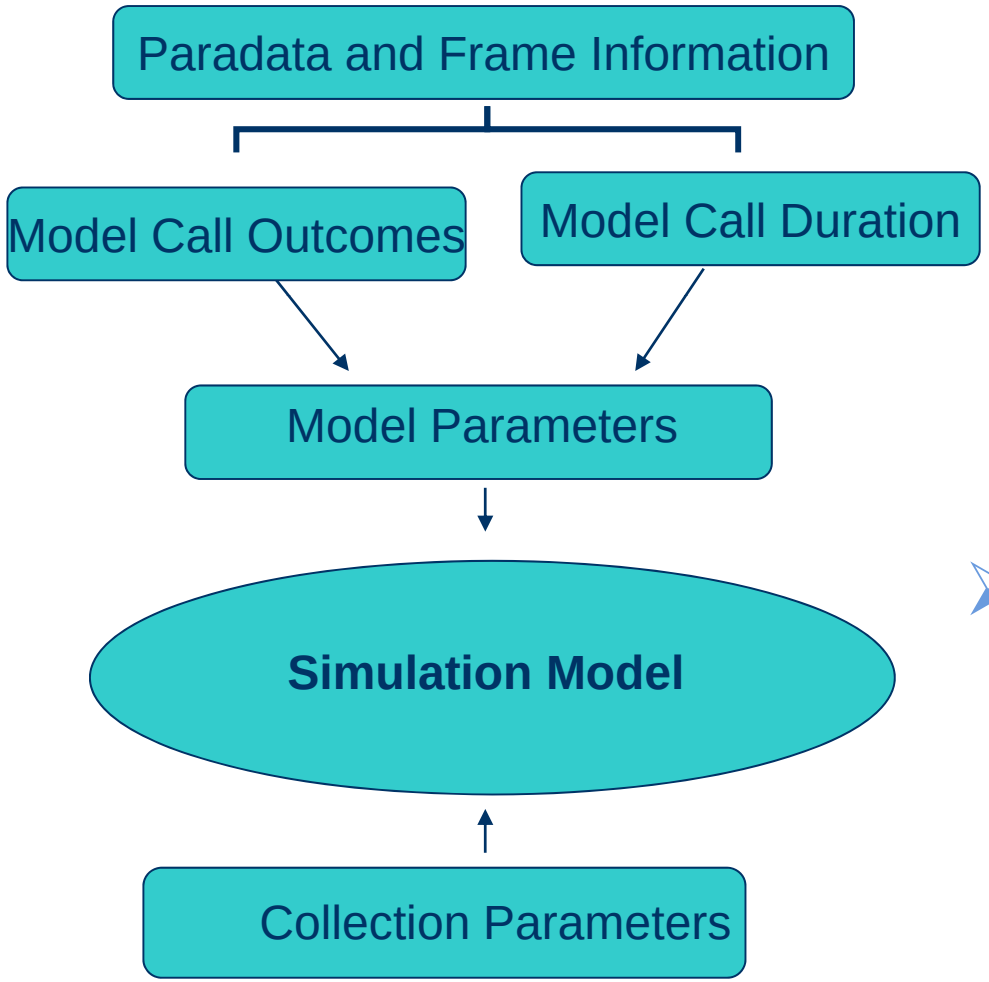
Modeling using Paradata: Call Outcome

- Multinomial Logistic Regression Model

$$\log \frac{p_j}{p_{k+1}} = \sum_{i=1}^{n} \beta_{ij} x_i \quad \text{for } j = 1, \ldots, k; \sum_{j=1}^{k+1} p_j = 1$$

  - Model probability of outcomes
  - $k+1 = 5$ outcomes
    - Unresolved, out of scope, refusal, other contact, respondent
  - $x_i = 7$ explanatory variables from paradata
    - Time of call(3); afternoon, evening, weekend
    - Residential status(1)
    - Call history(3): unresolved, refusal, contact
  - p_j = probability of outcome j
  - β_{ij} = parameters from logistic regression model
    - Estimated parameters from model entered into simulation

Microsimulation



- **Examples of collection parameters**
  - Distribution of interviewers
  - Definition of time slices
    - Distribution of calls
  - Call scheduler rules
    - Flows and priorities

