An Approach to Predict Final Yield Among Interim Cases

Rui Jiao[1], Andrea Piesse[1]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract
There are surveys for which meeting the target sample size or yield, possibly also for subgroups, is very important. Survey statisticians need to predict the final number of completes among cases that have been released but do not yet have a final response status (interim cases). Some interim cases may have been in the field longer than others and the nature of interim statuses may vary. Methodologists have developed statistical models to estimate the likelihood of a case being finalized or completed on the next contact attempt. However, the problem of predicting how many interim cases will result in completed interviews over the course of data collection has received less attention because this requires combining predictions across all future contact attempts, where the prediction at each contact attempt is conditional on the previous contact attempts that have been made.

This paper examines response propensity based on number of contact attempts, and proposes an approach using survival analysis to predict the proportion of current interim cases that will respond by the end of data collection. Finally, results of implementing the approach on synthetic data are presented.

Key Words: adaptive design, interim cases, response status, sample design, survival model

1. Introduction and Background

Many surveys rely on probability-based sampling to select individuals. Although this design permits population inference, final yield is not known until data collection ends because repeated attempts to obtain a completed interview are made over time and individuals' response propensities are unknown. Often, survey researchers closely monitor the response results on a regular basis after the survey is launched. Cross-functional teams then work together to ensure the overall sample yield is on target. Currently, with the increasing interest in obtaining sample yield by subpopulations, and with the development of adaptive design, survey researchers are looking for ways to adjust sampling rates midway during data collection in an effort to meet targets for different subpopulations by the end. This adaptive design is often considered for surveys that cover a long period of data collection because first, the observation time during an initial phase of data collection needs to be at least long enough to inform key design decisions for later phases; second, the later phase needs to span sufficient time to react to the design changes. An important piece of information driving decision-making is knowing the estimated final yield at any time during data collection.

For this purpose, survey researchers need to predict the final yield among cases that either have not been worked (NWK cases) or have been released but do not yet have a final response status (interim cases). Some interim cases may have been in the field longer than others and the nature of interim statuses may vary. The growth of computer-assisted data collection methods has provided rich paradata, including records of contact attempts. Methodologists have developed statistical models utilizing paradata to estimate the probability that the next call on a sample case will produce an interview (Groves and Heeringa, 2006). Some other studies have also proved that elements of paradata such as the number of contact attempts and/or the number of effective contacts have important predictive effects on response (Matsuo et al., 2006). However, the problem of predicting how many interim cases and NWK cases will result in completed interviews over the course of data collection has received less attention.

In this paper, we consider the full contact histories among the cases for whom an interviewer has obtained a final status during a study, as well as the partial contact histories among interim cases, through a sequence of survival functions and hazard functions. Then, using the developed survival models, we estimate the response propensities at each future contact attempt for each sample case until the maximum number of contact attempts is reached. Finally, we combine the predictions across all future contact attempts, where the prediction at each contact attempt is conditional on the previous contact attempts that have been made.

The paper proceeds as follows: Section 2 introduces the method of calculating response propensities by first providing a brief overview of survival analysis techniques, then mapping the concepts and notations to the survey context. The subsequent sections use synthetic data to demonstrate the proposed approach. Section 3 describes the characteristics of the synthetic data. Section 4 discusses the detailed steps for developing the survival functions and the hazard functions. Section 5 presents the developed models and the results of applying these to predict yield. Finally, Section 6 discusses the conclusions and considerations of the study.

2. Methods

2.1 The Notion of Using Survival Analysis

In survival analysis, subjects are usually followed over a time period to study the time until the event of interest occurs. The survival function is the probability of observing a survival time greater than some stated value t, denoted $S(t)$ (Hosmer and Lemeshow, 2008). The hazard function estimates the instantaneous rate of the event occurring at time t, denoted $h(t)$, conditional on the subject's survival up to time t. If there are cases for which the event has not occurred during the observation period, their observations are considered incomplete—a situation referred to as right censoring. Survival analysis models make use of both complete observations and incomplete observations or censored data.

The features of survival analysis that are especially beneficial to this study are first, it studies the influence of time to event; second, it not only provides but also cumulates the outcome at each time through the end of the observation period; and third, it can make use of interim cases as a form of incomplete observations. In the context of survey data collection, time can be substituted with the number of contact attempts. Consider a data collection lifecycle for each case—it begins with the first contact attempt and continues until the case completes the interview or the maximum number of contact attempts has been reached. At each contact attempt, the survival analysis model can estimate the probability the case still needs to be worked by an interviewer and the conditional probability the case will yield a response given that a contact attempt is to be made. Then the product of these two probabilities produces the estimated response propensity for that case at that contact attempt. Once the response propensity is expressed as a function of number of contact attempts, the final yield for the interim cases and NWK cases can be estimated by the sum of the response propensities across all future contact attempts and across all cases.

