Working Paper, University of Southampton

Sequence Analysis as a Tool for Investigating Call Record Data

Gabriele B. Durrant, Olga Maslovskaya and Peter W.F. Smith

Department of Social Statistics and Demography

School of Social Sciences

University of Southampton, UK

(g.durrant@southampton.ac.uk)

Address for correspondence:

Gabriele Durrant

Department of Social Statistics and Demography

School of Social Sciences

University of Southampton

SO17 1BJ, Southampton

g.durrant@southampton.ac.uk

Abstract

Researchers have become increasingly interested in better understanding the survey data collection process in interviewer-administered surveys. However, tools for analysing paradata capturing information about field processes, also called call record data, are still not yet fully explored. This paper introduces sequence analysis as a simple tool for investigating such data with the aim of better understanding and improving survey processes. A novel approach is to use sequence analysis within interviewers, which allows the identification of unusual interviewer calling behaviours, and may provide guidance on interviewer performance. Combining the technique with clustering, optimal matching and multidimensional scaling, the method offers a way of visualising, displaying and summarising complex call record data. The method is introduced to inform survey management and survey monitoring. Sequence analysis is applied to call record data from the UK Understanding Society survey. The findings inform further modelling of call record data to increase efficiency in call scheduling.

Key Words: interviewer call record data, multidimensional scaling, optimal matching, paradata, responsive and adaptive survey designs, survey cooperation.

Acknowledgement

We thank Frauke Kreuter for her helpful comments on the research presented in this manuscript. The research was funded by the UK Economic and Social Research Council (ESRC), 'The Use of Paradata in Cross-Sectional and Longitudinal Research', grant number: RES-062-23-2997, grant holders: Gabriele Durrant, Peter W. F. Smith and Frauke Kreuter, and the ESRC National Centre for Research Methods, Research Workpackage 1, grant number ES/L008351/1.

1. Introduction

Many survey agencies nowadays routinely collect survey process data, so-called paradata (Couper, 1998; Kreuter, 2013). For interviewer administered surveys, including both telephone and face-to-face surveys, data about the fieldwork process, often termed call record or call history data, have received more and more attention (Bates *et al.*, 2008; Laflamme, 2008; Blom *et al.*, 2010). Such data may contain information about the outcome of the call or visit and the day and time of the call attempts. Several outcomes may be distinguished such as non-contact, contact, ineligible, refusal, appointment made and any interviewing done. This string of outcomes of all call attempts to a household is referred to as a *call sequence*, an ordered collection of activities or states (Piccarreta and Lior, 2010). A large number of different sequences are possible even if the number of positions (i.e. total length) is relatively small. As an example, Table 1 indicates a selection of short call sequences observed in the UK survey Understanding Society.

[Table 1 about here]

In recent years, the analysis of call record data has received increasing attention (Groves and Heeringa, 2006; Bates *et al.*, 2008; Kreuter *et al.*, 2010a; Wagner, 2013a and 2013b; Durrant et al. 2011; Durrant et al., 2013a and b). Survey agencies hope that analysis of call record data may inform best interviewer calling practices, identify more difficult cases as well as unusual interviewer behaviour and may provide strategies for improved survey nonresponse adjustment methods (Kreuter *et al.*, 2010b; Biemer *et al.*, 2013; Hanly, 2013 and 2014). Although survey researchers have become increasingly interested in understanding and improving the process of data collection, it is often not clear how best to analyse such data, in particular since call record data can be large and may exhibit complex data structures, such as time dependencies and multilevel clustering (Durrant *et al.*, 2011; Durrant *et al.*, 2013a; Durrant *et al.*, 2013b; Hanly, 2014; Sinibaldi, 2014). It can also be of significantly lower quality requiring editing and cleaning checks. Often ad-hoc methods are used and usually summary measures are applied to describe outcomes of call sequences (Groves and Heeringa, 2006; Bates *et al.*, 2008), such as the number of non-contact calls or the total number of calls. As noted in Fasang and Liao (2014, p. 644) sequences of categorical states are much more complex than simple numerical variables and cannot be easily summarized as categorical variables with a limited number of categories. Also in the survey methods literature, it has been recognised more recently, that it may be important to analyse the contact sequence as a whole (Kreuter and Kohler, 2009; Hanly, 2013). It may be the actual interplay between several call outcomes that is informative. A particular call outcome may have a different meaning if analysed separately or if seen as part of a longer sequence. For example, a non-contact after an appointment may be regarded as a 'hidden refusal' whereas otherwise it may simply be interpreted as a period of absence (e.g. in Table 1 compare household 4 with 1 and 6). Consequently, modelling procedures, such as discrete time event history analysis, have been developed to analyse not just the final nonresponse outcome but the process leading to contact, cooperation or refusal as a whole and to recognise the entire contact history (Durrant *et al.*, 2011; Durrant *et al.*, 2013b; Wagner, 2013b). However, a full modelling approach may not always be necessary or desirable to analyse sequences as a whole.

This paper introduces the use of sequence analysis for investigating call record data to inform survey monitoring and management processes. 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. The technique may help to identify unusual cases, which then may require further investigation. In particular, it can build the basis for further statistical modelling, as already demonstrated in Durrant et al. (2015). Sequence analysis offers a potentially powerful tool for visualising, displaying and summarising call record data and for exploring and reducing the complexity of such data structures. It represents a relatively simple descriptive method which can be easily implemented in practice, not requiring any modelling techniques or distributional assumptions. It has the potential to provide survey researchers and field managers with a simple graphical tool to investigate interviewer calling patterns either during or post data collection. Findings from the analysis may then inform future routine monitoring as well as further, more sophisticated statistical modelling. It should be stressed that sequence analysis should only be a first start in analysing interviewer behaviour. Further investigations, for example of unusual cases, and further modelling informed by the findings from sequence analysis is required.

