A Discrete-Time Microsimulation Model of the 2016 Canadian Census Collection Operations: An Innovative Method for Responsive Design and Cost Control

Vincent Martin*

Abstract

Collection operations of a national census require careful planning and monitoring. With the mounting pressure to control survey costs, there is growing interest in methods to livecast the remainder of operations throughout collection activities. The problem of forecasting survey operations costs and response rates is one that has been discussed at great length already. However, models in the literature are generally planning tools that are not designed to update predictions over the course of collection operations. For the 2016 Canadian Census, Statistics Canada developed a microsimulation model designed to produce updated forecasts of subnational costs and response rates on a regular basis that reflected the most recent collection data available. The model built to meet this operational requirement produces reasonable outputs at small geography levels and was used to identify and address collection issues early during field non-response follow-up activities. In this paper, we discuss the general design of the model, methods to estimate parameters, upcoming improvements to the model that could not be implemented in time for 2016 and provide a qualitative evaluation of the utility of the model during the 2016 Canadian Census collection activities.

Key Words: Responsive design, microsimulation, survey planning, census, response model, operation model, survey cost, response rate

1. Introduction

Monte Carlo simulation models have long been used to explore phenomena too complex for classical modeling techniques. In the context of survey collection operations, microsimulation models have been used to evaluate route optimization algorithm (Chen 2008), survey response rates under different collection strategies (Doherty 2011), interviewer assisted collection (Couture, Bélanger and Neusy 2010, Karr, Cox and Kinney 2012) and several other applications. On another front, collection paradata have increasingly been used in the management and monitoring of survey collection operations (Kreuter, Couper and Lyberg 2010). Yet, examples of microsimulation models used for live collection forecasting or livecasting that integrate collection paradata on an ongoing basis are far more difficult to find, mainly because most surveys have too short a collection period or too small a sample to efficiently use such an approach. One exception is a population census which, in Canada for example, is the result of several collection activities spanning over a three months period.

Given the costs of collection operations of a national census, such methods offer the potential for high reward, especially since the resources required to develop and maintain a livecasting model are small whereas any potential gain in efficiency for collection operations that may result from the method scale with the size of the collection operation. Thus, it was sufficiently appealing for Statistics Canada to develop a microsimulation framework to forecast end of collection response rates and costs at subnational levels as a mean to improve our ability to monitor and manage collection activities for the 2016 Census. This paper describes the discrete-time stochastic process framework that was developed for the 2016 Canadian Census collection operations in such a way that it could be used to simulate

*Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, Canada K1A 0T6

the remainder of collection given the current progress while learning, where possible, from the available collection data in order to improve the forecast. This is an initial step towards livecasting survey collection operations that can also be used as a planning tool or as a responsive design tool in the monitoring and management of collection operations.

This paper begins with an overview of the 2016 Canadian Census collection operations and geography. Then, we describe the stochastic framework for the simulation and its constituents while providing methods for estimating the model parameters. Next, we elaborate on methods that did not make it into the 2016 model but will improve parameter estimation and assumptions for future iterations of the model. Finally, we provide a qualitative evaluation of the impact of the model on the collection operations of the 2016 Canadian Census.

2. Overview of the Canadian Census Collection Operations

The 2016 Canadian Census was conducted between May 02, 2016 and July 29, 2016. The collection strategy consisted of a wave methodology to promote self-response and Internet response during the month of May, which was followed by field and telephone non-response follow-up activities in June and July. The wave methodology (Mathieu 2017) is a sequence of treatments such as reminder letters, a paper questionnaire package and automated telephone reminders that are aligned with a general communication strategy through various media. There were two dwelling occupancy verification operations conducted on a subset of dwellings during the month of May.

The census collection frame is a mixture of an address-based dwelling frame and an area frame to be listed by field interviewers. Approximately 98% of the dwelling frame units in the country are in scope for the forecasting model and they can be categorized according to their collection methodology as belonging either to mail-out areas, a partition of the address-based frame, or list/leave areas, a partition of the area frame.

The Canadian Census collection operations management is decentralized with hierarchical reporting tied to the collection geography. The collection geography is broken down hierarchically into Canada, 5 Regional Offices (RO), 25 Local Census Offices (LCO), 1866 Crew Leader Districts (CLD) and 45,133 Collection Units (CU). At the lowest level of supervision, Crew Leaders supervise a small team of interviewers and are responsible to achieve response objectives in their assigned CLD. This results in interviewer mobility being frequent within a CLD but less so across CLDs.

A key objective of collection operations is to achieve as high a response rate as possible while maintaining homogeneity of response rates across collection areas in Canada. To ensure resources are used appropriately, we use a tolerance process designed to halt costly operations such as field and telephone non-response follow-up in areas that have met certain criteria. The initial tolerance strategy consists of progressively relaxing tolerance criteria during collection but it can be updated at any time in reaction to the observed collection progress and remaining resources available. Therefore, the tolerance strategy acts as a responsive design tool as it enables Statistics Canada to modify the collection strategy on the fly in order to optimize response rates and homogeneity objectives.

