Sequence analysis of call record data: exploring the role of different cost settings

Mark Hanly

University of Bristol

Paul Clarke

University of Essex

Fiona Steele

London School of Economics and Political Science

Published online in the Journal of the Royal Statistical Society: Series A (Statistics in Society)
http://onlinelibrary.wiley.com/doi/10.1111/rssa.12143/abstract

Sequence analysis of call record data: exploring the role of different cost settings

Summary. Sequence analysis is widely used in lifecourse research and more recently has been applied by survey methodologists to summarise complex call record data. However, summary variables derived in this way have proved ineffective for post-survey adjustments, due to weak correlations with key survey variables. We reflect on the underlying optimal matching algorithm, and test the sensitivity of the output to input parameters or "costs", which must be specified by the analyst. The results illustrate the complex relationship between these costs and the output variables which summarise the call record data. Regardless of the choice of costs, there was a low correlation between the summary variables and the key survey variables, limiting the scope for bias reduction. The analysis is applied to call records from The Irish Longitudinal Study of Ageing, a nationally representative, face-to-face household survey.

Keywords: Sequence Analysis; Paradata; Call Record Data; Nonresponse Bias; Adjustment Weights

1. Introduction

Interviewer call records are now routinely collected as part of large-scale household surveys. Potentially, these data present us with a rich source of information for respondents and nonrespondents which can help address the problem of survey nonresponse. Sequence analysis has been explored as a means of summarising these complex records (Kreuter and Kohler, 2009; Pollien and Joye, 2011, 2014; Maslovskaya et al., 2014; Durrant et al., 2014).

While applications of sequence analysis are growing in the social sciences (see Aisenbrey and Fasang (2010) for a recent review), it remains unclear how this technique, typically used to analyse long, episodic lifecourse trajectories, will perform when applied to call record sequences, which are relatively short but with high variability in length. In this article we provide theoretical and empirical insights on this issue.

Our interest in fully observed auxiliary variables is motivated by the need to correct for differential response rates. Post-survey adjustments typically rely on one or more auxiliary measures Z, which are available for each sampled unit and correlated with both the response outcome and survey variables (Kalton and Flores-Cervantes, 2003; Little and Vartivarian, 2005; Bethlehem et al., 2011; Brick, 2013). The search for suitable Z variables has recently turned to paradata, which refer to data generated as a by-product of the survey process (Couper, 1998; Blom, 2009; Kreuter et al., 2010; Kreuter and Casas-Cordero, 2010; Olson, 2013; Conrad et al., 2013; Kreuter and Olson, 2013). These data are attractive because they are available for all sampled units at little extra cost (Groves, 2006). Call record data are a particularly rich subcategory of paradata which present a useful area for exploration. In household surveys, these data are typically recorded by the interviewer at each visit, and include the number, time and outcome of each attempt.

Post-survey weighting adjustments are typically based on the assumption that the survey variables are missing at random (MAR) with respect to the auxiliary variables used to construct the weights (Little and Rubin, 2002). The motivation for including variables derived from call record data is that the MAR assumption becomes more plausible if the auxiliary variables include call record information. By doing this, we are assuming that the pattern of calls made to a household is associated with the characteristics of this household as measured by the survey variables.

More formally, let Y_i be the survey variable of interest for sampling unit i, and R_i be the binary response-outcome indicator for the same unit, where $R_i = 1$ if i was successfully interviewed and 0 otherwise. We use Z_i^F to denote the fixed auxiliary variables typically used for nonresponse adjustment, such as frame data or interviewer observations. However, we would not believe the MAR assumption to be plausible if these were the only Z variables available to us. Hence, we introduce S_i as the set of available call record information about the sequence of calls made to i. MAR now corresponds to the more plausible assumption that

$$\Pr(R_i = 1 \mid Y_i, Z_i^F, S_i) = \Pr(R_i = 1 \mid Z_i^F, S_i), \tag{1}$$

that is, only the call record information and fixed auxiliary variables together can explain the association between Y_i and R_i. While we would not expect (1) to hold perfectly in practice, we do expect the inclusion of S_i to lead to substantial bias reduction.

The practical problem is how to include S_i when calculating the weights. The set of raw call record information is difficult to model because, generally, it comprises a diverse mix of variables, with the number of variables needed to represent the call sequences varying between units. As such, the aim is to create a fixed number of scalar variables $Z_i^S = Z^S(S_i)$ to summarise the call record information for each unit, but which still satisfy MAR:

$$\Pr(R_i = 1 \mid Y_i, Z_i^F, S_i, Z_i^S) = \Pr(R_i = 1 \mid Z_i^F, Z_i^S). \tag{2}$$

A commonly used choice for Z_i^S is the total number of calls made to a household (Beaumont, 2005; Blom, 2009; Wagner et al., 2013). But the number of calls is a rather crude summary of the recruitment process which may lose some of the information contained in S_i vital for nonresponse adjustment. It was against this background that Kreuter and Kohler (2009) first suggested using sequence analysis to summarise call records in a manner that minimises the loss of information. Examining sequences of household-level call outcomes collected from 14 countries across three waves of the European Social Survey (ESS), the authors created six continuous measures to summarise the complex patterns of calls. The results showed that the extracted indicators were predictive of the response outcome (R_i), but only weakly associated with key survey variables (Y_i). This suggests a limited scope for effective post-survey adjustment. However, Kreuter and Kohler (2009) relied on the default input costs in their application when it has been previously shown that the choice of input costs can strongly influence the output from sequence analysis (Bison, 2009; Gauthier et al., 2009).

In this paper, we explore whether different choices of input costs can improve the properties of sequence analysis-derived summary variables for weighting adjustments. Input costs C for sequence analysis must be chosen subjectively by the analyst, but in general there is little theoretical guidance as to how this should be done (Wu, 2000). In theory, we would like to find $C = C^*$ such that the association between Y_i and $Z_i^S = Z_i^S(C^*)$ for respondents and nonrespondents is the maximum among the $Z_i^S(C)$ satisfying equation (2), but this is impossible mainly because the available data are incompletely observed. In the absence of a robust method, we search across an array of plausible choices for C to assess the sensitivity of the resulting summary variables. Finally, we return to the central question for practice, namely, whether using sequence analysis-derived Z_i^S makes any difference to the weighted estimates of our survey variables. To do this, we take the 'most promising' of the summary variables from our sensitivity analysis and compare the resulting weighted estimates to those obtained using simple summaries such as the number of calls. Our choice is not optimal in any formal sense, but is promising in the spirit of Little and Vartivarian (2005) and Kreuter et al. (2010) in having the highest correlation with Y_i among the respondents. The analysis is carried out on call record data from Wave 1 of The Irish Longitudinal Study on Ageing (TILDA) (Whelan and Savva, 2013).

2. Sequence Analysis

In this section, we review the basic principles of sequence analysis, with a particular focus on the optimal matching metric. An important objective of this review is to draw attention to features of the technique which are problematic for its application to call record data. Sequence analysis comprises a set of tools for describing and summarising longitudinal or sequential data. Useful descriptive outputs include calculation of transition rates, identification of common subsequences, and graphical representations of trajectories. Sequence data may be summarised by a matrix, which defines the 'distance' between each pair of uniquely observed sequences. This distance matrix can be further analysed using traditional data reduction techniques so that the rich sequential information can be represented by a manageable number of dimensions.

