1

Investigating call record data using sequence analysis to inform adaptive survey designs

Gabriele B. Durrant[1,2], Olga Maslovskaya[2] and Peter W.F. Smith[1,3]

[1] Department of Social Statistics and Demography
Social Sciences
University of Southampton, UK

[2] ESRC National Centre for Research Methods (NCRM)
Social Sciences
University of Southampton, UK

[3] ESRC Administrative Data Research Centre for England (ADRC-E)
Social Sciences
University of Southampton, UK

Word count:
Abstract: 173 words
Main text (without figures, tables and references): 6,791 words
Main text including tables, graphs and references: 8,694 words
Number of references: 55
Number of tables: 2
Number of figures: 5

Address for correspondence:
Professor Gabriele Durrant
Department of Social Statistics and Demography
School of Social Sciences
University of Southampton
SO17 1BJ, Southampton
g.durrant@soton.ac.uk
Tel: +44-(0)23-8059 5481
Fax : +44-(0)23-8059 5763

1   **Investigating Call Record Data Using Sequence Analysis**

2
3
4
5   **Abstract**

6   Researchers have become increasingly interested in better understanding the survey data

7   collection process in interviewer-administered surveys. However, tools for analysing paradata

8   capturing information about field processes, also called call record data, are still not yet fully

9   explored. This paper introduces sequence analysis as a simple tool for investigating such data

10   with the aim of better understanding survey processes and improving survey monitoring.

11   Combining the technique with optimal matching, clustering and multidimensional scaling, the

12   method offers a way of visualising and summarising complex call record data. A novel

13   approach is to use sequence analysis within interviewers, which allows the identification of

14   unusual interviewer calling behaviours, and may provide guidance on interviewer

15   performance. The method may hence be informative for adaptive survey designs and will help

16   to identify unusual behaviour and outliers. The technique is applied to call record data from

17   the UK Understanding Society survey. The findings inform further modelling of call record

18   data to increase efficiency in call scheduling and for the improvement of call outcome

19   predictions.

20

21   **Data Statement**

1 Introduction

2 Many survey agencies nowadays routinely collect survey process data, so-called

3 paradata (Couper, 1998; Kreuter, 2013). For interviewer administered surveys, including both

4 telephone and face-to-face surveys, data about the fieldwork process, often termed call record

5 or call history data, have received much attention (Bates, Dahlhamer and Singer, 2008;

6 Laflamme, Maydan and Miller, 2008; Blom, Jäckle and Lynn, 2010; Durrant, D'Arrigo and

7 Steel, 2011 and 2013b). Such data may contain information about the outcome of the call or

8 visit and the day and time of the call attempts. Several outcomes may be distinguished such as

9 non-contact, contact, ineligible, refusal, appointment made and any interviewing done. This

10 string of outcomes of all call attempts to a household is referred to as a *call sequence*, an

11 ordered collection of activities or states (Piccarreta and Lior, 2010). A large number of

12 different sequences are possible even if the number of positions (i.e. total length) is relatively

13 small. As an example, Table 1 indicates a selection of short call sequences observed in the

14 UK survey Understanding Society (http://dx.doi.org/10.5255/UKDA-SN-6614-5).

15 **Table 1:** Examples of call sequences (taken from Understanding Society).
16

House-hold	Call 1	Call 2	Call 3	Call 4	Call 5
1	Non-contact	Non-contact	Appointment	Interview	-
2	Non-contact	Non-contact	Refusal*	-	-
3	Contact	Non-contact	Non-contact	Non-contact	Non-contact
4	Any other status	Appointment	Non-contact	Non-contact	Non-contact
5	Non-contact	Ineligible*	-	-	-
6	Non-contact	Non-contact	Contact	Appointment	Interview

17
18 * The coding of the call outcome in Understanding Society does not distinguish between refusal
19 and ineligibles. However, for some cases, as possible in this cell, one can use the final response
20 outcome to draw a conclusion about the outcome at a particular call (if ineligible or refusal).
21
22
23

24 This paper introduces the use of sequence analysis for investigating call record data to

25 inform survey monitoring and management processes. The technique is a powerful tool for

visualising and summarising call record data, which can be large and complex (see Background section). It may help to identify unusual cases or outliers, which then may require further investigation. In particular, it can build the basis for further statistical modelling. A novel approach here is to use sequence analysis within interviewers, which allows the identification of unusual interviewer calling behaviours, and may provide guidance on interviewer performance.

Here, sequence analysis plots are combined with optimal matching, cluster analysis, and multidimensional scaling (Kruskal and Wish, 1978; Bartholomew, Steele, Moustaki and Galbraith, 2008; Piccarreta and Lior, 2010). This allows finding similarities across the contact histories and identifying groups of sequences that are homogeneous. The technique allows the identification of unusual calling behaviours and provides survey managers with tools to display such cases. The paper provides some practical guidance on how to analyse call record data using sequence analysis, for example regarding details on how to implement the method into practice including the use of software, how to detect unusual calling behaviours, outliers, unproductive call sequences and coding errors, and highlights possibilities for cost and efficiency savings.

In this paper, the method of sequence analysis is applied to data from the UK Understanding Society survey (Wave 1), a large scale household survey, for which extensive call record data have been collected. Although the method is illustrated using data from a face-to-face survey, it can be employed in the same way to telephone surveys. Other sequences arising in survey methodology may also be analysed with this approach, such as sequences of mouse movements in web surveys (Olson and Parkhurst, 2013).

The remainder of the paper is organised as follows. The following section (Section 2) provides further background and reviews the use of sequence analysis and call record data modelling. The methodology section, Section 3, introduces the basic principles of sequence

1 analysis in the context of survey process data. In Section 4 the method is applied to call

2 sequences from Understanding Society, including a separate analysis per interviewer. The

3 paper concludes with a summary of the main findings, limitations and implications for survey

4 practice.

5

6 **Background**

7   Sequence analysis methods are frequently used in a range of disciplines such as in

8 medicine and biology to detect DNA sequencing (Smith et al., 1986; Miyazawa et al., 1989),

9 but also in the social sciences, for example in demography to study family-life trajectories

10 (Elzinga and Liefbroer, 2007) and life course trajectories (Halpin, 2003; Aassve, Billaria and

11 Piccarreta, 2007; Wiggins, Erzberger, Hyde, Higgs and Blane, 2007; Martin, Schoon and

12 Ross, 2008; Picarreta and Lior, 2010), and in economics to study employment histories

13 (Pollock, Antcliff and Ralphs, 2002; Malo and Munoz-Bullon, 2003) and transitions from

14 education into work (McVicar and Anyadike-Danes, 2002). For a recent review of the method

15 in the social science context see Fasang and Liao (2014). Researchers have also started to use

16 sequence analysis in the context of survey methodology with the aim of informing

17 nonresponse adjustment methods based on call histories (Kreuter and Kohler, 2009; Hanly et

18 al., 2015). Kreuter and Kohler (2009) and Hanly (2013) found that variables derived from call

19 record data - in part informed by sequence analysis - did not improve nonresponse adjustment

20 methods. Although not as useful as hoped, they point out in their discussions, that the method

21 may have the potential for use in survey management field practice. This paper aims to

22 address this.

23   Call record histories were already collected in the time when data were not

24 automatically available through technologies. Durbin and Stuart (1954) discussed the issue of

25 efficacy of call-back strategies. The field of survey practice has changed significantly since

1   then. Technology has driven the data collection process forward so it is now possible to

2   monitor fieldwork processes using visualisation tools and methodologies that allow the

3   identification of patterns and potentially unusual or erroneous interviewer behaviours.

4   A number of authors have already analysed call record data (Groves and Heeringa,

5   2006; Bates et al., 2008; Kreuter et al., 2010a; Wagner, 2013a and 2013b; Durrant et al. 2011;

6   Durrant et al., 2013a and 2013b). Survey agencies hope that analysis of these data may inform

7   best interviewer calling practices and identify more difficult cases as well as unusual

8   interviewer behaviour (Kreuter et al., 2010b; Biemer, Chen and Wang, 2013; Hanly, 2014;

9   Hanly, Clarke and Steele, 2015). They are also interested in increasing efficiency, reducing

10 costs and identifying effective interviewer calling patterns (Bates et al. 2018; Durrant et al.,

11 2011; Lipps, 2011; Purdon, Campanelli and Sturgis, 1999; Swires-Hennessy and Drake, 1992;

12 Weeks, Jones, Folsom and Benrud, 1980; Weeks, Kulka and Pierson 1987). For example, it

13 may be important to identify sequences with a large number of unsuccessful call attempts

14 such as non-contact calls.