Here, sequence analysis plots are combined with the results of optimal matching, clustering, and multidimensional scaling (Kruskal and Wish, 1978; Bartholomew *et al.*, 2008; Piccarreta and Lior, 2010). This allows finding similarities across the contact histories and identifying groups of sequences that are homogeneous. For example, from a survey practice perspective aiming to increase efficiency and to reduce costs, it may be important to identify sequences with a large number of unsuccessful call attempts such as non-contact calls. The technique allows the identification of unusual calling behaviours and provides survey managers with tools to display such cases. The paper also provides some practical guidance on how to analyse call record data using sequence analysis, for example regarding details on how to implement the method in 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.

Sequence analysis methods are frequently used in a range of disciplines such as in medicine and biology to detect DNA sequencing (Smith *et al.*, 1986; Miyazawa *et al.*, 1989), but also in the social sciences, for example in demography to study family-life trajectories (Elzinga and Liefbroer, 2007) and life course trajectories (Halpin, 2003; Aassve *et al.*, 2007; Picarreta and Lior, 2010), and in economics to study transitions in and out of employment (Malo and Munoz-Bullon, 2003) and transitions from education into work (McVicar and Anyadike-Danes, 2002). For a recent review of the method in the social science context see

Fang and Liao (2013). Researchers have also started to use sequence analysis in the context of survey methodology with the aim of informing nonresponse adjustment methods based on call histories (Kreuter and Kohler, 2009; Hanly, 2013; Pollien and Joye, 2014). Kreuter and Kohler (2009) and Hanly (2013) found that variables derived from call record data - in part informed by sequence analysis - did not improve nonresponse adjustment methods. Although not as useful as hoped, they point out in their discussions, that the method may have the potential for use in survey management field practice. This paper aims to address this shortcoming. It focusses on the use of sequence analysis for the investigation of call record data and provides an additional tool to inform survey management and to guide field practice.

The method 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.

The remainder of the paper is organised as follows. The methodology section, Section 2, introduces the basic principles of sequence analysis in the context of call record data. In Section 3 the method is applied to call sequences from Understanding Society, including a separate analysis per interviewer. The paper concludes with a summary of the main findings, implications for survey practice and a discussion of further research.

2. Methodology: Using Sequence Analysis

Sequence analysis consists of a series of routines, plots and outputs. Simple sequence plots show the distribution of sequences. Transition rate matrices indicate the propensity for the next outcome options following a particular call outcome. Ideally one would like to find similarities among the contact histories to create groups that are homogeneous in their outcome variables and to summarise the different sequence patterns. The low frequency count of sequences may specify each individual sequence but makes summarising sequences complicated (Piccarreta and Lior, 2010). Given the large number of possible outcomes and patterns no single characteristic that fully describes the sequences will be available. We do not expect an individual summary measure to describe the sequence as a whole. Rather, a combination of multiple characteristics will be necessary to establish an adequate measure of similarity across sequences. We first describe how a distance matrix can be constructed to order the sequences based on a number of criteria. This distance matrix can then inform clustering and multidimensional scaling. The combination of sequence analysis and multidimensional scaling has been proposed by Piccerreta and Lior (2010) and this approach is extended here to call record data.

2.1 Optimal Matching

One way of constructing a distance matrix is optimal matching (Levenshtein, 1966; Sankoff and Kruskal, 1983; Abbott and Hrycak, 1990; Abbott and Tsay, 2000; Hollister, 2009). This method computes the distance between pairs of sequences by counting the number of basic operations that are necessary to transform one sequence into another (Levenshtein distance). In simple terms, the more operations are required, the larger the differences. Each type of operation may be associated with a weight or cost such that a weighted distance matrix may be constructed. The basic operations considered are insertion, deletion and substitution. The insertion and deletion operations are also referred to as 'indel' operations since a deletion of an element in one sequence is equivalent to the insertion of an element in the other sequence. A substitution operation implies the direct substitution of one element in the sequence with another. Let us briefly consider an example of an indel operation and a substitution that transform the call sequence of household 1 in Table 1 into the call sequence of household 2. Both have two noncontact calls (N) in common but household 1 responds with an interview (I) after an appointment (A) whereas household 2 refuses (R) at the first contact-call. By applying one deletion and one substitution we have converted one sequence into the other. A typical default option for a substitution cost is 2 and for an indel it is 1, although other cost settings may be advocated (see section 3.4; also Abbott and Tsay, 2000; Wu, 2000; Hanly, 2013; Hollister, 2009).

Household 1:		**N N A I**
Deletion	(~~I~~)	N N A
Substitution	(A↔R)	N N R
Household 2:		**N N R**
Resulting distance = 3		

The operations are performed between all possible pairs of sequences. This procedure creates a distance matrix with the dimension equivalent to the number of sequences. In total, since the matrix is symmetrical, $(n^2 - n) / 2$ distance measures need to be calculated. An advantage is that the method takes account of multiple dimensions or characteristics of the sequences, not just of one or two characteristics as is the case for conventional summary measures of call record data. As indicated, a potential difficulty is the adequate specification of the costs (Wu, 2000; Abbot and Tsay, 2002). Results from prior analysis or different datasets may provide an empirical solution. A theoretical solution with the researcher deciding on the costs for the different types of required operations is also possible. For further discussions and options on cost settings see Hollister (2009).

2.2 Cluster Analysis and Multidimensional Scaling

The distance matrix can be used to carry out cluster analysis, to find similarities between groups of sequences based on multiple dimensions (Kaufman and Rousseeuw, 1990; Bartholomew *et al.*, 2008; Everitt *et al.*, 2011). Different clustering algorithms can be used, for example centroid-based clustering, distribution-based clustering and density-based clustering. It should be noted that in principle different clustering methods may give different results. In this paper we use the optimized 'partitioning around medoids' (PAM) algorithm (Studer, 2013), which differs from hierarchical algorithms and uses a predefined number of *k* groups to obtain the best partitioning of the dataset. It aims to identify the *k* best representatives of groups, called medoids, where a medoid is defined as the observation of a group with the smallest weighted sum of distances from the other observations in that group. The algorithm then aims to minimize the weighted sum of distances from the medoids.