2.1. States and state-spaces

States are the individual elements of which sequences are composed. The set of possible states, referred to as the state-space, is specified by the analyst and should be coded to reflect the underlying process of interest. The states should be exhaustive and mutually exclusive. Previous applications to call record data have focused on the outcome of each interviewer's visit, such that the ordered sequence of states for a household describes the full recruitment trajectory from first call till last. Other codings are possible; for example, if interviewers' work patterns are the subject of interest, a state-space comprising call times may be more appropriate than call outcomes.

Our choice of state-space is motivated by a key tenet of nonresponse research, namely, noncontact and unwillingness are distinct causes of nonresponse bias (Groves and Couper, 1998). For example, Lynn and Clarke (2002) found that hard-to-contact households were younger, healthier and more likely to be employed compared to easy-to-reach sample members. Reluctant householders, on the other hand, had less savings and lower housing costs. Thus we distinguish between calls that result in contact and those where no contact is made. In the case of successful contacts, reluctant and willing outcomes are differentiated.

2.2. The distance metric

A distance metric is used to quantify the dissimilarity between two sequences. Numerous metrics are available, and different metrics are sensitive to different properties of sequential patterns. Robette and Bry (2012) and Halpin (2012) review several common distance metrics. We focus on a metric based on a technique known as optimal matching (OM). We choose OM because it is by far the most widely used metric, and has been used in all previous applications to call record data.

The OM distance metric is defined in terms of the number of operations required to match the states of one sequence to those of another. This matching process proceeds through a series of edits to the states of a sequence. There are two types of edits possible: *substitution* where one element of a sequence is directly swapped for another; and *insertion* and *deletion*, where missing elements are inserted or superfluous elements are deleted. Note that as a deletion in one sequence corresponds to an insertion in the sequence to which it is being matched, both operations are equivalent and referred to collectively as *indels*. A cost, or penalty, is assigned to each substitution and indel, and the distance between any two sequences is the sum of the costs of all the operations needed to match the pair.

Before calculating the distance between a pair of sequences their respective states must be optimally aligned to return the minimal distance, that is, the 'cheapest' combination of substitutions and indels to complete the transformation. This optimal alignment is not necessarily obvious but can be obtained using dynamic programming methods (Needleman and Wunsch, 1970). Of course, the cheapest alignment of two sequences will be dictated by the costs assigned to the substitution and indel operations. These costs are the input values under the researcher's control mentioned in the Section 1, and their assignment plays a fundamental role in the method.

2.3. Substitution and indel costs

Substitution and indel costs interact in complicated ways to determine the OM distance between two sequences. Consequently, the best way to set these costs has been a contentious issue for sequence analysis in the social sciences (Stovel et al., 1996; Levine, 2000; Wu, 2000; Abbott and Tsay, 2000; Aisenbrey and Fasang, 2010). Ideally, some theory or insight about the system under study should be used to make an informed judgement about which choice of costs is most sensible. It is also possible to determine the costs empirically (Brzinsky-Fay et al., 2006; Gabadinho et al., 2011; Gauthier et al., 2009).

Costs play different roles depending on the operation in question. Substitutions are used to match elements which occur at the same point in a sequence, so the magnitude of the cost should reflect the relative similarity of the elements to be swapped. Consider, for example, an interviewer's visit which leads to an interview. In terms of the householder's underlying willingness to respond, an interview is more similar to an appointment than to a refusal. Therefore, a low substitution cost should be applied to match interview and appointment, and a high cost for matching interview with refusal. Costs can also be chosen to allow for classification error, or where outcomes are difficult for the interviewer to code. For example, it may be difficult to determine whether a noncontact should be interpreted as a polite refusal; applying a low substitution cost between noncontact and refusal can allow for this uncertainty by effectively making the states interchangeable in a contact sequence.

Indel operations insert additional states or delete superfluous ones. This allows the number and relative position of states within a sequence to change. The role of indels is particularly important when matching sequences of different lengths, because indel operations are necessary to make up for the difference. In the lifecourse literature, this is not an issue because analyses are usually restricted to events occurring within a given age range which leads to sequences of the same length (Macindoe and Abbott, 2004; Lesnard, 2010; Barban and Billari, 2012). However, the length of call record sequences can vary considerably, and so the role of the indel operation becomes more important. For example, choosing a high indel cost would impose a large distance between hard-to-reach households with long sequences and easy-to-reach households with short sequences. Standardisation, which rescales the distance between two sequences by dividing by the length of the longest of the pair, is useful when dealing with sequences of different length (Abbott and Hrycak, 1990).

The ratio of the indel cost to the substitution cost is a key property of the optimal matching metric (Bison, 2009; Lesnard, 2010). Setting the indel cost to less than half the lowest substitution cost means that any two non-matching states can be aligned using two indels rather than a substitution. This is because a deletion followed by an insertion will always be 'cheaper' than a substitution. Below this critical ratio, the OM algorithm will rely only on the indel operation and the coding of the states is effectively ignored, leading Hollister (2009, p.13) to apply the term 'pseudo-substitution' to such matches. When the indel cost is greater than half the highest substitution cost, indels will only be used to account for differences in sequence length. As a thought experiment, suppose the indel cost is set to 100 times the highest substitution cost. Under this scenario, the indel costs incurred when matching sequence lengths would dwarf costs accrued when matching states. The distance matrix would be dominated by variation in length and summarising this matrix would return a single dimension equivalent to the sequence length.

Substitution costs for a state-space of size k are represented as a symmetric matrix of order $k \times k$, which contains the $(k-1)k/2$ costs for each possible substitution. Traditionally only one indel cost is defined and often set relative to the substitution costs.

2.4. Summarising the distance matrix

Sequence analysis produces a $m \times m$ zero-diagonal, symmetric matrix, which reflects the relative distance between each of the m uniquely observed sequences in the dataset under analysis. In isolation, the distance matrix is not particularly informative, and some further multivariate technique is necessary to reduce the information contained in it into a manageable number of variables. The two most commonly applied data-reduction methods are cluster analysis and multidimensional scaling (MDS) (Brzinsky-Fay et al., 2006; Abbott and Tsay, 2000). The aim of cluster analysis is to assign units to a small number of homogeneous groups or clusters. MDS on the other hand produces scale summary variables, allowing the units to be plotted in one or more dimensions to aid interpretation (Bartholomew et al., 2008).

3. Data and Methods

3.1. Data

Before setting out how our investigation will proceed, we first describe the survey and call record data we will be using. The call record data were gathered during the first wave of The Irish Longitudinal Study on Ageing (TILDA). TILDA is a prospective study of the residential population over the age of 50 living in the Republic of Ireland (Whelan and Savva, 2013). As a large scale, interviewer-mediated, face-to-face household survey, its paradata are similar to the ESS and Understanding Society, whose call records have previously been analysed using sequence analysis (Kreuter and Kohler, 2009; Pollien and Joye, 2011, 2014; Maslovskaya et al., 2014; Durrant et al., 2014). Call record information was recorded by interviewers on a pen-and-paper form. The call record dataset comprises contact history information for over $24,000$ addresses, approached by interviewers between October 2009 and February 2011. Ineligible sample addresses (non-residential or with no occupant over the age of 50) were removed and the contact information was cleaned. This involved removing duplicate entries and trimming contact attempts beyond the tenth call. Contact attempts proceeded beyond the tenth call at 411 households and resulted in $1,206$ additional calls (2.8% of all calls) to valid addresses. Truncating this long tail at 10 calls greatly speeded up the sequence algorithm without losing much information. The following analysis is based on $41,353$ calls made to $10,074$ households. The number of visits to a household ranged between one and ten and the modal number of visits was two.

