Generating Nonresponse Adjustment Variables using Sequence Analysis of Call Record Data.

Mark Hanly
Paul Clarke
Fiona Steele

University of Bristol
May 2013

Abstract

Kreuter and Kohler (2009) propose the use of sequence analysis as a means to generate nonresponse adjustment variables \( z \) from rich call record data. We extend their work in an application to sequences of interviewer calls from Wave One of The Irish Longitudinal Study of Ageing. By testing the sensitivity of the derived \( z \) variables to the algorithm input settings and by incorporating time of call information into the sequences, we show that the properties of the derived \( z \) variable can be influenced by the researcher. Comparing unadjusted and weighted survey estimates to gold-standard Census figures, we show that the useful household information contained in call record sequences does not exceed that captured by time-invariant or call aggregated paradata.
1. **Introduction**

This paper explores the use of sequence analysis of call record data as a method to generate variables for nonresponse adjustment purposes. Previously proposed by Kreuter and Kohler (2009), sequence analysis classifies household based on the series of interviewer calls made during data collection. We extend the work of Kreuter and Kohler in three ways. First we reflect on the sequence analysis algorithm in order to determine how best to apply it in a fieldwork context. Second, we perform a sensitivity analysis on the effect of the algorithm input values on the extracted variables. Third, we incorporate information about call timing in the sequences. The aim of these extensions is to produce variables which are available for all households and which maximise the correlation with both the probability of response and with key survey variables. More fundamentally, we are interested in the suitability of these sequence-derived variables for nonresponse adjustments. We develop a series of nonresponse weights incorporating the variables from our sensitivity analysis and compare the unadjusted and weighted survey estimates. We also compare to census estimates for these variables so that the extent of nonresponse bias reduction can be gauged.

2. **Previous Research Using Call Record Data**

In a survey environment characterised by low and declining response rates (Battaglia et al 2007; de Leeuw and de Heer 2002), nonresponse bias remains a threat to the validity of statistical analyses. Fieldwork efforts such as incentivisation (Jäckle and Lynn 2007; Singer et al 1999), targeted inducement (Lynn 2013) or responsive designs (Groves and Heeringa 2006) diminish the potential for bias by increasing the response rate, or more importantly, decreasing the respondent – nonrespondents gap for key survey estimates. However extended fieldwork efforts are costly, and do not guarantee bias reduction (Hall et al 2011). An alternative tactic is to employ post-survey adjustments which reweight estimates to reflect differential response rates across subgroups (Bethlehem, Cobben and Schouten 2011). Kalton and Flores-Cervantes (2003) describe several weighting approaches. Common to each is some auxiliary measure \( z \), available for each sampled unit and correlated with both the response outcome \( R \) and survey variables \( y \) (Little and Vartivarian 2005; Sarndaland Lundstrom 2008; Blom 2009; Kreuter et al 2010).

In the search for post-survey adjustments recent attention has turned to paradata. Paradata is ancillary information that is collected in conjunction with the main survey, such as observations on the sampled unit, call records and respondent utterances (Olson 2013; Kreuter and Casas-Cordero 2010; Kreuter et al 2010; Blom 2009). The availability and richness of paradata have made them an attractive hunting ground for \( z \) variables (Olson 2013, Kreuter et al 2010). However, findings to date on their usefulness for nonresponse adjustment have been mixed. While Blom (2009) demonstrates that including weights based on interviewer observations reduced relative bias in estimates from the European Social Survey (ESS), this result was limited when post-stratification weights were utilised, suggesting that paradata did not add much beyond more traditional sample frame variables. Kreuter et al (2010) present three examples of surveys which recorded auxiliary information on householders in order to generate suitable adjustment variables. In each instance, the \( z \) variables were selected to reflect the subject matter under study to ensure a correlation between \( z \) and \( y \). For
example in the National Survey of Family Growth, which focuses on fertility and family life, interviewers recorded a guess on respondents’ sexual activity and evidence of children in the household. However the weights generated from response models incorporating these variables produced no appreciable impact on key survey estimates. This result is ascribed to low correlations between z and R.

Call record data are a particularly rich subcategory of paradata which present a useful area for exploration. These are observations at a call level such as the time and outcome of call. In order to analyse longitudinal data of this type suitable analytical techniques are required. One option is event history analysis, which analyses household outcomes on a call-by-call basis (Durrant et al 2011). An alternative approach proposed by Kreuter and Kohler (2009) and also explored by Pollien and Joye (2011) is sequence analysis (SA), a tool for visualising and summarising sequential or longitudinal data (Abbot and Tsay 2000). While the event history model employed by Durrant et al (2011) focuses on individual call level outcomes, the sequence approach seeks to extract information captured in the entire series of calls made to a household (Kreuter and Kohler 2009). This is an attractive alternative and it is easy to imagine how the full sequence of events may be more revealing than summary statistics such the number of calls or the proportion of noncontacts. For example, a successful contact preceded by three noncontacts suggests a different dynamic than a successful contact followed by three noncontacts. SA offers the potential to unlock important information which is lost when aggregating across all contacts at a household.