15   Although survey researchers have become increasingly interested in understanding

16 data collection processes, it is still often not clear how best to analyse such data, since call

17 record data can be large and may exhibit complex data structures, such as time dependencies

18 and multilevel clustering. For example, calls are made over time requiring time dependent

19 analyses such as discrete time event history analysis; also calls are normally nested within

20 households and cross-classified within interviewers and areas requiring multilevel cross-

21 classified models (Durrant et al., 2011; Durrant et al., 2013a; Durrant et al., 2013b; Hanly,

22 2014; Hanly et al. 2015; Sinibaldi, 2014). Call record data can also be of significantly lower

23 data quality, since these data are regarded as a by-product and often do not undergo the same

24 editing and cleaning checks and may exhibit larger proportions of measurement error and

25 missing data than the actual survey data. Hence they require further cleaning before analyses.

1 Often ad-hoc methods are used and usually summary measures are applied to describe

2 outcomes of call sequences (Groves and Heeringa, 2006; Bates et al., 2008), such as the

3 number of non-contact calls or the total number of calls. As noted in Fasang and Liao (2014,

4 p. 644) sequences of categorical states are much more complex than simple numerical

5 variables and cannot be easily summarized as categorical variables with a limited number of

6 categories. In the survey methods literature it has been recognised more recently that it may

7 be important to analyse the contact sequence as a whole (Kreuter and Kohler, 2009; Hanly et

8 al., 2015). It may be the actual interplay between several call outcomes that is informative. A

9 particular call outcome may have a different meaning if analysed separately or if seen as part

10 of a longer sequence. For example, a non-contact after an appointment may be regarded as a

11 'hidden refusal' whereas otherwise it may simply be interpreted as a period of absence (e.g. in

12 Table 1 compare household 4 with 1 and 6). Consequently, modelling procedures, such as

13 discrete time event history analysis, have been developed to analyse not just the final

14 nonresponse outcome but the process leading to contact, cooperation or refusal as a whole and

15 to recognise the entire contact history (Durrant et al., 2011; Durrant et al., 2013b; Wagner,

16 2013b). However, a full modelling approach may not always be necessary or desirable to

17 analyse sequences.

18 Sequence analysis offers a potentially powerful tool for visualising and summarising

19 call record data and for exploring and reducing the complexity of such data structures. It

20 represents a relatively simple descriptive method which can be easily implemented in

21 practice, not requiring any modelling techniques or distributional assumptions, which should

22 be attractive to many survey agencies. It has the potential to provide survey researchers and

23 field managers with a simple, additional graphical tool to investigate interviewer calling

24 patterns either during or post data collection. Findings from the analysis may then inform

25 future routine monitoring as well as further, more sophisticated statistical modelling.

1   This benefit has already been illustrated in Durrant, Maslovskaya and Smith (2015;

2   2017). They demonstrated the use of the findings from this analysis in survey practice by

3   modelling sequence length and sequence outcome jointly, using a multinomial model thereby

4   improving on standard response modelling, which only models response outcome. In the past,

5   researchers used length of call sequences only as an explanatory variable in nonresponse

6   models. Here, they extended the modelling to include length of call sequences as an outcome

7   variable. Furthermore, their analysis suggests that modelling the joint outcome improves the

8   fit of the model in comparison to the two separate models for either length or final outcome.

9   **Methodology: Using Sequence Analysis**

10   Sequence analysis consists of a series of routines, plots and outputs. Simple sequence plots

11   show the distribution of sequences. Transition rate matrices indicate the propensity for the

12   next outcome options following a particular call outcome. Ideally one would like to find

13   similarities among the contact histories to create groups that are homogeneous in their

14   outcome variables and to summarise the different sequence patterns. The low frequency count

15   of sequences may specify each individual sequence but makes summarising sequences

16   complicated (Piccarreta and Lior, 2010). Given the large number of possible outcomes and

17   patterns no single characteristic that fully describes the sequences will be available. We do not

18   expect an individual summary measure to describe the sequence as a whole. Rather, a

19   combination of multiple characteristics will be necessary to establish an adequate measure of

20   similarity across sequences. In this section, we first describe how a distance matrix can be

21   constructed using optimal matching to order the sequences based on a number of criteria. This

22   distance matrix obtained through optimal matching can then inform cluster analysis and

23   multidimensional scaling. The combination of sequence analysis and multidimensional

1   scaling has been proposed by Piccerreta and Lior (2010) and this approach is extended here to

2   call record data and discussed at the end of this section.

3

4   ***Creating a Distance Matrix through Optimal Matching***

5   A distance matrix is a square symmetrical matrix (i.e. where the element in row i and column

6   j is equal to the element in row j and column i) that indicates distances between a number of

7   pairwise objects (the distance between object i and object j appears in the ith row and the jth

8   column) (Bartholomew et al., 2008). A distance matrix helps to summarise important features

9   of the data and to explore relationships. One way of constructing such a distance matrix is

10   optimal matching (Levenshtein, 1966; Sankoff and Kruskal, 1983; Abbott and Hrycak, 1990;

11   Abbott and Tsay, 2000; Hollister, 2009). This method computes the distance between pairs of

12   sequences by counting the number of basic operations that are necessary to transform one

13   sequence into another (Levenshtein distance). In simple terms, the more operations are

14   required, the larger the differences. Each type of operation may be associated with a weight or

15   cost such that a weighted distance matrix may be constructed. The basic operations

16   considered are insertion, deletion and substitution. The insertion and deletion operations are

17   also referred to as 'indel' operations since a deletion of an element in one sequence is

18   equivalent to the insertion of an element in the other sequence. A substitution operation

19   implies the direct substitution of one element in the sequence with another.

20   It is important to think carefully through the cost settings of each of the operations

21   involved as these costs may have an effect on the resulting distance matrix (Pollock et al.,

22   2002). A typical default option for a substitution cost is 2 and for an indel it is 1, although

23   other cost settings may be advocated (see also Abbott & Tsay, 2000; Wu, 2000; Hanly et al.

24   2015; Hollister, 2009; Pollock et al., 2002). For example, Pollock et al. (2002) calculated

25   costs on the basis of the likelihood of transitions from one employment status to another.

1   According to Abbot and Tsay (2000), the area of cost setting observes real diversity and some

2   theoretical questions about cost settings have not yet been answered. (However, this is

3   outside the scope of this paper.)

4   Let us briefly consider an example of an indel operation and a substitution that

5   transform the call sequence of household 1 in Table 1 into the call sequence of household 2.

6   Both have two noncontact calls (N) in common but household 1 responds with an interview

7   (I) after an appointment (A) whereas household 2 refuses (R) at the first contact-call. By

8   applying one deletion and one substitution we have converted one sequence into the other.

9

10

11	**Household 1:**		**N N A I**
12	Deletion	(‒)	N N A
13	Substitution	(A↔R)	N N R
14	**Household 2:**		**N N R**
15	Resulting distance = 3		

16   The operations are performed between all possible pairs of sequences. This procedure creates

17   a distance matrix with the dimension equivalent to the number of sequences. In total, since the

18   matrix is symmetrical, $(n^2 - n)/2$ distance measures need to be calculated. An advantage is

19   that the method takes account of multiple dimensions or characteristics of the sequences, not

20   just of one or two characteristics as is the case for conventional summary measures of call

21   record data. As indicated, a potential difficulty is the adequate specification of the costs (Wu,

22   2000; Abbot and Tsay, 2000). Results from prior analysis or different datasets may provide an

23   empirical solution. A theoretical solution with the researcher deciding on the costs for the

24   different types of required operations is also possible. For further discussions and options on

25   cost settings see Abbott and Tsay (2000) and Hollister (2009). Optimal matching is an

1   intermediate stage which is necessary to perform sequence analysis techniques such as cluster

2   analysis and multidimensional scaling discussed below.

3

4   *Cluster Analysis and Multidimensional Scaling*

5   The distance matrix can be used to carry out cluster analysis, to find similarities between

6   groups of sequences based on multiple dimensions (Kaufman and Rousseeuw, 1990;

7   Bartholomew et al., 2008; Everitt, Landau, Leese and Stahl, 2011). Different clustering

8   algorithms can be used, for example centroid-based clustering, distribution-based clustering

9   and density-based clustering. It should be noted that in principle different clustering methods

10   may give different results. In this paper we use the optimized 'partitioning around medoids'

11   (PAM) algorithm (Studer, 2013), which differs from hierarchical algorithms and uses a

12   predefined number of k groups to obtain the best partitioning of the dataset. This algorithm

13   was first introduced by Kaufman and Rousseeuw (1997). It aims to identify the k best

14   representatives of groups, called medoids, where a medoid is defined as the observation of a

15   group with the smallest weighted sum of distances from the other observations in that group.

16   The algorithm then aims to minimize the weighted sum of distances from the medoids.

17   It is important to bear in mind that the number of classes generated by cluster analysis

18   is decided by a researcher and is therefore to some extent arbitrary. According to Halpin and

19   Chan (1998), cluster analysis will always provide a solution even when there is no meaningful

20   structure in the data. Pollock et al. (2002) and Piccarreta and Lior (2010) suggest that it is

21   important to inspect different cluster solutions and sequences in clusters closely when

22   deciding on the most optimal solution for cluster analysis and evaluating the quality of a

23   partition. Pollock et al. (2002) argue that the best solution would have substantive meaning

24   which could be ascribed to each cluster membership.