3. A discrete-time stochastic process framework for collection operations

During collection activities, census management receives separate reports on the current level of expenses for various pay elements and the state of collection progress described through a tolerance rate, a proxy for the response rate that is used in the tolerance process. This tolerance rate accounts for preliminary information that may be reverted later such as

Table 1: Dwelling Frame Unit Collection Statuses

Value Code	Status Label	Status Description
0	Unresolved	Unresolved is the default or initial status of dwelling frame units
1	Occupied	Dwelling is occupied and provided a valid response
2	Unoccupied	Dwelling is unoccupied
3	Cancelled	Frame unit is not a dwelling
4	Other resolved	Frame unit was occupied on Census Day but a valid response cannot be obtained (e.g. respondent moved, refusal)
5	Tolerance Met	Unresolved unit for which interviewer led collection activities were halted by the tolerance process

treating a paper response as valid as soon as Canada Post receives it only to find out during processing that it was completely empty.

The microsimulation model described hereafter is designed to forecast subnational end-of-collection tolerance rates and expenses for a subset of pay elements that constitute a large share of the operational budget, are highly correlated with collection progress and are therefore more likely to differ from planned values. For example, field interviewers account for a very large proportion of the overall collection budget and may be asked to work more hours or additional interviewers may be hired if the observed self-response rates are worse than planned in their area. By contrast, regional managers are few in numbers and are required to work about the same hours no matter the staff on strength under their supervision. Therefore, the time claimed by field interviewers is included in the model but the time claimed by regional managers is not. Put differently, pay elements that account for a small share of the overall budget or to be easily predicted represent a small risk and the increased model complexity from their inclusion outweighs the benefit of including them.

Collection progress at the dwelling frame unit level, is tracked through a classification variable with a finite set of possible outcomes called the (dwelling) collection status. For the 2016 Canadian Census forecasting model, we consider an aggregated and re-indexed subset of the collection statuses over $m = 6$ possible outcomes as described in Table 1.

To develop the microsimulation framework, we consider the probability space $\{\Omega, \Sigma, \mathbb{P}\}$ with sample space

$$\Omega = \{0, 1, \ldots, m-1\}^{\eta} = \{(x_1, x_2, \ldots, x_n) | x_i \in (0, 1, \ldots, m-1) \; \forall \; i = 1, 2, \ldots, n\}$$

corresponding to assigning a collection status to each of the η dwelling frame units in scope for the model, Σ a σ-algebra of subsets of Ω and $\mathbb{P}$ a probability measure on this space and define the stochastic process of the collection status of a single dwelling frame unit i at the end of collection day t by $X_i = \{X_i(t) : t = 0, 1, \ldots\}$ on this space. A possible method to forecast collection progress from the current state corresponds to simulating $X_i(t_{end}) | X_i(t_{now})$ where t_{end} corresponds to the end of collection operations and t_{now} the most recent completed collection day with available data. In order to build a model that can produce a forecast for any given $t_{now} < t_{end}$, we ought to define a model such that $(X_1(t), \ldots, X_\eta(t)) \,|\, \{X_i(0), \ldots, X_i(t-1) : i \in (1, \ldots, \eta)\}$ can be simulated. That is, we ought to design a model such that simulating the collection process corresponds to simulating the conditional process day by day for all remaining collection days from some

t_{now}. Modeling the process of a dwelling collection status directly is not an easy task as it involves a number of distinct activities. Therefore, we define the following processes for a dwelling frame unit i:

$$
\begin{array}{rl}
R_i = \{R_i(t) : t = 0, 1, \dots\} & \text{the self-response process} \\
N_i = \{N_i(t) : t = 0, 1, \dots\} & \text{the field non-response follow-up process} \\
C_i = \{C_i(t) : t = 0, 1, \dots\} & \text{the telephone non-response follow-up process} \\
Y_i = \{Y_i(t) : t = 0, 1, \dots\} & \text{the apartment occupancy verification process} \\
Z_i = \{Z_i(t) : t = 0, 1, \dots\} & \text{the dwelling occupancy verification process} \\
Q_i = \{Q_i(t) : t = 0, 1, \dots\} & \text{the tolerance process}
\end{array}
$$

Each process is defined on a subset of the outcome statuses for $X_i(t)$ and is initialized as unresolved at $t = 0$. No assumptions are made about the independence of the processes for a given unit or about the independence of the units for any given process at this point.

