Paper 196-2011

An application of SAS® Simulation Studio: The Microsimulation of a Computer Assisted Telephone Interviewing System

Yves Bélanger, Statistics Canada, Ottawa, Canada
Kristen Couture, Statistics Canada, Ottawa, Canada
Elisabeth Neusy, Statistics Canada, Ottawa, Canada

ABSTRACT

In order to carry on large-scale telephone surveys, statistical agencies commonly use Computer Assisted Telephone Interviewing (CATI). For example, Blaise is the CATI system used at Statistics Canada. One of the features of Blaise is a call scheduler that regulates how interviews are scheduled and conducted, depending on a variety of parameters. It is not easy to find optimal values for these parameters, as field tests in real time are typically costly and take time to conduct and analyze. It is important, however, to try to find strategies that decrease the costs of interviewing, given that data collection usually accounts for the largest part of a survey's overall budget.

In this paper we show how microsimulation can be used to help find strategies to improve the efficiency of data collection. More specifically, we describe the characteristics of a prototype of a microsimulation system for CATI surveys built using SAS Simulation Studio.

INTRODUCTION

Of all the activities associated with surveys, data collection represents one of the largest segments of the global budget. For a number of years now, a constant increase in collection costs, combined with the gradual decrease in response rates has been observed at Statistics Canada. Various strategies have been employed during the collection phase in an attempt to reverse these trends. Among them are: the adoption of a limit on the number of calls (which is intended to make better use of resources) and the establishment of time slices for calls (which leads to a better distribution of calls throughout the day). Other strategies being investigated include using the best time to call (provided by the respondent or obtained through modelling), as well as experimenting with calling priorities. The common goal for all these strategies is to achieve a more efficient collection process.

New measures introduced in the collection process are often evaluated in the context of field tests conducted in real time, which are typically costly and time consuming. Moreover, uncontrollable circumstances happening in the field may also jeopardize the tests and make their results difficult to interpret.

As an alternative to real-time field experiments, and in order to address their limitations, a prototype of a Computer Assisted Telephone Interviewing (CATI) survey collection process was built using SAS Simulation Studio. Microsimulation is a modelling technique that operates at the level of individual units and is used to simulate large representative populations of these units. In the context of CATI surveys, units are sampled telephone numbers, also known as cases. The prototype built involves the following elements: the cases, the servers (interviewers), the waiting queues for cases yet to be interviewed, the calls and their results, the rules governing priorities, and the flow of cases in the system.

A FEW WORDS ABOUT SIMULATION STUDIO

SAS Simulation Studio is a Java-based application that uses discrete-event simulation to model and analyze systems. Simulation Studio provides a graphical user interface (GUI) and a batch interface. A model is built using blocks that communicate with each other through ports. Each block corresponds to a well-defined and specialized functionality. For more information, please refer to the documentation: SAS Institute Inc. (2009).

Throughout the description of the application in this paper, reference will be made to some of the blocks available in Simulation Studio, such as Numeric Source block, Delay block, Formula block, etc. However, for reasons of simplicity, we will not go into the details of the construction of the model. Instead, we will focus on the description of its most important characteristics.

Figure 1 shows an example of what a Simulation Studio model or project may look like, in this case, a partial view of the model described in this paper. It is a good illustration of how visually different a model built using Simulation Studio can be from a typical SAS ”program” or application. It also illustrates the various blocks used in a model and how they are inter-connected.