2.2 Survival Function $\boldsymbol{S}(\boldsymbol{t})$

Let t be the number of contact attempts that have been made, and define the event as the case receives a final status. The observation period covers from the first day that data collection begins to the day the snapshot of the field results is created. The interim cases at the end of the observation period are considered censored data.

The survival function $S_i(t)$ describes the probability of a case surviving after t contact attempts (i.e., requiring further field work) beyond a beginning point i, where i denotes the sequence number of the next contact attempt to be made in a case's data collection lifecycle. Because the survival function is essentially a transformation of the cumulative distribution function, the estimated probability of surviving starts at 1 and decreases as t increases. $S_1(t)$ is modeled for each observation period. Then, to predict yield for the interim cases and NWK cases, the next contact attempt serves as the beginning point for the new survival function, which is derived using $S_1(t)$.

The Kaplan-Meier estimator, also known as the Product-limit estimator, is used for this study to estimate the survival function at contact attempt t, calculated as:

$$\hat{S}_1(t) = \prod_{k=1}^{k \leq t} \frac{n_k - d_k}{n_k}, \tag{1}$$

where n_k is the number of cases surviving at the beginning of each contact attempt, d_k is the number of cases finalized at each contact attempt, and k runs from the first contact attempt up to the t^{th} contact attempt. Each term in the product estimates the conditional probability of survival past contact attempt k ($k \leq t$), then the unconditional probability of survival past the t^{th} contact attempt is obtained by multiplying together these terms up to t. The survival functions $\hat{S}_i(t)$ where $i > 1$ are obtained through algebraic derivation from $\hat{S}_1(t)$.

2.3 Hazard Function $\boldsymbol{h}(\boldsymbol{t})$

The event for the hazard function is defined differently from that for the survival function, and is defined as the case responds to the interview, i.e., the final status is a complete. So the hazard function $h(t)$ estimates the conditional probability a case responds at the t^{th} contact attempt, given that the interviewer has made $t - 1$ contact attempts.

The empirical hazard rate at contact attempt t is calculated as:

$$\hat{h}(t) = \frac{d'_t}{n'_t}, \tag{2}$$

where n'_t is the number of cases subject to the t^{th} contact attempt and d'_t is the number of cases completing an interview at the t^{th} contact attempt.

2.4 Response Propensity

The response propensity for a NWK or interim case l at the next future contact attempt i is estimated by the hazard rate at i because we are sure such a case will survive to the next contact attempt, expressed as:

$$P_{case\ l,\, contact\ t=i} = \hat{h}(t = i) \tag{3}$$

The response propensity for a NWK or interim case l at another future contact attempt t $(t > i)$ is estimated by the probability the case survives after $t - 1$ contact attempts times the probability the case responds at contact attempt t, expressed as:

$$P_{case\ l,\, contact\ t>i} = \hat{S}_i(t - 1) \times \hat{h}(t) \tag{4}$$

Then the final yield is calculated by adding the sum of the estimated response propensities across all future contact attempts for NWK and interim cases to the number of final completes to date, expressed as:

$$\begin{aligned} Estimated\ Final\ Yield = \left(\sum_{l \in NWK,\, Interim} \sum_{t=i}^{Maximum} P_{case\ l,\, contact\ t}\right) \\ + NumComplete_{to\ date}, \end{aligned} \tag{5}$$

where $NumComplete_{to\ date}$ denotes the number of final completes among the snapshot data.

3. Synthetic Data

Synthetic data were created to demonstrate the survival analysis approach. The scenario assumes a survey fielded over a 1-year period with a total sample size of n=100,000, and an overall response rate of 40 percent. This study considers nine snapshots of the field results taken on day 30, 60, 90, 120, 150, 180, 240, 300, and on the last day of data collection, which is on day 365. Each snapshot contains the same number of sample cases with the following information recorded at each time:

- Current response status—this identifies if the case (1) has completed the interview, called final complete, (2) has been finalized as nonresponse, called final not complete, (3) has been worked in the field but has no final status yet, called interim, or (4) has not been worked yet, called NWK.
- Number of contact attempts—for final complete and final not complete cases, this is the number of contact attempts the interviewer made to finalize the case. For interim cases, this is the number of contact attempts made to date. For NWK cases, this variable is undefined.
- Last known interim result code—interviewers usually code the detailed reason why a case remains interim (e.g., appointment made, refusal, not located, language problem, break-off). For final cases, this records the result of the last contact attempt before the case was finalized. For interim cases, this reflects the result code obtained on the most recent contact attempt. For NWK cases and cases finalized at the first contact attempt, this variable is undefined. The last known interim result code is used to group interim cases and final cases that required at least two contact attempts when developing the survival and hazard functions for interim cases.

