# Extracting and Rendering Representative Sequences*

Alexis Gabadinho, Gilbert Ritschard, Matthias Studer, and Nicolas S. Müller

Department of Econometrics and Laboratory of Demography, University of Geneva
40, bd du Pont-d'Arve, CH-1211 Geneva, Switzerland
alexis.gabadinho@unige.ch
http://mephisto.unige.ch/TraMineR

**Abstract.** This paper is concerned with the summarization of a set of categorical sequences. More specifically, the problem studied is the determination of the smallest possible number of representative sequences that ensure a given coverage of the whole set, i.e. that have together a given percentage of sequences in their neighbourhood. The proposed heuristic for extracting the representative subset requires as main arguments a pairwise distance matrix, a representativeness criterion and a distance threshold under which two sequences are considered as redundant or, identically, in the neighborhood of each other. It first builds a list of candidates using a representativeness score and then eliminates redundancy. We propose also a visualization tool for rendering the results and quality measures for evaluating them. The proposed tools have been implemented in our TraMineR R package for mining and visualizing sequence data and we demonstrate their efficiency on a real world example from social sciences. The methods are nonetheless by no way limited to social science data and should prove useful in many other domains.

**Keywords:** Categorical sequences, Representatives, Pairwise dissimilarities, Discrepancy of sequences, Summarizing sets of sequences, Visualization.

## 1 Introduction

In the social sciences, categorical sequences appear mainly as ordered list of states (employed/unemployed) or events (leaving parental home, marriage, having a child) describing individual life trajectories, typically longitudinal biographical data such as employment histories or family life courses. One widely used approach for extracting knowledge from such sets consists in computing pairwise distances by means of sequence alignment algorithms, and next clustering the sequences by using these distances [1]. The expected outcome of such a strategy is a typology, with each cluster grouping cases with similar patterns (trajectories). An important aspect of sequence analysis is also to compare the patterns of cases grouped according to the values of covariates (for instance sex or socioeconomic position in the social sciences).

A crucial task is then to summarize groups of sequences by describing the patterns that characterize them. This could be done by resorting to graphical representations

[*] This work is part of the Swiss National Science Foundation research project FN-122230 "Mining event histories: Towards new insights on personal Swiss life courses".

A. Fred et al. (Eds.): IC3K 2010, CCIS 128, pp. 94–106, 2011.

such as sequence index plots, state distribution plots or sequence frequency plots [2]. However, relying on these graphical tools suffers from some drawbacks. The summarizing task is mainly subjective and is rapidly complicated when there is a great number of distinct patterns, as is often the case.

Hence, we need more appropriate tools for extracting the key features of a given subset or data partition. We propose an approach derived from the concept of representative set used in the biological sciences [3,4]. The aim in this field is mainly to get a reduced reference base of protein or DNA sequences for optimizing the retrieval of a recorded sequence that resembles to a provided one. In this setting, the representative set must have “maximum coverage with minimum redundancy” i.e. it must cover all the spectrum of distinct sequences present in the data, including “outliers”.

Our goal is similar regarding the elimination of redundancy. It differs, however, in that we consider in this paper representative sets with a user controlled coverage level, i.e. we do not require maximal coverage. We thus define a representative set as a set of non redundant “typical” sequences that largely, though not necessarily exhaustively covers the spectrum of observed sequences. In other words, we are interested in finding a few sequences that together summarize the main traits of a whole set.

We could imagine synthetic — not observed — typical sequences, in the same way as the mean of a series of numbers that is generally not an observable individual value. However, the sequences we deal with in the social sciences (as well as in other fields) are complex patterns and modeling them is difficult since the successive states in a sequence are most often not independent of each other. Defining some virtual non observable sequence is therefore hardly workable, and we shall here consider only representative sets constituted of existing sequences taken from the data set itself.

Since this summarizing step represents an important data reduction, we also need indicators for assessing the quality of the selected representative sequences. An important aspect is also to visualize these in an efficient way.Such tools and their application to social science data are presented in this paper. These tools are included in our TraMineR library for mining and visualizing sequences in R [5].

## 2 Data

To illustrate our purpose we consider a data set from [6] stemming from a survey on transition from school to work in Northern Ireland. The data contains 70 monthly activity state variables from July 1993 to June 1999 for 712 individuals. The alphabet is made of 6 states: EM (Employment), FE (Further education), HE (Higher education), JL (Joblessness), SC (School) and TR (Training).