Examining sequences of household level call outcomes collected from three waves of the European Social Survey, Kreuter and Kohler (2009) show that the extracted z variables were predictive of the response outcome but not correlated with key survey variables. Pollien and Joye (2011) apply a similar methodology to call records from the ESS in Switzerland with comparable results. Using the SA algorithm in combination with cluster analysis, a ten-category typology of contact sequences is generated. Many of the extracted subgroups are characterised by particularly high or low response rates, but are less predictive of survey variables. It is unsurprising that z variables generated through SA are most predictive of cooperation given the nature of call records which are “characteristics of the fieldwork process and thus, by their very nature, related to nonresponse” (Blom 2009: 5). Sequence analysis of call record data may yet be useful for nonresponse adjustment if the method can be adapted to produce summary measures which maintain the power to predict response but are also related to the survey outcomes.

3. An Introduction to Sequence Analysis

3.1 States, Alphabets and Distance Metrics

Sequence analysis (SA) is a descriptive technique which finds patterns in complex data by grouping together individuals or units which have a similar composition or trajectory. The basic structure of SA can be described as follows. To begin, data is represented as a sequence of mutually exclusive, exhaustive states drawn from a finite alphabet. A metric is defined which quantifies the distance between any pair of sequences. Applying this metric to all possible pairwise combinations of sequences produces a distance matrix which can be analysed using traditional multivariate
Sequence Analysis of Call Record Data

techniques such as cluster analysis or multidimensional scaling. The resulting classifications or
dimensions may in turn be used as predictive variables in further analyses.

States are the individual elements or events which string together to create sequences. The
researcher must decide upon a set of codes, or an alphabet, which reflect the domain under
investigation. This alphabet must be finite and the states mutually exclusive. An intuitive example of
a state alphabet is the set of all possible labour market activities adopted by McVicar and Anyadike-
Danes (2000) in their study of transitions between school and work. Their classification comprises
six categories (jobless, training, employment, further education, higher education and school) and
reflects the spectrum of possible labour market statuses for their sample of young people. In the
context of call record data from household surveys we are interested in sequences of visits that an
interviewer makes to a sampled household. Previous applications have focused on one characteristic
of each visit: the outcome of the call. Thus an alphabet of call outcomes must cover all possible
results of an interviewer’s visit. A simple alphabet might consist of the following four states:
Noncontact (N), Contact but no interview (C), Refusal (R), Interview (I).

In order to quantify the dissimilarity between two sequences a distance metric is required.
By far the most common approach to generate a metric is optimal matching analysis (OMA). The
OMA distance metric is defined in terms of the number of operations required to transform one
sequence into another. The transformation is achieved through a combination of edits to the states
of a sequence. There are two types of edits possible: substitutions where one element of a sequence
is directly swapped for another, or insertions and deletions 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 compared, insertions and deletions are effectively
equivalent and are referred to collectively as "indels". A cost or penalty is associated with each
substitution or indel required to match two sequences. Summing these costs forms the basis of the
distance between two given sequences.

As an example, consider the following two sequences drawn from the simple state alphabet
suggested above.

Household 1: NNNR
Household 2: NCI

At household 1 three noncontacts are followed by a refusal. At Household 2 a noncontact is followed
by a successful contact and then interview.

To quantify the distance between these households using OMA we count the number of
operations needed to transform one sequence of call outcomes to the other.

Example of Distance Calculation

N N N R We begin with the sequence at Household 1
N [C] N R “Contact” is substituted for a “Noncontact” in position two.
N C N R “Noncontact” in the third position is deleted
N C [I] “Interview” is substituted for “Refusal” in the final position
The transformation is completed with two substitutions and one indel. Appropriately weighting these operations will return the raw distance between the two sequences, effectively a measure of their dissimilarity.

The above example is somewhat simplified. Before calculating the distance between a pair of sequences their respective elements must be optimally aligned to return the minimal distance, that is, the “cheapest” combinations of substitutions and indels to complete the transformation. This optimal alignment is not necessarily obvious but can be easily obtained using dynamic programming methods (Needleman and Wunsch 1970). Of course, the cheapest alignment of two sequences will be dictated by the weights, or costs, assigned to the substitution and indel operations. These costs are the input values under the researchers control and as we will see, their assignment plays a fundamental role in the method.

3.2 Substitution and Indel Costs

Substitution and indel costs, and importantly, their interaction, impact directly on the distance metric produced through OMA (Bison 2009; Lesnard 2010). The setting of substitution and indel costs has been a contentious issue throughout the development of sequence analysis in the social sciences (Aisenbrey and Fasang 2010; Abbott and Tsay 2002; Levine 2000; Wu 2000; Stovel et al 1996). Typically, some theory or insight into the system under study can be used to make an informed judgement on the choice of sensible costs. An alternative is to determine the costs empirically; either based on observed transition rates between states, or on the observed frequency of co-alignment of states (Gauthier et al 2009).

Before selecting a method to define edit costs an understanding of the role of substitutions and indels in the alignment process is desirable. Substitution costs and indel costs operate in different ways. Substitutions allow for the swapping of elements, thus the magnitude of the cost typically reflects the relative similarity between the elements to be swapped. Indel costs, on the other hand, are closely linked to temporality (Lesnard 2010). Insertions stretch or decelerate time, while deletions imply a compression or acceleration. The role of indels is particularly pertinent in the alignment of sequences of different length, where as many indels as the disparity in the length of the sequences under alignment will always be required. Standardisation, which rescales the distance between two sequences by dividing by the length of the longest of the pair, is also useful in this situation (Abbott and Hrycak 1990).