The data were created to mimic the data collection protocol of a real survey as much as possible: (1) cases are released in batches, (2) cases released or worked early in the field period receive more contact attempts on average, (3) cases that ultimately respond have fewer contact attempts on average. Figure 1 summarizes the characteristics of the synthetic data. Each bar presents the field results from the snapshot taken on that day. The top portion of each bar from day 30 to day 240 displays the number of NWK cases at that point in time, which decreases as the number of days in field increases. By day 300, all cases have been released and contacted at least once; and by day 365, the last day, all cases have been finalized as either final complete or final not complete, and no interim cases remain.

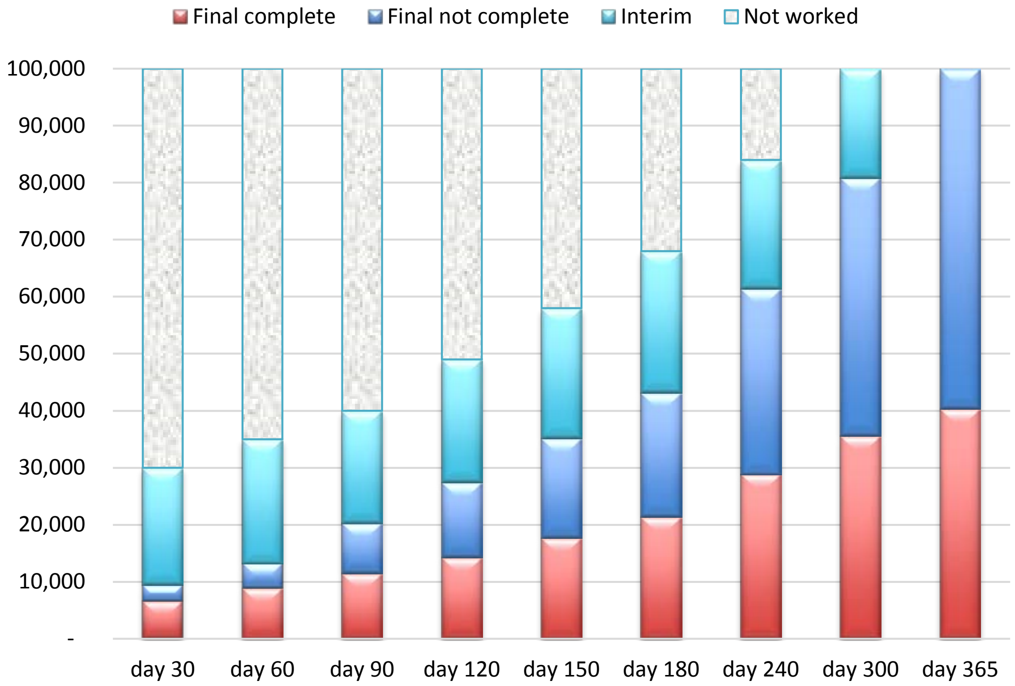

Figure 1: Synthetic data summary characteristics

In an adaptive survey design, changes are usually made during the first half of data collection so that enough time remains for the design changes to take effect. To serve that purpose, the procedure for predicting final yield in this study is conducted on a monthly basis in the first 180 days the survey has been in the field, so snapshots of field results on day 30, 60, 90, 120, 150, and 180 are used. The estimated final yield reveals whether the current sampling rates will lead to expected yields close to the desired targets, and therefore if the sampling rates for specific subpopulations need to be adjusted up or down.

4. Steps for Training the Model and Predicting Yield

4.1 Training the Model

Full contact histories are available for the final complete cases and the final not complete cases. The contact attempts made up to day *g* for the interim cases are considered the partial contact histories. Both types of contact history are used in developing survival functions and hazard functions.

The data from the snapshot for day g are at the case level. The number of contact attempts made (NumContacts) is the time dimension for this study, also denoted as t. The maximum number of contact attempts is assumed to be 15. In general, the maximum can vary from one survey to another, and setting this threshold will require knowledge of the data collection mode and field procedures.

Separate models are trained for the NWK cases and the interim cases because the available information for NWK cases is very limited, whereas the model for the interim cases can exploit information about the outcome of previous contact attempts.

