SOCIOLOGICAL METHODOLOGY
ASA AMERICAN SOCIOLOGICAL ASSOCIATION

00

MULTICHANNEL SEQUENCE ANALYSIS APPLIED TO SOCIAL SCIENCE DATA

*Jacques-Antoine Gauthier**
Eric D. Widmer†
Philipp Bucher‡
*Cédric Notredame***

Applications of optimal matching analysis in the social sciences are typically based on sequences of specific social statuses that model the residential, family, or occupational trajectories of individuals. Despite the broadly recognized interdependence of these statuses, few attempts have been made to systematize the ways in which optimal matching analysis should be applied multidimensionally—that is, in an approach that takes into account multiple trajectories simultaneously. Based on methods pioneered in the field of bioinformatics, this paper proposes a method of multichannel sequence analysis (MCSA) that simultaneously extends the usual optimal matching analysis (OMA) to multiple life spheres. Using data

We thank the editor and the anonymous reviewers for helpful comments and suggestions. Direct correspondence to Jacques-Antoine Gauthier, University of Lausanne, Faculty of social and political sciences (SSP), research center Methodology, Inequalities and Social Change (MISC), Bâtiment de Vidy, CH – 1015 Lausanne, Switzerland. Email: Jacques-antoine.gauthier@unil.ch.

*University of Lausanne
†University of Geneva
‡Swiss Institute of Bioinformatics and the École Polytechnique Fédérale de Lausanne
**Centre for Genomic Regulation, Barcelona, and Centre National de la Recherche Scientifique, Marseille

from the Swiss household panel (SHP), we examine the types of trajectories obtained using MCSA. We also consider a random data set and find that MCSA offers an alternative to the sole use of ex-post sum of distance matrices by locally aligning distinct life trajectories simultaneously. Moreover, MCSA reduces the complexity of the typologies it allows to produce, without making them less informative. It is more robust to noise in the data, and it provides more reliable alignments than two independent OMA.

1. INTRODUCTION

Most multivariate analyses using longitudinal data are based on hard causal models in which one or several independent variables predict the future actualization of some state of a dependent variable. Optimal matching analysis (OMA) offers a more descriptive perspective, that does not emphasize the causal priority of some variables over the others but aims at elaborating a systemic view on the social phenomena that develop over time. However, most applications of OMA have been limited to one dimension at a time, a serious shortcoming for empirical analyses. This paper develops a multichannel sequence analysis (MCSA) which enables researchers to describe individual trajectories on several dimensions simultaneously.

Various empirical studies (Elder 1985; Clausen 1986; Kohli 1986; Levy 1991, 1996; Giele and Elder 1998; Heinz and Marshall 2003; Mortimer and Shanahan 2003; Levy et al. 2005; Macmillan 2005) emphasize the multidimensionality of life trajectories based on social, psychological, and biological factors that interact over time (Wetzler and Ursano 1988; Spruijt and de Goede 1997; Repetti, Taylor, and Seeman 2002; Lesesne and Kennedy 2005). A major problem with research on life trajectories, however, lies in the fact that the researcher is confronted with a variety of unequally linked sequences unfolding at various speeds (Abbott 1992). Life course studies therefore require the integration of seemingly heterogeneous trajectories into a unique empirical model (Levy et al. 2005), an ambitious task that even regression-based models cannot accomplish (Esser 1996). In this perspective, Abbott (2001:151) insists on using sequential data as multicase narratives to uncover patterns of careers rather than looking for causal models.

Many social scientists have used OMA to model life trajectories. Nevertheless, since its emergence in the social sciences, OMA has neglected the multidimensionality of life trajectories. Social scientists have

always lacked a standard approach for undertaking multidimensional sequence analysis of life trajectories. To fill this gap, the present study proposes multichannel sequences analysis (MCSA), a computational approach that makes practical improvements to optimal matching algorithms at two levels.[1] First, it systematizes approaches for dealing with multidimensionality using OMA. Second, it accounts simultaneously for local interdependencies among different social statuses present at each point of the alignment process for all channels.[2] Third, it offers practical improvements toward visualizing parallel processes occurring in various life spheres, a key element to describe and interpret sets of individual trajectories (Tufte 1997; Müller et al. 2008; Piccarreta and Lior 2010).

In this study, we first present the quantitative methods available at the moment in the social sciences for dealing with the multidimensionality of the life course and describe in detail how the method works. To this end we also briefly present an example of substantive results produced by MCSA using social science data. We then illustrate the potential of MCSA by testing its validity and reliability using various formal criteria. Finally, we use random data to compare several multidimensional approaches using OMA.

2. QUANTITATIVE APPROACHES TO LIFE HISTORIES

There are a few methodological options for dealing with the multidimensionality of life trajectories. The most often used is event history analysis (EHA; Blossfeld and Rohwer 1995) and sequence analysis (SA; e.g., Sankoff and Kruskal 1983; Abbott and Hrycak 1990), while some attempts have been made with latent class methods (Macmillan and Eliason 2003) and life history graphs (Butts and Pixley 2004).

The latter focuses on internal configurations of the life course to reveal general sociological patterns. It uses a formal definition of life

[1] The computations presented in this paper are encapsulated in the program SALTT (Search Algorithm for Life Trajectories and Transitions), an open-source freeware program written in C (Notredame, Bucher, Gauthier, and Widmer 2005). The package and its documentation can be downloaded from: http://www.tcoffee.org/saltt/. Recently, TraMineR, a package of the R software environment, has allowed performing OMA and MCSA (Gabadinho et al. 2008). Otherwise, computations are made using SAS (Sas Institute 2004).

[2] By "channel," we mean each sequence of statuses that constitute the multidimensionality under study.

history that applies social network analysis at an intrapersonal level. Individual histories are expressed as structural relationships between life spells, such as centrality, betweenness, or closeness (Wasserman and Faust 1994). Life history graphs (LHG) take multidimensionality into account, but they put little emphasis on time as it focuses on the overlap of life spells for a given individual. The use of latent class methods (LCM) is based on transition probabilities. It allows identifying subsamples characterized by typical (i.e., most probable) family and occupational roles configurations over time. Unfortunately, methodological limitations make it difficult for LCM to consider more than a few time points and to represent life courses at an individual level.

Event history analysis estimates time-to-event or risk functions concerning, for instance, the occurrence of specific events, such as divorce or entry into the labor market, that are then used as dependent variables in different regression models (e.g., Kaplan-Meier or Cox). EHA provides strong information at a population level. However, the information concerning the unfolding of individual life history is limited to a dichotomous variable (the occurrence or not of a given event). A major strength of these methods is to allow statistical testing of the model, but they show limited sensitivity to temporality. Overall, LHG, LCM, and EHA are insufficient to account for the multidimensionality of life trajectories because they fail to "take a narrative approach to social reality" (Abbott 2001:185). In contrast, sequence analysis techniques and, more specifically, optimal matching analysis take the entire sequences of statuses held by individuals over a given period of time (e.g., family, occupational) as the analytical unit to find chronological patterns of stability and change (George 1993). Thus, each individual life course is modeled as a specific sequence of social statuses that may be expressed as a specific character string.

For instance, the sequence *aaaabbcccc* may describe the family trajectory of an individual over ten years (e.g., between the ages of 18 and 27), with *a* standing for living with both biological parents, *b* for living alone, and *c* for living in a couple. Basically, OMA involves making pairwise comparisons between individual sequences of statuses to evaluate how similar they are.[3] This is accomplished by counting the minimal (weighted) number of elementary operations (known as

[3] There are promising techniques for multiple sequence alignment, whereby all sequences are simultaneously compared to all others, but these tech-

"costs") of insertion, deletion,[4] and substitution that are necessary to transform one sequence into another (Sankoff and Kruskal 1983; Abbott and Hrycak 1990).[5] For instance, in Figure 8, shown later in this paper, one has to delete one *m* and then substitute two times *n* for *m* in the sequence $A^c = mmllm$ to transform it into $B^c = nlln$. Among all possible ways to transform sequence A^c in sequence B^c, the one associated with the smallest cost is obtained through dynamic programming (Needleman and Wunsch 1970) and is called the *optimal distance* between two sequences.[6] The alignment of two life-as-sequences takes into account both the relative position of specific statuses in each individual trajectories and the process of their unfolding over time. Moreover, as the modeling of the sequences is only limited by the number of time points and that of possible characters, the possible individual variability of sequence rapidly becomes huge. Thus, the distance computed by OMA summarizes in an elegant manner the extent to which life courses are similar. The smaller the distance between two life trajectories, the more similar they are. Once all pairwise alignments are computed, the researcher performs a cluster analysis on the resulting distance matrix to reveal types of individual trajectories.[7] Eventually, the typologies stemming from the two latter steps may be used as categorical variables in secondary analyses (cross-tabulations, regressions, and so forth).

We now turn to the more general issue of the extent to which OMA can be systematically and straightforwardly applied to multidimensional trajectories. Our goal is to evaluate the ability of OMA to adequately model two main tenets of the life course paradigm. The first one (the principle of linked lives) states that individuals participate

niques are poorly suited to large samples and divergent sequences (Claverie and Notredame 2003).

[4] Insertion and deletion are equivalent and are referred to as *indel*.

[5] The question of costs necessary to align sequences is a central methodological debate in the use of OMA by social scientists (e.g., Abbott and Tsay 2000; Levine 2000; Wu 2000). Recently, significant advances toward empirical, data-based cost-setting offer objective means of defining the relationships between elements to be compared (Gauthier et al. 2009; Aisenbrey and Fasang 2010).

[6] For a closer description of the algorithm, see, for example, Kruskal (1983).