The three first sequences of this data set represented as distinct states and their associated durations (the so called State Permanence Format) look as follows

```
    Sequence
[1] EM/4-TR/2-EM/64
[2] FE/36-HE/34
[3] TR/24-FE/34-EM/10-JL/2
```

We consider in this paper the outcome of a cluster analysis of the sequences based on Optimal Matching (OM). The OM distance between two sequences $x$ and $y$, also known as edit or Levenshtein distance, is the minimal cost in terms of indels — insertions and deletions — and substitutions necessary to transform $x$ into $y$. We computed the distances using a substitution cost matrix based on transition rates observed in the data set and an indel cost of 1. The clustering is done with an agglomerative hierarchical method using the Ward criterion. A four cluster solution is chosen. Table 1 indicates some descriptive statistics for each of them. The clusters define four subsets grouping sequences with "similar" patterns, but to interpret the results we need to summarize their content, that is to do cluster labelling.

The sequence frequency plots in Fig. 1 display the 10 most frequent sequences in each cluster and give a first idea of their content. The bar widths are proportional to the sequence frequencies. The 10 most frequent sequences represent about 40% of all the sequences in cluster 1 and 2, while this proportion is 27% and 21% for clusters 3 and 4 due to a higher diversity of the patterns.

## 3 Extracting Representative Subsets

Our main aim is to find a small subset of non redundant sequences that ensures a given coverage level, this level being defined as the percentage of cases that are within a given neighbourhood of at least one of the representative sequences. We propose an heuristic for determining such a representative subset.

It works in two main steps. In the first stage it prepares a sorted list of candidate representative sequences without caring for redundancy and eliminates redundancy within this list in a second stage. It basically requires the user to specify a representativeness criterion for the first stage and a similarity threshold for evaluating redundancy in the second one. The strategy for selecting the sequences that will form the representative set can be summarized as follow:

1. Compute a representativeness score for each distinct sequence in the data set according to the selected criterion;
2. Sort all distinct sequences according to their score, in ascending order if the criterion increases with representativeness or in descending order otherwise;
3. Select a rule for possibly limiting the size of the candidate list;
4. Remove iteratively from the candidate list each sequence that has dissimilarity below a given threshold with any already retained sequence. This ensures that the final set of representative sequences does not contain any pair of sequences that are closer from each other than the predefined threshold.

### 3.1 Sorting Candidates

The initial candidate list is made of all distinct sequences appearing in the data. Since the second stage will extract non redundant representative sequences sequentially starting with the first element in the list, sorting the candidates according to a chosen representativeness criterion ensures that the “best” sequences given the criterion will be included. We present here four alternatives for measuring the sequence representativeness. The first three measures, *neighbourhood density, centrality* and *frequency* are directly obtained from the distance matrix itself, while the fourth is obtained by statistical

**Table 1.** Number of cases, distinct sequences and discrepancy within each cluster

| | N | Dist. seq. | Discr. |
|---|---|---|---|
| Cluster 1 | 265 | 165 | 18.3 |
| Cluster 2 | 153 | 88 | 23.5 |
| Cluster 3 | 194 | 148 | 27.9 |
| Cluster 4 | 100 | 89 | 37.2 |

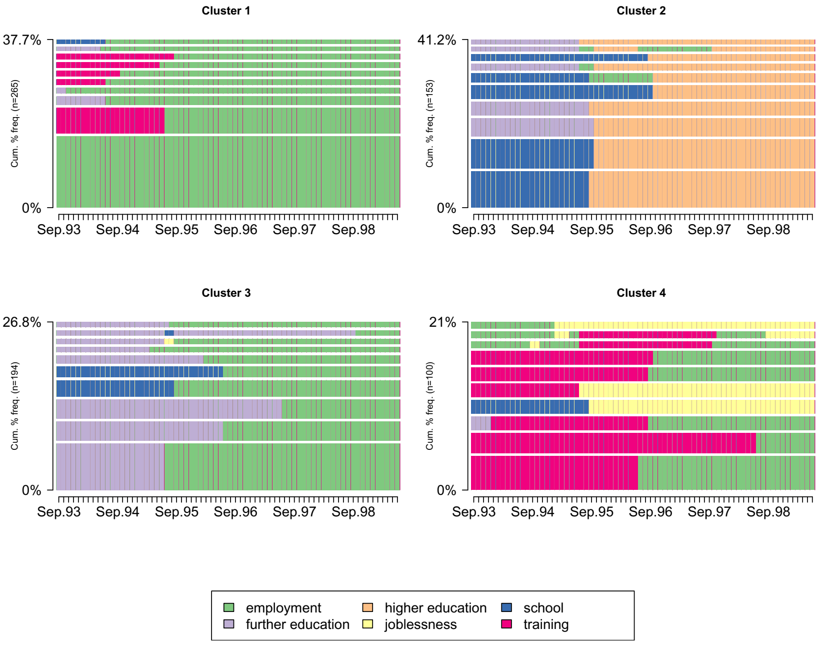

**Fig. 1.** 10 most frequent sequences within each cluster

modeling. Alternative vector of representativeness scores can also be provided by the user.