While the setting of substitution costs has traditionally received the most attention in the literature, more recent work has focussed on the role of indel costs in finding patterns in sequential data (Halpin 2008, Bison 2009; Hollister 2009; Lesnard 2010). Using a simple simulation, Bison (2009) identifies that it is the ratio between substitution and indel costs, rather than the patterns in the data, which dictate the results obtained through OMA. This study has two important implications: firstly, it is not the indel cost that is important but rather the magnitude of the indel cost relative to substitution cost(s) and, secondly, cost settings impose an “artificial reality” (Bison 2009: p65) on the sequences under study. Consequently, in order to maintain what Halpin (2009) refers to as the “sociological meaningfulness” of sequences, careful consideration must be given to these settings.
3.3 Combining Time-of-Call Information

The components discussed so far have all been with reference to the situation where one outcome, or domain, is under analysis. It is often preferable to analyse multiple domains simultaneously, allowing the broader context of a transition to be considered (Gautier et al 2010; Pollock 2007). In the context of call record data, Kreuter and Kohler (2009) suggest combining information about both the outcome and the time of the call as a means of producing richer sequences. As previously mentioned there is existing evidence of an interaction between contact and time of call and furthermore that this relationship varies by the profile of the household (Durrant et al 2011; Lipps 2010). Thus we would hope that to some extent households may be characterised by examining the patterns of call timing together with the outcome of the calls.

Gauthier et al (2010) propose the use of Multi-Channel Sequence Analysis (MCSA) as a method to analyse contemporaneous social trajectories. Adopted from the field of bioinformatics, this approach combines multiple sequences of information for each individual optimal alignment process. A separate sequences alphabet is defined for each strand of information, or “domain”. The authors show that the method produces concise, informative typologies.

4. Sequence Analysis in a Fieldwork Context

4.1 Appropriate Cost Settings

The social science literature on sequence analysis teaches us that substitution and indel costs play a key role in the output produced from sequence analysis. Bearing this in mind we now examine the costs used in previous applications to call record data. The OMA derived z variable presented by Kreuter and Kohler (2009) relies on a substitution cost of 2 and an indel cost of 1. Pollien and Joye (2011) employ the transition rate method to generate substitutions costs which do not exceed the value of 2. Their indel cost was also set at 1. Here we argue that neither approach is ideal.

We begin with the settings employed by Kreuter and Kohler. This approach assigns the same cost to the substitution of any two states. The implication here is that all of the defined states are equally similar to each other from a substantive perspective. This is a difficult assumption to justify. While we may view contact with a respondent as reasonably similar to contact with someone else, the same cannot be said comparing the outcomes of noncontact and interview, which are clearly meaningfully different from a nonresponse perspective. Another issue arises from setting the indel cost to exactly half the substitution cost. In this scenario the substitution of any two states can equivalently be achieved by a deletion followed by an insertion. This is what Hollister (2009: 13) describes as a “Pseudo Substitution”. Allowing the algorithm to align any two states in this manner effectively ignores much of the substantive information contained in the call record data.

The transition based costs employed by Pollien and Joye avoid pseudo-substitutions as the substitution costs are always strictly less than twice the indel costs. Nonetheless, we argue that this
Sequence Analysis of Call Record Data

approach is inappropriate in the context of call records. To support this stance we will briefly describe the method. The transition costs approach assigns a low substitution cost to states which are frequently observed contiguously. So if across all the available sequences, State A is frequently followed or preceded by State B, a low substitution cost will be assigned between A and B, implying they are substantively similar. This is inappropriate in the context of call records for two reasons. First, unlike life course sequences such as careers or marital status, sequences of call outcomes have less inherent structure. While Divorced must necessarily be preceded by Married or Separated in sequences of marital statuses, there is no such hierarchy in call outcomes. Most outcomes can happen at most visits regardless of what has come before. Secondly, and more importantly, in many contexts transitions do not reveal anything in terms of meaningful similarity or dissimilarity between two states. While transitions between marriage and divorce may be frequent, the state of being married is clearly different to that of being divorced. In the context of call outcomes, a Noncontact household will frequently experience a transition into being a contacted one, but of course this does not necessarily represent anything about the distinction between contacted and uncontacted households.

For the reasons discussed above we opt for user set substitution costs over those derived empirically from the data. While this approach is subjective, the costs can be selected such that the resulting distance measure reflects substantive differences in sequences of calls. Sets of substitution costs are traditionally presented as a $n \times n$ diagonal cost matrix, where $n$ is the number of states defined in the alphabet. Before deciding on cost settings, we must query what the underlying construct we wish to measure is in the context of call record data. An attractive option is that invoked by Pollien and Joye (2011), who generate their call outcome alphabet based on two constructs, the ease of access to a sampled individual and their willingness to be interviewed. This appropriately reflects a key tenet of nonresponse research, that noncontact and refusal are distinct mechanisms and bias samples in different ways (Lynn and Clarke 2001). Adopting this perspective, our sequence distance measure should effectively differentiate between households which are easily contacted and those which are hard-to-reach, and also between reluctant and cooperative households. In this context, the role played by sequence length is significant. The sequence length, or total number of calls to a household, may reflect either or both of the above dimensions. That hard-to-reach households generate long sequences as a result of multiple call-backs is obvious. In terms of willingness to be interviewed, evidence and theory suggest that effective interviewers will disengage from negative interactions and return at a later date, in order to avoid an outright refusal (Lynn and Clarke 2001; Groves and Couper 1998; Morton-Williams and Young 1987) resulting in a larger number of calls at reluctant households. Consequently, the indel cost must be set reasonably high, such that the substantive difference between long and short sequences is reflected in the distance metric.