At each visit, the interviewer recorded the exact time and date of the call, and indicated one of 25 outcome categories specified on the contact form. Table 1 contains the distribution of these call outcomes. As per the discussion in Section 2.1, the call outcomes were then divided into five states, differentiating between contact and noncontact, and degrees of amenability to cooperate among contacts. Any call failing to achieve face-to-face contact with an occupant was labelled as a noncontact (40.2%). Individual- and household-level refusals made up 12.5% of all calls made. Calls where face-to-face contact was established, but neither a refusal nor appointment was recorded, were coded as neutral (20.3%). Positive outcomes (13.0%) were calls where an appointment was made, or where partial interviewing took place. Successful interviews, which include visits where one or multiple household members were interviewed, comprised 14.1% of the total number of calls.

Using these codings, the data comprise $2,315$ unique recruitment sequences. Similar to the patterns reported by Kreuter and Kohler (2009) and Pollien and Joye (2011, 2014), the most common recruitment trajectories were short sequences leading to interview. Appointment at the first call followed by interview at the second was the single most frequent pattern (16% of eligible sampled addresses).

Table 1. Call outcome coding

Outcome Category	**n**	**%**	**State**	**n**	**%**
No contact/still chasing	15,642	37.8	Noncontact	16,621	40.2
Occupied no contact at address after 5 calls	342	0.8			
Unable to access block / apartments	283	0.7			
Occupier in but not answering after 5 calls	143	0.3			
Property not found	114	0.3			
Unsure if occupied, no contact	75	0.2			
Property vacant	12	0.0			
Non-residential property	4	0.0			
Entry refused by warden	3	0.0			
Property derelict/demolished	3	0.0			
Household refusal	4,959	12.0	Refusal	5,177	12.5
Individual refusal	218	0.5			
Some contact but no appointment	6,373	15.4	Neutral	8,377	20.3
Appointment broken	631	1.5			
Too ill to participate	577	1.4			
Other	361	0.9			
Contact made but unable to assess eligibility	150	0.4			
Away during fieldwork	90	0.2			
No one aged 50+ in the household	70	0.2			
Withdrawn by head office	47	0.1			
Mother tongue require	45	0.1			
Partial interview - refused to continue	33	0.1			
Partial interview - to be completed	104	0.3	Positive	5,363	13.0
One half of an eligible couple cooperated	229	0.6			
Appointment made	5,030	12.2			
Successful interview	5,815	14.1	Success	5,815	14.1
Total	41,353	100			

3.2. Sensitivity analysis

As discussed in Section 2.3, the choice of costs applied to the sequence matching algorithm has an impact on the output of the analysis. However there is little guidance about how costs should be set, and previous applications to call record data have relied on the default cost settings without considering how this decision may have influenced results. In the absence of any theory with which to set the costs directly, we take the MAR assumption (2) to hold throughout, and test the sensitivity of the summary variables Z^S to different choices of costs.

We examined 100 scenarios generated by the following cross-classification of 10 indel costs and 10 substitution matrices. As this is a computationally intensive procedure, the analysis was restricted to a random 20% subset of eligible sampled addresses, reducing the data to $2,053$ households. Households were selected in such a way that the overall proportion of cooperating households in the full sample was maintained in the subset for analysis. The edit cost settings were systematically varied as follows:

Substitution Costs: For values of s from 0 to 1.8 in increments of 0.2, the substitution cost matrix was defined as:

	Nc	Rf	Nt	Po	Sc
Noncontact	0				
Refusal	2	0			
Neutral	2	$2-s$	0		
Positive	2	2	$2-\frac{s}{2}$	0	
Success	2	2	$2-\frac{s}{2}$	$2-s$	0

We refer to the value s as the substitution cost parameter. The purpose of varying this parameter is to distinguish between close and disparate call outcomes, as outlined in Section 2.3. Setting $s = 0$ produces a constant substitution matrix, i.e. every possible substitution is assigned the same value, and no distinction is made between call outcomes. When $s = 1.8$, the substitution costs vary depending on the substantive relationships between the outcomes being matched. For example, as s increases, the cost of matching a *Neutral* call to a *Refusal* call will diminish, reflecting that the former may be a polite version of the latter. Two dependencies on s were defined. For the two pairs of states we consider most similar (*Neutral*/*Refusal* and *Positive*/*Success*) the cost ranges between 2 and 0.2. For the other variable substitutions, the cost ranges between 2 and 1.2. Note that the choice of 2 as the highest possible substitution cost is arbitrary, what is important is the magnitude of the highest cost relevant to the other substitution and indel costs.

Indel Costs: Holding each of these substitution matrices constant, the indel cost was increased from 0.2 to 2 in incremental steps of 0.2. As the indel cost increases relative to the substitution cost, the distance between two sequences returned by the OM algorithm will increasingly depend on the disparity between the sequence lengths.

Sequence analysis using the OM metric was applied to the call record data for each of the 100 scenarios we consider. Each distance matrix is summarised using the two most informative dimensions obtained using MDS. We then examine the correlation between the summary measures and two characteristics of the call record sequences. Specifically, the characteristics are the response indicator R and the number of calls, or sequence length, L. Examining these correlations reveals how different cost settings affect the dimensions being captured by the OM distance metric. Following this, correlations between the MDS summary variables and 10 key survey variables were examined, to identify the cost settings which maximise this key relationship.

3.3. Adjustment weighting

We are also wish to find out whether the Z^S variables based on sequence analysis can give different results to those obtained using more straightforward aggregates of the call records. To answer this question, we compare unweighted survey estimates to the weighted estimates obtained using two different sets of inverse probability weights. Both sets of weights are derived from a multilevel logistic regression model of the household response outcome incorporating a combination of time-invariant paradata and variables based on the call record data. The first weight included straightforward summaries of the call record data, while the alternative weight incorporated sequence-based summaries. A multilevel model was used to allow for potential inter-household dependencies in response behaviour within areas, arising from the two-stage clustered sample (Hox, 1998). The multilevel structure consisted of $10,074$ households nested within 634 neighbourhoods.

In more detail, the nonresponse weights were calculated as follows. To begin we define the response indicator $R_{ij} = 1$ if household i in neighbourhood j cooperated, and 0 otherwise. The probability of response is defined as $\pi_{ij} = Pr(R_{ij} = 1)$. We model the log odds of response using a multilevel logistic model

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + \beta_F Z_{ij}^F + \beta_S Z_{ij}^S + u_j \tag{3}$$

where α is the intercept, Z_{ij}^F is a vector of fixed auxiliary variables with coefficients β_F, Z_{ij}^S is a vector of call record variables with coefficients β_S, and $u_j \sim N(0, \sigma_u^2)$ are neighbourhood-level random effects.