[7] The general principle of cluster analysis consists of grouping individuals according to a systematic rule. In this paper, we use the hierarchical Ward's algorithm, which aims to minimize the intragroup and maximize the intergroup variance of interindividual distances.

simultaneously in various social spheres and that the corresponding positions they hold in each are interdependent, as is the case between family and occupational careers (e.g., Heinz 2003). The second tenet (lifelong development) emphasizes the fact that the interdependence of multiple social participation at an individual level may vary continuously over time (Elder et al. 2003). To develop a methodology that corresponds to the two tenets, we investigate the options currently available and define the prerequisites for simultaneously modeling distinct life sphere alignments, while also taking their interdependence into account.

3. OPTIMAL MATCHING ANALYSIS

Due to methodological or computational limitations, sequence analysis in the social sciences has until recently focused mainly on (1) one-dimensional social trajectories; (2) recoded statuses belonging to different social spheres prior to data processing; or (3) summed interindividual distances measured independently on distinct one-dimensional trajectories. In doing this, measurement of the similarity between two pairs of trajectories does not take full account of the possible interactions that may occur at some points of these linked sequences during the alignment process.

Three different strategies have been used in OMA to measure life trajectories along several dimensions. The first consists of using typologies from one-dimensional analyses (e.g., occupational trajectories) as response variables in a logistic regression model that includes indicators of other trajectories (e.g., number of children) as predictor variables (Widmer et al. 2003; Levy, Gauthier, and Widmer 2006). This approach to a large extent disregards the longitudinal information provided by the predictor variables.

A second strategy involves retrospectively combining the results obtained from various independent OMA into distinct types of trajectories (e.g., Han and Moen 1999). Since this approach sums interindividual distances from consecutive OMA, it is akin to cross-analyzing typologies stemming from distinct one-dimensional analyses of the same individuals. The main problem with such an approach is that it does not accurately take into account the local or temporal interdependence of the trajectories under study, because the respective types they

belong to are modeled independently of one another. Moreover, this combination of typologies produced by cross-tabulating the categorical variables stemming from one-dimensional OMA may lead to an overestimation of the number of relevant types, with many types being poorly populated and therefore noninformative. In particular, the approach suffers from a lack of parsimony and potential sensitivity to noisy data, as we demonstrate below. Furthermore, as each dimension is analyzed and clustered independently, it is impossible to use regular clustering quality estimates to decide on the number of types present in the data (Mojena 1977; Milligan and Cooper 1985, 1987). Moreover, the data in each dimension may not be equally reliable or informative. While the combination of alternative channels may compensate for this inequality, the separate treatment of dimensions will lead to spurious alignments, which may then result in the creation of artificial typologies.

A third and more interesting strategy is based on combining two or more alphabets (e.g., Stovel, Savage, and Bearman 1996; Abbott and Hrycak 1990; Blair-Loy 1999; Pollock 2007; Dijkstra and Taris 1995;[8] Elzinga 2003). An alphabet is a collection of characters bijectively associated with an ensemble of distinct statuses, and characteristic of a given life course dimension (e.g., family, occupational, residential).

For this purpose, an extended alphabet is generated by combining individual alphabets associated with specific channels. There is, however, a problem associated with this strategy: since it allows many possibilities for estimating the substitution costs associated with the extended alphabet, it becomes increasingly difficult to justify the choice of a given cost scheme as the number of categories grows larger and more heterogeneous.

Furthermore, depending on the number of channels, the extended alphabet becomes uncomfortably large (Han and Moen 1999). Take, for instance, two channels with three statuses. Family statuses are given a specific code for singlehood, marriage, or divorce/widowhood. Occupational status is recorded as "at home," "part-time," or "full-time." In this scenario, there is no rationale for deciding *a priori* how to set costs stemming from the combination of "at home"/"marriage" versus "single"/"part-time or other statuses." Moreover, each dimension's local contribution to the overall interindividual distance, as well

[8] These authors use a different algorithm from that of Needleman and Wunsch (1970), on which many OMA are based.

as the particular unfolding of each set of linked trajectories, remains hidden or unknown.

4. MULTICHANNEL SEQUENCE ANALYSIS

The multidimensional approach we have developed is based on the assumption that taking this local contribution into account differs from using an extended alphabet, since each dimension differentially influences the final alignment. Therefore, a systematic approach is needed in order to deal with these issues. We propose a multidimensional approach in which (1) the dimensions under study are used simultaneously during the alignment process; (2) no enumeration of an extended alphabet is needed; and (3) the combination of cost estimations is as explicit as possible and is dealt with using a standard parameter. Given an alphabet containing a finite number of characters, take two sequences I and J based on a finite number of characters belonging to the alphabet.[9] Consider the costs associated with insertions and deletions (henceforth called *indel* and abbreviated d), as well as with the substitution costs given by a cost matrix C, where $C_{s_i s_j}$ is the cost for aligning S_i, the i^{th} character of I against S_j, the j^{th} character of J.[10] In this paper, for simplification purposes, we set a cost of one to all substitutions involving two different characters. The substitution of a character with itself yields a cost of zero, and the costs associated with *indel* are set to the half of that of a substitution.[11] The optimal alignment score can then be computed using the following recursion:

[9] In practice, most algorithms are based on existing sets of characters, as is, for instance, the English alphabet (26 characters) or the ASCII characters table (127 characters that may be complemented with 128 extended ASCII codes). Taking into account a greater number of characters is not a limitation per se but may require some programming.

[10] In the context of this paper, we consider that the matching of identical characters yields a null score and that mismatches are associated with same sign, nonzero, finite costs, although other cost schemes may be found, notably in biology (Durbin et al. 2002).

[11] We choose this option to simplify the exposition. Many other weighting schemes for substitutions and/or insertion/deletion may be considered (Thompson et al. 1994; Durbin et al. 2002; Widmer et al. 2003). We propose elsewhere a method that estimates differentiated costs on an empirical basis (Gauthier et al. 2009)

$$F(i, j) = \min \begin{cases} F(i-1, j-1) + C_{s_i s_j} \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases} . \quad (1)$$

Each line in equation (1) defines a possible optimal match score of two subsequences, whether it is less costly at this point to insert, delete, or substitute characters to fully align the subsequences. For instance, $F(i-1, j-1)$ corresponds to the optimal match score of a subsequence containing the 1 to $i-1$ characters of sequence I against a subsequence containing the symbols 1 to $j-1$ in sequence J. As such, this equation defines a recursion in which the score of any alignment $F(i, j)$ can be estimated by considering an optimal extension of the three shorter alignments $F(i-1, j)$, $F(i-1, j-1)$, and $F(i, j-1)$. Considering that each of these shorter alignments was already an optimal matching of associated substrings, $F(i, j)$ will also be optimal (Durbin et al. 2002:20).[12]

We take the OMA concept a step further and extend it to the use of different information sources associated with individual trajectories. We name it multichannel sequence analysis (MCSA). In MCSA, each individual is associated with two or more distinct channels, each tapping a distinct life trajectory within a specific sphere (e.g., occupation, family, housing, location, health) by means of a specific alphabet. Channels associated with a given individual are synchronized so that, for example, the x^{th} character of the family channel and the y^{th} character of the occupational channel correspond to the same year for a given individual. For instance, given two individuals A and B, one can express the MCSA example given in Figure 8 as two bidimensional sequences:

$$A = \{(m, z), (m, t), (l, t), (l, t), (m, t)\} \quad \text{and}$$
$$B = \{(n, y), (l, z), (l, z), (n, z)\},$$

where each doublet in parentheses characterise the situation at a given time point; the first and second positions in the doublet correspond to the channels of family and occupational participation respectively. Once

[12] Of course, this strategy relies on the assumption that each position is independent and that the alignment scores are additive.

defined for an individual, these doublets remain the same throughout the alignment procedure. Optimal matching analysis is based on a recursive algorithm that parses a pair of sequences from the first to the last element in an array and then estimates an optimal score *at each point* of the alignment (Sankoff and Kruskal 1983). A given optimal solution for two substrings of the sequences A^c and B^c does not imply that the optimal solution will be the same for any extension of these substrings. The optimal distance is given only after the algorithm has been applied to the entire sequences to be aligned.[13] Therefore, our goal is to analyze multiple social participations while taking into account what each pair of nested individual statuses contributes over time to the overall similarity between two individual life courses. The method is general in the sense that it can use as many channels as needed, with the only condition being their synchronization.

In practice, taking into account synchronized channels within an OMA framework, as defined by equation (1), is relatively straightforward and only requires adapting the substitution costs $C_{s_i s_j}$ and the *indel* terms so that they reflect the relationship between equivalent channels. The multichannel version of these terms can be expressed as follows:

$$C_{s_i s_j} = \frac{\sum_{C=1}^{Nc} C^c_{s_i s_j}}{Nc}. \tag{2}$$

While a single cost matrix is used to match two individual life trajectories in standard OMA, our approach considers two or more channels per individual and uses one cost matrix for each channel. These cost matrices are standard and can be generated using any appropriate strategy, such as unitary, knowledge-based, or data-based (Gauthier et al. 2009;[14] Aisenbrey and Fasang 2010). For instance, in equation (2), a channel-specific cost matrix (C^c) is associated with each channel. This matrix controls the cost of matching any character in the

[13] For instance, using the cost schemes presented above, aligning $A^c = \{m\}$ with $B^c = \{1\}$ implies either a substitution (mismatch) or two insertion/deletions, whereas aligning $A^{c\prime} = \{ml\}$ – where $A^{c\prime}$ is equal to A^c plus character l – with $B^{c\prime} = \{l\}$ calls for an insertion/deletion followed by a match.