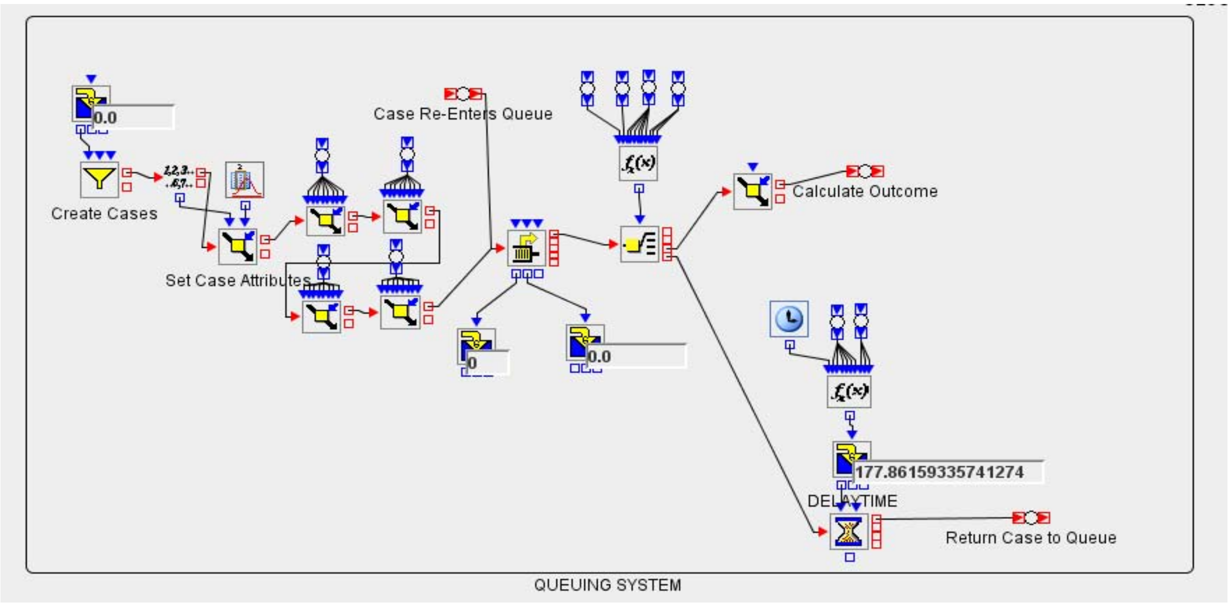

Figure 1 : Partial view of simulation model built using Simulation Studio

Simulation Studio is one of several products on the market that can be used to simulate the daily operations of a call center. This is a specific application of microsimulation that seems to be widespread, judging by the amount of information available on the Internet. There are basic differences, however, between a typical call center which receives inbound calls, and a collection agency such as Statistics Canada that relies on outbound calls. Fortunately, there are also many similarities, such as agents (or interviewers) and queues, which make the creation of a simulation model for survey collection possible.

OVERVIEW OF THE SYSTEM

There are several components that go into constructing a microsimulation of CATI collection, as shown in Figure 2. A central component is the computer simulation model, which requires input parameters in order to replicate the collection process as accurately as possible. As shown in the figure, two types of parameters are entered into the simulation: model parameters and user-defined parameters. Model parameters are determined prior to performing any simulation runs and are calculated from information available on the survey frame (a list of elements of the population of interest) as well as data obtained through collection of a previous survey (also called paradata). The calculated model parameters are used to assign call outcomes and call duration. In comparison, user-defined parameters, such as the maximum number of calls allowed and the distribution of interviewers, can be changed prior to each simulation run in order to control and manage the collection process.

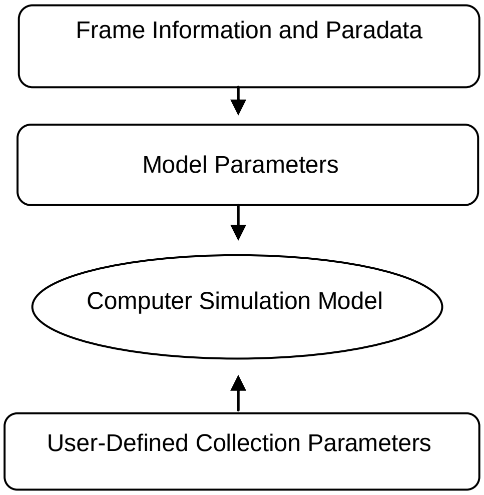
Frame Information and Paradata
Model Parameters
Computer Simulation Model
User-Defined Collection Parameters

Figure 2: Overview of Building a Microsimulation of Data Collection

DESCRIPTION OF THE SIMULATION MODEL

As mentioned previously, the simulation model is built to replicate CATI collection. The logical sequence of the simulation model is as follows.