The collection status $X_i(t)$ is the result of a deterministic set of rules on the above processes which is equivalent to saying that there exists some function γ such that

$$
\begin{aligned}
&\Pr\left(X_i(t) = \gamma(x, r, n, c, y, z, q) \middle| \right. \\
&\quad \left. X_i(t-1) = x, R_i(t) = r, N_i(t) = n, C_i(t) = c, Y_i(t) = y, Z_i(t) = z, Q_i(t) = q\right) = 1
\end{aligned}
$$

That is, the random process $X_i(t)$ is conditionally constant. The dwelling frame unit resolution process described in this way shares similarities with a survival process with competing causes of resolution from the various survival subprocesses it depends on.

Even though the microsimulation model is defined at the dwelling frame unit level, the forecast statistics of interest are mostly geospatial or spatiotemporal aggregates. To express the desired statistics, we define the collection of sets $\{\Lambda^j : j \in (\text{RO}, \text{LCO}, \text{CLD}, \text{CU})\}$ where each collection of sets Λ^j defines a partition of Canada $\Lambda^j = \{\Lambda^j_k : k \in (1, 2, \dots, |\Lambda^j|)\}$. That is, Λ^{CU} is the set of the 45,133 CU sets Λ^{CU}_k each containing all units in the k^{th} CU.

The forecast statistic of interest for collection progress is the tolerance rate. A simplified version of the tolerance rate for a set U is $TOL(U) = \frac{\sum_{i \in U} \mathbb{1}_{\{X_i(t_{end})=1\}}}{\sum_{i \in U} \mathbb{1}_{\{X_i(t_{end}) \in (0,1,4,5)\}}}$ where $|\cdot|$ denotes the cardinality of a set. The tolerance rate is defined over the subset of private occupied dwellings. It is slightly different from response rates in the treatment of the unresolved cases and in the use of temporary information left to be confirmed that is not explicitly defined in the simplified version of the tolerance rate used in this paper. The objective is for the microsimulation model to produce reasonable forecasts for $\{TOL(\Lambda^j) : j \in (\text{RO}, \text{LCO})\}$ but the statistic is defined for all geography levels under the model design and was extensively used at the CLD level during collection.

3.1 The self-response process

The self-response process $R_i(t)$ is defined over the subset of collection statuses $\{0, 1, 2, 3\}$. It can be integrated to the global process by adding a condition on $X_i(t)$ to deal with censoring of the self-response process resulting from competing processes. In the context of a model for the time to self-response, resolution outcomes from other processes may be interpreted as simply preventing us from observing the self-response process from that day forward, effectively censoring the self-response process. This type of interpretation of competing causes is extensively used in survival models theory and practice. Among the collection statuses, the tolerance met status that is generated by the tolerance process to be described in Section 3.5 censors interviewer led activities but not self-response. As such,

the distribution of $R_i(t)$ can be defined by the transition matrices

$$\Pr(R_i(t)|R_i(t-1), X_i(t-1) \in \{0,5\}) = \begin{bmatrix} p^R_{i,0,0}(t) & p^R_{i,0,1}(t) & p^R_{i,0,2}(t) & p^R_{i,0,3}(t) \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\Pr(R_i(t)|R_i(t-1), X_i(t-1) \notin \{0,5\}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where a matrix element in position (j,k) is the probability $p^R_{i,j,k}(t)$ that a dwelling frame unit i with a self-response status j at time $t-1$ has a self-response status k at time t. The diagonal elements with probability 1 simply mean that a resolved dwelling with outcome j will not go unresolved nor change resolution outcome afterwards. In practice, the condition $X_i(t-1) \in \{0,5\} \implies R_i(t-1) = 0$ but the diagonal block is added for completeness as the derivation rules expressed by the γ function were not explicitly defined.

If we define the random variable for the time to self-resolution of unit i by T_i, and note that it corresponds to the first time that the process $R_i(t) \neq 0$, then it follows that $T_i \sim \min_v(R_i(v) \neq 0)$ and that the survival process of the time to self-resolution taken over discrete intervals of time can be described as a function of the $p^R_{i,0,0}(t)$.

$$\begin{aligned} S_{R_i}(t) &= \Pr(T_i \geq t) \\ &= \Pr(\min_v(R_i(v) \neq 0) \geq t) \\ &= \Pr(R_i(v) = 0 \; \forall \; v \in \{0, 1, \ldots, t-1\}) \\ &= \textstyle\prod_{v=1}^{t-1} \Pr(X_i(v) = 0 | X_i(v-1) = 0) \\ &= \textstyle\prod_{v=1}^{t-1} p^R_{i,0,0}(v) \end{aligned}$$

Thus, a conditional process where dwellings first decide if they will respond at time t and conditionally establish their resolution outcome provides a general method to estimate the various probabilities. A possible model to estimate the time to self-resolution which is appropriate for the wave methodology time varying effects is an extension of the Cox proportional hazards model with time varying covariates and coefficients $\log(h(t)) = \alpha(t) + \sum_{k=1}^{K} \beta_k(t) x_k(t)$ which can be estimated with the complementary log-log model $\log(-\log(1 - p^R_{i,0,0}(t))) = \alpha_t + \sum_{k=1}^{K} \beta_k x_{i,k}(t)$ after making appropriate transforms of the time varying coefficients into the corresponding covariates. The complementary log-log model follows from time interval grouping of a continuous survival process (Agresti 2002, Allison 2010). The conditional distribution of resolution outcomes is a simple multinomial distribution.