[14] In this paper, we present an empirical method for defining substitution costs using a data-based iterative procedure.

channel in question with any counterpart character for another individual. Formally, given two individuals A and B, each associated with two channels c and d, $C^c_{s_i s_j}$ will be the cost associated with matching the i^{th} character of channel c for individual A with the j^{th} character of channel c for individual B. $C^d_{s_i s_j}$ will be the cost of matching the i^{th} character of channel d for individual A with the j^{th} character of channel d for individual B. Eventually, the contribution of channels c and d is averaged to yield the final cost associated with the matching of positions i and j for the two individuals, where Nc stands for the number of channels.[15] Costs for the insertion/deletion (*indel*) of each channel are averaged the same way. An alternative used below is to define *indel* as the average off-diagonal value (AOD) of the corresponding substitution matrix (Thompson et al. 1994). This procedure can be extended to any number of channels. In the above example, as cost matrices are unitary, matching the doublets (m, t) with (l, z) is more costly than matching (l, t) with (l, z) as the latter doublets share a common character. Hence, the optimal MCSA alignment presented in Figure 8 inserts a doublet of *indels* in order to match the most similar doublets.[16] Eventually, the raw score of this bidimensional alignment is computed as $2 * indels + 2 *$ (mismatch/mismatch) $+ 2 *$ (match/mismatch) $= 2 * 0.5 + 2 * 2 + 2 * 1 = 7$.

There are several ways to compute the distance from an optimal pairwise alignment. We may use the raw score provided by the algorithm,[17] or the percentage of identity (PID) between the two sequences (National Centre for Biotechnology Information 2004; May 2004). PID corresponds to the number of aligned identical characters, divided by the length of the longer sequence (see examples in Figure 8). It is an interesting measure, as it is approximately normally distributed (Doolitle 1981) and gives a useful indication concerning the common structure of two sequences (Raghava and Barton 2006).

[15] To simplify the exposition, we have set the combination of the substitution costs at one point of the aligned sequences at the average value of the two substitution costs involved at this point. Future developments should implement some alternative ways of dealing with the relationship between local scores.

[16] Matching two doublets leads to either two matches, one match and one mismatch, or two mismatches. Following the cost scheme used, the resulting costs may be quite differentiated.

[17] When the length of the sequences to align differ, the resulting distance between them is normalized by dividing it through the length of the longer sequence.

5. EMPIRICAL ILLUSTRATION

To test this method and illustrate its strength with an empirical example, we use data from the Swiss Household Panel (Tillmann and Zimermann 2004). It includes in its third wave a retrospective questionnaire that asks respondents to provide information on their educational, family, and occupational status from birth to the year of the interview. Each change in status is therefore associated with a starting date and an ending date. Every year, occupational trajectories are coded using a seven-category code scheme: full-time employment; part-time employment; positive interruption such as a sabbatical or trip abroad; negative interruption, such as unemployment or illness; full-time housework; retirement; and full-time education.

A ten-category code scheme is used for family trajectories: living with a biological father and mother; living with only one biological parent, either mother or father; living with one biological parent and her or his partner; living alone; living with a partner; living with a partner and one's own biological child; living with a partner and a nonbiological child; living with one's own biological child without a partner; living with friends; other situations. A 12-category scale is used to describe education-to-work trajectories.[18] Given that individual life trajectories require substantial time to differentiate from one another, and to build sequences that are as complete and informative as possible, we consider here only the individuals aged 45 and older who answered the retrospective questionnaire ($N = 2{,}212$). As we further restrict our sample to individuals whose trajectories contain less than 50% missing data, our final data set contains 1,847 individuals (54% women, 46% men) characterized by two sequences of statuses describing their family and occupational lives from birth to age 45.[19] Technically, MCSA may be applied to any number of sequences, without restriction regarding censored or incomplete data. However, from a sociological point of view, it makes sense to compare life course sequences that have about

[18] This scale is a combination of the seven categories of educational attainment following the classification by the Swiss Federal Statistical Office, which range from compulsory education to university degree, and five post-educational occupational statuses (full-time, part-time, household, unemployment, other).

[19] When comparing education-to-work trajectories, we use sequences ranging from age 0 to 25, with 2,153 individuals aged 25 or older.

the same size, which results here in the fact that the youngest cohorts are not taken into account.[20]

Our choice for an empirical example on which to test MCSA reflects the importance of debates about the influence of sociostructural factors on the divergent occupational and family trajectories of women and men (Moen 1985; Höpflinger and Debrunner 1991; Sheridan 1997; Levy, Gauthier, and Widmer 2006), as well as the availability of high-quality data on family and occupational status over entire lives. We use one channel to describe occupational statuses and another to model family trajectories over time. The interindividual distance matrix is computed by means of MCSA.[21] We then run a cluster analysis on that matrix using Ward's hierarchical method to reveal coherent types of individual trajectories. We use *stopping rules* in order to estimate the relevant number of clusters to retain (e.g., Milligan and Cooper 1985; SAS Institute 2004).[22] We eventually decide to keep three clear-cut, bidimensional types of individual trajectories, in addition to one residual category not presented here. Figures 1, 2, and 3 present three contrasted illustrations of the visualization potentialities offered by a multichannel approach to individual life courses—namely, simultaneous local analysis and parallel visualization of interdependent social trajectories (e.g., see Tufte 1997).

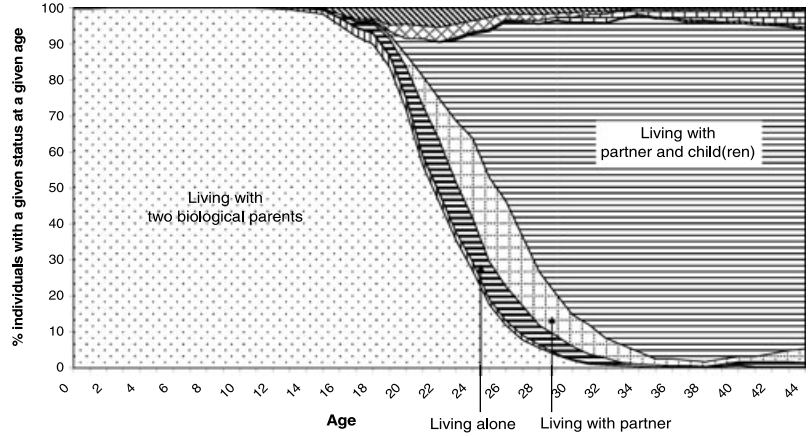
The first bidimensional type of trajectories (Figure 1, 26% of respondents) includes individuals that experience a quick transition to parenthood. After a long stay with their two biological parents, they live a few years alone or with a partner before entering a long and stable

[20] More generally, we do not know the extent to which missing cases are missing at random or not. As occurs frequently with survey data, we may expect slight selection biases toward, for example, age, sex, occupation, or nationality (e.g. see Groves et al. 2004).

[21] We use a unitary substitution cost matrix for both channels; insertion or deletion costs are set to half of one substitution cost.

[22] We retain three criteria among those tested by Milligan and Cooper: (1) pseudo *F,* which represents an approximation of the ratio between the intercluster and intracluster variance of sequences and measures the separation between all clusters at the current level; (2) *Je(2)/Je(1)* (Duda and Hart 1973), which may be transformed into a pseudo *T2,* an index that measures the separation between the two most recently joined clusters; and (3) *R squared*, that expresses the size of the experimental effect. It is reasonable to look for consensus among the three criteria (Nargundkar and Olzer 1998; SAS Institute 2004). In the present study, a given cluster solution was retained for analysis only if at least one of these three criteria supported its validity.

Dimension 1: Family Trajectories

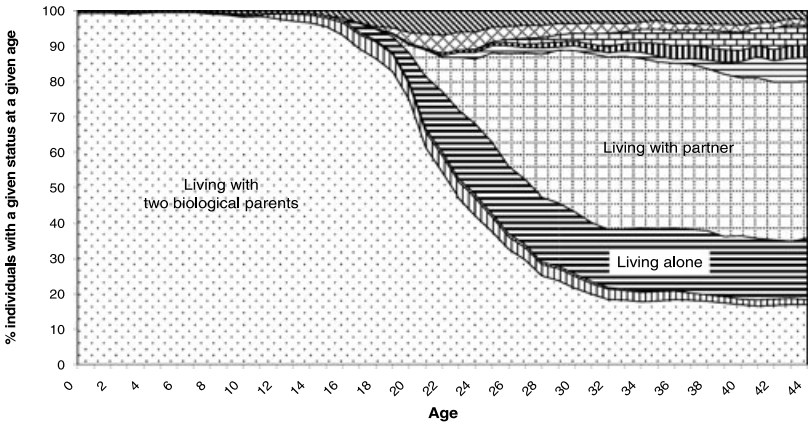


Dimension 2: Occupational Trajectories

FIGURE 1. "Parental and non–full-time employment" bidimensional trajectories (26%).

period of parental life in a nuclear family. The associated occupational trajectories of the same individuals show a short period of full-time work after completing education, followed by a long period out of the job market or working part-time. Women are significantly over-represented in this type, which we label "parental and non-full-time employment" trajectories; indeed, 92% of individuals belonging to this type are female.