Here Z_{ij}^F comprises interviewer observations on the location, type and condition of the surrounding area. These household-level auxiliary variables are representative of the paradata routinely collected by large-scale survey organisations and have been previously linked to nonresponse outcomes (Groves and Couper, 1998). Two models were fitted, with the following competing choices for Z^S:

(a) Aggregates of the call record data

- (i) The number of calls (four categories)
- (ii) The proportion of calls achieving contact (four categories)

(b) Variables based on the sequence analysis distance matrix

- (i) The first MDS dimension (quartiles)
- (ii) The second MDS dimension (quartiles)

The effects of the Z^S variables were parameterised using dummy variables to improve the model fit: in the case of the MDS dimensions quartiles were employed; for the aggregated variables, sensible divisions which provided groups of reasonably equal size were chosen. See Table 2 for more details.

The inverse probability weights are the reciprocal of the estimated response probabilities obtained from fitting these models, using empirical Bayes estimates of u_j. The weights derived from each model were trimmed at 10 to avoid excessive values and limit variance inflation. We compare unweighted point estimates and standard errors for ten TILDA variables to estimates based on these weights. For the weighted estimators, we used an approximate bootstrap method to account for both the sampling scheme and uncertainty in estimation of the nonresponse weights. Further details of this method can be found in Supplementary Information.

Data preparation and analysis was performed in Stata 12. Sequence analysis was carried out using the “SQ-Ados” program (Brzinsky-Fay et al., 2006).

4. Results

4.1. Sensitivity analysis

In order to understand how the candidate Z^S variables are influenced by different cost choices, we examined the correlations between the extracted MDS dimensions and two summary indicators of each call sequence: the sequence length L and the response outcome R. Figure 1 summarises the results; each box represents a summary of 100 correlations, one from each of the 100 combinations of costs. The range of these correlations indicates the extent to which properties of the sequence-derived Z^S variables can be influenced by the cost settings.

The first box depicts the distribution of correlations observed between Dimension 1 and L. The length of the box indicates the interquartile range $(0.40 - 0.51)$, with the median value (0.42) at the midline. The whiskers display the minimum (0.38) and maximum (0.62) correlations observed. The spread of this distribution indicates that the relationship between the first MDS dimension and sequence length varies with the indel and substitution cost settings.

Table 2. Summary of paradata and survey variables used in the analyses

Variable	Type	Description
Interviewer Observations (Z^F Variables)		
House Type	Binary	Detached home, incl. farm (1) other house type (0)
State of Dwelling	Binary	Physical state of the buildings in the area:
		Very good, Good (1); Satisfactory, Bad, Very Bad (0)
Dublin Indicator	Binary	Household in Dublin (1) or outside (0)
Call Records (Z^S Variables)		
Number of Calls	Ordinal	1, 2–3, 4–5, 6+
Proportion of Contacts	Ordinal	0–.499, .5–.667, .668–.999, 1
MDS 1	Ordinal	Quartiles of the first MDS dimension
MDS 2	Ordinal	Quartiles of the second MDS dimension
Survey Outcomes (Y Variables)		
Poor Physical Health	Binary	Self-reported physical health:
		Fair, Poor (1); Excellent, Very Good, Good (0)
Degree	Binary	Highest level of education:
		Tertiary (1); Primary or Secondary (0)
Sick/Disabled	Binary	Principle economic status:
		Unable to work due to permanent sickness or disability (1); Other (0)
Home/Family Care	Binary	Principle economic status:
		Looking after home or a family member (1); Other (0)
Single	Binary	Marital status: Single, never married (1); Other marital status (0)
Separated/Divorced	Binary	Marital status: Separated or divorced (1); Other marital status (0)
Chronic Pain	Binary	Self-report of chronic pain: In pain (1); No pain (0)
Polypharmacy	Binary	Use of multiple medications: 5+ Medications (1); 0 5 Medications (0)
Poor Mental Health	Binary	Self-reported mental health:
		Fair, Poor (1); Excellent, Very Good, Good (0)
Loneliness	Binary	Frequency of loneliness:
		Sometimes, Moderately, Always (1); Rarely or Never (0)

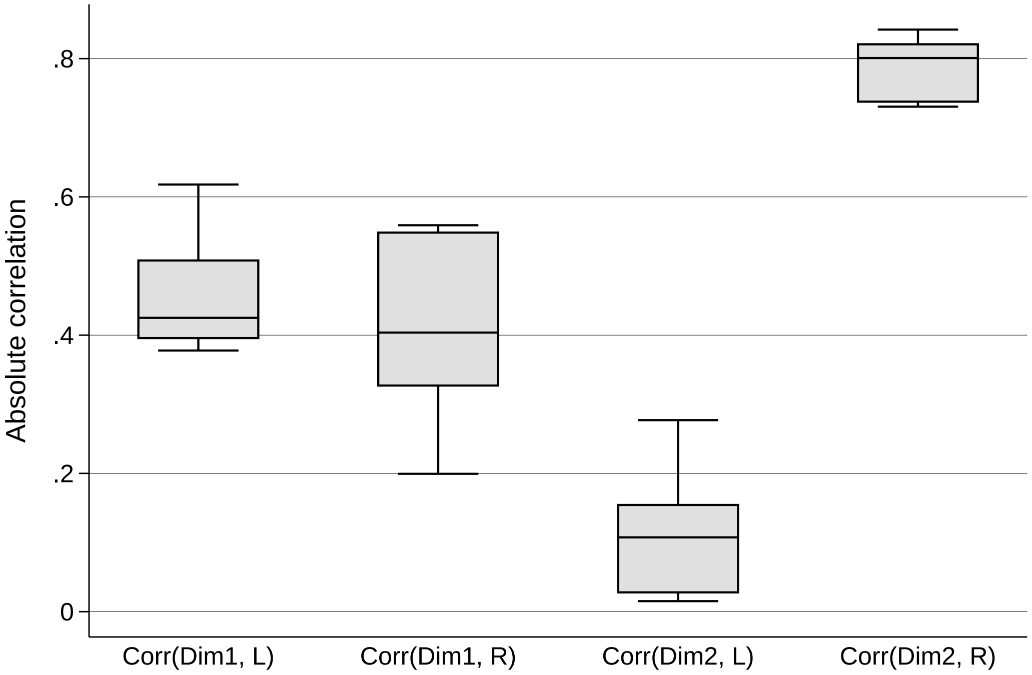

Fig. 1. Correlations between MDS dimensions and survey summary variables (Sequence length L and Response outcome R) across 100 cost scenarios. Correlations shown as absolute values.

There is a similarly large spread of correlations observed between Dimension 1 and R, which range from 0.20 to 0.56 depending on the combination of costs employed. For Dimension 2, correlations with L range from 0.02 to 0.28 while correlations with R range from 0.73 to 0.84. Regardless of the cost settings employed, the first extracted dimension tends to capture information about both the number of calls and response outcome, while the second dimension is dominated by R.

Figure 2 contains a plot of the correlations between the different MDS dimensions and the sequence characteristics across different indel and substitution settings. This reveals the complicated interactions between costs which moderate the information captured by the OM distance metric.