1   Another possibility is to analyse the distance matrix by multidimensional scaling

2   (Kruskal and Wish, 1978; Bartholomew et al., 2008). This is a multivariate technique that

3   aims to reveal the structure of the distance matrix by representing the sequences in a small

4   number of dimensions, such as on a 1-dimensional scale or in a 2-dimensional map or plot.

5   Then, each sequence is identified with a location in that dimension based on the distance

6   matrix. Groups of sequences may be identified based on low, medium or high values in that

7   dimension. In a 2- or 3-dimensional scatterplot, groups of sequences may be identified

8   visually. Part of the interest in the analysis is to try to uncover which attributes of the

9   sequences appear to carry weight in the similarity measure, i.e. which attributes are key

10   features that determine the distances between sequences. The multidimensional scaling scatter

11   plot allows the identification of groups of sequences as well as potential outliers.

12   A combination of the two techniques (cluster analysis and multidimensional scaling)

13   was proposed by Piccarreta and Lior (2010) in order to assure the accuracy of the results

14   obtained through cluster analysis and the visual representation of the clusters. The

15   multidimensional scaling can help to examine graphically the quality of the cluster

16   partitioning and of the cohesion within clusters (Piccarreta and Lior, 2010).

17   **Application of Sequence Analysis to the UK Understanding Society Survey Data**

18   *Design and Fieldwork of Understanding Society*

19   Sequence analysis is applied to call record data from the UK Understanding Society survey

20   data (Wave 1). Understanding Society is the UK Household Longitudinal Study of

21   approximately 40,000 responding households in the United Kingdom, covering topics on

22   health, work, education, income, family and social life to help understand the long term

23   effects of social and economic change as well as policy interventions. The study has many

24   advantages over previously existing datasets in the UK by being exceptionally large and

1   comprehensive. In particular, the study was designed to collect a range of paradata, including

2   call record data. Only interviewers with above average experience and ability were selected

3   for the study.

4   Data collection for each wave is scheduled across a 24 months period, with interviews

5   taking place annually. Wave 1 data collection took place between January 2009 and March

6   2011. All interviews at Wave 1 were carried out face-to-face in respondents' homes by trained

7   interviewers using computer-assisted personal interviewing (CAPI). All adult household

8   members (age 16 and older) were asked to respond. A household also needed to respond to a

9   household questionnaire in addition to all individual interviews. A minimum of six calls were

10   made at each sampled address before it was considered unproductive but interviewers were

11   encouraged to make further calls if possible. Interviewers had one month to contact

12   households allocated to them (McFall, 2012; McFall and Garrington, 2011).

13   The main survey consists of the two main sample components in Wave 1: the general

14   population sample (GP) and the ethnic minority boost sample (EMB). The survey has a multi-

15   stage sample design and households are clustered within interviewers and within primary

16   sampling units (PSUs). The details of the sample selection can be found in Lynn (2009). The

17   study benefits from rich paradata which include call record data and interviewer observation

18   variables (Knies, 2014).

19

20   *Call Record Data in Understanding Society and Analysis Sample*

21   Call record data in Understanding Society contains information about the outcome of calls.

22   The *call outcome at each call*, the key variable of interest here, is recorded by the survey

23   agency as a 5-categorical variable: 'non-contact', 'contact made', 'appointment made', 'any

24   interviewing done' and 'any other status'. A limitation is that the last category combines a

25   range of possible call outcomes, ranging from different types of refusals to different

1   ineligibility statuses. It should be stressed that this coding was provided by the survey agency

2   and hence it is not possible for us to distinguish the different types of categories contained in

3   the 'any other status' categories. The *final outcome* variable recorded at the household level

4   has about 50 outcome codes, split into six broad groups, containing 'ineligible', 'refusal',

5   'contact made but no interviewing', 'any interviewing but not completed' (at least one

6   individual interview completed), 'case completed' (i.e. household questionnaire and all

7   individual interviews from all member of the household have been completed). An additional

8   category in the call outcome variable, 'interviewing process completed', was also created, if it

9   was the last occurrence of the category 'any interviewing done' in the sequence and the final

10   outcome call indicated 'case completed', which means that all household members have

11   responded individually and the household questionnaire has been completed.

12   The analysis sample includes all households from Wave 1 with at least one call,

13   including those which later were classified as ineligible. Cases from the Ethnic Minority

14   Boost sample are excluded as rules for the selection of the boost sample differ from the rules

15   for the main sample (Berthoud, Fumagalli, Lynn and Platt, 2009). Calls with no recorded

16   outcomes are also excluded (1.2% of all calls). The final analysis sample contains a total of

17   255,778 calls with 11,143 distinct sequences, clustered within 47,899 households (including

18   both responding and nonresponding households) and 741 interviewers. The number of non-

19   contact calls is 142,705 calls (representing 56% of all calls) which is relatively high. The

20   minimum length of a sequence is one and the maximum is 30 (mean length 5.34, median 4).

21

22   ***Application: Basic Sequence Plots and Transition Rates***

23   Figure 1 shows a basic sequence plot which displays the sequences across calls for every

24   household, colour coded according to the final outcome of each call. Each horizontal line in

25   the plot represents a call record for one household, i.e. one sequence. About 10% of

1 households experience only one call, and these end primarily in interview or 'any other status'

2 indicating either ineligibles or refusals (as coded by the survey agency). Surprisingly, a small

3 proportion of households experience a contact call with no further outcome or even a non-

4 contact call with no further follow-up visits, which does not show adherence to the

5 interviewer guidelines. This trend continues with the proportion of interviews steadily

6 declining and the longer call sequences being predominantly driven by non-contact calls. Just

7 over 70% of all sequences have a length between 1 and 6 calls. After 10 calls 88% of all

8 households have been completed and after 15 calls this has increased to 98%. About 8% of

9 all call attempts are still being made after the 10th call (20,032 calls). Figure 1 clearly shows

10 that there are a number of households that do not receive the required minimum of 6 calls,

11 although the households have been coded as a contact with no further outcome or a non-

12 contact. This is the case for almost 8% of households.

13 It should be noted that a basic frequency plot such as in Figure 1 suffers from the

14 problem of overplotting giving the large number of sequences. This may hide particular

15 features and, if not carried out carefully, may even be misleading. Solutions have been offered

16 in the literature such as plotting only a subgroup of sequences, unusual cases and outliers and

17 ordering sequences (see Fasang and Liao (2014) for further discussions). We aim to overcome

18 overplotting by plotting subgroups and by ordering sequences.

19 **Figure 1**: Basic Sequence Plot ordered according to length of sequence and outcome of the last
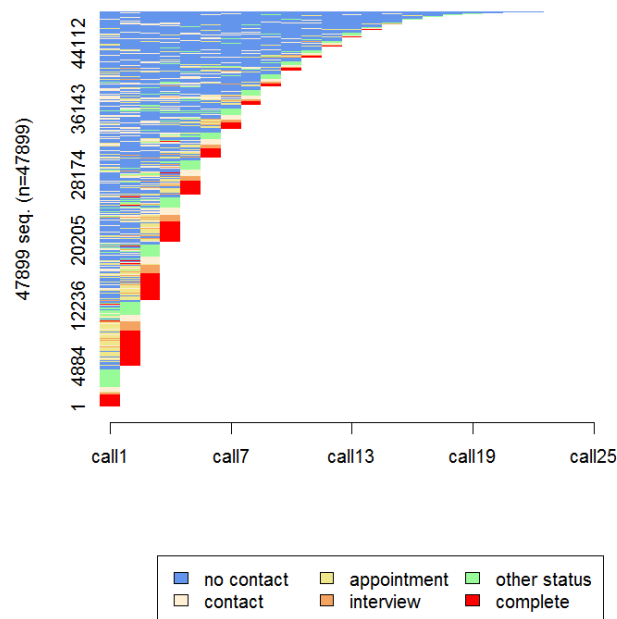20 call (n=47899 call sequences).

21

Figure 2 displays the ten most frequent sequences, sometimes referred to as a sequence frequency plot (Fasang and Liao, 2014). The most frequent sequence with only 6% contains two calls with an appointment made at the first call and complete interview at the next call. The second most frequent sequence (4.5%) contains one call with outcome 'any other status' (ineligible or refusals). Interview at the first call accounts for only 3% of all sequences. Sequences resulting in 'non-contact and interview' or 'non-contact, appointment and interview' account for 6% in total.

Analysing call record data from the European Social Survey (ESS), Kreuter and Kohler (2009) find a high proportion of short sequences resulting in interview at the first call (22%) or contact and interview (18%); third most frequent is immediate refusals (9%), and fourth is no contact then interview (6%). In their analysis, the UK tends to have longer sequences than the other countries in the sample.