Relying on indels alone, it is not easy to disambiguate between long sequences caused by noncontact and long sequences caused by reluctance. This aspect can be drawn out by applying substitution costs which reflect differences in these dimensions. Substituting a noncontact with a call where contact is made should incur a relatively high penalty, as should substituting a call where a positive outcome is achieved with one where a negative outcome is achieved. This approach in turn relies on an appropriate coding of the states. The defined states must first distinguish between calls
resulting in contact and those where no contact is made. Secondly, in the case of successful contacts, positive and negative outcomes should be differentiated.

Another important consideration is the ratio between indel and substitution costs (IS ratio). As discussed in Section 2, this relationship dictates to what extent the temporality of sequences is maintained in the optimal alignment. In order to avoid pseudo-substitutions, which ignore the contextual meaning of the sequence elements, it is desirable to set the indel cost to greater than half the highest substitution cost. As discussed above, a relatively high indel cost is also necessary to reflect the meaningfulness of different sequence lengths. However care is needed as setting a high ratio (i.e. indel costs many times greater than substitutions costs) would produce a distance metric which is dominated entirely by disparity in sequence length. In other words, the proportion of the distance between two sequences due to their dissimilarity in length would dwarf the proportion penalising the substantive differences in the sequence elements - the call outcomes.

To summarise, three principles should be adhered to when selecting a cost scenario in the context of call outcome sequences. First, substitution costs should vary to distinguish between calls resulting in contact and noncontact and between positive and negative call outcomes. Second, the indel cost should be relatively high to preserve the meaningfulness of sequence length. Third, the ratio between indel and substitution costs should be moderated such that sequence length does not become the dominating driver of the distance metric. Within these guidelines it is still difficult to decide exactly what cost settings will be most appropriate. As Wu (2000) correctly asserts, “even the most mathematically inclined” would have difficulty in predicting how changes in substitution costs will change the sequence analysis output. To understand this relationship, Abbott and Tsay (2000) suggest the use of sensitivity analyses; that is examining the variation in distance structure under different substitution and indel costs. To some extent we must rely on this approach, as well as the interpretability of the output produced, to validate our cost settings decision.

4.2 Incorporating Time-of-Call Information

The above section outlines how sequences of call outcomes may capture the underlying dimensions of accessibility of a household and their willingness to participate in the survey. However we are not interested in these response outcomes alone. This is especially true given our aim to extract variables which correlate with household characteristics as well as the response outcome. For this reason we extend our sequences to include information about the time of call, as suggested by Kreuter and Kohler (2009). This extension is motivated by existing research which suggests an interaction between household characteristics and the optimal call time (Durrant et al 2011; Lipps 2010). Thus the calling strategy adopted by the interviewer may reveal something inherent about a household’s profile. The results of Durrant et al (2011) reveal that pensioner households are easier to contact during the daytime on weekdays and less so during the weekend. The probability of contact was also reported to be significantly lower between 12pm and 5pm on weekdays at households with at least one adult in employment. On the other hand, households containing an occupant with a limiting long-term illness were significantly more likely to be contacted during these times, as were households where the dwelling was described as being in a poor state of repair.
If our hypothesis regarding the relationship between calling strategy and household characteristics holds true, then call sequences which are dominated by weekday afternoon calls will be focused on households containing more retired and physically limited occupants, while households predominately visited on evenings and weekends will contain employed and healthier residents. The underlying dimension here is part economic activity and part activeness in general. Thus, we will differentiate between these important timing slots, both in the construction of the time alphabet and through appropriate substitution costs.

5. Data and Methods

5.1 Data

We analyse call record data gathered during the first wave of The Irish Longitudinal Study of Ageing (TILDA). TILDA is a prospective study of the residential population over the age of fifty living in the Republic of Ireland. As a large scale, interviewer-mediated, face-to-face household survey, this study is similar in design to the ESS paradata previously analysed with SA. Call record information was recorded by interviewers on a pen-and-paper interviewer contact sheet similar in design to the contact form employed in the ESS (Stoop et al 2003). The call record dataset comprises contact history information for over 24,000 addresses, approached by interviewers between October 2009 and February 2011. Ineligible sampled addresses (non-residential address/no occupant over the age of 50) were removed and the contact information was cleaned (duplicate entries were dropped as were contact attempts beyond the tenth call). Following this preparation, 41,353 contacts events nested within 10,265 households were available for analysis. The number of visits to a household ranged between one and ten and the modal number of visits was four.

At each visit the interviewer recorded the exact time and date, and indicated one of twenty-five outcome categories specified on the contact form. Motivated by the discussion in Section 3, outcome categories were divided into five states, differentiating between contact and noncontact, and degrees of amenability within contacts (Table 1). Time slots were characterised as Morning (Mon – Thurs, 9am – 12pm), Afternoon (Mon – Thurs, 12pm – 6pm), Evening (Sun – Thurs, 6pm+) and Weekend calls (Fri – Sun 6pm). Interviewer calling patterns indicated that Fridays were more similar to weekends than weekdays, while Sunday evenings were more similar to weekday evenings. This latter observation also reflects distinctions made by Durrant et al (2011).
Table 1. Call outcomes and State Alphabet for the Outcome Domain