Dimension 1: Family Trajectories



Dimension 2: Occupational Trajectories

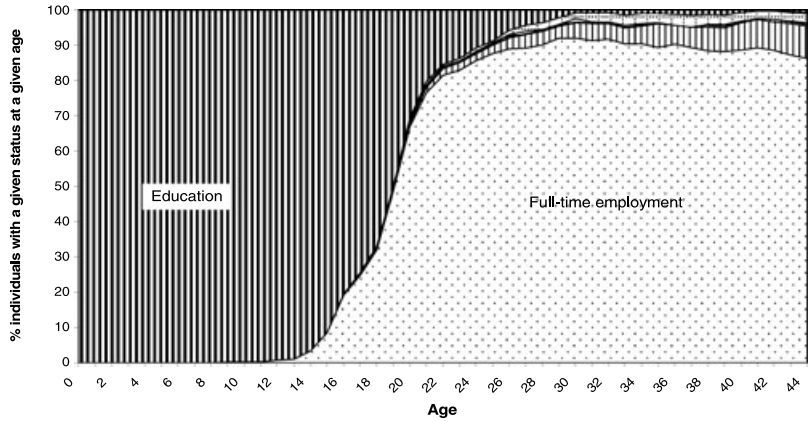
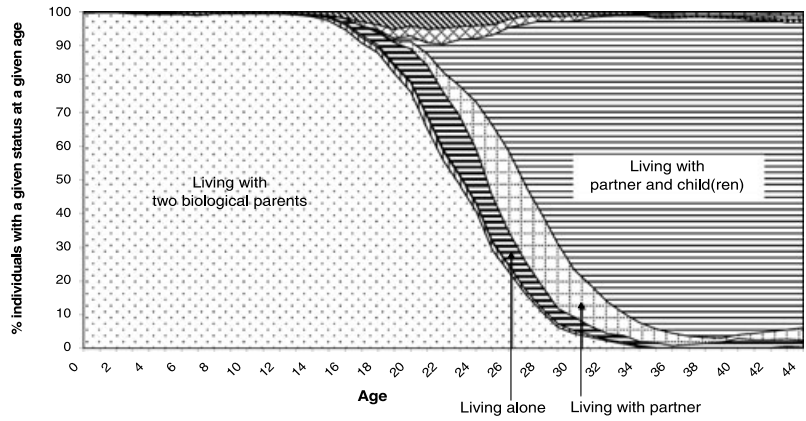


FIGURE 2. "Nonparental and full-time employment" bidimensional trajectories (24%).

The second type (Figure 2, 24% of the sample) brings together people who experienced a long stay in a family of orientation composed of two biological parents, followed by a relatively long period of predominantly single living and/or childless conjugal life. The occupational trajectories of this type consist nearly exclusively of full-time activity. We name this nongendered type "nonparental and full-time employment" trajectories. In contrast to the first type, the proportions of men and women in this type are roughly equal.

Dimension 1: Family Trajectories



Dimension 2: Occupational Trajectories

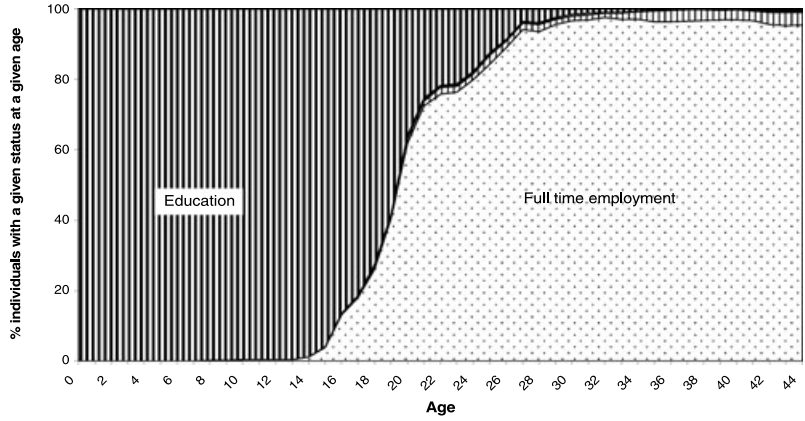


FIGURE 3. "Parental and full-time employment" bidimensional trajectories (30%).

The third type (Figure 3, 30% of the sample) comprises a large majority of men (92%) who follow family trajectories similar to those presented in Figure 1, and whose employment activity is stable and full-time.

Further decomposition of the residual category (not presented here) reveals interesting minority patterns, such as conjugal trajectories associated with non–full-time occupational activities (7%, women over-represented), or parental trajectories combined with long-term full-time

paid work of individuals who experienced their own parents' separation during childhood (7%, no gender bias).

6. VALIDATION

In the following sections, which use the data from the Swiss Household Panel, we test the extent to which MCSA produces more consistent results than regular OMA according to three criteria: (1) parsimony (reduces complexity), (2) reliability (takes advantage from channels interdependence), and (3) robustness (resists noise and distortion).

6.1. *Reduction of Complexity*

Based on two distinct sequences of statuses for each individual in our data sets, three distance matrices are produced. Two of them correspond to one-dimensional analyses performed separately on family and occupational trajectories, whereas the third stems from MCSA applied simultaneously to both trajectories and corresponds to the empirical example presented above.[23] We then run a cluster analysis on these matrices, using Ward's hierarchical method (Wards 1963). The number of clusters actually present in the data is estimated using the stopping rules presented above. For both one-dimensional types of trajectory (family and occupational), the presence of three or five clusters is supported in the data. The same procedure suggests the presence of four clusters in the distance matrix resulting from MCSA. If we cross-combine solutions stemming from the one-dimensional sequence analysis to build *ex post* multidimensional trajectories, we find typologies ranging from nine to 25 types each,[24] whereas an MCSA performed on the same data drastically reduces complexity, as the stopping rules indicate the presence

[23] For this exploratory analysis, to focus on the specific features of MCSA, we use two unitary matrices of substitution, and the cost of insertion/deletion is set to the half of that of substitution.

[24] Cross-combinations of three or five types of family trajectories with three or five types of occupational trajectories form, respectively, nine, 15, 15, and 25 distinct types of family-and-occupational trajectories.

in the data of only four types of bidimensional trajectories.[25] In the following, we will not consider the respective semantic value of these typologies but will focus instead on the extent to which this reduction is associated with a loss of information.

To measure the ability of multichannel analysis to both reduce complexity and preserve information, we cross-tabulate the four clusters' multidimensional typology stemming from MCSA with the corresponding cross-combinations of one-dimensional OMA described above. The Goodman-Kruskal statistic is a measure of "proportionate reduction in error" (PRE), which reflects the percentage by which knowledge of the independent variable reduces errors in predicting the dependent variable (Goodman and Kruskal 1979; Siegel and Castellan 1988; Olzak and Ritschard 1995; Confais, Grelet, and Le Guen 2005). This statistic varies between 0 (absolute independence) and 1 (perfect association). When lambda (R|C)[26] has a value of 1, it means that each row of the table has only one cell different from zero. To efficiently reduce the complexity of a contingency table, we should capture the maximum information available in the rows, with minimum overlapping from one row to another in the same column, as schematically presented in Figure 4.

This is exactly what we get from cross-tabulating MCSA with the combined OMA—that is, many cells with no cases or very few cases and many cells with high-column percentages and no cells in the same column with comparably high scores.[27] Table 1 shows the degree of association (lambda and contingency coefficients) between family and occupational types of trajectories computed either with MCSA (four clusters) or cross-combined monochannel solutions (three or five clusters, respectively).

The contingency coefficients in Table 1 show a strong association between multichannel and cross-combined monochannel

[25] The stopping rules reveal also a seven-types solution for the MCSA. Its association with the cross-combined monochannels is very similar to the four clusters solution.

[26] This is called asymmetric lambda, which predicts the rows distribution (*R*) under the condition that one knows the columns distribution (*C*).

[27] Due to the size of the contingency tables used in the tests, we decided to provide a schematic example of the situation (Figure 4) and to summarize the results by only indicating the value of the lambdas and the contingency coefficients (Table 1).

Cross-Combined Monochannel Typologies	F_1*O_1	F_1*O_2	F_1*O_3	F_2*O_1	F_2*O_2	F_2*O_3	F_3*O_1	F_3*O_2	F_3*O_3
MCSA Typology									
$MCSA_1$	*X*		*X*						
$MCSA_2$		*X*		(*Y*)					
$MCSA_3$				(*Y*)		*X*			*X*
$MCSA_4$					*X*		*X*	*X*	

X= most cases of a column are concentrated in a single cell;
Y= important proportion of cases are distributed on more than one cell of the same column.

FIGURE 4. Schematic representation of the association between cross-combination of family (F) and occupational (O) monochannel typologies containing three types each (respectively F_1, F_2, F_3, O_1, O_2, O_3) and MCSA typology containing four types based on the same data.

solutions. The asymmetric lambda R|C is systematically higher than the asymmetric lambda C|R, indicating that knowing the distribution of combined monochannel solutions allows for better predictions of the multichannel solution distribution. Put another way, MCSA efficiently reduces the complexity of the data while conserving most of the relevant information. More than 80% of the MCSA solution may be predicted by the cross-combined one-dimensional OMA distributions, whereas the reduction in complexity (i.e., the difference between the number of cells in the cross-combined and multichannel solutions, divided by the number of cells in the cross-combined solution) is, respectively, 56%, 73%, and 84%. The asymptotic standard error (ASE) values are much lower than the lambda values. This means here that the 95% confidence interval limits of the lambdas do not contain zero (data not shown), suggesting that these results may be considered statistically significant (SAS technical support, private communication, 2006).