A contributing factor in Understanding Society may be the allowance of long call sequences (at least for the first wave where survey researchers are keen to keep as many

1   sample members in the sample as possible) rather than the specification of a maximum

2   number of calls. In our sample the sixth group is also interesting as it finishes with the status

3   'any interview done' but there is no follow-up to complete the interviewing process or a

4   coding to indicate 'completed status'. Both graphs are helpful in visualising sequences and

5   can help assessing compliance with the survey protocol which prescribes what interviewers

6   should be doing.

7   **Figure 2**: Sequence frequency plot: basic sequence plot of the 10 most frequent sequences.

8

9
10
11
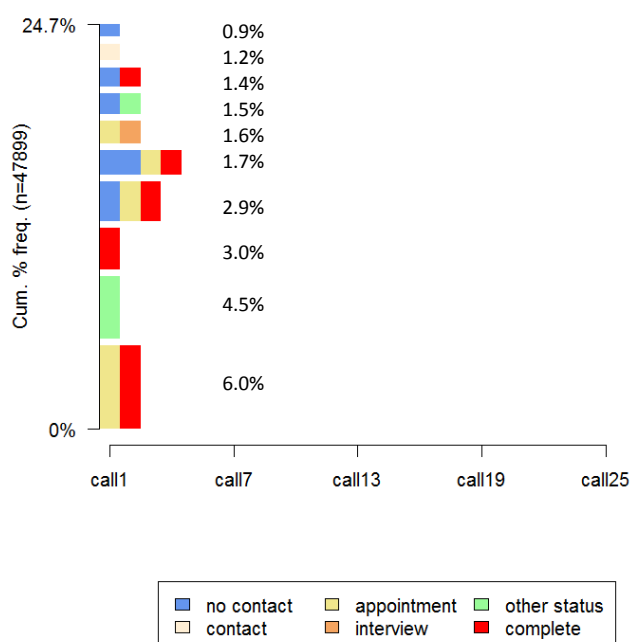
12   Table 2 contains a transition rate matrix, indicating the likely outcome at the next call

13   given a particular call outcome at the current call. Rates in rows add up to 1. Table 2 indicates

14   that a non-contact is very likely to lead to another non-contact (65% of the time) and to a

15   contact call without a further outcome (12%). A contact call is likely to lead to a non-contact

16   (38%), to another contact (20%) and to the end of the sequence (18%), possibly indicating a

17   refusal but with the interviewer not coding it as such. 'Any other status' is either the end of

18   the sequence (56%) or it leads to a non-contact (23%), or another 'any other status' outcome

1   (11%). An appointment leads in almost 24% of cases to an interview. However, 18% of cases

2   also result in a non-contact indicating a broken appointment and therefore possibly a hidden

3   refusal. The matrix indicates again some unusual interviewer behaviour: for example, in 6%

4   of completed cases, somewhat surprisingly, further calls are being made after the entire

5   interviewing process has already been completed (2.4% of all sequences). These cases require

6   further investigations as they may imply unnecessary costs. In total 18% of cases with a

7   contact and 1% of cases with an appointment are not followed-up. Whilst there may be

8   legitimate reasons for these interviewer behaviours, including coding errors of outcomes, it

9   seems nevertheless worthwhile investigating these unusual calling strategies, including the

10   characteristics of interviewers who conduct such calls. To summarise, outcomes such as

11   noncontact, contact and 'any other status' are likely to lead to non-productive call sequences,

12   whereas an appointment may indicate a high likelihood for an interview at the next call.

13   **Table 2**: Transition rate matrix indicating likely outcome of the next call given a particular call
14   outcome at the current call.

15

Current Call Outcome	Next call outcome						
	Non-Contact	Contact	Other status	Appoint-ment	Interview	Complete	End of sequence
Non-Contact	0.65	0.12	0.07	0.08	0.01	0.02	0.04
Contact	0.38	0.20	0.09	0.09	0.03	0.04	0.18
Other status	0.23	0.07	0.11	0.02	0.01	0.01	0.56
Appointment	0.18	0.06	0.05	0.05	0.24	0.42	0.01
Interview	0.08	0.09	0.03	0.04	0.06	0.17	0.53
Complete	0.01	0.04	0.01	0.00	0.00	0.00	0.94

16

17

18   Another advantage of basic sequence plots is the identification of potential coding

19   errors. For example, initial analysis of the sequence plots for the Understanding Society data

20   (results not shown) found more than 13% of cases with calls after completed interviews. After

21   analysing the call record data further, it became apparent that a number of calls had

22   erroneously been entered in the wrong order. This was subsequently corrected by the survey

23   agency, and also for further waves.

1

2   *Cluster Analysis and Multidimensional Scaling*

3   Next, cluster analysis based on the optimal matching distance matrix is performed. To

4   implement optimal matching the cost settings need to be specified. The 'constant' method in

5   R is the default option for optimal matching in the *TraMineR* library (substitution cost 2, indel

6   cost 1). With this method the substitution costs are the same for all possible call outcomes and

7   the substitution operation is equivalent to 2 indel operations. However, it is more intuitive to

8   set the costs according to the transition rates, i.e. the probability of moving from one call

9   outcome to another. The TRATE method in R implements this (Gabadinho et al., 2010 and

10   2011), and this is the method used here. A number of different cost settings were explored,

11   but the overall conclusions remained the same. Cluster analysis was performed exploring

12   different number of clusters (k = 3, 4, 5 and 6 clusters).

13   Figure 3 displays the four cluster solution which was believed to be the most

14   appropriate choice due to it being the most interpretable and parsimonious. The first cluster

15   (17,705 households) contains mainly successful sequences with interviews. These are

16   primarily shorter sequences (around 2 to 3 calls) but interestingly also contain a small number

17   of longer sequences (around 7-13 calls). These longer sequences are also characterised by

18   appointments and interviews, and may belong to households with several household members

19   where longer call sequences may be expected. The sequences from cluster 1 hardly contain

20   any non-contacts. Cluster 2 (12,768 cases) contains again shorter sequences, with around 2-5

21   calls, but with predominantly unsuccessful outcomes, including ineligibles, refusals, non-

22   contacts and non-productive contact calls. Cluster 3 (11,406 households) contains medium to

23   long call sequences (around 5 to 8 calls mostly) with a number of successful calls or an

24   interview during the call sequence but also many non-contact calls. Cluster 4 (6,020 cases)

25   contains long sequences with 10 and more calls, predominately driven by non-contact calls,

1 and mostly unsuccessful call outcomes. To summarise, the cluster analysis seems to represent

2 a categorisation that is driven primarily by call length and outcome. Cluster analysis proves to

3 be useful in summarising rich call record data into meaningful and interpretable dimensions

4 which can be used for further analysis (see as an example Durrant et al., 2015 and 2017).

5 **Figure 3**: Sequence plots resulting from cluster analysis with 4 clusters based on optimal
6 matching.

7

8

9
10
11

12
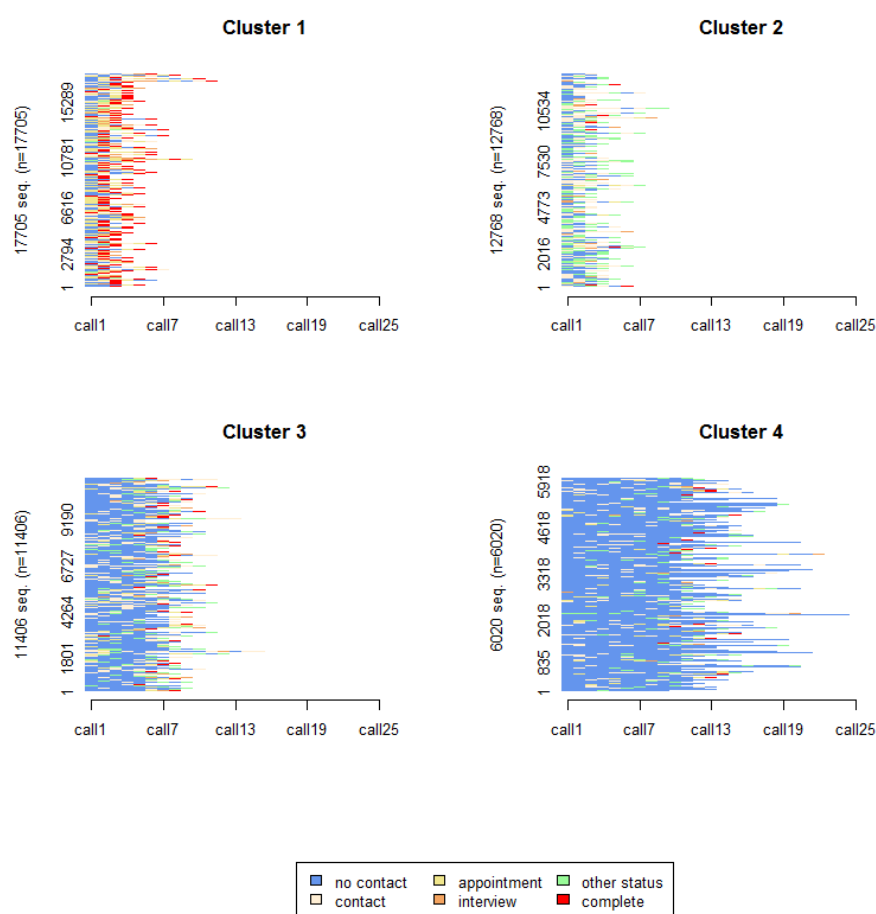
13 We now turn to the results from the multidimensional scaling analysis, based on the