<table>
<thead>
<tr>
<th>Outcome Category</th>
<th>N</th>
<th>%</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>No contact/still chasing</td>
<td>15,642</td>
<td>37.8</td>
<td>Noncontact N = 16,621 (40.2%)</td>
</tr>
<tr>
<td>Occupied no contact at address after 5 calls</td>
<td>342</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>Unable to access block / apartments</td>
<td>283</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>Occupier in but not answering after 5 calls</td>
<td>143</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Property not found</td>
<td>114</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Unsure if occupied, no contact</td>
<td>75</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Property vacant</td>
<td>12</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Non-residential property</td>
<td>4</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Unable to access block / apartments</td>
<td>283</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>Occupier in but not answering after 5 calls</td>
<td>143</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Property derelict/demolished</td>
<td>3</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>House hold refusal</td>
<td>4,959</td>
<td>12.0</td>
<td>Refusal N = 5,177 (12.5%)</td>
</tr>
<tr>
<td>Individual refusal</td>
<td>218</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Some contact but no appointment</td>
<td>6,373</td>
<td>15.4</td>
<td>Neutral N = 8,377 (20.3%)</td>
</tr>
<tr>
<td>Appointment broken</td>
<td>631</td>
<td>1.50</td>
<td></td>
</tr>
<tr>
<td>Too ill to participate</td>
<td>577</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>361</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Contact made but unable to assess eligibility</td>
<td>150</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Away during fieldwork</td>
<td>90</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>No one aged 50+ in the household</td>
<td>70</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Withdrawn by head office</td>
<td>47</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Mother tongue require</td>
<td>45</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Partial interview - Refused to continue</td>
<td>33</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Partial interview - to be completed</td>
<td>104</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>One half of an eligible couple cooperated</td>
<td>229</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Appointment made</td>
<td>5,030</td>
<td>12.2</td>
<td>Positive N = 16,621 (13.0%)</td>
</tr>
<tr>
<td>Successful interview</td>
<td>5,815</td>
<td>14.1</td>
<td>Cooperation N = 5,815 (14.1%)</td>
</tr>
<tr>
<td>Total</td>
<td>41,353</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>
5.2 Cost Settings

The domains of call outcome and call time were analysed individually to identify suitable cost settings before combining information from both domains using MCSA. Within each domain, three substitution cost matrices and three IS ratios were cross-classified to generate nine cost scenarios per domain (Table 2).

Table 2. Cross-classification of Substitution and Indel Costs

<table>
<thead>
<tr>
<th>Call Outcome Domain</th>
<th>Time of Call Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Substitution Cost Matrix</strong></td>
<td><strong>Substitution Cost Matrix</strong></td>
</tr>
<tr>
<td><strong>Constant Cost Matrix</strong></td>
<td><strong>Constant Cost Matrix</strong></td>
</tr>
<tr>
<td>( N_c )</td>
<td>( k = 1 )</td>
</tr>
<tr>
<td>( R_f )</td>
<td>Indel Cost = 1</td>
</tr>
<tr>
<td>( N_t )</td>
<td>Indel Cost = 1 5</td>
</tr>
<tr>
<td>( P_o )</td>
<td>( k = 4 )</td>
</tr>
<tr>
<td>( C_o )</td>
<td>( k = 4 )</td>
</tr>
<tr>
<td>Indel Cost = 1 5</td>
<td>Indel Cost = 3</td>
</tr>
<tr>
<td>Average off-diagonal value = 2</td>
<td>Average off-diagonal value = 2</td>
</tr>
<tr>
<td><strong>Moderate Cost Matrix</strong></td>
<td><strong>Moderate Cost Matrix</strong></td>
</tr>
<tr>
<td>( N_c )</td>
<td>( k = 7 )</td>
</tr>
<tr>
<td>( R_f )</td>
<td>( k = 7 )</td>
</tr>
<tr>
<td>( N_t )</td>
<td>( k = 7 )</td>
</tr>
<tr>
<td>( P_o )</td>
<td>( k = 7 )</td>
</tr>
<tr>
<td>( C_o )</td>
<td>( k = 7 )</td>
</tr>
<tr>
<td>Average off-diagonal value = 3 6</td>
<td>Average off-diagonal value = 3 6</td>
</tr>
<tr>
<td><strong>High Cost Matrix</strong></td>
<td><strong>High Cost Matrix</strong></td>
</tr>
<tr>
<td>( N_c )</td>
<td>( k = 10 )</td>
</tr>
<tr>
<td>( R_f )</td>
<td>( k = 10 )</td>
</tr>
<tr>
<td>( N_t )</td>
<td>( k = 10 )</td>
</tr>
<tr>
<td>( P_o )</td>
<td>( k = 10 )</td>
</tr>
<tr>
<td>( C_o )</td>
<td>( k = 10 )</td>
</tr>
<tr>
<td>Average off-diagonal value = 3 6</td>
<td>Average off-diagonal value = 3 6</td>
</tr>
</tbody>
</table>

**Note:** \( N_c \) = Noncontact, \( R_f \) = Refusal, \( N_t \) = Neutral, \( P_o \) = Positive, \( C_o \) = Cooperative

Within the call outcome domain, the first scenario \( (k = 1) \) employs the common default cost settings: a constant substitution cost of 2 and an indel cost of 1 (Brzinsky-Fay et al 2006; Gabadhino et al 2011). Maintaining the same constant substitution costs, the indel cost is increased to 3 and 6 (or in other words the IS ratio is increased from 0.5 to 1.5 then 3) respectively producing the
scenarios \( k = 2 \) and \( k = 3 \). The substantive interpretation of these increases is a stronger weight being placed on the sequence length. That is, as the IS ratio increases, the distance between two sequences returned by the OM algorithm will increasingly depend on the disparity in their length.