- Software
  - SAS Simulation Studio

Various Scenarios

- Common parameters
  - Three time periods for interviewer distribution:
    - 9:00-13:00; 13:00-17:00; 17:00-21:0
  - 10,000 cases (about 2/3 residential)
  - 40 days of collection
  - Fixed total of 4,800 interviewer-hours
  - Limit number of calls to three refusals
  - Interviewer distribution kept fixed throughout data collection
  - Time slice strategy kept fixed throughout data collection

Results

- Need to have more interviewers in the evening
- Time slice strategy needs to be aligned with interviewer distribution
- No time slice (X) approach seems to have best performance

Scenario	Interviewer Distribution	Cap on Calls	Time Slices	Response Rate	% Complete	% Finalized	% Capped	Interviewer-Hrs Utilization
A	10-10-10	X	X	66.0%	41.8%	78.5%	10.4%	100.0%
B	10-10-10	5, 20	X	64.0%	41.7%	76.6%	23.4%	98.4%
C	12-9-9	X	X	66.0%	41.3%	78.7%	11.4%	100.4%
D	12-9-9	5, 20	prop	63.2%	41.1%	76.1%	16.1%	97.4%
E	9-9-12	X	X	67.5%	42.6%	79.5%	11.2%	100.5%
F	9-9-12	5, 20	X	63.8%	41.9%	76.3%	23.8%	97.3%
G	9-9-12	5, 20	equal	60.4%	39.8%	73.9%	13.3%	91.7%
H	9-9-12	5, 20	prop	63.5%	41.8%	76.0%	17.2%	96.8%

Notes: 1. Response rate = 100 * Complete / (Total cases - OOS)

2. % Finalized = 100 * (Complete + OOS) / Total cases

3. % Capped includes capped refusals

Optimization project

- Minimize CATI costs for given response rate
  - Model built at aggregated level within each regional office
  - Two parts: modelling and optimization at regional office level
- Advantage of macro level approach
  - Provides guidelines of how assignment of CATI interviewers can be improved
  - Uses operation research procedures for optimization, and can be adapted to more complex configurations and operational constraints

Modeling using Paradata

- Use existing BTH record - SLID
  - Call outcome success summarized by time slice as probability of completing a questionnaire
  - Simple logistic regression or regression
- Output estimated model parameters
  - Estimated parameters from logistic regression model
  - Smoothed probabilities of completing a questionnaire
- Input smoothed probabilities into optimization
  - Create optimal CATI mix by time slice subject to cost and operational constraints

Modeling Probabilities using Paradata

- Regression
  - Simple: $p_s = \beta_0 + \sum_{i=1}^{n} \beta_i x_{is} + e_s$ for $s = 1, \ldots, S;$
  - Logistic: $\text{logit}\left(\frac{p_s}{1-p_s}\right) = \beta_0^* + \sum_{i=1}^{n} \beta_i^* x_{is}$ for $s = 1, \ldots, S;$
  - x_{is} = explanatory variables from paradata
    - Time of call: morning, afternoon, early and late evening
    - Average cumulative number of calls **or** unit cost up to and including time slice *s*
  - p_s = probability of a completed questionnaire within time slice *s*