6.2. *Interdependence*

Starting from the results presented in Table 1, we now turn to the extent by which statistical association between individual trajectories unfolding in distinct social spheres influences the quality and reliability of the MCSA features described above. We therefore first cross-tabulate the categorical variables corresponding to the one-dimensional

TABLE 1
Association Between Categorical Variables (Asymmetric Lambda) Corresponding to Types of Trajectories Stemming from Either MCSA or Cross-Combined One-Dimensional OMA

Cross-Combined One-dimensional OMA		MCSA (4 clusters)	
		Value	ASE
Combination 1: 9 clusters	Lambda C|R	0.4641	0.0132
Trajectories:			
Family (3 clusters) * Occupational (3 clusters)	Lambda R|C	0.7975	0.0118
Dimension of contingency table 1: $9 * 4 = 36$	Contingency coefficient	0.8197	
Combination 2: 15 clusters	Lambda C|R	0.3436	0.0128
Trajectories:			
Family (5 clusters) * Occupational (3 clusters)	Lambda R|C	0.8237	0.0113
Dimension of contingency table 2: $15 * 4 = 60$	Contingency coefficient	0.8237	
Combination 3: 15 clusters	Lambda C|R	0.3772	0.0128
Trajectories:			
Family (3 clusters) * Occupational (5 clusters)	Lambda R|C	0.8006	0.0117
Dimension of contingency table 3: $15 * 4 = 60$	Contingency coefficient	0.8238	
Combination 4: 25 clusters	Lambda C|R	0.2659	0.0124
Trajectories:			
Family (5 clusters) * Occupational (5 clusters)	Lambda R|C	0.8463	0.0110
Dimension of contingency table 4: $25 * 4 = 100$	Contingency coefficient	0.8320	

ASE = Asymptotic standard error.

typologies of family trajectories with those of occupational trajectories (Table 2).[28]

Table 2 shows that family and occupational types of trajectories have strong statistical association. The value of the likelihood ratio chi-square is larger when the number of clusters is greater; the

[28] According to our stopping rules, we consider for both trajectories a three- and a five-type typology.

TABLE 2
Association Between Categorical Variables (Likelihood Ratio Chi-square) Corresponding to Types of Family and Occupational Trajectories Stemming from One-dimensional OMA

Cross-Tabulated Types of Trajectories	df	LR χ^2	p Value
Family (3 types) ∗ Occupational (3 types)	4	27.5998	<.0001
Family (3 types) ∗ Occupational (5 types)	8	32.2821	<.0001
Family (5 types) ∗ Occupational (3 types)	8	28.1240	0.0005
Family (5 types) ∗ Occupational (5 types)	16	35.4739	0.0034

Family (3 types) = three types of family trajectories;
Occupational (5 types) = three types of occupational trajectories;
LR χ^2 = likelihood ratio chi-square; df = degree of freedom. N = 1847.

significance level stays under the threshold of 0.01 but decreases slightly as the number of types of trajectories increases. From this result we hypothesize that the use of MCSA provides better results when the types of one-dimensional trajectories are statistically associated. Two life spheres are considered interdependent when the types stemming from OMA performed independently on each of the corresponding trajectories are associated.[29] As mentioned earlier, it is the common information implied by interdependence that allows MCSA to reduce the complexity of multidimensional typologies by locally "deducing" a channel's missing or hidden information. Therefore, in order to test this hypothesis, we focus on other multiple social participations over time—namely, family and education-to-work trajectories. To differentiate education-to-work from occupational trajectories, we limit the period of observation from birth to age 25.[30] One-dimensional OMA performed on these trajectories along with usual stopping rules indicate a two- or five-cluster solution for the first channel (family trajectories), a three- or five-cluster solution for the second one (education-to-work trajectories), and a four-cluster solution for the typology stemming from

[29] Association is measured using the likelihood ratio of chi-square and asymmetric lambda.

[30] To support the comparison with the results presented in Table 2, we measure the association between family and occupational trajectories over a 25-year period and still find similarly high degrees of association. We conclude that the absence of association between family and education-to-work trajectories is therefore not due primarily to the length of the trajectories.

TABLE 3
Association Between Categorical Variables (Likelihood Ratio Chi-square and Asymmetric Lambda) Corresponding to Types of Family and Education-to-Work Trajectories from Either MCSA or Cross-Combined One-Dimensional OMA

Cross-Tabulated Types of Trajectories	Likelihood Ratio Chi-square	*p* Value	Lambda R|C
Family (2 types) ∗ Educ-work(3 types)	2.7324	0.2551	0.0000
Family (2 types) ∗ Educ-work (5 types)	4.2019	0.3794	0.0000
Family (5 types) ∗ Educ-work (3 types)	5.8219	0.6672	0.0000
Family (5 types) ∗ Educ-work (5 types)	8.1966	0.9428	0.0000
MCSA (4 types) ∗ [Family (2) ∗ Educ-work (3)]	833.3803	<.0001	0.3257
MCSA (4 types) ∗ [Family (2) ∗ Educ-work (5)]	836.3845	<.0001	0.3257
MCSA (4 types) ∗ [Family (5) ∗ Educ-work (3)]	1669.0510	<.0001	0.4397
MCSA (4 types) ∗ [Family (5) ∗ Educ-work (5)]	1675.4887	<.0001	0.4397

Family = family trajectories; Educ-work = trajectories of the transition between education and work; MCSA = multichannel sequence analysis of these trajectories. The number of types considered is indicated in parentheses.

MCSA.[31] Cross-tabulations of the categorical variables based on each one-dimensional typology are also created (Table 3).

The results from Table 3 show that the categorical variables representing these one-dimensional types of trajectories are not statistically associated with one another, whereas the MCSA based on family and education-to-work sequences is logically and significantly correlated with the cross-combination of these types. Lambda values in this case, however, are much lower, in comparison to the results stemming from the significantly correlated one-dimensional trajectories described above. This means that the percentage reduction in error in predicting the dependent variable given by MCSA in this case is two to four times lower than it is when the lambda values are obtained with higher correlated trajectories. These results confirm to a certain extent that MCSA is more efficient at reducing data complexity when the considered

[31] The stopping rules also suggest a six-cluster solution for the MCSA. Its association with the cross-combined monochannels is quite similar to the four-cluster solution.

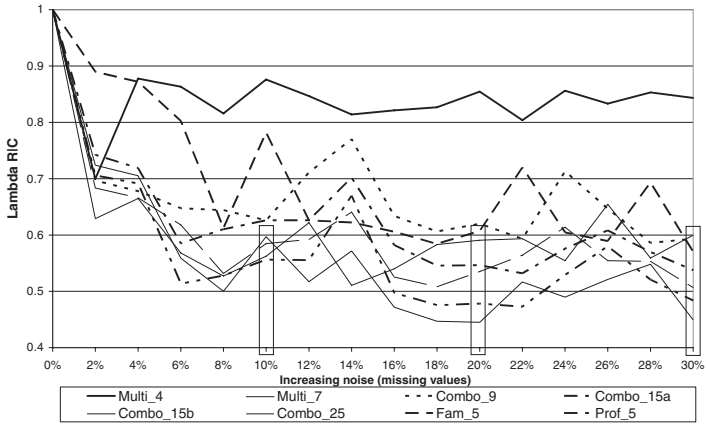
trajectories are interdependent—that is, when they share a certain amount of information.

6.3. *Resistance to Noise*

The third test comparing MCSA and unidimensional OMA concerns the ability of the two approaches to "resist" noise in the data. In other words, we aim at testing the extent to which these procedures are able to identify the same structure in the data when characters in sequences are progressively and randomly replaced by characters that do not belong to the alphabet building the original sequences. From our original data set of family and occupational life trajectories which contain only valid values, we generate 15 alternative data sets for each type of trajectory. Each of these data sets contains a progressively greater proportion of a randomly assigned unknown status compared to the original data set (from 2% to 30%, in increments of 2%).[32] The unknown status is associated with the same unitary substitution cost as the other statuses. The size of the sequences remains the same after the noising process. We then run cluster analyses on each of the distance matrices produced by MCSA and OMA for these data sets and then cross-tabulate the typologies stemming from the original data set with those obtained using the increasingly noisy versions of that same data set. For a given type of trajectory, the number of clusters is held constant and corresponds to the types presented above (cf. Section 4.1.). The degree of association between typologies (lambda coefficient) is computed for each solution and plotted in Figure 5. It compares the ability of MCSA and cross-combined OMA to identify the original data structure from its degraded signal.

Figure 5 illustrates that the four-clusters multichannel typology resists noise much better than do the other typologies. The lambda values for the former remain stable at approximately 0.85, which indicates a rather strong association with the original solution. For one-dimensional types of trajectories, the lambda values decline rapidly and show greater variation than the four-clusters multichannel solution.

[32] The "noising" of the data is a random procedure that is made by SALTT on each individual sequence.



Multi = multichannel analysis; Combo = cross combination of monochannel analyses; Fam = types of family trajectories as categorical variables; Prof = types of professionnal trajectories as categorical variables; _n= number of clusters retained.

FIGURE 5. Value of asymmetric lambda, by increasing amount of missing values on eight types of trajectories.

As the noising of the data occurs before the clustering procedure, each noise level considered on the abscissa of Figure 5 will produce a specific cluster solution. This may explain the substantial variation from one noise level to another that is visible in Figure 5, and the peak lambda value in Figure 6 for the three-cluster solution. In the latter case, the noisy sequences lead to a splitting of the two-cluster solutions, which is not the case when clustering the original sequences that serve as references for both graphs.

To focus specifically on the behavior of MCSA regarding noised data, we compute the values of the asymmetric lambda for the two to 25 clusters solutions of the original multichannel trajectories for three levels of noise in the same data (10%, 20%, and 30%) and plot them in Figure 6.