**Neighbourhood Density.** This criterion is the number — the density — of sequences in the neighbourhood of each candidate sequence. This requires to set the neighbourhood radius. We suggest to set it as a given proportion of the maximal theoretical distance between two sequences. Sequences are sorted in decreasing density order. This criterion leads indeed to sort the candidates according to their potential coverage. The neighbourhood density for each sequence in the set is obtained from the distance matrix by counting by row or column the number of distances that are less than a defined threshold (the neighbourhood radius).

**Centrality.** A classical representative of a data set used in cluster analysis is the *medoid*. It is defined as the most central object, i.e. the one with minimal sum of distances to all other objects in the set [7]. This leads to use the sum of distances to all other sequences, i.e. the centrality as a representativeness criterion. The smallest the sum, the most representative the sequence. It may be mentioned that the most central observed sequence is also the nearest from the 'virtual' true center of the set [8]. The centrality of each sequence in the set is obtained from the distance matrix by summing the distances by row or column.

**Frequency.** Sequences may also be sorted according to their frequency. The more frequent a sequence the more representative it is supposed to be. Hence, sequences are sorted in decreasing frequency order. This criterion makes sense in sets where some or all sequences appear more than once. This is indeed the density criterion with the neighbourhood radius set to 0.The frequency of each sequence in the set is obtained from the distance matrix by counting by row or column the distances that are equal to 0 (a distance of 0 between two sequences meaning that they are identical).

**Likelihood.** The sequence likelihood $P(s)$ is defined as the product of the probability with which each of its observed successive state is supposed to occur at its position. Let $s = s_1 s_2 \cdots s_\ell$ be a sequence of length $\ell$. Then

$$P(s) = P(s_1, 1) \cdot P(s_2, 2) \cdots P(s_\ell, \ell)$$

with $P(s_t, t)$ the probability to observe state $s_t$ at position $t$. The question is how to determinate the state probabilities $P(s_t, t)$. One commonly used method for computing them is to postulate a Markov model, which can be of various order. Below, we just consider probabilities derived from the first order Markov model, that is each $P(s_t, t)$, $t > 1$ is set to the transition rate $p(s_t|s_{t-1})$ estimated across sequences from the observations at positions $t$ and $t - 1$. For $t = 1$, we set $P(s_1, 1)$ to the observed frequency of the state $s_1$ at position 1. The likelihood $P(s)$ being generally very small, we use $-\log P(s)$ as sorting criterion. The latter quantity is minimal when $P(s)$ is equal to 1, which leads to sort the sequences in ascending order of their score.

### 3.2 Eliminating Redundancy

Once a sorted list of candidates has been defined, the second stage consists in extracting a set of non-redundant representatives from the list. The procedure is as follows:

1. Select the first sequence in the candidate list (the best one given the chosen criterion);
2. Process each next sequence in the sorted list of candidates. If this sequence is similar to none of those already in the representative set, that is distant from more than a predefined threshold from all of them, add it to the representative set.

The threshold for redundancy (similarity) is defined as a proportion of the maximal theoretical distance between two sequences and is the same as the neighbourhood radius that is used for computing the coverage (see below) or the neighbourhood density. For the OM distance between two sequences $(s_1, s_2)$ of length $(\ell_1, \ell_2)$ this theoretical maximum is

$$D_{max} = \min(\ell_1, \ell_2) \cdot \min\big(2C_I, \max(S)\big) + |\ell_1 - \ell_2| \cdot C_I$$

where $C_I$ is the indel cost and $\max S$ the maximal substitution cost.

### 3.3 Controlling Size/Coverage Trade-off

Limiting our representative set to the mere sequence(s) with the best representative score may lead to leave a great number of sequences badly represented. Alternatively, proceeding the complete list of candidates to ensure that each sequence in the data set is well represented may not be a suitable solution if we look for a small set of representative sequences.

To control the trade-off between size and representativeness, we use a threshold *trep* for the *coverage* level, that is the percentage of sequences having a representative in their neighbourhood. The coverage is recomputed each time that a sequence is added to the representive set and the selection process stops when the coverage threshold is reached.

Alternatively we can set the desired number of representatives and let the coverage unspecified. Fo example selecting the medoid as representative is done by choosing the *centrality* criterion and setting the number of representatives to 1.

## 4 Measuring Quality

A first step to define quality measures for the representative set is to assign each sequence to its nearest representative according to the considered pairwise distances. Let $r_1...r_{nr}$ be the $nr$ sequences in the representative set and $d(s, r_i)$ the distance between the sequence $s$ and the $i$th representative. Each sequence $s$ is assigned to its closer representative. When a sequence is equally distant from two or more representatives, the one with the highest representativeness score is selected. Hence, letting $n$ be the total number of sequences and $na_i$ the number of sequences assigned to the $i$th representative, we have $n = \sum_{i=1}^{nr} na_i$ . Once each sequence in the set is assigned to a representative, we can derive the following quantities from the pairwise distance matrix.