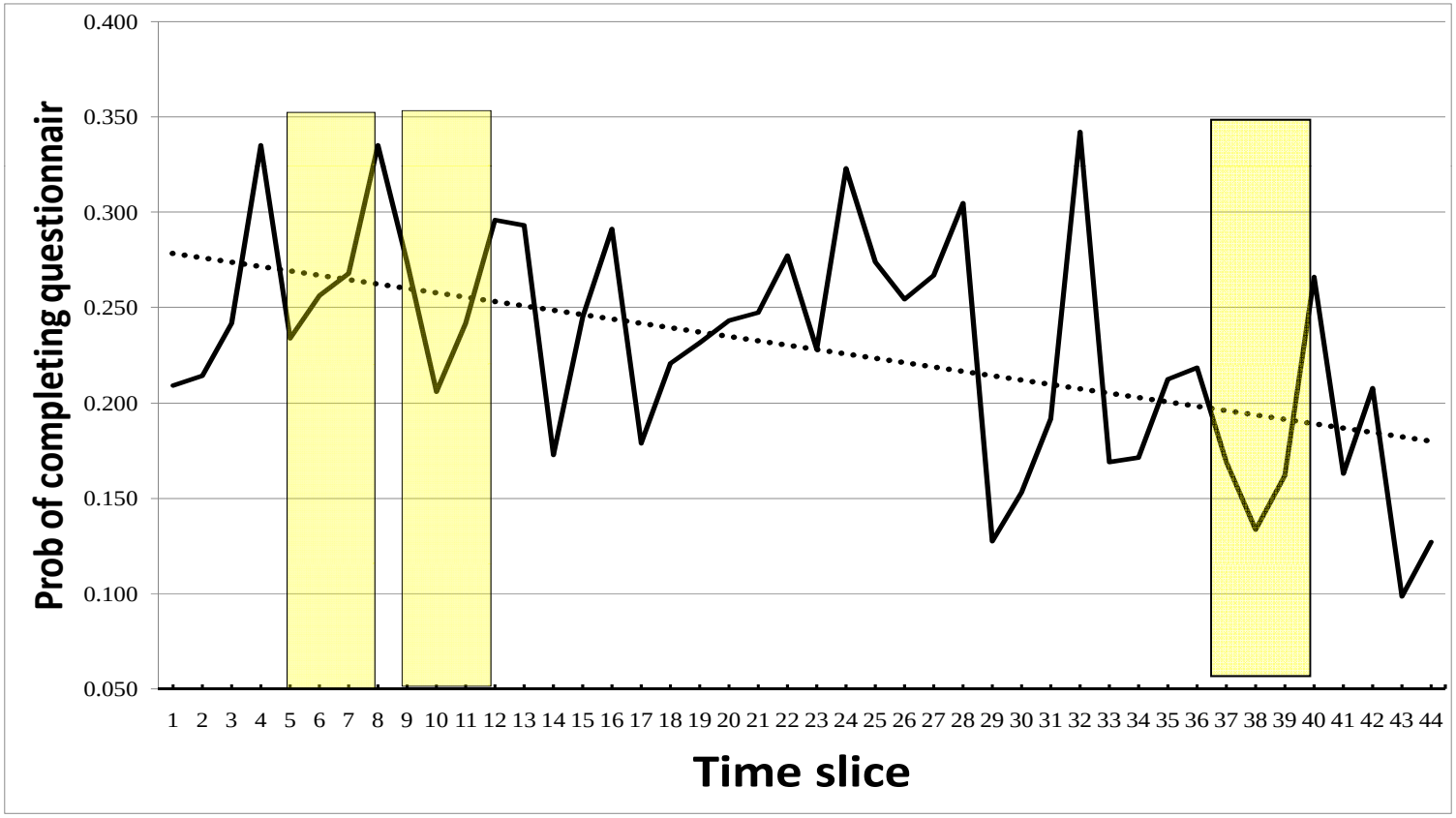
Optimizing CATI schedule

- The total data collection cost is given by

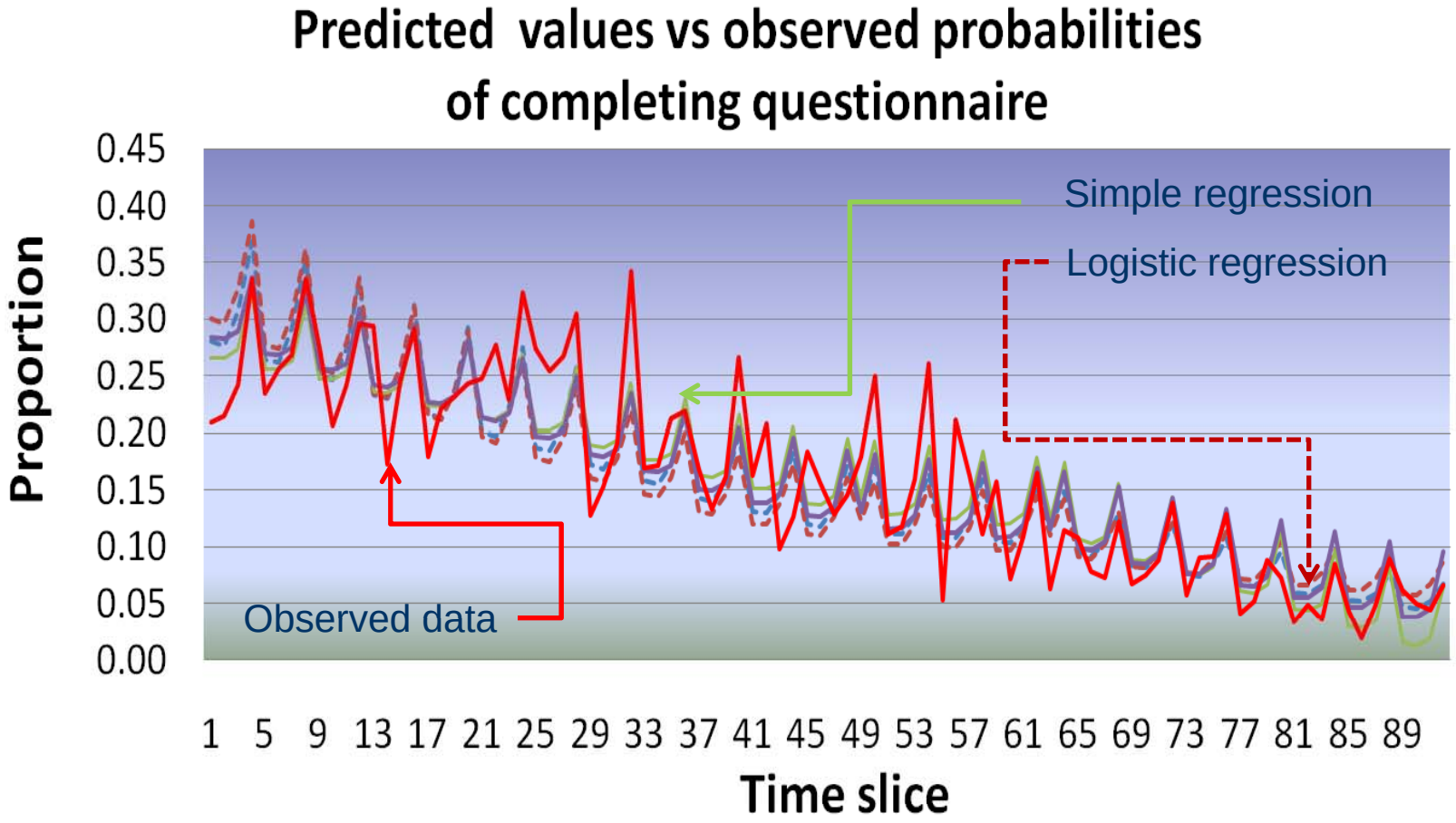
$$f = \sum_{s=1}^{S} \left\{ t_1 \tilde{p}_s \, c_s + t_2 \left(1 - \tilde{p}_s\right) c_s \right\}$$

  - t_1 and t_2: costs for success / failure
  - $\tilde{p}_s$: smoothed p's
- The "call" vector $\boldsymbol{c} = \left(c_1, c_2, ..., c_S\right)$ minimizes f subject to
  - The number of calls for each time slice is greater than or equal to zero, and
  - The expected response rate $\sum_{s=1}^{S} \tilde{p}_s c_s \,/\, n$ is equal to a pre-specified response rate R.

Summary of results for SLID



Summary of results for SLID



Summary of results for SLID

- **Regression**
  - Not much difference between using simple or logistic regression for this data set: fits are pretty good
  - Intercept and continuous variable significant for all regional offices
  - Best time period for calls depends on regional office:

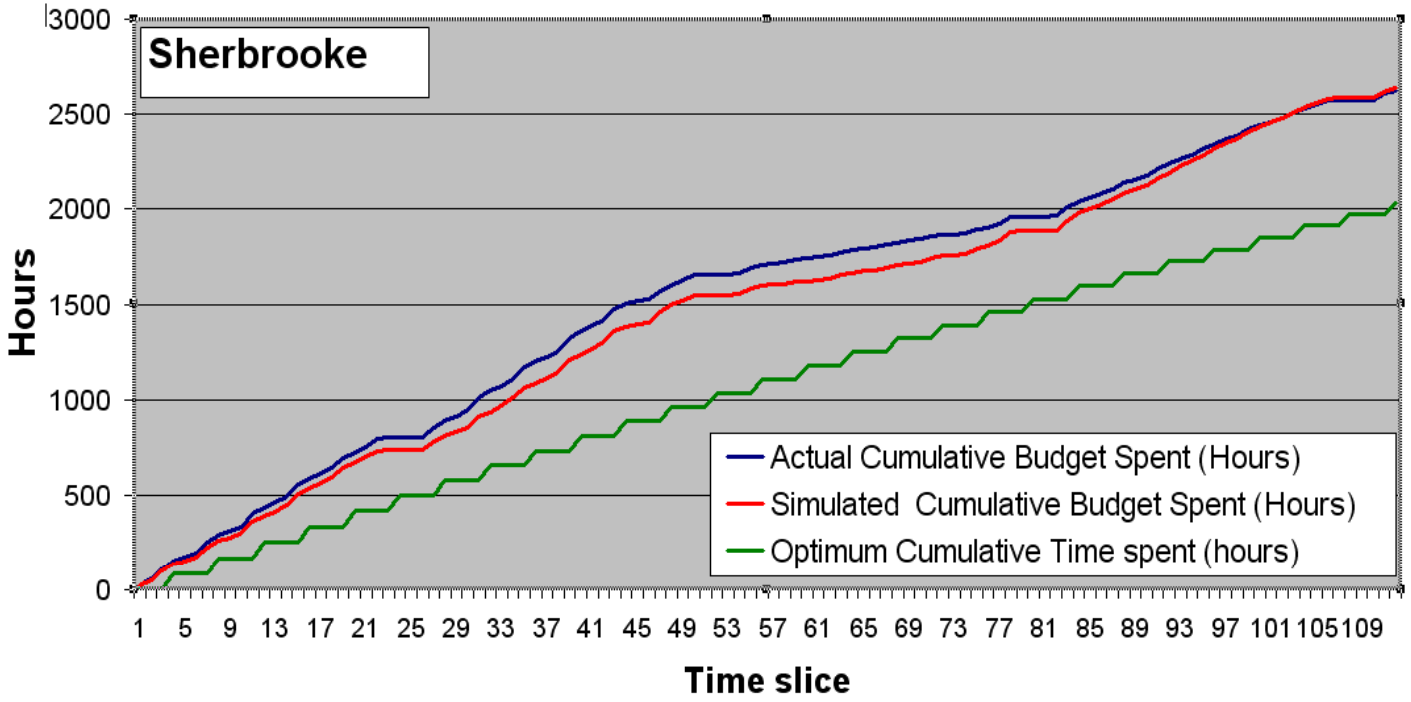
    Late evening

    Morning, early and late evening

    All time periods good

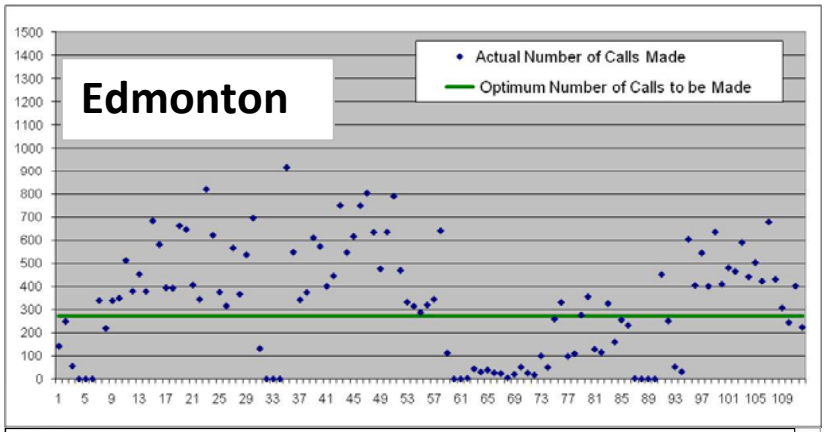
Summary of results for SLID

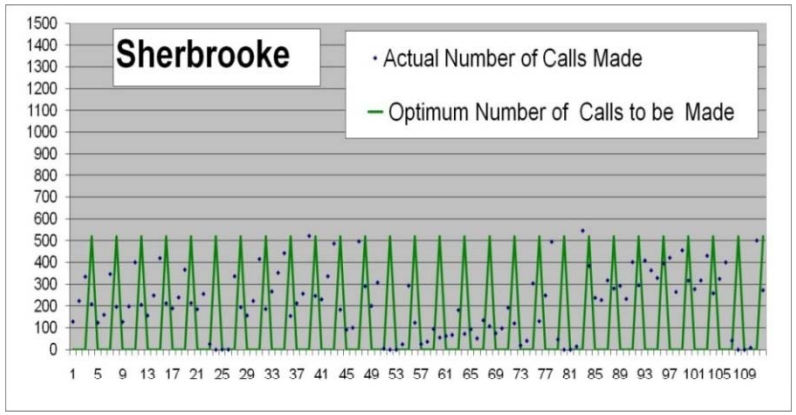
- **Optimization**
  - Gains between 9% to 22% possible **given** no restrictions on interviewer allocation to achieve same target response rate

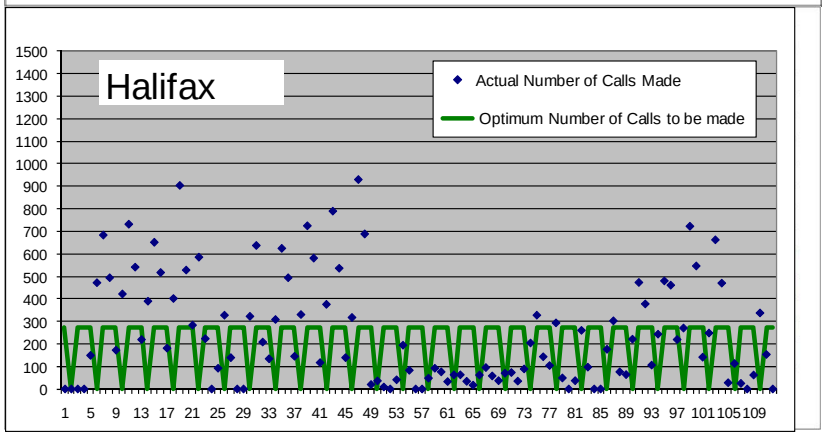


Summary of results for SLID

- **Optimization:** Allocation of calls throughout day varies







- Work flow should be uniform throughout collection period
- Calls should be made during appropriate periods within the day

Conclusions and Future Work

Microsimulation and Optimization

- Improve logistic models by adding more auxiliary variables
- Include more complicated collection procedures in the model such as interviewer characteristics
- Simulate and optimize collection with multiple surveys and interviewers carried out concurrently to gage impact on costs
- Run simulation for a survey to predict outcome and compare with actual results from field

Conclusions and Future Work

- Individual assignment of interviewers to time shifts within the day needs to reflect :
  - Legal , ergonomic, and operating constraints
  - Minimum and maximum number of days that interviewers work within the week
  - Shift duration per day (no more or less than a fixed number of hours), including starting time range of each interviewer
  - Number of shifts within a day should be reasonable

Conclusions and Future Work

- How do we do above?
  - Extend optimization to include mix of interviewers and surveys
  - Translate the number of calls within each shift and survey into number of required interviewers
  - Use commercial software such as XIMES to account for constraints, and schedule each interviewer by time shift (Gartner, Musliu, and Slany 2001)

Gartner, J. Musliu, N., and Slany W. (2001). Rota: a research project on algorithms for workforce scheduling and shift design optimization, *AI Communications*, 14, 83-92

For more information, please contact

- Mike Hidiroglou
  - mike.hidiroglou@statcan.gc.ca
- François Laflamme
  - francois.laflamme@statcan.gc.ca
- Yves Bélanger
  - yves.belanger@statcan.gc.ca