Figure 6 shows that noise resistance is weakened by the increasing number of clusters and by the level of noise in the sequences. It appears, however, that more noise is systematically associated with a weaker lambda for any given cluster solution. We must also address the extent to which the resistance to noise of a given cluster solution suggests the reliability of that solution. For instance, to what extent can we use such a result to select one cluster solution over another?

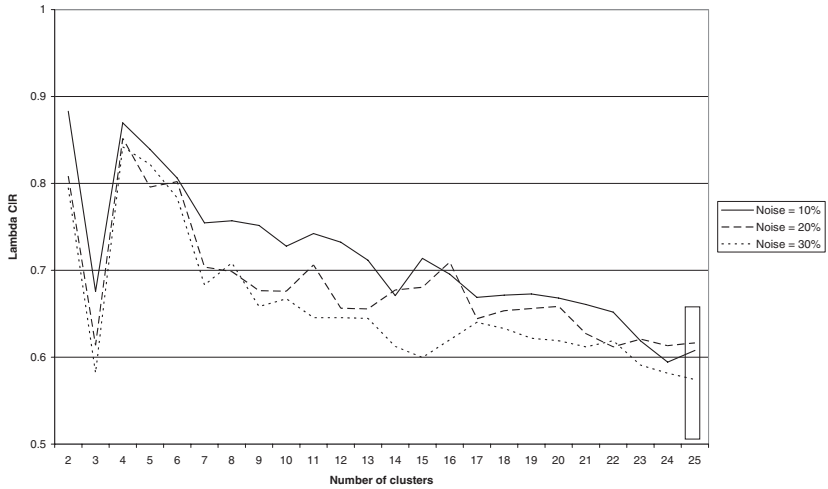


FIGURE 6. Asymmetric lambda value between a given multichannel cluster solution and its corresponding noised solutions (10%, 20%, and 30% missing)

The best solution is also the one that is more resistant to internal variations, which suggests a more stable and informative data structure. Comparing the 25-cluster solutions for the typologies based on MCSA (Figure 6) and one-dimensional OMA (Figure 5) shows that the noisy multichannel solution predicts the original solution better than does the combination of one-dimensional OMA, although this difference is small at a noise level of 10%.

6.4. *Minimizing the Distortion of Alignments*

Considering two distinct dimensions of the individual life course, we use the length variation resulting from pairwise alignment as an indicator of distortion. Minimizing this variation is of special interest because each position in a sequence represents a year of life, which corresponds to a specific age. Given the fact that some statuses and some transitions are more common at certain ages than at others, alignments with greater length variation bias the actual relations between age and social statuses. For instance, Figure 8 exemplifies how MCSA contributes to limiting distortions, since the optimal alignment of Channel d results in a length of six; whereas when both channels are aligned simultaneously

```
Position:       0123456789

seq1 aligned:   A-BBBBA-CA      (seq1 original:     ABBBBACA)

seq2 aligned:   AABBB-AB-A      (seq2 original:     AABBBABA)
```

FIGURE 7. Measuring the distortion resulting from a pairwise alignment.

(MCSA), the length of the final alignment is five for both dimensions. In this way, MCSA keeps the chronological order of both trajectories as close to the original as possible without using *indels*, which allows for a better structural conservation of sequences than do systematic substitutions. The distortion due to an alignment is defined as the sum of the products of the number of character(s) shifted, multiplied by the size of the shift (time units), and divided by the total, number of aligned character pairs. This is a standardized measure that may be used to align sequences of different lengths, although the ones used here are of equal lengths. Figure 7 gives an example of distortion measurement resulting from a pairwise alignment. Considering the aligned sequence seq1, the three characters 'B' from the original sequence seq1 (positions 2–4) and character 'C' (position 8) are shifted by one position (time unit) to the right. In this case, there are six aligned character pairs in the alignment. The value of the distortion resulting from the pairwise alignment of seq1 and seq2 is 0.66 [$((3 * 1) + (1 * 1)) / 6 = 4/6 = 0.666$].

Our aim is to test whether MCSA provides less distorted alignments than one-dimensional OMA does. Therefore, using SHP data, we compare the age distortion stemming from two separate monochannel alignment procedures for each individual—one for family and the other for occupational trajectories—with age distortion computed using MCSA based on the same trajectories.

From three data sets containing 1,847 family, occupational, and multidimensional trajectories, we obtain 1,704,781 possible alignments for each of them.[33] A distortion score is computed for each alignment. To compare the alignments produced by cross-combined one-dimensional OMA and MCSA, we subtract for each individual the largest distortion score stemming from either one-dimensional

[33] Number of alignments = N * (N-1)/2 = 1847 * 1846/2 = 1'704'781

TABLE 4
Difference in Distortion Between Multichannel (Reference) and Family, Occupational as well as Max(Family, Occupational) Pairwise Alignments

Sequences Aligned	Family	Occupational	Max(fam., occup.)
Multichannel is better (−)	28%	26%	48%
No difference (0)	44%	50%	46%
Multichannel is worse (+)	28%	24%	6%
Total	100%	100%	100%
	N = 1'704'781	N = 1'704'781	N = 1'704'781

Max(fam., occup.) = larger distortion resulting from the alignment of a pair of either family or occupational trajectories for the same individual.

alignments to that produced by MCSA.[34] A resulting negative value indicates that the distortion stemming from MCSA is smaller than that resulting from one-dimensional alignments. Table 4 shows the distortion differences between MCSA (reference) and one-dimensional OMA based on family and occupational trajectories. For each individual, we also consider the larger distortion produced by either alignment.

Table 4 shows that in the majority of cases, MCSA provides less or equally distorted alignments than does regular one-dimensional OMA. Since MCSA represents a combination of two alignments, it would be understandable if MCSA produces more distorted alignments than one-dimensional OMA. Actually, MCSA clearly produces better results than the one-dimensional OMA performed on the channel associated with the most distorted alignments. MCSA generates less distorted alignments in approximately 50% of cases. Distortion from MCSA is greater than that from one-dimensional OMA in only 6% of alignments. In other words, MCSA's distortions are almost always smaller than, or equal to, those of two one-dimensional OMA applied separately. By reducing sequences' distortion in the alignment process, MCSA offers a better conservation of structural and temporal patterns (Lesnard and Saint Pol 2004).

Table 5 presents the paired *t*-test values for these comparisons and shows that MCSA significantly reduces the structural and

[34] Comparison of individual distortion scores equals distortion score measured on MCSA (the largest distortion score measured on the alignment of either family or occupational trajectories).

TABLE 5
Paired *t*-Test on Distortion's Value Resulting from Either MCSA or One-dimensional OMA

	Mean	Standard Deviation	*t*-Test Value	*p*
Family—MCSA	0.2911	2.4628	154.33	<.0001
Occupational—MCSA	0.1659	1.7331	124.98	<.0001
Max (fam., occup.)—MCSA	1.0948	2.4066	594.00	<.0001

N = 1'704'781.

temporal distortion of aligned sequences ($p < 0.0001$). This reduction is greatest when comparing MCSA to regular OMA performed on sequences that produce the greatest distortions. Despite the large number of cases, which improves performance in significance tests, the relatively high standard deviations indicate substantial variability in the data. Moreover, non-paired *t*-tests on the same data (not shown in Table 5) indicate that MCSA produces significantly smaller standard deviations than does one-dimensional OMA ($p < 0.0001$).

7. FURTHER VALIDATION ON RANDOM DATA

Having already shown the favorable properties of MCSA compared to regular OMA performed on existing social science data, this section aims at assessing the extent to which MCSA also provides qualitatively similar results when used on random data. To compare various multidimensional approaches using OMA, we use two simulated data sets ($N = 2001$ pairs of sequences), each corresponding to a specific channel. In this simulated data, the alphabet and length of sequences are kept constant. In each data set, the first sequence has a length of five characters and the second a length of four. The alphabets of the first and the second channel contain three and four characters, respectively (cf. Figure 8).[35] We first use the simulated data to evaluate whether different approaches to multidimensional sequence analysis produce similar results. We compare four ways of computing multidimensionality: ex-post sum of the distance matrices produced by two independent OMA, MCSA,

[35] To create the sequences, we use the Perl's function *rand( )*, which produces uniformly distributed pseudo-random numbers (Wall et al. 2000).

and two ex-ante recoding of both channels into one unique channel (called "extended 1 and 2" in Figure 8). For aligning pairs of sequences (nested or separately), we use unitary substitution costs matrices.[36] For the extended alphabet, we consider the fact that, when comparing two characters of the recoded sequences, the cost may be one unit if both recoded characters have a character in common. The cost is set to two units if they have no characters in common.[37] The cost is zero when both pairs of characters are identical for two given individuals. In the first case, the value of *indel* is set to the average off-diagonal value (AOD) of the substitution cost matrix, while in the second case, *indel* is set to half of this value[38] (see Figure 8). We compare these different approaches using the degree of similarity between all pairs of sequences, which is given by either the raw score of the alignment or the PID.

Using the simulated data sets and the cost schemes described above, we compute linear coefficient correlations among alignment scores stemming from different ways of assessing the distance between multidimensional sequences, as shown in Table 6, where the distances produced using either extended alphabets, ex-post sum of monochannel distances, or MCSA are strongly associated, although not identical. According to the two latter methods, the use of either percent identity or raw score produces the same correlations with the other measures of multidimensional distances. Since it otherwise brings the most differentiated correlations, we retain PID to estimate the distance between sequences (May 2004). The five measures of multidimensional distances between individual trajectories are all based on some linear function of one-dimensional OMA distances. They differ essentially in the timing of the contribution of each channel—that is, either before, during, or after the alignment process. As one can read from Table 6, results produced by MCSA appear here as a representative denominator to the other measures. This means that they are at the same time as variable

[36] This means that substituting any character with another one has a cost of 1, whereas substituting a character with itself has a cost of 0.