An important challenge is that survival model estimation methods assume that survival is homogeneous given the variables accounted for in the model. In our context, several covariates that might be of interest in a parametric or semi-parametric survival model are unknown until the unit is resolved. One such important covariate is the dwelling occupancy status which is especially problematic for the conditional model rationale discussed previously. Since dwelling frame covariates are not available for the list/leave areas, the only covariates readily available for all dwellings for the forecasting model are wave methodology specific event times and those are essentially identical for all dwellings with the same collection method. As such, the chosen estimation method was the non-parametric Kaplan-Meier estimator of the survivor function (Hosmer and Lemeshow 1999, Allison 2010). The conditional multinomial distribution $R_i(t)|R_i(t-1) \neq 0, R_i(t-1) = 0, X_i(t-1) \in \{0,5\} \sim M_t(p^*_{1,t}, p^*_{2,t}, p^*_{3,t})$ was estimated for each time t by the maximum likelihood

estimator $\hat{p}^*_{j,t} = \frac{\text{\# units at risk for self-response at t-1 with outcome } j \text{ at t}}{\text{\# units at risk for self-response at t-1 resolved at t}}$. Certain assumptions may be required if the length of future survey collection cycles are longer than that of the cycle used for model estimation. Separate estimates of the survival and multinomial models were obtained for mail-out areas and list/leave areas and parameters for unit i were those of their collection method estimate. We note that the required forecasting model parameters for the self-resolution process $R_i(t)|R_i(t-1), X_i(t-1)$ are $\hat{p}^R_{i,0,0}(t) = \frac{\hat{S_{R_i}}(t+1)}{\hat{S_{R_i}}(t)}$ and $\hat{p}^R_{i,0,j}(t) = \hat{p}^*_{j,t}(1 - \hat{p}^R_{i,0,0}(t))$.

If the collection methodology changes between two cycles, a parametric complementary log-log model restricted to covariates available on the collection frame such as those defining the collection strategy provides a general framework to incorporate methodology changes in the forecasting model. Additionally, a more complete parametric model may be used if the model framework is used for planning only and not for livecasting. For planning, the conditional distribution of the dwelling characteristics are not needed and past cycle data or some augmented version of it may be used as a simulation frame.

The chosen model clearly violates the homogeneity assumption of the survival self-resolution process. In Section 4, we propose a Bayesian approach using an occupancy propensity model to mitigate the issue.

3.2 The field non-response follow-up process

The field non-response follow-up process (NRFU) $N_i(t)$ is similar to $X_i(t)$ in that it is better understood as the combined process of the time when attempts are made and the outcome of said attempts. Thus, we define the field attempt outcome process $F_i = \{F_i(a) : a = 0, 1, \dots\}$ over the subset of collection statuses $F_i(a) \in \{0, 1, 2, 3, 4\}$ and the attempt counting process $\{A^F_i(t) : t > 0\}$ where $A^F_i(t)$ is the cumulative number of field non-response follow-up attempts made on dwelling i at time t and set $N_i(t) = F_i(a)$ where $A^F_i(t) = a$ for the derivation of $X_i(t)$. The distribution of $F_i(a)|F_i(a-1)$ can be described by the transition matrix

$$\Pr(F_i(a)|F_i(a-1)) = \begin{bmatrix} p^F_{i,0,0}(a) & p^F_{i,0,1}(a) & p^F_{i,0,2}(a) & p^F_{i,0,3}(a) & p^F_{i,0,4}(a) \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where a matrix element in position (j, k) is the probability $p^F_{i,j,k}(a)$ that a dwelling frame unit i with a field attempt outcome status j after attempt $a-1$ has a field attempt outcome status k after attempt a. The binary NRFU resolution process described by $p^F_{i,0,0}(t)$ and $(1 - p^F_{i,0,0}(t))$ is also a survival process. Unlike the self-resolution survival process, it is a true discrete-time process. Thus, a parametric estimation of the corresponding model with appropriate transforms of the time varying coefficients could be obtained using the logit model $\log(\frac{p^F_{i,0,0}(t)}{1-p^F_{i,0,0}(t)}) = \alpha_t + \sum_{k=1}^K \beta_k x_{i,k}(t)$ instead of the complementary log-log model (Agresti 2002, Allison 2010).