14 optimal matching distance matrix. Figure 4 shows two graphs, presenting a) the first

15 dimension only and b) the first and the second dimensions together. Multidimensional scaling

16 orders sequences according to a criterion. In any particular area of the vertical axis of Figure

4a the omitted sequences are similar to the ones plotted. The graph implicitly uses the method by Fasang and Liao (2014) of using the middle sequence or medoid. Displaying the sequences according to their ranking of the multidimensional scaling analysis in the first dimension (Figure 4a) indicates an ordering primarily according to length, with successful call sequences at the bottom of the vertical axis, characterised by appointments, interviews and fully completed interviews with no non-contacts. These are predominantly short sequences with only some longer ones, similarly to cluster 1 discussed above. Next, short sequences are observed with 'any other status' outcome (ineligible or refusal) without non-contacts. Then, calls with non-contacts are displayed in increasing order from the bottom to the top of the vertical axis with calls driven by non-contacts at the very top. The one-dimensional graph therefore seems to be displaying sequences according to length and to some extent outcome. Although this first graph has strong similarities with the simpler Figure 1, it is different in that it groups sequences together according to *all* call outcomes during the whole sequence not just the outcome of the last call and it is not simply based on length. In Figure 4b each point inside the graph represents a sequence in the dataset, including its position according to the first and second dimension. The first dimension, presented on the horizontal axis, orders the sequences according to length and to some extent outcomes with length being the driving factor. The vertical axis displays the sequences according to the second dimension with sequences below the horizontal line representing mostly successful sequences and above the line mostly unsuccessful sequences. The further away the sequences are from the horizontal line the larger the number of non-contacts. The second dimension therefore displays sequences as a mixture of outcome and length with outcome being the driving factor. Figure 4b displays a number of outliers or unusual cases, such as those with long calls and predominantly non-contact calls. In survey practice these could be displayed separately to investigate further mechanisms leading to such patterns.

1   **Figure 4**: Multidimensional scaling plots

2

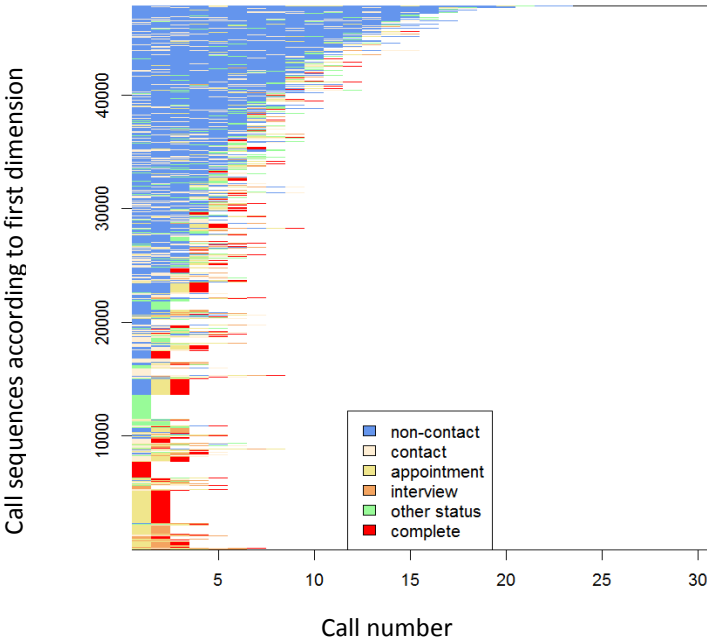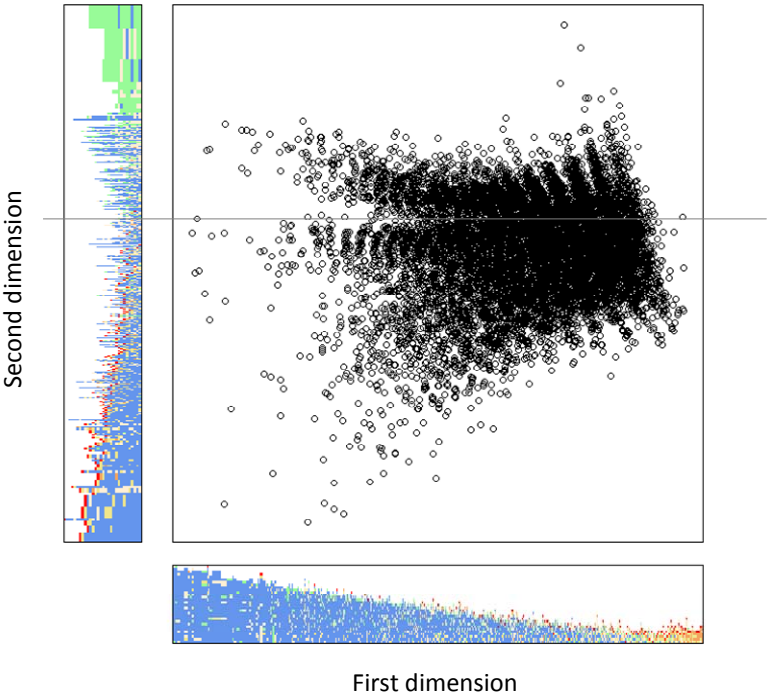3   Figure 4a: Multidimensional scaling plot: first dimension

4
5
6
7
8

9   Figure 4b: Multidimensional scaling plot: first and second dimensions (the colour legend for the
10   first and second dimensions is as in Figure 4a)

11
12

1

2   ***Multidimensional Scaling Within Interviewers***

3

4   It is well known that interviewers can have significant influences on response

5   outcomes (Pickery, Loosbeldt and Carton, 2001; Hox and De Leeuw, 2002; Durrant, Groves,

6   Staetsky and Steel, 2010; Durrant and D'Arrigo, 2014; Vassallo, Durrant, Smith and

7   Goldstein, 2015). Of particular interest is an analysis of the calling behaviour per interviewer,

8   which provides an easy and intuitive tool for investigating interviewer performance and

9   adherence to interviewing protocols and guidelines. This is of relevance from both a

10   methodological perspective (analysis of sequences within subroups) and from a substantive

11   perspective, since interviewers play a crucial role in scheduling calls and making contact and

12   establishing cooperation with sample members (Durrant et al. 2010; Durrant and D'Arrigo,

13   2014).

14   Figure 5 shows multidimensional scaling plots for two selected interviewers. The axes

15   are defined as previously in Figure 4. The call sequences are colour-coded to indicate short

16   and long call sequences (defined as up to 6 calls and more than 6 calls respectively to

17   represent the interviewer guidelines) and successful and unsuccessful sequences (defined as at

18   least one interview in the household versus no interviews respectively). The plots reveal a

19   clear distinction between short and long, and successful and unsuccessful calling sequences as

20   well as significant differences between interviewers. For example, interviewer A has made

21   mostly long call sequences (in fact 57%) and has experienced a relatively high proportion of

22   unsuccessful calls (63%), with the plot indicating also a cluster of short and medium-to-long

23   successful calls. Interviewer B has made mostly short call sequences (88%) with a relatively

24   high proportion of successful call sequences (62%). Hence, such sequence plots allow the

25   visualisation and comparison of calling patterns between interviewers. However, it should be

26   noted that any interviewer effect in face-to-face surveys can be confounded with area effects

1 (Campanelli and O'Muircheartaigh, 1999; Durrant et al., 2010) and differences in interviewer

2 performances found may not represent a causal effect. Hence, the plots do not allow a direct

3 evaluation of interviewer performances since differences may also reflect the difficulty of the

4 cases to be interviewed or the type of area the interviewer works in. Sequence analysis, as

5 with all descriptive statistics tools, should therefore be followed up by further investigations

6 on the underlying reasons for the calling behaviour. Nevertheless, the findings may point

7 towards some tailored treatment such as incentives or to more specialised training for

8 interviewers in certain areas.