**Mean Distance.** Let $SD_i = \sum_{j=1}^{na_i} d(s_j, r_i)$ be the sum of distances between the $i$th representative and its $na_i$ assigned sequences. A quality measure is then

$$MD_i = \frac{SD_i}{na_i}$$

the mean distance to the $i$th representative.

**Coverage.** Another quality indicator is the number of sequences assigned to the $i$th representative that are in its neighbourhood, that is within a distance $dn_{max}$

$$nb_i = \sum_{j=1}^{na_i} \Big( d(s_j, r_i) < dn_{max} \Big) \ .$$

The threshold $dn_{max}$ is defined as a proportion of $D_{max}$. The total coverage of the representative set is the sum $nb = \sum_i^{nr} nb_i$ expressed as a proportion of the number $n$ of sequences, that is $nb/n$.

**Distance Gain.** A third quality measure is obtained by comparing the sum $SD_i$ of distances to the $i$th representative to the sum $DC_i = \sum_{j=1}^{na_i} d(s_j, c)$ of the distances of each of the $na_i$ sequences to the center of the complete set. The idea is to measure the gain of representing those sequences by their representative rather than by the center of the set. We define thus the quality measure $Q_i$ of the representative sequence $r_i$ as

$$Q_i = \frac{DC_i - SD_i}{DC_i}$$

which gives the relative gain in the sum of distances. Note that $Q_i$ may be negative in some circumstances, meaning that the sum of the $na_i$ distances to the representative $r_i$ is higher than the sum of distances to the true center of the set. A similar measure can be used to assess the overall quality of the representative set, namely

$$Q = \frac{\sum_i^{nr} DC_i - \sum_i^{nr} SD_i}{\sum_i^{nr} DC_i} = \sum_{i=1}^{nr} \frac{DC_i}{\sum_j DC_j} Q_i \ \ .$$

Representing all sequences by a sequence located exactly at the center of the set yields $Q = 0$.

**Discrepancy.** A last quality measure is the sum $SC_i = \sum_{j=1}^{na_i} d(s_j, c_i)$ of distances to the true center $c_i$ of the $na_i$ sequences assigned to $r_i$, or the mean of those distances $V_i = SC_i/na_i$, which can be interpreted as the discrepancy of the set [8].

## 5 Visualizing Representative Sequences

A graphical tool for visualizing the selected representative sequences together with information measures is included in the TraMineR package. A single function produces a "representative sequence plot" (Figure 2) where the representative sequences are plotted as horizontal bars with width proportional to the number of sequences assigned to them. Sequences are plotted bottom-up according to their representativeness score. Above the plot, two parallel series of symbols associated to each representative are displayed horizontally on a scale ranging from 0 to the maximal theoretical distance $D_{max}$. The location of the symbol associated to the representative $r_i$ indicates on axis $A$ the (pseudo) variance ($V_i$) within the subset of sequences assigned to $r_i$ and on the axis $B$ the mean distance $MD_i$ to the representative.

**Key Patterns.** The set of representative sequences extracted using the neighbourhood density criterion and a coverage threshold of 25% is displayed in Figure 2 for each of the four clusters of our example. The plots give clearly a more readily interpretable view of the content of the clusters than the frequency plots displayed in Figure 1. Detailed statistics about these sets are presented in Table 2 and overall statistics in Table 3.

The pairwise distances used are the optimal matching distances that we used for the clustering. The threshold $dn_{max}$ for similarity (redundancy) between sequences was set as 10% of the maximal theoretical distance $D_{max}$. The sequence length being

**Table 2.** Representative sequences by cluster, density criterion, coverage=25%

| | $na$ | (%) | $nb$ | (%) | $MD$ | $V$ | $Q$ |
|---|---|---|---|---|---|---|---|
| Cluster 1 | | | | | | | |
| $r_1$ | 265 | 100.0 | 99 | 37.4 | 27.6 | 18.3 | -50.7 |
| Cluster 2 | | | | | | | |
| $r_1$ | 153 | 100.0 | 48 | 31.4 | 39.1 | 23.5 | -66.7 |
| Cluster 3 | | | | | | | |
| $r_1$ | 194 | 100.0 | 55 | 28.4 | 40.2 | 27.9 | -44.3 |
| Cluster 4 | | | | | | | |
| $r_1$ | 54 | 54.0 | 16 | 16.0 | 33.7 | 22.7 | -0.0 |
| $r_2$ | 21 | 21.0 | 7 | 7.0 | 30.7 | 21.3 | 4.8 |
| $r_3$ | 25 | 25.0 | 6 | 6.0 | 39.9 | 23.2 | 18.8 |