From the upper left-hand plot it can be seen that the correlation between Dimension 1

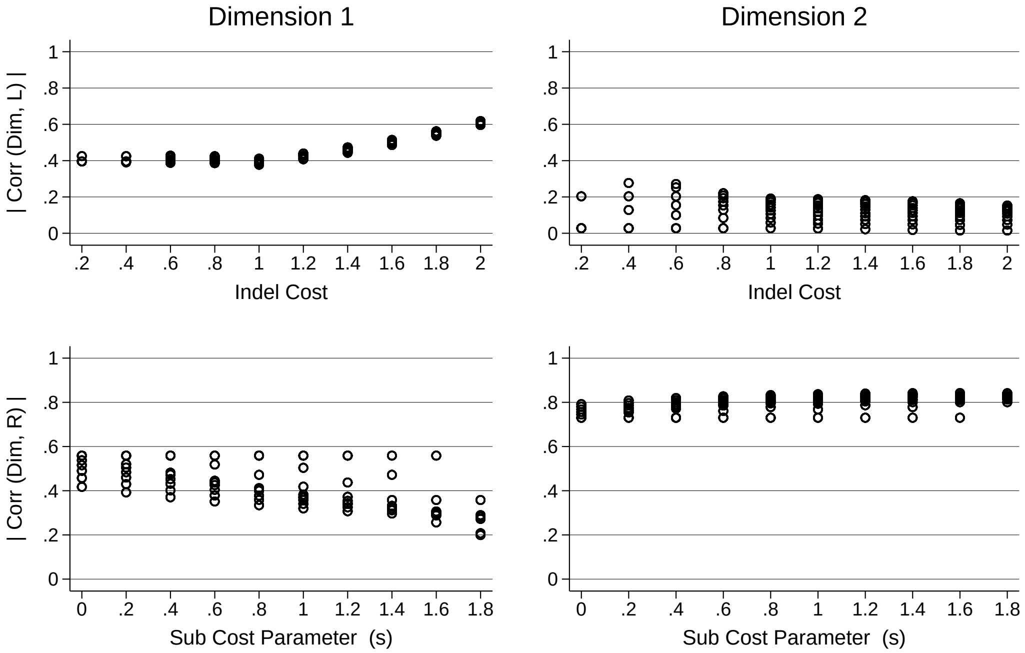

Fig. 2. Correlations between MDS dimensions and sequence length (L) and response (R), by indel cost and substitution cost parameter.

and L increases linearly once the indel cost exceeds the critical value of 1, i.e. half the largest possible substitution cost. Allowing substitution costs to vary has a limited impact on this relationship, regardless of the indel setting. As the first dimension becomes increasingly aligned with the sequence length, the opposite effect is observed for the second dimension: as the indel cost increases, the extent to which the second dimension captures information about length diminishes. The effect of changing the substitution cost parameter s can be seen in the lower plots. As s increases, thereby introducing variability in the substitution costs, the correlation between the first dimension and R decreases. The second dimension consistently displays a high correlation with R, which is maximised when the most variable substitution costs are employed. Thus, by appropriately setting the input costs, the analyst can to some extent control which features of the call sequence are being captured by the

OM distance. However the variation described above is only with respect to properties of the sequences themselves.

In order to understand which characteristics of the *household* are being reflected in its call sequence, and how the cost settings influence this, we examine the correlation between the extracted dimensions and key survey variables. We also compare these to the correlations between survey variables and other paradata variables, to explore the potential added value of sequence-based summaries over simpler paradata such as aggregates of call records or time-invariant observations on the household. Table 3 summarises the correlations between the available paradata measures and ten survey variables, for the subset of respondents. For the MDS dimensions the minimum and maximum correlations observed across the 100 cost scenarios are presented. It is clear that the MDS dimensions are not highly correlated with the survey variables examined here. The difference between the maximum and minimum absolute correlations observed is small, suggesting that the sensitivity of this relationship to the cost settings is low. This means that there is little potential to fine tune costs to maximise the correlations of interest. Neither MDS dimension out-performs the time-invariant or call-aggregated paradata variables. The observed correlations between these paradata and survey variables are in the same range as those presented by Kreuter and Kohler (2009) and Kreuter et al. (2010).

There are two clear outcomes from this analysis. First, cost settings do have an impact on which aspects of the sequences inform the OM distance structure and the resulting summary dimensions. In particular, the relative size of the indel cost moderates the extent to which disparity in sequence length is influential. Second, regardless of the cost settings employed, the dimensions summarising the information in the call record data do not correlate with substantive survey variables. This approach has allowed us to quickly assess 200 candidate summaries of the call record data. That none of these produce a suitable summary adjustment variable indicates that there is very much a limit to what the pattern of call outcomes at a household can reveal about the characteristics of individual occupants.

Table 3. Summary of correlations between paradata variables and survey variables.

Variable	Dimension 1		Dimension 2		State	Detached	Dub	# Calls	% Con
	min	max	min	max					
Poor physical health	0.02	0.03	0.00	0.02	0.12	0.06	0.03	0.01	0.01
Degree	0.00	0.03	0.06	0.08	0.08	0.08	0.10	0.04	0.03
Econ status: Disabled	0.02	0.04	0.05	0.07	0.08	0.07	0.06	0.01	0.00
Econ status: Homemaker	0.00	0.02	0.09	0.11	0.04	0.06	0.05	0.01	0.01
Mar status: Single	0.12	0.15	0.00	0.08	0.10	0.03	0.03	0.02	0.14
Mar status: Sep/Div	0.03	0.04	0.00	0.03	0.11	0.09	0.01	0.02	0.05
Chronic pain	0.03	0.05	0.02	0.05	0.03	0.11	0.05	0.02	0.05
Polypharmacy	0.05	0.06	0.02	0.04	0.07	0.05	0.01	0.02	0.03
Poor mental health	0.01	0.01	0.00	0.01	0.10	0.11	0.04	0.00	0.02
Loneliness	0.02	0.04	0.00	0.02	0.13	0.06	0.02	0.05	0.02

Note: Correlations presented as absolute values. State = State of household; Detached = Detached home; Dub = Dublin indicator; # Calls = Number of calls to household; % Con = Proportion of calls to household where face-to-face contact was established.

4.2. Impact on weighted estimates

In this section, we explicitly examine whether supplementing post-survey adjustment weights with Z^S variables derived using sequence analysis can lead to different results. The sensitivity analysis indicated that there was no combination of costs which led to a large correlation between the summary variables and survey items. The maximum association, to the extent there was one, was achieved when setting the indel and substitution cost parameters to 1 and 1.4 respectively. Therefore, we choose the summary variables based on these costs. Rather than use a subsample here, sequence analysis was repeated on the full dataset. Table 4 contains three sets of estimated proportions and standard errors for ten TILDA survey variables. The estimates are unweighted and weighted using two different schemes which differ according to the Z^S variables used (as described in Section 3.3). The first weight (weight 1) is based on simple aggregates of the call record data; the second weight (weight 2) uses the sequence analysis summaries. In both cases the shift from the unweighted estimate, measured in unweighted standard error units, is also presented.