9   **Figure 5:** Multidimensional scaling plots for two different interviewers: interviewer A
10   (anonymised interviewer number: 11002071) and interviewer B (anonymised
11   interviewer number: 11006065) (the colour legend for the first and second dimensions
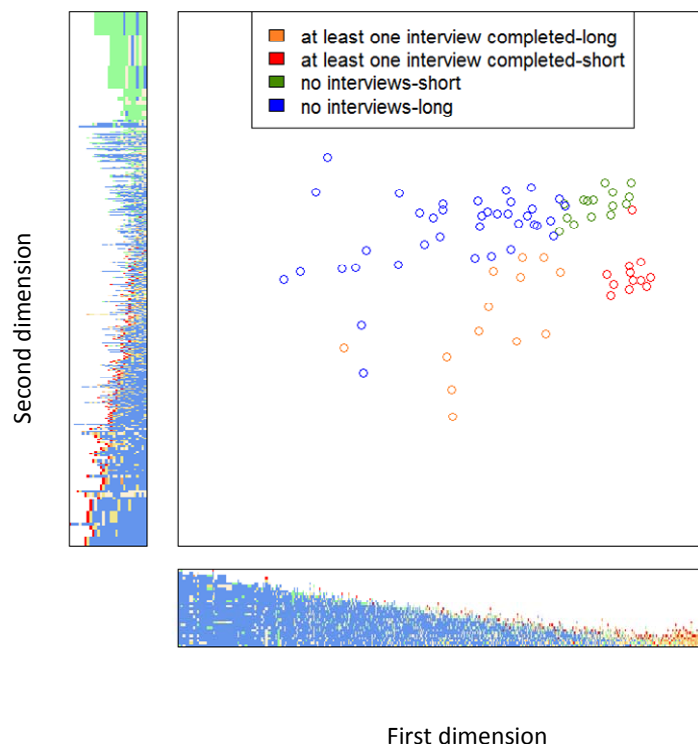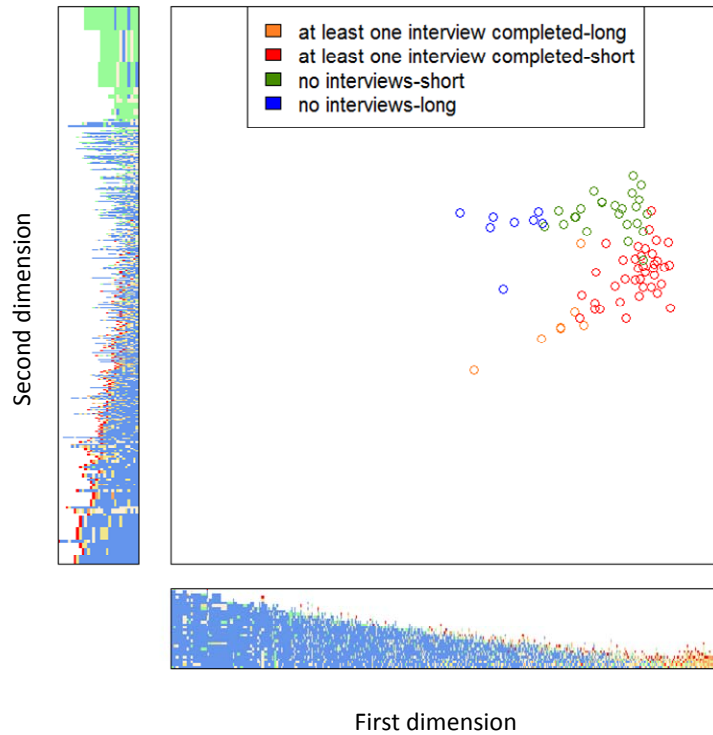12   is as in Figure 4a)
13

Figure 5a: Interviewer A

14
15
16

Figure 5b: Interviewer B

First dimension

1
2

3   **Conclusions and Implications for Survey Practice**

4   The paper presents sequence analysis as a tool for investigating call record data to better

5   understand and improve survey processes. It offers a potentially powerful tool for displaying

6   and summarising record data, which can be large and complex, and for the detection of

7   outliers and unusual cases and subgroups. The method proposed can be used for both face-to-

8   face and telephone, and for cross-sectional and longitudinal surveys. A novel approach is to

9   introduce sequence analysis within interviewers, allowing the identification of unusual

10   interviewer behaviour and an initial analysis of interviewer performance. This contributes to

11   the growing body of literature on interviewer performance and evaluation (e.g. Pickery et al.,

12   2001; Durrant et al., 2010; Durrant and D'Arrigo, 2014; West and Groves, 2013).

Basic sequence plots show the distribution of sequences across households. Transition rate matrices indicate the likelihood of a particular call outcome given a previous outcome. Combining sequence analysis with cluster analysis and multidimensional scaling allows grouping of sequences with similar features. Multidimensional scaling defines a rank order, based on a distance matrix, and sequences can be plotted in one or two dimensions, which helps to reveal key features of the sequences that drive that ordering. Sequence analysis, in particular the multidimensional scaling plots, may identify groups of sequences, outliers and unusual or unexpected calling behaviour, which may require further investigation by fieldwork managers. Although some of the standard statistical software can nowadays implement the method, the routines are still mainly limited to a relatively small number of short sequences. Here, we overcame this problem by using the *WeightedCluster* and *Vegan* libraries in R.

We first list the specific findings from the application of sequence analysis to the Understanding Society data. Although the findings are not the main focus of this paper, they support the general implications below:

1. Despite clear guidance on the minimum number of calls per address, a number of households are identified that received significantly less than 6 calls, despite the fact that they were neither coded response, refusal, nor ineligible (a total of 8% of households). Some of those call sequences consisted of non-contact calls throughout. For future rounds, we recommend monitoring of such cases, establishing the reasons and, if appropriate, provide further guidance to interviewers to avoid such occurrences. In a small number of cases further calls are being made although a case was already coded as completed. Careful consideration and guidance of interviewer work in such circumstances may help to avoid unnecessary calls in the future to improve survey efficiency and to reduce costs.

2. The results from the transition rate matrix indicate that a non-contact is likely to be followed by another non-contact. A contact call is likely to lead to a non-contact or another non-productive call. Hence, findings from transition rate matrices may help to predict response outcomes and might help designing tailored treatment, for example, regarding decisions of stop calling.

3. The intuitive notion that sequences are characterised by length and outcome are supported by the substantive findings from both the cluster and the multidimensional scaling analyses. The multidimensional scaling plots showed a clear distinction between short and long, and successful and unsuccessful calling sequences.

4. The sequence analysis reveals the significance of non-contact calls. In fact, in this dataset a large proportion of all calls are non-contact calls (56%). If the aim is to increase efficiency and to reduce the number of unproductive calls, it seems advisable to investigate methods to reduce the large number of non-contact calls, e.g. to apply methods for the reduction of non-contact calls such that targeting time slots with a high likelihood of establishing contact and to provide such guidance to interviewers. (For examples of such techniques, see Weeks, Jones, Folsom and Benrud, 1980; Weeks, Kulka and Pierson, 1987; Kulka and Weeks, 1988; Durrant et al., 2011.)

Following on from the specific findings, the more general implications of the method for survey practice are wide ranging:

1. Sequence analysis may be used to inform future (routine) monitoring to identify unusual calling behaviours and outliers and to assess the adherence to interviewing guidelines either during data collection or retrospectively once data collection has finished. In a further step, the methods may inform the design of automated flag systems that indicate unusual calling behaviours, to inform the choice of which summary measures to use in routine monitoring or to derive summary statistics and indicators for future use. Although

1   often used, simple summary measures on their own are generally not sufficient to analyse

2   sequences as a whole.

3 2. More specifically, sequence analysis of call record data is another tool to identify

4   problems or editing errors in the dataset. For this dataset in fact, sequence analysis

5   identified a number of editing and coding errors that were subsequently corrected by the

6   survey agency in this and future waves. These errors had not been picked up initially by

7   standard summary measures.

8 3. Related to 1 and 2, the findings from sequence analysis may be used to inform responsive

9   and adaptive survey designs, which represent data-driven tailoring of data collection

10   procedures to different sample members. In a responsive design the survey is modified

11   after the initial design and after the survey has been fielded, with modifications based on

12   the early phases of the fieldwork. If a survey has been conducted a number of times

13   previously, it may be possible to determine different optimal designs for different

14   subgroups on the basis of this past experience (adaptive survey design). (For a range of

15   applications of such designs, including call record data, see Kreuter, 2013, Schouten,

16   Peytchev and Wagner, 2017, Luiten, 2013 and Tourangeau, Brick, Lohr and Li, 2017.)

17 4. Sequence analysis informs further (more sophisticated) modelling going beyond simple

18   summary measures. As a demonstration, Durrant et al. (2015 and 2017) model for the

19   first time sequence length and sequence outcome jointly, informed by the outcome of the

20   sequence analysis that both indicators simultaneously are important for describing survey

21   processes. As a result, they were able to improve on standard response modelling using a

22   multinomial model. More specifically, having identified sequence length and outcome as

23   (joint) key features of the calling patterns, it is of interest to investigate the correlates and

24   determinants of both. Certain call outcomes early on in the sequence are predictive of

25   later call outcomes and sequence length. Durrant et al. (2015 and 2017) identify cases

1   with a high likelihood of long unsuccessful calls early on in the data collection process to

2   inform more efficient calling strategies.