One may ask whether the probability of attempt outcomes on the same dwelling really changes as a function of the attempt number. It is likely that the aggregate behavior follows from the change in distribution of the unobserved heterogeneity among units with constant hazard. That is, a collection of heterogeneous units consisting of a mixture of easy and hard cases, each with constant probabilities across attempts would result in a process with probabilities varying across attempts. However, even if the true model is constant hazard

across attempts given some covariates, is does not mean that a multinomial logistic model for attempt outcome that does not vary by the attempt number is more appropriate for our purpose. In particular, such a model would be much less appropriate as a tool to evaluate the impact of different staffing scenarios. Thus, by repeating the estimation method described for the conditional self-response model using the logit model instead of the complementary log-log model the estimated stochastic process parameters do vary across attempt numbers due to the unobserved heterogeneity but are more suitable than attempt invariant parameters for our purpose.

In order to mitigate the problem of unobserved heterogeneity for live forecast, the following method was adopted such that the distribution used to simulate the residual collection period for units in an area would more likely have been generated from a cluster with a similar mixture of the unobserved covariates:

1. Produce empirical estimates of $F^k(a)|F^k(a-1), a \leq a_k$ for each Λ_k^{CLD} crew leader district using past cycle data and generalize the process for $a > a_k$ assuming that $F^k(a)|F^k(a-1) \sim F^k(a_k)|F^k(a_k-1)\ \forall\ a > a_k$. The parameter a_k is chosen such that stable estimates can be produced.

2. Assign each of the $|\Lambda_k^{CLD}|$ distributions to C clusters. The clusters were obtained using Ward's method for hierarchical clustering and the Hellinger distance measure

$$H(G^j, G^k) = \sqrt{\sum_{i=1}^{\infty} \frac{\left(\sqrt{g_a^j} - \sqrt{g_a^k}\right)^2}{2}},$$

   where $G^k \sim \min_a(F^k(a) \neq 0)$ is the probability distribution of the number of attempts to obtain a resolution outcome and $g_a^k = \Pr(G^k = a)$.

3. Produce empirical estimates of $F^c(a)|F^c(a-1), a \leq a_c$ for each cluster and an additional $c = C + 1$ cluster using Canada-wide data and generalize for $a > a_c$ as in step 1 above.

4. Assign an initial cluster from the $C + 1$ clusters to each of the $|\Lambda_k^{CLD}|$ CLDs using past cycle attempt outcome data mapped onto the current cycle geography using a maximum likelihood classifier. Assign areas made almost exclusively of dwelling growth between the two cycles to the Canada wide cluster $c = C + 1$.

5. At the beginning of the microsimulation for a live forecast at t_{now}, reclassify each area based on the attempt outcome data observed to date using the maximum likelihood classifier under certain regulatory conditions.

Field collection operations procedures are such that crew leaders are responsible to provide each interviewer they supervise with a list of unresolved cases for them to work on. There are soft rules in the system aiming for attempts to be captured and lists to be renewed on a daily basis while accommodating collection in remote areas with poor Internet access. The list making and the implicit ordering that will manifest through the ordering of the attempts are manual processes. That is, no system exists to optimize the creation or ordering of the assignment list. If such a tool existed, it would be possible to integrate it in our microsimulation framework.

The list assignments also lock cases in the centralized collection system such that cases that are on an active field interviewer list assignment are rendered unavailable for telephone non-response follow-up. It was possible to account for this dependency because

of the model framework as detailed later in the telephone non-response follow-up attempt counting process.

The stochastic analog of the list and attempt making processes used in the microsimulation consists of two sequential simple random samples. The list-making sample was drawn from the subset of unresolved cases ($X_i(t) = 0$) within the crew leader district and the attempt allocating sample was drawn from the list. Censoring of the NRFU process from other processes is handled by the list-making mechanism and the attempt counting process. That is, $\Pr(A_i(t) > a | A_i(t-1) = a, X_i(t-1) \neq 0) = 0$ due to the sampling condition which implies that $\Pr(N_i(t) = n | N_i(t-1) = n, X_i(t-1) \neq 0) = 1$. This sampling process naturally introduces dependence between units but it is assumed that the attempt-outcome process $F_i(a)$ remains independent between dwelling frame units. To address the situation where more than one attempt per case was made on any given day, it was assumed that no more than two attempts per case per day would be done and that all cases would first be attempted once and the residual number of attempts would be sampled from units not resolved by the first attempt to remain consistent. While this may strike as challenging to implement in the discrete time framework, it can be easily implemented by observing that the attempt-outcome process can be fully simulated upon initialization leaving only the attempt counting process to be simulated daily basis. We do not claim that the sequential SRS results in an appropriate model for what is being done in the field but the distribution of the number of attempts produced by this method was sufficiently similar to the distribution observed in past cycles for our purpose.

Before we discuss how list and attempt sample sizes are established, we briefly describe the field staffing model. The model considers a staffing capacity and the effective staff. They are both defined as a number of interviewer hours for a given day and CLD. This capacity can be fully specified as a model input or derived using recent pay claim data and a simple model that is beyond the scope of this paper. An underlying assumption of the aggregation is that interviewer productivity is homogeneous within a CLD but some variation is implicitly accounted for in the attempt-outcome process estimation method with the rationale that interviewer characteristics are unobserved covariates of the dwellings they attempted.