When weight 1 is applied, notable point estimate shifts are observed for the marital status dummies: the proportion of single households and the proportion of separated/divorced households. Both of these estimates increase by approximately one percentage point, from 11.9% to 12.9% for the former and 8.8% to 9.8% for the latter. Based on complete enumeration from the contemporaneous Irish 2011 census, the true proportions of single and separated/divorced households are 14.9% and 11.2% respectively for the population of interest. So for both proportions, the unweighted estimates are downwardly biased and weight 1 reduces this bias. When weight 2 is applied these estimates shift in the opposite direction, although they are not substantively different from the unweighted estimates.

The estimated proportions for the remaining variables are reasonably stable, regardless of the weight employed. The standard errors are inflated for the weighted estimates, as would be expected (Little and Vartivarian, 2005). This is especially true for the weight incorporating sequence analysis summaries of the call records (weight 2).

Table 4. Point estimates and standard errors for ten survey variables

	Unweighted		Weighted (weight 1)			Weighted (weight 2)		
Variable	Mean	SE	Mean	SE	SE Shift	Mean	SE	SE Shift
Poor physical health	23.7	0.597	24.0	0.624	0.5	23.8	0.747	0.2
Degree	29.1	0.746	29.1	0.765	0.0	28.2	0.883	−1.2
Econ status: Disabled	4.8	0.318	5.2	0.344	1.3	5.2	0.435	1.3
Econ status: Homecare	15.7	0.582	15.4	0.567	−0.5	16.1	0.708	0.7
Mar status: Single	11.9	0.441	12.9	0.495	2.3	11.4	0.553	−1.1
Mar status: Sep/div	8.8	0.406	9.8	0.475	2.5	8.6	0.491	−0.5
Chronic pain	38.3	0.800	38.7	0.801	0.5	38.3	0.923	0.0
Polypharmacy	20.9	0.544	21.0	0.577	0.2	21.4	0.716	0.9
Poor mental health	10.6	0.433	10.9	0.453	0.7	10.4	0.548	−0.5
Loneliness	21.2	0.612	21.9	0.641	1.1	21.0	0.732	−0.3

Note: Weighted standard errors calculated using an asymptotic bootstrap (details in supplementary material). All standard errors account for stratification and clustering of sampled units due to the complex sample design.

5. Discussion

Recently, the use of call records as a source of auxiliary variables for post-survey nonresponse-adjustment weighting has received considerable attention. However, the question remains open as to whether these records contain useful information for nonresponse adjustment and, if so, how best to exploit this information. In this article, we considered the use of sequence analysis for summarising call records in a way that preserves vital information which might be lost using simple summaries like the total number of calls to a household. Applying sequence analysis to any dataset requires a series of decisions to be made, which are often subjective and lacking theoretical motivation (Wu, 2000). We have evaluated some of the choices made when applying sequence analysis to call record data.

In the absence of a robust method for choosing costs, we proposed a sensitivity analysis to assess the impact of costs on the resulting summary measures. The results showed that varying the costs assigned to the OM edit operations does influence which aspects of the contact sequences are most important when calculating the similarity between recruitment trajectories. Higher indel costs increase the dependency on disparity in the number of calls made; substitution costs which distinguish between substantively different call outcomes will increase the distance between cooperative and unwilling trajectories. Importantly, variation in the number of calls will always dominate the sequence analysis output, regardless of the coding of the call. This result is relevant in other contexts where there is high variability in sequence length. However, despite this sensitivity, the correlations between the sequence summaries and the survey variables were consistently low regardless of the costs we considered. This is why the weighted results were not substantially different from those obtained using a simple summary because, to a large extent, nonresponse adjustments are driven by the strength of association between the auxiliary and survey variables. Including sequence analysis summary variables in the nonresponse model increased the standard errors with little change to point estimates.

Our results broadly align with those of Kreuter and Kohler (2009) and Pollien and Joye (2011, 2014) who all found that summary variables derived through sequence analysis of call record data were predictive of response outcomes, but less so of survey variables. While this previous work did not explore the role of costs, we could not find a choice of costs in our application that markedly increased the association with any of the survey variables. This indicates that the role of costs may not be crucial, but we cannot be certain because we only considered one survey. Thus, we would advise others to assess the sensitivity of their results to different choices of costs. If the results *were* sensitive to the choice, the question would then be about which costs to choose. We proposed choosing the costs which lead to the summary variables most highly correlated with the survey variables, but this intuition presumes that (2) holds to an acceptable degree, which cannot be verified.

Of course, it remains entirely possible that call records are simply uninformative about the characteristics of the survey units. However, it is difficult to give a definitive answer to this questions because it will vary between surveys. The promise of call records is based on the assumption that the call record information is indicative of household characteristics being related to at-home patterns, such as age, employment status and family size. These associations were weak for the set of survey variables we examined, but may be different in other survey contexts.

Two other explanations for the weakness of association should also be taken into account. The first is data quality. Call records may be prone to measurement error; for example, it may be difficult for interviewers to distinguish between noncontact and hidden refusal (Nicoletti and Peracchi, 2005). Lack of motivation and issues with technology may also decrease the quality of interviewer coding (West and Sinibaldi, 2013). Another explanation is that the response outcome and call records are the outcomes of a joint data collection process (Beaumont, 2005). Thus, the contact data are not a fixed property of a household but rather they are subject to variation over theoretical replications of the same survey protocol

The collection of complex call records presents an administrative burden, and these data are only useful insofar as they can be used to improve surveys, be it in terms of cost reduction, bias reduction, fieldwork management or any other application. This investigation suggests that, in terms of generating nonresponse adjustment variables, more straightforward paradata can adequately capture all the available information about a household. We acknowledge that there are some limitations to this study. Measurement error properties of interviewer records are only recently coming under scrutiny (Sinibaldi et al., 2013; West and Sinibaldi, 2013), and errors in the recorded number of calls may bias results (Biemer et al., 2013). Call dynamics which might be present in a general population survey may be lost in this sample of over 50s. Hazard models such as those explored by Durrant et al. (2011, 2013) may be more appropriate for call record data which do not have the long, episodic patterns present in say, career or family trajectories.

Acknowledgements

We would like to thank the team at The Irish Longitudinal Study on Ageing for providing access to the paradata necessary to carry out this study. We would also like to thank three anonymous reviewers whose comments greatly improved this paper.

References

Abbott, A. and A. Hrycak (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *Americal Journal of Sociology 96*(1), 144 – 185.

Abbott, A. and A. Tsay (2000). Sequence analysis and optimal matching records in sociology: Review and prospect. *Sociological Methods and Research 29*(3), 3 – 33.

Aisenbrey, S. and A. E. Fasang (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the lifecourse. *Sociological Methods and Research 38*, 420 – 462.

Barban, N. and F. C. Billari (2012). Classifying life course trajectories: A comparison of latent class and sequence analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(5), 765–784.

Bartholomew, D., F. Steele, I. Moustaki, and J. Galbraith (2008). *Analysis of Multivariate Social Science Data* (2^{nd} ed.). Chapman and Hall/CRC.

Beaumont, J. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology 31*(2), 227–231.

Bethlehem, J., F. Cobben, and B. Schouten (2011). *Handbook of Nonresponse in Household Surveys*. Wiley.

Biemer, P. P., P. Chen, and K. Wang (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*(1), 147–168.

Bison, I. (2009). OM matters: The interaction effects between indel and substitution costs. *Methodological Innovations Online 4*(2), 53 – 67.