[37] For instance, if "recoded *f*" stands for "*m*" and "*z*" at the same position in channels 1 and 2, "recoded *j*" stands for "*m*" and "*t*," and recoded "*g*" for "*n*" and "*z*," the cost of substituting "*f*" and "*j*" is two-fold lower than the cost of substituting "*j*" and "*g*."

[38] At this point, we did not consider the differentiation between gap opening penalty (GOP) and gap extension penalty (GEP; Thompson et al. 1994), or between internal and external gaps.

```
Channel c : Alignment: 40 Percent Identity (PID) / Raw score = 2.5 / Distortion = 1.00

Sequence Ac             mmllm
Sequence Bc             -nlln

Channel d : Alignment: 20 Percent Identity (PID) / Raw score = 3.5 / Distortion = 1.00

Sequence Ad             -ztttt
Sequence Bd             yz--zz

MCSA : Alignment: 20 Percent Identity (PID) / Raw score = 7 / Distortion = 1.50

Channel c
Sequence Ac             mmllm
Sequence Bc             -nlln

Channel d
Sequence Ad             ztttt
Sequence Bd             -yzzz

Extended 1 : Alignment: 0 Percent Identity (PID) / Raw score = 7.55 / Distortion = 1.00

Sequence Acd            fjiij
Sequence Bcd            -ceeg

Extended 2 : Alignment: 0 Percent Identity (PID) / Raw score = 5.85 / Distortion = 1.00

Sequence Acd            fj-iij-
Sequence Bcd            --cee-g
```

A^c = sequence corresponding to channel c for individual A
B^{cd} = sequence corresponding to the extended alphabet built on the combination of Channels c and d for individual B

FIGURE 8. Comparing MCSA to summing two distance matrices or using extended alphabet (random data sets).

as the five others but less sensitive to the computation options, which is a first indication toward its robustness.

8. DISCUSSION

This paper explores two key points regarding the methodological potential of multichannel sequence analysis (MCSA). First, MCSA offers an overall advantage over conventional OMA, since it allows for the simultaneous analysis of multiple social trajectories without prior recoding of the data. MCSA produces an extended alphabet that corresponds to the combination of two or more alphabets defining different types of sequences used in the analysis. The main advantage of MCSA over other extended alphabet methods (Dijkstra and Taris 1995; Stovel et al. 1996) is that it avoids defining, coding, and weighting all combinations prior to the analysis, and it therefore allows for the use of weighting strategies specific to each dimension (family, occupation, and so forth) considered separately, such as the data-based training procedure

TABLE 6
Similarity Matrix (Linear Coefficient Correlation) Between Six Measures of Two-dimensional Pairwise Distances Using Random Data Sets (N = 2001)

	PID_MCSA	Sum_PID_OMA	Score_ext1	Score_ext2	PID_ext1	PID_ext2
PID_MCSA	1.000					
Sum_PID_OMA	0.803	1.000				
Score_ext1	0.898	0.773	1.000			
Score_ext2	0.960	0.748	0.826	1.000		
PID_ext1	0.754	0.517	0.704	0.798	1.000	
PID_ext2	0.739	0.540	0.577	0.898	0.736	1.000

(Gauthier et al. 2009). Keeping the specific codification of each trajectory distinct allows for better interpretation of MCSA's typologies, since nested trajectories are represented as parallel processes associated with substitution matrices that are in themselves informative.

Applied to social science data from the Swiss household panel, the illustrative application of MCSA shows that it produces more convincing results than does independent OMA. Moreover, it provides semantically and graphically straightforward patterns of the ways multidimensional social participations unfold over time, a feature that represents one of the central developing fields of sequence analysis.

Second, our results on the same data show that MCSA performs best when the dimensions under study are interdependent. Comparing the analysis of correlated versus uncorrelated monochannels, we find that MCSA leads to a greater reduction of complexity when the trajectories are statistically associated and when the number of clusters is relatively small. This outcome provides a first indication toward MCSA's range of applications: It is precisely when nested trajectories are interdependent—that is, when they share information—that MCSA is required. Additionally, the results show the ability of MCSA to simplify the outputs obtained from regular OMA. This simplification is achieved by dramatically reducing the number of categories involved, while retaining a high proportion of the original information. This means that considering interdependence increases the complexity of nested trajectories under study, but at the same time it reduces the number of relevant combinations, compared to cross-combining results from independent one-dimensional OMA.

We also test the resistance of MCSA to noisy data. It appears that MCSA is less sensitive to increasing noise in the data than are combinations of regular one-dimensional OMA. MCSA uses the interdependence that exists between nested trajectories as an additional source of information to identify relevant multidimensional patterns, even when some of the data are missing.

This paper also examines the issue of sequence distortion produced by the use of insertions and deletions that change sequence length. Carrying out an alignment modifies the correspondence between actual age and the position in the sequence prior to alignment. In measuring the distortion resulting from MCSA and conventional OMA, we find that MCSA was nearly always superior or equivalent to conventional OMA in minimizing this distortion; that is, MCSA performed better

by keeping the length of aligned sequences as close as possible to that of the original sequences. We demonstrated the ability of MCSA to produce less distorted alignments and take the timing of episodes more accurately into account than combinations of conventional OMA. This feature is particularly important when considering not only relative duration but also dimensions such as social age, which take into account the fact that some social statuses or transitions are more common at certain points in life than at others. In other words, MCSA produces alignments that optimize the relationship between age and social statuses over time.

Finally, using random data, we demonstrate that distances produced by MCSA differ from those produced by either summing pairwise distances in independent OMA or by recoding the data prior to analysis. Furthermore, MCSA yields the strongest correlations with the results of alternative measures of multidimensional distances. It therefore appears to be the most representative technique among those that we examined.

Given the number of dimensions that may play a role in the variability of results obtained through either method, this paper provides initial guidance on the potential advantages of MCSA. Further developments of the method, along with in-depth testing, are needed to continue improving MCSA. Our main expectation regarding MCSA is to significantly reduce the "signal differences" between channels when channels are related. It is precisely the correlation between channels that allows the alignment procedure to benefit from the information contained in one sequence but not another, and ultimately to produce multichannel alignments that reduce the complexity, distortion, and loss of signal due to such noise or to missing values. This informational asymmetry between channels may vary over time (i.e., it is position-specific), and it may depend on specific stages of the life course or on specific features of social age. In some stages, for instance, occupational status is poorly or not at all informative (e.g., during school years). In such cases, information from the other channel(s) should be given preference. In other words, if one channel is more informative than another at a given point in the sequence, we should rely more heavily on the more informative channel to compute the multichannel alignment. In the same way, if there are missing values on one channel, we should "let the other channel talk" by giving it more weight. Future developments should implement heuristic procedures to systematize methods for dealing with such information asymmetries between channels. In

this paper, we have set the combination of substitution costs at one point to the average value of the two substitution costs involved at that point. An alternative would be to follow some weighting scheme based in theory (e.g., costs set to the highest or the lowest value), or to rely on empirically determined costs of substitution.

Overall, MCSA presents two main advantages over one-dimensional OMA. It allows both for the discovery of regularities within multidimensional trajectories and for the reduction of the effects of noise, whether due to missing data, poorly recorded information, or heterogeneous information content.

APPENDIX

The computations presented in this paper are encapsulated in the program SALTT (Search Algorithm for Life Trajectories and Transitions), an open source freeware written in C (Notredame, Bucher, Gauthier, and Widmer 2005). It can be compiled and installed on any UNIX-like platform including Linux, Cygwin, and MacOSX. The package and its documentation can be downloaded from: http://www.tcoffee.org/saltt/.

REFERENCES

Abbott, A. 1992. "From Causes to Events. Notes on Narrative Positivism." *Sociological Methods and Research* 20 (4):428–55.

———. 2001. *Time Matters: On Theory and Method*. Chicago, IL: University of Chicago Press.

Abbott, A., and A. Hrycak. 1990. "Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers." *American Journal of Sociology* 96 (1):144–85.

Abbott, A., and A. Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology." *Sociological Methods and Research* 29 (1):3–33.

Aisenbrey, S., and A. E. Fasang. 2010. "New Life for Old Ideas: The 'Second Wave' of Sequence Analysis Bringing the 'Course' Back Into the Life Course." *Sociological Methods and Research*, 38 (3):420–62.

Blair-Loy, M. 1999. "Career Patterns of Executive Women in Finance: An Optimal Matching Analysis." *American Journal of Sociology* 104 (5):1346–97.

Blossfeld, H.-P., and G. Rohwer. 1995. *Techniques of Event History Modeling*. Mahwah, NJ: Lawrence Erlbaum.

Butts, C., and J. Pixley. 2004. "A Structural Approach to the Representation of Life History Data." *Journal of Mathematical Sociology* 28 (2):81–124.

Clausen, J. A. 1986. *The Life Course: A Sociological Perspective*. Toronto: Prentice-Hall.

Claverie, J.-M., and C. Notredame. 2003. *Bioinformatics for Dummies*. New York: Wiley.

Confais, J., Y. Grelet, and M. Le Guen. 2005. "La Procédure FREQ de SAS. Tests d'independance et mesures d'association dans un tableau de contingence." *La Revue Modulad* 33:188–224.

Dijkstra, W., and T. Taris. 1995. "Measuring the Agreement Between Sequences." *Sociological Methods and Research* 24 (2):214–31.

Doolitle, R. F. 1981. "Similar Amino Acid Sequences: Chance or Common Ancestry." *Science* 214 (4517):149–59.