Another possibility is to analyse the distance matrix by multidimensional scaling (Kruskal and Wish, 1978; Bartholomew *et al.*, 2008). This is a multivariate technique that aims to reveal the structure of the distance matrix by representing the sequences in a small number of dimensions, such as on a 1-dimensional scale or in a 2-dimensional map or plot. Then, each sequence is identified with a location in that dimension based on the distance matrix. Groups of sequences may be identified based on low, medium or high values in that dimension. In a 2- or 3-dimensional scatterplot, groups of sequences may be identified visually. Part of the interest in the analysis is to try to uncover which attributes of the sequences appear to carry weight in the similarity measure, i.e. which attributes are key features that determine the distances between sequences. The multidimensional scaling scatter plot allows the identification of groups of sequences as well as potential outliers.

2.3 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 *et al.*, 2010 and 2011) and STATA. In STATA two approaches have been implemented, SADI (Halpin, 2014) and STATA SQ (Brzinsky-Fay *et al.*, 2006. In R the library *TraMine*R (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 and 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 and 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).

3. Application of Sequence Analysis to the UK Understanding Society Survey

3.1 Design and Fieldwork of Understanding Society

The method is applied to call record data from the UK Understanding Society Survey (Wave 1). Understanding Society is the UK Household Longitudinal Study of approximately 40,000 responding households in the United Kingdom, covering topics on health, work, education, income, family and social life to help understand the long term effects of social and economic change as well as policy interventions. The study has many advantages over previously existing datasets in the UK by being exceptionally large and comprehensive. In particular, the study was designed to collect a range of paradata, including call record data. Only interviewers with above average experience and ability were selected for the study.

Data collection for each wave is scheduled across a 24 months period, with interviews taking place annually. Wave 1 data collection took place between January 2009 and March 2011. All interviews at Wave 1 were carried out face-to-face in respondents' homes by trained interviewers using computer-assisted personal interviewing (CAPI). All adult household members (age 16 and older) were asked to respond. A household also needed to respond to a household questionnaire in addition to all individual interviews. A minimum of six calls were made at each sampled address before it was considered unproductive but interviewers were encouraged to make further calls if possible. Interviewers had one month to contact households allocated to them (McFall, 2012; McFall and Garrington, 2011).

3.2 Call Record Data in Understanding Society and Analysis Sample

Call record data in Understanding Society contains information about the outcome of calls. The *call outcome at each call*, the key variable of interest here, is recorded by the survey agency as a 5-categorical variable: 'non-contact', 'contact made', 'appointment made', 'any interviewing done' and 'any other status'. A limitation is that the last category combines a range of possible call outcomes, ranging from different types of refusals to different ineligibility statuses. The *final outcome* variable recorded at the household level has about 50 outcome codes, split into six broad groups, containing 'ineligible', 'refusal', 'contact made but no interviewing', 'any interviewing but not completed' (at least one individual interview completed), 'case completed' (i.e. household questionnaire and all individual interviews from all member of the household have been completed). An additional category in the call outcome variable, 'interviewing process completed', was also created, if it was the last occurrence of the category 'any interviewing done' in the sequence and the final outcome call indicated 'case completed', which means that all household members have responded individually and the household questionnaire has been completed.

The analysis sample includes all households from Wave 1 with at least one call, including those which later were classified as ineligible. Cases from the Ethnic Minority Boost sample are excluded as rules for the selection of the boost sample differ from the rules for the main sample. Calls with no recorded outcomes are also excluded (1.2% of all calls). The final analysis sample contains a total of 255,778 calls with 11,143 distinct sequences, clustered within 47,899 households (including both responding and nonresponding households) and 741 interviewers. The number of non-contact calls is 142,705 calls (representing 56% of all calls) which is relatively high. The minimum length of a sequence is one and the maximum is 30 (mean length 5.34, median 4).

3.3 Application: Basic Sequence Plots and Transition Rates

Figure 1 shows a basic sequence plot which displays the sequences across calls for every household, colour coded according to the final outcome of each call. Each horizontal line in the plot represents a call record for one household, i.e. one sequence. About 10% of households experience only one call, and these end primarily in interview or 'any other status' indicating either ineligibles or refusals. Surprisingly, a small proportion of households experience a contact call with no further outcome or even a non-contact call with no further follow-up visits, which does not show adherence to the interviewer guidelines. This trend continues with the proportion of interviews steadily declining and the longer call sequences being predominantly driven by non-contact calls. Just over 70% of all sequences have a length between 1 and 6 calls. After 10 calls 88% of all households have been completed and after 15 calls this has increased to 98%. About 8% of all call attempts are still being made after the 10th call (20,032 calls). Figure 1 clearly shows that there are a number of households that do not receive the required minimum of 6 calls, although the households have been coded as a contact with no further outcome or a non-contact. This is the case for almost 8% of households.

It should be noted that a basic frequency plot such as in Figure 1 suffers from the problem of overplotting giving the large number of sequences. This may hide particular features and may even be misleading, if not carried out carefully. Solutions have been offered in the literature such as plotting only a subgroup of sequences, unusual cases and outliers and ordering sequences (see Fasang and Liao (2013) for further discussions). We aim to overcome this by plotting subgroups and by ordering sequences.

[Figure 1 about here]

Figure 2 displays the ten most frequent sequences, sometimes referred to as a sequence frequency plot (Fasang and Liao, 2013). 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.

For comparison, for 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 this 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 sample members in the sample as possible) rather than the specification of a maximum number of calls. In our sample the sixth group is also interesting as it finishes with the status 'any interview done' but there is no follow-up to complete the interviewing process or a coding to indicate 'completed status'. Both graphs are helpful in visualising sequences and can help assessing compliance with the survey protocol which prescribes what interviewers should be doing.