To train the model for predicting final yield among the NWK cases, we use all finalized cases and interim cases to develop an overall survival function $S_1(t)$ with the beginning point being the first contact attempt, and an overall hazard function $h(t)$.

To train the model for prediction among the interim cases, the survival models are stratified by using the last known interim result code because to some extent this reveals the difficulty of obtaining a response (Groves and Couper, 1998; Matsuo et al., 2006). CaseGroup was created using these result codes to distinguish cases requiring different levels of effort, as follows:

- CaseGroup 1—Contact established (e.g., appointment, breakoff);
- CaseGroup 2—Contact not established (e.g., not located, located but no appointment yet); and
- CaseGroup 3—Known impediments (e.g., refusal, physical impediments).

Then for each CaseGroup j, a survival function with the beginning point being the first contact attempt $S_1^{CaseGroup\ j}(t)$ was constructed. The hazard functions were initially also stratified by CaseGroup but exploratory analysis showed that the hazard rates were not significantly different between CaseGroup 2 and CaseGroup 3 for these data. As a result, the two groups were combined when developing hazard functions, yielding two hazard functions $h^{CaseGroup\ j\prime}(t)$, where j' = CaseGroup 1 or CaseGroup 2&3. To predict final yield for interim cases, a series of survival functions $S_i^{CaseGroup\ j}(t)$ was algebraically derived from the survival function $S_1^{CaseGroup\ j}(t)$, where the beginning point i denotes the sequence number of the next contact attempt, which varies among the interim cases. The survival functions $S_i^{CaseGroup\ j}(t)$ and the hazard functions $h^{CaseGroup\ j'}(t)$ were applied to predict final yield among the interim cases.

4.2 Predicting Yield

To predict final yield among the NWK cases, the estimated propensity to respond by the end of data collection was calculated for each NWK case l using $\hat{S}_1(t)$ and $\hat{h}(t)$, as shown in equation (6).

$$\begin{aligned} P_{l,NWK} = {} & \hat{h}(t=1) + \hat{S}_1(t=1) \times \hat{h}(t=2) + \hat{S}_1(t=2) \times \hat{h}(t=3) \\ & + \cdots + \hat{S}_1(t=14) \times \hat{h}(t=15) \end{aligned} \tag{6}$$

For interim case l, the propensity to respond by the end of data collection was estimated using $\hat{S}_i^{CaseGroup\ j}(t)$ and $\hat{h}^{CaseGroup\ j'}(t)$, as shown in equation (7).

$$\begin{aligned} P_{l,Interim} = {} & \hat{h}^{CaseGroup\ j'}(t = i) \\ & + \hat{S}_i^{CaseGroup\ j}(t = i) \times \hat{h}^{CaseGroup\ j'}(t = i + 1) \\ & + \hat{S}_i^{CaseGroup\ j}(t = i + 1) \times \hat{h}^{CaseGroup\ j'}(t = i + 2) \\ & + \cdots + \hat{S}_i^{CaseGroup\ j}(t = 14) \times \hat{h}^{CaseGroup\ j'}(t = 15), \end{aligned} \tag{7}$$

where i denotes the sequence number of the next contact attempt, and case l is a member of level j and level $j'$of the three-level (survival function) CaseGroup and the two-level (hazard function) CaseGroup, respectively.

Finally, as shown in equation (8), the final yield was estimated by adding the sum of P_l across all NWK cases and interim cases to the number of final completes to date, expressed as:

$$Estimated\ Final\ Yield = \left(\sum\nolimits_{l \in NWK,\ Interim} P_l\right) + NumComplete_{to\ date} \tag{8}$$

5. Results

5.1 Survival Models

Survival functions and hazard functions were trained repeatedly using snapshots of the field results from day 30 to day 180. In the figures below, we intentionally present the estimated survival and hazard functions $\hat{S}_1^{CaseGroup\ j}(t)$ and $\hat{h}^{CaseGroup\ j'}(t)$ based on the snapshots corresponding to the lower bound and upper bound of the first half of the data collection period (day 30 and day 180) to show by example how the survival models can change as contact histories accumulate.