Duda, R. O., and P. E. Hart. 1973. *Pattern Classification and Scene Analysis*. New York: Wiley.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 2002. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, England: Cambridge University Press.

Elder, G. H., ed. 1985. *Life Course Dynamics: Trajectories and Transitions, 1968–1980*. Ithaca, NY: Cornell University Press.

Elder, G. H., M. Kirkpatrick Johnson, and R. Crosnoe. 2003. "The Emergence and Development of Life Course Theory." Pp. 3–19 in *Handbook of the Life Course*, edited by J. T. Mortimer and M. J. Shanahan. New York: Kluwer.

Elzinga, C. H. 2003. "Sequence Similarity: A Non-Aligning Technique." *Sociological Methods and Research* 31(4):3–29.

Esser, H. 1996. "What Is Wrong with 'Variable Sociology'?" *European Sociological Review* 12 (2):159–66.

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller. 2008. *Mining Sequence Data in R with the TraMineR Package: A User's Guide*. University of Geneva. Retrieved January 21, 2010. (http://mephisto.unige.ch/traminer).

Gauthier J.-A., E. D. Widmer, P. Bucher, and C. Notredame 2009. "How Much Does It Cost? Optimization of Costs in Sequence Analysis of Social Science Data." *Sociological Methods and Research* 38 (1):197–231.

George, L. K. 1993. "Sociological Perspectives on Life Transitions." *Annual Review of Sociology* 19:353–73.

Giele, J. Z., and G. H. Elder Jr., eds. 1998. *Methods of Life Course Research: Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage.

Goodman, L. A., and W. H. Kruskal. 1979. *Measures of Association for Cross Classification*. New York: Springer-Verlag.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2004. *Survey Methodology*. Wiley Series in Survey Methodology. New York: Wiley.

Han, S.-K., and P. Moen. 1999. "Clocking Out: Temporal Patterning of Retirement." *American Journal of Sociology* 105 (1):191–236.

Heinz, W. R. 2003. "From Work Trajectories to Negotiated Careers." Pp. 185–204 in *Handbook of the Life Course*, edited by J. T. Mortimer and M. J. Shanahan. New York: Kluwer.

Heinz, W. R., and V. W. Marshall, eds. 2003. *Social Dynamics of the Life Course: Transitions, Insitutions, and Interrelations*. New York: Aldine de Gruyter.

Höpflinger, F. C., and A. Debrunner. 1991. *Familenleben und Berufsarbeit*. Zurich, Switzerland: Seismo.

Kohli, M. 1986. "The World We Forgot: A Historical Review of the Life Course." Pp. 271–303 in *Later Life: The Social Psychology of Aging*, edited by V. W. Marshall. London: Sage.

Kruskal, J. 1983. "An Overview of Sequence Comparison." Pp. 1–44 in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. Kruskal. Don Mills, Ontario: Addison-Wesley.

Lesesne, C. A., and C. Kennedy. 2005. "Starting Early: Promoting the Mental Health of Women and Girls Throughout the Life Span." *Journal of Women's Health* 14 (9):754–63.

Lesnard, L., and T. de Saint Pol. 2006. "Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis)." *Bulletin of Sociological Methodology* 90:5–25.

Levine, J. H. 2000. "But What Have You Done for Us Lately?" *Sociological Methods and Research* 29 (1):34–40.

Levy, R. 1991. "Status Passages as Critical Life Course Transition: A Theoretical Sketch." Pp 87–114 in *Theoretical Advances on Life Course Research*, edited by W. R. Heinz. Weinheim, Germany: Deutscher Studien Verlag.

———. 1996. "Toward a Theory of Life Course Institutionalization." Pp. 83–108 in *Society and Biography*, edited by A. Weymann and W. R. Heinz. Weinheim, Germany: Deutscher Studien Verlag.

Levy, R., J. A. Gauthier, E. D. Widmer. 2006. "Entre contraintes institutionnelle et domestique: les parcours de vie masculins et féminins en Suisse." *Canadian Journal of Sociology* 31 (4):461–89.

Macmillan, R., ed. 2005. *The Structure of the Life Course: Standardized? Individualized? Differentiated?*, Vol. 9. Amsterdam: JAI Press.

Macmillan, R., and S. R. Eliason. 2003. "Characterizing the Life Course as Role Configurations and Pathways. A Latent Structure Approach." Pp. 529–54 in *Handbook of the Life Course*, edited by J. T. Mortimer and M. J. Shanahan. New York: Kluwer Academic.

May, A. C. W. 2004. "Percent Sequence Identity: The Need to Be Explicit." *Structure* 12:737–38.

Milligan, G. W., and M. C. Cooper. 1985. "An Examination of Procedures for Determining the Number of Clusters in a Data set." *Psychometrika* 50 (2):159–79.

———. 1987. "Methodology Review: Clustering Methods." *Applied Psychological Measurement* 11 (4):329–54.

Moen, P. 1985. "Continuities and Discontinuities in Women's Labor Force Activity." Pp. 113–55 in *Life Course Dynamics: Trajectories and Transitions, 1968–1980*, edited by G. H. Elder. Ithaca, NY: Cornell University Press.

Mojena, R. 1977. "Hierarchical Grouping Methods and Stopping Rules: An Evaluation." *The Computer Journal* 20:359–63.

Mortimer, J. T., and M. J. Shanahan, eds. 2003. *Handbook of the Life Course*. New York: Kluwer Academic.

Müller, N. S., A. Gabadinho, G. Ritschard, and M. Studer. 2008. "Extracting Knowledge from Life Courses: Clustering and Visualization." Pp. 176–85 in *Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery*. Turin, Italy: Springer-Verlag.

Nargundkar, S., and T. J. Olzer. 1998. "An Application of Cluster Analysis in the Financial Services Industry." Presented at the Sixth annual meeting of the South East SAS Users Group (SESUG), Norfolk, Virginia.

National Centre for Biotechnology Information (NCBI) 2004. *Glossary*. Retrived October 15, 2004 (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html).

Needleman, S. B., and C. D. Wunsch. 1970. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." *Journal of Molecular Biology* 48:443–53.

Notredame, C., P. Bucher, J.-A. Gauthier, and E. Widmer. 2005. *T-Coffee/SALTT: User Guide and Reference Manual*. Retrieved October 15, 2005. (http://www.tcoffee.org/saltt)

Olzak, M., and G. Ritschard. 1995. "The Behaviour of Nominal and Ordinal Partial Association Measures." *The Statistician* 44 (2):195–212.

Piccarreta, R., and O. Lior. 2010. "Exploring Sequences: A Graphical Tool Based on Multi-Dimensional Scaling." *Journal of the Royal Statistical Society*, Series A: *Statistics in Society* 173 (1):165–84.

Pollock, G. 2007. "Holistic Trajectories: A Study of Combined Employment, Housing, and Family Careers Using Multiple Sequence Analysis." *Journal of the Royal Statistical Society*, Series A*: Statistics in Society* 170:167–83.

Raghava, G. P. S., and G. Barton. 2006. "Quantification of the Variation in Percentage Identity for Protein Sequence Alignments." *BMC Bioinformatics* 7 (1):415.

Repetti, R. L., S. E. Taylor, and T. E. Seeman. 2002. "Risky Families: Family Social Environments and the Mental and Physical Health of Offspring." *Psychological Bulletin* 128(2):330–66.

Sankoff, D., and J. Kruskal. 1983. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Don Mills, Ontario: Addison-Wesley.

SAS Institute. 2004. *SAS/STAT® 9.1 User's Guide*. Cary, NC: SAS Institute Inc.

Sheridan, J. T. 1997. *The Effects of the Determinants of Women's Movement into and out of Male-Dominated Occupations on Occupational Sex Segregation*. Madison: Department of Sociology, Center for Demography and Ecology, University of Wisconsin.

Siegel, S., and N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioural Sciences*, 2nd ed. New York: McGraw-Hill.

Spruijt, E., and M. de Goede. 1997. "Transitions in Family Structure and Adolescent Well-Being." *Adolescence* 32(128):897–911.

Stovel, K., M. Savage, and P. Bearman. 1996. "Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890–1970." *American Journal of Sociology* 102(2):358–99.

Thompson, J., D. G. Higgins, and T. Gibson. 1994. "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice." *Nucleic Acids Research* 22:4673–80.

Tillmann, R., and E. Zimermann. 2004. "Introduction: The Swiss Household Panel and the Nature of This Book." Pp. 1–25 in *Vivre en Suisse 1999–2000 [Living in Switzerland 1999–2000]*, edited by R. Tillmann and E. Zimermann. Bern, Switzerland: Peter Lang.

Tufte, E. R. 1997. *Visual Explanation, Images and Quantities, Evidence and Narrative*. Cheshire, CO: Graphic Press.

Wall, L., T. Christiansen, and J. Orwant. 2000. *Programming Perl*, 3rd ed. Sebastopol, CA: O'Reilly.

Ward, J. H. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58(301):236–44.

Wasserman, S., and K. Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, England: Cambridge University Press.

Wetzler, H. P., and R. J. Ursano. 1988. "A Positive Association Between Physical Health Practices and Psychological Well-Being." *The Journal of Nervous and Mental Disease* 176 (5):280–83.

Widmer, E., R. Levy, A. Pollien, R. Hammer, and J.-A. Gauthier. 2003. "Une analyse exploratoire des insertions professionnelles et familiales: Trajectoires de couples résidant en Suisse." *Revue suisse de Sociologie* 29(1):35–67.

Wu, L. L. 2000. "Some Comments on 'Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect'." *Sociological Methods and Research* 29(1):41–64.