**Table 3.** Comparing criterions with coverage of 25%, 50% and 75%

| | $nr$ | $COV$ | $MD$ | $Q$ | $nr$ | $COV$ | $MD$ | $Q$ | $nr$ | $COV$ | $MD$ | $Q$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster 1 | | | | | | | | | | | | |
| Density | 1 | 37.4 | 27.6 | -50.7 | 2 | 51.7 | 21.9 | -19.8 | 12 | 75.1 | 14.3 | 22.1 |
| Frequency | 1 | 28.3 | 30.5 | -66.6 | 3 | 53.2 | 15.6 | 15.0 | 20 | 75.1 | 8.2 | 55.3 |
| Likelihood | 1 | 28.3 | 30.5 | -66.6 | 3 | 53.2 | 15.6 | 15.0 | 14 | 75.1 | 9.9 | 45.9 |
| Centrality | 2 | 38.9 | 23.0 | -25.5 | 4 | 61.1 | 16.1 | 11.8 | 17 | 75.1 | 11.9 | 34.8 |
| Cluster 2 | | | | | | | | | | | | |
| Density | 1 | 31.4 | 39.1 | -66.7 | 3 | 55.6 | 17.8 | 24.0 | 8 | 75.2 | 12.0 | 48.8 |
| Frequency | 2 | 40.5 | 18.7 | 20.5 | 4 | 52.3 | 15.0 | 35.9 | 9 | 75.8 | 8.5 | 63.6 |
| Likelihood | 2 | 40.5 | 18.7 | 20.5 | 4 | 51.0 | 14.7 | 37.3 | 9 | 75.2 | 8.4 | 64.1 |
| Centrality | 2 | 26.1 | 31.1 | -32.6 | 10 | 64.7 | 13.4 | 43.1 | 15 | 78.4 | 9.5 | 59.6 |
| Cluster 3 | | | | | | | | | | | | |
| Density | 1 | 28.4 | 40.2 | -44.3 | 5 | 51.0 | 24.0 | 13.9 | 28 | 75.3 | 11.5 | 58.9 |
| Frequency | 2 | 32.0 | 34.4 | -23.5 | 6 | 51.5 | 19.3 | 30.6 | 34 | 75.3 | 9.3 | 66.6 |
| Likelihood | 2 | 30.4 | 32.1 | -15.2 | 6 | 51.5 | 21.9 | 21.3 | 31 | 75.3 | 9.9 | 64.5 |
| Centrality | 2 | 33.0 | 35.4 | -26.8 | 13 | 51.5 | 24.7 | 11.5 | 48 | 75.3 | 11.4 | 59.2 |
| Cluster 4 | | | | | | | | | | | | |
| Density | 3 | 29.0 | 34.6 | 7.0 | 10 | 51.0 | 22.7 | 38.9 | 33 | 75.0 | 10.5 | 71.7 |
| Frequency | 3 | 26.0 | 34.0 | 8.6 | 18 | 50.0 | 19.2 | 48.4 | 37 | 75.0 | 9.2 | 75.4 |
| Likelihood | 3 | 27.0 | 34.7 | 6.9 | 11 | 50.0 | 22.1 | 40.5 | 34 | 75.0 | 10.3 | 72.3 |
| Centrality | 14 | 35.0 | 30.9 | 17.0 | 26 | 51.0 | 21.2 | 42.9 | 45 | 75.0 | 10.6 | 71.6 |

$\ell = 70$, the indel cost 1 and the maximal substitution cost 1.9995, we get $D_{max} = 70 \cdot \min(2, 1.9995) = 139.96$.

The first cluster is represented by a sequence begining with a short spell of training followed by employment during the rest of the period. This single representative covers (within 10% of $D_{max}$) 99 sequences (37%) of the cluster (Table 2). Hence, this cluster is characterized by patterns of rapid entry into employment. The overall quality measures