At the start of the simulation, a batch of cases is created, representing telephone numbers which need to be called to conduct interviews. Once created, the cases are sent into a “calling” queue where they will wait until an interviewer resource becomes available. When an interviewer resource becomes available, the next available case is taken from the queue and a call outcome is assigned to the case. The interviewer resource will remain busy for a simulated duration of time depending on which call outcome has been assigned to the case.

The outcome assigned to the case determines whether the case receives a “finalized” or “in progress” status. Cases that receive a finalized status are sent out of the system; otherwise, the case receives an in progress status and is returned to the calling queue. Calls that result in a successfully completed questionnaire or in an out of scope are examples of outcomes that produce a finalized status. Calls that result in no contact or a busy signal are examples of outcomes that produce an in progress status.

The simulation will continue to run until either the collection period has finished or all cases have resulted in a finalized status. At this point, a file is produced containing the history of each call attempt.

In the next sections some important features of the simulation model are described in more detail.

CALL OUTCOMES

A first set of model parameters is used by the simulation to assign outcomes to call attempts. During actual collection, telephone calls result in a variety of possible outcomes, such as completed questionnaire, out of scope, answering machine, or busy signal. The probability of the various call outcomes depend on a number of different factors. In order to simulate data collection as realistically as possible, the probabilities of the various call outcomes are modeled using actual data. These data may be variables available a priori on the sampling frame, or data obtained through the actual collection process of a real survey previously conducted, and similar to the survey that is being simulated. The latter data are referred to as paradata.

The probabilities for the call outcomes are modeled using a multinomial logistic regression model. In general, suppose that there are k+1 possible outcomes, and that the probability of each outcome is denoted by p_1, p_2, … p_{k+1}. As well, suppose that each call attempt has n characteristics denoted by variables x_1, $x_{2,}$…, x_n. The outcome probabilities are modeled using the following multinomial logistic regression model:

$$\log\left(\frac{p_j}{p_{k+1}}\right) = \sum_{i=1}^{n} \beta_{ij} x_i, \quad \text{for } j = 1,\ldots,k, \quad \text{where} \sum_{j=1}^{k+1} p_j = 1.$$

The LOGISTIC procedure is used for that purpose. The parameters estimated from the model, $\hat{\beta}_{ij}$, *i*=1,…,*n*, and *j*=1,…,*k* become inputs used in the simulation. When the outcome of a call needs to be generated during simulation, first the probability of each outcome, the $\hat{p}_j$, are calculated, and then a random number from a uniform distribution on [0,1] is used to randomly select one of the possible outcomes. For a more complete description of the regression model and how its results are used, please refer to Couture, Bélanger and Neusy (2010).

As the values of the parameters $\hat{\beta}_{ij}$ remain unchanged during the simulation run, they are entered into the system prior to the start of the simulation. This is done easily using Numeric Source blocks, a feature of Simulation Studio, one block for each column of a SAS data set created by PROC LOGISTIC and containing the $\hat{\beta}_{ij}$. Note that there are a total of *k* times *n+1* such parameters, where *k* is also the number of columns of the input data set.

The calculation itself of the probabilities $\hat{p}_j$ and of the final outcome is done within a number of Formula blocks, which allow complex calculations to be performed.

DURATION OF CALLS

A second set of model parameters is used by the simulation to assign the duration of the call. During actual collection, the duration depends on the call outcome. For example, an attempt which resulted in a fully completed questionnaire will tend to have a longer duration than an attempt where no contact was made.

Once again, paradata are used to model call duration. The distribution of the call durations obtained during collection for each outcome is fit to a theoretical distribution supported by Simulation Studio, such as a normal distribution. The parameters of the distribution are imported into the simulation system and used to randomly assign the duration of each call attempt.

Once the duration has been determined, the case is set aside (in a Delay block) for that duration of time, after which it re-enters the system.

As with the parameters that determine the call outcomes, parameters for call durations are read into the system prior to the simulation run, using a Numeric Source block. The duration of a call is generated once its outcome is known, using a Formula block.

CAP ON CALLS AND TIME SLICES

In order to control costs, as well as respondent burden, a limit on the number of call attempts is used in most of Statistics Canada household surveys conducted by telephone. That limit may depend on characteristics known a priori for each case, and on the survey being considered. Once a case reaches its limit, upon completion of the call, it leaves collection and is considered finalized. A limit on the number of calls is also applied in the case of refusals. Typically, for Statistics Canada household surveys, only two additional conversion calls are allowed after a first refusal is encountered.