[Figure 2 about here]

Table 2 contains a transition rate matrix, indicating the likely outcome at the next call given a particular call outcome at the current call. Rates in rows add up to 1. Table 2 indicates that a non-contact is very likely to lead to another non-contact (65% of the time) and to a contact call without a further outcome (12%). A contact call is likely to lead to a non-contact (38%), to another contact (20%) and to the end of the sequence (18%), possibly indicating a refusal but with the interviewer not coding it as such. 'Any other status' is either the end of the sequence (56%) or it leads to a non-contact (23%), or another 'any other status' outcome (11%). An appointment leads in almost 70% of cases to an interview. However, 18% of cases also result in a non-contact indicating a broken appointment and therefore possibly a hidden refusal. The matrix indicates again some unusual interviewer behaviour: for example, in 6% of completed cases, somewhat surprisingly, further calls are being made after the entire interviewing process has already been completed (2.4% of all sequences). These cases require further investigations as they may imply unnecessary costs. In total 18% of cases with a contact and 1% of cases with an appointment are not followed-up. Whilst there may be legitimate reasons for these interviewer behaviours, including coding errors of outcomes, it seems nevertheless worthwhile investigating these unusual calling strategies, including the characteristics of interviewers who conduct such calls. To summarise, outcomes such as noncontact, contact and 'any other status' are likely to lead to non-productive call sequences, whereas an appointment may indicate a high likelihood for an interview at the next call.

[Table 2 about here]

Another advantage of basic sequence plots is the identification of potential coding errors. For example, initial analysis of the sequence plots for the Understanding Society data found more than 13% of cases with calls after completed interviews. After analysing the call record data further, it became apparent that a number of calls had been swapped due to initial data entering errors. This was subsequently corrected by the survey agency, and also for further waves.

3.4 Cluster Analysis and Multidimensional Scaling

Next, cluster analysis based on the optimal matching distance matrix is performed. To implement optimal matching the cost settings need to be specified. The 'constant' method in R is the default option for optimal matching in the *TraMine*R library (substitution cost 2, indel cost 1). With this method the substitution costs are the same for all possible call outcomes and the substitution operation is equivalent to 2 indel operations. However, it is more intuitive to set the costs according to the transition rates, i.e. the probability of moving from one call outcome to another. The TRATE method in R implements this (Gabadinho *et al.*, 2010 and 2011), and this is the method used here. A number of different cost settings were explored, but the overall conclusions remained the same. Cluster analysis was performed exploring different number of clusters (k = 3, 4, 5 and 6 clusters). All specifications explored led to very similar results and interpretations.

Figure 3 displays the four cluster solution which was believed to be the most appropriate choice. The first cluster (17,705 households) contains mainly successful sequences with interviews. These are primarily shorter sequences (around 2 to 3 calls) but interestingly also contain a small number of longer sequences (around 7-13 calls). These longer sequences are also characterised by appointments and interviews, and may belong to households with several household members where longer call sequences may be expected. The sequences from cluster 1 hardly contain any non-contacts. Cluster 2 (12,768 cases) contains again shorter sequences, with around 2-5 calls, but with predominantly unsuccessful outcomes, including ineligibles, refusals, non-contacts and non-productive contact calls. Cluster 3 (11,406 households) contains medium to long call sequences (around 5 to 8 calls mostly) with a number of successful calls or an interview during the call sequence but also many non-contact calls. Cluster 4 (6,020 cases) contains long sequences with 10 and more calls, predominately driven by non-contact calls, and mostly unsuccessful call outcomes. To summarise, the cluster analysis seems to represent a categorisation that is driven primarily by call length and outcome.

[Figure 3 about here]

We now turn to the results from the multidimensional scaling analysis, based on the optimal matching distance matrix. Figure 4 shows two graphs, presenting a.) the first dimension only and b.) the first and the second dimensions together. Multidimensional scaling 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 (2013) 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.

[Figure 4 about here]

3.5 Multidimensional Scaling Within Interviewers

It is well known that interviewers can have significant influences on response outcomes (Pickery et al. 2001; Hox and De Leeuw, 2002; Durrant et al. 2010; Durrant and D'Arrigo, 2014; Vassallo et al., 2015). Of particular interest is an analysis of the calling behaviour per interviewer, which provides an easy and intuitive tool for investigating interviewer performance and adherence to interviewing protocols and guidelines. This is of relevance from both a methodological perspective (analysis of sequences within subroups) and from a substantive perspective, since interviewers play a crucial role in scheduling calls and making contact and establishing cooperation with sample members (Durrant et al. 2010; Durrant and D'Arrigo, 2014).

Figure 5 shows multidimensional scaling plots for two selected interviewers. The axes are defined as previously in Figure 4. The call sequences are colour-coded to indicate short and long call sequences (defined as up to 6 calls and more than 6 calls respectively to represent the interviewer guidelines) and successful and unsuccessful sequences (defined as at least one interview in the household versus no interviews respectively). The plots reveal a clear distinction between short and long, and successful and unsuccessful calling sequences as well as significant differences between interviewers. For example, interviewer A (anonymised interviewer number 11002071) has made mostly long call sequences (57%) and has experienced a relatively high proportion of unsuccessful calls (63%), with the plot indicating also a cluster of short and medium-to-long successful calls. Interviewer B (anonymised interviewer number 11006065) has made mostly short call sequences (88%) with a relatively high proportion of successful call sequences (62%).

It should be noted that the plots do not allow a direct evaluation of interviewer performance. The results may, at least in part, reflect the difficulty of the cases the interviewers have been allocated to or the type of area they work in. Sequence analysis, as with all descriptive statistics tools, should therefore be followed up by further investigations on underlying reasons for the calling behaviour. For example, interviewer A is a British male of around 60 years of age and working in London and the South East. Interviewer B is also a British male in approximately the same age group but working in Yorkshire and the Humber, and an interviewer will find it more difficult to get a response from households in London than in areas, such as Yorkshire.