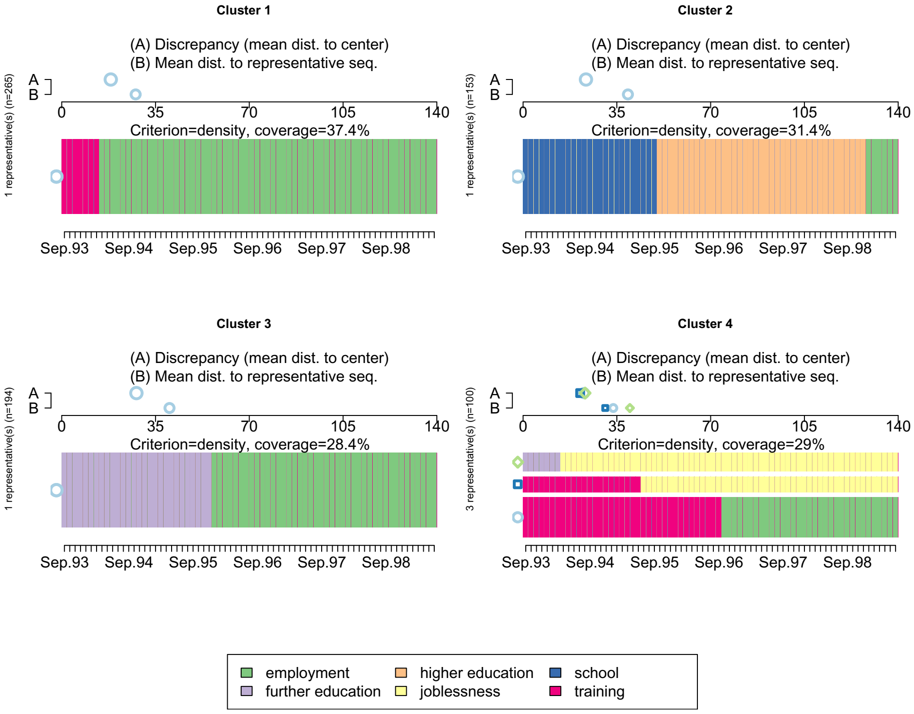

**Fig. 2.** Representative sequences selected with the density criterion, within each cluster - 25% coverage threshold

for the representative set are indeed the same as those for the single representative it contains.

The second cluster is described by a pattern leading to higher education and then to employment, starting with a spell of school. This pattern covers (have in its neighbourhood) 31% of the sequences. In cluster 3, the pattern is a transition to employment preceded by long (compared to Cluster 1) spells of further education.

The key patterns in cluster 4 were less clear when looking at the sequence frequency plot (Figure 1). The diversity of the patterns is high in this cluster which leads to the extraction of three non redundant sequences from the candidate list to achieve the 25% coverage: one is a long spell of training leading to employment and the two others are long spells of joblessness preceeded by either a short spell of further education or a long spell of training. Hence these trajectories can be characterized as less successful transitions from school to work. The overall quality measure reaches its highest level ($Q = 7\%$). The discrepancy is high in this group ($V = 37.2$) and the three selected representatives cover the sequence space so that representing the sequences with their assigned representative rather than by the center of the set significantly decreases the sum of distances.

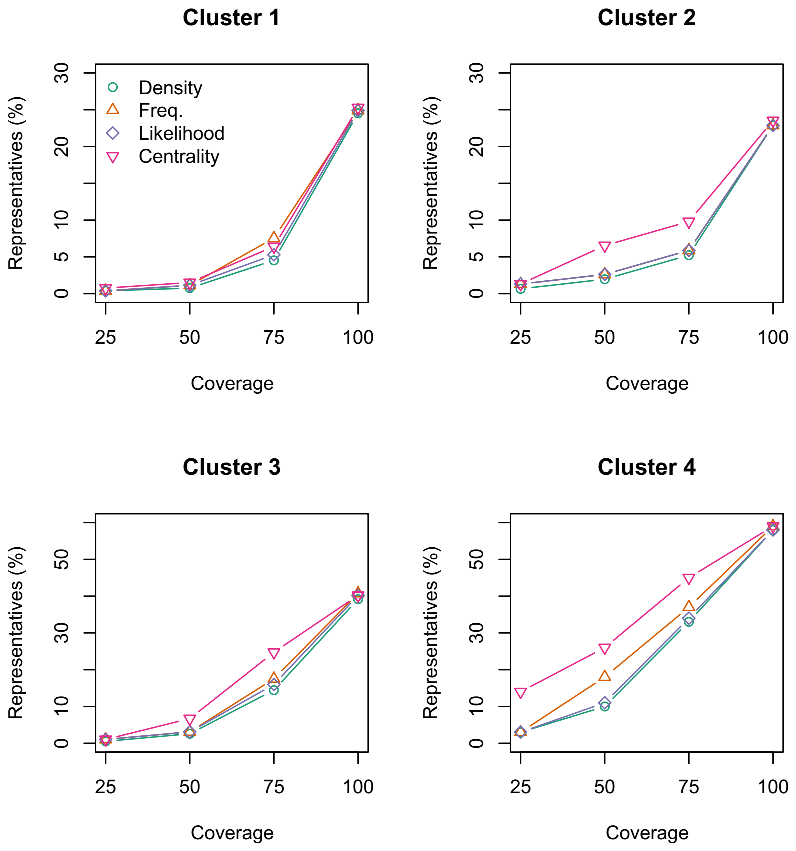

**Fig. 3.** Number of representative sequences (as the percentage of all sequences in the cluster) selected with several criterions, with trep of 0.25, 0.50, 0.75 and 1.0

## 6 Comparing Sorting Criterions

Sorting the candidate list according to the distance to the center yields poor results in many cases, as measured by the number of representatives needed for achieving a given coverage level. Indeed selecting the objects closest from the center of the group leads to poor representation of objects that are far from it, as shown in Figure 5. However, the centrality criterion may yield better overall quality measures with reduced coverage (50% and below). Depending on the spatial distribution of the sequences as defined by the distance matrix, uncovered sequences may indeed be much more distant from their attributed representative than from the center of the set.