A cap on calls is typically used in conjunction with time slices, as follows. A typical interviewing day is partitioned into a number of non-overlapping periods of time called time slices. Simple examples of time slices are morning, afternoon and evening. Time slices are used to ensure that calls are well distributed during the day, which is especially important when the cap on calls is low. This is achieved by distributing the maximum number of calls among the time slices, so that each time slice has its own cap on calls. For example, a limit of 20 calls could be partitioned as follows: 5 calls in the morning, 5 calls in the afternoon and 10 calls in the evening.

Once again the definition of time slices is entered into the system prior to the simulation run using SAS data sets, read using Numeric Source blocks. Time slice parameters are considered to be user-defined parameters, as they may typically be changed before each simulation run, in order to determine the best time slice scenario.

During the simulation, when an interviewer resource becomes available, the case with the highest priority (e.g., the one with the minimum number of call attempts) will leave the queue to be held by the interviewer resource. At this point, the case is checked to see whether it has already reached its maximum number of attempts within the current time slice. If it has, then the interviewer resource releases the case, which is sent to a Delay block, and the interviewer resource returns to the resource pool where it will pick up the next available case in the queue. Otherwise, the case and its corresponding interviewer resource will continue through the simulation to receive a call outcome.

INTERVIEWERS

The number of interviewers and their distribution during the day and throughout the collection period are other examples of user-defined parameters that are used in simulation runs. Simulation Studio has many features to determine how interviewers are defined and scheduled during a simulation. Interviewers are represented by resource entities, and features such as the Resource Agenda and the Resource Scheduler are used together to tell the resource pool how many interviewer resources are available at any time during a simulation run.

It is important to note that interviewer shifts need not be defined to be similar to time slices (and vice-versa). However, when testing various scenarios involving changes in interviewer allocation as well as changes in allocation of call attempts to time slices, it is often easier to interpret the results when interviewer shifts and times slices are the same. This will be illustrated in the example given later.

SIMULATION CLOCK

In order to accurately simulate the collection process, there needs to be a clock that continuously keeps track of the hour, day, and total days of collection that have passed. Simulation Studio has a built-in clock which keeps track of the simulation time in minutes. The current simulation time is used to calculate various attributes such as those that determine call outcomes and time slices, for example.

OUTPUT FILE

At the end of the simulation, each call attempt that has been made is output to a SAS data set so that statistics on the simulated collection process can be gathered. This is done using a Bucket block.

AN EXAMPLE USING A RANDOM DIGIT DIALLING SURVEY

Random Digit Dialling (RDD) surveys are widely used at Statistics Canada and elsewhere in the world, and are often conducted with the help of a CATI system. They use a frame of telephone numbers, and samples are created by randomly selecting or generating numbers from the frame. Phone numbers are classified before the start of collection as residential (when an address can be associated with the phone number), or unknown (status unknown). Phone numbers that are found during collection to be not in service or to correspond to a business are classified as out of scope. Cell phones may also be considered as out of scope.

THE SAMPLE

In this example, a sample of 10,000 phone numbers is generated. About 2/3 are randomly assigned a residential status, according to the proportion encountered in practice, and the remainder are considered unknown. Numbers with a residential status are assigned a cap on calls of 20, and those with an unknown status a cap of 5. The reason for the lower cap for unknown numbers is that it usually takes fewer calls to determine the real status (in scope or not) for these numbers. However, if during collection an unknown number is found to be in scope (that is, is found to reach a household), the cap is raised to 20 to increase the chances for a completed interview.

In addition to the usual cap on calls, a limit on the number of calls is also applied in the case of refusals. Typically, for Statistics Canada household surveys, only two additional conversion calls are allowed after a first refusal is encountered. Accordingly, a cap of three call attempts is used in the example.