Figures 2 and 3 display the Kaplan-Meier estimate of survival probabilities as a function of NumContacts, stratified by the three-level CaseGroup, based on snapshots for day 30 and day 180, respectively. In both figures, the survival function with the lowest probabilities is for the contact-established cases. This reveals that on average the contact-established group received the least number of contact attempts and obtained final status faster than the other two groups. The survival functions based on day 30 for the contact-not-established cases and the known-impediments cases decrease at a much slower rate than those based on day 180. Such a pattern might be observed if, for example, interviewers are more inclined to finalize interim cases as the midpoint of data collection approaches (day 180) than at the beginning of the field period. This illustrates the importance of predicting final yield at various points in time, because data observed early on may not capture enough information to develop robust survival models and it may take some time for data collection procedures to stabilize.

Figure 4 displays the empirical estimate of the hazard rate as a function of NumContacts, stratified by the two-level CaseGroup, when the contact-not-established cases and the known-impediments cases are combined into a single group. The hazard functions based on day 30 are presented in solid lines and those based on day 180 are shown in dashed lines. Recall that the hazard rate for this study refers to the conditional probability to respond. The contact-established CaseGroup has consistently higher probability to respond at each contact attempt than the other CaseGroup. Within a CaseGroup stratum, the hazard functions based on day 30 and day 180 are very close, at least for the first ten contact attempts. The hazard rates for the contact-established cases after the tenth contact attempt are unstable due to small sample size.