3   Sequence analysis can also have limitations. Although it does not depend on explicit

4   distributional assumptions and does not require modelling techniques, a range of choices,

5   such as regarding metric and costs and number of clusters or even regarding a cluster analysis

6   technique, have to be made. Here a sensitivity analysis was carried out exploring the different

7   settings of the algorithms. There is also the potential problem of overplotting with many

8   sequences displayed in a plot, which may lead to misrepresentation of data if plotting is not

9   carried out carefully. There is a technical limit how thin each sequence line can be displayed

10   and what is visible to the human eye (see also Fasang and Liao, 2014). This may be overcome

11   by only plotting a subgroup of sequences or only unusual cases. A number of suggestions

12   have been made, for example using relative frequency sequence plots (Fasang and Liao,

13   2014), and some of those techniques have been explored here. Software package capabilities

14   represent another limitation for sequence analysis. The R software package is not able to

15   conduct optimal matching or multidimensional scaling on a large number of relatively long

16   sequences as the limit of a vector in R is $2^{31}-1$ elements. Consequently, optimal matching can

17   be applied in R to a dataset with up to around 35,000 sequences (depending on the length of

18   the sequences). Other statistical software packages such as STATA and CHESA, suffer from

19   the same limitation. However, to overcome this problem we used the *WeightedCluster*

20   (Studer, 2013) and *Vegan* (Oksanen et al., 2013) libraries in the R software package. Also, a

21   function is available in the *WeightedCluster* library for performing cluster analysis around

22   medoids (PAM algorithm), which is computationally efficient and is therefore appropriate for

23   analysing very large datasets (Studer, 2013).

24   To conclude, sequence analysis is a useful technique for summarising and visualising

25   rich call record data. It can be used as a technique for monitoring field work in survey practice

as well as for summarising complex call record data. These summaries can then be employed for further analysis of different aspects of survey research, e.g. to inform modelling for the prediction of call outcomes (see Durrant et al., 2015 and 2017).

References

Aassve, A., Billari, F.C., & Piccarreta, R. (2007). Strings of adulthood: A Sequence analysis of young British women's work-family trajectories. *European Journal of Population, 23*, 369-388. DOI: http://dx.doi.org/10.1007/s10680-007-9134-6.

Abbott, A. & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology, 96*, 144-185. DOI: 10.1086/229495.

Abbott, A. & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology review and prospect. *Sociological Methods and Research, 29*, 3-33. DOI: 10.1177/0049124100029001001

Bartholomew, D., Steele, F., Moustaki, I., & Galbraith, J. (2008). *Analysis of multivariate social science data*. London: CPC Press.

Bates, N., Dahlhamer, J., & Singer, E. (2008). Privacy concerns, too busy, or just not interested: Using doorstep concerns to predict survey nonresponse. *Journal of Official Statistics, 24*, 591-612.

Berthoud, R., Fumagalli, L., Lynn, P., & Platt, L. (2009). *Design of the Understanding Society Ethnic Minority Boost Sample*. Understanding Society Working Paper Series No. 2009-02. Institute for Social and Economic Research, University of Essex.

Biemer, P. P., Chen, P., & Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of Royal Statistical Society: Serie A, 176*, 147–168. DOI: 10.2307/23355181

Blom, A., Jäckle, A., & Lynn, P. (2010). The use of contact data in understanding cross-national differences in unit non-response. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts,* edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith, 335-354. New York: Wiley and Sons.

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

Campanelli, P. & O'Muircheartaigh, C. (1999). Interviewers, interviewer continuity, and panel survey nonresponse. Quality & Quantity, 33, 59-76.

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. In *Proc. of the Joint Statistical Meeting, Section of Survey Research Methods*, 743-772. Dallas, Texas, USA.

Durbin, J. & Stuart, A. (1954). Call-backs and clustering in sample surveys. *Journal of Royal Statistical Society, Series A, 117(4)*, 387-428.

Durrant, G. B., D'Arrigo, J., & Müller, G. (2013a). Modeling Call Record Data: Examples from Cross-Sectional and Longitudinal Surveys." In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter, 281-308. New Jersey: Wiley and Sons.

Durrant, G. B., D'Arrigo, J., & Steele, F. (2011). Using Field Process Data to Predict Best Times of Contact Conditioning on Household and Interviewer Influences. *Journal of Royal Statistical Society: Series A, 174*, 1029-1049. DOI: 10.1111/j.1467-985X.2011.00715.x

Durrant, G. B., D'Arrigo, J., & Steele, F. (2013b). Analysing Interviewer Call Record Data by Using a Multilevel Discrete-Time Event History Modelling Approach. *Journal of Royal Statistical Society: Series A, 176*, 251-269. DOI: 10.1111/j.1467-985X.2012.01073.x

Durrant, G. B., Groves, R.M., Staetsky, L., & Steele, F. (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly, 74*, 1-36. DOI: 10.1093/poq/nfp098

Durrant, G. B., Maslovskaya, O., & Smith, P.W.F. (2015). Modelling Final Outcome and Length of Call Sequence to Improve Efficiency in Interviewer Call Scheduling. *Journal of Survey Statistics and Methodology, 3*, 397-424. DOI:10.1093/jssam/smv008.

Durrant, G.B. & D'Arrigo, J. (2014). "Doorstep Interactions and Interviewer Effects on the Process Leading to Cooperation or Refusal." *Sociological Methods and Research, 43*, 490-518. DOI: 10.1177/0049124114521148.

Durrant, G.B., Maslvskaya, O., & Smith, P.W.F. (2017). Using Prior Wave Invormation and Paradata: Can They Help to Predict Response Outcomes and Call Sequence Length in a

Longitudinal Study? *Journal of Official Statistics, 33 (3), 1-19*. DOI: https://doi.org/10.1515/jos-2017-0037.

Elzinga, C. H. & Liefbroer, A.C. (2007). De-standardization of family-life trajectories of young adults: A Cross-national comparison using sequence analysis. *European Journal of Population, 23*, 225-250. DOI: http://dx.doi.org/10.1007/s10680-007-9133-7.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. Chichester: Wiley.

Fasang, A.E. & Liao, T.F. (2014). Visualizing Sequences in the Social Sciences: Relative Frequence Sequence Plots. *Sociological Methods and Research, 43 (4)*, 643-676. DOI: 10.1177/0049124113506563

Gabadinho, A., Ritschard, G., Mueller, N.S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40*, 1-37. DOI: 10.18637/jss.v040.i04.

Gabadinho, A., Ritschard, G., Studer, M., & Mueller, N.S. (2010). Mining sequence data in R with the TraMineR package: A User's guide. *Technical Report*. Geneva: University of Geneva. Available at: http://mephisto.unige.ch/pub/TraMineR/doc/1.4/TraMineR-1.4-Users-Guide.pdf (accessed March 2016).

Groves, R. M. & Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of Royal Statistical Society: Series A, 169*, 439-459. DOI: 10.1111/j.1467-985X.2006.00423.x

Halpin, B. (2003). *Tracks through time and continuous processes: transitions, sequences and social structure*. Conference paper. Frontiers in Social and Economic Mobility, Cornell University.

Halpin, B. (2014). *SADI: Sequence Analysis tools for Stata*. Working Paper WP2014-03, Department of Sociology, University of Limerick. Available at: http://www.ulsites.ul.ie/sociology/sites/default/files//SADI.pdf (accessed March 2016).

Hanly, M. (2014). *Improving Nonresponse Bias Adjustments with Call Record Data*. Conference paper. 25th International Workshop on Household Survey Nonresponse, Iceland.

Hanly, M., Clarke, P., & Steele, F. (2015). Sequence Analysis of Call Record Data: Exploring the Role of Different Cost Settings. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 179(3),* 793-808. DOI: 10.1111/rssa.12143.

Hollister, M. (2009). Is Optimal Matching Suboptimal? *Sociological Methods and Research, 38 (2),* 235-264. DOI: 10.1177/0049124109346164.

Hox, J. & De Leeuw, E. (2002). The Influence of Interviewers' Attitude and Behavior on Household Survey Nonresponse: An InternationalCcomparison. In *Survey Nonresponse,* edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, & R.J.A. Little, 103-119. New York: Wiley.

Kaufman, L. & Rouseeuw, P.J. (1987). Clustering by means of Medoids. In *Statistical Data Analysis Based on the L1- Norm and Related Methods*, edited by Y.Dodge, 405-416, North-Holland.

Kaufman, L. & Rousseeuw, P.J. (1990). *Finding Groups in Data. An Introduction to Cluster Analysis*. New York: Wiley and Sons.

Knies, G. (ed.) (2014). *Understanding Society-The UK Household Longitudinal Study Waves 1-4, 2009-2013, User Manual*. Colchester: University of Essex.

Kreuter, F. & Kohler, U. (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*, 203-226.

Kreuter, F. (ed.) (2013). *Improving surveys with paradata: Analytic uses of process information*. New Jersey: Wiley and Sons.

Kreuter, F., Couper, M., & Lyberg, L. (2010a). The use of paradata to monitor and manage survey data collection. In *Proc. of the Joint Statistical Meeting, Section of Survey Research Methods*, 282-296. Vancouver, Canada.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., & Raghunathan, T.E. (2010b). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys. *Journal of Royal Statistical Society: Series A, 173*, 389–407. DOI: 10.1111/j.1467-985X.2009.00621.x.