THE OUTCOME MODEL

As explained previously, probabilities for call outcomes have to be calculated before the start of the simulation. To this end, paradata from a real RDD survey, the 2004 Canada Survey of Giving, Volunteering and Participating is used. The multinomial logistic model used is fairly basic as it only has five different outcomes:

1. Unresolved (e.g., busy signal, no answer)
2. Out of scope (e.g., cell phone, business)
3. Refusal
4. Other contact – questionnaire not completed (e.g., answering machine, appointment)
5. Questionnaire completed.

Seven explanatory variables are used in the model:

1. Afternoon (= 1 if call made between 12 and 5; 0 otherwise)
2. Evening (= 1 if call made between 5 and 9; 0 otherwise)
3. Weekend (= 1 if call made on weekend; 0 otherwise)
4. Residential (= 1 if initial status was residential; 0 otherwise)
5. Unresolved (= 1 if call history is only unresolved; 0 otherwise)
6. Refusal (= 1 if call history shows at least one refusal; 0 otherwise)
7. Contact (= 1 if call history shows at least one contact; 0 otherwise).

Note that the first three explanatory variables are instant attributes of the call being made, the fourth one is a characteristic of the sampled unit and the last three are cumulative attributes of all calls previously made for that unit.

This combination of outcomes and explanatory variables (plus an intercept) yields a total of (5-1) x (7+1) = 32 parameters. These are input into the system as explained previously.

With respect to call durations, the following table shows distributions that are used, as determined by the analysis of paradata.

Table 1: Distributions of Call Durations for Possible Outcomes

Outcome	Distribution in minutes
Unresolved	Exponential(1.6)
Out of scope	Exponential(1.7)
Refusal	Exponential(3.3)
Other contact	Exponential(2.7)
Completed	Normal(29.8,10.8)

THE TIME SLICES

Parameters governing the use of time slices also need to be determined in advance. For this example three time slices are created: morning (from 9:00 to 13:00), afternoon (from 13:00 to 17:00) and evening (from 17:00 to 21:00). Note that each of these time slices represents a work shift of four hours in duration. Using time slices of equal length makes it easier to compare scenarios with a different distribution of interviewers to these time slices, when the total number of interviewers remains fixed.

A total of 30 interviewers are used every day, distributed for simplicity among the time slices described in the previous paragraph. The total collection time covered by the simulation is fixed at 40 days. The total interviewing capacity is calculated as 40 days x 30 interviewers x 4 hours = 48,000 interviewer-hours.

THE RESULTS

Table 2 shows the results of various simulation scenarios. In these scenarios the only parameters that were changed were the interviewer distribution and the strategy defining the use of cap on calls and time slices. It was found that using a sample size as large as 10,000 produced results with very little variability. That is, there was little difference between two distinct simulation runs generated for the same scenario (using different starting seeds). It was deemed necessary to run more than one simulation only when the results of two scenarios were very close. In those cases five simulations were ran for each of the two scenarios, and the median result for each (in terms of proportion of completed cases) was retained and shown in Table 2.

The first four columns of Table 2 describe the characteristics of each scenario. “Scenario” identifies the scenario that was run, from A to H. “Inter Distr” represents the interviewer distribution; 10-10-10, for example, means that there were 10 interviewers in each of the three time slices: morning, afternoon and evening. “Cap Calls” indicates if a cap on calls was used or not; when it was used, the cap was set at 5 for cases of unknown status and 20 for residential cases, as explained previously. “Time slices” shows if a time slice scenario was used, and which one; “equal” means that calls were distributed fairly equally among time slices, while “prop” means that calls were distributed among time slices proportional to the number of interviewers.

The last four columns show the actual results of the simulations. The column “Complete” shows the proportion of cases that returned a complete questionnaire, while “Finalized” shows the proportion of cases that returned a complete questionnaire or were solved as out of scope. The column “Capped” represents the proportion of cases that were capped and finally, “Inter Util” shows the proportion of the time that interviewers were busy.

Note that in all scenarios a maximum of three refusals was allowed, which explains why the column “Capped” shows cases being capped even when a limit on “regular” call attempts was not used. For example, in Scenario A which did not have such a limit, 10.4% of cases were still terminated when they reached their third refusal.