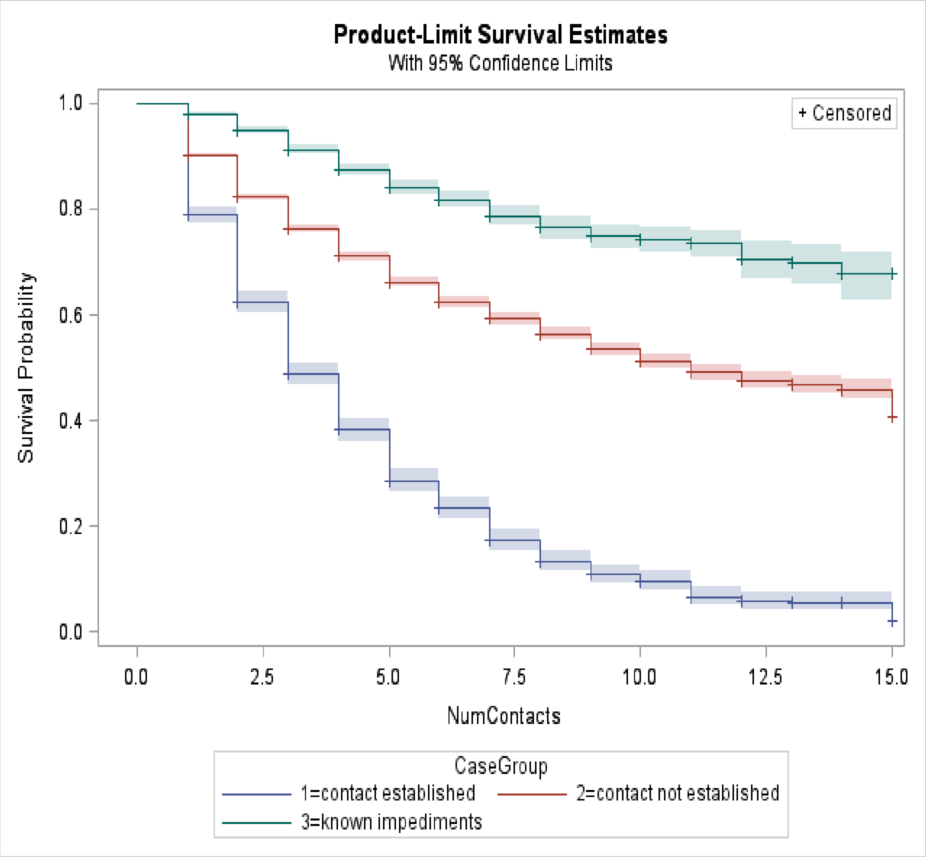

Figure 2: $\boldsymbol{S_1^{CaseGroup\,j}(t)}$ on day 30

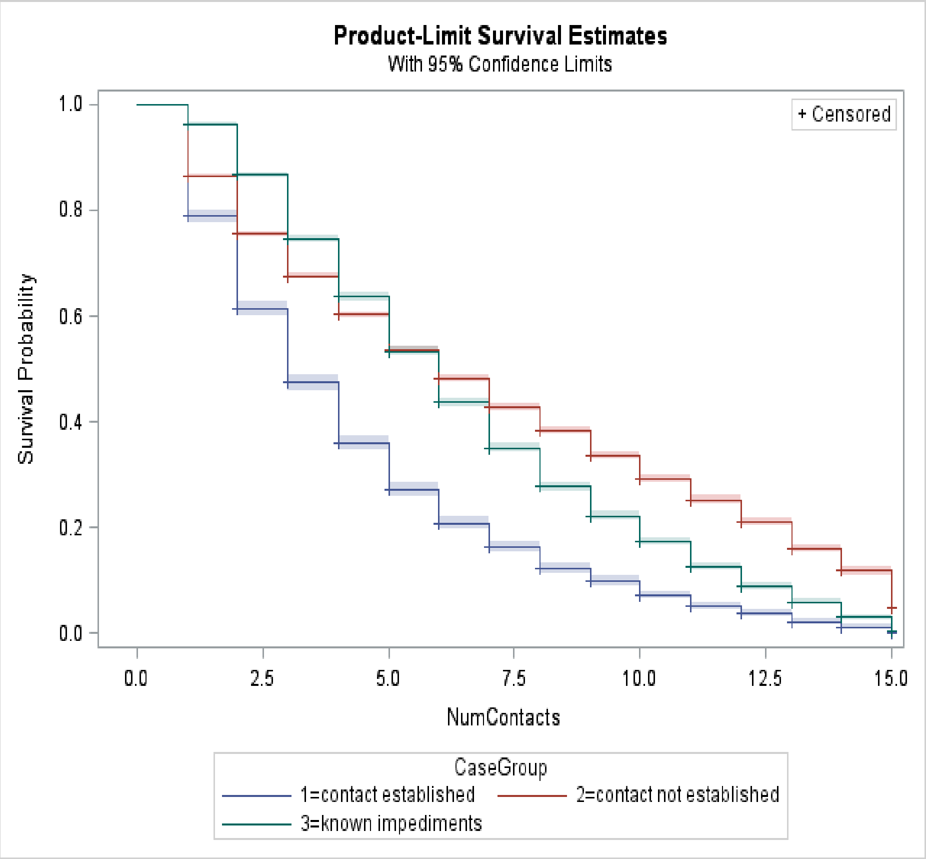

Figure 3: $S_1^{CaseGroup\ j}(t)$ on day 180

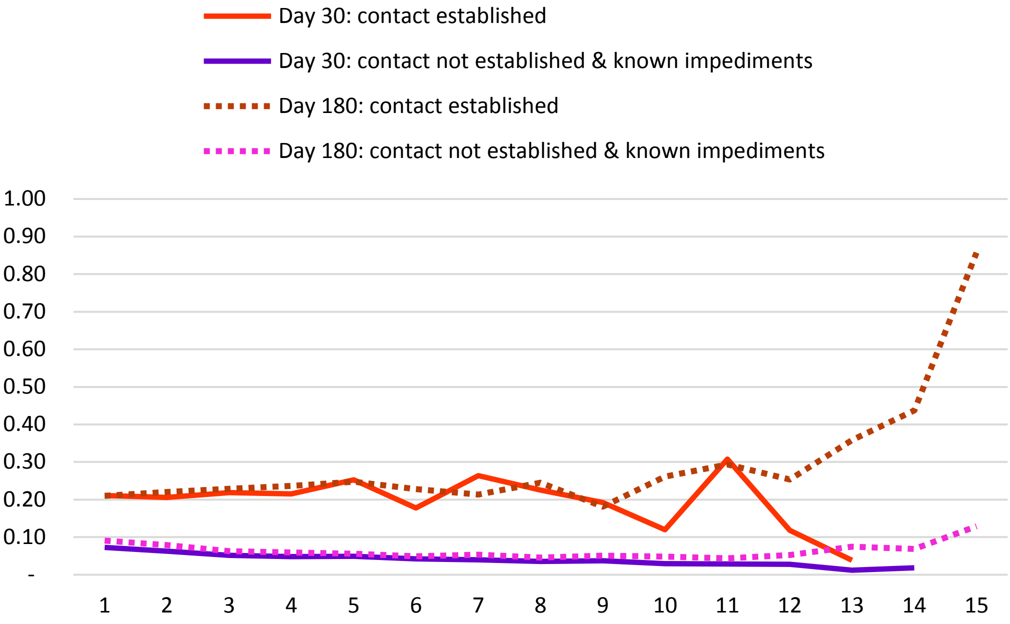

Figure 4: $h^{CaseGroup\ j'}(t)$ on day 30 and day 180

5.2 Predicted yield

Figure 5 displays the predicted number of final completes among NWK cases for each snapshot taken at a different number of days in field, shown in the shaded areas of the bars in the figure. The actual number of final completes among these cases is identified by overlaying a solid diamond shape. The yields predicted based on day 30, 60, and 90 are overestimated. In practice, this might be attributable to several factors: (1) not enough contact histories had been accumulated; (2) the field strategy was still not stable; and/or (3) the cases used to train the survival models were mainly early respondents. Predicting yield for the NWK cases is different from that for the interim cases because no history is available for these cases, and the results appear to be very sensitive to the snapshot data used to train the models. As more full contact histories are observed, by day 180, a little over 30,000 cases remain not worked and the predicted final yield among them is very close to the actual result.