The systematic variation in the IS ratio is repeated for moderate and high substitution cost matrices, generating scenarios \( k = 4 \) to \( 6 \) and \( k = 7 \) to \( 9 \) respectively. The purpose of the variable substitution costs is to distinguish between close and disparate call outcomes, as outlined in Section 3. In the moderate substitution matrix, the cost for replacing a “Refusal” call with a “Neutral” call is reduced to 1, as is the substitution cost between “Positive” and “Cooperation” outcomes, reflecting their relative similarity. In the same vein, the cost of replacing “Refusal” with “Cooperation” is increased to 3, reflecting the dissimilarity between these outcomes. The replacement cost between “Noncontact” calls and all other outcomes is maintained at 2. The second user-set cost matrix further inflates the substitution costs. Note that, the IS ratio is relative to the average off-diagonal substitution cost for each matrix.

The cross-classification of three substitution cost matrices and three IS ratios is repeated in the time-of-call domain to generate nine further cost scenarios (\( k = 10 \) to \( 18 \)). Here the interpretation of the IS ratio is the same as in the outcome domain; higher ratios place a stronger weight on the length of the sequences over the sequence elements. The substitution costs differentiate between calls made during the morning, afternoon, evening and weekends.

5.3 Analysis

Under each of the cost scenarios presented above, the sequence data was submitted to OMA and the resulting distance matrix was summarised using cluster analysis with Ward's linkage (Bartholomew et al 2002). Single domain analysis was carried out using the SQ-Ados in STATA (Brzinsky-Fay et al 2006). We selected a four cluster solution in each instance in order to facilitate comparison and because exploratory analysis indicated that four groups were sufficient to generate distinct, interpretable subgroups of suitable size. Adopting the terminology of Pollien and Joye (2011), we refer to these cluster solutions as typologies. We view these as auxiliary \((z)\) variables and therefore denote the typology generated from the cost scenario \( k \) as \( z_k \).

In order to understand the impact of cost settings on the generated typologies we present the Cramér’s V measure of association with the number of calls \((N)\) and the underlying construct of interest in each domain. For sequences of call outcomes this is a binary indicator for cooperation \((C)\). For sequences of call times, an ordinal measure of the proportion of weekend and evening calls \((W)\) is used. The Cramér’s V statistic is a measure of association between two nominal variables. Ranging from 0 to 1, 0 indicates no association while 1 indicates perfect association. Selecting the most effective costs scenarios, the two domains of call outcome and call time were combined using MCSA. MCSA was performed using the TraMineR package in R (Gabadinho et al 2011). Three MCSA analyses were performed, one giving equal weight to both domains, and two which moderated the weights placed on the time and outcome domains, first to 2:1 and then 1:2.

Our final aim is to quantify the added value of the sequence derived \(z\) variables in terms of nonresponse bias reduction. To begin we define Cooperation \(C_{ij} = 1\) if household \(i\) in
neighbourhood $j$ cooperates, and 0 otherwise. We model the probability of cooperation $\pi_{ij} = \Pr(C_{ij} = 1)$ using a multilevel logistic model incorporating traditional paradata variables $x$ only (1). The multilevel structure consists of 10074 households nested within 634 neighbourhoods.

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = X'_i \beta_x + u_j$$  \hspace{1cm} (1)

where $u_j \sim N(0, \sigma_u^2)$

The vector $X'_i$ is a set of interviewer observations and aggregated call record statistics (Table 3) while $u_j$ denote the neighbourhood fixed effects. Next, we repeatedly refit the model including the sequence-derived auxiliary variables $Z_{k}$, ($k = 1 . . 21$).

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = X'_i \beta_x + Z'_{(k)ij} \beta_z + u_j$$ \hspace{1cm} (2_k)

where $u_j \sim N(0, \sigma_u^2)$

Inverse probability weights were generated by dividing the predicted probability of response from each model into quintiles, and setting the weight for each quintile to the inverse of the mean response probability within that quintile. The weight generated from the nonresponse model containing only traditional paradata (Model 1) is denoted $w_x$. The weight generated from the model incorporating the typology $Z_k$ is denoted $w_{Z_k}$. We compare the shift in estimated proportions for six TILDA variables using each of these $k + 1$ inverse probability weights. The survey variables which were selected for comparison were also measured in the Irish 2011 Census, allowing the potential nonresponse bias reduction to be gauged (Table 3).
### Table 3. Summary of Paradata and Survey Variables used in the Analyses

<table>
<thead>
<tr>
<th>Variable</th>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Traditional Paradata (X Variables)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Calls</td>
<td>Ordinal</td>
<td>1 - 2, 3 - 5, 6 - 7, 8+</td>
</tr>
<tr>
<td>Contact Proportion</td>
<td>Scale</td>
<td>The proportion of calls at a household where contact was successful</td>
</tr>
<tr>
<td>House type</td>
<td>Binary</td>
<td>Detached home, incl. farm (1) other house type (0)</td>
</tr>
<tr>
<td>State of area around the household</td>
<td>Ordinal</td>
<td>Physical state of the buildings in the area</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Very good (1), Good, Satisfactory, Bad, Very Bad (5)</td>
</tr>
<tr>
<td>Dublin Indicator</td>
<td>Binary</td>
<td>Household in Dublin (1) or outside (0)</td>
</tr>
<tr>
<td><strong>Survey Outcomes (Y Variables)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degree</td>
<td>Binary</td>
<td>Highest level of education</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tertiary (1); Primary or Secondary (0)</td>
</tr>
<tr>
<td>Sick/Disabled</td>
<td>Binary</td>
<td>Principle economic status</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unable to work due to permanent sickness or disability (1); Other (0)</td>
</tr>
<tr>
<td>Home/Family Care</td>
<td>Binary</td>
<td>Principle economic status</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Looking after home or a family member (1); Other (0)</td>
</tr>
<tr>
<td>Single</td>
<td>Binary</td>
<td>Marital status</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single, never married (1); Other marital status (0)</td>
</tr>
<tr>
<td>Separated/Divorced</td>
<td>Binary</td>
<td>Marital status</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Separated or Divorced (1); Other marital status (0)</td>
</tr>
</tbody>
</table>
6. Results