Table 2: Results of Various Simulation Scenarios

Scenario	Interv Distr	Cap Calls	Time Slices	Complete	Finalized	Capped	Interv Util
A	10-10-10	X	X	41.8%	78.5%	10.4%	100.0%
B	10-10-10	5, 20	X	41.7%	76.6%	23.4%	98.4%
C	12-9-9	X	X	41.3%	78.7%	11.4%	100.0%
D	12-9-9	5, 20	prop	41.1%	76.1%	16.1%	97.4%
E	9-9-12	X	X	42.6%	79.5%	11.2%	100.0%
F	9-9-12	5, 20	X	41.9%	76.3%	23.8%	97.3%
G	9-9-12	5, 20	equal	39.8%	73.9%	13.3%	91.7%
H	9-9-12	5, 20	prop	41.8%	76.0%	17.2%	96.8%

Each simulation took about 15 minutes of real time to run, and generated between 65,000 to 75,000 call attempts and their outcomes. It was very interesting, and reassuring, to see that many of the simulation results matched those seen over the years with real surveys at Statistics Canada since the introduction of a cap on calls and time slices. In the following paragraphs a few results are highlighted and briefly discussed.

The impact of a cap on calls

This is seen when comparing scenarios A and B, or E and F. The proportion of “good” outcomes, i.e. cases completed or finalized, decreases slightly with the cap on calls (scenarios B and F), while the proportion of capped cases doubles. Also, the interviewer utilization decreases with the cap on calls, which in this context just reflects the fact that collection is finished before the simulation run is over (i.e. before the end of the 40 days of collection), all cases being either finalized or capped. The slight loss in quality with the cap on calls is therefore offset by the savings in interviewer costs, provided interviewers can be re-assigned to another survey.

The impact of interviewer distribution

This is seen when comparing scenarios A, C and E, which differ only in the interviewer distribution during the day. The proportion of completed cases for scenario C is slightly lower than for scenario A, but the proportion of finalized cases is higher. This is due to a higher number of cases finalized as out of scope for scenario C, as more calls are made in the morning, a good time to identify out of scope cases such as businesses. Scenario E has a higher proportion of completed (and also finalized) cases than both scenarios A and C, which reflects the fact that in general it is easier to reach respondents during the evening.

The impact of time slices

Compare first scenarios F and G, which differ only in the application of time slices for scenario G. The impact of having time slices is clearly negative as fewer cases get completed or finalized for scenario G, while interviewer utilization decreases and fewer cases reach their cap. This is due to cases prevented from being called in a certain time slice because they have reached their call limit (i.e. their cap) in that time slice. When all in progress cases face the same situation in the same time slice, collection comes to a halt and interviewers become idle until the beginning of the next time slice. This is a phenomenon that usually happens towards the end of the collection period.

Compare now scenarios G and H, which differ only in their time slice strategy. It is scenario H with the proportional call allocation to time slices that provides the best results, which clearly shows the importance of choosing the right time slice allocation. Finally, a comparison between scenarios D and G, where scenario D shows overall better results, also illustrates how a better interviewer distribution (for scenario G) can be negated by a weaker time slice allocation. In other words, this last comparison shows how it may be as important to choose the right allocation of calls to time slices as it is to choose the right interviewer distribution.

CONCLUSION AND FUTURE WORK

We have described in this paper a prototype of a simulation model of a CATI collection process, built using SAS Simulation Studio. Although the prototype is still fairly simple, results obtained to date are very promising as they replicate similar results obtained through experimentation with real-life surveys.

In the near future, our plan is to continue the development of the prototype on two fronts. First, new outcomes and explanatory variables will be added to the outcome model in order to get a better representation of the collection process. Second, new features will be added to the simulation model itself in order to better replicate the capabilities of the call scheduler, such as how it deals with case priorities. We are also looking into expanding the prototype in order to simulate more than one survey at a time, a common situation which is a challenge especially for interviewer allocation, not only during the day but also between surveys.

REFERENCES

Couture, Kristen, Y. Bélanger, E. Neusy (2010). *Modelling and Simulation of Survey Collection Using Paradata.* Proceedings of the 2010 Joint Statistical Meetings in Vancouver, Canada.

SAS Institute Inc. (2009). *SAS© Simulation Studio 1.5: User's Guide.* Cary, NC: SAS Institute Inc.