4. Conclusions and Implications for Survey Practice

Sequence analysis tools have recently been introduced to survey methodology in the context of nonresponse adjustment methods, although without much success in improving the adjustments (Kreuter and Kohler, 2009; Hanly, 2013; Pollien and Joye, 2014). A greater potential of the method lies in the use of the technique to inform survey management. The paper here presents sequence analysis as a tool for investigating call record data to better understand and improve survey processes. Although often used, simple summary measures on their own are not sufficient to analyse sequences as a whole and this is where sequence analysis can make an additional contribution. A novel approach of this paper is to introduce sequence analysis within interviewers, allowing the identification of unusual interviewer behaviour and an initial analysis of interviewer performance. This contributes to the growing body of literature on interviewer performance and evaluation (e.g., Pickery et al., 2001; Durrant et al., 2010; Durrant and D'Arrigo, 2014; West and Groves, 2013). Sequence analysis offers a potentially powerful tool for visualising, displaying and summarising record data, which can be large and complex, and for the detection of outliers and unusual cases and subgroups. The method proposed can be used for both face-to-face and telephone surveys, and for cross-sectional and longitudinal surveys. In this paper, it is applied to call record data from Understanding Society, a large-scale longitudinal survey in the UK.

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.

Although many of the findings could have been also discovered by using certain summary statistics, it is the potential of sequence analysis to inform the choice of such summary statistics that then can be used in future (routine) monitoring and for the identification of unusual cases and outliers. The main substantive findings from the analyses are:

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.
2. The results from the transition rate matrix indicate that a non-contact is likely to be followed by another non-contact and that an appointment is very likely to lead to an interview. A contact call is likely to lead to a non-contact call or another non-productive call.
3. In a small number of cases further calls are being made although a case was already coded as completed. Unless simply a coding error, careful consideration and guidance of interviewer work in such circumstances may help to avoid unnecessary calls in the future to improve survey efficiency.
4. The intuitive notion that sequences are characterised by length and outcome are clearly 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.

   Informed by the findings of this sequence analysis Durrant et al. (2015) define further joint modelling of both sequence length and outcome which had not been done before.

More specifically, having identified sequence length and outcome as key features of the calling patterns, it is of interest to investigate the correlates and determinants of both. Certain call outcomes early on in the sequence may be predictive of later call outcomes and sequence length. Durrant et al. (2015) identify cases with a high likelihood of long unsuccessful calls early on in the data collection process to inform more efficient calling strategies. Hence, this work provides a good example for the added value of sequence analysis and the type of further modelling the analysis can inform.

5. The sequence analysis reveals the significance of non-contact calls. 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. It may be advantageous to identify for example good times to establish contact (Weeks *et al.*, 1980; Weeks *et al.*, 1987; Kulka and Weeks, 1988; Durrant *et al.*, 2011) and to provide such guidance to interviewers.
6. The analysis of sequences by interviewers allows evaluation of interviewer performance and the identification of unusual interviewer behaviour. Sequence plots can then help investigate poor performers identified through this analysis and may identify key performance indicators that can be used in routine monitoring. Further analysis is then needed to identify the reasons for potentially different behaviours. For example, the effect could be primarily an area effect rather than an interviewer effect, where cases in an area are particularly hard to contact or to persuade.

Sequence analysis can also have limitations. Although it does not depend on explicit distributional assumptions and does not require modelling techniques, a range of choices, such as regarding metric and costs, have to be made. Here a sensitivity analysis was carried out exploring the different settings of the algorithm. There is also the potential problem of overplotting with many sequences displayed in a plot, which may lead to misrepresentation of data if not carried out carefully. There is a technical limit how thin each sequence line can be displayed and what is visible to the human eye (see also Fasang and Liao, 2013). This may be overcome by only plotting a subgroup of sequences or only unusual cases. A number of suggestions have been made, for example using relative frequency sequence plots (Fasang and Liao, 2013), and some of those techniques have been explored here.

Implications of the methods and the substantive findings for survey practice may be wide ranging. 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, for example, as part of a responsive survey design procedure, or retrospectively once data collection has finished to inform future survey designs. In a further step, the methods may inform the design of automated flag systems that indicate unusual calling behaviours and to derive summary statistics and indicators for future use. Sequence analysis of call record data also helps identifying problems or editing errors in the dataset. For this dataset in fact, sequence analysis helped to identify a number of editing and coding errors that were subsequently corrected in this and future waves. Sequence analysis offers survey managers and survey researchers a good starting point for further modelling work and for research on intervention methods (for an example see Durrant et al. 2015).

Extensions of this work may include considering different algorithm settings for the distance matrix (Hanly, 2013). Although not encountered here, results could potentially depend on the settings of the algorithm. Another interesting area of work may be the incorporation of further characteristics, such as time and date of the call, into the sequence analysis routine. A natural extension of the research presented is the use of sequence analysis for longitudinal data, where the wealth of data from previous waves, both in terms of previous calling patterns and survey data, may help to predict future call sequences and outcomes.

5. References

Aassve, A., Billari, F. C. and Piccarreta, R. (2007) Strings of adulthood: A Sequence analysis of young British women's work-family trajectories. *Eur. J. Popul.*, **23**, 369-388.

Abbott, A. and Hrycak, A. (1990) Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *Am. J. Sociol.*, **96**, 144-185.

Abbott, A. and Tsay, A. (2000) Sequence analysis an d optimal matching methods in sociology review and prospect. *Sociol. Method. Res.*, **29**, 3-33.

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

Bates, N., Dahlhamer, J. and Singer, E. (2008) Privacy concerns, too busy, or just not interested: Using doorstep concerns to predict survey nonresponse. *J. Off. Statist.,* **24**, 591-612.

Biemer, P. P., Chen, P. and Wang, K. (2013) Using level-of-effort paradata in non-response adjustments with application to field surveys. *J. R. Statist. Soc.* A, **176**, 147–168.

Blom, A., Jäckle, A. and 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* (eds. J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, B.-E. Pennell and T. W. Smith), 335-354. New York: Wiley and Sons.

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

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.