The strength of the microsimulation framework stems from the effective staffing component, which establishes the staff working for a given day and CLD as a function of the capacity and the maximum available workload. The maximum available workload for a CLD k at time t is defined as $\sum_{i \in \Lambda_k^{CLD}} \mathbb{1}_{\{X_i(t)=0\}} \cdot TPA_k(t) \cdot \min(2, G_i - A_i^F(t))$ where $TPA_k(t)$ is an average time per attempt parameter for CLD k estimated for different intervals of time from past cycle data and G_i is the attempt number of the first resolution for unit i obtained from the principle that the attempt-outcome process can be fully simulated upon initialization.The effective staff is thus defined as $\min(\text{max workload}, \text{staff capacity})$. The total size of the lists follows a similar rationale also using average list sizes from past cycle data. Stochastic improvements to the use of such simple averages as models of case density or through route optimization instead of using an estimate for different intervals of time is left for future work.

Finally, the key forecast statistic of interest for the field operation costs for an appropriately defined set U is:

$$FC(U) = \sum_{(k:\Lambda_k^{CLD} \subset U)} \sum_{t=t_{now}}^{t_{end}} \sum_{l=1}^{2} (H(t,k,l) R_H(k,l) + K(t,k) R_K(k))$$

where $H(t,k,l)$ is the number of hours claimed at time t in CLD k by interviewers ($l = 1$) or supervisors ($l = 2$), $R_H(\cdot,\cdot)$ the hourly rate, $K(\cdot,\cdot)$ the kilometers claimed and $R_K(\cdot)$

the kilometer rate.

3.3 The telephone non-response follow-up process

The telephone non-response follow-up operations were conducted from 25 Collection Support Units (CSU) located in each of the 25 Local Census Offices (LCO). The telephone non-response follow-up process $C_i(t)$ is similar to the field non-response follow-up process $N_i(t)$ in that it is broken down into a telephone attempt outcome process and an attempt counting process. We define the telephone attempt outcome process $D_i = \{D_i(a) : a = 0, 1, \dots\}$ over the subset of collection statuses $D_i(a) \in \{0, 1, 2, 3, 4\}$ and the attempt counting process $\{A_i^D(t) : t > 0\}$ where $A_i^D(t)$ is the cumulative number of telephone non-response follow-up attempts made on dwelling i at time t. The stochastic process $D_i(a)$ is described by the transition matrix for $D_i(a)|D_i(a-1)$

$$\Pr(D_i(a)|D_i(a-1)) = \begin{bmatrix} p^D_{i,0,0}(a) & p^D_{i,0,1}(a) & p^D_{i,0,2}(a) & p^D_{i,0,3}(a) & p^D_{i,0,4}(a) \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where a matrix element in position (j, k) is the probability $p^D_{i,j,k}(a)$ that a dwelling frame unit i with a telephone attempt outcome status j after attempt $a-1$ has a telephone attempt outcome status k after attempt a. The telephone attempt-outcome process parameters can be estimated following the same methods as those described for the field attempt-outcome process. The quality of the telephone attempt outcome paradata that was available was of worse quality than that of the field attempt outcome. As a result, a single Canada-wide estimation was done with various assumptions about the data leaving no place for the clustering and classification methods used for the field non-response follow-up process. The field attempt-counting process is assumed to be non-informative of the telephone attempt-outcome process and vice versa. That is, we assume that $D_i(a+1)|\left[D(a), A_i^D(t) = a, A_i^F(t) = b\right] \sim D_i(a+1)|\left[D(a), A_i^D(t) = a\right]$. This is a strong assumption for our estimation methods but the quality of the telephone attempt data for the 2011 Census offered no better alternative.

The key differences with the field process is that the CSU attempt-counting process $A_i^D(t)$ can make up to 3 attempts per day, does not have a list making step, the attempt sample is drawn nationally from the subset of unresolved dwelling that were not sampled on a field interviewer list, have a valid telephone number and are flagged as eligible for telephone follow-up. The dwelling level telephone eligibility flag could be updated from head office through an overnight process and was used as a responsive design tool to focus telephone follow-up in areas that needed it the most. This overnight process was also implemented in the microsimulation model.

Since the telephone non-response follow-up was conducted from LCOs, the staffing model for telephone follow-up was aggregated in terms of interviewer hours at the LCO level, everything else discussed for NRFU staffing holds in this case as well.