Blom, A. G. (2009). Nonresponse bias adjustments: What can process data contribute? ISER Working Paper 2009-21, University of Essex. Available at http://www.econstor.eu/bitstream/10419/92033/1/2009-21.pdf.

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics 29*(3), 329–353.

Brzinsky-Fay, C., U. Kohler, and M. Luniak. (2006). Sequence analysis with Stata. *Stata Journal 6*(4), 435 – 460.

Conrad, F. G., J. S. Broome, J. R. Benkí, F. Kreuter, R. M. Groves, D. Vannette, and C. McClain (2013). Interviewer speech and the success of survey invitations. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*(1), 191–210.

Couper, M. (1998). Measuring survey quality in a casic environment. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 41–49.

Durrant, G., J. D'Arragio, and F. Steele (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influences. *Journal of the Royal Statistical Society: Series A 174*(4), 1029 – 1049.

Durrant, G. B., J. D'Arrigo, and F. Steele (2013). Analysing interviewer call record data by using a multilevel discrete time event history modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*(1), 251–269.

Durrant, G. B., O. Maslovskaya, , and P. W. Smith (2014). Sequence analysis as a tool for investigating call record data. Working paper, University of Southampton. Available at `http://eprints.soton.ac.uk/375810/1/paper_Durrant%20et%20al_Sequ%20anal_vs%205.pdf` [accessed 04/05/15].

Gabadinho, A., G. Ritschard, N. S. Mueller, and M. Studer. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software 40*(4), 1 – 37.

Gauthier, J.-A., E. Widmer, P. Bucher, and C. Notredame (2009). How much does it cost?: Optimization of costs in sequence analysis of social science data. *Sociological Methods and Research 38*(1), 197 – 231.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly 70*(5), 646–675.

Groves, R. M. and M. P. Couper (1998). *Nonresponse in Household Surveys.* New York: John Wiley and Sons, Inc.

Halpin, B. (2012). Sequence analysis of life-course data: A comparison of distance measures. Working paper WP2012-02, Department of Sociology, University of Limerick. Available at `http://www.ul.ie/sociology/pubs/wp2012-02.pdf`.

Hollister, M. (2009). Is optimal matching suboptimal? *Sociological Methods and Research 38*(2), 235–264.

Hox, J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, and M. Schader (Eds.), *Classification, data analysis, and data highways*, pp. 147–154. Springer.

Kalton, G. and I. Flores-Cervantes (2003). Weighting methods. *Journal of Official Statistics 19*(2), 81 – 97.

Kreuter, F. and C. Casas-Cordero (2010). Paradata. Working paper, German Council for Social and Economic Data. Available at `http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_136.pdf`.

Kreuter, F. and U. Kohler (2009). Analyzing contact sequences in call record data. Potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. *Journal of Official Statistics 25*(2), 203–226.

Kreuter, F., K. Olsen, J. Wagner, T. Yan, T. E. Rice, C. Casas-Cordero, M. Lemay, A. Peytchev, R. Groves, and T. Raghunathan (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. *Journal of the Royal Statistical Society: Series A 173*(2), 389–407.

Kreuter, F. and K. Olson (2013). Paradata for nonresponse error investigation. In F. Kreuter (Ed.), *Improving Surveys with Paradata: Analytic Uses of Process Information.* Wiley.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods and Research 38*(3), 389 – 419.

Levine, J. H. (2000). But what have you done for us lately?: Commentary on Abbott and Tsay. *Sociological Methods and Research 29*(1), 34 – 40.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data.* Hoboken, New Jersey: Wiley.

Little, R. J. and S. Vartivarian (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology 31*(2), 161–168.

Lynn, P. and P. Clarke (2002). Separating refusal bias and non-contact bias: Evidence from UK national surveys. *Journal of the Royal Statistical Society: Series D (The Statistician) 51*(3), 319–333.

Macindoe, H. and A. Abbott (2004). Sequence analysis and optimal matching techniques for social science data. In M. Hardy and A. Bryman (Eds.), *Handbook of Data Analysis.* London: Sage.

Maslovskaya, O., G. B. Durrant, and P. W. Smith (2014). Sequence analysis as a graphical tool for investigating call record data. In *Paradata Conference: From Survey Research to Practice, London, GB, 26th Jun 2014.*

Needleman, S. B. and C. D. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology 48*(3), 443 – 453.

Nicoletti, C. and F. Peracchi (2005). Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 168*(4), 763–781.

Olson, K. (2013). Paradata for nonresponse adjustment. *The ANNALS of the American Academy of Political and Social Science 645*(1), 142 – 170.

Pollien, A. and D. Joye (2011). A la poursuite du répondant? Essai de typologie des séquences de contact dans les enquêtes. In E. Seismo (Ed.), *Parcours de Vie et Insertions Sociales*, pp. 189–212. Zürich.

Pollien, A. and D. Joye (2014). Patterns of contact attempts in surveys. In P. Blanchard, B. Felix, and J.-A. Gauthier (Eds.), *Advances in Sequence Analysis, Theory, Methods, Applications*. London: Springer.

Robette, N. and X. Bry (2012). Harpoon or bait? A comparison of various metrics in fishing for sequence patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 116*(1), 5–24.

Sinibaldi, J., G. B. Durrant, and F. Kreuter (2013). Evaluating the measurement error of interviewer observed paradata. *Public Opinion Quarterly 77*(S1), 173–193.

Stovel, K., M. Savage, and P. Bearman (1996). Ascription into achievement: Models of career systems at Lloyds Bank, 1890-1970. *American Journal of Sociology*, 358–399.

Wagner, J., R. Valliant, F. Hubbard, and C. Jiang (2013). Level-of-effort paradata and nonresponse adjustment models for a national face-to-face survey. working paper, University of Michigan. Available at `http://hrsonline.isr.umich.edu/sitedocs/userg/HRS_Weights_Wagner_etal.pdf`.

West, B. T. and J. Sinibaldi (2013). The quality of paradata: A literature review. In F. Kreuter (Ed.), *Improving Surveys with Paradata: Analytic Use of Process Information*, pp. 339–359. Wiley.

Whelan, B. J. and G. M. Savva (2013). Design and methodology of The Irish Longitudinal Study on Ageing. *Journal of the American Geriatrics Society 61*(s2), S265–S268.

Wu, L. L. (2000). Some comments on sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological methods and research 29*(1), 41 – 64.

Supplementary Information for 'Sequence Analysis for Call Record Data: Exploring the Role of Different Cost Settings'

by Mark Hanly (University of Bristol), Paul Clarke (University of Essex) and Fiona Steele (London School of Economics & Political Science)

1. Introduction

In Section 3.3 of the article, we describe an approximate bootstrap procedure for standard error estimation that accounts for the complex sampling design and imprecision in our estimates of the response propensities. While this simple but general procedure may not estimate the standard errors particularly accurately, it does allow us to assess the relative precision of the different weighted estimators in our study.