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

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

Durrant, G. B., D'Arrigo, J. and Steele, F. (2011) Using Field Process Data to Predict Best Times of Contact Conditioning on Household and Interviewer Influences. *J. R. Statist. Soc.* A, **174**, 1029-1049.

Durrant, G. B., D'Arrigo, J. and Steele, F. (2013b) Analysing Interviewer Call Record Data by Using a Multilevel Discrete-Time Event History Modelling Approach. *J. R. Statist. Soc.* A, **176**, 251-269.

Durrant, G. B., Groves, R. M., Staetsky, L. and Steele, F. (2010) Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys, *Publ. Opin. Q.*, **74**, 1-36.

Durrant, G. B., Maslovskaya, O. and 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* (forthcoming).

Elzinga, C. H. and Liefbroer, A. C. (2007) De-standardization of family-life trajectories of young adults: A Cross-national comparison using sequence analysis. *Eur. J. Popul.*, **23**, 225-250.

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

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

Gabadinho, A., Ritschard, G., Mueller N. S. and Studer, M. (2011) Analyzing and visualizing state sequences in R with TraMineR. *J. Stat. Software*, **40**, 1-37.

Gabadinho, A., Ritschard, G., Studer M. and Mueller, N. S. (2010) Mining sequence data in R with the TraMineR package: A User's guide. *Technical Report.* Geneva: University of Geneva.

Groves, R. M. and Heeringa, S. G. (2006) Responsive design for household surveys: tools for actively controlling survey errors and costs. *J. R. Statist. Soc.* A*,* **169**, 439-459.

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, http://www.ul.ie/sociology/pubs/wp2014-03.pdf.

Hanly, M. (2013) Generating Nonresponse Adjustment Variables using Sequence Analysis of Call Record Data. *Conference paper*. (nominated runner-up in the young researcher paper competition), European Survey Research Association, Slovenia.

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

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

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

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

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

Kreuter, F., Couper, M. and 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. and Kohler, U. (2009) Analyzing contact sequences in call record data. Potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. *J. Off. Statist.*, **25**, 203-226.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and Raghunathan, T. E. (2010b) Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys. *J. R. Statist. Soc.* A, **173**, 389–407.

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

Kulka, R. A. and Weeks, M. F. (1988) Towards the development of optimal calling protocols for telephone surveys: A conditional probabilities approach. *J. Off. Statist.*, **4**, 319-358.

Laflamme, F., Maydan, M. and 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. *Cybernet. Control Theor.,* **10**, 707-710.

Malo, M. A. and Munoz-Bullon, F. (2003) Employment status mobility from a life-cycle perspective: A sequence analysis of work-histories in the BHPS. *Demographic Res.*, **9**, 119-162.

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

McFall, S. L. and 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.

McVicar, D. and Anyadike-Danes, M. (2002) Predicting successful and unsuccessful transitions from school to work by using sequence methods. *J. R. Statist. Soc.* 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. and Kitamura, N. (1989) Molecular cloning and sequence analysis of cDNA for human hepatocyte growth factor. *Biochem. Bioph. Res. Co.*, **163**, 967-973.

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

Piccarreta, R. and Lior, O. (2010) Exploring sequences: a graphical tool based on multi-dimensional scaling. *J. R. Statist. Soc.* A, **173**, 165-184.

Pickery, J., Loosveldt, G. and Carton, A. (2001) The effects of interviewer and respondent characteristics on response behavior in panel surveys. A multilevel approach. *Sociol. Method. Res.*, **29**, 509-523.

Pollien, A. and Joye, D. (2014) Patterns of contact attempts in surveys. In *Advances in Sequence Analysis, Theory, Methods, Applications* (eds. P. Blanchard, F. Bühlmann and J.-A. Gauthier), 285-309. London: Springer.

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

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. and Hood, L. E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**, 674-679.

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.

Vassallo, R, Durrant, G.B., Smith, P.W.F. and Goldstein, H. (2014): Interviewer Effects on Nonresponse Propensity in Longitudinal Surveys: A Multilevel Modeling Approach, Journal of the Royal Statistical Society, Series A (issue to be assigned).

Wagner, J. (2013a) Adaptive Contact Strategies in Telephone and Face-to-Face Surveys. *Surv. Res. Methods*, **7**, 45-55.

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* (ed. F. Kreuter), 145-170. New Jersey: Wiley and Sons.

Weeks, M. F., Jones, B. L., Folsom, R. E. and Benrud, C. H. (1980) Optimal times to contact sample households. *Publ. Opin. Q.*, **44**, 101-114.

Weeks, M. F., Kulka, R. A. and Pierson, S. A. (1987) Optimal call scheduling for a telephone survey. *Publ. Opin. Q.,* **51**, 540-549.

West, B. T. and Groves, R. M. (2013) A propensity-adjusted interviewer performance indicator. *Publ. Opin. Q.*, **77**, 352-374.

Wu, L. L. (2000) Some comments on" Sequence analysis and optimal matching methods in sociology: Review and prospect". *Sociol. Method. Res.*, **29**, 41-64.

Tables and Graphs

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

Household	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

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

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

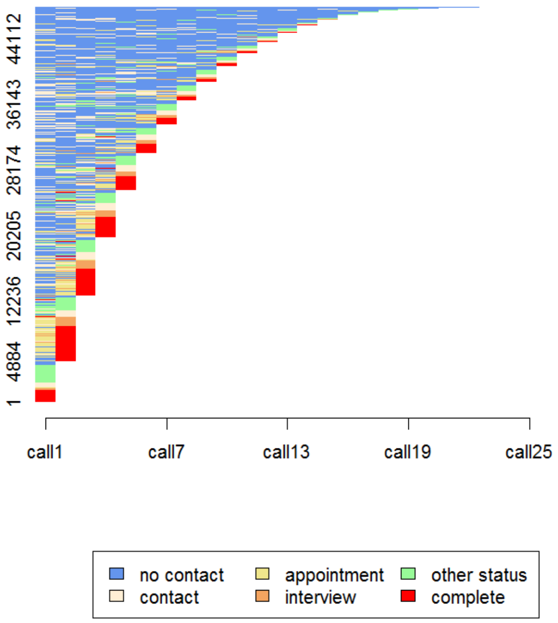

Figure 2: Sequence frequency plot: basic sequence plot of the 10 most frequent sequences.