3.4 The occupancy verification processes

The Dwelling Occupancy Verification (DOV) and the Apartment Occupancy Verification (AOV) processes are activities conducted on subsamples of dwellings and large apartment buildings respectively. The samples are drawn exclusively in the mail-out collection areas

where we have a good quality survey frame and are based on an occupancy propensity model. The DOV is a field operation resembling field NRFU excepted that a single attempt is made for any given case and slightly different procedures are used for these attempts. The AOV is a telephone operation conducted by the Collection Support Units that is very similar to the telephone non-response follow-up. The key difference is that an apartment building manager is contacted to obtain the occupancy status of each apartment unit. Their objective is to improve the efficiency in resolving unoccupied and canceled dwelling frame units by making attempts closer to the Census Day and using procedures that are more appropriate for unoccupied and canceled dwellings.

Their modeling and estimation are analogous to those for the field and telephone non-response follow-up respectively, therefore, we refer the reader to Subsections 3.2 and 3.3 for details.

3.5 The tolerance process

The tolerance process is an overnight scheduled process that calculates tolerance rates and a number of quality indicators, and compares them with established thresholds. A plan of daily thresholds for the entire collection period is established before collection but may be subject to ad hoc changes. The rates and indicators are estimated at the collection unit level (Λ_i^{CU}) and the objective of this process is to halt costly field and telephone operations in CUs with high tolerance rates to ensure resources are spent elsewhere in an attempt to achieve more homogeneous response rates across Canada. There are three possible outcomes from the tolerance process:

1. If tolerance thresholds and quality conditions are met, the collection unit area now meets tolerance and all unresolved collection statuses become tolerance met statuses indicating that field and telephone activities are halted in the area.

2. If tolerance thresholds or quality criteria are not met but a relaxed set of conditions are met, the collection unit is sent for a manual review.

3. Otherwise, collection operations continue in the area.

Of course, introducing the manual process in a microsimulation model is unreasonable. Different options were considered, but the effect of treating relaxed criteria as not meeting them was marginal given that the manual process took one or two days to address and the regular criteria were often met by then. The key purpose of this manual process is to ensure that appropriate quality control is performed by supervisors in areas that cannot meet quality conditions after they reached a high tolerance rate. For example, if a large proportion of the dwelling frame units in an area had a collection status of other resolved, certain quality conditions possibly could not be met even after resolving every single case. The quality review could result either in some erroneous outcome codes to be identified and corrected, generally in the form of reseting a case status to unresolved to return it to collection, or, if the observed data was confirmed as correct, in meeting tolerance through a manual procedure despite failing the quality conditions. This quality control process was not modeled but it was partially taken into account in the estimation of the various attempt-outcome processes by treating the first resolution outcome for cases that were reset from quality control procedures as an unresolved outcome.

The tolerance process $Q_i = \{Q_i(t) : t = 0, 1, \dots\}$ defined on the subset of collection statuses $\{0, 5\}$ is deterministic when conditioned on all other observed processes for all units within the same collection unit. That is, the conditional distribution

$$Q_i(t) | \left[Q_i(t-1) = 0, \{X_i(t-1), R_i(t), N_i(t), C_i(t), Y_i(t), Z_i(t) : i \in \Lambda_k^{CU}\} \right]$$

is constant.

4. Towards a Bayesian Occupancy model

As discussed several times in section 3, the conditional model for the type of resolution outcome likely has more to do with unobserved heterogeneity than with the randomness of the attempt outcome and therefore constitutes a violation of the survival model assumptions. A model design feature that did not make it into the 2016 Census model but would go a long way towards improving this issue is the use of a dwelling occupancy propensity model to simulate dwelling occupancy. The occupancy propensity is already modeled for the occupancy verification processes and can improve the forecasting model if treated as a probability. This improvement is scheduled to be added to the existing model for the next census cycle.

The notation developed so far is not adequate to provide an explicit formula for the posterior probability at time t_{now} that a dwelling is occupied given the observed processes. We only note that the joint distribution of the attempt counting processes, when appropriately conditioned on the dwelling being unresolved for each attempt is independent of the dwelling occupancy status.

This Bayesian posterior probability that a dwelling is occupied given the observed processes would be used to simulate dwelling occupancy upon initialization of a simulation such that the day-by-day stochastic processes can be simulated according to their respective conditional estimates.

5. Results and Model Utility

Due to the nature of the model outputs, we cannot share the key results that were the end-of-collection tolerance rates and costs compared with their true realization. We instead provide a qualitative assessment of the utility of the model in the collection of the 2016 Canadian Census. Also, we argue that evaluating forecast results in relation to the true realization of the process is inappropriate without context as the model was also used as a responsive design tool in such a way that it worked against its own assumptions. That is, as a forecasting tool, it was not designed to account for responsive design micromanagement decisions that were taken after study of the simulation results.

The model was scheduled to enter production around the 6th week of collection, shortly after field operations had started. The very first cost forecast provided a quantitative figure for the surplus that resulted from observed self-response rates that exceeded our original planning. This figure impacted the general strategy for the remainder of collection.