Kruskal, J. B. and M. Wish. (1978). *Multidimensional scaling*. Beverly Hills: Sage.

Kulka, R. A. & Weeks, M.F. (1988). Towards the development of optimal calling protocols for telephone surveys: A conditional probabilities approach. *Journal of Official Statistics, 4*, 319-358.

Laflamme, F., Maydan, M., & Miller, A. (2008). Using paradata to actively manage data collection survey process. In *Proc. of the Joint Statistical Meeting, Section of Survey Research Methods*, 630-637. Denver, Colorado, USA.

Levenshtein,V.I. (1966). Binary codes capable of correcting deletions, insertions and reversal. *Soviet Physics Doklady, 10*, 707-710.

Lipps, O. (2011) A note on improving contact times in panel surveys, Field Methods, published online in Oct 2011.

Luiten, A. (2013). Improving survey fieldwork with paradata, Statistics Netherlands, The Hague, Heerlen.

Lynn, P. (2009). *Sample Design for Understanding Society*. Understanding Society Working Paper Series 2009–01. http://research.understandingsociety.org.uk/publications/working-paper/2009-01.

Malo, M. A. & Munoz-Bullon, F. (2003). Employment status mobility from a life-cycle perspective: A sequence analysis of work-histories in the BHPS. *Demographic Research, 9*, 119-162. DOI: http://dx.doi.org/10.4054/DemRes.2003.9.7.

Martin, P., Schoon, I, & Ross, A. (2008). Beyond transitions: Applying optimal matching analysis to life course research. *International Journal of Social Research Methodology, 11 (3)*, 179-199. DOI: 10.1080/13645570701622025.

McFall, S. L. & Garrington, C. (eds.) (2011). *Early findings from the first wave of the UK's household longitudinal study*. Colchester: Institute for Social and Economic Research, University of Essex.

McFall, S. L. (ed.) (2012). *Understanding Society: Findings 2012*. Colchester: Institute for Social and Economic Research, University of Essex.

McVicar, D. & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of Royal Statistical Society: Series A, 165*, 317-334.

Miyazawa, K., Tsubouchi, H., Naka, D., Takahashi, K., Okigaki, M., Arakaki, N., Nakayama, H., Hirono, S., Sakiyama, O., Takahashi, K., Gohda, E., Daikuhara, Y., & Kitamura, N. (1989). Molecular cloning and sequence analysis of cDNA for human hepatocyte growth factor. *Biochemical and Biophysical Research Communications, 163*, 967-973.

Oksanen, J. (2013). *Multivariate analysis of ecological communities in R: vegan tutorial.* Available at : http://www.cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf (accessed March 2016).

Olson, K. & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In *Improving Surveys with Paradata: Analytic Use of Process Information,* edited by F. Kreuter, 37-68. New Jersey: Wiley and Sons.

Piccarreta, R. & Lior, O. (2010). Exploring sequences: a graphical tool based on multi-dimensional scaling. *Journal of Royal Statistical Society: Series A, 173,* 165-184. DOI: 10.1111/j.1467-985X.2009.00606.x.

Pickery, J., Loosveldt, G., & Carton, A. (2001). The effects of interviewer and respondent characteristics on response behavior in panel surveys. A multilevel approach. *Sociological Methods and Research, 29,* 509-523. DOI: 10.1177/0049124101029004004.

Pollock, G., Antcliff, V., & Ralphs, R. (2002). Work orders: Analysing emplpyment histories using sequence data. *International Journal of Social Research Methodology, 5(2),* 91-105. DOI: 10.1080/13645570110062432.

Purdon, S., Campanelli, P. and Sturgis, P. (1999). Interviewer's Calling Strategies on Face-to-Face Interview Surveys. J. Off. Statist., 15, 2, 199-216.

Sankoff, D. & Kruskal, J.B. (1983). *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Reading: Addison-Wesley Publication.

Schouten, B., Peytchev, A. & Wagner, J. (2017) Adaptive Survey Design, Chapman and Hall/CRC, London, New York.

Sinibaldi, J. (2014). *Using Call-Level Interviewer Observations to Improve Response Propensity Models, Chapter 4 in: Evaluating the Quality of Interviewer Observed Paradata for Nonresponse Applications*. PhD Thesis. München: Ludwig-Maximilian-Universität.

Smith, L. M., Sanders, J.Z, Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B.H., & Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature, 321*, 674-679. DOI:10.1038/321674a0.

Studer, M. (2013). WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working Papers, 24*, 1-33. Geneva: University of Geneva.

Swires-Hennessy, E. and Drake, M. (1992) The optimum time at which to conduct interviews, International Journal of Market Research (Journal of the Market Research Society) 34, 1, 61-72.

Tourangeau, R., Brick, J.M., Lohr, S., & Li, J. (2017). Adaptive and Responsive Survey Designs: A Review and Assessment. *Journal of the Royal Statistical Society, Series A, 180 (1)*, 203-223. DOI**:** 10.1111/rssa.12186.

Vassallo, R, Durrant, G.B, Smith, P.W.F, & Goldstein, H. (2015). Interviewer Effects on Nonresponse Propensity in Longitudinal Surveys: A Multilevel Modeling Approach. *Journal of the Royal Statistical Society, Series A, 178*, 83-99. DOI: 10.1111/rssa.12049

Wagner, J. (2013a). Adaptive Contact Strategies in Telephone and Face-to-Face Surveys. S*urvey Research Methods, 7*, 45-55. DOI: http://dx.doi.org/10.18148/srm/2013.v7i1.5037

Wagner, J. (2013b). Using Paradata-Driven Models to Improve Contact Rates in Telephone and Face-to-Face Surveys. In *Improving Surveys with Paradata: Analytic Use of Process Information,* edited by F. Kreuter, 145-170. New Jersey: Wiley and Sons.

Weeks, M. F., Jones, B.L., Folsom, R.E., & Benrud, C.H. (1980). Optimal times to contact sample households. *Public Opinion Quarterly, 44*, 101-114. DOI: 10.1086/26856.

Weeks, M. F., Kulka, R.A, & Pierson, S.A. (1987). Optimal call scheduling for a telephone survey. *Public Opinion Quarterly, 51*, 540-549.

West, B. T. & Groves, R.M. (2013). A propensity-adjusted interviewer performance indicator. *Public Opinion Quarterly, 77*, 352-374. DOI: 10.1093/poq/nft002.

Wiggins, R.D., Erzberger, C., Hyde, M., Higgs, P, & Blane, D. (2007). Optimal matching analysis using ideal types to describe the lifecourse: An Illustration of how histories of work, partnerships and housing relate to quality of life in early old age. *International Journal of Social Research Methodology, 10(4)*, 259-278. DOI: 10.1080/13645570701542025.

Wu, L. L. (2000). Some comments on" Sequence analysis and optimal matching methods in sociology: Review and prospect". *Sociological Methods and Research, 29*, 41-64. DOI: 10.1177/0049124100029001003.

Appendix

Note: Software for Implementing Sequence Analysis

Nowadays, a number of standard statistical software packages can implement sequence analysis including optimal matching, clustering and multidimensional scaling. These include R (Gabadinho, Ritschard, Studer, & Mueller, 2010; Gabadinho, Ritschard, Mueller, & Studer, 2011) and STATA. In STATA two approaches have been implemented, SADI (Halpin, 2014) and STATA SQ (Brzinsky-Fay, Kohler, & Luniak, 2006. In R the library *TraMineR* (Gabadinho et al., 2010 and 2011) has been specifically designed to carry out sequence analysis. SAS produces sequence index plots but cannot perform optimal matching. Software, such as SPSS, cannot implement the method at present. In addition, some specialised software packages such as CHESA also exist to conduct sequence analysis (Piccarreta & Lior, 2010).

However, potential limitations still exist. The R software package is not able to conduct optimal matching or multidimensional scaling on a large number of relatively long sequences as the limit of a vector in R is $2^{31}-1$ elements. Consequently, optimal matching can be applied in R to a dataset with up to around 35,000 sequences (depending on the length of the sequences). Other statistical software packages such as STATA and CHESA, suffer from the same limitation. In STATA the optimal matching procedure is capable of working with a moderate number of relatively short sequences; it has been tested using around 2,000

sequences with a maximum length of 100 positions (Brzinsky-Fay et al., 2006). Similarly in the CHESA package optimal matching can be applied to datasets with up to around 2000 cases (again depending on the length of the sequences) (Piccarreta & Lior, 2010). To overcome this problem we use the *WeightedCluster* (Studer, 2013) and *Vegan* (Oksanen et al., 2013) libraries in the R software package. We conduct optimal matching and multidimensional scaling using the unique sequences only and then applying the appropriate weights to all sequences in the datafile. Also, a function is available in the *WeightedCluster* library for performing cluster analysis around medoids (PAM algorithm), which is computationally efficient and is therefore appropriate for analysing very large datasets (Studer, 2013).