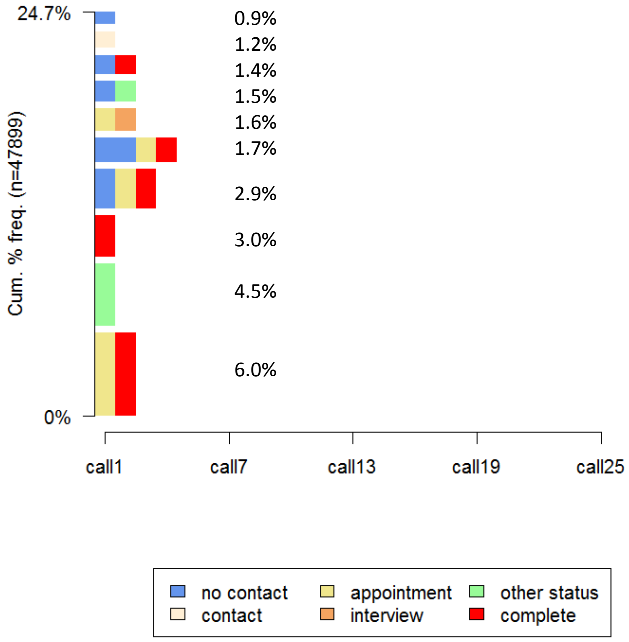

Table 2: Transition rate matrix indicating likely outcome of the next call given a particular call outcome at the current call.

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

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

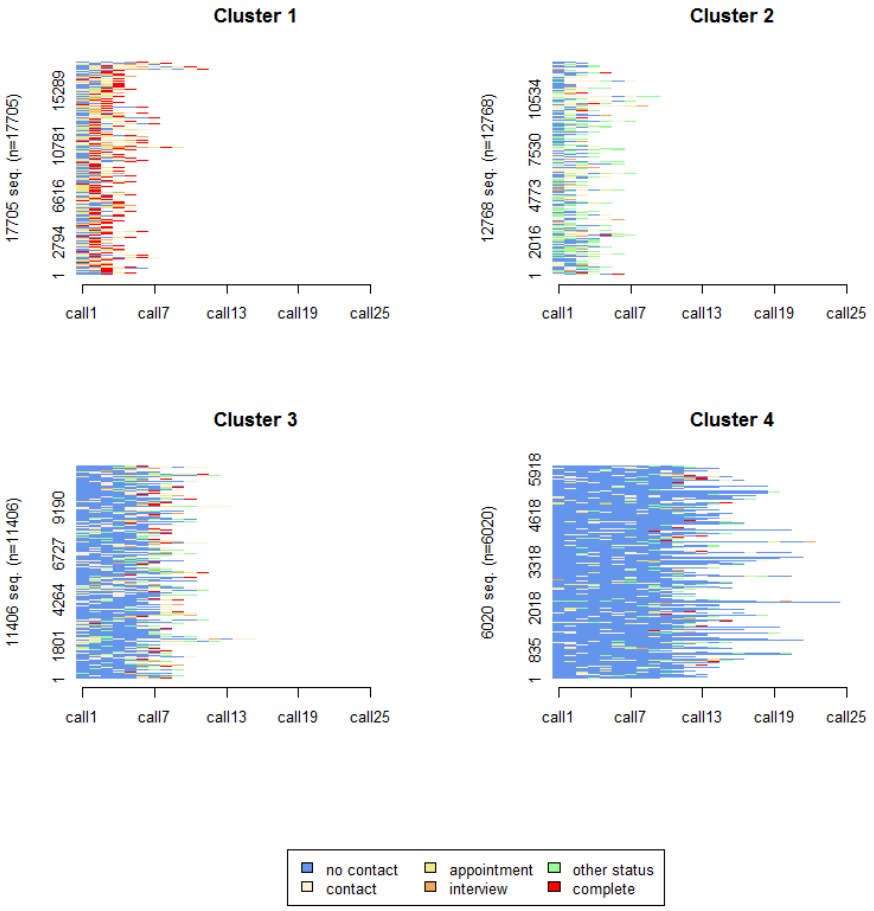

Figure 4: Multidimensional scaling plots

Figure 4a: Multidimensional scaling plot: first dimension

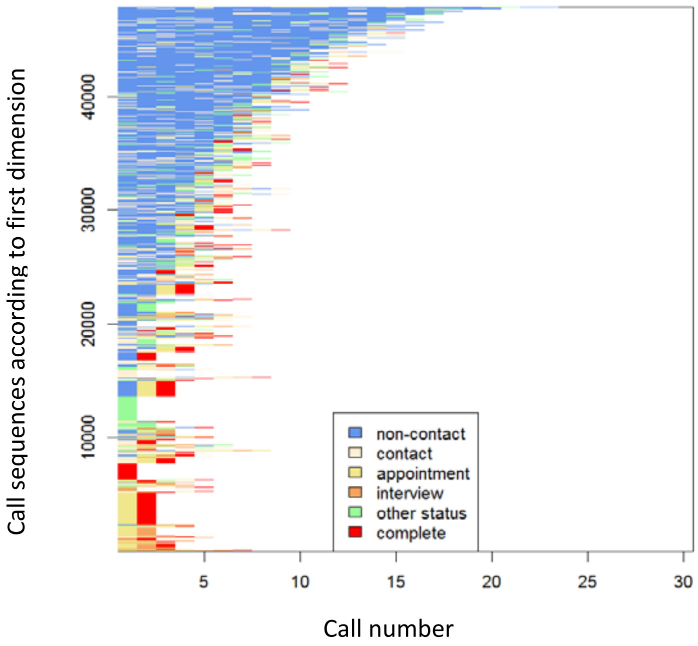

Figure 4b: Multidimensional scaling plot: first and second dimensions (the colour legend for the first and second dimensions is as in Figure 4a)