Beyond its planned purpose, the model was used to estimate staffing deficit by Crew Leader District by incrementally increasing the initial staffing levels drawn from pay claims data until tolerance objectives were met for each CLD. The difference in effective staffing between the initial model and the final iteration model were used as an estimate of the deficit. A list of areas with the largest modeled staffing deficit was produced and regional managers were required to provide corrective plan to census management. In past census cycles, such corrective plans would have been performed much later in the collection cycle because traditional progress reports do not account for staffing levels at lower geographic levels. As a result, it was possible to use cost effective solutions such as hiring or moving staff on shorter distances to improve response rate homogeneity.

Anecdotally, the model was used to forecast the date at which collection support units would begin to run out of telephone follow-up cases in the system. It was brought up only to highlight the flexibility of the model design in predicting timeliness but it turned out to

be only two days off from the observed situation with the difference explainable from an inappropriate use of the system.

Ultimately, the performance of the model lies more in how it improved the central office's ability to support regional offices in managing collection in smaller areas. The model output was used as evidence of problems to be addressed and became an integral part of the dialog between head office and regional offices. It played such a key role in the 2016 Census collection operations that additional outputs were designed and produced to support regional offices such as heat maps of the end-of-collection tolerance rates predictions. The regional offices even requested that comprehensive model output be shared with them directly such that they could act beyond the list of areas they were required to act upon as resources allowed.

6. Concluding Remarks

The forecasting model that was developed for the 2016 Census fills an important gap in Statistics Canada's collection operations management toolbox. It provides a link between resources and responses at a granular level. It also provides a tool to estimate staffing deficit for relatively small areas early in collection operations.

The model discussed was an initial take on a general method to forecast costs and response rates with a number of strong assumptions, several of which will see improvements in future versions. The framework developed is sufficiently general that it could be used for any lengthy collection operation and for most combination of collection activities. It may even be sufficiently general to eventually be extended to the collection operations of multiple concurrent surveys.

This work opened up a wide range of applied research projects in time to event variants of traditional logistic regression models. Parametric time to event models could not only improve the forecasting model itself but also our ability to evaluate the impacts of potential changes to the wave methodology. The model is scheduled to replace a number of separate models produced over the 5 year cycle and will improve coherence of the different planning models in doing so.

7. Future Work

Future work includes the Bayesian framework for the dwelling occupancy discussed in Section 4. It also includes parametric estimation of the self-response or self-resolution model. We also plan to improve various parameters estimates at smaller geography levels and explore the feasibility of modeling staff mobility if not in the model itself, at least in the form of output processing from the current framework using staffing capacity surpluses and deficits. Finally, a lot of program optimization will be done to increase the number of iterations that can be performed within a short time frame for 2021 production to reduce the variability due ot the simulation itself for small area forecast statistics.

8. Acknowledgements

The author would like to thank their colleagues Cilanne Boulet, David Dolson, Patrice Mathieu, Jean-Pierre Morin, Sander Post and Marie-Hélène Toupin for their valuable feedback in reviewing this paper.

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis, Second Edition.* John Wiley & Sons, Inc.

Allison, P. D. (2010). *Survival Analysis Using SAS: A Practical Guide, Second Edition.* Cary, NC: SAS Institue Inc.

Chen, B. C. (2008). "Simulation Modeling of Field Operations in Surveys," In *JSM Proceedings*, Survey Research Methods Section. Denver, CO: American Statistical Association. 2272–2288.

Couture, K., Bélanger, Y., and Neusy, E. (2010). "Modelling and Simulation of Survey Collection Using Paradata." In *JSM Proceedings*, Government Statistic Section. Vancouver, BC: American Statistical Association. 3923–3933.

Doherty, S., Whitehead, J. & Cheng, R. (2008). "A Simulation Model for 2011 UK Census Field Operations." SimTecT 2008 Simulation Conference: Simulation Maximising Organisational Benefits. Melbourne Australia.

Hosmer, D. W. and Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data.* John Wiley & Sons, Inc.

Karr, A. F., Cox, L. H. and Kinney, S. K. (2012). "The World's Simplest Survey Microsimulator (WSSM)." *Proceedings of the 2012 Federal Committee on Statistical Methodology Research Conference*. Washington, DC.

Kreuter, F., Couper, M. and Lyberg, L. (2010). "The Use of Paradata to Monitor and Manage Survey Data Collection" In *JSM Proceedings*, Survey Research Methods Section. Vancouver, BC: American Statistical Association. 282–296.

Mathieu, P. (2017). "The 2016 Canadian Census: An Innovative Wave Collection Methodology to Maximize Self-Response and Internet Response." European Survey Research Association Conference, Lisbon, Proceedings to be published

Piegorsch, W. W. (1992). "Complementary Log Regression for Generalized Linear Models." *The American Statistician*, Vol. 46, No. 2, 94–99.

Statistics Canada (2015). *Demosim: An Overview of Methods and Data Sources.* Statistics Canada Catalogue no. 91-21-x, Ottawa.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model.* Springer.