2. The Algorithm

Recall that the response probability $\Pr(R_{ij} = 1) = \pi_{ij}$ for household i in neighbourhood j was specified to follow the two-level logistic model

$$\pi_{ij} = \text{logit}^{-1}\left(\alpha + \beta_F Z_{ij}^F + \beta_S Z_{ij}^S + u_j\right), \quad (\text{A}.1)$$

where $\text{logit}^{-1}(a) = e^a/(1+e^a)$ is the inverse logistic function, Z_{ij}^F is the vector of fixed auxiliary variables, Z_{ij}^S is the vector of sequence-based summaries, and u_j is the normally distributed neighbourhood-level random effect. Herein dropping the neighbourhood subscript to simplify notation, we denote by $\hat{\pi}_i$ the estimated response probability for household i.

The approximate bootstrap can now be defined as follows:

Bootstrap-Replicate Phase:

Step 1: Independently generate response outcomes from $R_i^* \sim \text{Bernoulli}(\hat{\pi}_i)$ for each sample member (i.e. from each respondent and each nonrespondent in the actual data).

Step 2: Denoting the bootstrap sample by $\left\{\left(z_i^F, z_i^S, r_i^*\right)\right\}$, refit the response probability model (A.1) to the bootstrap sample to obtain π_i^* and weighted $\theta^* = \sum_{i=1}^n w_i y_i \, r_i/\pi_i^*$ (where w_i is the appropriate set of design weights).

Step 3: Repeat Step 1 and Step 2 B times to obtain bootstrap replicates $\left\{\theta^{*(b)} : b = 1, \ldots, B\right\}$.

Variance-Component Phase:

Step 4: Create the survey 'variable' $y_i r_i/\hat{\pi}_i$ (which always takes the value zero for sample nonrespondents) and obtain the design-consistent estimate of the total/mean of this survey variable using the appropriate set of design weights: denote the result by v_1. We did this using the `svy` commands in Stata.

Step 5: Calculate the bootstrap variance of $\theta^{*(b)}$: denote the result by v_2.

Step 6: Combine the two estimates to obtain the estimated standard error $\sqrt{v_1 + v_2}$.

3. A Heuristic Justification of the Approximate Bootstrap Method

Consider the following weighted estimate

$$\hat{\theta} = \sum\nolimits_{i=1}^{n} w_i y_i \, r_i / \hat{\pi}_i, \qquad (\mathrm{A.2})$$

where the survey variable is $y_i r_i$ rather than y_i, which takes the value 0 for nonrespondents and leads to a sample size of n rather than the number of respondents $\sum_{i=1}^{n} r_i = r$. (For example, the weights for a simple random sampling design are $w_i = 1/n$.)

Estimate (A.2) is a realisation from the estimator

$$\hat{\theta}_{n,N} = \sum\nolimits_{k=1}^{N} w_k y_k \, R_k S_k / \hat{p}_k, \qquad (\mathrm{A.3})$$

where we have changed the index from i to k to make clear we are summing over the entire population. Without loss of generality, we can take the sample members to be indexed by $k = 1, \ldots, n$ and those not in the sample by $k = n + 1, \ldots, N$.

Note that (A.3) has the form of an estimator of the population total, but we could have equivalently written it as a population-mean estimator of the form

$$\hat{\theta}_{n,N} = \frac{\sum_{k=1}^{N} w_k^* y_k \, R_k S_k / \hat{\pi}_k}{\sum_{k=1}^{N} w_k^* S_k}.$$

This differs from (A.3) in the design weights; the two sets of design weights are connected by the relationship $w_k = w_k^* / \sum_{m=1}^{N} w_m^* S_m$ (for example, $w_k^* = N/n$ for simple random sampling).

When considering the properties of estimator (A.3), we treat all values of the auxiliary and survey variables as fixed constants. Any random variation comes from the sampling indicator S_k and response outcome indicator R_k. The former is determined by the sampling design; the latter is assumed to follow the model

$$R_k \sim \mathrm{Bernoulli}(\pi_{k0}), \qquad (\mathrm{A.4})$$

conditional on being sampled. Each sample member is taken to respond or nonrespond independently of the others, and π_{k0} is the true response propensity.

For large n, we can approximate (A.3) by

$$\hat{\theta}_{n,N} \simeq \sum\nolimits_{k=1}^{N} w_k y_k \, R_k S_k / \pi_{k0}, \qquad (\mathrm{A.5})$$

from which $\hat{\theta}_{n,N}$ is easily shown to be approximately unbiased under the joint sampling design and response model (A.4), given the appropriate choice of design weights.

Our variance estimator is based on the decomposition

$$\mathrm{var}(\hat{\theta}_{n,N}) = \mathrm{var}_{S_1,\ldots,S_N}\{E(\hat{\theta}_{n,N}|S_1,\ldots,S_N)\} + E_{S_1,\ldots,S_N}\{\mathrm{var}(\hat{\theta}_{n,N}|S_1,\ldots,S_N)\}, \quad (\mathrm{A.6})$$

where the inner moments are with respect to the conditional distribution of $R_1,\ldots,R_n$ given $S_1,\ldots,S_N$.

In hypothetical situations where y_i is known for each sample member, we would estimate both inner moments using a parametric bootstrap by taking draws from

$$R_i^* \sim \mathrm{Bernoulli}(\hat{\pi}_i),$$

for $i = 1,\ldots,n$, to create a series of bootstrap samples. Then for each bootstrap sample, we would estimate π_i^*, and obtain the bootstrap replicates

$$\theta^{**} = \sum\nolimits_{i=1}^{n} w_i y_i\, r_i^*/\pi_i^*.$$

However, we cannot calculate θ^{**} because y_i is missing for nonrespondents in the observed sample; instead, we calculate

$$\theta^* = \sum\nolimits_{i=1}^{n} w_i y_i\, r_i/\pi_i^*, \quad (\mathrm{A.7})$$

which is why we refer to this as an 'approximate' bootstrap.

For the first component on the right hand side of (A.6), we could estimate $E(\hat{\theta}_{n,N}|S_1,\ldots,S_N)$ using

$$\hat{E}^*(\hat{\theta}_{n,N}|S_1,\ldots,S_N) = \sum\nolimits_{i=1}^{n} w_i y_i r_i\, \overline{1/\pi_i^*}, \quad (\mathrm{A.8})$$

where $\overline{1/\pi_i^*}$ is the mean of the bootstrap replicates for household i. For simplicity, however, we use $1/\hat{\pi}_i$ because it is the probability limit of $\overline{1/\pi_i^*}$ as B increases. In practice, we estimate (A.8) using an appropriate design-based variance estimator (of a total or mean, depending on which design weights we choose) with $y_i\, r_i/\hat{\pi}_i$ as the fixed 'survey variable' and remembering to treat the 0 values for nonrespondents as genuine data values.

For the second component on the right hand side of (A.6), we estimate $\mathrm{var}(\hat{\theta}_{n,N}|S_1,\ldots,S_N)$ by

$$\widehat{\mathrm{var}}^*(\hat{\theta}_{n,N}|S_1,\ldots,S_N) = \frac{1}{B-1}\sum\nolimits_{b=1}^{B}\left(\theta^{*(b)} - \overline{\theta^*}\right)^2, \quad (\mathrm{A.9})$$

where $\overline{\theta^*}$ is the mean of the bootstrap replicates. This completes estimation of the second variance component if we further assume that $\mathrm{var}(\hat{\theta}_{n,N}|S_1,\ldots,S_N)$ is equal for all samples.