Figure 6 displays the estimated final yield among interim cases. The predicted final yields by days in field presented in the shaded areas of the bars are consistently close to the actual results.

To complete this study, the predicted final yields among NWK cases and interim cases were summed with the number of final completes to date, to estimate the total final yield. Figure 7 presents the estimated number of final completes from each of these sources. Note that on the last day of data collection, all cases have been finalized. The last bar in Figure 7 consists of all final completes observed by the end of the survey. In comparing this to the earlier estimates, the total estimated final yield converges to the actual result starting around day 120.

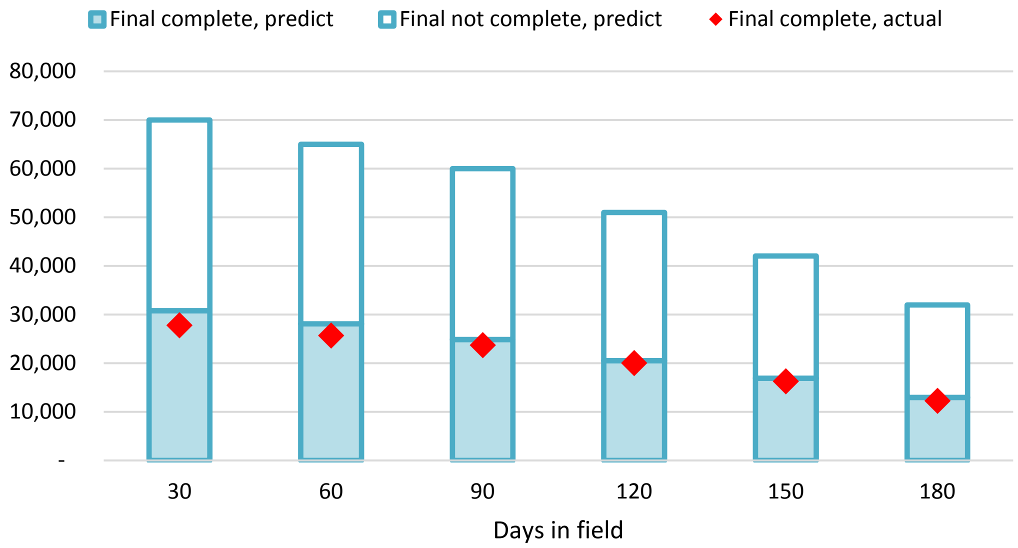

Figure 5: Estimated final yield among NWK cases

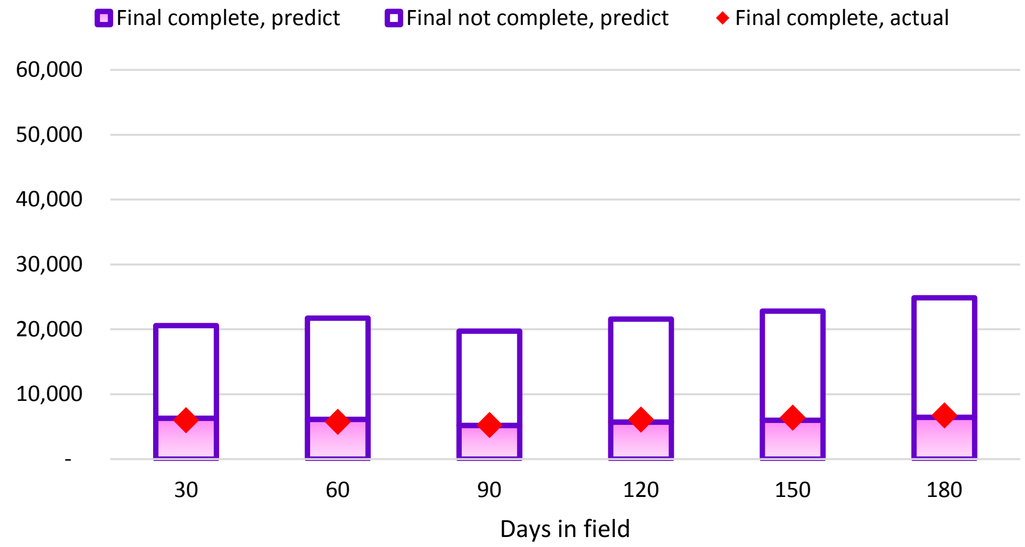

Figure 6: Estimated final yield among interim cases

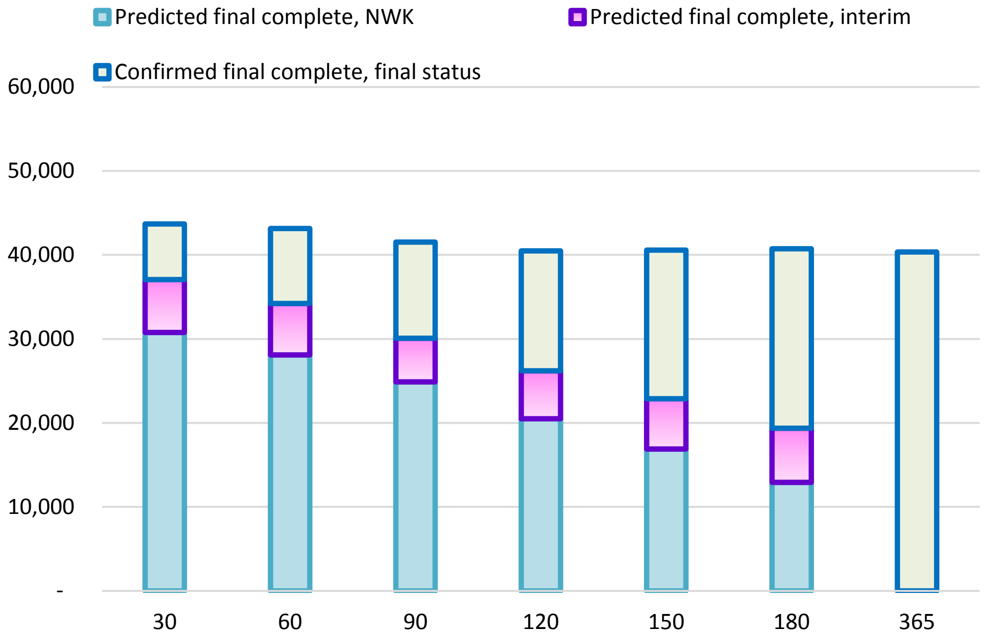

Figure 7: Total estimated final yield

6. Discussion

The proposed approach utilizes contact information that interviewers have observed to date, then predicts the survey's final yield while data collection is ongoing. Based on results using synthetic data (created to mimic an actual survey application) there are several conclusions to draw about this approach: (1) it improves as contact information accumulates; (2) it works better for interim cases than for not worked cases, especially at very early stages of data collection; and (3) it captures changes in level of interviewer effort that have occurred over time, assuming those changes are reflected in the snapshot data used to develop the survival models.

Some aspects require further consideration. The maximum number of contact attempts needs to be chosen carefully, based on the survey researcher's knowledge of the number of contact attempts an interviewer will make before finalizing a case as not complete. This maximum number will likely vary by mode of data collection. Use of different values will affect the results to some degree.