The third part of Table 3 presents the results obtained after increasing the *trep* coverage threshold to 75%. As a consequence the proportion of well represented sequences

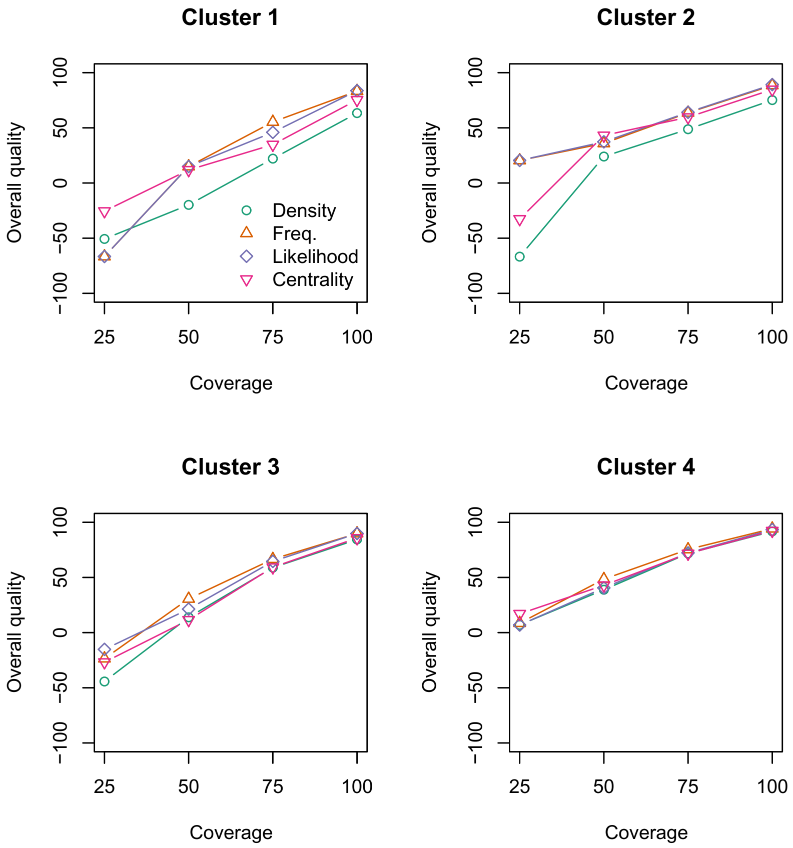

**Fig. 4.** Overall quality obtained with several criterions, with coverage of 25%, 50%, 75% and 100%

is now at least 75%. This gain comes however at the cost of a considerable increase in the number $nr$ of selected representative sequences. Full coverage is achieved with about one quarter of the sequences in clusters 1 and 2, while 60% of the sequences are needed in cluster 4 (Figure 3). Table 3 and Figure 4 show how increasing coverage leads to a decrease in mean distance to representative and an increase in overall quality. The mean distance to representative approaches 0 and is below the neigbourhood radius when full coverage is reached, while overall quality approaches 100%. Table 3 and Figure 3 confirm that the neighbourhood density criterion yields systematically the smallest number of representatives for each cluster and coverage level. The best results for the overall quality measure is obtained with the frequency criterion for three of the four clusters. Indeed with the frequency criterion the representatives that have the most sequences having a null distance to them (the highest frequency) are selected first, impacting favourably the overall quality measure.