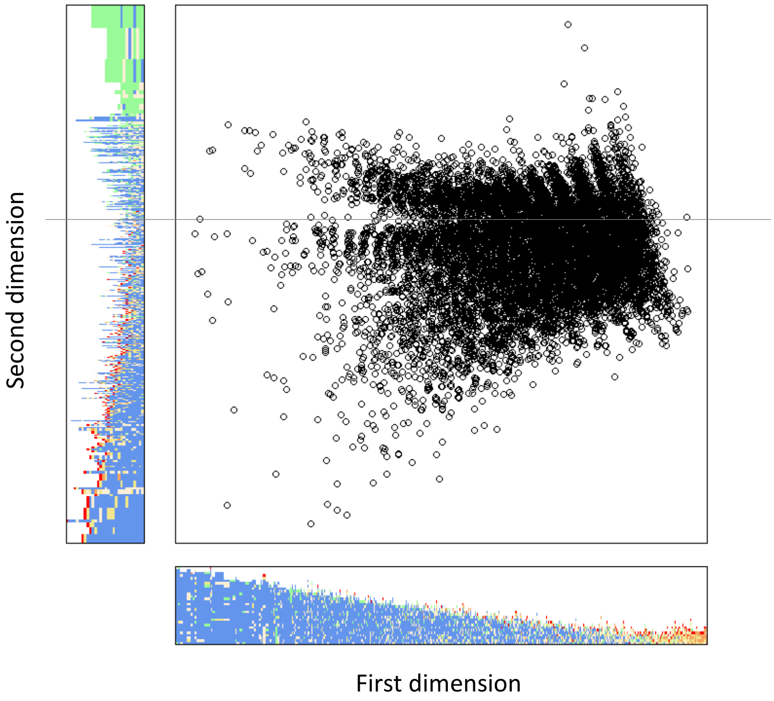

Figure 5: Multidimensional scaling plots for two different interviewers: interviewer A (anonymised interviewer number: 11002071) and interviewer B (anonymised interviewer number: 11006065) (the colour legend for the first and second dimensions is as in Figure 4a)

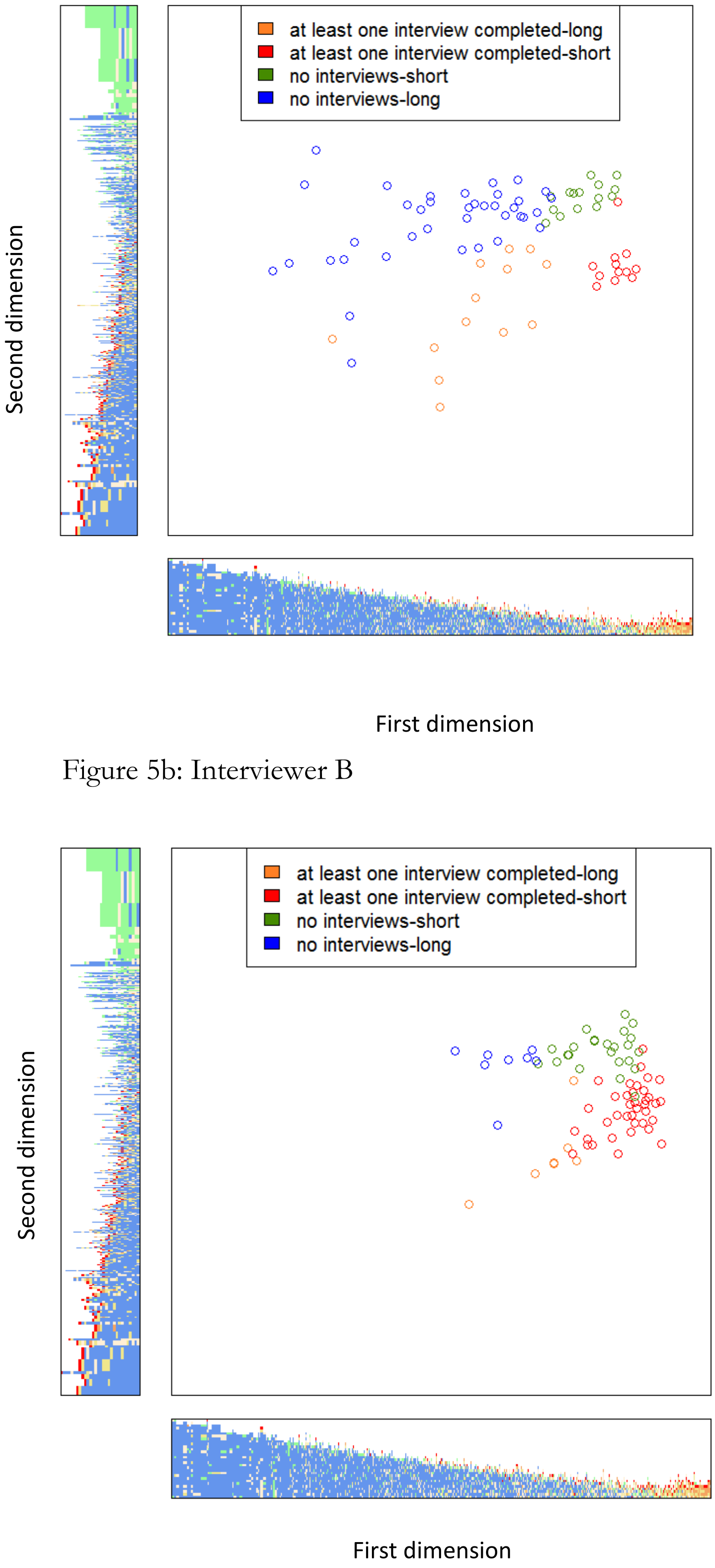

Figure 5a: Interviewer A

Figure 5b: Interviewer B