6.1 Impact of Cost Settings

Figure 1 below presents Cramér’s V measure of association between the typologies $z_k$ and $N$ and $C$ for the call outcome domain and $N$ and $W$ for the time-of-call domain. To understand Fig. 1, consider the upper plot focusing on the outcome domain. When $k = 1$ (constant substitution costs and an IS ratio of 0.5) the Cramér’s V measure of association between the extracted typology $z_1$ and Cooperation ($C$) is $V(z_1, C) = 0.65$. The relationship between $z_1$ and the Number of Calls ($N$) is $V(z_1, N) = 0.35$. Maintaining the same substitution costs and increasing the IS ratio to 1.5 gives $V(z_2, C) = 0.34$ and $V(z_2, N) = 0.56$. Increasing the IS ratio to 3 gives $V(z_3, C) = 0.18$ and $V(z_3, N) = 0.65$.

The impact of increasing the IS ratio is clear. As the size of the indel cost increases relative to the average substitution cost, the resulting typologies are driven more by the sequence lengths (i.e. number of calls) than by the information contained in the sequence states. The impact of user-set versus constant substitution costs on the typologies is weaker, although the effect is generally in the direction we would expect. Increasing or reducing costs to reflect substantive differences and similarities between the sequence states increases the measure of association between the
extracted typologies and the underlying constructs of $C$ and $W$. For example in the outcome domain, shifting substitution costs from constant to moderate while maintaining an IS ratio of 1.5 increases the measure of association with the binary cooperation indicator from 0.34 ($k=2$) to 0.58 ($k=5$).  

6.2 Multichannel Sequence Analysis

Having individually examined the separate domains of outcome and time, the two sequences are analysed simultaneously, using multichannel sequence analysis (MCSA). The cost scenarios employed are those identified above which maximised the ability of the OM algorithm to produce distance structures relating to the relevant underlying dimensions. For the outcome domain the scenario combining the most punitive substitution matrix with an IS ratio of 1.5 was selected ($k=8$). For the time domain, the default cost setting was chosen ($k=10$).

Before submitting to MCSA the substitution costs for the outcome domain were rescaled such that the average substitution cost matched the average cost in the time domain (i.e. 2). The indel was adjusted such that the ratio remained at 1.5 (i.e. an indel cost of 3). Within the outcome domain these scaled and un-scaled scenarios are equivalent. The purpose of this procedure is to ensure that when combined in MCSA, the contribution to the substitution costs from the two domains are on the same scale. The weight given to each domain can in turn be adjusted within the MCSA algorithm. Three MCSA distance matrices were generated: one which gave equal weight to both domains, one which weighted up the time domain and one which weighted up the outcome domain, both at a ratio of 2:1. As before, cluster analysis with Ward’s linkage was used to generate three further typologies ($z_k, k = 19, 20, 21$) summarising the distance matrix structure generated under each cost scenario.

The first of the multichannel outputs ($k=19$), which applies an equal weight to call outcome and call time, produces a typology which is strongly related to the Number of Calls ($V(z_{19}, N) = 0.59$) but less so to the substantive outcomes of Cooperation ($V(z_{19}, C) = 0.65$) and Call Timing ($V(z_{19}, T) = 0.26$). Increasing the weight on the time domain produces a measure ($k=20$) with a stronger association with the time of call and a weaker association with cooperation. As would be expected, increasing the weight on call outcome ($k=21$) shifts these associations in the opposite direction.

6.3 Impact on Weighted Estimates

The sensitivity analysis above demonstrates that it is possible to shape the characteristics of the sequence typologies by adjusting the cost settings and combining information from multiple domains. We are more interested in the power of these typologies to reduce nonresponse bias. Figure 2 below presents the weighted estimates for six survey variables plotted as points, with labels indicating the typology $k$ incorporated in the derivation of the weight $w_{zk}$. The baseline estimates of $\bar{y}$ under $w_{zk}$, are highlighted (long dash line). The unadjusted survey estimate (short dash line) and Census 2011 estimate (solid line) are also displayed.
Fig 2. The estimate of $\bar{y}$ weighted by $w_{x_k}$. Short dash line denotes the unadjusted estimate. Long dash line denotes the adjustment under the baseline weight $w_{x_k}$. Solid line denotes the value from the Irish Census 2011.
To understand Fig. 2 consider the first plot in the upper panel which depicts the weighted estimates of the proportion of persons over fifty reporting poor general health. The unadjusted survey estimate (23.9%) underestimates the census figure (27.2%). This bias is reduced when weighted by \(w_x\) and further reduced when weighted by \(w_{z_1}\). The greatest bias reduction is offered by \(w_{z_2}\), which produces an estimate of 26.3%.