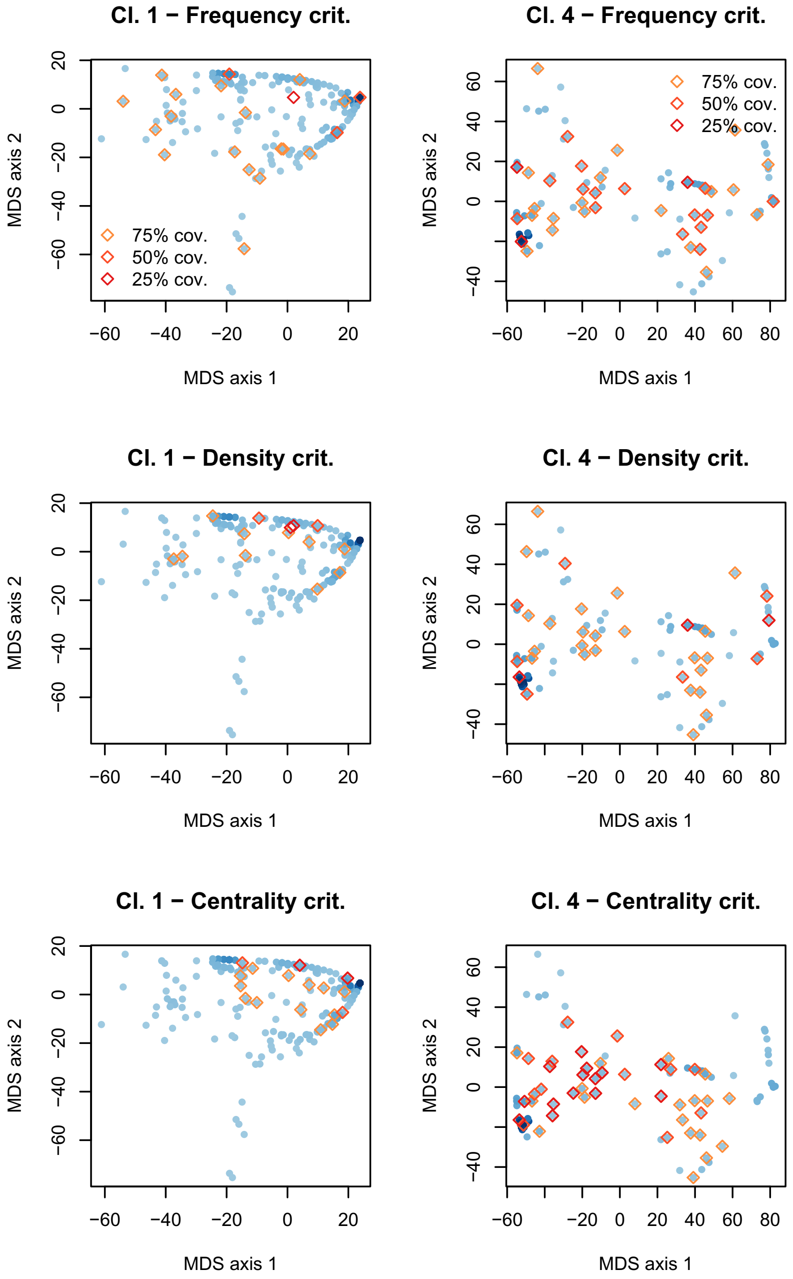

**Fig. 5.** Multidimensional scaling (MDS) representation of the pairwise distance matrix and selected representatives, with coverage of 25%, 50% and 75% - Cluster 1 and 4

## 7 Conclusions

We have presented a flexible method for selecting and visualizing representatives of a set of sequences. The method attempts to find the smallest number of representatives that achieve a given coverage. Different indicators have been considered to measure representativeness and the coverage can be evaluated by means of different sequence dissimilarity measures. The heuristic can be fine tuned with various thresholds for controlling the trade-off between size and quality of the resulting representative set. The experiments demonstrated how good our method is for extracting in an readily interpretable way the main features from sets of sequences. The proposed tools are made available as functions of the TraMineR R-package for categorical sequence analysis but are indeed not limited to sequence data sets and can be applied to dissimilarity matrices representing distances between any object type.

## References

1. Abbott, A., Tsay, A.: Sequence analysis and optimal matching methods in sociology, Review and prospect. Sociological Methods and Research 29(1), 3–33 (2000) (With discussion, pp. 34–76)
2. Müller, N.S., Gabadinho, A., Ritschard, G., Studer, M.: Extracting knowledge from life courses: Clustering and visualization. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 176–185. Springer, Heidelberg (2008)
3. Hobohm, U., Scharf, M., Schneider, R., Sander, C.: Selection of representative protein data sets. Protein Sci. 1(3), 409–417 (1992)
4. Holm, L., Sander, C.: Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 14(5), 423–429 (1998)
5. Gabadinho, A., Ritschard, G., Studer, M., Müller, N.: Mining sequence data in R with the TraMineR package: A user's guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva (2009)
6. McVicar, D., Anyadike-Danes, M.: Predicting successful and unsuccessful transitions from school to work by using sequence methods. Journal of the Royal Statistical Society. Series A (Statistics in Society) 165(2), 317–334 (2002)
7. Kaufman, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis. John Wiley and Sons, New York (1990)
8. Studer, M., Ritschard, G., Gabadinho, A., Müller, N.S.: Discrepancy analysis of complex objects using dissimilarities. In: Guillet, F., Ritschard, G., Briand, H., Zighed, D.A. (eds.) Advances in Knowledge Discovery and Management. Studies in Computational Intelligence. Springer, Berlin (2010) (forthcoming)
9. Clark, R.D.: Optisim: An extended dissimilarity selection method for finding diverse representative subsets. Journal of Chemical Information and Computer Sciences 37(6), 1181–1188 (1997)
10. Daszykowski, M., Walczak, B., Massart, D.L.: Representative subset selection. Analytica Chimica Acta 468(1), 91–103 (2002)