The stratification of cases when developing the survival models can be different for different surveys. Although the categorization of CaseGroup used in this study was supported by literature review on the topic of response propensity, a different choice might be more appropriate for some surveys. Note that stratification is also an attempt to make censored observations incomplete due to random factors within a stratum. As such, some explanatory analysis is needed at the beginning.

For interim cases, final yield is estimated based on the CaseGroup corresponding to the most recent contact attempt, even though these cases might receive different interim result codes at future contact attempts. This assumes that their likelihood to respond to the survey is adequately predicted by the survival models for the CaseGroup strata to which they belong based on the snapshot of field results.

Finally, although this approach captures the effects of changes in field operations that have already occurred, it assumes that the same data collection protocol reflected in the snapshot data applies to the remainder of the data collection period.

References

Groves, R.M. and Couper, M.P. 1998: Nonresponse in household interview surveys. New York: Wiley.

Groves, R.M. and Heeringa, S.G. 2006: Responsive design for household surveys: tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society, Series A, 169 (3), 439-457.

Hosmer, D. and Lemeshow, S. 2008: Applied survival analysis, second edition. New York: Wiley.

Matsuo, H., Loosveldt, G., and Billiet, J. 2006: The history of the contact procedure and survey cooperation – Applying demographic methods to European Social Survey contact forms round 2 in Belgium. Louvain-la-Neuve, Belgium. Paper presented at the Quetelet conference.