This example is encouraging: including sequence typologies in the nonresponse model decreases bias, and the bias reduction is maximised when variable substitution costs are used. However it is clear from Figure 2 that the effect of the weights is inconsistent across variables. For some variables there is very little change in the estimate regardless of the weight employed. Some weights in fact increase bias on certain variables. Weights based on the time sequences (10 – 18) or those incorporating time information through MCSA (19 – 21) generally have little impact on an estimate beyond what is achieved by \(w_x\), suggesting that time-of-call information does not offer any information that is not captured with more traditional paradata. The largest estimate shifts are generally observed under \(w_{z_1}\) and \(w_{z_2}\). These are the household typologies derived from sequences of call outcomes which combine high substitution costs with a constant (7) or moderate (8) IS ratio.

Figure 3 below reveals the effect of adding typology \(z_k\) to the baseline nonresponse model on the correlations with \(C\) and \(Y\). Within each plot, the y-axis displays the correlation between the predicted probability of response and the cooperation indicator \(\hat{\rho}(\hat{r}_{zk}, C)\). The x-axis shows the correlation between the predicted probability of response and the survey variable \(\hat{\rho}(\hat{r}_{zk}, Y)\). The corresponding correlations under the baseline model (Model 1) are included for reference. Incorporating \(z\) variables derived from the call domain with user set substitution costs and a low to moderate IS ratio (\(k = 4, 5, 7, 8\)) increases the correlation with the response outcome. However this increase is frequently at the expense of the correlation with the survey variable.

Many of the sequence derived \(z\) variables, including all of those based on time sequences, have a negligible impact on the correlation between the response propensity and both the cooperation indicator and survey variables. The nonresponse weights which induce the greatest shifts in variable estimates are those which incorporate sequence typologies which increase the predictive power of the nonresponse model. However corresponding decreases in the correlation survey variables limit the capacity to reduce bias.
Fig. 3 The correlation between the Probability of Response and (a) Cooperation (y–axis) and (b) the Survey Variable (x–Axis) when including sequence typology $z_k$ in the nonresponse model. The horizontal and vertical solid lines in the plot region indicate the correlations with $C$ and $Y$ under the baseline models ($\rho(\hat{\pi}_X, C)$ and $\rho(\hat{\pi}_X, Y)$.)
7. Discussion

The availability and richness of paradata have made them an attractive hunting ground for post-survey nonresponse adjustment variables. However, the question of how best to exploit these data remains open. Many approaches focus on time invariant household observations (e.g. presence of children or the presence of litter) or aggregated call records (number of calls or proportion of noncontacts). Individual call level records such as the time and outcome of calls have also been examined, either alone, or included in more sophisticated event history models. The application of sequence methods to these data represents a shift away from individual call outcomes, and towards the entire recruitment process. This shift is motivated by the premise that important information is potentially contained in the full sequence which cannot be captured either by aggregating the call records or analysing them on a call-by-call basis.

Our results broadly align with those of Kreuter and Kohler (2009) and Pollien and Joye (2011) who found that variables derived through sequence analysis of call record data were predictive of response outcomes but less so of survey variables. Beyond this, we have shown that the characteristics of sequence-derived typologies can be shaped by thinking about the how the cost settings and the underlying construction of the sequence states operate. However, while we have begun to understand the relationship between the inputs and outputs of the sequences analysis method, we have been less successful in our ultimate aim of producing auxiliary variables suitable for post-survey adjustment.

In order to extract the maximum amount of information from the TILDA wave 1 call records we adjust the inputs of the sequence analysis algorithm: the substitution and indel costs and the domains to be analysed. Our sensitivity analysis on the effects of these cost settings suggests that to some extent the properties of sequence typologies can be shaped by appropriately setting the substitution and indel costs. Substitution costs which distinguished between substantively different call outcomes produced typologies driven by the household’s amenability to being surveyed. Low to moderate indel ratios ensured that the typologies did not simply become a function of the sequence length, i.e. the number of calls to a household. Thus, in terms of predicting nonresponse, it was possible to extract additional information from the full sequence of calls that was not adequately captured by time-invariant or call-aggregated variables. However, while we were able to improve the correlation between our typologies and cooperation, we were less successful in increasing the correlation with survey variables. Time-of-call sequences were correlated with household characteristics, but when added to the nonresponse model they did not increase the magnitude of the correlation beyond what was achieved from more straightforward paradata variables. So regardless of the cost settings or domain in use, little additional information pertaining to household characteristics could be extracted from sequences of call records from this survey. As a result, the impact of sequence derived typologies on nonresponse bias is limited here.

The collection of complex call records presents a burden to interviewers, and these data are useful only insofar as they can be used to improve surveys; be that 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 many limitations to this study. The quality and other properties of interviewer recorded call records are
poorly understood. Call dynamics which might be present in a general population survey may be lost in this sample of over fifties. Sequence analysis may not be the best technique to extract information from these data which do not have the strong underlying sequential structure that is present in say, career or family trajectories.

Observing that sequence indicators derived from ESS call records were related to response outcomes but not to survey variables, Kreuter and Kohler (2009) suggest that the most useful application may be to fieldwork management. The identification of subgroups to be targeted with specific inducements (Lynn 2013) or differentiating between design phases as part of a responsive design (Groves and Heeringa 2006) are potential application in this sphere. Further applications of sequence analysis, such as analysis of interviewer’s calling patterns are also being explored (Durrant et al 2013; Pollien 2013).
Sequence Analysis of Call Record Data

References


Pollien, A. (2013). *Contact’s strategies: is there only one way that lead to interview?* Paper presented at the 5th Conference of the European Survey Research Association, Ljubljana.


